Currents in Computational Molecular Biology 2004

Edited by:

Apostol Gramada
Philip E. Bourne

FOREWORD

This book contains the abstracts of the posters presented at the eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB 2004), an ACM SIGACT. The conference was organized by the San Diego Supercomputer Center (SDSC), University of California San Diego (UCSD) and the International Society for Computational Biology (ISCB). There were 292 posters accepted within the following areas of research:

Combinatorial libraries and drug design (A)	Molecular evolution (H)
Computational genomics (B)	Molecular sequence analysis (I)
Computational proteomics (C)	Protein structure (J)
Gene expression (D)	Recognition of genes and regulatory elements (K)
Gene networks (E)	Structural and functional genomics (L)
Genomics (F)	Complex systems and simulation (M)
Microarray design and data analysis (G)	Methods (N)

The conference held two poster sessions. The even numbered posters were presented on March 28-29 and the odd numbered on March 30-31. The poster abstracts were limited to two pages. They are listed in the book according to the section letter and the number assigned within each section. Due to withdrawals subsequent to the initial poster id assignment, occasional gaps appear in the sequence of numbers that form the poster id. Poster abstracts were published without detailed peer review or copy editing. The conference organizers take no responsibility for the originality of the material.

These abstracts can be found on-line at *http://recomb04.sdsc.edu./poster_sched.html*.

The poster editorial team would like to thank first the authors for their effort in submitting the poster abstract in the format that we requested. We also thank a number of people who assisted us in setting and managing the poster database and the whole submission process: Lynn Fink, Kyle Wright, Wayne Townsend-Marino and Dana Jermanis. Finally, we wish to thank Applied Biosciences Inc. and Microsoft Inc. for their support with the organization of the poster session, printing the poster book and providing funds for a “best poster” prize.

Apostol Gramada
Philip E. Bourne

San Diego, March 2004

CONTENTS

Table of Contents

Table of Contents

Table of Contents

A1. Pharmacophore Multiplets in Combinatorial Library Design: A Novel Approach to Generation and Storage

Edmond Abrahamian[1], Peter Fox[1], Inge Thøger Christensen[2], Henning Thøgersen[2], Robert D. Clark[1]

Keywords: pharmacophore multiplets, molecular fingerprints, molecular similarity

1 Introduction.

Pharmacophore triplets and quartets have been used by many groups in recent years, primarily as a tool for molecular diversity analysis [1, 2]. In most cases, slow processing speeds and the very large size of the bitsets generated have forced researchers to compromise in terms of how such multiplets were stored, manipulated and compared, *e.g.*, by using simple unions to represent multiplets for sets of molecules. Here we report using bit*maps* in place of bit*sets* to reduce storage demands and to improve processing speed. Here, a bit*set* is taken to mean a fully enumerated string of zeros and ones, from which a compressed bit*map* is obtained by replacing uniform blocks ("runs") of digits in the bitset with a pair of values identifying the content and length of the block (run-length encoding compression). High resolution multiplets involving four features are enabled by using 64 bit executables to create and manipulate bitmaps, which "connect" to the 32 bit executables used for database access and feature identification *via* an extensible mark-up language (XML) data stream. The encoding system used also supports simple multiplets, quartets in which a privileged substructure is used as an anchor point, and augmented triplets in which the fourth feature represents a hydrogen bond donor or acceptor extension point linked to the complementary acceptor or donor atom in a base triplet. It can readily be extended to larger, more complex multiplets as well.

The system is set up to support hypothesis generation, which entails creation of consensus bitmaps built up from active ligands identified in preliminary screening. Rather than a simple union across bitmaps, a weighting scheme is used that gives greater weight to multiplets expected to be more discriminating. Multi-conformer bitmaps can be obtained from pre-generated conformations or by random perturbation on-the-fly.

2 Methods.

The first step in encoding multiplets is to sweep through the molecules of interest and identify pharmacophoric features specified in the multiplet definition file. These may be point features or centroids of substructures or extension points. The features found are compiled into an ASCII file using the XML format, which is then passed on to a separate executable for generation of the desired bitmaps. The intermediate XML files constitute a "feature stream" which makes it possible to efficiently generate different "flavors" of multiplets in parallel, since feature identification is a relatively slow step in the overall process.

An unambiguous, two-way, one-to-one mapping must exist between every possible pharmacophore multiplet and some position in the corresponding bitset. For triplets, the distances d_i are sorted in descending order and the vertex falling between the longest and shortest edges is assigned to the

[1] Tripos, Inc., 1699 S. Hanley Rd., St.Louis MO 63117 USA. E-mail: edmond@tripos.com
[2] Novo Nordisk A/S, Protein Structure Research, Novo Nordisk Park, 2760 Måløv, Denmark

central position (f_2) in the feature index. The first vertex (f_1) is then set to match the feature connected to f_2 by the longest edge (d_1), whereas the third feature index (f_3) is determined by the type of the feature connected to f_2 by the shortest edge (d_3; see Fig.1). In the event that different edges fall in the same distance bin, the tie is broken lexicographically, where feature priority is taken as the order in which features are specified in the configuration file. Mapping for a quartet is more elaborate and takes chirality into account.

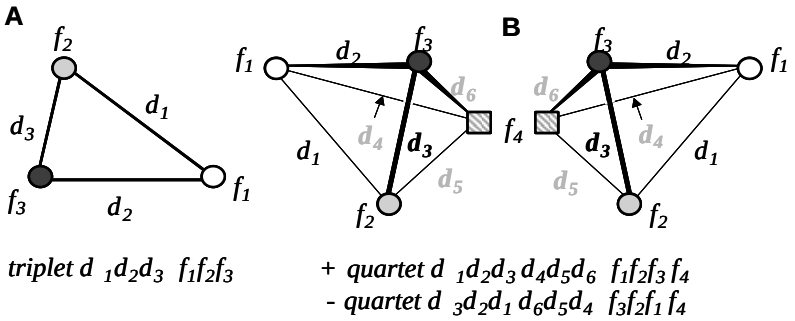


Figure 1: Multiplet encoding scheme (**A**) Triplet , (**B**) Quartet

3 Results

Analysis was carried out on a set of 96 glycogen synthase kinase-3 (GSK3) inhibitors and analogs originally compiled for a quantitative structure-activity relationship (QSAR) analysis [3]. This included 70 active compounds and 26 inactive analogs included in the study for the sake of completeness. An additional 91 were drawn from the NCI anticancer screening database that exhibited a Tanimoto similarity of 0.65 or greater to the compounds in the training set with respect to standard UNITY [4] substructural fingerprints. Twenty-five conformers were produced for each of the 187 compounds, and the triplet bitmaps so obtained were used for hierarchical clustering. The actives all cluster together to the right in the dendrogram, whereas the less active analogs are distributed among the unhighlighted nodes corresponding to decoy compounds from the NCI database. Results from other data sets are available as well [5].

4 References and bibliography.

[1] Mason, J.S. and Beno, B.R., 2000. Library design using BCUT chemistry - space descriptors and multiple four-point pharmacophore fingerprints: simultaneous optimization and structure-based diversity. *J.Mol. Graphics Mod.*, 18:438-451.

[2] Good, A. C. and Mason, J. S. et al. 2001. In Ghose, A. K. and Viswanadhan V. N., eds. *Pharmacophore-Based Approaches to Combinatorial Library Design and Evaluation*, New York, NY: Marcel Dekker. pp 399-428.

[3] Nærum, L.; Nørskov-Lauritsen, L.; Olesen, P.H. 2002. Scaffold Hopping and Optimization towards Libraries of Glycogen Kinase-3 Inhibitors. *Bioorg. Med. Chem. Lett.*, 12:1525-1528.

[4] UNITY is distributed by Tripos, Inc., 1699 S. Hanley Rd., St.Louis, MO 63144.

[5] Abrahamian, E. et. al. 2003. Efficient Generation, Storage, and Manipulation of Fully Flexible Pharmacophore Multiplets and Their Use in 3-D Similarity Searching. *J. Chem. Inf. Comput. Sci*. 43:458-468.

A2. The Docking Mesh Evaluator

Roummel F. Marcia,[1] Susan Lindsey,[2] J. Ben Rosen,[3] Julie C. Mitchell[4,5]

Keywords: protein docking, Poisson-Boltzmann equation, global optimization

1 Introduction

The Docking Mesh Evaluator (DoME) is a software for predicting a bound protein-ligand docking configuration by determining the global minimum of a potential energy function. Our present energy model is based on solvent effects defined implicitly using the Poisson-Boltzmann equation, as well as a pairwise Lennard-Jones term.

2 Software

Our approach consists of two phases. The first involves scanning the energy landscape for favorable configurations. This phase can be done once as a preprocessing step and need not be done again. The second phase involves the iterative underestimation of successive collections of local minima with convex quadratic functions, using the configurations from the first phase as initial seed points for optimization. The minima of the underestimators are then used as predicted values for the global minima. Both serial and parallel versions of this "coupled" optimization have been successfully implemented. Preliminary results are reported in [2].

Currently, our research is focused on optimizing parameters in the energy function, in order to obtain the best accuracy in predicting known docking configurations. In particular, we consider the benchmarking set of Chen et al. [1] for testing protein-protein docking algorithms. Of the 59 test cases it contains, 22 are enzyme-inhibitor complexes, 19 are antibody-antigen complexes, 11 are various diverse complexes, and 7 are difficult test cases whose solutions have significant conformational changes. These optimized parameters are expected to yield realistic results for biological problems whose solutions are unknown.

Flexibility in the protein-ligand model is being implemented using a hybrid of global optimization and rotamer search. Near the surface interface, subtle side-chain rearrangements are often necessary to model induced fit between the receptor and the ligand. These rearrangements can be modeled using candidate residue conformations, called rotamers. Using this approach, the protein backbone is held fixed while residues are allowed to take on various configurations. Such pseudo-flexibility is a more viable alternative to full backbone and side-chain flexibility, which requires inordinately many free variables, thus making the computational cost prohibitively expensive. Local shape complementarity analysis performed using the Fast Atomic Density Evaluator [3] will provide added efficiency by highlighting regions in which shape mismatches occur.

[1]University of Wisconsin - Madison, Departments of Mathematics and Biochemistry, Madison WI, USA E-mail: marcia@math.wisc.edu

[2]University of California - San Diego, San Diego Supercomputer Center, San Diego CA, USA E-mail: lindsey@sdsc.edu

[3]University of California - San Diego, Department of Computer Science and Engineering, San Diego CA, USA E-mail: jbrosen@ucsd.edu

[4]University of Wisconsin - Madison, Departments of Mathematics and Biochemistry, Madison WI, USA E-mail: mitchell@math.wisc.edu

[5]Author for correspondence

References

[1] R. Chen, J. Mintseris, J. Janin, and Z. Weng, *A protein-protein docking benchmark*, Prot. Struct. Fun. Gen., 52, pp. 88–91, 2003.

[2] R. F. Marcia, J. C. Mitchell, and J. B. Rosen, *Iterative convex quadratic approximation for global optimization in protein docking*, Comput. Optim. Appl., Submitted, 2003.

[3] J. C. Mitchell, R. Kerr, and L. F. Ten Eyck, *Rapid atomic density measures for molecular shape characterization*, J. Mol. Graph. Model., 19(3), pp. 324–329, 2001.

A3. Using MEGA to Predict Molecular Bio-Activity

Arun Qamra, [1] **King-Shy Goh,** [2] **Edward Y. Chang,** [3]

Keywords: drug design, structure-activity relationships, machine learning, MEGA

1 Introduction

The discovery of a new drug typically requires over 10 years and up to a billion dollars. Machine learning algorithms have the potential to reduce experimental time and cost by intelligently guiding the discovery process. Drug molecules work by binding to protein molecules in the body and modulating their actions. Hence, the first step in drug design is to find compounds that bind with the desired "target" protein. Traditionally this is done by experimentally evaluating activity of a large number of compounds against the target. It is known that the chemical behavior is (largely) dictated by a compound's structure. Machine learning methods can thus be used to learn structure-activity relationship models, which can then be used to virtually screen compounds for activity. Recently, techniques such as Neural Networks, Bayesian Networks [2], and SVMs [3] have been applied to do so. However, this problem presents unique challenges to machine learning, such as the large number of features, limited training data, and significant positive/negative imbalance. We propose the use of MEGA [1] for learning activity models, demonstrate that MEGA can accurately predict activity, and using intelligent sampling, it can do so with much less training data. Another significant advantage is that the model MEGA learns can be interpreted.

2 The MEGA Algorithm

MEGA models concepts in k-CNF. A k-CNF expression, $c_1 \wedge \cdots \wedge c_L$, is a conjunction of terms c_i, where each c_i is a disjunction of at most k predicates (all features and their negations are used as predicates). MEGA also maintains, and refines at each iteration, a k-DNF expression (disjunction of k-predicate conjunctions) that represents the candidate sampling space. MEGA starts with the most specific k-CNF and the most general k-DNF, and for each training instance, removes terms so as to generalize the k-CNF or specialize the k-DNF, depending on whether the instance is positive or negative. The k-CNF expression left after training is used as the learnt model to classify unseen data. MEGA can perform Active Learning by iteratively selecting for labeling, the most informative unlabeled samples. MEGA selects these points based on the candidate sampling space and the learnt concept. This allows an accurate concept to be learnt in a few iterations from a few most informative training instances. Please refer [1] for details. MEGA is appropriate for the drug discovery problem for a number of reasons. The active learning approach used by MEGA is very suitable for the typically iterative drug discovery process, and allows learning from few training instances, thus drastically reducing experimental costs. A significant advantage of MEGA is that the model learnt by MEGA is interpretable since it is a logical expression. Interpretability of the learnt model can provide valuable insights into molecular bio-activity and aid the design and discovery of appropriate drugs. Another advantage is that MEGA can start learning even in the absence of positive instances (unlike most other methods), since

[1] Computer Science, Univ. of California Santa Barbara, `arun@cs.ucsb.edu`
[2] Electrical & Computer Eng., Univ. of California Santa Barbara, `kingshy@engineering.ucsb.edu`
[3] Electrical & Computer Eng., Univ. of California Santa Barbara, `echang@ece.ucsb.edu`

negative instances can shrink the sampling space and increase the probability of a positive instance being sampled in future iterations. This again is well suited to drug discovery since finding initial active compounds may not be easy.

3 Experiments, Results and Discussion

We evaluated MEGA's performance with the Thrombin dataset from the KDD Cup 2001 competition [2], and compared it with that of the KDD Cup winner. The task is to predict binding to thrombin. The training set contains 1909 instances, including 42 actives, while the test set contains 634 instances. 139, 351 binary features are used to describe structural and physical properties of each compound. The problems of high dimensionality, training data scarcity, and positive/negative imbalance, are evident here.

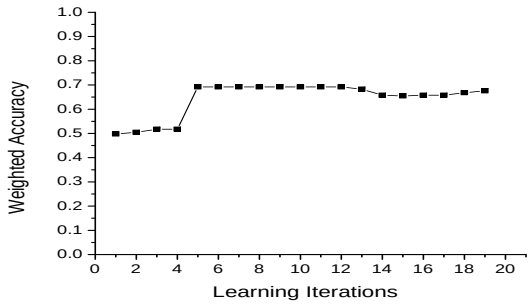


Figure 1: Weighted Accuracy vs Learning Iterations

Before using MEGA to learn the concept from the training data, we used Mutual Information to chose the 100 most important features. For the test set, which contains 150 positives and 484 negatives, the classifier learnt classified with an overall Accuracy of 76.5% and a Weighted Accuracy of 67.6%, generating 75 false positives and 74 false negatives. In comparison, the KDD Cup winner (a Bayesian classifier) gave an overall Accuracy of 71.1% and a Weighted Accuracy of 68.4%, generating 128 false positives and 55 false negatives. Weighted Accuracy is defined as the average of prediction accuracy for positives and that for negatives. Space does not permit us to present comparisons with other techniques, but performance comparable to the KDD Cup winner is encouraging, and given the advantages enumerated above, we can say that MEGA is definitely a promising technique. We next conducted experiments to use MEGA's intelligent sampling to iteratively select samples from the training set for active concept learning. Figure 1 shows the Weighted Accuracy achieved at each iteration, where 100 samples were sampled at each iteration. From the graph, we see that the classifier shows high accuracy after just 5 iterations. With intelligent sampling, we can learn an equally good classifier with just 500 instances instead of the 1909 originally used, thus resulting in drastically reduced experimental costs. Learning here is limited by the size of the given training dataset. In a real scenario, further experiments could be conducted to create more informative training data based on MEGA's recommendations, and thus potentially achieve even higher accuracies at low cost.

References

[1] Chang, E., and Li, B. 2003. MEGA — The Maximizing Expected Generalization Algorithm for Learning Complex Query Concepts. In: *ACM Transactions on Information Systems (TOIS)*

[2] Cheng, J., Hatzis, C., Hayashi, H., Krogel, M.-A., Morishita, S., Page, D., and Sese, J. 2001. KDD Cup 2001 Report. In: *ACM SIGKDD Explorations*, 3(2), pp.47-64

[3] Warmuth, M. K., Ratsch, G., Mathieson, M., Liao, J., and Lemmen, C. 2002. Active Learning in the Drug Drug Discovery Process. In: *Advances in Neural Information Processing Systems*

A4. Structure-Based Design of HIV Entry Inhibitors

Hepan Tan,[1] Jiang Zhu,[2] Wayne A. Hendrickson[3]

Keywords: de no design, flexible docking, genetic algorithm, multiple copy stochastic molecular dynamics, gp120-CD4 interactions, lead discovery, entry inhibitor, receptor flexibility

1 Introduction.

The acquired-immunodeficiency syndrome (AIDS) has evolved into one of the major epidemics worldwide, Significant effort has been made in the development of antiviral therapy. Actually all the compounds that are currently used or under advanced clinical trial can be classified according to the step of the viral life cycle which they target, including those well known inhibitors of HIV reverse transcriptase and protease. However, drug resistance, toxicity, tolerability, latent viral reservoirs are the imminent problems that need to overcome. A new strategy for vaccine and drug design that complements existing cocktail therapy recipes is to target HIV entry. It provides the advantages of interfering with multiple intermediates in this multi-step process. Consequently, viral attachment, co-receptor binding, and viral-cell membrane fusion provide promising targets for anti-HIV drug discovery. Our efforts are focused on the possible strategy to interfere with viral attachment, the very first step of viral entry. The enormous effort to develop small ligands that binds specifically to gp120 and trap it into an early fusion-inactive state has not been very successful until recently.

We therefore carried out a comprehensive computational study on the de novo design and the virtual screening of gp120 inhibitors. Genetic Algorithm controlled de novo design was carried out to propose potential gp120 binders; Virtual screening of NCI-3D compound was carried out using incremental construction algorithm and the resulting top hits were ranked and visualized for further binding assay; also we applied MCSMD approached to design ligands from common medicinal chemistry building blocks.

2 Materials and Methods.

LigBuilder was used to design gp120 inhibitors [1]. Binding site analysis was carried out first to derive a 3D-pharmacophore. Seed structures based on the phenyl ring of Phe43 from CD4 was used to initiate the building process. The best binders from the calculations were clustered based on their theoretical LogP values. FlexX [2] was used to dock NCI-3D database to gp120-CD4 binding site, the binding site residues included in the calculation were within a sphere of 20 Å of which the center is the geometric average of Phe43 phenyl ring and the side chain of Arg59 of CD4. The top 300 hits were clustered and visualized and 15 compounds were obtained for further study. DycoBlock and F-DycoBlock [3] were applied to carry out multiple copy stochastic molecular dynamics (MCSMD) simulation to search for binding sites of building blocks within gp120-CD4 interaction site, the resulting fragments were linked and refined on-the-fly, solvent accessible surface area (SASA) was used to evaluate the best fit ligands. GOLD (Genetic Optimization of Ligand Docking) [4] and MCSMD methods mentioned above were also used to study ligand binding mode with gp120.

[1, 2, 3] Department of Biochemistry and Molecular Biophysics & HHMI, College of Physicians and Surgeons, Columbia University, 630 West 168th Street, BB259, New York, NY 10032. E-mail: ht149@columbia.edu , jz2106@columbia.edu , wayne@convex.hhmi.columbia.edu

3 Results and Discussions.

A panel of compounds were designed with LigBuilder and DycoBlock, and they were compared with the virtual screening results.

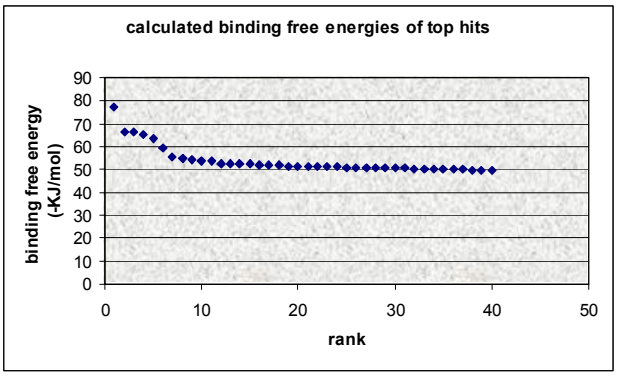


Figure 1: the best-ranked hits from the virtual screening of NCI-3D

Compd	Tot-E	Bind-E	eel	elj	sasa-b	sasa-f
1843	-159.8	-474.5	-318.6	-155.8	128.7	335.2
2095	-421.8	-463.3	-295.1	-168.2	139.3	386.7
612	-490.3	-430.5	-276.3	-154.2	145.2	374.4
2444	-475.1	-427.2	-246.6	-180.6	136.9	402.6
555	-349.0	-419.1	-265.0	-154.1	131.5	372.3
1144	-217.4	-413.6	-334.5	-79.1	226.4	361.0
1219	126.8	-390.6	-231.4	-159.2	126.7	330.4
1808	-214.6	-389.8	-219.1	-170.7	212.6	429.5
2372	-597.4	-371.0	-265.1	-105.9	264.5	414.4
1115	-216.2	-364.3	-206.5	-157.8	222.8	427.1

Table 1: Top 10 MCSMD designed ligands in terms of binding energy

4 References.

[4] Jones G, Willett P, Glen RC, Leach AR, Taylor R. 1997. Development and validation of a genetic algorithm for flexible docking. J Mol Biol. 267(3): 727-48.

[2] Rarey M, Kramer B, Lengauer T, Klebe G. 1996. A fast flexible docking method using an incremental construction algorithm. J Mol Biol. 261(3): 470-89.

[1] Wang R., Gao Y., and Lai L. 2000. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. J. Mol. Model. 6: 498-516.

[3] Zhu J, Fan H, Liu H, Shi Y. 2001. Structure-based ligand design for flexible proteins: application of new F-DycoBlock. J Comput Aided Mol Des. 15(11): 979-96.

A5. *Shape Signature*, a New Approach to Computer-aided Ligand- and Receptor-Based Drug Design

Lifeng Tian[1], Randy J. Zauhar[1]

Keywords: molecular surface, ray tracing, Shape Signatures

1 Introduction.

A unifying principle of rational drug design is the use of either shape similarity or complementarity to identify compounds expected to be active against a given target. Shape similarity is the underlying foundation of ligand-based methods, which seek compounds with structure similar to known active, while shape complementarity is the basis of most receptor-based design, where the goal is to identify compounds complementary in shape to a given receptor. These approaches can be extended to include molecular descriptors in addition to shape, such as lipophilicity or electrostatic potential.

We introduce a new technique, which we call Shape Signatures, for describing the shape of ligand molecules and of receptor site. *Shape Signatures* is an approach for compactly encoding the shapes of molecules and receptors [1]. This is achieved by a method much like ray tracing. We begin with either a ligand molecule or a receptor site in a protein. In either case, we enclose the molecule in its molecular surface [2,3]. A ray is initiated at a randomly chosen point on this surface, and is allowed to propagate by the rules of optical reflection. If we are considering a ligand molecule, then the ray is directed into the molecular interior, while for a receptor the ray is directed to the exterior of the molecule and propagates in the volume of the binding site. For further analysis we retain the positions of the reflection points, as well as the ray tracing *segments* (defined as the line segments that connect reflection points). In addition, we may associate the reflection points with various properties computed on the surface, such as the molecular electrostatic potential (MEP).

Given a completed ray tracing, we compute probability distributions based on the geometry of the ray, and perhaps also properties computed on the molecular surface. Our "Shape Signatures" are just these distributions, described as histograms. The simplest signature is the distribution of segment lengths observed in the ray tracing. This distribution has a one dimensional domain (namely segment length), and we refer to it as a "1D Signature". Signatures with higher dimensional domains can be defined as joint probability distributions that involve both ray tracing geometry and surface properties, such as the molecular electrostatic potential (MEP). A "2D-MEP" signature encodes information about both molecular shape and charge distribution.

Shape signatures can be generated from ray tracing of receptor sites in a manner identical to ligand ray tracing. Where a match between 1D signatures of two ligands indicates they may have similar shape, a match between a receptor-based query and a potential ligand indicates that they are *complementary* in shape.

2 Application.

[1] Department of Chemistry & Biochemistry, University of the Sciences in Philadelphia, 600 S 43rd street, Philadelphia, PA 19104. E-mail: lt0000@usip.edu, r.zauhar@usip.edu

Searching Chemical Libraries for Lead Compounds. Shape Signatures can be used to scan libraries for compounds similar in shape to a query compound (presumably a known active). Matches can be found on the basis of shape alone (using 1D signatures), or shape combined with electrostatics (using 2D-MEP signatures). The first approach casts a wider net, while the 2D approach will locate compounds very similar to the query in both shape and electrostatic field.

Scanning Libraries for Compounds Complementary to a Protein Receptor. A special strength of the Shape Signatures approach is that the handling of receptors and ligands is completely symmetrical – to generate Shape Signatures for receptor sites or sub sites all that is required of the user is to select the atoms that define the site, save these as a file in the query structure database, and use an alternate script designed for receptor-based signature generation.

References

[1] Zauhar, R.J., Moyna, G., Tian, L.-F., Li, Z.-J., and Welsh, W.J. *Shape signatures*, a new approach to computer-aided ligand- and receptor-based drug design, *J.Med. Chem.*, accepted, 2003.
[2] Richards, F. M. Areas, Volumes, Packing and Protein Structure. *Annual Review in Biophysics and Bioengineering*. **1977**, *6*, 151-176.
[3] Zauhar, R. J. SMART: a solvent-accessible triangulated surface generator for molecular graphics and boundary element applications. *J Comput.-Aided Mol. Des.* **1995**, *9*, 149-159.

B1. Distributions of time to coalescence under stochastic population growths: application to MRCA dating

Krzysztof A. Cyran[1] and Marek Kimmel[2]

Keywords: coalescence distributions, MRCA dating, Markov branching processes, mtDNA, HVR

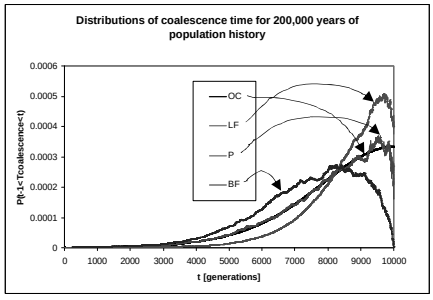
1 Introduction

In the last decade a lot of effort has been spent on inferring human population history from genetic diversity data [6, 7]. The majority of methods were based on the Wright-Fisher (W-F) model of genetic drift which assumes multinomial sampling scheme and thus (for large population) Poisson distribution of the number of progeny for any particular locus. Since this model is not always accurate, the question arises: What is the influence of the departure from W-F model on the distribution of the coalescence time and further analysis of genetic variation? To answer it we performed an extensive (forward direction) simulation study estimating the coalescence distribution for populations evolving according to various stochastic scenarios. We compared coalescence distributions of W-F type models and of the O'Connell (OC) model [4] (corrected in [1]) and the results allowed us to estimate the time to the most recent female common ancestor (MRFCA) of modern humans. For this purpose we used genetic data from HVRI and HVRII of mtDNA of modern humans and Neanderthal fossils.

2 Coalescence distributions in Markov branching processes

We modeled the population trajectories by the slightly supercritical Markov branching process. For such processes, the OC model predicts the asymptotic coalescence distribution to be independent of the type of offspring distribution with given mean and bounded variance. We used the OC model as a standard and checked how well the distributions obtained in various W-F (time-homogeneous) models would match it (Fig. 1a.). We also performed the Kolmogorov-Smirnov test for statistical comparison of distributions obtained with the standard (Table 1, last column), and performed simulations for (time-inhomogeneous) branching processes in variable environment (Fig. 1b.).

a.



b.

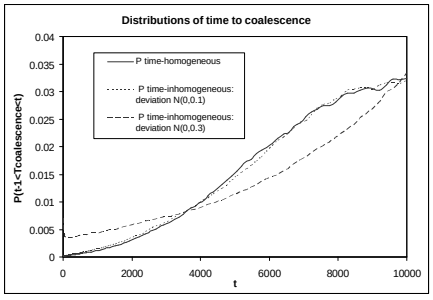


Figure 1: Distributions of time to coalescence for different population histories: a. Distribution for the case of stochastic time homogeneous growth (progeny distributions: P- Poisson, LF-linear fractional, BF- binary fission), b. Coalescence distribution for the cases of stochastic time-homogeneous vs. time-inh. Poisson growth.

[1] (a) Department of Statistics, Rice University, Houston, Texas, USA. E-mail: chrisc1@rice.edu
(b) Institute of Computer Science, Silesian University of Technology, Gliwice, Poland
[2] Department of Statistics, Rice University, Houston, Texas, USA. E-mail: kimmel@rice.edu

Population trajectory scenario	g = $E(T_c/T \mid N_0=1)$	s^2 = $Var\,(T_c/T \mid N_0=1)$	s	*Final pop. size*	*Equal to OC distribution?*
OC	0.801	0.0253	0.159	10^7	–
W-F with P	0.802	0.0253	0.159	10^7	Yes
W-F with BF	0.735	0.0289	0.17	0.5×10^7	No
W-F with LF	0.844	0.0243	0.156	2×10^7	No
W-F with P, s_{e1}	0.794	0.0289	0.17	10^7	Not sure
W-F with P, s_{e2}	0.699	0.0724	0.269	2×10^7	No

Table 1: Relative time to coalescence g and its variance s^2 for a pair of alleles for various demographic scenarios starting from a common ancestor. Two last rows are for randomly changing environment with std. dev. $s_{e1} < s_{e2}$.

3 Estimation time to mtEve

In order to be able to date the MRCFA based on data from Table 1, we have to know the average genetic distance between modern humans d_{avg}, divergence rate d and the duration of one human generation λ. After a series of successful sequencings of *H. neanderthalensis* mtDNA [2, 3] dated to live until about 40, 000 years ago [5], d no longer had to be estimated from human-chimpanzee divergence. Assuming the infinite site model, relying on data from [2] and using *H. neanderthalensis* as an outgroup, we estimated d to be about 1.2×10^{-7}. For d_{avg}=1.8% [2] and λ=20 years this results in times to mtEve given in Table 2. For all stochastic trajectories we analyzed, the resulting time falls into the 95% confidence interval of the estimate reported in [2]. However, our results, with the average of 193×10^3 years, indicate a systematic shift of 30×10^3 years towards the past. We also showed in this paper that after changing the outgroup from chimpanzee to Neanderthals, genetic models with different assumptions tend to give similar (therefore mutually supporting) predictions. However, interestingly, the estimates in the deterministic growth models are systematically higher than those in the stochastic model. The computer program used for calculations of the coalescence distribution can be downloaded from web: www.stat.rice.edu/~kimmel/software/coalescence.

	W-F time-homogeneous			W-F time-inh. P.		W-F exponential growth			
OC.	P	BF	LF	s_{e1}	s_{e1}	$Z_T=10^9$	$Z_T=10^8$	$Z_T=10^7$	$Z_T=10^6$
187	187	204	178	189	215	223	239	266	311

Table 2: Estimates of the time to mtEve $E(T_a)$ in [10^3 years] for various population history scenarios.

References

[1] Kimmel, M. and Axelrod D. E. 2002. *Branching Processes in Biology*. NewYork: Springer-Verlag.

[2] Krings, M., Geisert, H., Schmitz, R., Krainitzki, H. and Pääbo S. 1999. DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen, *Proceedings of the National Academy of Sciences USA*, 96: 5581-5585.

[3] Krings, M., Capelli, C., Tschentscher, F., Geisert, H., Meyer, S. von Haeseler, A., Grossschmidt, K., Possnert, G., Paunovic, M. and Pääbo, S. 2000. A view of Neandertal genetic diversity, *Nature Genetics*, 26: 144-146.

[4] O'Connell, N. 1995. The genealogy of branching processes and the age of our most recent common ancestor, *Advances in Applied Probability* 27: 418-442.

[5] Schmitz, R., Bonani, G., and Smith, F. H. 2002. New research at the Neandertal type site in the Neander Valley of Germany. *Journal of Human Evolution* 42(3):A32-A32.

[6] Wooding, S. and Rogers A. 2000. A Pleistocence population X-plosion? *Human Biology* 72: 693-695.

[7] Wooding, S. and Rogers A. 2002. The matrix coalescence and an Application to Human Single-Nucleotide Polymorphisms, *Genetics* 161: 1641-1650.

B3. Analysis of Sorting by Transpositions based on Algebraic Formalism [1]

Cleber V. G. Mira, [2] **Joao Meidanis,** [3]

Keywords: Genome Rearrangements, Computational Biology

1 Statemet of the Problem.

Genome rearrangements analysis focus on the relative positions of the same block of genes at two or more distinct genome sequences. Some mutational events, such as *transpositions*, affect the genome sequences solely in their ordering of blocks of genes. Given a permutation representing a genome $\pi = (\pi_1\ \pi_2\ \ldots\ \pi_n)$, a *transposition* $\tau(i,j,k)$ for $1 \leq i < j \leq n$ and $1 \leq k \leq n$, but $k \notin [i,j]$, is the following operation on π.

$$\tau(i,j,k)\pi = (\pi_1\ \pi_2\ \ldots\ \pi_{i-1}\ \pi_j\ \ldots\ \pi_{k-1}\ \pi_i\ \ldots\ \pi_{j-1}\ \pi_k\ \ldots\ \pi_n)\,,$$

if $i < j < k$.

The *problem of transposition distance* consists in finding the minimum number of transpositions to transform one genome into another. That is:

$$\sigma = \tau_t\ \tau_{t-1}\ \ldots\ \tau_1\ \pi$$

The number t is the transposition distance $d_\tau(\pi,\sigma)$ between two genomes π and σ. For example, consider the following sequence of transpositions which order the permutation (4 3 2 1 5):

$$\begin{aligned} \tau(1,5,6)\pi &= (1\ 4\ 3\ 2\ 5) \\ \tau(2,5,6)\tau(1,5,6)\pi &= (1\ 2\ 4\ 3\ 5) \\ \tau(3,5,6)\tau(2,5,6)\tau(1,5,6)\pi &= (1\ 2\ 3\ 4\ 5) \end{aligned}$$

2 Algebraic Formalism

The permutations can be analyzed through a graph representation called *cycle graph* [1]. However, we represent the permutations and transpositions by means of a new algebraic formalism developed by Meidanis and Dias [2]. In this approach a genome is described as a permutation on the symmetric group over $\{0,\ 1\ \ldots\ n\}$. But we are interested in the cycle decomposition of the permutations in S_n. Since the transposition event does not change the orientation of a block, only one of the strands is considered in its cycle decomposition representation. A genome in the Algebraic Formalism is usually represented as:

$$\pi = (0\ \pi_1\ \pi_2\ \pi_3\ \ldots\ \pi_n)(\overline{\pi_n}\ \ldots\ \overline{\pi_3}\ \overline{\pi_2}\ \overline{\pi_1}\ -0)$$

Observe that the "dummy block" zero is used in this representation. The earlier permutation is a product of two disjoint cycles, each one representing a strand of the genome. As the

[1] Research supported by grants from FAPESP.

[2] Institute of Computing, University of Campinas (UNICAMP), Sao Paulo, Brazil E-mail: cleber@ic.unicamp.br

[3] Scylla Bioinformatics, Sao Paulo, Brazil E-mail: meidanis@scylla.com.br

transposition event does not change the orientation of the blocks of genes, we will not consider the strand which has the block -0.

This permutation is seen as a function which induces a circular order of its elements, such that $\pi_{i+1} = \pi(\pi_i)$. The *identity permutation*,1, in the permutation group is $(1)(2)(3)\ \ldots\ (n)$. Each element in an 1-cycle in the cycle decomposition of a permutation is called a *fixed element*. Fixed elements are usually omitted in the cycle decomposition representation. The *support*, $Supp(\pi)$, of a permutation π is the subset of elements not fixed in π.

The product of permutations, $\pi\sigma$, is performed in this way: for each element $x \in [n]$ is applied the composition $(\pi\sigma)(x) = \pi(\sigma(x))$. For instance, consider this example: $(3\ 2\ 5\ 1)(6\ 4\ 2) = (1\ 3\ 2\ 6\ 4\ 5)$. The *inverse permutation* of π is the permutation π^{-1}, such that $\pi\pi^{-1} = 1$. To obtain the inverse permutation of a cycle π is easy — the inverse of $\pi = (\pi_1\ \pi_2\ \ldots\ \pi_n)$ is $\pi^{-1} = (\pi_n\ \pi_{n-1}\ \ldots\ \pi_1)$. A permutation τ *divides* a permutation π, $\tau|\pi$, if and only if $|\pi\tau^{-1}| = |\pi| - |\tau|$, where $|\pi|$ is the *norm* of π; i.e. the minimum 2-cycle decomposition of π.

A transposition in this new approach is the permutation $\tau(\pi_u, \pi_v, \pi_w) = (\pi_u\ \pi_v\ \pi_w)$. To apply a transposition in the genome π is to perform the product $\tau\pi$. For instance: $(4\ 2\ 5)(0\ 1\ 4\ 3\ 2\ 5) = (0\ 1\ 2\ 4\ 3\ 5)$.

A transposition τ is *applicable* to π if $\tau\pi$ is a strand. Also, a transposition τ is applicable to π if and only if $\tau|\pi$. There exists transpositions which are not applicable to a genome π. For example: $(4\ 5\ 2)(0\ 1\ 4\ 3\ 2\ 5) = (0\ 1\ 5)(2)(4\ 3)$. The length of a cycle α in the cycle decomposition of a permutation π is $|Supp(\alpha)|$. A cycle is *odd*, if its length is odd.

3 Transposition Distance Bounds.

Let $|\pi|_3$ denotes the minimum number of 3-cycles $\tau_1, \tau_2, \ldots, \tau_k$, where $k = |\pi|_3$, such that $\pi = \tau_1\tau_2\ldots\tau_k$. The algebraic approach provides the following lower bound to the transposition distance. Notice that given a genome π and $\tau_1\ \tau_2\ \ldots\ \tau_k\ \pi = \sigma$, such that k is minimum, then $\tau_1\ \tau_2\ \ldots\ \tau_k = \sigma\pi^{-1}$. Therefore:

Proposition 3.1 (Lower Bound) $d_\tau(\pi, \sigma) \geq |\sigma\pi^{-1}|_3$.

The formula $\sigma\pi^{-1}$, which is called *Quotient*, is very important in the algebraic theory because it straightforwardly provides lower bounds to others rearrangement problems [2] and gives an algebraic relationship between the genomes π and σ. Next we state that the previous lower bound is equivalent to the best known lower bound [1].

Proposition 3.2 $|\pi|_3 = \frac{(n - c_{odd}(\pi))}{2}$

A *split* is a transposition not applicable to π. If we permit splits besides transpositions, then the split+transposition distance, $d_{st}(\pi, \sigma)$, is:

Proposition 3.3 (Split+Transposition Distance) $d_\tau(\pi, \sigma) = |\sigma\pi^{-1}|_3$.

References

[1] V. Bafna and P. A. Pevzner, 1995. Sorting by Transpositions. In: *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*,San Francisco, USA, pp. 614–623

[2] J. Meidanis and Z. Dias 2000.An Alternative Algebraic Formalism for Genome Rearrangements. In: *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and Evolution of Gene Families* "D. Sankoff and J. H. Nadeau",editors, Kluwer Academic Publishers. pp 213–223.

B5. An Integrated Tool for Investigating Genetic Disorder-Relevant Tandem Repeats in Human Genome

Feng-Mao Lin[1], Ming-Yu Chen[2], Hsien-Da Huang[3], Jorng-Tzong Horng[4]

Keywords: genetic disorder, tandem repeats, single nucleotide polymorphisms

1 Introduction.

Tandem repeats (TRs) and single nucleotide polymorphisms (SNPs) are associated with human inherited diseases, play an important role in evolution and in regulatory processes [1-3]. Because the biologists, who are investigating in genetic disorders, are also interested in TRs and SNPs with particular limits, and they will design primer sequence for experiment. Hence, the goal in this work is to develop an effective tool for observing the information about TRs, SNPs and genetic disorders. We establish a database to store gene information, TRs, SNPs and OMIM [4] data. The system facilitates the analysis of genetic disorders. We develop a primer design tool for identifying specific TRs or SNPs when users interested in specific disease features. Web user interfaces and graphical interfaces are designed and implemented. Accordingly, the relationship of genes and genetic disorders recorded in OMIM are found. The main contribution of this work is to provide a user-friendly and effective tool for genetic disorders in the research of human inherited diseases.

2 Materials and methods.

The present system contains two parts. They are data processing and result display. From GenBank [5] the DNA sequences of Homo sapiens have been used for identifying tandem repeats, and the gene coding regions from GenBank have been used for retrieving gene location data. The relationship between human genes and genetic disorders from OMIM [6] has been retrieved. From the above steps, the information of tandem repeats mapping genes and genetic genes mapping disorders has been used for generating the association of tandem repeats with disease phenotypes. The part of result representation has been developed with a user-friendly interface in PHP. Users can access the database via graphical web pages. The mechanism of query-refinement helps users for browsing in tandem repeats efficiently. To provide primer sequence for experiment on user-interesting genomic sequence regions, the primer design tool has been integrated into our system, and the primer3 has been applied to fit this purpose. Primer3 pick primers from a DNA sequence [7], and it can avoid choosing primers in transposable elements and can pick oligonucleotide for probe or primers.

3 Results.

[1] Department of Computer Science and Information Engineering, National Central University, Taiwan. E-Mail: meta@db.csie.ncu.edu.tw

[2] Department of Computer Science and Information Engineering, National Central University, Taiwan. E-Mail: puenny@db.csie.ncu.edu.tw

[3] Department of Biological Science and Technology & Institute of Bioinformatics, National Chiao-Tung University, Hsin-Chu, Taiwan. E-Mail: bryan@mail.nctu.edu.tw

[4] Department of Computer Science and Information Engineering and Department of Life Science, National Central University, Taiwan. E-Mail: horng@db.csie.ncu.edu.tw

The possible tandem repeats have been searched in the human genome by Tandem Repeat Finder [8], and the occurrence of repeats in the genomic regions has been identified based on the annotation of the human genome sequence in the GenBank database. The expansion of repeats in the coding region of the gene has been found that being associated with genetic disorders in the OMIM database. These data was stored in our database, and a web site was designed for accessing these information. There are 24 chromosomes, which contain 26,179 candidate genes, 1,246,831 distinct tandem repeat patterns and 34,263,072 tandem repeat sites. Most of these tandem repeat sites represented in non-genomic regions, only nearly 2% of them represented in such genomic regions as exons, introns, upstream and downstream of genes. Table 1 shows the distribution of these genes associated tandem repeat sites.

Region	Occr.	Ratio
Exon	127,537	1.83%
Intron	6,574,552	94.37%
Upstream	127,519	1.83%
Downstream	137,212	1.97%
All	6,966,820	100.00%

Table 1. The distribution of tandem repeat sites which represented in various genomic regions.

The existence of genetic disorders associated with the expansion of tandem repeats raises the interests of clinicians and experts who research in inherited diseases. As a tool, it may help clinicians and experts' studies in observation of the relationship between tandem repeats and genetic disorders. With a graphical user interface, it integrates not only the information about tandem repeats, genes, and genetic disorders but also primer design tool. Generations of the genomic regions and tandem repeats sites have been done and they have been mapped to genetic disorders in this study. Observing on these data as users' wish for retrieving tandem repeats that are associated with disorder-mapped genes can be reached via our tool, and then it provides primer sequences for experiment to verify the suspect.

References

[1] Cummings, C. J. and Zoghbi, H. Y. 2000. Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum Mol Genet* 9:6 909-16.
[2] Stallings, R. L. 1994. Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases. *Genomics* 21:1 116-21.
[3] Subramanian, S., Madgula, V. M., George, R., Mishra, R. K., Pandit, M. W., Kumar, C. S., and Singh, L. 2003. Triplet repeats in human genome: distribution and their association with genes and other genomic regions. *Bioinformatics* 19:5 549-52.
[4] Parton, M. J. 2003. Online Mendelian Inheritance in Man OMIM: www.ncbi.nlm.nih.gov/entrez. *J Neurol Neurosurg Psychiatry* 74:6 703.
[5] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. 2003. GenBank. *Nucleic Acids Res* 31:1 23-7.
[6] Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V. A. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 30:1 52-5.
[7] Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365-86.
[8] Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 27:2 573-80.

B6. Search Space Reduction via Clustering for Haplotype Reconstruction

Jinghua Hu, [1] **Weibo Gong,** [2] **Patrick A. Kelly** [3]

Keywords: Haplotype Reconstruction, Genotype, Space Reduction, Clustering, Algorithm

1 Introduction.

The problem of haplotype reconstruction from phase unknown genotypes has been formulated into two major categories. One is the Haplotype Identification(HI) problem that identifies the haplotype set to explain the genotypes. Related algorithms include Clark's parsimony algorithm [1] and Gusfield's Perfect Phylogeny [2]. The other is the Haplotype Frequency Estimation(HFE) problem that seeks the optimal estimation on haplotype frequencies for resolving the genotypes. Related algorithms include EM-based algorithm [3], Gibbs sampling based algorithm [4], Partition-Ligation [5], etc.

One of the challenges for large scale haplotype reconstruction is the problem size that grows exponentially with the number of heterozygous sites in genotype. Divide-and-Conquer strategies have been applied in [5] and [6] to address the problem. Genotypes are broken into short segments where haplotyping is easily performed, and then the partial solutions are combined together to recontruct the full sequences.

This poster introduces our study on a new approach to problem size reduction for large scale haplotyping. The goal of this approach is to systematically reduce the haplotype search space while preserving the most possible solutions in the reduced space. The reduced search space will then serve as the starting point for existing algorithms, such as EM, to complete the haplotype inference.

The motivation is to examine the inter-correlations within the sample population. We present the framework of applying clustering techniques on genotype data based on the compatibility rules and proximity measures derived from genotype patterns. We examine the quality of clustering results through the reduced haplotype pools constructed from the clusters. Comparison is carried out on several performance indices, such as set coverage rate, size reduction ratio, etc. Results from different clustering algorithms, as well as results integrated from multiple clustering runs are collected for analysis.

Through experiments on simulated data sets, we demonstrate that our approach to search space reduction, in combination with existing algorithms, would enable us to handle large scale haplotyping problems more efficiently.

2 Search Space Reduction via Clustering.

The input of the algorithm is the genotype data matrix of size $N \times L$, where N is the population size, and L the length of genotype. The output is the haplotype pool constructed from the clusters, i.e., the *reduced pool*. The framework consists of three stages.

[1]Department of Electrical and Computer Engineering, University of Massachusetts at Amherst, Amherst, MA 01003, USA. E-mail: jhu@ecs.umass.edu

[2]Department of Electrical and Computer Engineering, University of Massachusetts at Amherst, Amherst, MA 01003, USA. E-mail: gong@ecs.umass.edu

[3]Department of Electrical and Computer Engineering, University of Massachusetts at Amherst, Amherst, MA 01003, USA. E-mail: kelly@ecs.umass.edu

1. Build compatibility and proximity matrices from input data.
2. Apply clustering algorithms on genotypes. Genotypes are assigned into clusters based on compatibility and proximity.
3. Construct haplotype pools from clusters by keeping the haplotypes shared by all members in a cluster. All partial pools are merged together as the final pool.

We adopt two indices for evaluating the quality of clustering results. Here the *original pool* refers to the haplotype set created by enumerating all possible haplotypes from genotypes, and the *truth pool* refers to the correct haplotype set.

- Set Coverage Rate: Percentage of haplotypes in the *truth pool* correctly preserved in the *reduced pool*.
- Size Reduction Ratio: Ratio between the size of the *original pool* and the size of the *reduced pool*.

The choice of proximity measures, clustering criteria, and clustering algorithms should aim at the goal of constructing reduced haplotype pools of high coverage rates as well as large reduction ratios.

3 Experimental Results.

The experiments are carried out on simulated data sets with the following observations.

- Efficient space reduction: A basic implementation of the sequential clustering algorithm combined with randomized assignment of genotypes would yield an average set coverage rate of above 80%, with an average size reduction ratio of 60 on data sets $L = 32, N = 32$.
- Merged pools perform better: Merged pools from multiple runs of clustering yield better quality at limited extra cost. Clustering algorithms may produce complementary results that work better together.
- The quality of clustering results also depend on input data and the settings of clustering algorithms. Further study on the relationship between input data patterns and computational complexity is desired.

4 References and bibliography.

References

[1] Clark, A. G. 1990. Inference of haplotypes from PCR-amplified samples of diploid populations. *Molecular Biology and Evolution*, 7:111-122.

[2] Gusfield, D. 2002. Haplotyping as Perfect Phylogeny: Conceptual Framework and Efficient Solutions. *Proceedings of 6th ACM International Conference on Computational Biology (RECOMB)*, ACM Press, pp. 166-175.

[3] Excoffier, L. and M. Slatkin. 1995. Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population. *Molecular Biology and Evolution*, 12(5):921-927.

[4] Stephens, M., J. J. Smith and P. Donnelly. 2001. A New Statistical Method for Haplotype Reconstruction from Population Data. *American Journal of Human Genetics*, 68:978-989.

[5] Niu, T., Z. S. Qin, X. Xu and J. S. Liu. 2002. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *American Journal of Human Genetics*, 70:157-169.

[6] Eskin, E., E. Halperin and R. M. Karp. 2003. Large Scale Reconstruction of Haplotypes from Genotype Data. *Proceedings of 7th International Conference on Computational Biology (RECOMB)*, pp. 104-113.

B7. Reconstructing Phylogenetic Trees from Dissimilarity Maps

Dan Levy, [1] **Francis E. Su,** [2] **Ruriko Yoshida** [3]

Keywords: Phylogenetic tree, dissimilarity map, four point condition

Algorithms for reconstructing a phylogenetic tree from sequence data typically rely on first finding a distance matrix; a set of values indicating the dissimilarity for each pair of elements [1] [2] [9] [11]. However, one may choose to consider instead the dissimilarity of groups of elements. If we let $[n]$ denote the set $\{1, 2, ..., n\}$ of leaves and $\binom{[n]}{m}$ denote the set of all m-element subsets of $[n]$, we may define an *m-dissimilarity map* as a function $D : \binom{[n]}{m} \to \mathbf{R}_{\geq 0}$. This may be thought of as a measure of how dissimilar a set of m elements are or, in terms of the tree, as the sum of the edge weights of the subtree spanned by those leaves. Pachter and Speyer gave theoretical conditions under which a tree may be constructed from its dissimilarity map; in particular a tree may be reconstructed from its m-dissimilarity map iff $n > 2m - 2$ [5]. Their method proceeds by first finding the splits, recovering the topology from the splits using Buneman indices, and then using the distances and topology to determine the edge weights.

In contrast, our method constructs the topology of the tree and determines the edge weights simultaneously. We achieve this result through an analogue of the four point condition [3] [9] for m-dissimilarity maps. We say that a pair of leaves (i, j) is a *sub-cherry* in the subtree T iff i and j are in T and there is only one vertex of degree 3 on the unique path from i to j in T. We call this intermediate vertex the *sub-cherry node*. Our four point condition for m-dissimilarity maps locates sub-cherries in subtrees with $m + 2$ leaves and determines the distance from the sub-cherry node to the rest of the subtree. By locating sub-cherries and determining sub-cherry node distances, we may reconstruct both the tree and the edge weights provided that $n > 2m - 2$.

In the case that $m = 3$, we have exploited certain symmetries to provide a fast algorithm for reconstructing phylogenetic trees from 3-dissimilarity maps. This algorithm has a time complexity of $O(n^2)$. We have also written and tested a C++ implementation of this algorithm.

Phylogenetic tree reconstruction from m-dissimilarity maps may provide a more accurate approach to determining the maximum likelihood tree by utilizing the more accurate m-subtree weights as opposed to pairwise distances [6] [10]. As demonstrated by the case $m = 3$, these algorithms are computationally competitive with distance based methods and may serve as a viable alternative to neighbor joining or quartet reconstructions.

References

[1] Bruno, W. J., Socci, N.D., and Halpern, A.L. 2000. Weighted Neighbor Joining: A Likelihood-based approach to distance-Based phylogeny reconstruction. *Molecular Biology and Evolution 17*. pp. 189 - 197.

[1]Department of Mathematics, University of California, Berkeley, CA. E-mail: levyd@math.berkeley.edu

[2]Department of Mathematics, Harvey Mudd College, Claremont, CA 91711. E-mail: su@math.hmc.edu

[3]Department of Mathematics, University of California, Davis, CA. E-mail: ruriko@math.ucdavis.edu

[2] Buneman, P. 1971. The recovery of trees from measures of dissimilarity. In: *Mathematics in the Archaeological and Historical Science, (ed. F. R. Hodson, D. G. Kendall, and P. Tautu),* Edinburgh University Press, Edinburgh. pp. 387-395.

[3] Buneman, P. 1974. A note on the metric property of trees. *Journal of Combinatorial Theory.* Series B, 17. pp.48-50.

[4] Hakimi, S.L. and Yau, S.S. 1965. Distance matrix of a graph and its realizability. *Quart. Appl. Math. 22.* pp. 305-317.

[5] Pachter, L. and Speyer, D. 2004. Reconstructing trees from subtree weights. *Applied Mathematics Letters*, in press.

[6] Ranwez, V. and Gascuel, O. 2002. Improvement of distance-based phylogenetic methods by a local maximum likelihood approach using triplets. *Molecular Biology and Evolution, 19.* pp.1952-1963.

[7] Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4.* pp.406-425.

[8] Semple, C. and Steel, M. 2003. Phylogenetic, *Oxford Lecture Series in Mathematics and its Applications, 24.* Oxford University Press.

[9] Simöes-Pereira, J.M.S. 1969. A note on the tree realizability of a distance matrix. *Journal of Combinatorics Theory, 6.* pp.303-310.

[10] Strimmer, K. and von Haessler, A. 1996. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution, 13.* pp.964-969.

[11] Zaretskii, K.A. 1965. Constructing trees from the set of distances between pendant vertices. *Uspehi Matematicekih Nauk, 20.* pp. 90-92.

B8. Global optimization in QTL analysis

Kajsa Ljungberg[1], **Sverker Holmgren**[1], **Örjan Carlborg** [2]

Keywords: genetic mapping, quantitative traits, QTL, epistasis, global optimization

1 Introduction and problem formulation.

Many phenotypes of medical, economical and general scientific importance can be measured quantitatively. Quantitative traits are often affected by the joint effect of multiple genes and the environment, and one way to dissect the underlying genetic architecture is to identify quantitative trait loci, QTL, in the genome [2]. A QTL is a chromosomal region, locus, harboring one or several genes that affect the trait under study. Ten or more QTL can influence a single trait. Real examples show that the QTL can interact in nonlinear ways, and therefore it is desirable to simultaneously model their effects. A simultaneous search for n QTL can be regarded as a global optimization problem in n dimensions.

We consider QTL mapping in experimental populations. The exact form of the optimization problem depends on the choice of mapping method. With the widely used linear regression parametric method, searching for n QTL is equivalent to finding the n-element vector $\bar{x}$, representing the combination of n QTL positions, that minimizes

$$f(\bar{x}) = \min_b \|A(\bar{x})b - y\|_2^2,$$

where A is a matrix of indicator variables depending nonlinearly on $\bar{x}$ and of fixed effects coefficients used to remove the influence of non-QTL factors. The number of rows in A equals the number of individuals in the population. The vector y contains the phenotype values, and b is a vector of regression variables. Other common parametric methods result in a generalized least squares problem for every $\bar{x}$ [4].

2 Methods.

There are two parts to the computational problem. The first is the kernel problem, i.e. to solve the (generalized) least squares problem of the objective function for a given $\bar{x}$. In [4] we show how the special structure of the least squares problem can be efficiently exploited in an updating algorithm. We compare our kernel algorithm with the library least squares solver routines used in standard software. The relative gain in arithmetic operations, compared to the most suitable of the library routines, is roughly proportional to $(k_{fix}/k)^2$, the square of the ratio of the number of fixed columns in A to the total number of columns.

The second part of the computational problem is the global problem, i.e. finding the $\bar{x}$ that minimizes $f(\bar{x})$. The standard way of finding the global optimum of the objective function, i.e. the most likely positions of the QTL given a mapping method and model, is to perform an exhaustive search in a dense grid covering the search space. This ensures that the global optimum is found, but the method is computationally slow, and in practice infeasible for analyses in more than two dimensions. Already in two dimensions it is very

[1]Dept. of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden. E-mail: kajsa.ljungberg@it.uu.se, sverker.holmgren@it.uu.se

[2]Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, United Kingdom. E-mail: contact@orjancarlborg.com

time-consuming to perform the thousands of searches required to obtain robust empirical significance thresholds for statistical evaluation of the results. In higher dimensions a more efficient global optimization algorithm is essential. In [5] we adapt an optimization algorithm called DIRECT [3] to the QTL search problem. We compare DIRECT with exhaustive search and with a genetic optimization algorithm previously used for QTL search [1].

3 Results.

We test DIRECT in 2-6 dimensional searches. To verify that the global optimum is found when testing on real data sets, an exhaustive search is performed in two and three dimensions. We also analyze simulated data with known QTL locations in up to six dimensions.

Global search method	Kernel algorithm	2D search, 191 pigs $(k_{fix}/k)^2 \approx 0.78$ [seconds]	3D search, 850 chickens $(k_{fix}/k)^2 \approx 0.03$ [seconds]
Exhaustive search	Library routine G02DAF	150,000	1,140,000,000
Exhaustive search	Library routine SQRDC	4,800	14,800,000
Exhaustive search	New updating algorithm	1,500	11,400,000
Genetic algorithm	New updating algorithm	40	8600
DIRECT	New updating algorithm	4	360

Table 1: Approximate CPU time in seconds required for one search to find the global optimum.

In Table 1 we show examples of results for searches in two and three dimensions using two farm animal real data sets. The reported CPU times are approximate, and the times for exhaustive search using G02DAF in 2 and 3 dimensions and SQRDC in 3 dimensions have been extrapolated from shorter runs. The kernel updating algorithm gives a substantial gain when the number of fixed columns in A is large compared to the total number of columns. The library routine G02DAF, used in standard software, is very slow due to several extra computations not needed for QTL mapping purposes.

Using DIRECT for the global optimization results in a one order of magnitude speed-up compared to the genetic optimization algorithm using a carefully tuned parameterization. DIRECT is 2-3 orders of magnitude faster than exhaustive search in two dimensions, and 4-5 orders of magnitude faster in three dimensions. This enables routine searches in at least three dimensions, including derivation of empirical significance thresholds.

References

[1] Carlborg, Ö., Andersson, L. and Kinghorn, B. 2000. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics* 155:2003–2010.

[2] Doerge, R. 2002. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* 3:43–52.

[3] Jones, D., Perttunen, C. and Stuckman, B. 1993. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications* 79:157–181.

[4] Ljungberg, K., Holmgren, S. and Carlborg, Ö. 2002. Efficient algorithms for quantitative trait loci mapping problems. *Journal of Computational Biology* 9:793–804.

[5] Ljungberg, K., Holmgren, S. and Carlborg, Ö. 2003. Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. Technical Report 2003-043, Department of Information Technology, Uppsala University, Sweden. Submitted to *Bioinformatics*.

B9. The Portable Cray Bioinformatics Library

James Long[1]

Keywords: benchmarking, bioinformatics, compression, endian, library

1 Introduction.

The original Cray Bioinformatics Library (CBL) is a low level set of library routines using proprietary Cray hardware to implement some common nucleotide/protein sequence manipulations typical in a bioinformatics context. Written in Fortran and Cray assembly language (most are callable from C), the original CBL was coded and optimized on a Cray SV1 vector machine. Cray also has a port for their new X1™.

The Portable CBL is the open source version [1] written in C that implements the computational primitives in a generic fashion with little regard to specific hardware. The CBL routines facilitate performance by operating on compressed data whenever possible. In the case of nucleotide data, for example, it is sufficient to represent each of the four nucleotides with only two bits, and thus a 64-bit word can contain a sequence of 32 nucleotides instead of the normal 8. The CBL search routine compares whole words of a compressed query against a compressed database, realizing a significant performance increase. In addition to 2-bit compression, CBL supports 4 bit and 5 bit levels for larger alphabets. The CBL will continue to grow as additional biological computational primitives are identified and implemented [2].

2 Version 1.0 Routines.

cb_amino_translate_ascii - translate nucleotides to amino acids
cb_compress - compresses nucleotide or amino acid ASCII data
cb_copy_bits - copy contiguous sequence of memory bits
cb_countn_ascii - counts A, C, T, G, and N characters in a string
cb_fasta_convert - restructure the memory image of a FASTA file
cb_free - frees memory allocated with cb_malloc in Cray version
- simply calls free() in portable version
cb_irand - generates an array of random bits
cb_malloc - allocate block aligned memory region in Cray version
- simply calls malloc() in portable version
cb_read_fasta - loads data from a FASTA file into memory arrays
cb_repeatn - find short tandem repeats in a nucleotide string
cb_revcompl - reverse complements compressed nucleotide data
cb_searchn - gap-free nucleotide search allowing mismatches
cb_uncompress - uncompress nucleotide or amino acid data to ASCII
cb_version - returns the version number of libcbl

[1] Arctic Region Supercomputing Center, PO 756020, Fairbanks, AK 99775-6020
E-mail: jlong@arsc.edu

3 Performance.

A benchmark option in v1.0 exercises seven of the routines. Platforms used:

800 MHz Cray X1, running in both MSP and SSP mode
1.3 GHz Intel Itanium 2, 1.5 MB L3, intel 7.1 (icc) and gcc 2.96 compilers
1.4 GHz AMD Athalon MP 1600+, 256 KB cache, intel 7.1 and gcc 3.2.2 compilers
1.7 GHz IBM P4, 32 MB shared L3, 64-bit mode
2.8 GHz Intel Xeon, 512 KB cache, intel 8.0 and gcc 3.2.2

	Cray CBL		Portable CBL								
CBL Function	**800 MHz X1**		**800 MHz X1**		**1.3 GHz Itanium2**		**1.4 GHz AMD**		**1.7 GHz IBM**	**2.8 GHz Xeon**	
	MSP	**SSP**	**MSP**	**SSP**	**icc**	**gcc**	**icc**	**gcc**	**P4**	**icc**	**gcc**
cb_amino_tran	8	27	90	156	31	46	63	87	36	25	64
cb_compess/un	5	10	44	56	24	59	64	70	38	31	39
cb_copy_bits	3	4	1	1	8	27	45	45	10	19	18
cb_count_ascii	4	15	5	23	16	44	59	62	23	23	27
cb_repeatn	45	55	122	142	42	55	48	49	26	25	27
cb_revcompl	3	12	19	33	20	80	148	138	35	53	63
cb_searchn	23	85	36	65	92	94	129	167	40	78	127

Table 1: Benchmark times in seconds.

4 Roadmap.

The Portable CBL will follow the roadmap for Cray's implementation (now at 2.0). Developers interested in contributing to the roadmap should consult the author.

Coming in version 1.1:
cb_swa_fw - compute Smith-Waterman cell scores with ASCII input

Coming in version 1.2
cb_isort & cb_isort1 - unsigned integer radix sort with and w/o index array
cb_cghistn - histograms of cg density in a string
cb_swn_fw & cb_swn4_fw- same as cb_swa_fw, except with 2- or 4-bit nucleotide input
cb_nmer - creates up to 64-bit-length short sequences from each starting point in the input string.

Coming in version 2.0
cb_sort – multi-pass sort routine for compressed data

References

[1] http://cbl.sourceforge.net
[2] Long, J. 2003. The Portable Cray Bioinformatics Library. *Proceedings of the 45th Cray User Group Conference,* http://www.arsc.edu/support/technical/html/200305.OpenCBL/jlong_cbl.htm

B10. Before SNP mapping: Data preprocessing by fixed length genomic sequence patterns

Chia-Hao Ou , Ming-Jing Hwang

Keywords: SNP mapping, genomic sequence analysis, sequence patterns, data preprocessing

1 Introduction.

How to map SNP sequences onto their genomic positions and evaluate the mapping reliability is an issue of interest in current bioinformatics researches. The conventional approach to this problem is to use sequence alignment tools, such as BLAST [1], to evaluate the mapping results [2], but sequence alignment process demands a great deal of computational power, and repetitive DNA, such as repetitive elements and segmental duplications, could significantly complicate sequence alignment. Our previous work, the UniMarker (UM) method [3], employed an alignment-free sequence mapping program and achieved a high agree rate, 99.73%, with NCBI mapping results on dbSNP. To further improve on the efficiency and reliability of this method, in this work, we used hits of fixed length (14 base pairs) genomic sequence patterns as a data cleaning indicator to assign each SNP record to its most likely chromosomes and delegate only those of a low hit ratio to be mapped against all the chromosomes. This simple prescreening method was shown to identify some erroneous assignments of NCBI.

2 Methods.

When a SNP record is completely aligned to a genomic fragment, both sequences should share common subsequences. If there are mutations or indels in the SNP record, its hit rate (number of identical subsequent fragments) should be much lower than records that can be completely aligned. Since we can find unique sequences in most of the SNP records, we can use the hit ratio of each chromosome to find each SNP record's chromosome quickly and reliably.

In this work, we used a 14-mer overlapping window to generate 14-mer sequence patterns of each chromosome, with 50 arrays to store the segment patterns from all positive and negative strands of chromosomes. After that, we scanned each SNP record by the same 14-mer overlapping window and keep a chromosome-specific pattern count when there is a hit. Subsequences that contain an alphabet other than ATCG, such as N, will be counted by not_ATCG counter. After scanning a SNP sequence, each chromosome's 14-mer hit ratio is calculated by the following equation:

$$\text{Hit ratio (\%)} = \text{Chromosome hits} * 100 / (\text{SNP sequence length} - \text{not\_ATCG count} - 13)$$

3 Results.

It takes 30 minutes to generate the 50 arrays that contain every chromosome's 14-mer patterns and another 4 hours to generate the hit ratio results for 3,365,561 SNP records (dbSNP build 115, the files named rs_chNotOn.fas, rs_chMulti.fas and rs_chMasked.fas were excluded). We use the

Institute of Biomedical Science, Academia Sinica, Taipei 116, Taiwan. E-mails: ouch@gate.sinica.edu.tw, mjhwang@ibms.sinica.edu.tw

following strategies to process the results: (1) Keep the chromosome identity for the SNP when the hit ratio is higher than 99%. (2) Keep the chromosome identity with the highest hit ratio (larger than 70%) when each hit ratio is lower than 99%. The results showed that 99.8% SNP records contain the same chromosome identity with the NCBI assignment. For SNP records satisfying either condition, the difference in the rate of the best and the second hit rate larger than 10% is 83.9%. Only 275 records (out of total 2,824,219 records) have chromosome assignments different from NCBI when the hit rate difference is larger than 10%. Checking with BLAST, BLAT, and SSAHA, we have found, for many of these records, the results of these three alignment based methods agreed with our pre-screening assignment and not with the assignment of NCBI.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990. Basic local alignment search tools. J. Mol. Biol. 215:403-410.

[2] Kitts, A. and Sherry, S. The SNP Database of Nucleotide Sequence Variation. The NCBI Handbook.

[3] Chen, Y.Y., Lu, S.H., Shih, S.C. and Hwang, M.J. 2002. Single Nucleotide Polymorphism Mapping Using Genome-Wide Unique Sequences. Genome Research 12-7:1106-1111.

B11. Efficient Method for Inferring Hierarchy of Clonal Complexes from Multi-Locus Sequence Types

Wasinee Rungsarityotin [1], Mark Achtman [2], Homayoun Bagheri-Chaichian [3], Alexander Schliep [4]

Keywords: genetic structure, molecular sequence types, evolution, graph algorithms

1 Introduction.

In 1998 multi-locus sequence typing (MLST) was proposed as a nucleotide sequence based approach that could be applied to many bacterial pathogens [4]. In brief, MLST consists of identifying specific loci on the genome that code for neutral (and hence conserved) house-keeping genes. For each locus, a fragment of approximately 500bp is sequenced, and each unique sequence is assigned an arbitrary allelic label. Hence, given m loci, each individual MLST entry consists of a vector $\mathbf{S}$ of length m (for example $m = 7$ for E. coli), whereby each vector component s_i is an integer corresponding to the allele number. An MLST data set consists of an ordered set of vectors of type $\mathbf{S}$. Each unique vector S is also given a label and referred to as a sequence-type or ST (eg. ST1). High-throughput sequencing technology facilitates large scale collection of MLST data and thus causes a need for a portable, reproducible, and scalable typing system that reflects the population and evolution of bacterial species. Perfect phylogeny may not work on MLST data because in practice there are not enough loci. Even if nucleotide sequences of isolates are available, it is still difficult to reconstruct a perfect phylogeny due to a high rate of recombination in bacterial pathogens [1, 5].

2 Algorithm.

Existing programs such as BURST [2], though simple to implement and visualize, do not provide an analytical method to infer relationship between groups of sequence types — clonal complexes. In this paper we examine an algorithm for finding k-way partitioning of a fully connected weight undirected graph which can be efficiently approximated with a generalized eigenvalues problem of the Laplacian matrix [3]. This methodology allows us to infer groups of clonal complexes, using only pairwise similarity. In brief, the algorithm is testing some objective functions for splitting, perform recursive bipartition of a vertex set of a graph. To decide which partition can be cut further, we order splits by their significance, considering the cost of the cut and number of clusters. The order in which the successive splits occur imposes a hierarchy of groupings, allowing to infer relations among complexes at varying levels of resolution. To confirm this hypothesis, more comparative result between simulation data and real data from bacterial species will appear in a future paper.

[1]Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63–73, D-14195 Berlin, Germany. E-mail: rungsari@molgen.mpg.de

[2]Max Planck Institute for Infection Biology, Schumann Straße 21/22, D-10117 Berlin, Germany. E-mail:achtman@mpiib-berlin.mpg.de

[3]Max Planck Institute for Infection Biology, Schumann Straße 21/22, D-10117 Berlin, Germany.

[4]Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63–73, D-14195 Berlin, Germany.

References

[1] Achtman, M. 2002. A phylogenetic perspective on molecular epidemiology. *Molecular medical microbiology* 1:485–509. London: Academic Press.

[2] The Multi Locus Sequence Typing website. http://www.mlst.net.

[3] Shi, J. and Malik, J. 2000. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8):888–905.

[4] Urwin, R. and Maiden, C.J. Martin. 2003. Multi-locus sequence typing: a tool for global epidemiology. *TRENDS in Microbiology*, 11(10):479–487, October, 2003.

[5] Wang, L., Zhang, K. and Zhang, L. 2001. Perfect phylogenetic networks with recombination. *J. of Comp. Biology*, 8:69-78, 2001.

B12. Description of Haplotypes and their Ancestry Structure from SNP Data

Jonathan Sheffi[1], Itsik Pe'er[2], David Altshuler[3], Mark J. Daly[4]

Keywords: haplotypes, perfect phylogeny, hidden Markov models, ancestral recombination graph

1 Background.

Understanding the patterns into which SNP alleles combine along the genome is one of the major challenges in contemporary genetics. Such a description of haplotype variation holds the promise for more powerful association studies. Many current approaches to this modeling problem devise a Hidden Markov Model (HMM) that steps along the chromosome, emitting a symbol (allele) at each polymorphic site. States of this HMM correspond to haplotype fragments that give rise to the observed samples, with transitions representing recombination of these fragments.
Distinct implementations of the HMM paradigm differ in the meaning of the model haplotypes, in the manner they are manifested as samples, and the assumption ob haplotype blocks. Standard software [4] uses a block-free model, in which model haplotypes are the actual haplotypes observed in the sample data. Block-wise HMMs that optimize the Minimum Description Length (MDL) of the data, limit the model haplotypes to be single-block fragments. Such models are becoming more and more complex, starting from a model that just seeks to list common haplotypes within each block [3], through a model that attempts to minimize entropy of inter-block transitions [1], to a model that also tries to assign a realistic meaning to emission probabilities as chances for mutation.
We devise a more general HMM-type approach for description of haplotype variation that tries to mimic real biological phenomena and entities by model components. Specifically, model haplotypes are explicitly parts of ancestral chromosomes. As time passed, these haplotypes underwent divergence, mutation, and recombination, all of which are explicit in our model. This expressive description aims at reconstructing the major lineages in the ancestral recombination graph, thus better capturing the statistical properties of the data, and better aiding in designing cost-effective genetic studies.

2 Model for Haplotype Variation.

In the simplest case, data are haploid, i.e., comprise of a set $O = \{ c_1, \ldots, c_n \}$ of chromosomes, each typed for each of $|S|$ bi-allelic SNPs. $X[c,s]$ is the observed bit at chromosome $c \in O$ for SNP $s \in S$. Each SNP is labeled by its chromosomal location $l[s]$. Observed chromosomes are time-labeled as contemporary, i.e. $\tau[c] = 0$ for $c \in O$. We allow also diploid (phased or unphased) or trio (partially phased) data.
The model includes a set of ancestral (hidden) chromosomes, or haplotypes, $H = \{ c'_1, \ldots, c'_n \}$. $X[c',s]$ is the ancestral bit at haplotype $c' \in H$ for each SNP s along the interval $[s_{\text{first}}(c')..s_{\text{last}}(c')] \subseteq S$

[1] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA USA. Whitehead Institute & Broad Institute, Cambridge, MA USA. E-mail: jsheffi@broad.mit.edu
[2] Whitehead Institute & Broad Institute, Cambridge, MA USA. E-mail: peer@broad.mit.edu
[3] Broad Institute, Cambridge, MA, USA & Massachusetts General Hospital, Boston, MA USA E-mail: altshuler@molbio.mgh.harvard.edu
[4] Whitehead Institute & Broad Institute, Cambridge MA USA. E-mail: mjdaly@broad.mit.edu

where c' is defined. Each ancestral chromosome $c' \in H$ is time-labeled in the past $\tau[c'] < 0$, and frequency-labeled $0 < f[c'] < 1$.
At each SNP s, all observed chromosome and haplotypes defined at s except one have a local parent haplotype $P_s(c)$, defined at s. $X[c,s]$ is identical to $X[P_s(c),s]$ with probability $1-\theta_{cs}$ where $\theta_{cs}=\mu_s(\tau[c]-\tau[P_s(c)])$, μ_j being a position-dependent mutation probability. This sets the emission probabilities.
If $P_s(c)$ is defined at $s+1$, then $P_s(c)=P_{s+1}(c)$, and with probability $1-r_{cs}$ where $r_{cs}=\rho_s(l_{s+1}-l_s)(\tau[c]-\tau[P_s(c)])$, ρ_s being a position-dependent recombination rate. With probability r_{cs}, $P_{s+1}(c)$ is randomly selected from the ancestral chromosomes that are defined for the current position based upon their frequency. If $P_s(c)$ is undefined at $s+1$, $P_{s+1}(c)$ is randomly selected according to a prescribed probability table.

3 Learning the Model.

Algorithmically, we optimize the likelihood of the bottom model layer only (only between samples and their parents). Parameters are estimated using standard EM for HMM inference, while HMM topology (boundaries between which each haplotype is defined) are inferred by a local search.
The haplotypes in the upper layers of the model are blockwise computed by perfect phylogeny reconstruction. They in turn, contribute to the next iteration of inference of the bottom layer by inducing a prior based on their phylogenetically-based MDL.

4 Advantages of the Model.

By explicitly assigning time and frequency labels to haplotypes, they acquire the meaning of ancestral haplotypes, and allow optimizing parameters for the real genetic processes, of mutation and recombination. Incorporation of local phylogenies to the MDL score realistically handles the process of divergence as well. Finally, our model is built over haplotypes that are defined along an arbitrary interval. These elements generalizes both blockwise and full-length models, overcoming the limitations of the formers, which ignore haplotype block structure, as well as the latter methods, which blindly rely on this structure.

References

[1] Anderson E.C. and Novembre J. 2003. Finding haplotype block boundaries by using the minimum-description-length principle. *American Journal of Human Genetics*. 73(2):336-354.

[2] Greenspan S. and Geiger D. 2003. Model-based inference of haplotype block variation. In: *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB 03)*, Berlin: ACM. pp. 131-137.

[3] Koivisto M., Perola M., Varilo T., Hennah W., Ekelund J., Lukk M., Peltonen L., Ukkonen E. and Mannila H. An MDL method for finding haplotype blocks and for estimating the strength of haplotype block boundaries. *Proceedings of the Eighth Pacific Symposium on Biocomputing (PSB '03)*, Lihue, Hawaii pp.:502-513.

[4] Stephens M., Smith N.J. and Donnelly P. 2001.A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*. 68(4):978-989.

B13. FROM RESOURCE TO RESEARCH: MGI AND GO

Mary E. Dolan, Joel E. Richardson, Janan T. Eppig, Martin Ringwald, Carol J. Bult, James A. Kadin, Judith A. Blake[1]

Keywords: Gene Ontology, computational genomics

The Mouse Genome Informatics (MGI) (http://www.informatics.jax.org) system provides a comprehensive public resource about the laboratory mouse that integrates information on sequences, maps, genes and gene families, expression, alleles and variants, strains and mutant phenotypes. The integration of such diverse data depends upon quality determinations of object identities and relationships and upon the use of defined, structured vocabularies (ontologies).

The Gene Ontology (GO) Project provides structured, controlled vocabularies in the domain of molecular biology, which have been incorporated in various bioinformatics resources to aid biological annotations. The GO project model has been extended for the development of other ontologies and has fostered standardization among model organism database systems. MGI also incorporates several different classification schemes that enable a variety of query capabilities for our users. The Mouse Anatomical Dictionary and Phenotype Classifications provide the mechanism for annotation of aspects of gene expression, QTL analysis, and incorporation of information about experimental mouse mutants.

The challenge for ontology developers is to construct fully documented, easily maintained ontologies that are accessible to the larger scientific community. Moreover, an important goal is to develop a semantic framework that not only adds meaning to the data but also, with the hierarchical structure of the vocabulary, can be used to create new information through inference. For example, the GO structure is a directed acyclic graph (DAG) in which each annotation node can have one or more child nodes and must have one or more parent nodes. Gene products are annotated at varying levels of detail based on experimental and computational evidence. Due to inference from the structure of the GO vocabulary, a gene product annotated to a finer level of detail is also annotated to any ancestor (coarser) level as well.

We have developed and adapted a number of software tools to facilitate the use of MGI resources in a GO context and take advantage of the GO structure to use this valuable annotation resource as a research tool. A MGI-GO browser provides tree views and links to annotated data sets. Thus experimentally annotated sets of genes can be analyzed by function, process or component via their GO representations. We also provide several MGI GO tools, which permit a user to explore what any set of annotated genes, for example, a cluster of up-regulated genes, have in common. The MGI GO TermFinder tool compares annotations of the gene set with overall MGI gene annotations to find statistically significant overrepresentations of GO classifications. Similarly, the MGI GO_Slim Chart tool compares the distribution of the gene set in a predefined classification scheme with the overall MGI gene distribution.

MGI is funded by grants from NIH/NHGRI, NIH/NICH and NCI. The GO project is funded by NIH/NHGRI and by the European Union RTD program.

[1] Mouse Genome Informatics, The Jackson Laboratory, 600 Main Street, Bar Harbor, ME 04609 USA. E-mail: mdolan@informatics.jax.org, jer@informatics.jax.org, jte@informatics.jax.org, ringwald@informatics.jax.org, cjb@informatics.jax.org, jak@informatics.jax.org, jblake@informatics.jax.org

B14. A Pattern Discovery-Based Method for Detecting Multi-Locus Genetic Association

Zhong Li[1], Aris Floratos[1], David Wang[1], Andrea Califano[2]

Keywords: pattern discovery, multi-locus, genetic association, power calculation

1 Introduction.

Methods to effectively detecting global multi-locus genetic associations are becoming increasingly relevant in the genetic dissection of complex trait. Current approaches typically only consider a limited number of hypotheses, most of which are related to the effect of a single locus or of a relatively small number of loci co-localized on a chromosomal region, therefore do not accommodate the full range of genetic mechanisms that may contribute to a complex trait. We have developed a novel association analysis methodology (enGENIOUS) that is specifically designed to detect genetic associations involving multiple disease-susceptibility loci. Our approach relies on the efficient discovery of patterns comprising spatially unrestricted polymorphic markers and on the use of appropriate test statistics to evaluate pattern-trait association. Because markers within a pattern are not required to be in linkage (or even on the same chromosome), this method represents a truly global multi-locus association test. Power calculations using multi-locus disease models confirm the superior performance of out approach when compared to a frequency-based single marker analysis method. When applied on a real dataset, our method was successful in localizing a previously verified two-locus/gene interaction associated with Schizophrenia. In addition, we also identified a novel and less conspicuous association involving different markers on the same two genes, suggesting a role for genetic heterogeneity and population substructure in Schizophrenia.

2 Results.

enGENIOUS demonstrated better power than a single marker association test

A simulation-based power calculation was performed to evaluate the power of the pattern discovery-based method on multi-locus association analysis. Power was computed for two disease models: dominant-recessive and recessive-recessive, each involving two loci (markers M1 and M2). For each disease model, we considered five genotype frequency settings for disease-affected markers and two sample sizes (250 cases/250 controls or 500 cases/500 controls) (Figure 1). For each combination of disease model, frequency setting, and sample size, 500 simulated datasets were generated. As shown in Figure 1, the pattern discovery-based method consistently outperformed the single marker χ^2 test under both disease models (Figure 1A and 1B). As expected, power for both tests improved with larger sample size and higher genotype frequency. The power difference between these two methods demonstrates the advantage of using the pattern discovery-based method on association detection, especially when affected genotype frequency is low. Because no constraint has been established on the physical distance between markers M1 and M2, the same results would have been obtained even if M1 and M2 were on different chromosomes, thus providing direct evidence that the pattern discovery-based method is indeed a global association test.

Discovery of a novel gene-gene interaction in Schizophrenia patients

enGENIOUS was applied to a dataset collected from a case/control association study on Schizophrenia. The dataset contained genotypes for 28 SNP markers spanning 115 kb at 12q24 (co-localized with the DAO gene) and 266 kb at 13q34 (co-localized with the G72 gene). With the

[1] First Genetic Trust Inc., 201 Route 17 North, Suite 902, Rutherford, NJ 07070, USA. E-mail: zli@firstgenetic.net

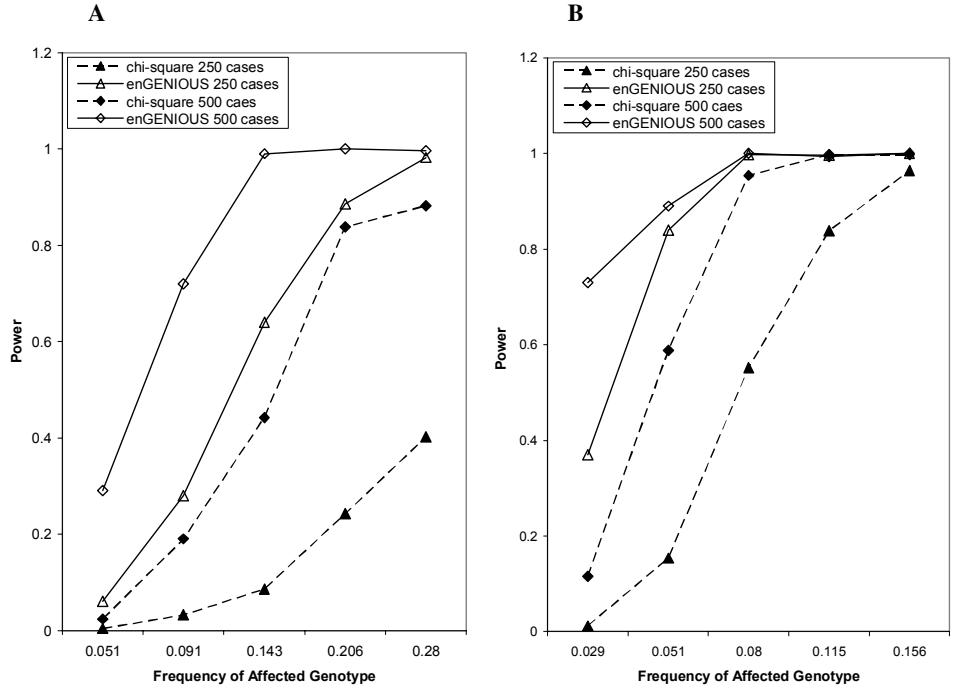
[2] Department of Biomedical Informatics, Columbia University, 1150 St. Nicholas Ave., New York, 10032, USA. E-mail: califano@dbmi.columbia.edu

association test rejection level for individual pattern as 9.49e-5 at the support constraint of 24, only two out of 5,952 identified patterns were found to be significantly associated with Schizophrenia. The most significant pattern (P=4.64e-5) included two markers, marker MDAAO-6 on chromosome 12, and M-22 on chromosome 13, the same two markers found to be associated with Schizophrenia in both genetic analysis and bioassay [1]. The second most significant pattern (P=7.33e-5) included seven markers, one marker (marker B-7) on chromosome 12, and markers B-1, B-2, B-3, B-4, B-5, and B-6 (B-1~B-6) on chromosome 13. None of the markers in this pattern was significant by itself when evaluated with the single-marker χ^2 test. To calculate the combined relative risk on DAO and G72 loci as characterized by Pattern A or Pattern B for Schizophrenia, we constructed a χ^2 test in which individuals carrying either of the two patterns or none of the two patterns were counted in cases and in controls. A highly significant P value (P=8.67E-11) indicated that the genetic interaction between DAO and G72 is strongly associated with Schizophrenia and multiple molecular mechanisms might be responsible for the susceptibility on those two loci.

3 Figures and tables.

Figure 1 Power Comparisons between a Single Marker χ^2 test vs. enGENIOUS.

Powers were compared between a single marker χ^2 test and the pattern discovery-based multi-locus association test under two disease models (dominant-recessive in A, recessive-recessive in B). A and B showed the power curves for both methods with two sample sizes (250 cases/250 controls and 500 cases/500 controls).



References

[1] Chumakov I, Blumenfeld M, Guerassimenko O, Cavarec L, Palicio M, Abderrahim H, Bougueleret L, et al. (2002) Genetic and physiological data implicating the new human gene G72 and the gene for D-amino acid oxidase in schizophrenia. *Proc Natl Acad Sci U S A* 99:13675-13680.

B15. Algebraic Statistical Genetics
Affected Sib-Pair Linkage Analysis

Ingileif Bryndís Hallgrímsdóttir [1]

Keywords: Algebraic statistics, linkage analysis, affected sib-pairs, IBD probabilities.

Linkage analysis, or gene mapping, is concerned with finding the chromosomal location of disease genes. The genes for hundreds of Mendelian (one gene) disorders have already been successfully mapped. Most common diseases are however not caused by one gene and the challenge today is to solve *complex diseases*, diseases that are caused by many, possibly interacting, genes and environmental factors.

The emerging field of algebraic statistics consists of the application of algebra and geometry to statistical models. We use algebraic methods to study the models used in linkage analysis, hoping to gain new insights that will provide better understanding and interpretation of existing statistical tests, and help develop new tests for two-locus linkage analysis. However the first step is to describe the model in this new language, and that is what we are concerned with here.

The central observation of the emerging field of algebraic statistics is that many statistical models are algebraic varieties, since the expressions for many discrete distributions are polynomials in the parameters of the model. Although usually not presented as such, the expressions for the identity by descent (IBD) probabilities, z_0, z_1, and z_2 are polynomials in the parameters of the model. Here z_i is the probability that an affected sib-pair shares i alleles IBD and the parameters of the model are the frequency, p, of the disease gene (allele) in the population and the penetrances f_0, f_1 and f_2, where f_i is the conditional probability for an individual of getting the disease given that (s)he carries i copies of the disease gene.

Using the tools of computational algebra we have derived an invariant describing the one-locus model for affected sib-pairs. The invariant is a polynomial expression in the parameters of the model and the IBD probabilities z_0, z_1 and z_2, it has 13 terms and is a quartic in the z's. For any triplet z_0, z_1, z_2 in the model the invariant vanishes. The first two terms are given below:

$$(\text{-64 f}_0 f_1^5 f_2^2 - 32 f_0^2 f_1^6 + 64 f_0 f_1^6 f_2 - 16 f_1^4 f_2^4 + 64 f_0^3 f_1^5 - 64 f_0^2 f_1 f_2^5 - 64 f_0^5 f_1 f_2^2 - 32 f_0^5 f_2^3 - 16 f_0^4 f_1^4 + 64 f_0^3 f_1 f_2^4 + 32 f_0^4 f_2^4 - 32 f_0^3 f_2^5 + 64 f_0^4 f_1 f_2^3 + 16 f_0^2 f_2^6 - 32 f_1^6 f_2^2 + 64 f_1^5 f_2^3 + 16 f_0^6 f_2^2 + 160 f_0^4 f_1^2 f_2^2 + 160 f_0^2 f_1^2 f_2^4 - 96 f_0^3 f_1^4 f_2 + 224 f_0^2 f_1^4 f_2^2 - 64 f_0^2 f_1^5 f_2 - 96 f_0 f_1^4 f_2^3 - 320 f_0^3 f_1^2 f_2^3) \cdot z_0^3 z_2 +$$

$$(\text{-16 f}_1^5 f_2^3 + 4 f_1^4 f_2^4 + 32 f_0^3 f_1^4 f_2 + 32 f_0 f_1^7 - 16 f_1^8 + 32 f_1^7 f_2 + 16 f_0^5 f_1 f_2^2 + 4 f_0^4 f_1^4 - 4 f_0^2 f_2^6 + 16 f_0^3 f_1 f_2^4 - 8 f_0^4 f_2^4 + 32 f_0 f_1^4 f_2^3 - 32 f_0^2 f_1^3 f_2^3 - 16 f_0^3 f_1^5 + 16 f_0^2 f_1 f_2^5 - 32 f_0^3 f_1^3 f_2^2 - 16 f_0^2 f_1^5 f_2 + 88 f_0^2 f_1^4 f_2^2 - 32 f_0^2 f_1^2 f_2^4 - 16 f_0 f_1^5 f_2^2 - 32 f_0^4 f_1^2 f_2^2 - 4 f_0^6 f_2^2 + 16 f_0^4 f_1 f_2^3 - 64 f_0 f_1^6 f_2) \cdot z_0^2 z_1^2 + ...$$

The term invariant is not used in the literature on linkage analysis and we have borrowed it from phylogenetics. If f_0, f_1, f_2 are non-decreasing the invariant given above vanishes on all points that lie within the possible triangle [1]. However the invariant can be specialized to certain models, e.g. for the strictly recessive model the penetrances are $f_0 = 0$, $f_1 = 0$ and $f_2 = f$, where $f \in [0, 1]$ and we recover the classical invariant:

[1]Department of Statistics, University of California, Berkeley. E-mail: ingileif@stat.berkeley.edu

$$z_1^2 - z_0 z_2$$

,

known as the Hardy-Weinberg curve. For an additive model $f_0 = 0$, $f_1 = f/2$ and $f_2 = f$ and the invariant becomes:

$$z_0 - z_1 + z_2$$

which is equivalent to $z_1 = 1/2$ since $z_0 + z_1 + z_2 = 1$. These two invariants can be easily derived from the equations for z_0, z_1 and z_2. The invariant that describes a strictly dominant model is more complicated and is presented here:

$$z_0 z_1^2 + 4z_0^2 z_2 - 8z_0 z_1 z_2 + 4z_0 z_2^2 + 4z_1 z_2^2 - 4z_2^3$$

.

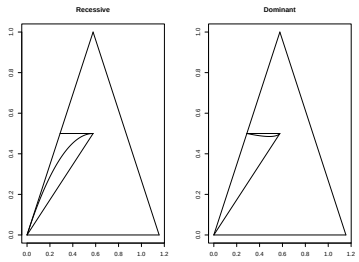


Figure 1: Holmans' triangle. The larger triangle is the probability simplex, $z_0 + z_1 + z_2 = 1$ and the smaller triangle is the possible triangle for affected sib-pair IBD probabilities. The curves on the left shows the recessive model and the one on the right the dominant model.

The invariants can be used to determine which disease model fits the data best, if we plug the proportions of 0, 1 and 2 IBD sharing of affected sib-pairs into a model specific invariant the invariant will vanish if the data fits the model (confirmed with simulations). Similar invariants can be derived for the two-locus case, although, one obtains not one but many invariants. Invariants for certain two-locus disease models have been derived and up to 30 invariants are needed to describe a model. We do not yet have a list of general invariants for the two-locus case.

An attractive aspect of the algebraic point of view, is that the methodology provides the natural setting for generalizing a number of well-known linkage results. Examples of this are Holmans triangle [1], which admits natural higher dimensional analogues. Also, known invariants for the simple disease models in the one locus case are easily recovered from the general formalism.

References

[1] Peter Holmans, 1993. Asymptotic Properties of Affected-Sib-Pair Linkage Analysis. *American Journal of Human Genetics* 52:362-374.

C1. BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria

Frode S. Berven[1], Kristian Flikka[2]. Harald B. Jensen[1], Ingvar Eidhammer[3]

Keywords: Beta-barrel protein prediction, functional genomics, proteomics, outer membrane

1 Introduction.

The integral OM proteins (OMPs) of Gram-negative bacteria generally consist of β-structures, and form monomeric, dimeric or trimeric transmembrane (TM) β-barrels containing between eight and 22 TM β-strands [1]. These proteins have proven difficult to predict mainly due to very short TM stretches of amino acids with highly variable properties [2]. Recent publications have described different approaches to recognize β-barrel OMPs from polypeptide sequences [3-5], but none of the programs have been made available for public use. This work describes the development of a program that predicts whether or not a polypeptide sequence from a Gram-negative bacterium is an integral β-barrel outer membrane protein. The precision of the predictions was found to be 91% with a recall of 88% when tested on the proteins with SwissProt annotated subcellular localisation in *Escherichia coli* K 12 (788 sequences) and *Salmonella typhimurium* (366 sequences). BOMP is available at http://www.bioinfo.no/tools/bomp.

2 Program components

The program, called the β-barrel Outer Membrane protein Predictor (BOMP), is based on two separate components to recognize integral β-barrel proteins. The first component is a C-terminal pattern typical for many integral β-barrel proteins. The second component calculates an integral β-

[1]Department of Molecular Biology, ; University of Bergen, 5020 Bergen, Norway. E-mail: Frodeb@ii.uib.no
[2]Computational Biology Unit, Centre for Computational, University of Bergen, 5020 Bergen, Norway. E-mail: Flikka@ii.uib.no
[3]Department of Informatics, University of Bergen, 5020 Bergen, Norway. E-mail: Ingvar.Eidhammer@ii.uib.no

barrel score of the sequence based on to which extent the sequence contains stretches of amino acids typical for transmembrane β-strands. In order to limit the number of wrongly predicted integral OMPs (false positives), we developed a final filtering procedure based on a reference set containing polypeptide sequences with known subcellular localization. The relative abundance of Asparagine and Isoleucine was used as a discriminator between the integral OMPs and the sequences with other localization. When an unknown protein is run through the filter, it is compared to the reference set by using a k-nearest-neighbor method with k=5 [7] to determine if the candidate is a true integral OMP. The input sequence is considered to be an integral OMP when having at least three integral OMPs as the nearest neighbors. As a supplement to the prediction methods outlined above, we added the possibility to include an automated BLAST search to be performed on the input sequence. The input sequence is used in a BLAST search against a database containing 10618 Gram-Negative polypeptide sequences with given subcellular localization in SwissProt release 42, in order to find the highest scoring alignment with an E-value above 10 e^{-10} and a length of between 80 and 120% of the input sequence [8]. The localization of the best database hit will either support or contradict the result from the prediction part of BOMP, and provide additional information about the input sequence to the user.

3 References

1. Tamm, K.L., A. Arora, and H.J. Kleinschmidt, *Structure and Assemly of beta-barrel membrane proteins.* The Journal of Biological Chemistry, 2001. **276**: p. 32399-32402.
2. Koebnik, R., K.P. Locher, and P. Van Gelder, *Structure and function of bacterial outer membrane proteins: barrels in a nutshell.* Mol Microbiol, 2000. **37**(2): p. 239-253.
3. Casadio, R., et al., *Fishing new proteins in the twilight zone of genomes: The test case of outer membrane proteins in Escherichia coli K12, Escherichia coli O157:H7, and other Gram-negative bacteria.* Protein science, 2003. **12**: p. 1158-1168.
4. Wimley, C.W., *Towards genomic identification of beta-barrel membrane proteins: Composition and architecture of known structures.* Protein science, 2002. **11**: p. 301-312.
5. Zhai, Y. and H.M.J. Saier, *The beta-barrel finder (BBF) program, allowing identification of outer membrane beta-barrel proteins encoded within procaryotic genomes.* Protein science, 2002. **11**: p. 2196-2207.
6. O'Connel, M.J., *Search Program for Significant Variables.* Computer Physics Communications, 1974. **8**: p. 49-55.
7. Ripley, B.D., *Pattern Recognition and Neural Networks.* Cambridge University Press, 1996.
8. Gardy, L.J., et al., *PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria.* Nucleic Acid Research, 2003. **31**(13): p. 3613-3617.

C2. Motif Finding and Multiple Alignment through Vector-Space Embedding of Protein Sequences

Arnab Bhattacharya, Tamer Kahveci, Ambuj K. Singh
Department of Computer Science
University of California, Santa Barbara, CA 93106
{*arnab,tamer,ambuj*}*@cs.ucsb.edu*

Keywords: motif, multiple alignment, vector space, dependency graph

1 Algorithm

We introduce a vector-space embedding of protein sequences which will allow us to find the motifs in a set of proteins. Our method can also be used for the multiple alignment of more than two proteins. It is superior to the existing methods that depend on the order of proteins since we consider all the proteins at once.
Our motif finding method consists of the following steps:

1. Protein subsequences are mapped to points in a multi-dimensional space.
2. Spatially tight clusters of these points are found such that most or all of these proteins are represented in each cluster.
3. A *dependency graph* is constructed with the clusters as vertices and directed edges between the *non-conflicting* vertices.
4. The longest path in the graph is chosen as the motif.

Our multiple alignment algorithm adds another step:

5. Use the motif found in step 4 as the backbone of the alignment and recursively invoke the motif finding algorithm for each unaligned region.

We will now explain each of these steps in more detail.
Step 1: A window of length w is slid along the protein sequence. Each positioning of the window produces a subsequence of w residues. A *score vector* is computed for each such subsequence as the concatenation of the score vectors of each residue. Each row of a score matrix is considered to be the score vector of the amino acid for that row. A score vector maps to a point in a multi-dimensional space. A protein of length n will thus have $n - w + 1$ points in the vector space. We typically choose $w = 3$. We refer the reader to [1] for further details on vector space embedding.
Step 2: All the clusters of points in the vector space are identified. A cluster is defined as a set of points (at most one from each protein) which are within a radius of r from each other and which represents at least $p\,\%$ of the total number of proteins. Here, r and p are the *distance* and the *membership* (the percentage of proteins in the cluster) thresholds respectively. Typically, $r =$ 1-2 % of the dimensions of the search space and $p =$ 80-100 %.
Step 3: A directed dependency graph is built on the clusters as follows. Each cluster is considered to be a vertex in the graph. A directed edge from vertex i to vertex j is added if all of the protein residue positions in i are strictly less than those in j. Such vertex pairs are called non-conflicting. A weight is assigned to each vertex based on how tight it is. A vertex (i.e., a cluster) with less inter-point distances gets a larger weight. Also, a vertex with a higher membership gets a larger weight. A weight is assigned to each edge as well. Each edge corresponds to a pair of points from each protein. The edges get a higher weight if the differences between the residue positions of each pair of points are 1) small and 2) not much varied. These conditions demote the number of gaps.

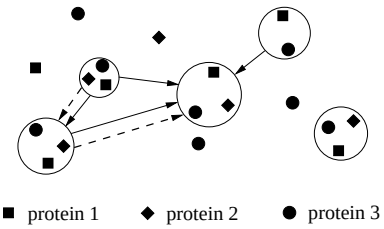


Figure 1: An illustration of the points of three proteins in a two-dimensional space. The circles show the clusters (nodes of the dependency graph). The arrows show the edges of the dependency graph. The dashed arrows show the largest weighted path.

Table 1: The motifs that we find for five proteins from the *FMN-linked oxidoreductases* superfamily for $w = 3$. The bold letters show the backbone. The residue positions for the motifs are (7, 141, 220, 270, 288) for `1al7:-`, (81, 225, 284, 346, 348) for `1d3g:A`, (9, 55, 108, 116, 118) for `1huv:A`, (7, 97, 131, 178, 184) for `1icp:A`, and (113, 184, 252, 336, 365) for `1lco:B`.

PDB id	Motifs											
1al7:-	···	V**N**E	···	L**V**R	···	L**Q**T	···	L**E**E	···	G**V**R	···
1d3g:A	···	Y**K**M	···	L**V**K	···	L**S**T	···	L**E**A		**L**L	···
1huv:A	···	V**E**D	···	L**V**D	···	L**S**T	···	I**E**D		**L**A	···
1icp:A	···	V**E**E	···	I**V**D	···	I**S**C	···	I**E**A	···	G**V**E	···
1lco:B	···	L**K**S	···	M**L**G	···	Y**Q**L	···	I**E**E	···	G**V**S	···

Step 4: Once the directed graph is built, each path on the dependency graph defines a motif. This is because each vertex corresponds to a set of matching residues from the proteins, and the directed edges of a path ensures that these matching residues do not conflict. We find the largest weighted path in the graph. The weight of a path is defined as the sum of the weights of its vertices and edges. Figure 1 illustrates the algorithm developed so far in two-dimensions.
Step 5: The motif found in Step 4 defines the backbone of the multiple alignment. We partition the sequences by clipping the proteins from the residues in the motif. This produces sets of subsequences whose end points are the two consecutive motif residues. Each of these sets are then realigned using Steps 1 to 4 recursively until the length of the subsequences drop below a threshold. These subsequences are then aligned using multiple alignment.

2 Results

In order to demonstrate the effectiveness of our method, we ran it on a number of proteins. Table 1 shows the motifs found for the proteins `1al7:-`, `1d3g:A`, `1huv:A`, `1icp:A` and `1lco:B` of the *FMN-linked oxidoreductases* superfamily for $w = 3$. In this example, we find five motifs which are shown in bold letters. The letters next to the bold ones are also similar with a high probability since they are in the same window as the bold letters. For multiple alignment, the subsequences between consecutive bold letters are aligned similarly.

References

[1] A.Bhattacharya, T. Can, T. Kahveci, A.K. Singh, and Y.-F. Wang. ProGreSS: Simultaneous Searching of Protein Databases by Sequence and Structure. In *PSB*, pages 264–275, 2004.

C3. Long-Duration Molecular Dynamics Simulation on Constructed Nacrein Structure

Frank Chang[1]*, Samson Cheung[2], Ming Wong[1], Cathy Bitler[3], Andrew Palma[4]

Keywords: nacrein, molecular dynamics, biomineralization

1 Introduction

Nacrein (pearl protein) is a water soluble protein and present in the mantle tissue of some mollusc species. Its function is believed to regulate aragonite or calcite polymorphism during biomineralization as in the processes of pearl formation [Miyamoto 1996]. Primary structural analysis of nacrein shows high homology to carbonic anhydrase II (CAII), a protein that regulates intracellular pH homeostasis during osteoclasts bone resorption [Teitelbaum 2000]. A distinct difference between nacrein and CAII is that nacrein has a long glycine-rich repeated sequence between its two carbonic anhydrase domains. The repeated sequence is homologous to that found in proteins of the collagen family.

The role of nacrein in biomineralization has been hypothesized [Miyamoto 1996]; however, its tertiary structure has yet to be determined. We applied long-duration molecular dynamics (MD) simulation to construct the tertiary structure of nacrein protein from *Pinctada fucata* (Japanese pearl oyster). The constructed nacrein structure shows that spatial active sites are similar to CAII (Figure 1 and Figure 2).

We are also interested in identifying the catalytic sites and calcium binding sites (EF-hand like motif) in the nacrein protein. To further investigate this nacrein protein, we will apply MD in an aqueous environment, and this simulated tertiary structure would be the foundation for an *in silico* approach in developing nacrein applications in biomaterial, bone morphogenesis, and pearl formation.

2 Methods

The primary sequence of nacrein was obtained from an earlier report [Miyamoto 1996]. X-ray diffraction structure of carbonic anhydrase II (PDB ID 5CAC) was obtained from the Protein Data Bank. The tertiary structure of nacrein was constructed by a comparative modeling method using ROSETTA [Alm 1999].

Molecular dynamics simulation of both tertiary protein structures were carried out with the NAMD version 2.5b2 and VMD 1.8.1 [Laxmikant 1999 and Humphrey 1999] on Origin 2000 and SGI Altix supercomputers at NAS facility in NASA [NASA-NAS]. Hydrogen atoms were added by using the guesscoord command in NAMD. Force field parameters were adopted from CHARMM22. The van

[1] Changene Inc. NASA Research Park, Moffett Field, CA 94035. * To whom correspondence should be addressed. E-mail: ftchang@changene.com
[2] NASA Advanced Supercomputing Division, Ames Research Center, Moffett Field, CA 94035. E-mail: cheung@nas.arc.nasa.gov
[3] Ecco Biotech, Menlo Park, CA 94025
[4] MDL Information System Inc. San Leandro, CA 94577, E-mail: apalma@mdli.com

der Waals interactions were cut off at 12 Å and 10^7 steps (10 nanoseconds) of energy minimization were performed at 310 K. The nonbonded interaction was recorded every 1000 time steps. The trajectories at given conditions were considered to be the probable structures and were analyzed in details.

3 Results

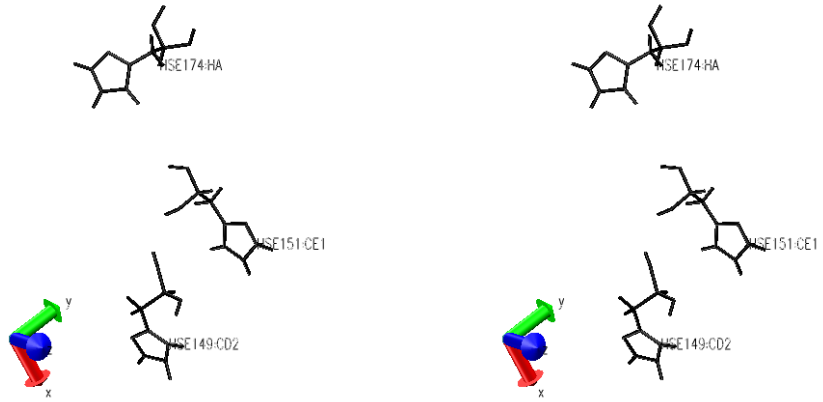


Figure 1. Tertiary structure of nacrein is constructed and applied with MD (entire structure not shown). Nacrein has three domains; Domain 1 (residue 1 – 235) and Domain 3 (residue 314 – 447) are similar to CAII, and Domain 2 (residue 236 – 313) has similarities is similar to collagen family proteins. The figure above is a stereo view of Domain 1 active site in the constructed nacrein structure. The residues shown in the Domain 1 active site are HIS-149, HIS-151, and HIS-174.

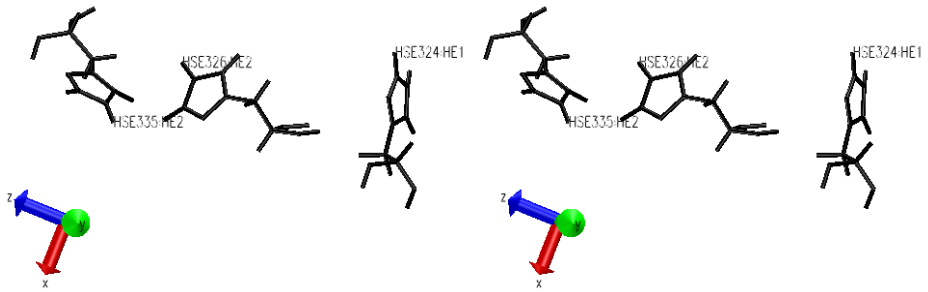


Figure 2. Stereo view of Domain 3 active site in constructed nacrein structure. The residues shown in the Domain 3 active site are HIS-324, HIS-326, and HIS-335. Note: variation of coordinates in Figure 1 and Figure 2 is to enhance visualization of the active sties.

4 References

[1] Alm, E. *et al*. 1999. *Proceedings of the National Academy of Sciences USA* 96:11305-11310.
[2] Humphrey, W. *et al*. 1999. *Journal Molecular Graphics*, 14:33-38.
[3] Laxmikant, K. *et al*. 1999. *Journal of Computational Physics*, 151:283-312.
[4] Miyamoto, H. *et al*. 1996. *Proceedings of the National Academy of Sciences USA* 93:9657-9660.
[5] NASA-NAS: http://www.nas.nasa.gov
[6] Teitelbaum, S. 2000. *Science*, 289:1504-1508.

C4. Gene Finding with Proteomic Data

Kristen K. Dang[1], Edward J. Collins[2], Michael C. Giddings[3]

Keywords: proteomics, mass spectrometry, genome annotation, *Francisella tularensis*

1 Introduction and Methods.

The sequencing technologies developed during the Human Genome Project and since have been successful in speeding the process of genomic sequencing. As of January 2004, NCBI lists 149 completed microbial genomes, as well human, mouse, rat and several other genomes. Annotation methods have also made major advances, but are still a limiting step in genome research. Additionally, current methods still have some significant limitations, especially in their ability to identify genes in diverse organisms. Many leading gene-finding methods use Hidden Markov Models or other machine-learning techniques that require training on a set of known data. These methods are suitable for phylogenetically related organisms, but their success rate on diverse organisms is likely to be lower due to the differences in the genomes of the training and test data. Additionally, genes produced by non-standard splicing are likely to be missed by annotation methods that rely on known genes.

A method of gene finding that is faster and more reliable for diverse organisms than current methods would allow researchers to use unfinished sequence, without waiting for annotation. We have previously reported a method for identifying genomic origins of proteins using only peptide mass fingerprint data as input [1]. In the current work, we used this method to identify and characterize protein samples from the bacterium *Francisella tularensis*, for which the genome is only partially sequenced and annotated [2]. The microbe is also lacks well-studied close phylogenetic relatives [3]. The ability to identify genes directly from proteomic data, without prior computational interpretation or annotation, provides a complementary tool to cDNA sequencing in the hunt for genes.

We analyzed 38 protein samples from *F. tularensis* using 2D SDS-PAGE, tryptic digestion, and MALDI-TOF-TOF mass spectrometry. Resulting mass lists were submitted to our Genome-Fingerprint Scanning method, which identifies a genome locus from which the observed spectrum likely originated. The method performs an *in silico* simulation of a tryptic digestion, then computes peptide weights for the resulting fragments and matches them against the observed mass list. If many matching weights are found within close genomic proximity of each other, that region becomes a candidate hit. The method also identifies ORFs that completely or partially encompass the hit region. A multi-part scoring method that considers number of hits in a window, adjacency of peptides, number of missed trypsin cleavage sites, number of in-frame stop codons, and duplicate mass matches is used to evaluate hit quality. The genome hits were searched against the NCBI microbial database using BLAST for a functional characterization.

2 Results.

[1] Department of Biomedical Engineering, University of North Carolina at Chapel Hill. E-mail: kamerath@email.unc.edu

[2] Department of Microbiology and Immunology, University of North Carolina at Chapel Hill. E-mail: edward_collins@med.unc.edu

[3] Departments of Microbiology and Immunology and Biomedical Engineering, University of North Carolina at Chapel Hill. E-mail: giddings@unc.edu

The GFS software produced well-scoring in-ORF hits for 29 of the 38 samples, some of which were corroborated by tandem-MS data (not shown). Tandem MS data was not available for all samples, and the scoring algorithm for such data in GFS is still in development. Software analysis was conducted at mass tolerances of 100 and 200 ppm; results from both parameters tended to corroborate the results of the other.

Number of samples	38
Samples with strong genome locus hits	29
Top hits with encompassing ORFs	30
ORFs with strong BLAST hits	30

Table 1: Results summary for samples submitted to GFS.

Based on these encouraging preliminary results, it appears that GFS can be useful in proteomics projects where only a preliminary sequence or annotation is available. Several refinements are under development for GFS, including probability-based scoring, faster digestion methods, and more efficient computing distribution for faster results. GFS will be available to the public for use at http://gfs.unc.edu.

3 References.

[1] Giddings, M.C., Shah, A.A., Gesteland, R., and Moore, B. 2003. Genome-based peptide fingerprint scanning. *PNAS* 100: 20-25.

[2] Prior, R.G. et al. .2001. Preliminary analysis and annotation of the partial genome sequence of *Francisella tularensis* strain Schu 4. *Journal of Applied Microbiology* 91: 614-620.

[3] Titball, R.W., Johansson, A., Forsman, M. 2003. Will the enigma of *Francisella tularensis* virulence soon be solved? *Trends in Microbiology* 11: 118-123.

C5. A Unified Representation of Multi-Protein Complex Data for Modeling Interaction Networks

Chris Ding,[1] Xiaofeng He,[1] Richard F. Meraz,[1] Stephen R. Holbrook[1]

Keywords: protein interactions, supercomplex, bipartite graph, MinMaxCut clustering

The protein interaction network presents one perspective for understanding cellular processes. Recent experiments employing high-throughput mass spectrometric characterizations have resulted in large datasets of physiologically relevant multi-protein complexes. We present a unified representation of such datasets based on an underlying bipartite graph model that is an advance on existing models of the network. Our unified representation allows for weighting of connections between proteins shared in more than one complex as well as addressing the higher level of organization that occurs when the network is viewed as consisting of protein complexes that share components. This representation also allows for the application of the rigorous MinMaxCut graph clustering algorithm for the determination of relevant protein modules in the networks. Statistically significant annotations of clusters in the protein-protein and complex-complex network using terms from the Gene Ontology suggest that this method will be useful for posing hypothesis about uncharacterized components of protein complexes or uncharacterized relationships between protein complexes.

Protein Complex Data Modeled as a Bipartite Graph

A bipartite graph has two types of nodes: p-nodes that denote proteins and c-nodes that denote protein complexes. A protein complex (c-node) connects to each of its constituent proteins (p-nodes). A bipartite graph is specified by its adjacency matrix $B = (b_{ij})$ where

$$b_{ij} = \begin{cases} 1 & \text{if protein } p_i \text{ is in protein complex } c_j \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Thus, a protein complex is represented by a column in B and a protein is represented by a row in B. We call the relation between proteins and complexes, represented by the bipartite graph, as the p-c network. Starting from the p-c network, we can naturally obtain the following two networks.

Protein - Protein Interactions (p-p network)

Our unified representation goes beyond the conventional uniformly-weighted protein interactions. The interaction strength between two proteins p_i, p_j is

$$(BB^T)_{ij} = \#\text{ of protein complexes containing both } p_i \text{ and } p_j \quad (2)$$

$(BB^T)_{ii} = \sum_j b_{ij}$ = number of protein complexes containing protein p_i, called the weight of p_i.

Protein Complex - Protein Complex Associations (c-c network)

The interaction strength between two protein complexes c_i, c_j is

$$(B^T B)_{ij} = \#\text{ of proteins shared by } c_i \text{ and } c_j \quad (3)$$

$(B^T B)_{jj} = \sum_i b_{ij}$ = number of proteins contained in c_j, called the weight of c_j.

[1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720. E-mail: {chqding, xhe,rfmeraz,srholbrook}@lbl.gov

The p-c network B, the p-p network BB^T and the c-c network B^TB are the three main components of the unified representation framework.

MinMaxCut Clustering Result Analysis

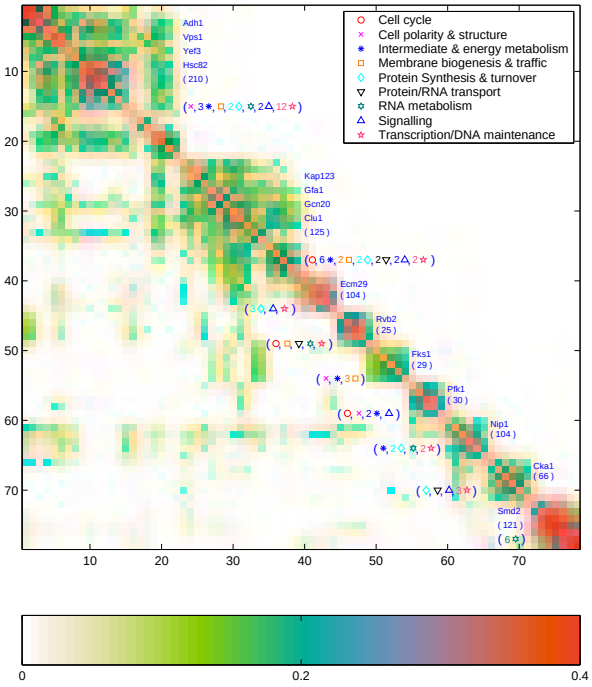


Figure 1: Predicted protein supercomplexes (clusters of the c-c network). Color represents normalized interaction strength.

Figure.1 shows the result of MinMaxCut clustering on c-c network. Clusters (called supercomplexes) are labeled with the most frequently occurring proteins and the number of TAP-MS protein complexes with related biological processes. Gene Ontology annotations show that supercomplexes represent the diversity of interconnected cellular processes. For instance, GO annotations on the largest supercomplex (the first cluster shown in figure) suggest that it encompasses complexes involved in chromatin dynamics, transcriptional regulation and initiation, cell cycle control, DNA replication and repair, and signal transduction.

References

[1] Ho, Y., et al. 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415:180-193.

[2] Gavin, A.-C., et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141-147.

[3] Ding, C. 2002. Analysis of gene expression profiles: class discovery and leaf node ordering. *Proc. 6th Int'l Conf. Comp. Mol. Bio. (RECOMB)*, pp. 127–136.

[4] Ding, C., He, X., Zha, H., Gu, M. and Simon, H. 2001 A min-max cut algorithm for graph partitioning and data clustering. *Proc. IEEE Int'l Conf. Data Mining (ICDM)*. pp. 107–114.

C7. Which pathways cannot be reconstructed using protein phylogenetic profiles?

Yohan Kim[1], Shankar Subramaniam[1,2]

Keywords: genomic context, protein networks, computational proteomics

1 Introduction

Phylogenetic profile methods have shown that those proteins that share similar profiles are more likely to have same functions than those that do not [1]. After fine-tuning of these methods, their applications on a number of complete genomes have uncovered novel cellular systems [2,3]. However, what has received relatively little attention in these studies is providing explanations for why phylogenetic profile based methods cannot confidently assign functional relationships to a significant number of those proteins that are annotated in the KEGG database [4]. There are three scenarios that fit this observation. One scenario is that the number of complete genomes from which profiles were derived was not sufficient. Consequently, even though two proteins were functionally related, their profiles did not have enough number of co-varying profile elements for them to be considered similar. In the second scenario, two proteins considered were universally shared across genomes and thus the method could not pick up strong signals from comparisons of their profiles to assign functional relationships. Finally in the third scenario, a profile of one protein did show enough variations in its elements but there were no proteins with similar profiles even with 'sufficient' number of sequenced genomes being used. In order to better assess the performance of protein phylogenetic profiles based methods, pathways for *E. coli* K12 in the KEGG database and those reconstructed using the methods are compared. It is our hope that by identifying which types or classes of proteins are less amenable to phylogenetic profile based methods, we are more likely to come up with a new generation of algorithms that can assign functions to them with greater confidence.

2 Figures and Tables

total # of proteins	4311
# of protein with at least one KEGG pathway entry	1153
total # of unique protein pairs that share at least one KEGG pathway entry	38751

Table 1: *E. coli* K12 statistics.

[1] Dept. of Chemistry and Biochemistry, UCSD, 9500 Gilman Dr., San Diego, California, USA. E-mail: ykim@ucsd.edu

[2] Dept. of Bioengineering, UCSD, 9500 Gilman Dr., San Diego, California, USA. E-mail: shankar@sdsc.edu

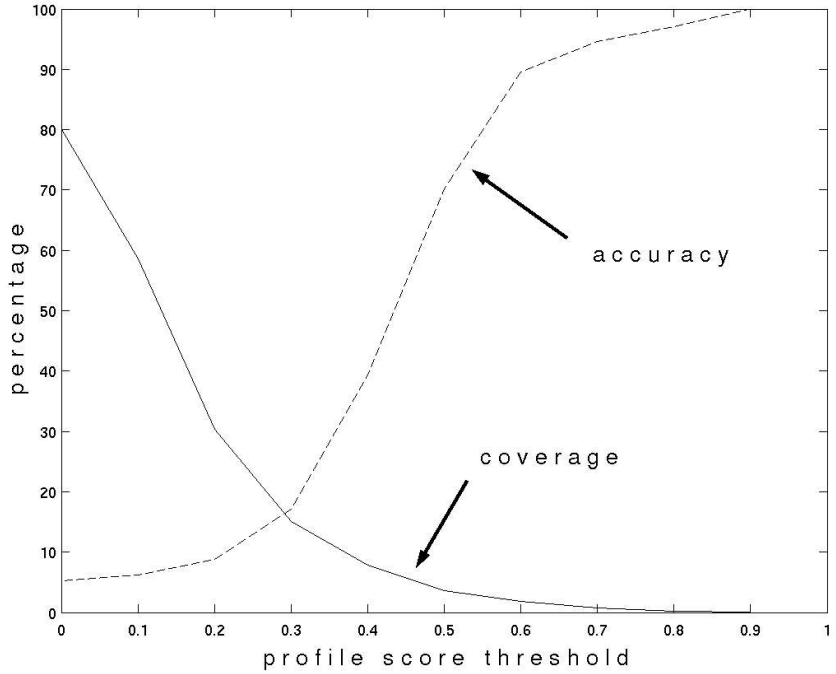


Figure 1: Accuracy and coverage of pathway predictions for *E. coli* K12 using protein phylogenetic profiles.

References

[2] Date, S.V. and Marcotte, E.M. 2003. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology* 21:1055-1062.

[4] Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto encyclopedia of genes and genomes.*Nucleic Acids Research* 28:27-30.

[1] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. 1999. Assigning protein functions by comparative analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences USA* 96:4285-4288.

[3] von Mering, C., Zdobnov, E.M., Tsoka, S., Ciccarelli, F.D., Pereira-Leal, J.B., Ouzounis, C.A., Bork, P. 2003. Genome evolution reveals biochemical networks and functional modules. *Proceedings of the National Academy of Sciences USA* 26:15428-15433.

C8. The Protein Mutant Resource: Visual and Statistical Analysis of Mutation with Implications for Homology Modeling

Werner G. Krebs[1], Philip E. Bourne[2]

Keywords: Protein Mutant Resource, PMR, PDB bias, structural bioinformatics, homology modeling, statistical inference, data-mining, functional annotation, GO, Gene Ontology, Encyclopedia of Life, EOL

1 Introduction.

Although databases of mutant gene products [1-3] as well as specialized databases of mutant protein structures have previously been developed [4-6], no comprehensive, PDB-wide database of mutant protein structures previously existed. The Protein Mutant Resource (PMR), a freely accessible database and associated tools, first addressed this need [7-9].The PMR systematically characterizes related artificially mutated structures from the Protein Data Bank (PDB) by grouping point mutations of the same structure. The PMR is available at http://pmr.sdsc.edu.

2 Database and algorithms.

The PMR illustrates the relationship between these mutated structures using morph technology [10,11] previously developed for the Macromolecular Motions Database (http://molmovdb.org). The PMR is intended to allow molecular biologists, protein engineers, and rational drug designers to analyze visually the apparent protein conformational change induced by mutation. In addition to accurately inferring mutant classifications in the Gene Ontology (GO) using an innovative, statistically rigorous data-mining algorithm with more general applicability [9], the PMR characterizes each entry by the number and type of artificially induced mutations found within the PDB. Comparison of the frequency of mutation in the PMR/PDB datasets against the accepted PAM250 natural amino acid mutation frequency indicates an inverse relationship, suggesting human efforts to engineer proteins with stable 3-dimensional structures involve processes statistically different from the exploration of structure space by evolution [7,8]. The PMR references nearly 20% of PDB chains and is updated via an automatic algorithm.

3 Implications for homology modeling.

Changes in PMR structures due to single-point mutations [7,8] are somewhat larger than is typically predicted by homology modeling, as the latter often does not take into account mutations' possible effects on linker flexibility [10,11] or secondary structure. We suggest that it may be possible to more accurately model mutations involving changes in backbone configuration by applying distributions based on PMR data to our existing Encyclopedia of Life (EOL) [12] technology (Figure 1).

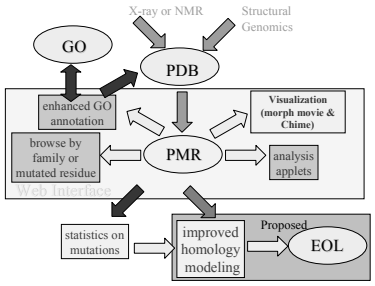


Figure 1: Diagram of PMR operation showing facilities, features, interactions with external databases, and proposed homology modeling facility via Encyclopedia of Life (EOL) project.

[1] San Diego Supercomputer Center Dept 0505, University of California, San Diego, La Jolla, CA 92093-0505, USA. E-mail: wkrebs@sdsc.edu

[2]School of Pharmacology and San Diego Supercomputer Center Dept 0505, University of California, San Diego, La Jolla, CA 92093-0505, USA. E-mail: bourne@sdsc.edu

References

[1] Beukers, M.W., et al. 1999. TinyGRAP database: a bioinformatics tool to mine G-protein-coupled receptor mutant data. *Trends Pharmacol Sci* 20:475-7.

[6] Gromiha, M.M., et al. 2002. ProTherm, Thermodynamic Database for Proteins and Mutants: developments in version 3.0. *Nucleic Acids Res* 30:301-2.

[5] Kawabata, T., Ota, M. and Nishikawa, K. 1999. The Protein Mutant Database. *Nucleic Acids Res* 27:355-7.

[7] Krebs, W.G. and Bourne, P.E. 2004. Statistical and Visual Morph Movie Analysis of Crystallographic Mutant Selection Bias in Protein Mutation Resource Data. *J. Bioinfo. Comp. Biol.,* in press.

[9] Krebs, W.G. and Bourne, P.E. 2004. Statistically Rigorous Automated Protein Annotation. *Bioinformatics,* in press.

[11] Krebs, W., et al. 2004. Studying Protein Flexibility in a Statistical Framework: Tools and Databases for Analyzing Structures and Approaches for Mapping this onto Sequences. *Methods Enzymol* 374.

[8] Krebs, W.G. and Bourne, P.E. 2003. Statistical and Visual Morph Movie Analysis of Crystallographic Mutant Selection Bias in Protein Mutation Resource Data. In W. Xu and P. Markstein, editors, *Computational Systems Bioinformatics: CSB2003*. Los Alamitos, CA: IEEE Computer Society Press. pp. 180-9

[10] Krebs, W.G. and Gerstein, M. 2000. The morph server: a standardized system for analyzing and visualizing macromolecular motions in a database framework. *Nucleic Acids Res* 28:1665-1675.

[2] Kristiansen, K., Dahl, S.G. and Edvardsen, O. 1996. A database of mutants and effects of site-directed mutagenesis experiments on G protein-coupled receptors. *Proteins* 26:81-94.

[12] Li, W.W., et al. 2003. The Encyclopedia of Life Project: Grid Software and Deployment. *New Gener. Comp.* (in press).

[3] Maurer, S.M., Firestone, R.B. and Scriver, C.R. 2000. Science's neglected legacy. *Nature* 405:117-20.

[4] Nishikawa, K., et al. 1994. Constructing a protein mutant database. *Protein Eng* 7:773.

C9. OrthoMCL: application of a graph cluster algorithm to comparative genomics and genome annotation

Li Li[1], Christian J. Stoeckert Jr[2], David S. Roos[1]

Keywords: OrthoMCL, comparative genomics, orthologous group, genome annotation, Markov cluster

1 Introduction

Comparative genomics has become a valuable approach in gene identification, functional annotation and evolutionary analyses. Orthology and paralogy are major concepts in molecular evolution and have been applied broadly in comparative genomics. Orthologs are genes from different species that derive from a common ancestor by speciation, while paralogs are genes that derive from a single gene that was duplicated within a genome [1]. As orthologs are likely to retain identical function over evolutionary time, the identification of orthologs is an important tool for gene annotation. Recently, two terms, inparalogs and outparalogs were introduced to distinguish paralogs derived from gene duplication before speciation and those from after speciation [2]. In comparative genomics, the clustering of orthologous genes provides a framework for data integration, highlighting the divergence and conservation of biological processes.

To cluster orthologous genes from eukaryotic genomes, however, complications arise from extensive gene duplication and functional redundancy, the multi-domain structure of many proteins and the predominance of incomplete eukaryotic genome sequencing. Previously, we have devised a scalable approach called OrthoMCL for the identification of orthologous groups in eukaryotic genomes, utilizing Markov Cluster algorithm, which is based on probability and graph flow theory, to delineate the many-to-many relationships between orthologous and paralogous genes [3]. Here we will present two applications of this approach in comparative genomics and genome annotation. Firstly, we applied this approach to comparative analysis of eukaryotic genomes and genome annotation for malaria parasites *Plasmodium*. Then we applied OrthoMCL to compare two separate gene finding efforts of human and mouse genomes, Ensembl [4] and Allgenes (http://www.allgenes.org). While Ensembl provides automated genome annotation from genomic sequences, Allgenes construct gene models from EST and mRNA sequences.

2 Results

We applied OrthoMCL on publicly available eukaryotic genomes including human, mouse, fly, worm, mosquito, *Arabidopsis*, yeast, malaria parasites *Plasmodium falciparum* and *Plasmodium yoelii* with *E. coli* as an outgroup. Data and results were stored in an object-oriented relational database, Genomic Unified Schema (GUS) (http://www.gusdb.org) and can be queried online (http://www.cbil.upenn.edu/gene-family/). We identified 26681 clusters of putative orthologs and inparalogs, 7519 of which are species-specific inparalogs, probably due to lineage-specific expansion. We then compared the orthologous genes of human and mouse identified by OrthoMCL with a curated dataset extracted from HomoloGene (http://www.ncbi.nlm.nih.gov/HomoloGene/). 91% of the 7328 curated orthologous pairs were found in the same ortholog group identified by

[1] 415 South University Avenue, Department of Biology, University of Pennsylvania, Philadelphia PA19104, USA. E-mail: {lili4,droos}@sas.upenn.edu

[2] Center for Bioinformatics, Blockley Hall, 423 Guardian Drive, University of Pennsylvania, Philadelphia, PA19104, USA. E-mail: stoeckrt@pcbi.upenn.edu

OrthoMCL. We also evaluated the consistency of the OrthoMCL clusters with EC annotation from the Enzyme Database (http://us.expasy.org/enzyme). Of the 1012 OrthoMCL clusters that contain at least two EC-annotated sequences, 909 (90%) are consistent with EC annotation. OrthoMCL clusters that contain *Plasmodium* sequences were incorporated into the *Plasmodium* genome database (http://www.plasmodb.org) to facilitate genome annotation (see example in Figure 1), identification of novel gene families, differentially expanded paralog groups and taxa-specific genes. 75% of *P. falciparum* proteome and 52% of *P. yoelii* proteome were found to be orthologous, while 1964 groups were also specific to *Plasmodium* genomes, providing candidates for candidates for drug and/or vaccine development.

gene	species	description
ENSANGP00000017463	*A. gambiae*	MANNOSE 6 PHOSPHATE ISOMERASE EC_5.3.1.8 PHOSPHOMANNOSE ISOMERASE PMI PHOSPHOHEXOMUTASE
At1g67070.1	*A. thaliana*	phosphomannose isomerase, putative / similar to phosphomannose isomerase GI:10834550 from [Arabidopsis thaliana]
At3g02570.1	*A. thaliana*	putative mannose-6-phosphate isomerase / similar to mannose-6-phosphate isomerase GB:NP_002426 from [Homo sapiens];supported by full-length cDNA: Ceres:40616.
CE07925	*C. elegans*	mannose-6-phosphate isomerase status:Partially_confirmed
CE33544	*C. elegans*	Mannose 6-phosphate isomerase status:Confirmed
CG8417-PA	*D. melanogaster*	gene symbol:CG8417 FBgn0037744 gene_boundaries:(3R:5,585,335..5,586,946 [+]) (GO:0004476 mannose-6-phosphate isomerase)
EG10566	*E. coli*	manA Mannosephosphate isomerase
ENSP00000318192	*H. sapiens*	MANNOSE-6-PHOSPHATE ISOMERASE (EC 5.3.1.8) (PHOSPHOMANNOSE ISOMERASE) (PMI) (PHOSPHOHEXOMUTASE). [Source:SWISSPROT;Acc:P34949]
ENSP00000318318	*H. sapiens*	MANNOSE-6-PHOSPHATE ISOMERASE (EC 5.3.1.8) (PHOSPHOMANNOSE ISOMERASE) (PMI) (PHOSPHOHEXOMUTASE). [Source:SWISSPROT;Acc:P34949]
ENSMUSP00000034856	*M. musculus*	MANNOSE 6 PHOSPHATE ISOMERASE EC_5.3.1.8 PHOSPHOMANNOSE ISOMERASE PMI PHOSPHOHEXOMUTASE
MAL8P1.156	*P. falciparum*	hypothetical protein
PY03463	*P. yoelii*	Phosphomannose isomerase type I, putative
YER003C	*S. cerevisiae*	PMI40;protein amino acid glycosylation*;mannose-6-phosphate isomerase activity;cellular_component unknown

Figure 1: A screen shot of ortholog group 762863 with *P. falciparum* gene MAL8P1.156 annotated as 'hypothetical protein', whose function maybe inferred from orthologs from other species.

To compare human and mouse gene models from Ensembl and Allgenes, we compared OrthoMCL clusters using human, mouse protein sequences from Ensembl with those using human, mouse protein sequences from Allgenes, using the same data from other species. With orthologs as supporting evidence, we identify gene models that are consistent or complementary between these two databases.

References

[1] Fitch, W.M. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* 19:99-113.

[4] Hubbard T, Barker D, Birney E, Cameron G, et al. 2002. The Ensembl genome database project. *Nucl Acids Res* 30: 38-41.

[3] Li, L., Stoeckert, C.J., Roos, D.S. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178-2189.

[2] Sonnhammer, E.L.L, Koonin, E.V. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics* 18: 619-620.

C10. Tandem MS Analysis and An Emerging Genome: The Sea Urchin Sperm Plasma Membrane Proteome

Anna T. Neill[1], Terry Gaasterland[2], John R. Yates III[3], Victor D. Vacquier[4]

Keywords: sea urchin, fertilization, membrane protein, tandem MS, whole genome sequence

1 Introduction.

Fertilization is a fundamental cellular process involving signaling and recognition at cell surfaces. Sperm plasma membrane proteins play key roles by mediating sperm motility, egg recognition, and gamete fusion. Phylogenetically, sea urchins lay at the base of the deuterostome lineage which leads to the vertebrates. Pragmatically, sea urchins provide an excellent model for the study of fertilization because large quantities of gametes are readily available. Moreover, sperm plasma membrane proteins are easy to isolate from the rest of the cell, and can be obtained separately from the head and the tail (flagellum). Thus it is straightforward to study proteins in the context of their subcellular location.

2 Approach.

We use micro liquid chromatography (□-LC) tandem mass spectrometry (MS/MS) to characterize the sea urchin sperm plasma membrane proteome. One way to interpret this type of MS data relies on prior knowledge of target protein sequences. Sea urchin genome sequencing is currently underway but in early stages. Consequently, the interpretation of sea urchin MS data requires methods to extract protein information from unassembled or partially assembled genomic sequence. Our approach is to translate protein sequence fragments from all open reading frames in the raw genome sequence data, and then to reduce and merge the fragments into a minimally redundant protein database, which is then searched with the collected spectra using mass-to-peptide assignment software.

[1] Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093-0202. E-mail: aneill@ucsd.edu

[2] The Rockefeller University, New York, NY 10021. E-mail: gaasterl@genomes.rockefeller.edu

[3] Department of Cell Biology, The Scripps Research Institute, La Jolla, CA 92037. E-mail: jyates@scripps.edu

[4] Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA 92093-0202. E-mail: vvacquier@ucsd.edu

C11. Characterizing protein function by integrating interaction data and domain information

Kinya Okada[1], Md. Altaf-Ul-Amin[2], Hirotada Mori[3], Shigehiko Kanaya[4], Kiyoshi Asai[5]

Keywords: interaction partners, domains

1 Introduction.

Comprehensive protein interaction data may provide clues to protein function in complex biological networks. However, these data contain a lot of 'noise' such as false positives and false negatives. This problem can be overcome by integrating other '-omic' resources, such as transcriptome profiling data [1]. In addition, we find in this study that the integration of protein interaction data and domain information is a rational approach not only to reducing error rates, but also to finding other aspects of protein function.

Domains are thought to be related to biologically important functions because they are evolutionally conserved units of proteins [2]. It is therefore likely that a functionally important domain contributes to deciding a protein's interaction partners. An example is reported in [3].

Similarity in the structural and functional properties of proteins is reflected by both domains and protein interaction partners. For characterizing protein function, we focus in this study on proteins which have both the same interaction partners and the same domains. Here we report the evaluation of this strategy and the functional characterization of proteins in *Escherichia coli*.

2 Method.

Protein interaction data were generated by pull-down assays. In this method, individual proteins in *E. coli* are tagged and allowed to be pulled down with interacting proteins, which are then analyzed by mass spectrometry. To obtain information about domain-protein relationships in *E. coli*, we downloaded Swisspfam, the domain structure database of SWISSPROT and TrEMBL proteins, from http://pfam.wustl.edu. Integrating protein interaction data and domain information, we extracted pairs of proteins which have the same protein interaction partners and the same domains by a graph-theoretical method. A typical protein pair satisfying these conditions is represented in Fig.1.

[1] NAIST 8916-5 Takayamacho, Ikoma, Nara, Japan. E-mail: kinya-o@is.aist-nara.ac.jp
[2] NAIST. Email: amin-m@is.aist-nara.ac.jp
[3] NAIST. Email: hmori@gtc.aist-nara.ac.jp
[4] NAIST. Email: skanaya@gtc.aist-nara.ac.jp
[5] CBRC, University of Tokyo and NAIST. Email: asai@cbrc.jp

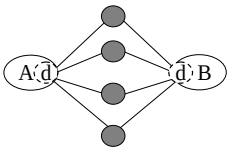


Figure 1: Proteins A and B have the common domain d and
the same interaction partners (filled circles). Lines represent interactions.

3 Result and Discussion.

We obtained 13,176 interaction pairs among 3,035 proteins of *E. coli*. Among the extracted paired proteins having the same interaction partners and the same domains, we found that protein pairs interacting with three or more common interaction partners do not occur randomly. We therefore examined the protein pairs satisfying this condition. All the pairs could be classified into 55 common partner networks consisting of 175 proteins. One of these networks is shown in Fig.2.

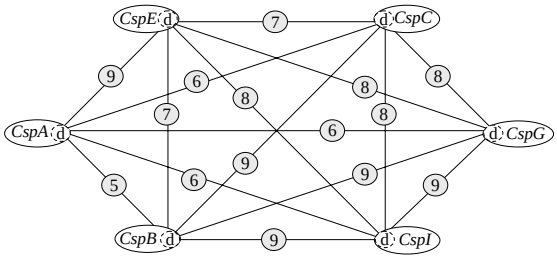


Figure 2: Common partner network constituted by the cold shock protein family. The proteins are linked via their PF00313 domains (d). Numbers in shaded circles represent the number of common interaction partners.

The paired proteins extracted in our strategy are expected to have common properties which are related to their common domains. Therefore, the function of previously uncharacterized partners in protein pairs can be predicted. Because other unshared domains of multi-domain proteins are ignored in our strategy, the difficulties of annotation for multi-domain proteins can be bypassed.

In some of the common partner networks, proteins with additional domains had a larger number of protein interaction partners or modified function, while in other cases proteins with fewer domains lacked some of their interaction partners. This suggested that the number of protein interaction partners has been strongly influenced by the addition or loss of domains during evolution. Another feature is bi-functional proteins, which can be seen in several common partner networks. These proteins seem to be linked with one domain for the first function and linked with another domain for the second function. In these cases, we may be able to ascribe roles to particular domains in a multi-domain protein.

References

[1] Ge H, Walhout AJ, Vidal M. 2003. Integrating ‘omic’ information: a bridge between genomics and systems biology. *Trends Genet* 19:551-60.

[2] Ponting CP, Russell RR. 2002. The natural history of protein domains. *Annu Rev Biophys Biomol Struct* 31:45-71.

[3] Morin F, Vannier B, Houdart F, Regnacq M, Berges T, Voisin P. 2003. A proline-rich domain in the gamma subunit of phosphodiesterase 6 mediates interaction with SH3-containing proteins. *Mol Vis* 9:449-59.

C12. The Encyclopedia of Life: A New Web Resource for Domain-Based Protein Annotation Data

Gregory B. Quinn[1], Mark A. Miller[1], Kim Baldridge[1], Ilya Shindyalov[1], WilfredLi[1], Dmitry Pekurovsky[1], Robert W. Byrnes[1], Kristine Briedis[1], Vicente Reyes[1], Adam Birnbaum[1], Coleman Mosley[1], Julia Ponomarenko[1], Yohan Potier[1], Celine Amoreira[1], Stella Veretnik[1], Philip Bourne[1,2]

Keywords: pipeline, domains, annotation, gui, visualization, datawarehouse, dhtml, svg, jboss

1 Introduction

The explosion in the availability of putative protein sequence data from DNA sequencing projects has provided a challenge to the research community: how to analyze this vast amount of data and then present these annotations in an easy-to-navigate web-based graphical user interface (GUI). Using a unique and highly benchmarked software pipeline, the Encyclopedia of Life (EOL) provides the researcher with domain-based annotation for all publicly available protein sequence data. Scientists will be able to uncover the prevalence of a given protein across all kingdoms of life, molecular interactions with that protein, and whether the function of the protein varies across species.

2 The Software Pipeline

Core to the EOL project is the integrated **Genome Annotation Pipeline (iGAP)**, a suite of programs that annotate protein sequences for their putative structure and biological function. Performing this analysis with iGAP on data from a whole genome is a CPU-intensive process which requires the use of powerful GRID computing resources, mainframe supercomputers and computer clusters. The iGAP analysis includes calculating three-dimensional models and assigning biological function for all recognizable proteins in all currently known genomes.

3 An Intuitive Web Interface

Key to EOL being the vital scientific resource intended is that its data is made available through an intuitive and highly functional web interface. The interface uses the JBoss application server to connect a datawarehouse which stores the EOL data in a query-optimized schema to the book metaphor web interface; this optimized design makes browsing and drilling down to data very fast and simple. Powerful keyword and protein sequence search functions enable the researcher to quickly locate data and annotations of interest. Dynamic HTML (HTML) is used extensively to provide the end user with a rich browsing and visual experience, and Scalable Vector Graphics (SVG) – based applets are used to present structural mapping of domain region matches.

4 A Collaborative Effort

EOL is an open collaboration led by the San Diego Supercomputer Center, and includes the work of researchers from the Singapore Bioinformatics Institute, Tokyo Institute of Technology, UFCG in Brazil, Belfast E-Science Center and Monash University in Australia. Other computational groups are invited to join in the effort to provide researchers with the best possible protein annotation resource.

[1]San Diego Supercomputer Center, University of California San Diego, La Jolla, CA 92093
[2]Department of Pharmacology, University of California San Diego, La Jolla, CA 92093

C13. Support Vector Machine approach to Active Sites Prediction using Local Sequence Information.

Dariusz Plewczynski[1], Adrian Tkacz[1], Leszek Rychlewski[1*]

Keywords: kinase substrate prediction, nearest neighbor, sequence similarity, database of active sites, Swiss-Prot database, support vector machine

1 Introduction.

The AutoMotif Server (AMS) predicts functional patterns in proteins. A list of possible functional motifs for a given query protein is predicted using only query protein sequence and the database of proteins annotated for certain types of biological processes by Swiss-Prot database [1]. All short segments of a query protein sequence sites are compared with the annotated sequence fragments using the support vector machine SVM approach [3]. Various methods are used here for building models based on different representations (in total 10 different embeddings) of known instances [4]. In order to estimate the efficiency of the classification for each type of functional site and the prediction power of the method the leave-one-out tests are used [2]. User can access all sites annotated by Swiss-Prot database (version 4.2), add new proteins with instances, or new annotation information. All data, constructed models and automatic predictor are updated after each major upgrade of the Swiss-Prot DB.

2 Software and files.

The method is available as an internet server at http://automotif.bioinfo.pl/. The whole database of annotated segments (positive instances), parent proteins (with detailed biological information included) is implemented as https://mysql.bioinfo.pl/ as the mySQL database with phpMyAdmin web interface.

3 Tables.

Table I. The prediction efficiency for various types of phosphorylation. The results are obtained using SVM learning with polynomial kernel ((s a*b+c)^d). Data is collected from Swiss-Prot DB annotation tables (without BY SIMILARITY, PREDICTED, PROBABLE, POTENTIAL or PARTIAL annotations). The first column in the table gives the number of positives and negatives for each type of activation process. The first row describes the dimension for each embedding method.
Results are collected for 8 different methods for preparing SVM input vectors representing each segment (with length 9 or 13 amino acids). The first one is the simplest BIN method uses binary representation of amino acids in the SVM input vector. The BIN+LOOKUP includes additional vector of 9 or 13 values (depending on the size of the segments) of frequency ratios between positives and negatives for these particular amino acid found in the input segment and the position in each predicted segment. The SPARSE method puts instead of 1 the value of frequency ratio between positives and negatives for these particular amino acid found in input segment and the position in each predicted segment. The SPARSE+LOOKUP includes also the frequency ratios for

[1] BioInfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland, Tel: +48-61-8653520, Fax: +48-61-8643350, E-mail: darman@bioinfo.pl

segments. The LOOKUP vector uses only frequency ratios for amino acids found in a query segment. The BLOSUM+LOOKUP method prints also values for various types of amino acids rescaling them by BLOSUM62 coefficients of the similarity between each type of amino acid and particular type of amino acid found in a query protein. The SUM_PROF uses only sum over the all frequency ratios (dot product of them), and the BLOSUM+SUM_PROF adds also BLOSUM62 similarity matrix. The last two methods uses the whole frequency information calculated on the both (positives and negatives) datasets with, or without separate LOOKUP information.
The most stable method is profile PROF+LOOKUP, SPARSE+LOOKUP or BLOSUM+LOOKUP methods. Other types of methods have lower efficiency (recall / precision).

Recall / precision	Number of positives/ negatives	BIN	BIN +LOOKUP	SPARSE	SPRASE +LOOKUP	BLOSUM +LOOKUP	LOOKUP	BLOSUM +SUM_ PROF	SUM_ PROF	PROF	PROF +LOOKUP
Dim (9/13 frag)		180 264	189 273	180 264	189 273	189 273	9 13	189 273	9 13	180 264	189 273
PKA (9)	86/14353	11.63%	43.02%	36.05%	37.21%	41.86%	41.86%	39.53%	37.21%	41.86%	41.86%
		76.92%	58.73%	55.36%	74.42%	69.23%	85.71%	80.95%	68.09%	75.00%	76.60%
PKC (9)	56/14368	1.79%	16.07%	14.29%	14.29%	17.86%	0%	0%	0%	17.86%	17.86%
		100%	42.86%	44.44%	40.00%	90.91%	0%	-	-	83.33%	62.50%
CDC2 (9)	41/14375	0%	29.27%	21.95%	24.39%	24.39%	21.95%	0%	0%	9.76%	17.07%
		-	31.58%	23.68%	33.33%	28.57%	69.23%	-	-	20.00%	28.00%
SULF (9)	83/6426	39.76%	39.76%	38.55%	39.76%	46.99%	38.55%	13.25%	7.23%	48.19%	57.83%
		97.06%	75.00%	74.42%	73.33%	76.47%	72.73%	100%	100%	86.96%	78.69%
ABL (13)	4/10846	0%	100%	100%	100%	100%	100%	75.00%	75.00%	50.00%	75.00%
		-	100%	100%	100%	100%	100%	100%	100%	100%	100%
CK2 (9)	62/11746	0%	17.74%	19.35%	20.97%	12.90%	14.52%	0%	0%	11.29%	12.90%
		-	47.83%	44.44%	39.39%	50.00%	100%	-	-	53.85%	53.33%
CK (9)	85/11739	0%	10.59%	11.76%	12.94%	8.24%	5.88%	0%	0%	9.41%	9.41%
		-	36.00%	35.71%	40.74%	63.64%	71.43%	-	-	57.14%	36.36%
ABLpept (13)	129/1304	0%	27.13%	6.98%	30.23%	11.63%	34.88%	0%	0%	3.10%	8.53%
		0%	61.40%	64.29%	63.93%	71.43%	77.59%	-	-	57.14%	61,11%

Table 1: The prediction efficiency for predicting phosphorylation by various types of kinases.

4 References.

[1] Bairoch A, Apweiler R. 1999. The Swiss-Prot protein sequence data bank and its supplement TrEMBL in 1999. Nucl Acids Res. 27, pp. 49-54.

[2] Joachims, T. (2000). Estimating the Generalization Performance of a SVM Efficiently. Proceedings of the International Conference on Machine Learning, Morgan Kaufman.

[3] Vapnik,V.N. (1998). Statistical Learning Theory. Wiley, New York.

[4] Zavaljevski, N, Stevens, F.J., Reifman, J. (2002). Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. Bioinformatics. vol. 18(5), pp. 689-696.

C14. Mechanisms for Antagonistic Regulation of AMPA and NMDA-D1 Receptor Complexes at Postsynaptic Sites

Gabriele Scheler, [1] **Johann Schumann**[2]

Keywords: synaptic plasticity, receptor upregulation, AMPA, dopamine, striatum, D1

1 The problem: Maintenance of stable states

The difficulty of clearly establishing a mechanism for LTP/LTD has been sometimes analyzed as stemming from its definition as an electrophysiological phenomenon corresponding to a number of different molecular components of both presynaptic and postsynaptic plasticity regulation.

Here we present a hypothesis on the pathways underlying AMPAR/D1R regulation derived from an ongoing project on modeling membrane receptor plasticity[3]. We assume that the maintenance of brief or repetitive input signals at membrane receptors necessary to achieve lasting receptor upregulation is mainly supported by the interactive dynamics of the intracellular system, rather than being located in individual switches. Here we suggest specifically a bifurcation into two stable states for a simulated cortico-striatal synapse (Fig. 1). This model incorporates the important role of calcium dynamics and the CaMKII autophosphorylation switch but emphasizes the increased computational power achieved by cAMP-calcium interaction and dopamine D1 receptors (or other G-protein coupled receptors located at postsynaptic sites).

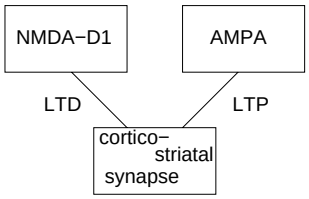


Figure 1: Two stable states for postsynaptic receptor regulation

2 cAMP-dependent modulation of calcium signals

The interaction of cAMP and calcium is a major factor in determining activation levels of a number of proteins critically involved in synaptic plasticity (see 3). Receptor complexes involving D1R and L-type calcium channels produce cAMP elevation at receptor activation and calcium influx through opening of the L-type calcium channels, where calcium further raises cAMP via the calcium-activated adenylyl cyclases AC1 and AC8. (The synergistic cAMP/calcium signal can be terminated by G_i coupled receptors, e.g. μ-opioid or dopamine D2 receptors.) NMDA receptor activation by high-frequency glutamatergic stimulation can produce sharp calcium signals sufficient to induce AMPA upregulation and LTP, e.g. in hippocampal CA1/CA3 but also at corticostriatal synapses. However, a different induction pattern, favoring longer lasting, weaker calcium oscillations induces AMPA LTD in hippocampus, which is also the dominant effect with dopamine D1 receptor stimulation at corticostriatal synapses.

A simulation of cAMP/calcium interaction shows that cAMP activation generates broadened oscillatory calcium signals (a) because of the sustained feedback between cAMP and calcium and (b) because cAMP is highly diffusible, and in spite of some compartmentalization of the signal will extend beyond the limits of a single synapse. Both effects counteract the generation of very high concentration gradients for calcium.

This means that the concurrent or temporally contiguous activation of cAMP during NMDA receptor mediated calcium entry is capable of profoundly altering the calcium signal and prevent the kind of persistent AMPA phosphorylation required for LTP-like AMPA upregulation.

[1]ISLE; Stanford, Ca, 94305 Email:scheler@stanford.edu
[2]RIACS/NASA Ames; Moffett Field, Ca, 94035
[3]References at URL: http://www.stanford.edu/∼scheler/plasticity.html

3 Regulation of AMPA and D1 receptor efficacy

The analysis of the kinase-phosphatase regulatory network stimulated by NMDA, AMPA and dopamine receptors shows the emergence of two internally consistent states of protein phosphorylation state and concentration (see Fig. 2A and Fig. 2B).

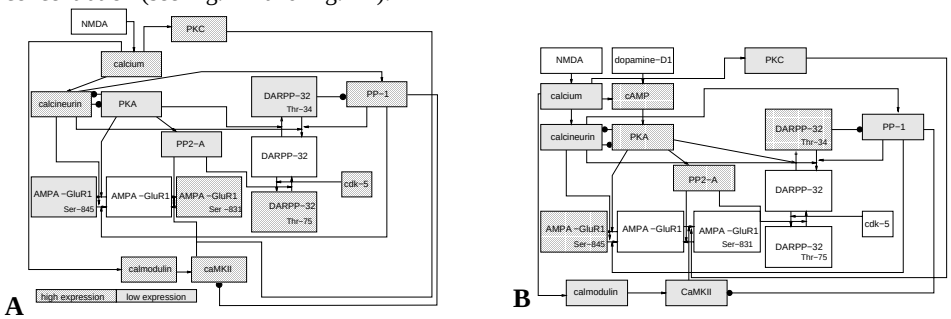


Figure 2: Intracellular Network for AMPA **(A)** and NMDA-D1 **(B)** dominated synapse

CaMKII is the autophosphorylated switch that gets turned on by high calcium. It promotes AMPA-GluR1-Ser-831 phosphorylation, which seems to be the dominant mode for the induction of AMPA upregulation and LTP. Calcineurin dephosphorylates AMPA-GluR1-Ser-845 and in this way lowers AMPA peak current. High calcineurin and low PP2-A concentrations also impair dopamine signaling pushing DARPP-32 into a Thr-75 phosphorylated state. This removes inhibition of PP-1 and allows PP-1 regulation of the CaMKII switch to operate. High cdk5 can inhibit NMDA currents, and thus may prevent further plasticity (Fig. 2A).

With cAMP activation, PKA and PP2-A are highly activated, while calcineurin/PP-1 is low. AMPA is predominantly phosphorylated by PKA at Ser-845 and dephosphorylated at Ser-831 which induces increase of peak current but may prevent lasting plasticity. In contrast the DARPP-32 switch at Thr-34, together with the slow calcium oscillation promotes D1 receptor insertion (Fig. 2B).

These states are both robust and stable, they do not contain major instabilities, thus they are suitable for the maintenance of input signals and induction of longer-term plasticity, such as receptor endo- and exocytosis. The proposed model relates to a number of attested observations:

- NMDA activation and prolonged calcium increase promotes D1 insertion
- D1 receptor activation promotes AMPA desensitization/downregulation (LTD), but increases AMPA peak current
- high-frequency stimulation induces LTP, requiring CaMKII autophosphorylation
- maintenance of LTP-state is associated with NMDA downregulation and reduction of calcium inflow.

4 Conclusion

From the analysis of these pathways we conclude that postsynaptic processes that regulate synaptic transmission undergo significant cross-talk with respect to glutamatergic and neuromodulatory (dopamine) signals. The main hypothesis is that of a compensatory regulation, a competitive switch between the induction of increased AMPA conductance by CaMKII-dependent phosphorylation and reduced expression of PP2A, and increased D1 receptor sensitivity and expression by increased PKA, PP2A and decreased PP-1/calcineurin expression. Both types of plasticity are induced by NMDA receptor activation and increased internal calcium, they require different internal conditions to become expressed. Specifically, we propose that AMPA regulation and D1 regulation are inversely coupled. The net result may be a bifurcation of synaptic state into predominantly AMPA or NMDA-D1 synapses. This could have functional consequences: stable connections for AMPA and conditional gating for NMDA-D1 synapses.

C15. Do Sense-Antisense Proteins Really Interact?

Ruchir R. Shah, [1] **Todd J. Vision,** [2] **Alexander Tropsha** [3]

Keywords: S-AS=Sense Antisense, complementary proteins.

1 Introduction

For each of the 64 codons there is a corresponding *antisense* codon, as shown for the four glycine codons in Table 1. Pairs of proteins encoded by reverse complementary codons, usually from opposite strands of the same coding region, have been dubbed sense-antisense, or S-AS, protein pairs [1]. As can also be seen in this table, there are often multiple antisense amino acids for a given sense amino acid due to the degeneracy of the genetic code [2]. Thus, two protein sequences may have recognizable sense-antisense alignments even when the protein pairs are encoded at different loci and the two coding sequences are not perfect reverse complements of one another (Figure 1).

Whereas specific interactions between complementary mRNA sequences are well known, similar interactions between sense and antisense protein sequences have been a matter of some debate. On the other hand the pool of experimental evidence supporting specific interaction between antisense proteins is growing; see [3] for the most recent comprehensive list. Here in, we address the question of whether sense-antisense protein pairs are of biological significance by seeking to identify all such potential pairs within the yeast proteome. We ask whether such pairs are more abundant than expected by chance and whether there is evidence for functional or physical interaction between the members of each pair. We find over 500 statistically significant Smith-Waterman alignments using a custom sense-antisense scoring matrix with a predicted 25% false positive rate. In addition, we find that sense-antisense pairs have more similar gene expression profiles than random pairs and that they are enriched in protein interaction/protein complex data. Most of the 28 pairwise interactions observed among sense-antisense pairs are due to a network of interactions among only 14 proteins (Figure 2). These results raise a myriad of structural, functional and evolutionary questions regarding sense-antisense proteins.

2 Figures and tables

Sense				Antisense
Gly	GGA	→	TCC	Ser
Gly	GGC	→	GCC	Ala
Gly	GGG	→	CCC	Pro
Gly	GGT	→	ACC	Thr

Table 1: The antisense amino acids of glycine are the corresponding amino acids encoded by the reverse complement of each codon.

[1]Laboratory for Molecular Modeling, School of Pharmacy, UNC-CH, Chapel Hill, NC. E-mail:ruchir@email.unc.edu

[2]Department of Biology, UNC-CH, Chapel Hill, NC. E-mail:tjv@biomass.bio.unc.edu

[3]Laboratory for Molecular Modeling, School of Pharmacy, UNC-CH, Chapel Hill, NC. E-mail: tropsha@email.unc.edu

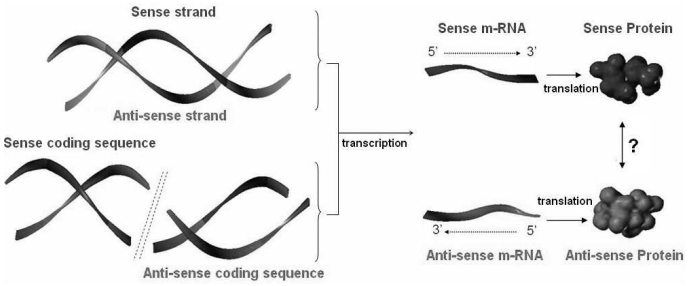


Figure 1: A sense and antisense protein pair can be encoded by genes at the same genetic locus (top left) or at different loci (bottom left). The possible interaction between proteins is indicated by a question mark.

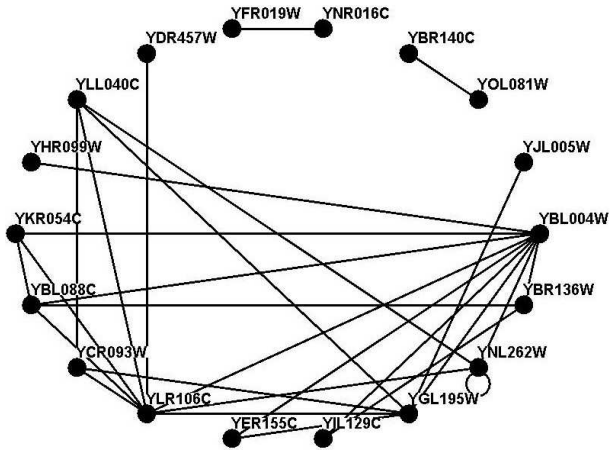


Figure 2: S-AS interacting protein network: All 28 interacting S-AS protein pairs were used to generate the network using the software *Osprey* [4]. Each node in the network represents a protein; An edge connecting two nodes means that the nodes are antisense with each other and are also found to be interacting.

3 References and bibliography

References

[1] Mekler, L.B. 1969. Specific selective interactions between amino acid residues of the polypeptide chains. *Biofizika* 14:581-584.

[2] Mekler, L.B. and Idlis, R.G. 1981. Constructions of models of three-dimensional biological polypeptides and nucleoprotein molecules in agreement with a general code which determines specific linear recognition and binding of amino acid residues of polypeptides to each other and to trinucleotides of polynucleotides. *Deposited Doc. VINIT* 1476-1481 (in Russian).

[3] Heal, J.R. *et al.* 2002. pecific interactions between sense and complementary peptides: The basis for the proteomic code. *ChemBioChem* 3:136-151.

[4] URL:http://biodata.mshri.on.ca/osprey/servlet/Index.

C16. Predicting Co-Complexed Protein Pairs Using Genomic and Proteomic Data Integration

Lan V. Zhang[1], Sharyl L. Wong[1], Oliver D. King[1], Frederick P. Roth[1]

Keywords: protein-protein interaction, protein complex, decision tree, data integration, machine learning

1 Introduction.

Identifying all protein-protein interactions in an organism is a major objective of proteomics. A related goal is to know which protein pairs are present in the same protein complex. High-throughput methods such as yeast two-hybrid (Y2H) and affinity purification coupled with mass spectrometry (APMS) have been used to detect interacting proteins on a genomic scale [1-4]. However, both Y2H and APMS methods have substantial false-positive rates. Aside from high-throughput interaction screens, other gene- or protein-pair characteristics may also be informative of physical interaction. Therefore it is desirable to integrate multiple datasets and utilize their different predictive value for more accurate prediction of co-complexed relationship.

2 Results.

Using a probabilistic decision tree approach, we integrated high-throughput protein interaction data with other gene- and protein-pair characteristics to predict co-complexed protein (CCP) pairs. Our predictions proved more sensitive and specific than predictions based on Y2H or APMS methods alone or in combination (Figure 1). Among the top predictions not annotated as CCPs in our reference set of protein complexes (obtained from the MIPS complex catalog), a significant fraction were found to physically interact according to a separate database (YPD, Yeast Proteome Database), and the remaining predictions may potentially represent unknown CCPs (Table 1).

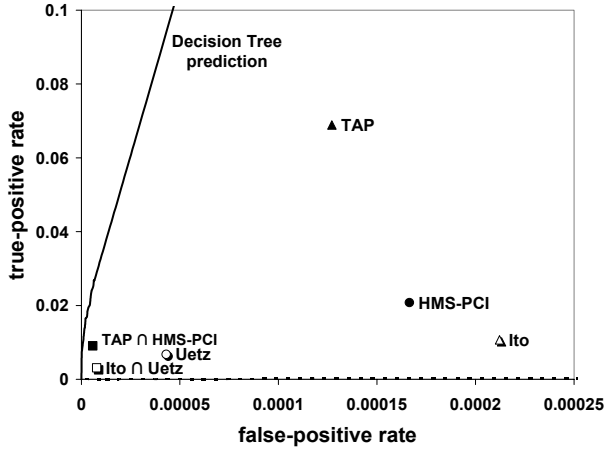


Figure 1: ROC (Receiver Operating Characteristic) curve of decision tree predictions in comparison with four high-throughput datasets (two YPD studies: Ito *et al.* and Uetz *et al*., and two APMS studies: Gavin *et al*. and Ho *et al*.), as well as their simple combinations (intersection of the two Y2H studies and intersection of the two APMS studies). Solid line represents decision tree predictions, while dotted line represents random predictions.

[1] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, MA 02115, USA. E-mail: fritz_roth@hms.harvard.edu

Rank	Protein 1	Protein 2	Score	YPD Complex Annotation
1	Rpl40Bp	Rps31p	0.943	
2	Rps31p	Rpl40Ap	0.938	
3	Smc1p	Smc3p	0.864	Cohesin
4	Gpt2p	Sec28p	0.857	
5	Pwp2p	Utp13p	0.844	Small subunit processome
5	Sgn1p	Pub1p	0.844	
7	Rdh54p	Rad5p	0.833	
7	Arp3p	Rvs167p	0.833	
7	Arp3p	Srv2p	0.833	
10	Spt5p	Rpb3p	0.800	Paf1p complex
10	Spt5p	Rpo21p	0.800	Paf1p complex
12	Pwp2p	Dip2p	0.776	Small subunit processome
12	Pwp2p	Ylr409C	0.776	
12	Sap190p	Sap155p	0.776	
12	Sap190p	Sap185p	0.776	
12	Pph21p	Pph22p	0.776	
12	Nop7p	Fpr4p	0.776	
12	Sap185p	Sap155p	0.776	
12	Sik1p	Cbf5p	0.776	
12	Nop2p	Ebp2p	0.776	Pre-60S ribosomal particle
12	Rpa135p	Ret1p	0.776	
22	Pwp2p	Asc1p	0.750	
22	Drs1p	Spb4p	0.750	
24	Rsm10p	Mrps5p	0.744	Mrp4p-associated complex (mitochondrial ribosome)
24	Mtr3p	Rrp45p	0.744	Exosome 3'-5' exoribonuclease complex
24	Rrp40p	Rrp46p	0.744	Exosome 3'-5' exoribonuclease complex
24	Rrp40p	Ski6p	0.744	Exosome 3'-5' exoribonuclease complex

Table 1: Top 27 predictions that are not annotated as CCPs in the reference set

3 Conclusion.

We demonstrated that the probabilistic decision tree approach can be successfully used to predict co-complexed protein (CCP) pairs from other gene- or protein-pair characteristics. Our top-scoring CCP predictions provide testable hypotheses for experimental validation.

References

[1] Ho, Y., Gruhler A., Heilbut A., Bader G.D., *et al.*, 2002. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature,* 415:180-183.

[2] Ito, T., Chiba T., Ozawa R., Yoshida M., *et al.*, 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proceedings of the National Academy of Sciences USA,* 98:4569-4574.

[3] Uetz, P., Giot L., Cagney G., Mansfield T.A., *et al.*, 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature,* 403:623-627.

[4] Gavin, A.C., Bosche M., Krause R., Grandi P., *et al.*, 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,* 415:141-147.

C17. Structal: A software for optimizing structural and sequence alignments

Hervé Seligmann[1] and Neeraja M Krishnan[1]

Keywords: structural alignment, homology, correspondence, gap penalty, secondary structure, empirical test of ancestral reconstruction.

1. Introduction

Comparative biology, whether morphological, physiological or molecular, is based on correspondences between organisms or parts of organisms. Homology is a type of correspondence that exists between more or less similar entities when they have common ancestry. In a morphological example, comparisons between wings and arms are considered homologous in spite of them being morphologically different because they diverged originally from the same evolutionary and developmental processes. But comparisons do not necessarily have to be based on common ancestry [1]. In proteins with similar functions [2] but different descent, alignments between sequences maximize structural similarities, independently of the number of evolutionary steps necessary to transform one sequence into the other. In contrast, sequence alignments optimize over sequence similarity and gap penalty. The assumption of common ancestry is introduced into the process by assuming parsimony of divergent events between the sequences. It presumes that, in molecules with simple structure and function, preserving sequence similarity also provides functional equivalence for homologously aligned sites. However, sometimes, highly similar homologous segments can have very different contributions in the structures and functions of their respective molecules, even in molecules with secondary structures as simple as tRNAs.

2. Materials and Methods

Ideally, sequence alignments contribute to the same homologous properties of their known functional structure. However, in some cases, sites considered homologous according to the sequence alignments do not contribute to the same specific local structure. For tRNAs, because of the simple and well defined function associated with the cloverleaf structure, the functional realism of sequences can therefore be easily estimated by using folding programs based on empirical thermodynamic properties of RNA [3] or information properties (tRNAscan-SE, [4]): the functional tRNA cloverleaf structures consist of two types of nucleotide sites, complementary ones forming stems, and those forming loops. This property was elegantly used to evaluate the accuracy of sequencing, because sequencing errors result in lower foldability of tRNAs into their classical functional structure [5]. We estimated the functional accuracy of all reconstructed ancestral tRNAs (results were qualitatively similar for parsimony, maximum likelihood and Bayesian reconstructions) in primate mitochondria by canonical complementarities between sites in helices.

3. Results

Complementarity between sites in helices for reconstructed ancestral sequences decreased as a function of each of the following (partly interdependent) measures of evolutionary divergence between the extant species (in increasing order of effect on tRNA functional realism): the resulting average similarity between the aligned sequences, the number of gaps introduced by the sequence alignment procedure, and the similarity between the secondary structures of the extant, aligned sequences. These tRNA sequences were aligned by mapping corresponding stems and loops (based only on the criteria of structural similarity).

Biological Computation and Visualization Center, Dept of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA.

4. Discussion

Structural and sequence similarity among extant tRNAs have the highest, and, respectively, the lowest effect on the functional realism of reconstructed sequences. Hence, alignments optimizing over both sequence and structure criteria, giving more weight to structural ones, should yield the functionally most accurate ancestral sequences. Our methods could assess optimal combinations of structural and sequence constraints in order to obtain the biologically most realistic alignment and create an off the shelf software to analyze nucleotide sequences, especially structural ones, such as tRNAs, rRNAs and control regions, and, at the very least, could hint at principles useful in solving similar, but much more complex problems associated with sequence versus structural similarities in amino acid sequences.

5. Figures and Tables

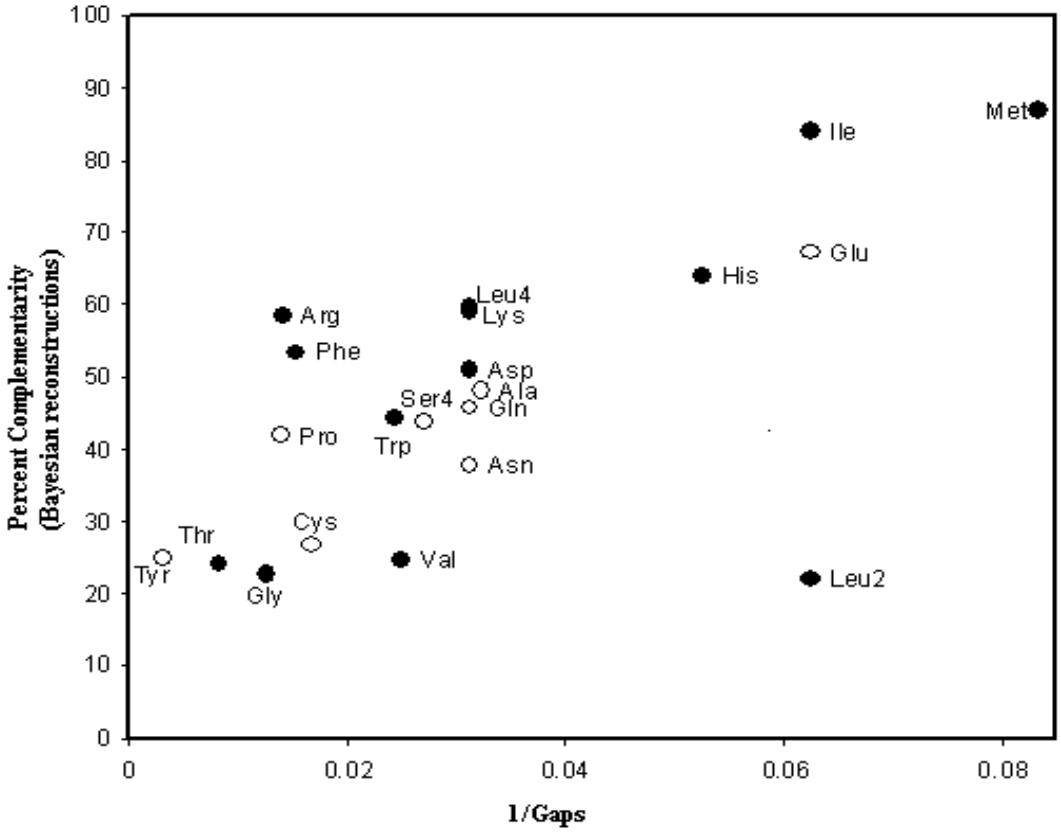


Fig. 1. Effect of number of gaps on functional realism of reconstructed tRNAs using a structural alignment. (Filled dots are for tRNAs coded on the heavy strand and empty dots are for tRNAs coded on the light strand)

6. References

[2] Friedberg, I., Kaplan, T. and Margalit, H. 2000. Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. Protein Science 9: 2278-84.

[4] Lowe, T.M. and Eddy, S.R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* 25: 955-64.

[5] Noor, M.A. and Larkin, J.C. 2000. A re-evaluation of 12s ribosomal RNA variability in *Drosophila pseudoobscura*. *Molecular Biology and Evolution* 17: 938-41.

[1] Rieppel, O. 1988. Fundamental of comparative biology. Basel: Birkhauser Verlag.

[3] Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research* 31: 3406-15

C18. An alternative to the SEQUEST cross-correlation scoring algorithm for tandem mass spectral identification through database lookup: the Luck scoring function, and the probability of an unrelated spectra match model

Tema Fridman, [1] **Jane Razumovskaya,** [2] **Nathan Verberkmoes,** [3] **Greg Hurst** [4] **and Ying Xu** [5]

Keywords: tandem mass spectrometry, scoring function

1 Introduction.

Mass spectrometry represents a leading technology for examining the functional states of proteins in cells [1], [3]. In a typical LC/MS/MS experiment, a protein mixture of interest is digested into peptides which are separated, ionized and introduced into mass spectrometer. Selected peptide ions are isolated and fragmented through collision-induced dissociation (CID). The resultant fragments are then measured for their m/z and intensity values. The fragmentation pattern is then used to identify the parent peptide. The most practical and widely used technique is peptide identification (peptides are then matched to corresponding proteins) through database search (SEQUEST, Mascot software programs). In such methods, peptides from a sequence database are compared to the experimental data. Specifically, theoretical mass spectra are first generated for a set of candidate peptides, and then compared with the experimental spectra using a matching function.

SEQUEST [2] represents one of the most widely used and accurate programs for peptide identification via database searches. Developed several years ago, SEQUEST was not specifically designed for large-scale applications. One of the limiting factors in meeting the needs for genome-scale applications is its (lack of) computing speed.

Here we present a new method with high discriminating power for searching protein sequence databases for peptide identification. The accuracy compares favorably to the SEQUEST scoring function, with better separation between correct and incorrect matches. The algorithm also runs significantly faster than the SEQUEST program, by roughly about two orders of magnitude.

2 The Model.

First we introduce a *Luck* function, that has physical sence of quantitative measure of luck to obtain a certain outcome $wish_k$, given any unimodal probability distribution of possible outcomes:

$$Luck(wish_k) = Sign(wish_k - mode) \log \frac{P(mode)}{P(wish_k)}, \quad (1)$$

[1]ORNL, PO Box 2008 MS 6164, Oak Ridge, TN 37831-6164, USA. E-mail: tfa@ornl.gov
[2]ORNL, PO Box 2008 MS 6164, Oak Ridge, TN 37831-6164, USA. E-mail: rzw@ornl.gov
[3]ORNL, PO Box 2008 MS6131 Oak Ridge, TN 37831-6131, USA. E-mail: verberkmoesn@ornl.gov
[4]ORNL, PO Box 2008 MS6131 Oak Ridge, TN 37831-6131, USA. E-mail:hurstgb@ornl.gov
[5]Biochemistry and Molecular Biology Department, University of Georgia, Athens, GA 30602, USA. E-mail: xyn@bmb.uga.edu

with *mode* being the result that occurs most frequently, and $Sign(x)$ being the sign function (1 for positive x, 0 for $x = 0$, and -1 for negative x).

We developed a model derives the probability distribution of degree of match between a given experimental spectrum (produced by the peptide of interest, the *true* peptide) and a theoretical spectrum from a database peptide, assuming that the theoretical spectrum is produced by a different, unrelated peptide. The resulting probability is a function of the experimental spectrum density, the length of the interval on which the experimental peaks are distributed, the number of theoretical peaks, and their distribution pattern. The model does not take into account the intensity of peaks and the size of the database.

Based on the probability distribution above, we calculate the *Luck* of the match between each experimental-theoretical spectral pair, and use it as a scoring function for the match. As the experimental spectrum has consistently higher degree of match with the theoretical spectrum of the true peptide than with that of an unrelated peptide, we get a substantially higher *Luck* score for the correct match (i.e., you are exceptionally lucky to see that degree of match, if you assume that the process, which generates the match, is random).

3 The experiment and the results.

We have tested our algorithm on a data set of 3771 experimental spectra that resulted from performing an LC-MS/MS experiment on a protein mixture of eight purified proteins.Our peptide database consisted of all possible tryptic peptides generated from the eight target proteins (having 803 peptides), and from the 2873 *S. oneidensis* proteins (having 289,166 peptides). The latter was used as a distractor dataset.

Among 1053 spectra of parent charge one, we identified 526 correctly versus 532 found by Sequest. Among the remaining 2721 spectra of parent charge two and three, we identified 496 versus SEQUEST's 505 for parent charge two, and 221 versus SEQUEST's 227 for parent charge three. The spectra, that SEQUEST identified and *Luck* function did not, have very low Xcorr score. In terms of sensitivity – specificity analysis, *Luck* function performs better than SEQUEST due to greater separation between correct and incorrect identifications.

4 References and bibliography.

References

[1] Aebersold,R., Mann,M. 2003 Mass spectrometry-based proteomics. In: *Nature*, 422(6928):198–207. Review.

[2] Eng,J.K., McCormack,A.L., and Yates III,J.R. 1994 An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. In: *J Am Soc Mass Spectrom*, 5: 976–989.

[3] Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K., Yang,L., Wolting,C., Donaldson,I., Schandorff,S., Shewnarane,J., Vo,M., Taggart,J., Goudreault,M., Muskat,B., Alfarano,C., Dewar,D., Lin,Z., Michalickova,K., Willems,A.R., Sassi,H., Nielsen,P.A., Rasmussen,K.J., Andersen,J.R., Johansen,L.E., Hansen,L.H., Jespersen,H., Podtelejnikov,A., Nielsen,E., Crawford,J., Poulsen,V., Sorensen,B.D., Matthiesen,J., Hendrickson,R.C., Gleeson,F., Pawson,T., Moran,M.F., Durocher,D., Mann,M., Hogue,C.W., Figeys,D., Tyers,M. 2002 Systematic Identification of Protein Complexes in Saccharomyces cerevisiae by Mass Spectrometry. In: *Nature*, 415:180–183.

C19. Cross-Species Peptide Mass Fingerprinting Database Searching and "Conserved-Domains"

Heming Xing[1], John Pirro[1]

Keywords: proteomics, mass spectrometry, database search, peptide mass fingerprinting, cross-species

Peptide mass fingerprinting provides rapid protein identification using mass spectrometer as the primary analytical technique coupled with bioinformatics. This relies on the presence of protein sequence in the current database. As genome sequencing projects continue to add more and more sequence data into the sequence database, proteins from poorly characterized organism will increasingly be identified using cross-species comparison to proteins from well characterized organisms.

In this study, the application of cross-species protein identification using peptide mass fingerprinting [1] has been investigated. More than 7000 human/mouse protein orthologous pairs [2] are used to study the performance of cross-species PMF database searching. When sequence identity of ortholog pairs drops below 70% virtually no tryptic peptide of molecular weight between 700 and 3000 were conserved, which is consistent with previous small-scale study[3]. MOWSE program was also tested for the performance of cross-species protein identification using PMF. Mouse proteins are theoretically digested using trypsin and PMF searches are done against human protein sequence database. In 23% of the cases, PMF search returns the human ortholog hit ranked as No. 1, and in 37% of the cases, human ortholog hits are ranked within top 5. As expected in 42% of the cases, human ortholog hits are ranked out of top 50.

To improve the performance of cross-species PMF searches, conserved domains are used as the basis of recognizing homologous proteins across species boundaries. Several strategies are being tested to put more weights on the "conserved-domains" when the final MOWSE score is calculated. Preliminary results suggest that the MOWSE score considering the domain conservation can improve the performance of cross-species protein identification using PMF.

References

[1] Papping, DJC, Hojrup, P, and Bleasby AJ. "Rapid identification of proteins by peptide-mass fingerprinting." Curr Biol 1993, 3: 327-332.

[2] http://www.informatics.jax.org/

[3] Wilkins, MR, Williams, KL. "Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass; A theoretical evaluation." J theor Biol 1997 186:7-15.

[1] Bioinformatics Research, Charles River Proteomic Services, Worcester, MA, USA. Email: heming.xing@dds.criver.com, john.pirro@criver.com

D2. GOMER: Predicting Gene Regulation by Modeling Binding Sites

Josh A. Granek[1], **Neil D. Clarke**[2]

Keywords: protein-DNA interaction, transcription, regulation

1 Introduction.

The information required for all transcriptional regulation is encoded in DNA sequences. Models of these regulatory sequences enables us to understand and predict the expression patterns of genes. To this end, we have developed GOMER (Generalizable Objective Model of Expression Regulation), a software package that predicts transcriptional regulation by modeling the binding of transcription factors to genome sequences. GOMER allows regulatory models to be defined and evaluated in terms of the model's ability to rationalize data on regulation. GOMER is ideally suited for modeling regulatory systems based on expression data from microarray experiments, as well as genome localization data from chromatin-IP experiments. GOMER can also be used to predict novel genes expected to be regulated by a modeled system.

2 Methods and Results.

GOMER is based on a simple thermodynamic model for binding site occupancy [4] which has been generalized (1) to allow for flexibility in the description of the distribution of binding sites. These descriptions of regulatory regions can be user defined through the use of extension modules written in Python.

$$\mathrm{Score}_{\mathrm{feature}} = 1 - \prod_{i=1}^{\mathrm{sites}} \prod_{s=(\mathrm{for,rev})} \left(\frac{1}{1 + [X] K_a^{X,i,s} \kappa_{\mathrm{site}}^{(i)}} \right) \tag{1}$$

For example, regulatory region functions can define a regulatory region as simple as the 600bp 5' to the ORF (2), or something more complex, such as a Gaussian distributed weighting of binding sites reflecting a preference for regulatory binding sites to be a certain distance from the start of transcription.

$$\kappa_{\mathrm{site}}^{(i)} = \begin{cases} 1, & 0 \leq \alpha \leq 600 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

This model has been further generalized to account for cooperative and competitive interactions among transcription factors. The characteristics of these cooperative and competitive interactions can be defined through extension modules, analogous to that used to define regulatory regions.

GOMER uses rank order metrics [1, 2] to evaluate a given model's ability to identify genes expected to be regulated, providing an objective criteria for improving model parameters. GOMER can be used to model regulation of any feature (or combination of features)

[1]School of Medicine, Johns Hopkins University, Baltimore, MD, E-mail: jgranek@jhmi.edu
[2]School of Medicine, Johns Hopkins University, Baltimore, MD, E-mail: nclarke@jhmi.edu

defined in a genome description file, or features defined by base coordinates in a genome (e.g. microarray features)

We have applied GOMER to several transcriptional regulation systems, for which we have built models that are able to identify the known regulated targets. Additionally, the flexibility of GOMER has allowed us to apply it to other types of systems where sequence specific binding plays a role. One example is a GOMER model we built to demonstrate a hypothesized artifact of microarray based chromatin-IP experiments.[3] This artifact is inherent to the design of these experiments, however, by using the model to account for this artifact, we have found that the results of these chromatin-IP experiments are better correlated with predictions for sequence specific binding.

3 References and bibliography.

References

[1] Clarke, N.D. and J.A. Granek. 2003. Rank order metrics for quantifying the association of sequence features with gene regulation. *Bioinformatics* 19:212–8.

[2] Hanley, J.A. and B.J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143:29–36.

[3] Lieb J.D., Liu X., Botstein D., Brown P.O. 2001. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nature Genetics* 28:327–34.

[4] Liu, X. and N.D. Clarke. 2002. Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. *Journal of Molecular Biology 323:1–8.*

D3. Discovery of Tumor-Specific Alternative Splicing Sites*

Fang Rong Hsu[1], Chia Yang Cheng[2]

Keywords: alternative splicing, tumor-specific

1 Introduction.

The splicing mechanism recognizes the boundaries between exons and introns. Alternative splicing can be a mechanism to generate more than one mRNA and protein from a single gene. Alternative RNA splicing functions in two ways. One is turn on/off control of gene expression; the other is in the formation of multiple protein isoforms [1]. It has been estimated that 35%-59% of human genes have more than one RNA isoforms [2].

Alternative splicing can display a family of different protein in different tissues [1]. In the other hand, cancer associated splice variants have been reported for many genes [3]. ~15% of diseases-causing mutations in human genes involve misregulation of alternative splicing and errors in mRNA processing have been associated with cancer and other human diseases. [4]

Currently, many alternative splicing databases were developed [5, 7], and some detected tissue-specific and tumor-specific alternative splicing sites [2, 3]. However, after being sifted by clustering and aligning to genomic sequence, few EST sequence could be used to find tumor-specific sites for each tissue.

AVATAR, a value added transcriptome data base, which was developed by aligning ESTs to genomic sequence directly [7, 8], offer more rich resource to analysis EST express in many tissues at each alternative splicing sites. In this study, we present a method to detect the tumor-specific alternative splicing isoforms from certain tissue.

2 Materials and Methods

Three steps were present to detect tumor-specific alternative splicing sites. First, 14,099 human alternative splicing sites were collected from AVATAR, which exons, introns and alternative splicing sites were identified by aligning five million ESTs [9] to human genomic sequence(Build 31) .

Second, five million ESTs from 8,431 libraries were categorized into 45 tissues and three types of histology, normal, tumor and unknown [6, 10]. ESTs were divided into four pools by tissue and histology (isoform 1 and tumor, isoform 2 and tumor, isoform 1 and normal, and isoform 2 and normal) at each alternative splicing site.

Third, we calculated theses data by Fisher's exact test, divided left tail of P-value by right tail of P-value as confidence C. The splicing sites with certain tissue, which Log odd ratio of C was greater than 2, were suggested as tumor-specific alternative splicing sites.

3 Result

We found 20 genes which's LOD greater than 2. For each gene, most ESTs dates in the four pools were from more than one library. 40% of tumor-specific isoforms were happened in Brain, and 50% were 5' type alternative splicing (see table 1). Only three alternative splicing sites were also

[1] Department of Bioinformatics, Taichung Healthcare and Management University, Taichung, Taiwan. E-mail: frshu@thmu.edu.tw

[2] Department of Bioinformatics, Taichung Healthcare and Management University, Taichung, Taiwan. E-mail: std91242006@ms1.thmu.edu.tw

* support in part by NSC 92-3112-B-468-001 and NSC 91-2662-E-468-001-CC3, Taiwan

tissues-specific sites. GNAS complex locus (GNAS), LOD score is 2.09, located on chromosome 20, is one of the three genes which's alternative splicing isoforms were published in literatures.

type	No. of tumor specific sites
cassette	3
3'	7
5'	10

Tissue	No. of tumor specific sites
Brain	8
Lung	4
Skin	3

Table 1: Tissue and types distribution of human tumor-specific alternative splicing sites

GNAS gene structure has been well investigated [11]; Gsα is encoded by exons 1-13 of GNAS. Exon 2-13 are also included in two additional overlapping transcripts, XLαs and NESP55, each with a distinct first exon. (Fig. 1) In brain, exon 3 skipping isoform (S) and exon 3 including isoform (S') were both expressed in tumor and normal EST pools. In S, tumor ESTs (7) are expressed more than normal ones (4), but oppositely tumor ESTs (5) were less than normal ESTs (12) in S'. (Fig. 1)
We could ratiocinate that GNAS exon 3 might associated with cancer, though, reversed from which one of the three mRNAs the ESTs were unknown.
We have good ground for thinking that it is worth to validate these 20 tumor-specific alternative splicing sites by traditional experiment.

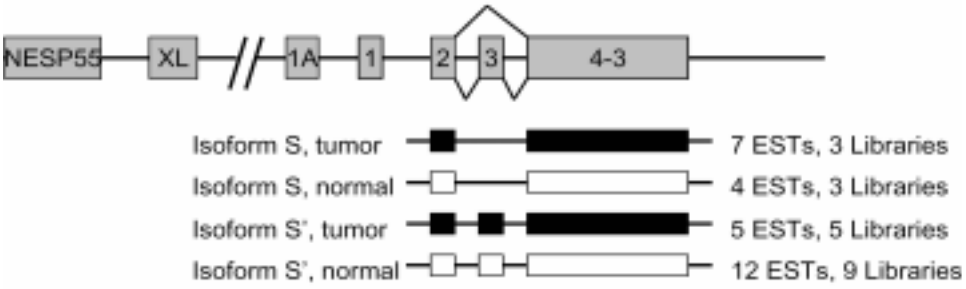


Figure 1: GNAS gene structure and EST expression in brain.

References

[1] A. R. Krainer 1997. Eukaryotic mRNA Processing, Oxford Univ., pp.242-279
[10] Cancer Genome Anatomy Project, http://www.ncbi.nlm.nih.gov/ncicgap/
[11] CancerGene Card GNAS1, http://caroll.vjf.cnrs.fr/cancergene/CG46.html
[7] F. R. Hsu, Ying Tsong Chen, Hwan-You Chang, Yaw-Lin Lin, Yin-Te Tsai, Hui-Ling Peng, Min Yao Shih, Chia-Hung Liu, and Jer Feng Chen, 2003, Proc. Of the 6th Conference on Engineering Technology and Chinese/Western Medicine Applications, Sep. 13, Taichung Taiwan, pp. 196-201.
[8] F. R. Hsu and J. F. Chen 2003, Proc. Of the IEEE Computational Systems Bioinformatics Conference (CSB2003), Aug., 11-14, San Francisco, USA, pp. 564-566.
[5] Lee, C., Atanelov, L., Modrek, B., Xing, Y., 2003, Nucleic Acids Res. 31: 101-5.
[6] Library Browser. http://www.ncbi.nlm.nih.gov/UniGene/lbrowse2.cgi?TAXID=9606
[9] NCBI dbEST, http://www.ncbi.nlm.nih.gov/dbEST
[4] Philips, A.V. and Cooper, T. A.,2000, Cell. Mol. Life Sci., 57,235-249
[2] Xu, Q., Modrek, B., Lee, C., 2002, Nucleic Acids Res. 30: 3754-66
[3] Xu, Q., Lee, C., 2003, Nucleic Acids Res. 31: 5635-5643

D4. Adaptive evolution of *E. coli* on lactate leads to convergent, generalist phenotypes

Andrew R. Joyce[1], Stephen S. Fong[2], Bernhard Ø. Palsson [3]

Keywords: adaptive evolution, microarray, gene expression, *E. coli*, MultiFun

1 Abstract.

Laboratory evolution has become a commonly used approach to address many of the fundamental long-standing questions about evolution [1] as well as to provide an experimental basis for delineating the mechanistic basis of adaptive evolution [2-3]. A recent review of this field [1] succinctly states key open questions in this field including 1) reproducibility of phenotypic changes during evolution, 2) mechanisms involved in eliciting these phenotypic changes, and 3) connection between observed phenotypes and underlying evolutionary changes. Here we answer these questions by demonstrating that laboratory adaptive evolution of *Escherichia coli* on lactic acid leads to a reproducible phenotype that exhibits improved fitness in multiple growth environments due to changes in a global regulatory mechanism that was discovered through detailed, quantitative phenotype testing and transcriptional profiling.

2 Introduction and Results.

A foundational concept in evolutionary biology is the notion that the evolutionary process brings organisms through a "fitness landscape" [4]. A fundamental question associated with this idea pertains to the reproducibility of the outcome of evolution. Traditionally, the fitness landscape is depicted as containing multiple peaks of improved fitness which implies the possibility of divergence during evolution. A contrasting perspective is correlated to computational modelling descriptions that imply a single global optimum phenotype [5]. To address these differing viewpoints, seven parallel evolution experiments (L2, L3, LA, LB, LC, LD, and LE) were conducted on lactate medium starting from a single parental colony of wild-type E. coli. Quantitative measurements of growth rate, substrate uptake rate (SUR), and oxygen uptake rate (OUR) were taken in replicate for the wild-type strain and all evolution strains at day 20 of evolution (approximately 300 generations) and at the end of evolution (day 60 or approximately 1,000 generations).

The evolutionary paths at day 20 and day 60 of evolution for all seven strains were traced from a common wild-type starting point, using the three measured parameters with growth rate as an indication of fitness (Figure 1A). Furthermore, multivariate ANOVA clustering (Figure 1B) reveals that 6 of the 7 evolved strain cluster at the end point of evolution. In the context of the "fitness landscape" concept, this indicates that six of the seven strains converged upon a single fitness peak by day 60 of evolution, while at day 0 (WT) and day 20 the strains functioned in different regions of the landscape.

[1] Bioinformatics Program, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0412, USA. E-mail: ajoyce@ucsd.edu

[2] Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0412, USA. E-mail: ssfong@ucsd.edu

[3] Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0412, USA. E-mail: bpalsson@be-research.ucsd.edu

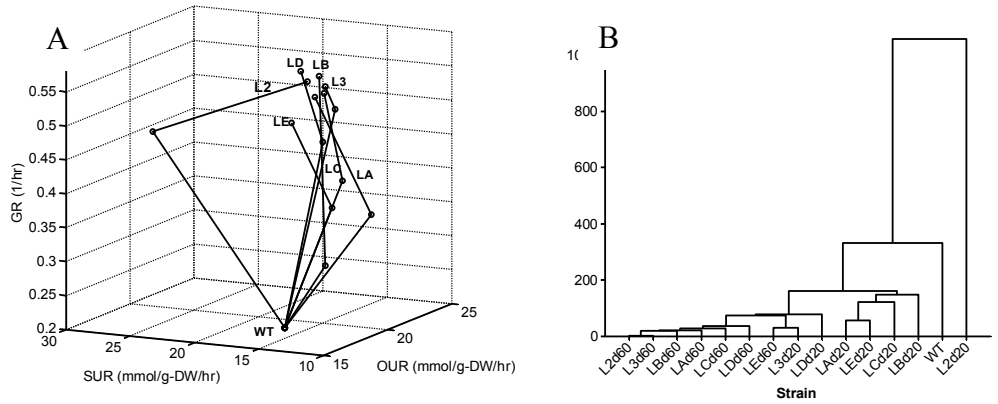


Figure 1: (A) Evolutionary Paths of Evolved Strains. (B) Multivariate ANOVA clustering of strains.

In an effort to probe the changes underlying these phenotypic improvements, we used Affymetrix E.coli Antisense Genome Arrays to measure global mRNA transcript levels in the wild-type and evolved strains. After normalization with dChip software, statistical filtering via ANOVA, and p-value cutoff selection by FDR [5], a pool of 1283 genes that showed significant differential expression in at least one strain at day 20 or day 60. The selected genes were then functionally characterized based on the MultiFun annotation scheme [6]. Interestingly, the strains exhibited similar expression patterns within day 20 and day 60 groups in that the proportion of genes changed relative to wild-type were similar in terms of functional classification. However, only 6 and 18 annotated genes were commonly differentially expressed across 6 of the 7 strains at day 20 and day 60, respectively.

3 Conclusions.

Based on growth, substrate uptake, and oxygen uptake rate measurements, parallel laboratory evolutions of *E. coli* can converge on a single reproducible phenotype, in agreement with existing computationally-based hypotheses [5]. Based on genome-wide transcriptional profiling, however, the evolved strains achieve improved fitness through apparently distinct mechanisms.

4 References.

[2] Buckling, A., M.A. Wills, and N. Colegrave. 2003. Adaptation limits diversification of experimental bacterial populations. *Science*, 2003. **302**(5653): 2107-9.

[5] Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* **57**: 289-300.

[1] Elena, S.F. and R.E. Lenski. 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet* **4**(6): 457-69.

[3] Elena, S.F. and R. Sanjuan, 2003. Evolution. Climb every mountain? *Science* **302**(5653): 2074-5.

[4] Ibarra, R.U., J.S. Edwards, and B.O. Palsson, 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature*. **420**(6912): 186-189.

[6] Serres, M.H., et al. 2001. A functional update of the Escherichia coli K-12 genome. *Gen Biol*, **2**(9): RES0035.

D5. Principal Component Analysis combined with probabilistic analysis of Gene Ontology as applied to neuroblastoma gene expression data

Alexei Krasnoselsky[1], Jun Wei, Sven Bilke, Quinrong Chen, Craig Whiteford and Javed Khan

Keywords: principal component analysis, gene expression, gene ontology, neuroblastoma

Gene expression profiling using micro-array technology has become an important tool in tumor classification and has been successfully applied to many human cancers. However, classifying subtypes of a cancer and establishing a link between the phenotype or known genomic abnormalities and the underlying gene expression is a more challenging task. Neuroblastoma (NB) presents one example of a highly heterogeneous cancer with multiple stages. NB is a childhood tumor of the sympathetic nervous system with a share of 8-10% of all childhood malignancies [1]. The clinical stage classification system defines stages 1 through 4 and a special case of 4S. There is a number of cytogenetic characteristics that distinguish NB, such as chromosomal abnormalities, or amplifications/deletions. Among the most well described cytogenetic markers is MYCN amplification. In this study we attempted to associate gene expression patterns with tumor phenotype, known genomic abnormalities and biological processes.

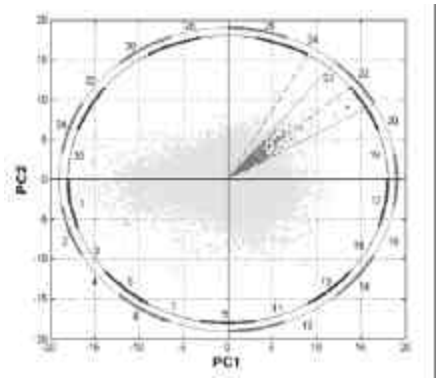
We carried out genome-wide expression profiling of 108 NB tumors (stages 1 trough 4, MYCN-amplified as well as non-amplified) using cDNA micro-arrays with 42293 clones. These tumors were of stages 1-4 with some of them with MYCN amplified. Quality filtered data consisting of 33680 clones corresponds to 22030 unique Unigene clusters. We employed PCA to reduce the complexity of the initial "tumor dimensions" from 108 to two principal components (PC) that are interpretable in terms of the tumor stage and MYCN amplification. In the subspace of these dimensions stage 1 (ST1) and stage 4 (ST4) tumors could be well separated, as well as MYCN-amplified from non-amplified tumors. Utilizing a "stepwise" PCA, whereby AMP tumors were gradually added to the dataset, PCA performed and the explained variance recorded, we were able to establish that MYCN amplification results in a different pattern of expression than in non-amplified tumors.

We performed further analysis in the PCA-reduced subspace by combining the expression data with the annotation information from the Gene Ontology (GO), as well as chromosomal location (cytoband assignment) for each gene. Our novel approach combines probabilistic analysis of GO and with gene expression profiles arranged in the PCA-reduced space of biologically interpretable PCs in attempt to extract important biological processes associated with amplification or stage progression in NB, as well as to associate known genomic abnormalities with the transcriptional activity in these tumors. In our algorithm, the gene projections on the first two PCs ("eigen-tumors"), representing stage and MYCN amplification dimensions are considered. The plane containing these projections is partitioned into segments as shown on Fig.1. Each segment is defined by the angle of inclusion and the angle of shift between the adjacent segments. Within each segment a sub-selection could be made for genes differentially-expressed between tumors of various phenotypes (e.g. stage 1 vs. 4, or MYCN not amplified vs. amplified). Thus, a bin is formed containing sub-selected genes from this segment. Next, for each gene in the segment GO annotations and chromosomal locations are obtained. Then, the probability of finding by random chance a chromosomal location or a specific node or leaf of GO is calculated for every Unigene cluster from the total number of occurrences in the bin and on the micro-array chip. This probability is further adjusted for multiple comparisons using the Bonferroni correction. The over-represented chromosomal locations are visualized in the PC subspace of eigen-tumors as shown on

[1]All authors are from the Advanced Technology Center, National Cancer Institute, NIH, Gaithersburg, MD, 20877
E-mail addresses in the authors order: krasnosa@mail.nih.gov; weij@mail.nih.gov; bilkes@mail.nih.gov; chenqi@mail.nih.gov; whitefoc@mail.nih.gov; khanjav@mail.nih.gov

Fig.2. These chromosomal locations are associated with specific patterns of expression and stage/MYCN amplification, based on their location in the eigen-tumor space.

Our analysis of NB tumors revealed associations between the known genomic abnormalities (such as chromosome 1p deletions) and the corresponding transcriptional activity, as well as major biological pathways differentiating stage and amplification. Fig.2 shows the chromosomal locations that are over-represented in the corresponding segments of genes differentially expressed between MYCN amplified and non-amplified tumors of ST4. In addition to known genomic abnormalities, we found new chromosomal locations that could be of potential interest.



.

Fig. Diagram of partitioning of PC subspace.

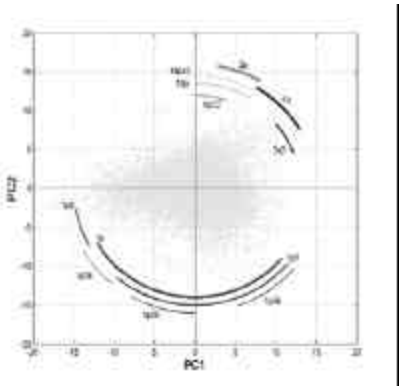


Fig.2 Cytobands that are over-represented in the MYCN amplified tumors.

The second component of our approach is the GO analysis of genes arranged in the eigen-tumor space as described above. Calculating and assigning probabilities of over-representation allowed us to select biological processes that are associated with gene expression patterns underlying a particular NB phenotype. We show that amplification of MYCN in ST4 strongly correlates with up-regulation of the genes involved in the ribosomal machinery, corroborating the results of several studies implicating MYCN in the regulation of ribosomal genes [2]. In contrast, the advanced stage (ST4) in the absence of MYCN amplification correlates with a significant increase in the expression of the genes involved in the cell cycle with no further increase in the expression of these genes upon MYCN amplification in ST4.

In conclusion, we developed an approach of analysis of gene expression data that combines *a priori* clinical-biological information with micro-array generated gene expression profiles. The combination of PCA and probabilistic GO analysis allows to extract biologically important features from the data and associate expression with genomic abnormalities and biological processes that may suggest a mechanism of tumor development. We applied this method to NB, where we validated our approach by confirming existing knowledge and providing new insights.

1. Brodeur, G.M., *Neuroblastoma: biological insights into a clinical enigma.* Nat Rev Cancer, 2003. **3**(3): p. 203-16.
2. Boon, K., Caron, H. N., van Asperen, R., Valentijn, L., Hermus, M. C., van Sluis, P., Roobeek, I., Weis, I., Voute, P. A., Schwab, M., Versteeg, R.., *N-myc enhances the expression of a large set of genes functioning in ribosome biogenesis and protein synthesis.* EMBO J, 2001. **20**(6): p. 1383-93.

D6. Discretization Methods for Expression Data

Sonia Leach, [1] [2]

Keywords: discretization, gene expression

1 Introduction

Gene expression data has become a popular source of genomic information. Often the real-valued measurements must be discretized. For example, a possible similarity metric in clustering is one based on mutual information for discrete data. [4] In learning regulatory networks, discrete models typically require fewer parameters to be estimated than their continuous counterparts, an important consideration given the small number of samples compared to the number of genes. [1, 5, 6]

In an earlier paper, we considered a number of discretization schemes for gene expression data, varying from fixed boundary techniques such as Fold-Change cut-offs (FC) or Equally Populated bins (EqPop), to per-gene variable boundary techniques which optimized some parameter of the original expression profile, such as Equal Range (EqRange), and Between-Gap-Minimizations (NN). [3] We also introduced two novel techniques that sought either to optimize the correlation between the real-valued gene expression profile of a gene with its discretized version (IDR), or sought to preserve the set of pairwise correlations of a gene's profile with all other genes (PDR). We discuss each of these briefly in the next section. Results from the earlier work are partially repeated here in Figure 1 and show that the two novel methods strongly outperform the other techniques on a variety of datasets which vary by organism, sample size, experimental condition, and microarray platform. [2] The performance is measured on the y-axis by the plotting the histogram of differences between the vector of pairwise correlation coefficients on the real-valued data and the vector of correlations on the discretized data.

2 Preserving correlation between Real Data and Discrete Data

The PDR and IDR algorithms were motivated by the observation that most clustering and modelling techniques try to capture the joint dependency structure between gene expression profiles. The IDR technique preserves the shape of the real-valued expression profile as much as possible by maximizing, over all possible 3-bin boundary assignments, the Pearson correlation coefficient between the resulting discretized levels and the real-valued levels. The PDR method preserves the joint pairwise correlation structure calculated on the real-valued data by considers all possible 3-bin boundaries for a single gene and minimizes, over all boundary choices, the Euclidean distance between the vector of pairwise correlations for the gene versus the other genes using the real valued data and the vector of pairwise correlations for the *discretized* version of the gene versus the real-valued profiles of the other genes. It differs from IDR since the optimal assignment for preserving profile shape might not be the optimal bin assignment for preserving pairwise correlational structure of the original data.

The two novel techniques, IDR and PDR, show strong performance over the other methods though at the expense of greater computation. Each method requires considering every

[1] University of Colorado Health Sciences Center, 4200 E. 9th, SOM C236, Denver CO, E-mail: sml@cs.brown.edu

[2] Brown University, 115 Waterman St, Providence RI

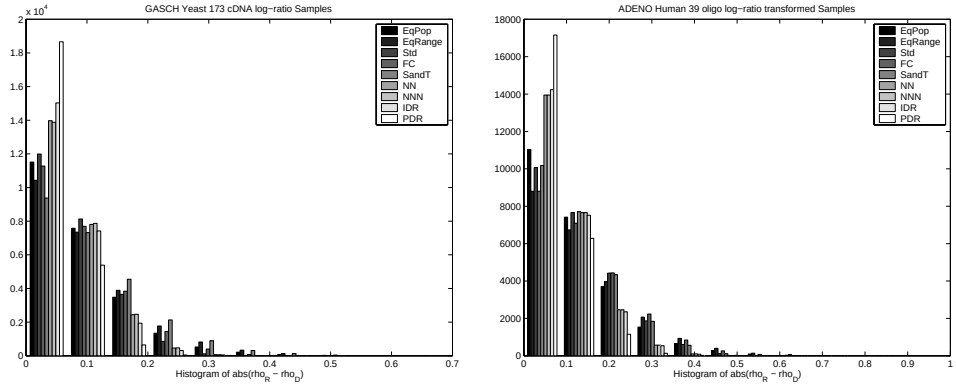


Figure 1: Histogram Of Differences of Pairwise Correlations in Real vs Discrete Data

possible 3-bin assignment for an expression profile, subject to the constraint that the median profile value be assigned the middle bin. For log-ratio expression data, the extra computation proves beneficial, and we found that the log-transformation in particular, not the ratio aspect, affords these two techniques the advantage (data not shown).

Of the two, IDR is less computational yet still offers improvement over the other methods. Moreover, unlike PDR, the bin assignments for a gene's profile are made independently of the profile data for other genes. In particular, the optimization problem of searching over bin assignments to preserve correlation between the real vector of expression levels and the discrete vector of expression levels lends itself to an efficient dynamic programming algorithm. In this poster, we will present the algorithm and the results of using the algorithm on an expanded collection of datasets, including more oligo-based datasets.

References

[1] Friedman,N., Linial, M., INachman, I., Pe'er, D. 2000. Using Bayesian Networks to Analyze Expression Data. *J Comp Biol* 7(3-4):601–620

[2] Gasch A.P., Spellman, P.T., Kao, C.M., Carmel–Harel, O., Eisen, M.B., Storz, G., Botstein, D., Brown, P.O. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell.*, 11(12):4241–57.

[3] Leach, S. 2004. Comparing Discretization Techniques for Gene Expression Data, *Proc of Intelligent Systems for Molecular Biology (ISMB-04)* submitted.

[4] Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X., and Somogyi, R. 1998. Cluster Anaysis and Data Visualization of Large-scale Gene Expression Data, *Pac Symp Biocomput PSB-98*, pp. 42-53.

[5] Pe'er, D. Regev, A., Elidan, G., and Friedman, N. 2001, Inferring subnetworks from perturbed expression profiles,*Bioinformatics* 17 Suppl.1:S215–S224.

[6] Yoo, C., Thorsson, V., and Cooper, G.F. 2002. Discovery of causal relationships in a gene regulation pathway from a mixture of the experimental and observational DNA microarray data, *Pac Symp Biocomput PSB-02*, pp. 498-509.

D7. Suitability of Spherical SOM for Gene Expression Analysis

Hirokazu Nishio, [1] **Ken-nosuke Wada,** [2] **Yoshiko Wada,**[2]
Md. Altaf-Ul-Amin,[1] **Shigehiko Kanaya**[1]

Keywords: Spherical Self-Organizing Maps, microarray, gene expression, dimension reduction, N-measure, distortion

1 Introduction.

Self-Organizing Maps(SOM)[1] are visualization methods by creating dimension-reduced feature maps, where units in low-dimensional represantaion space are associated with those in original high-dimentional data space. In the present study, we consider two-types of SOMs, one has planar representation space called 'Plain SOM' and the other has spherical space called 'Spherical SOM'[2]. In general, the profile vectors of gene expression are normalized to unity in length in order to focus not on the absolute quantity of expression but the similarity of the direction[3]. Since the normalized vectors are distributed on the surface of a supersphere, it tempts to think that the Spherical SOM is suitable for the data of gene expression. But this idea had no ground because the surface of a sphere is topologically different from that of a supersphere. In the present study, we propose the measure for suitablity of SOM to a given data set, and confirmed that the Spherical SOM was actually suitable for the data of gene expression.

2 Method.

If a SOM has the representation space which can express the structure of data without distorting, distance between units in representation space should be proportional to that in original data space. This is quantified by the linearity of 'd-d plot' as follows: All pairs of the units are plotted in the two-dimensional plane. Horizontal axis corresponds to the distance in representation space, and the vertical axis corresponds to the distance in data space. The linearity of a distribution is quantified by the coefficient of determination of regression analysis, which is called 'N-measure'.

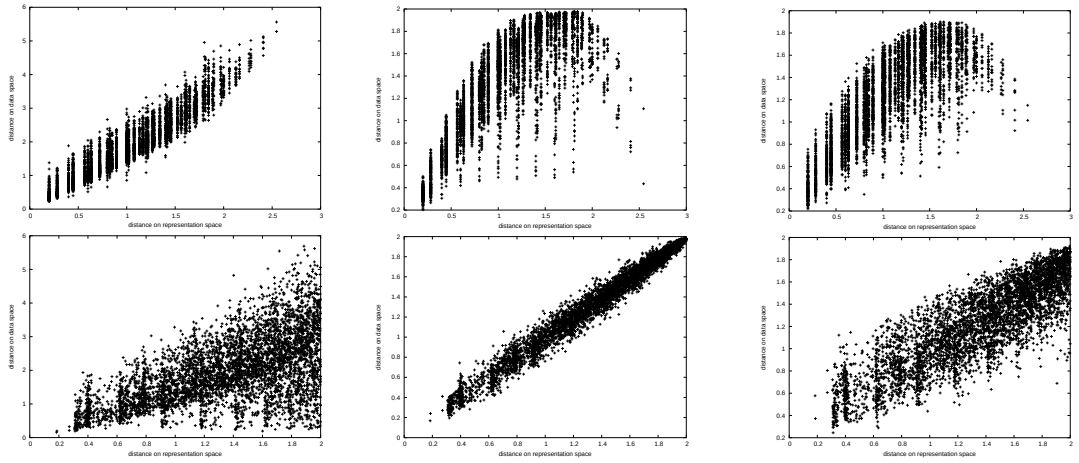
3 Result and Discussion.

In order to demonstrate the validity of N-measure, we examine three data sets as follows:

- Randomly distributed 1800 points on a plain surface consisting of two axes. Variance of each axis is unity.
- Randomly distributed points on a sphere surface consisting of three axes. Initially three dimensional 1800 vectors are randomly generated. Here, variance of each axis is unity. Then all the vectors are normalized to unity in length.
- Actual 8-dimensional normalized gene expression profiles of *B. subtilis*.

[1]Nara Institute of Science and Technology, Graduate School of Information Science, 8916-5 Takayama, Ikoma, Nara 630-0101. E-mail: hiroka-n@is.aist-nara.ac.jp
[2]UNTROD Inc. 4-4-123 Aoyama, Narashi, Nara-ken, 630-8101

Figure 1 shows the d-d plots for those data with Plain SOM (upper) and Spherical SOM (lower). In case of plain surface data, a linear relationship is observed in the d-d plot of Plain SOM (left upper). The N-measure of Plain SOM (0.8901: see Table 1) is greater than that of Spherical SOM (0.3212). On the other hand, in case of sphere surface data, a linear relationship is observed in the d-d plot of Spherical SOM (middle lower). The N-measure of Plain SOM (0.3782) is less than that of Spherical SOM (0.9764). Thus, suitablity of SOM to a given data set can be estimated by N-measure. The N-measure increases when a SOM has the representation space which can express the structure of data without distorting. In case of actual 8-dimensional gene expression data, the N-measure of Plain SOM (0.1682) is less than that of Spherical SOM (0.6044). We conclude the Spherical SOM is suitable for the analysis of the normalized data of gene expression.



data on plain surface　　data on sphere surface　　normalized gene expression data

Figure 1: d-d plot(upper:Plain SOM, lower:Spherical SOM)

	Plain SOM	Spherical SOM
data on plain surface	0.8901	0.3212
data on sphere surface	0.3782	0.9764
data of normalized gene expression	0.1682	0.6044

Table 1: coefficient of determination

References

[1] Teuvo Kohonen. Self-Organizing Maps. Springer Series in Information Sciences. Springer, second edition, 1997.

[2] Helge Ritter. Self-Organizing Maps on non-euclidean Spaces. Kohonen Maps, pp. 97-108, ed. E. Oja and S. Kaski, 1999

[3] Hirokazu Nishio, Md. Altaf-Ul-Amin, Tetsuo Sato, Ken-nosuke Wada, Yoshiko Wada, Kotaro Minato, Kazuo Kobayashi, Naotake Ogasawara, Shigehiko Kanaya. Visualization of Gene Classification Based on Expression Profile Using BL-SOM. Proc. WSOM'03, pp. 101-106, 2003.

D8. *In Silico* Identification and Analysis of Tissue-Specific Genes using the Database of Human Expressed Sequence Tags

Sheng-Ying Pao [1, 2], Win-Li Lin [1], Ming-Jing Hwang [2]

Keywords: tissue-specific genes, genome wide expression profile, EST

1 Introduction.

Genome wide transcriptome analysis with histological information can provide insights to identify candidate genes that are differentially expressed in certain tissues. At the transcriptome level, differential expression of genes plays key roles in maintaining and regulating cellular functions. Genes preferentially expressed can be characterized by their significantly different expression levels of transcripts in various tissues. In this study, tissue-specific genes were identified using human expressed sequence tags (EST).

2 Methods.

EST data, GenBank reports from dbEST, and UniGene build #161 were downloaded from NCBI (National Center for Biotechnology information). We extracted a description triplet for each EST library under the field "Title" (Lib. Name), "Tissue" and "Organ". According to the triplet, each library was classified into a corresponding category of tissue. The classification process is illustrated in fig.1 wherein libraries are automatically classified by tissue and organ unless they are different, both null, or inconsistent with the title; in these cases, manual classifying and checking with title was carried out. To mitigate variation due to unspecified tissue and artificially modified expression, libraries described as subtracted, differentially displayed, normalized, or coming from multiple tissues were excluded. Libraries without a clear description in the triplet were also discarded. By subjectively selecting category names of classified libraries, a "target_tissue list" consisted of interested tissue names was then generated.

For each category in the "target_tissue list", the corresponding EST gi numbers and the UniGene clusters to which they belong were retrieved automatically. For each target tissue in the "target_tissue list", we performed the differential expression evaluation according to [1]:

$$p(y|x) = \left(\frac{N_2}{N_1}\right)^y \frac{(x+y)!}{x!y!\left(1+\frac{N_2}{N_1}\right)^{(x+y+1)}}$$

where x and y are the number of ESTs matching the UniGene cluster from the target tissue and from all the other tissues, respectively. Likewise, N1 and N2 are the total number of ESTs from the target tissue and from all the other tissues, respectively. Clusters with $N_1<1000$, $N_2<1000$, $x \leqq 0.05*N_1$, or $y \leqq 0.05*N_2$ were excluded from the test for insufficient sample size.

[1]Institute of Biomedical Engineering, National Taiwan University College of Medicine, No.1 Jen-Ai Road Section 1, Taipei, Taiwan. E-mail: r91548018@ntu.edu.tw

[2] Institute of Biomedical Sciences, Academia Sinica, No.128 Yen-Chiu-Yuan Road Section 2, Taipei, Taiwan. E-mail: mjhwang@ibms.sinica.edu.tw

3 Results and discussion.

We discarded 1898 EST libraries as described in methods, leaving 6,247 libraries with 3,352,546 ESTs for analysis. Along with UniGene build #161, they provide sufficient data to identify tissue specific genes by *in silico* data mining. For theses libraries we constructed a tissue hierarchy and classified the libraries into 343 categories of tissues, of which we selected 113 categories to form a target_tissue list to detect differentially expressed genes. This "target_tissue list" provides a flexible framework for detecting differential expression in tissues of interest for various research purposes aiming to, for example, analyze the expression divergence of genes in one special cell type from multiple tissues, or to compare genes specific to diseased and normal states.

To evaluate our approach, we compared the placenta-specific genes identified in our results with those of Ref [2], in which 90 preferentially expressed genes were reported, of which 19 have subsequently been removed from UniGene and have become nonexistent in UniGene build #161. Our analysis yielded 508 genes preferentially expressed in placental libraries with $p<10^{-6}$, which included all of the 71 genes reported in [2] that remained existent in UniGene build #161.

In summary, we have constructed a tissue hierarchy for EST libraries and identified differential expressed genes in 113 tissues. The tissue-specific gene information derived from this study will be useful in functional genomics research.

4 Figures.

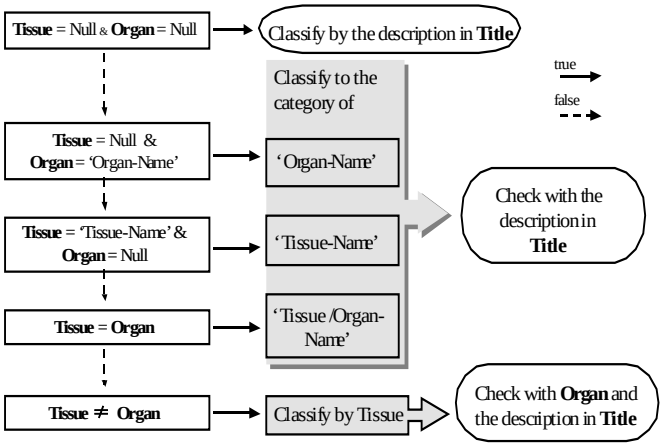


Figure 1. Procedure for EST library classification.

References

[1] Audic, S. and Claverie, J.M. 1997. The significance of digital gene expression profiles. *Genome Research*., 7: 986-995.

[2] Miner, D. and Rajkovic, A. 2003. Identification of expressed sequence tags preferentially expressed in human placentas by in silico subtraction. *Prenatal Diagnosis*., 23(5):410-419.

D9. Regulation of NF-κB responsive genes in a single cell

P. Paszek[1], T. Lipniacki[2], A. Brasier[3], B. Tian[3], B. Luxon[3], M. Kimmel[4]

Keywords: Stochastic regulation of transcription, early and late genes, NF-kappaB

1 Experiment and problem formulation.

Nuclear factor κB (NF-κB) regulates numerous genes important for pathogen or cytokine inflammation, immune response, cell proliferation and survival. In resting cells it remains inactive in cytoplasm, bond to its inhibitor IκBα. In response to extracellular signals such as TNF, IκBα is destroyed, NF-κB enters the nucleus, binds to specific regulatory sites and triggers gene transcription. Our observations [1] show (Fig. 1) that in HeLa cells, the NF-κB responsive genes can be grouped into 3 characteristic classes: early (such as IκBα, A20 or IL8) for which the amount of mRNA transcript has its maximum at about 1 hour, intermediate (such as NF-κB1 or TNFAIP2) with the maximum at 3 hours and late (such as NAF1 or NF-κB2) with the maximum at about 6 hours. In contrast to some other cell lines, in HeLa cells NF-κB is not effectively lead out of the nucleus by the newly synthesized IκBα, but rather, after entering the nucleus at 15 min from the beginning of TNF stimulation, it remains there for at least 6 hours [1]. This implies that some other cofactors are needed to initiate and terminate expression of genes belonging to these 3 groups. The aim of this work is to propose the mechanisms of gene regulation at a single cell level able to generate these 3 characteristic classes of expression profiles involving a small number of co-regulators. We will not try to identify these co-regulators, which at this point should be understood in broad sense as activating (e.g. histone acetylation) or repressing events, not necessarily connected with DNA protein binding.

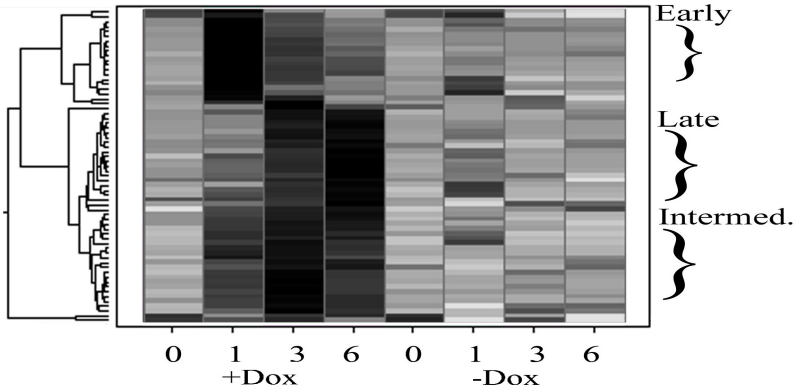


Figure 1: Kinetics of NF-κB-dependent gene expression in HeLa cells [1]. Data represent the mean of three independent time courses analyzed by high-density microarrays. In this experiment the NF-κB nuclear translocation is enabled by culturing cells in the presence of doxycycline (Dox). The expression profiles for selected genes of each group where confirmed by Northern blots.

[1] Department of Statistics, Rice University, Houston, TX, USA, E-mail: ppaszek@rice.edu

[2] IPPT PAN, Warsaw, Poland and Department of Statistics Rice University, E-mail: tomek@rice.edu

[3] University of Texas Medical Branch, Galveston, TX, USA

[4] Department of Statistics, Rice University, Houston, TX, USA, E-mail: kimmel@rice.edu

2 Model and results.

The observed profiles of expression may be explained by assuming that all 3 classes of genes are regulated by at most 2 co-activators (not including NF-κB) and 1 repressor. We assume that: (1) Each gene has two potentially active alleles, and the activation and repression of these alleles proceed independently. (2) Expression of any early gene is initiated by NF-κB binding, while to initiate expression of an intermediate gene additionally one co-activator is needed, and to initiate expression of late genes two co-activators are needed in addition to NF-κB. (3) Transcription of all genes is terminated by a repressor. (4) All genes have the same mRNA degradation halftime equal to 30 min. (5) Binding of activators and repressor occur in a stochastic way, with binding halftimes (chosen to fit experimental expression profiles) of 10min, 70min, and 4 hours, respectively, for activators and of 70min for the inhibitor.

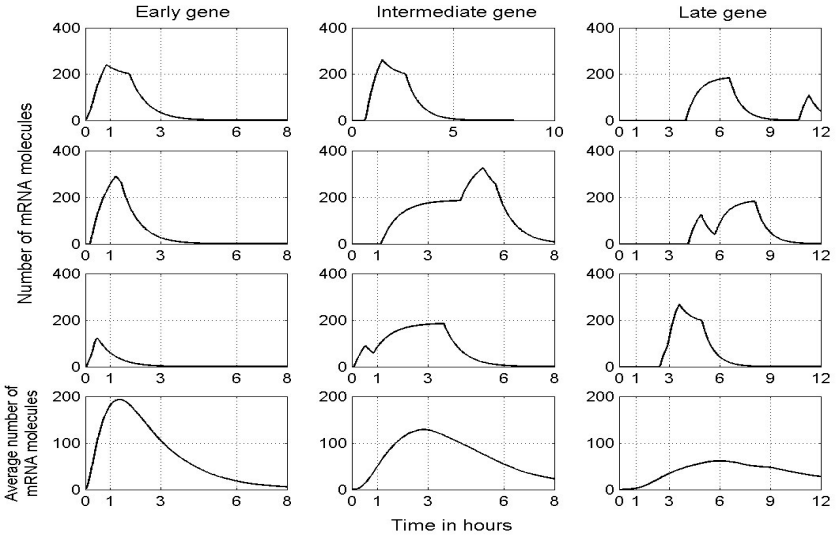


Figure 2: The mRNA profiles for early, intermediate and late genes. First 3 rows show mRNA profiles in single cells, while the profiles in last row result from averaging over population of 1000 cells. These latter profiles should be compared to experimental data in Fig. 1. The kinks visible on single cell profiles correspond to initiation or termination of expression in any of two homologous copies of the gene. The difference among single cell profiles is larger for late genes and as a result the averaged expression profile is broader, what is well confirmed in Northern blot data.

We have shown that the three classes of profiles can be explained by a relatively small number of regulatory factors. Moreover, our analysis shows that, especially for late genes, the single cell mRNA profiles may be very different from the averaged profile. The single cell experiments testing this hypothesis are currently underway.

References

[1] Tian, B., Brasier, A. R., unpublished data

D10. Stochastic Models Inspired by Hybridization Theory Improve Measures of Gene Expression in Affymetrix geneChip arrays

Zhijin Wu [1] **Rafael Irizarry** [2]

Keywords: oligonucleotide arrays, expression measure, empirical Bayes

1 Introduction.

High density oligonucleotide arrays have been widely used to measure gene expression in many areas of biomedical research. Affymetrix GeneChip arrays are among the most popular. In the GeneChip system, a fair amount of pre-processing and data reduction occur after the image processing step, including background adjustment, normalization and summerization.

In Affymetrix GeneChip arrays, each gene is probed by 11-20 short oligonucleotides that perfectly match to mRNA sequence of the gene, refered to as perfect match (PM) probes. In order to account for optical noise and cross-hybridization, Affymetrix pairs each PM probe with a mismatch probes (MM) that is identical to the PM except the middle base. The intensities observed on the MM probes are subtracted from the PM intensities to give background-corrected intensities in the default algorithm (MAS 5.0) for expression measure. The background-adjusted intensities from each probeset is summarized to define the expression measure of the corresponding gene.

Various alternative algorithms of obtaining expression measures have been developed and shown to outperform MAS 5.0 ([1], [2]). Among those the robust multi-array analysis (RMA) has become a popular substitute for MAS 5.0. RMA uses a global background adjustment that substabtial gains in precision with minor sacrifices in accuracy.

The accuracy of RMA can be improved when background is adjusted in a more specific fashion. The challenge is to do so without much sacrifice in precision. Since the release of the probe sequences by Affymetrix, its relationship with background hybridization has been observed by different groups. A stochastic model based on molecular hybridization theory has been proposed to describe various components of probe intensity([4]). In this poster we present an improved expression measure based on this sequence dependent model.

2 Method and Result

Our model for the PM intensity contains optical noise, non-specific binding (NSB) and gene-specific binding (GSB). For each PM/MM probe pair,

$$\begin{aligned} PM &= O + N_{PM} + S, \text{where} \log(S) = s + a \\ MM &= O + N_{MM} \end{aligned} \tag{1}$$

The optical noise O is considered to be a global effect identically distributed for each probe. The N_{PM}, N_{MM} refer to NSB for the PM and MM probes respectively. S stands for the observed specific intensity and is composed of probe effect a and specific signal s.

[1]Department of Biostatistics, Johns Hopkins University. E-mail:zwu@jhsph.edu
[2]Department of Biostatistics, Johns Hopkins University. E-mail:rafa@jhu.edu

Strong probe effect a in detecting signal was first observed by Li and Wong[1]. Naef and Magnasco [3] propose a simple yet effective model, in which the affinity of a probe is described as the sum of position-dependent base affinities, for predicting a. We propose that both specific binding and non-specific binding can be described with Naef and Magnasco's model.

Since the optical noise variance is negligable when comared to the variance due to the NSB component, we adjust for O by considering it as a constant. The NSB components $\log(N_{PM}), \log(N_{MM})$ are modeled as correlated bivariate normal random variables with means related to the probe affinities. We obtain the specific binding by minimizing the mean squared error of s: $\hat{s} = \mathrm{E}[s|PM, MM]$. For the expression measure presented in this poster we imposed a uniform distribution of s.

After extracting specific signal from each probe, we quantile normalized the adjusted intensities and obtained an expression measure using median polish on the normalized and background adjusted intensities. We denote the new expression measure GCRMA because it uses base (GCTA) composition of probes and same summerization and normalization as RMA.

Assessments on the Latin-square dataset provided by Affymetrix suggest that with comparable precision to RMA, GCRMA has much better accuracy than MAS 5.0 and some of the widely used alternativevs including RMA, dChip and PerfectMatch. The improvement in accuracy is most obvious in moderate expression levels, where the bulk of the non-spiked-in genes fall in the intensity distributoin.

The expression measures obtained using GCRMA can be used in further analysis, such as differential expression estimation. However, with model (1) in place we can obtain estimates of differential expression and standard errors for those estimates directly from the probe level data. In this poster we present some preliminary results related to this work.

References

[1] C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science U S A*, 98:31–36, 2001.

[2] R. Irizarry, F. C. B. Hobbs, Y. Beaxer-Barclay, K. Antonellis, U. Scherf, and T. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003. Summaries of affymetrix genechip probe level data. *Nucleic Acids Research*, 31, 2003.

[3] F. Naef and M. O. Magnasco. Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. *Physical Review E*, 68:011906, 2003.

[4] Wu, Z. and Irizarry, R.A. (To appear) Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays. *Proceedings of RECOMB 2004*.

[5] L. Zhang, L. Wang, A. Ravindranathan, and M. Miles.

D12. ESTmapper: Efficiently Clustering EST Sequences Using Genome Maps

Xue Wu, Woei-Jyh (Adam) Lee, Damayanti Gupta, Chau-Wen Tseng [1]

Keywords: Bioinformatics, EST clustering, genome map, suffix tree, high performance computing

1 Introduction

Expressed sequence tags (ESTs) are nucleotide sequences of transcribed genes that provide important information about gene discovery and gene expression in many organisms. Because individual ESTs are incomplete, many ESTs must be *clustered* to discover the full sequence for each gene. Many earlier EST clustering algorithms (TGICL [1], d2_cluster [2], GeneNest [3], CLOBB [4]) rely on performing pairwise comparisons between ESTs, with closely matching pairs put into a single cluster. But obviously these algirithms do not scale well because of the $O(n^2)$ number of pairwise comparisons needed. More recent approachs (PaCE [5], Xsact [6]) build a *suffix tree* of all ESTs to identify pairs of ESTs with long common substrings. But these algorithms are still pairwise comparison based, and the time complexity are $O(n \log n)$. In addition, since ESTs represent large numbers of incomplete, error-prone, and redundant fragments of genes, the above EST to EST comparison based algorithms cannot guaranteen the accuracy of the clustering results. Biologists' help are needed to correct and assure the precision of the final results. With the advent of high-throughput sequencing of the entire genomes of many species, an alternative approach becomes possible. In this paper, we describe ESTmapper, a new tool for clustering EST sequences based on efficiently mapping ESTs to the genome.

2 ESTmapper

The algorithm used by ESTmapper consists of two steps: preprocessing genome and clustering ESTs. It first builds a suffix tree for the genome, then searches for long common substrings between each EST and the genome. We use them to build gapped matching regions to account for sequencing errors and splicing, and use the longest overall matching region to map the EST to locations in the genome. ESTs mapped to overlapping locations are then placed in a cluster. Preliminary experiments show that ESTmapper is not only very efficient for its linear processing time, but also precise with respect to the input EST data when compared to other popular EST clustering tools.

3 Performance and Statistics

We compared the performance of ESTmapper, TGICL and PACE on a AMD Athalon PC cluster. The dataset are a selection of ESTs from Arabidopsis thaliana and its second chromosome. TGICL and ESTmapper are executed on a single PC node, but PaCE is executed on 8 nodes. Resutls are presented in the following table. We also evaluated the

[1] Computer Science Department, University of Maryland at College Park. E-mail: {wu,adamlee,dami,tseng}@cs.umd.edu

scalability of ESTmapper. Figure 1 shows its running time when mappig from 250K to 1.5 million human ESTs against human chromosome 21.

# ESTs	Execution time (seconds)			Memory use (MB)		
	TGICL	PaCE	ESTmapper	TGICL	PaCE	ESTmapper
2,980	14	131	72	77	103	661
5,960	34	282	85	79	106	661
11,921	73	686	111	81	112	661
23,842	142	872	174	85	181	661
47,685	313	1702	278	88	336	662
95,370	675	4285	460	102	603	662
190,740	1987	-	861	129	1192	663

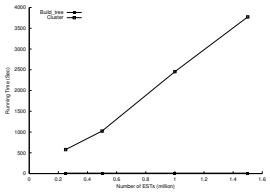


Figure 1: Scalability of ESTmapper

In addition, ESTmapper also provides various statistics about the mapping and clustering results. Table 1 shows several parameters that affect ESTmapper's clustering precision. The dataset are 2224 Arabidopsis ESTs and its second chromosome. At the same time the results also gives one example of statistics information provided by ESTmapper.

Min Common Substring length	Number Unmapped ESTs	Average Mapping Region length	mapping region/EST length ratio	average gap length	average gap/mapping length ratio	average number gaps	number cluster	number singleton
10	0	408	97.80%	1	0.04%	0.1	711	310
20	0	411	98.50%	1	0.07%	0.2	711	310
40	0	407	97.60%	3	0.04%	0.1	713	311
80	15	407	97.30%	1	0	0.02	708	309
160	91	416	97.20%	0	0	0.005	689	302
320	505	452	97.60%	0	0	0	568	263
640	2166	688	97.90%	0	0	0	50	53

Table 1: Impact of minimum common substring length

4 Conclusions

The measured performance results show that ESTmapper has constant memory usage and linear execution time with regard to the number of processed ESTs. Since the size of genome is relatively constant, our EST clustering is more scalable than other clustering algorithms. Besides, our EST clustering tool can also provide useful statistic information about the clustered ESTs, which we believe can help biologists with their research.

References

[1] G. Pertea, X. Huang, F. Liang and V. Antonescu et al, TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets *Bioinformatics* Vol 19(5), pp. 651 - 652; March, 2003

[2] J. Burke, D. Davison and W. Hide, d2_cluster: A Validated Method for Clustering EST and Full-Length cDNA Sequences *Genome Res.*, 9(11): 1135 - 1142; November, 1999

[3] S. Haas, T. Beissbarth, E. Rivals, A. Krause and M. Vingronu, GeneNest: automated generation and visualization of gene indices *Trends Genet.*, 16(11), 521-523; 2000

[4] J. Parkinson, D. Guiliano and M. Blaxteur, Making sense of EST sequences by CLOBBing them *BMC Bioinformatics*, Vol 3(1):31; 2002

[5] A. Kalyanaraman, S. Aluru, S. Kthari and V. Brendeul, Efficient clustering of large EST data sets on parallel computers *Nucleic Acids Research*, Vol 31(11), pp. 2963-2974; 2003

[6] K. Malde, E. Coward and I. Jonassen, Fast sequence clustering using a suffix array algorithm *Bioinformatics*, Vol. 19, pp. 1221-1226; 2003

E1. Improving Extreme Pathway Computations

Steven L. Bell[1] and Bernhard Ø. Palsson[2]

Keywords: metabolic networks, stoichiometric matrix, sparse matrices, convex cone

1 Introduction.

The abundance of genomic data available today allows for construction of genome-scale metabolic networks for many organisms. The topology of the type of networks considered here is determined by an $m \times n$ stoichiometric matrix, $\mathbf{S}$, whose rows and columns represent the system's metabolites and reactions, respectively. The dynamics of the system is given by $\dot{\mathbf{x}}(t) = \mathbf{Sv}$, where $\mathbf{x}$ is the m-dimensional vector of metabolite concentrations, " $\dot{}$ " denotes time-derivative, and $\mathbf{v}$ is a vector of fluxes which we assume is independent of concentrations and time.

Under the assumption that the system is in steady-state, we have that $\mathbf{Sv} = \mathbf{0}$, and to obtain biologically feasible solutions to this equation, we also impose the condition that $\mathbf{v} \geq \mathbf{0}$. The solution set is a so-called *convex cone* which can be generated by a finite (and unique up to a multiple) number of vectors, i.e., each biologically feasible flux vector (when the system is in steady state) can be expressed as a non-negative linear combination of these *extreme pathways* [2]. The extreme pathways are the edges of the convex cone, or more precisely, they are *conically independent*, i.e., no such vector can be expressed as a non-negative linear combination of any other vectors in the cone.

Given a metabolic network, where the metabolites are represented by the nodes and the edges represent the associated reactions, we compute the extreme pathways using an algorithm presented in [3] (see also [4]). The algorithm uses matrix operations similar to those used in the well-known Gaussian elimination algorithm. Such operations require frequent access to memory, significantly degrading performance of the algorithm if large matrices are stored. Furthermore, the computational time (and the number or extreme pathways) typically grows exponentially as the size of the network grows linearly. Existing implementations work well for relatively small networks, but are of limited use for genome-scale systems. Here, we propose two means to improve the performance of computing extreme pathways: an efficient sparse matrix storage and computational procedure and a scheme to select pivoting columns which we will refer to as the *exponential threshold method*.

2 Description of the algorithm

We now give a simplified version of the extreme pathway algorithm. The algorithm may be described as a sequence of tableaux $T^0, T^1, \ldots, T^N$, where the initial tableau is given by $T^0 = [\,\mathbf{I} \quad \mathbf{S}'\,]$, and the final tableau $T^N = [\,\mathbf{P} \quad \mathbf{0}\,]$. In the initial tableau, $\mathbf{S}$ is the $m \times n$ stoichiometric matrix, "prime" denotes transpose, and $\mathbf{I}$ is the $n \times n$ identity matrix (and hence T^0 is an $n \times (n+m)$ matrix). The final tableau, T^N, for some $0 < N \leq m$, consists of the matrix $\mathbf{P}$ whose rows are the extreme pathways, and the zero matrix, $\mathbf{0}$, which has m columns. Converting the right hand matrix $\mathbf{S}'$ to the zero matrix is done column by column

[1]Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, San Diego, CA. 92093-0412 E-mail: sbell@ucsd.edu

[2]Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, San Diego, CA. 92093-0412 E-mail: bpalsson@bioeng.ucsd.edu

using elementary row operations, each tableaux corresponding to a column. For $1 \leq i \leq N$, the tableau T^i is obtained from T^{i-1} by first choosing a pivoting column of the right hand matrix (originating from $\mathbf{S}'$) to zero out, column j, say. Suppose there are `pos` positive, `neg` negative, and `z` zero elements in column j. First, the `z` rows of T^{i-1} containing a zero in column j are copied to T^i. Then each of the `pos` rows which contain positive elements in column j is combined (using an elementary row operation) with each of the `neg` rows which contain negative elements in column j so that a zero is produced in column j of T^i. More precisely, if $T^{i-1}_{s,j} > 0$ and $T^{i-1}_{t,j} < 0$ for some s and t, then $|T^{i-1}_{t,j}|\, T^{i-1}_s + |T^{i-1}_{s,j}|\, T^{i-1}_t$ is the new row to be added to T^i. (Here, T^{i-1}_s denotes the s^{th} row, $T^{i-1}_{s,j}$ is the (s,j)-element in the tableau T^{i-1}, and $|x|$ is the absolute value of x.) Finally, all rows which are not conically independent are deleted from T^i. Hence, the number of rows in T^i is at most $\mathtt{z} + \mathtt{neg} * \mathtt{pos}$.

3 Sparse matrices and the exponential threshold method

The stoichiometric matrix contains few non-zero elements (about 5%) and although the final pathway matrix is less sparse (about 25% non-zero elements), we believe that sparse matrix methods used with success in similar algorithms, such as Gaussian elimination, can also benefit implementations of the extreme pathway algorithm. Memory is conserved since only non-zero entries of matrices are stored, and computational performance is improved since row operations use only non-zero elements of the vectors. There are many storage schemes for sparse matrices each with its own type of data structures, and the problem usually dictates which particular scheme is employed ([1], pg. 37). From Section 2, we see that for the extreme pathway algorithm it is important that individual column elements and whole rows can be accessed efficiently, so the storage method must be designed with these objectives in mind.

In Section 2 we saw that the aim of the extreme pathway algorithm is to zero out the columns of the tableaux using elementary row operations. Furthermore, for a fixed iteration, the number of rows to be added to the next tableau depends on `pos` and `neg`, the number of positive- and- negative elements, respectively, in the pivoting column. Our proposed method dictates that a column is processed only if $\mathtt{pos} * \mathtt{neg} < \mathbf{T}$, where $\mathbf{T}$ is some threshold. When all the columns of a tableau have been tested, the threshold is raised (exponentially) and the process starts anew, if there are any remaining columns to be zeroed out. The rationale for the threshold method is that the sparse columns are processed first since they require less computation, and postponing processing the denser columns may result in some of their non-zero elements being zeroed out by the elementary row operations performed in the earlier iterations, i.e., doing less work early may reduce the amount of work that has to be done later, resulting in less overall work. Preliminary results seem to indicate that this is indeed the case.

References

[1] Duff, I. S., Erisman, A. M. and Reed, J. K. 1986. *Direct Methods for Sparse Matrices*. New York: Oxford University Press.

[2] Rockafellar, R. 1970. *Convex Analysis*. Princeton: Princeton University Press.

[3] Schilling, C. H., Letscher, D. and Palsson, B. Ø. 2000. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology* 203:249–283.

[4] Schuster, R. and Schuster, S. 1993. Refined algorithm and computer program for calculating all non-negative fluxes admissible in steady states of biochemical reaction systems with or without some flux rates fixed. *Computational and Applied Bioscience* 9:79–85.

E2. Analysis of Heterogeneous Regulation in Biological Networks

Irit Gat-Viks [1] [2], Amos Tanay [1] [2], Ron Shamir [1]

Keywords: genetic and metabolic regulation, high-throughput data analysis, model learning algorithms.

1 Motivation.

Biological systems employ heterogeneous regulatory mechanisms that are frequently interrelated. For example, the rates of metabolic reactions are strongly coupled to the concentrations of their catalyzing enzymes, which are themselves subject to complex genetic regulation. Such regulation is in turn frequently affected by metabolite concentrations. Metabolite-mRNA-enzyme-metabolite feedback loops have a central role in many biological systems and call for modeling and learning of heterogeneous regulatory mechanisms.

2 Method.

In this work we study steady state behavior of biological systems that are stimulated by changes in the environment (e.g., lack of nutrients) or by internal perturbations (e.g., gene knockouts). Our model of the system contains variables of several types, representing diverse biological factors such as mRNAs, proteins and metabolites. Interactions between biological factors are formalized as regulation functions which may involve all variable types and complex combinatorial logic. Our model combines metabolic pathways (cascades of metabolite variables), genetic regulatory circuits (networks of mRNAs and transcription factors protein variables), protein networks (cascades of post-translational interactions among protein variables) and the relations among them (metabolites may regulate transcription, enzymes may regulate metabolic reactions). We show how such models can be built from the literature and develop computational techniques for their analysis and refinement given a collection of heterogeneous high-throughput experiments. We develop algorithms to learn novel regulation functions in lieu of ones that manifest inconsistency with the experiments.

Our approach is innovative in several aspects:

• We model a variety of variables types, extending beyond gene network studies, that focus on mRNA, and metabolic pathways methods, that focus on metabolites. Consequently, our model can express effects of translation regulation and post translational modifications.

• Our approach allows handling feedback loops as part of the inference and learning process. This is crucial for adequate joint modeling of metabolic reactions and genetic regulation.

• We build an initial model based on prior knowledge, and then aim to improve (expand) this model based on experimental data. We show that formal modeling of the prior knowledge allows the interpretation of high throughput experiments in new level of detail.

• Our algorithms learn new transcription regulation functions by analyzing together gene expression, protein expression and growth phenotypes data.

[1]School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. Email: {iritg,amos,rshamir}@post.tau.ac.il.

[2]These authors contributed equally to this work.

3 Experimental Results

Our methodologies and ideas were implemented in a new software tool called MetaReg. It facilitates evaluation of a model versus diverse experimental data, detection of variables that manifest inconsistencies between the model and the data, and learning optimized regulation functions for such variables. Graphical representations provide overview of results and highlight model inconsistencies. We used MetaReg to study the pathway of lysine biosynthesis in yeast. We performed an extensive literature survey and organized the knowledge on the pathway into a model consisting of about 150 variables. As part of model construction, we reviewed the results of many low throughput experiments and included in the model the most plausible regulation function of each variable. We assessed the model versus a heterogeneous collection of experimental results, consisting of gene expression, protein expression and phenotype growth sensitivity profiles. In general, model inference agreed well with the observations, confirming the effectiveness of our strategy. In several important cases, however, significant modeling discrepancies indicated gaps in the current biological understanding of the system. Using our learning algorithm we generated novel regulation hypotheses that bridge some of these gaps. We also showed that our method attains improved accuracy in comparison to extant network learning methods.

E3. CAMP: a computational system for Comparative Analysis of Metabolic Pathways

Chun-Yu Chen[1], Chuan-Hsiung Chang[2]

Keywords: pathway comparison, metabolism, algorithm

1 Introduction.

We present a systematic method for comparing the metabolic pathways based on KEGG (Kyoto Encyclopedia of Genes and Genomes) [1] reference pathway data. Our comparison of over 100 metabolic pathways simultaneously can help researchers to figure out what happened in the course of the evolution and adaptation in each well-defined metabolic pathway. The comparison results can be used to improve the genome annotation results by either identifying missing genes, finding alternative routes or reducing annotation errors. CAMP (Comparative Analysis of Metabolic Pathways) will be freely available for academic or nonprofit use at http://gel.ym.edu.tw/camp/.

2 Materials and Methods.

The KEGG provides metabolic pathway information for each completely sequenced organism in xml file format [1]. Each file contains organism-specific metabolic pathway information. We have used these xml files as our input source, and the algorithm of our computational system outputs the percentage of similarity and/or difference between any two metabolic pathways. We have stored the pathway information as a graph $G_X = (V, E_X)$ in the adjacency list [2] for organism X. V is the common set of vertices representing the compounds involved in reference pathway G. E_A is a set of edges representing all the possible metabolic reactions in organism A referenced to pathway G. So G_A and G_B represent the information of pathway G in organism A and B, respectively. The value of denominator here is the number of relations belongs to G_B, and, the value of similarity is the number of relations shared in both G_A and G_B. The percentage score (similarity over denominator) is returned at the end. Therefore, the critical point of this comparison is to figure out how to define the relation. Our current work indicates this relation can be appropriately determined by three parameters: length of reaction paths, whether to include those relations not existed in both organisms for similarity comparison, and type of comparisons either functional or topological. The last one is illustrated in Figure 1.

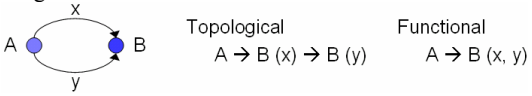


Figure 1: The difference between topological and functional comparison in the adjacency list. Where A and B represent the compounds (vertices), x and y represent the reactions (edges).

3 Results and Conclusion.

[1] Bioinformatics Program, Institute of Health Informatics and Decision Making, National Yang-Ming University, Taipei, Taiwan, 11221. E-mail: cyc@gel.ym.edu.tw

[2] Institute of Bioinformatics, National Yang-Ming University, Taipei, Taiwan, 11221. E-mail: cchang@ym.edu.tw

The metabolic pathway comparison has been implemented by Java with well-organized source code and comments, as well as JavaDoc API. A web-based query page and a Java applet are provided for displaying the comparison results. In the example shown below, we have compared the carbohydrate-related metabolic pathways of different Vibrio species. Our comparison results are consistent with the phylogeny analysis results currently available, i.e. the *Vibrio vulnificus* (VV) YJ016 is much similar to *Vibrio vulnificus* CMCP6, than with *Vibrio parahaemolyticus* (VP) and *Vibrio cholerae* (VC). This result is shown in Figure 2.

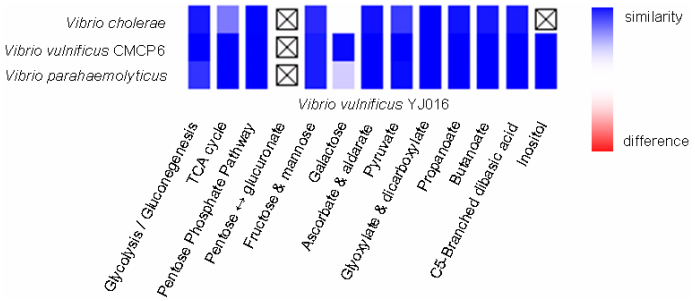


Figure 2: Comparisons of carbohydrate-related metabolic pathways in Vibrio species.

To know exactly what the differences are and why these Vibrio species are so different in the galactose metabolism pathway, the system will response with three pathway views. The first two can display those enzymes only specifically present in VC when referenced to those in VV YJ016, and vice versa. The last view shows the enzymes that are shared among these organisms. All the metabolic pathways of the target organism can also be compared with those of many other organisms as shown in the example of Figure 3. The output results can not only be presented along with the taxonomy information of these organisms, but also be sorted or clustered on both axes. Therefore, our CAMP system can be used to thoroughly compare metabolic pathways in many different ways with annotated genomic information.

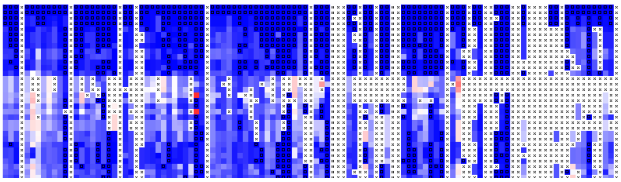


Figure 3: Comparison of γ Proteobacteria against *E. coli* K-12 MG1655 in 113 metabolic pathways.

4 Acknowledgement.

This work is supported by a grant (NSC 92-3112-B-010-020) from the National Science Council of the R.O.C.

5 References.

[1] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. 2004. The KEGG resource for deciphering the genome. Nucleic Acids Research, 2004, Vol. 32, Database issue D277-D280

[2] Hopcroft, J. E. and Tarjan, R. E. 1973. "Algorithm 447: Efficient Algorithms for Graph Manipulation," *Communications of the ACM*, 16, 372-378.

E4. Converting KEGG pathway database to SBML

Akira Funahashi,[*1,2] **Akiya Jouraku,**[*2] **Hiroaki Kitano** [1,2,3,4]

Keywords: pathway analysis, pathway database, systems biology, KEGG, SBML

1 Introduction

Systems biology is characterized by synergistic integration of theory, computational modeling, and experiment [1]. Though software infrastructure is one of the most critical component of systems biology research, there is no common infrastructure or standard to enable integration of computational resources. To solve this problem, the Systems Biology Markup Language (SBML) [2] was developed. SBML is an open, XML-based format for representing biochemical reaction networks. A number of simulation and analysis packages already support SBML Level 1/Level 2 or are in the process of being extended to support it[3].

Also, an identification of gene-regulatory logic and biochemical networks is a major challenge of systems biology. Several attempts are under way to create a large-scale, comprehensive database on gene-regulatory and biochemical networks. Converting such databases to SBML is quite important by following reasons.

1. Many applications which support SBML can directly use such huge database.
2. The feedback from developing the converter will suggest the new additional feature of the next level of SBML.

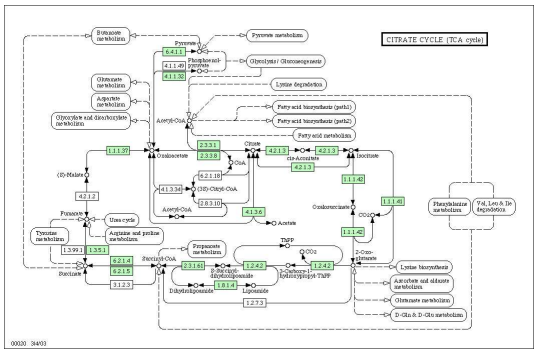


Figure 1: Citrate cycle (TCA cycle) of Caenorhabditis elegans.

2 KEGG to SBML Conversion

For the first step of converting such biochemical database, we've decided to convert KEGG (Kyoto Encyclopedia of Genes and Genomes) [4],[5] pathway database. KEGG contains more than 13,000 metabolic pathways for 150 species. Figure 1 shows an example pathway which is available from KEGG pathway database. As shown in Figure 1, each metabolic pathway of KEGG consist of compounds and enzymes (i.e. each enzyme acts as a regulator). By using these information from KEGG, we have implemented a converter(kegg2sbml),

[1] ERATO-SORST Kitano Symbiotic Systems Project, JST, M-31 6A 6-31-15 Jingumae Shibuya-ku, Tokyo 150-0001, Japan. E-mail: `funa@symbio.jst.go.jp`

[2] Graduate School of Science and Technology, Keio University, Japan. E-mail: `jouraku@am.ics.keio.ac.jp`

[3] The Systems Biology Institute,Japan.

[4] Sony Computer Science Laboratories, Inc. Takanawa Muse Bldg. 3-14-13, Higashigotanda Shinagawa-ku, Tokyo 141-0022, Japan.

* These authors contributed equally to this work.

which automatically converts KEGG pathway database files into SBML Level-1 and Level-2. Kegg2sbml requires following database files for conversion [5].

1. KEGG Pathway database files
2. KGML (KEGG Markup Language) files
3. LIGAND[6] database files

Furthermore, kegg2sbml can parse layout information from KEGG and add some layout information to SBML, which can be used in CellDesigner, a process network diagram editor developed by us [7]. Figure 2 shows a screenshot of CellDesigner, while opening the converted SBML document with layout information. Comparing Figure 2 with Figure 1, it is clear that the layout information of KEGG was successfully converted to SBML.

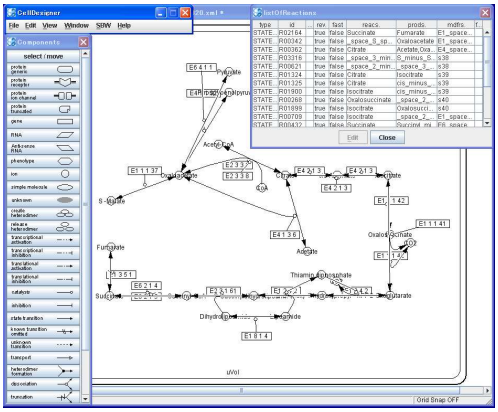


Figure 2: Screenshot of CellDesigner browsing converted SBML document.

3 Conclusion

We have converted KEGG pathway database to SBML. By implementing the converter (kegg2sbml), we have succeeded to convert 10,860 pathways into SBML Level-1 and Level-2. All converted SBML documents are freely available from `http://systems-biology.org/001/`, so that existing SBML compliant applications can directly use these pathways. Also kegg2sbml is available from `http://sbml.org/kegg2sbml.html` as an open source product.

References

[1] Kitano, H. 2002. Systems biology: a brief overview. *Science* 295:1662-1664.

[2] Hucka, M., Finney, A., Sauro, H. M., Bolouri, et.al. 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 19:524–531.

[3] The list of SBML compliant software is available from `http://sbml.org/`.

[4] Kanehisa, M. 1997. A database for post-genome analysis. *Trends Genet.* 13:375-376.

[5] Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28:27-30.

[6] Goto, S., Nishioka, T., and Kanehisa, M. 1998. LIGAND: Chemical Database for Enzyme Reactions *Bioinformatics* 14:591-599.

[7] Funahashi, A., Tanimura, N., Morohashi, M. and Kitano, H. 2003. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO* 1:159-162.

[5] These files are available from `ftp://ftp.genome.ad.jp/pub/kegg/`, but non-academic users are requested to obtain a license agreement through the licensor, Pathway Solutions Inc., for downloading the KEGG data.

E6. An Integrated Platform to Construct Transcriptional Network from Gene Expression Data

Tao-Wei Huang[1], Hwa-Sheng Chiu[1], Ming-Hong Lin[2], Han-Yu Chuang[1], Chi-Ying F. Huang[2], Cheng-Yan Kao[1]

Keywords: gene expression data, pathway, gene network, transcriptional network

1 Introduction.

In the post-genome era, systems biology is an emergent field for understanding of biological systems at whole system-level instead of single gene or protein. There are numbers of exciting and profound issues that are actively investigated such as the gene regulation and biochemical network. In this study, we focus on the transcriptional networks. The traditional clustering algorithm is applied to the gene expression data to acquire some groups of genes. The signature algorithm [2] can be used to refine and extend these groups of genes generated by clustering algorithm. When a group of genes with common *cis*-regulatory binding motif in promoter region, and we can assume these genes are co-regulated by the same *trans*-regulatory factor. Not only the binding motif information is investigated but also known pathway, protein interactions and cellular component of Gene Ontology information are referred [3]. We performed this idea and approach to find the most upstream regulators. Appling this approach to cancer research, these genes may be novel oncogenes. We integrated some biological databases and also provided some bioinformatics tools as web service for biologists to predict the transcriptional networks *in silico*. By this approach, we found some important transcriptional module and transcription factors, and we will exam it by RNAi experiment *in vitro* or *in vivo*. Besides, the more details of our result can be found on our website (http://insilico.csie.ntu.edu.tw:9999/RECOMB2004/).

2 Method.

In order to obtain more precise transcriptional networks, we integrated the some well-know biological databases locally, updated periodically and developed some bioinformatics software packages as followings:

(1) MotifFinder: This package can accept a list of genes of interest as input. The upstream promoter sequences of these genes are extracted from DBTSS database. The range of promoter sequences takes as parameter from -3000 to +1000. It will also process the retrieved result from TRANSFAC and find the common *cis*-regulatory elements and corresponding *trans*-regulatory factors as output.
(2) GOFinder: The given input is a list of genes of interest. The biological process, cellular component, and molecular function categories of Gene Ontology is shown. Besides, the biologists can search the gene ontology information with different ontology level from 1 to 5.

[1] Dept. of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan.
E-mail: {d90016, r91031, r90002, cykao}@csie.ntu.edu.tw
[2] Division of Molecular and Genomic Medicine, National Health Research Institutes, Taipei 115, Taiwan.
E-mail: {daycolin, chiying}@nhri.org.tw

(3) PathwayFinder: The pathway package can accept a list of genes of interest as input. All the BioCarta and KEGG pathways containing these genes will be reported in a descendant order according to the number of input genes in a pathway.
(4) ProteinInteractionFinder: The interaction tool can take a list of genes of interest as input. The protein-protein interactions information from DIP and the annotations of these interacting proteins retrieved from SWISS-PROT will be shown.

We use the bottom-up approach to construct the transcriptional networks from small transcriptional modules. The rectangle box stands for a transcriptional module. The gene or protein represented as white ellipse, and the black ellipse is the protein with interactions (Figure 1). The detail processes described as following:

Step1. As a first step for further analysis, we applied the k-means clustering algorithm to gene expression data. The genes in the same cluster *c* are possibly and potentially co-regulated.
Step2. After generating the k clusters, we applied the signature algorithm to each cluster *c* to obtain a new cluster *c*' as 'transcriptional module' initially. The signature algorithm can be used to extend and refine partial knowledge about a pathway module. The genes in the same module *c*', therefore, are more potentially co-regulated.
Step3. Some useful bioinformatics tools are provided to analyze these modules and more annotations are extracted from GeneCards, GO, SWISS-PROT, NCBI for biologists. Genomics information such as binding motif and Gene Ontology are provided. Proteomics information such as pathway and protein-protein interactions is provided. Combining the biological information, the biologists can infer the regulators and construct transcriptional networks more precisely.
Step4. Take the regulating genes or proteins as newer transcriptional module, and apply Step3 to these modules recursively.

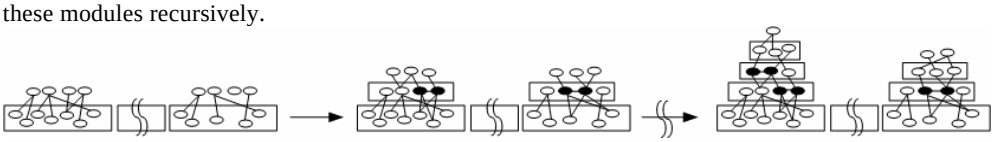


Figure 1: Construct transcriptional network from transcriptional module iteratively.

3 Results.

HCC microarray dataset, obtained from Stanford Microarray Database, contains 1648 differentially expressed genes in HCC vs. nontumor liver samples as analyzed by Chen *et al*. [1]. By our approach, we found some important transcription modules and some regulators can regulate these modules. From further literature review, we also found the regulators involve in hepatocyte regeneration upon partial hepatectomy. Besides, we will exam it by RNAi experiment *in vitro* or *in vivo*. This approach may provide novel targets involved in the carcinogenesis of HCC.

References

[1] Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, Ji J, Dudoit S, Ng IO, Van De Rijn M, Botstein D, Brown PO. 2002. Gene expression patterns in human liver cancers., *Molecular Biology of the Cell*, Jun;13(6): pp. 1929-1939

[2] Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. 2002. Revealing modular organization in the yeast transcriptional network., *Nature Genetics*,Aug;31(4): pp. 370-377

[3] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data., *Nature Genetics*, Jun;34(2) pp. 166-176

E7. Path Finding and Topology Correction in Biological Networks

Ryan Kelley[1], Astrid Haugen[2], Bennet Van Houten[2], Trey Ideker[1]

Keywords: biological networks, systems biology

1 Abstract

Systems biology requires the integration of diverse sets of biological data, such as microarray expression data and protein interaction studies. Viewing these data in a network context is one way to achieve this goal. In order to analyze data integrated in this manner, methods must be determined for discovering significant relations in these networks. Previous work focused on identifying significant subgraphs.[2] However, it is also useful to identify specific types of subgraphs, such as paths. Biologically, these paths may represent a causal chain of events. In this work, we present a method for identifying such paths in a network. This method builds on the previous work of Kelley *et al.*[3] Briefly, this technique uses linear programming to identify the highest scoring path of length L ending at each node, given the highest scoring path of length L-1. Extensions to this work include a scoring correction for the topology of the network. This is meant to correct the problem whereby nodes with many neighbors are more likely to belong to a high scoring path.

Fitness data as described by Giaever *et al.*[1] was obtained for the response of *Saccharomyces cerevisiae* to arsenic, which was then incorporated into a species-specific metabolic map. Our method was sucessfully used to identify signficant paths in this metabolic map. Specifically, 6 different paths containing 10 unique metabolic reactions were discovered. These reactions are involved in the synthesis/conversion of amino acids relevant to the arsenic response. These results suggest that this method may be useful in analyzing other data sources in conjuction with different types of biological networks.

References

[1] Giaever *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome *Nature* 418, 387-391.

[2] Ideker T, Ozier O, Schwikowski B, Siegel A (2002) Discovering regulatory and signalling circuits in molecular interaction networks *Bioinformatics* 18: S233-S240.

[3] Kelley, B.P. *et al.* Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci USA* 100, 11394-9.

[1]Department of Bioinformatics, University of California San Diego, 9500 Gilman Dr. 0412, La Jolla, CA 92093 rmkelley@ucsd.edu

[2]Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709

E8. Predicting cis-Regulatory Elements and Regulatory Networks

Stefan Kirov[1], Bing Zhang[2], Denise Schmoyer[3], Oakley Crawford[4], Jay Snoddy[5]

Keywords: regulatory networks, cis-regulatory elements, high throughput analysis

1 Introduction.

Presently, vast amount of sequence and expression data are available for several eukaryotic genomes. This provides us with the opportunity to actually predict both cis-regulatory elements (CREs) and putative regulatory networks (RN), whose expression is driven by a set of CREs both predicted or experimentally defined. And it has been shown previously that transcription regulation is often driven by multiple CREs, sometimes as much as 30[1, 2]

The abundance of expression and sequence information used separately is not sufficient because (a) the process of expression regulation in eukaryotes is remarkably complex (b) CREs discovery represents a serious computational problem due to their size and minimal conservation. Our approach to the problem is to combine evolutionary conserved upstream regions from sets of possibly co-regulated genes into a dataset we use to predict single or composite CREs. This strategy is based on the assumption that most eukaryotic CREs occur inside evolutionarily conserved regions[3, 4]. We extend this analysis by searching a database of upstream regions with the set of CREs predicted in the first step. By doing this, we prove the CREs under investigation are over represented in the initial dataset and second we elucidate the proposed regulatory network (the initial dataset) by extending and/or refining the list of genes that indeed do have matches to a part of or to the entire CRE set.

We created a pipeline, called Batch Sequence Analysis (BSA), which can accomplish the analysis process described above, without manual intervention. We present here the structure and logic of BSA and some preliminary results generated by BSA.

1 Materials and Methods.

We assembled a sequence database of upstream regions for 36389 genes (19093 human, 12988 mouse, and 3980 rat genes). We used for this purpose the ENSEMBL database (version 19, www.ensembl.org). We defined our clusters of orthologous groups by querying the GeneKeyDB database (genereg.ornl.gov/gkdb) and the ENSEMBL compara database. The conserved regions are established through additive pairwise alignment with the bl2seq, after which all non-conserved sequences are masked.

[1] Graduate School for Genome Science and Technology, Oak Ridge National Laboratory-University of Tennessee, PO Box 2008, MS6164, Oak Ridge TN 37831-6164, USA. E-mail: janeway@skirov@utk.edu

[2] Graduate School for Genome Science and Technology, Oak Ridge National Laboratory-University of Tennessee, PO Box 2008, MS6164, Oak Ridge TN 37831-6164, USA E-mail: bzhang@utk.edu

[3] Life Sciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831-6124, USA E-mail: dcs@ornl.gov

[4] Graduate School for Genome Science and Technology, Oak Ridge National Laboratory-University of Tennessee, PO Box 2008, MS6164, Oak Ridge TN 37831-6164, USA E-mail: ocs@ornl.gov

[5] Graduate School for Genome Science and Technology, Oak Ridge National Laboratory-University of Tennessee, PO Box 2008, MS6164, Oak Ridge TN 37831-6164, USA E-mail: v8v@ornl.gov

The pipeline uses currently the MEME/MAST (meme.sdsc.edu/meme/website/intro.html) system for CRE/RN prediction. Parsers for these programs as well for the TRANSFAC format are included in the current Bioperl version (bioperl.org). The images representing the motif matches are created using the BioPerl GD library. We use the weblogo to create images of each CRE consensus (weblogo.berkeley.edu). All results are stored in a relational ORACLE database called PSITE (genereg.ornl.gov/gkdb). The pipeline is written in Perl and can be installed under LINUX/UNIX.

2 Results and Discussion

The general structure of the BSA pipeline is described in Figure1. The data generated by the pipeline is stored in the PSITE database after the CRE and RN steps. The input data set format is a simple list of several popular database identifiers (LocusLink, ENSEMBL, REFSEQ, GO, gene symbol, etc.).

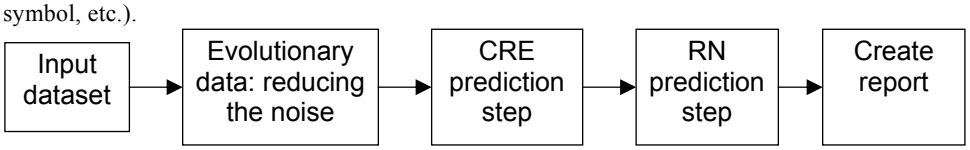


Figure 1: BSA pipeline structure.

During the CRE step, new over-represented motifs are predicted and can be compared to a previously identified set of Positions Scoring Matrices (PSMs), TRANSFAC for example. Such comparison can be done at the PSM sequence match level by first looking for possible matches within the defined conserved regions to known motifs and then mining the PSITE database for overlapping matches. An example is shown in Table 1.

Query motif ID	Gene	Start	Matching motif ID	Gene	Start
Transfac ID X	6607	-753	ps4_000000000496	6607	-750
ps4_000000000303	6789	-139	ps4_000000000491	6789	-145

Table 1: Overlapping sequence matches between motif, predicted during the analysis and previously predicted or experimentally discovered CRE.

The BSA report consists of all files created during the run. This includes the sequences that were retrieved, log files, graphic representation of the CRE and CRE matches. This report can aid experimentalists to formulate new hypothesis, concerning the transcription regulation of the genes that were analyzed and possible new interaction as new RN is predicted.

We used different data sources to test the pipeline: GO families, microarray expression data, QTLs (Quantitative Trait Loci), etc. We will present here some of our findings.

Though we have used BSA pipeline exclusively to search for CRE within the promoter regions, it can also be used to look for sequence motifs inside other sequence regions such as UTRs, introns, etc.

3 References and bibliography.

[1] Fickett, J.W. and W.W. Wasserman, *Discovery and modeling of transcriptional regulatory regions.* Curr Opin Biotechnol, 2000. **11**(1): p. 19-24.

[2] Yuh, C.H., H. Bolouri, and E.H. Davidson, *Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene.* Science, 1998. **279**(5358): p. 1896-902.

[3] Wasserman, W.W., et al., *Human-mouse genome comparisons to locate regulatory sites.* Nat Genet, 2000. **26**(2): p. 225-8.

[4] Dermitzakis, E.T., C.M. Bergman, and A.G. Clark, *Tracing the Evolutionary History of Drosophila Regulatory Regions with Models that Identify Transcription Factor Binding Sites.* Mol Biol Evol, 2003. **20**(5): p. 703-14.

E9. Towards Automated Explanation of Gene-Gene Relationships

Wacław Kuśnierczyk,[1] Astrid Lægreid,[2] Agnar Aamodt[1]

Keywords: microarrays, gene-gene relationships, knowledge-intensive problem solving, iterative search, public databases and tools.

1 Introduction.

During the recent decades research in molecular biology experienced several paradigm shifts that changed the researchers' approach to solving particular problems in the field. The invention of microarray technology in the last decade of the previous millennium can certainly be seen as one of those paradigm shifts [3]. However, although there appear first reports showing efforts to combine various resources of genomic data instead of investigating just one source (e.g., [2, 4]), the understanding and interpretation of the results—the key issue in any attempt to a discovery—is still entirely left to the human.

Microarray data represent a reasonable source of new hypotheses on gene function and between-gene relationships. In order to fully understand and explain these hypotheses, biological background information from many resources has to be explored and combined.

We propose a novel method intended to aid a researcher in understanding hypothetical relationships between genes, e.g., genes not previously known to be related. In order to justify a tentative link between genes, the sequences, promoter regions, protein structure, function and other properties may have to be investigated for these and possibly other related genes. Although a manual search for information that would link two genes is theoretically possible, in practice it may be a very tedious task. Our approach is an attempt to design and implement a high-level wrapper for existing databases and tools, providing an automated process of forming relevant human-readable explanations. The proposed solution draws from the achievements of research in artificial intelligence—knowledge representation and knowledge-intensive reasoning, non-deductive inference mechanisms, and machine-learning [1].

2 Methods.

The proposed system is an intelligent interface between the user on one side, and remote databases and publicly available tools on the other side, enhanced by background knowledge in molecular biology. Its modular architecture is illustrated in Fig 1. A typical question that can arise from a microarray experiment and may be asked to the system is of the form *How are genes g_1 and g_2 related?* or *What might be the causal relationship between genes g_1 and g_2?* The exact syntax of the query depends on the actual implementation of the query interface module (QI).

The query, translated into the internal representation language, is interpreted by the core reasoner (CR), which utilizes general domain knowledge (GDK) to construct an explanation chain that links the two investgated genes. The GDK is modelled as a multi-relational semantic network, a kind of ontology, where each concept and relation are represented as a

[1]Department of Information and Computer Science, Norwegian University of Science and Technology, Sem Sælandsv. 7, 7491 Trondheim, Norway. E-mail: {waku,agnar}@idi.ntnu.no

[2]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Olav Kyrresg. 3, 7489 Trondheim, Norway. E-mail: astrid.lagreid@medisin.ntnu.no

distinct object. The CR contains inheritance and propagation methods for reasoning over a combined set of relations (e.g. *causes*, *has-function*, *has-structure*, *has-subclass*, *has-part*, *triggers*, *inhibits*, etc.), assigning a specific 'explanatory strength' to each relation. Data of different types, related to the queried genes, are retrieved from databases (DB) and matched with the help of available tools (T). The connection between the system and various public resources goes through dedicated interfaces responsible for contacting a resource suitable for a particular task and in a way specified by the resource's query interface. After the concepts have been instatiated with specific data coming from the query and from databases, the system attempts to construct an explanation by searching for one or more paths in the semantic network that would connect the two genes. The retrieval of information needed to instatiate the concepts is repeated iteratively until a pathway is found, or specific search-limiting criteria are met.

An explanation is output through the explanation module (EI), and the user may or may not accept it, thus giving feedback for the search process. The explanation may be refined at a later time, and the result may be retained in the case-base (CB) for further reference and to boost a search similar to a previously completed one. An explanation is of the form *Gene g_1 regulates gene g_x which in turn produces a protein that may interfere with the action of gene g_2*, though typically it would be much more complex.

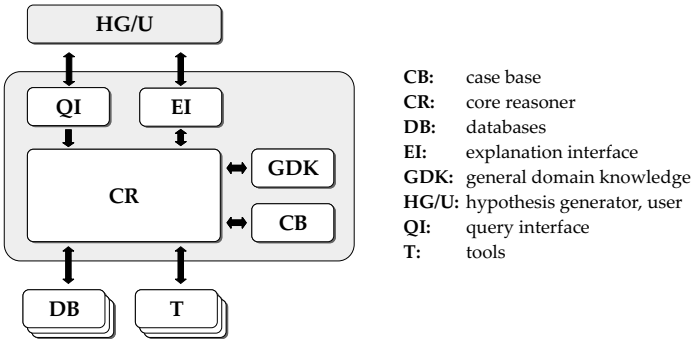


Figure 1: The architecture of an automated system for explanation of gene-gene relationships.

Note on the implementation The system is currently in the design phase. The specific modules will be implemented concurrently and a functional version of the system is planned to be released by the end of year 2005.

References

[1] Aamodt, A. 1991. *A knowledge intensive, integrated approach to problem solving and sustained learning.* PhD thesis, Norwegian University of Science and Technology.

[2] Bar-Joseph, Z. et al. 2003. Computational discovery of gene modules and regulatory networks. *Nature Biotechnology* 21(11):1337–1342.

[3] Brown, P.O. and Botstein, D. 1999. Exploring the new world of the genome with DNA microarrays. *Nature Genetics* 21:33–37.

[4] Köhler, J. and Schulze-Kremer, S. 2002. The Semantic Metadatabase (SEMEDA): Ontology based integration of federated molecular biological data sources. *In Silico Biology* 2.

E10. Linkage by context: Discovering functional linkages between proteins from their known interactions

Insuk Lee[1,*] and Edward Marcotte [1,2]

Keywords: protein functional network, network neighborhood, linkage by context

1 Introduction.

Proteins can be thought of as being organized into networks, with proteins connected by virtue of physical or functional interactions – such network models are becoming increasingly important in system biology and medicine. To date, many useful experimental and computational methods have been developed for finding functional linkages among proteins. Unfortunately, none of these are comprehensive, and all show limited accuracy and proteome coverage. Therefore we are, currently, able to detect only a fraction of the total linkages between proteins, with many left undetected.

Here, we propose a method to find new functional linkages between proteins based on the proteins' already known interaction partners in an existing network. The fundamental idea is that two proteins that are linked to similar groups of proteins (i.e., that share a network neighborhood) are more likely to be linked themselves. This general idea has been already used to verify the quality of protein interactions [1] or to suggest new linkages in a variety of biological networks [2].

In this analysis, we discovered new functional linkages between yeast proteins by analyzing the existing yeast functional network. In this preexisting network, pairs of yeast proteins were linked if they were observed to operate in the same cellular pathway/system via a variety of large-scale analyses [3-10]. Each link carries a real-valued measure of confidence. The method we describe here allowed us to discover several thousand additional linkages in yeast, and to improve the overall quality of the yeast protein network.

2 Method and Results.

Starting from a reasonably accurate yeast protein network (the result of an integration of an variety of functional genomics data [11]), we calculated a matrix of protein functional linkages. In this matrix, we have real-valued confidence scores instead of simple binary values for the linkages, thus, each row (or column) of the matrix represents the "context vector" of a protein, or the list of linkages it participates in, and their relative strength. Missing linkages were indicated by zero entries in the matrix. Using this matrix, we calculated the Pearson correlation coefficient between the genes' context vectors, including in the calculation only those entries in which at least one of the two genes had a non-zero value. The resulting correlation coefficients indicate the degree of similarity between the overall network neighborhoods of each pair of genes, regardless of whether the genes were previously linked. Protein pairs were sorted by correlation coefficient, and we calculated the likelihood that the pairs of proteins were in the same KEGG pathway [12] as a function of correlation coefficient.

We found that the quality of linkage by context depends strongly on the preexisting network. We compared linkages by context from three different groups of preexisting linkages, each with the same size but different quality, made by replacing high quality linkages with low ones (Fig A). Assessment of the quality of linkages derived only from network context (Fig B) shows that given a

[1] Center for Computational Biology and Bioinformatics, Institute for Cellular and Molecular Biology, and
[2] Department of Chemistry and Biochemistry 1 University Station A4800, Austin Texas 78712-1064, USA
* email: lee-micro@mail.utexas.edu

reliable starting network, this method produces additional protein functional linkages of reasonable accuracy. Thus, for the many ongoing attempts to integrate functional genomics data and protein interactions, this approach represents a simple but powerful to improve the final networks.

3 Figures.

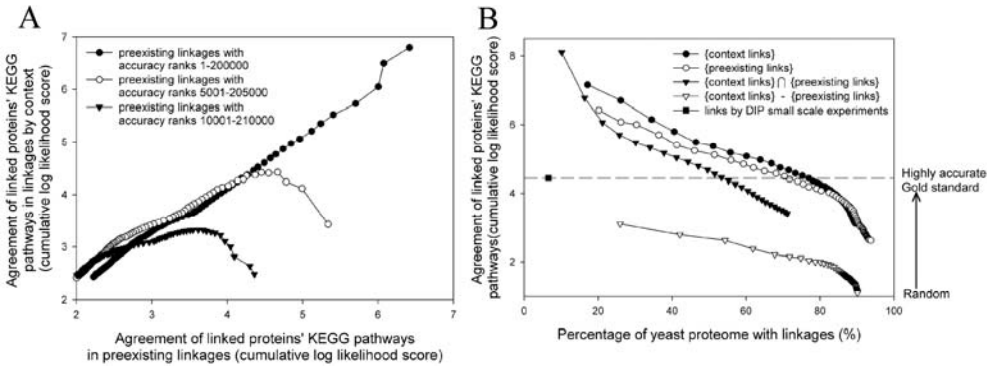


Figure (A). Dependence of the linkages derived from network context on the quality of the initial linkages. As the quality of initial network decreases, the quality of the context-derived linkages decreases correspondingly. (B) Assessment of linkages derived only from network context. The linkages derived from network context show better performance than preexisting linkages, linkages from both preexisting network and its context, or linkages from only network context. Linkages derived only from network context still show very significant likelihood of being correct (i.e. much better than random chance of proteins being in the same pathway).

4. References

[1] Goldberg, D.S. & Roth, F.P. Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A* **100**, 4372-4376 (2003).

[2] Schlitt, T. et al. From gene networks to gene function. *Genome Res* **13**, 2568-2576 (2003).

[3] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83-86 (1999).

[4] Uetz, P. et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-627 (2000).

[5] Tong, A.H. et al. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364-2368 (2001).

[6] Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-4574 (2001).

[7] Marcotte, E.M., Xenarios, I. & Eisenberg, D. Mining literature for protein-protein interactions. *Bioinformatics* **17**, 359-363 (2001).

[8]. Gavin, A.C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141-147 (2002).

[9] Ho, Y. et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180-183 (2002).

[10] Gollub, J. et al. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res* **31**, 94-96 (2003).

[11] Insuk Lee & Marcotte, E. Functional network of yeast is accurate, extensive, and modular. *Submitted* (2004).

[12] Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res* **30**, 42-46 (2002).

E11. Learning Context-sensitive Boolean Network from Steady-state Observations and Its Analysis

Huai Li[1], Jon Whitmore[1], Ed Suh[1], Mike Bittner[2], and Seungchan Kim[2]

Keywords: gene regulatory network, context-sensitive Boolean network, Markov chain simulation

1 Introduction.

Boolean network model [1] may provide useful insights for network dynamics at the coarse level. Recently Boolean network has been extended to cope with certain randomness inherent in biological system as Probabilistic Boolean network [4]. While Probabilistic Boolean network is a step forward toward a better mathematical model with capability to abstract uncertainty in biological system, it fails to describe context specific determinism of regulatory system. Context can be defined as a certain condition under which a limited number of genes are tightly regulated by each other via specific cellular mechanisms to perform a specific task [2]. This specific task can be a different developmental stage, or tissue specific function, resulting in a specific cell-type. The change of this context will result in the change in the set of genes that are highly interactive, and probably their connectivity and relationships. Different biological contexts can also correlate with different diseases or might be a reason why a certain group of patients respond to a therapy while others do not. We started to study this problem and have been developing a context-sensitive Boolean network (cBN) model that will abstract the following hypothesis; regulatory mechanism itself in cellular system is static and hard-coded in its genetic code (genomic information), but its activation and inactivation (transcriptomic information) is context sensitive.

While high-throughput gene expression profiling provide vast amount of data for cellular system, most of those measurements come from the stead state observation of the system. Suppose we infer rules from steady-state observations, would the network driven by these rules mimic behavior of biological process? For answering this question, one plausible way we can make sense out of a rule-making procedure is to see what it does in a case we understand the ground truth through the simulation of a small synthetic network driven by some artificial rules. In addition, how is the sensitivity and stability of the network to the methodology of rule formation? Although the stability of a large random Boolean network was well studied both analytically and mathematically [1,3], there has not been extensive study for their structures and the inference of model parameters based on steady-state observations, and their relevance to approximating certain biological systems behavior.

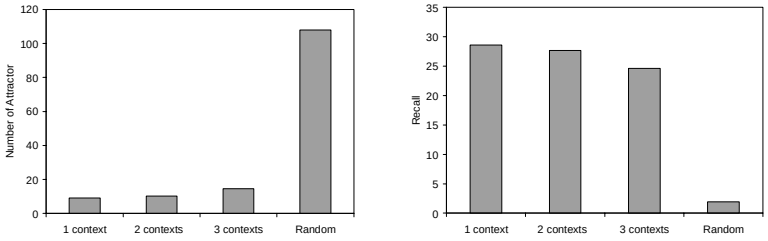
2 Results.

Fig. 1 shows the effect of varying the extent of data consistency on the dynamics of the network and the recall of rules based on simulated data. The recall is defined as the ratio of the number of relevant rules retrieved from the inferred rules to the total number of relevant rules originally constructed.

[1] Computational Biology, Translational Genomics Research Institute, 400 N. Fifth Street, Suite 1600, Phoenix, AZ 85004, USA. Email: hli@tgen.org

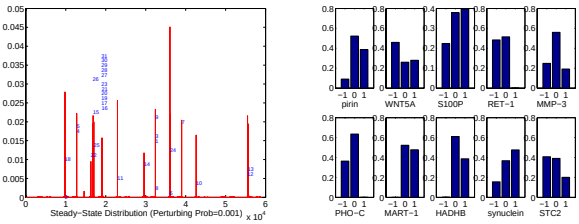
[2] Molecular Diagnostics and Target Validation Division, Translational Genomics Research Institute, 400 N. Fifth Street, Suite 1600, Phoenix, AZ 85004, USA. Email: dolchan@tgen.org

When applying cBN to microarray data, the simulation results showed that the total 31 melanoma sample states occupied 43% of the portion in the steady state distribution with perturbation probability $p = 0.001$, as shown in Fig. 2.



(a) (b)

Figure 1: Effect of varying the extent of data consistency on the recall of rules and the dynamics of the network.



(a) (b)

Figure 2: The estimated distribution after long run based on melanoma data.

3 Conclusion.

In this study, we tried to address if the rules inferred from the steady state observations and the network dynamics driven by those rules can provide us useful information by analyzing the sensitivity and stability of the network to the methodology of rule formation. We used cBN model constructed by a set of artificial rules/functions and inferred rules from steady state observations of this artificial network. By comparing various statistics estimated from the network reconstructed by the inferred rules against those estimated from the network originally constructed by the artificial rule set, even though in this very limited context, we conclude that the inference of rules from steady state observations and its analysis might be quite informative to understanding of cellular system. We also conclude that the more consistent the data is the more stable the network is and the more useful information the network can provide. When applied to microarray data, we observed the rules inferred in fact effectively drive cell states into cancer states.

References

[1] Kauffman, S.A. 1990. Requirements for Evolvability in Complex Systems: Orderly Dynamics and Frozen Components. *Physica D* **42**:135-152.

[2] Kim, S., et al. 2002. Can Markov chain mimic biological regulation? *J. Biol Systems* **10**(4):337-358.

[3] Lynch, J.F. 1993. A Criterion for Stability in Random Boolean Cellular Automat. *Ulam Quarterly* **2**:32-44.

[4] Shmulevich, I., et al. 2002. Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* **18**(2):261-74.

E12. Stochastic regulation of NF-κB pathway

T. Lipniacki[1], P. Paszek[2], A. Brasier[3], B. Luxon[3], M. Kimmel[2]

Keywords: Stochastic regulation of transcription, molecular pathways, NF-kappaB

1 Model formulation.

Typically, in an animal cell there are tens or hundreds mRNA molecules of a given species and tens of thousand of corresponding protein molecules. Therefore processes such as mRNA translation, formation and degradation of protein complexes, and catalytic and spontaneous degradation, which involve a large number of molecules, may be modeled by ordinary differential equations (ODEs). In contrast, in a single cell, the regulation mRNA transcription may be discrete, governed by stochastic events of binding and dissociation of transcription factors. The stochasticity in regulation of transcription leads to large variability among cells. Since a given cell reacts to its own mRNA and protein levels, and not to the average levels in the population, the information about this variability is very important. In this work we apply our model on the NF-κB regulatory module [1], Fig. 1, to the single cell by modeling the transcriptional part of the regulatory network using a stochastic switch. The mathematical representation of the model consists of 14 ODEs accounting for: formation of complexes and their degradation, transport between nucleus and cytoplasm, and transcription and translation, together with 4 equations accounting for binding and dissociation probabilities of NF-κB molecules to regulatory sites in A20 and IκBα promoters. The simulation time is split into small time intervals Δt. Within Δt's, the ODEs are solved using the fourth order MATLAB solver. At the end of each interval, the

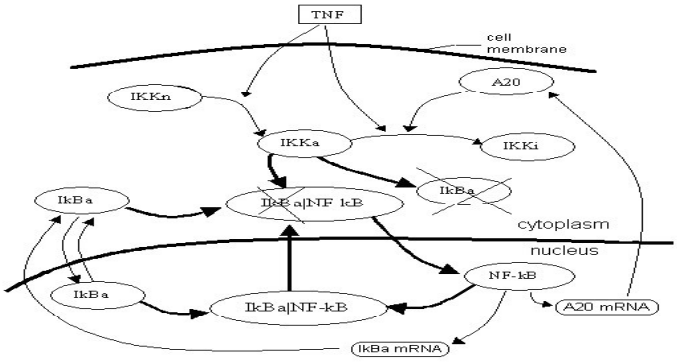


Figure 1: The model involves two-compartment kinetics of the activators IKK and NF-κB, the inhibitors A20 and IκBα, and their complexes. In resting cells, the unphosphorylated IκBα binds to NF-κB and sequesters it in an inactive form in the cytoplasm. In response to extracellular signals such as TNF, IKK is transformed from its neutral form (IKKn) into its active form (IKKa), capable of phosphorylating IκBα, leading to IκBα degradation. Degradation of IκBα releases NF-κB, which enters the nucleus and triggers transcription of the two inhibitors and numerous other genes. The newly synthesized IκBα leads NF-κB out of the nucleus and sequesters it in the cytoplasm, while A20 inhibits IKK converting IKKa into the inactive form IKKi, a form different from IKKn but also not capable of phosphorylating IκBα. Bold arrows stand for very fast kinetics.

[1] IPPT PAN, Warsaw, Poland and Department of Statistics, Rice University, E-mail: tomek@rice.edu
[2] Department of Statistics, Rice University, Houston, TX, USA
[3] University of Texas Medical Branch, Galveston, TX, USA

binding and dissociation probabilities are calculated, and the status of the two promoters, which may be ON or OFF, is evaluated and kept constant during the next time interval.

2 Results.

The stochasticity of the model implies that simulations, which mimic the behavior of single cells, performed with the same model parameters and the same initial conditions, are different. The averaged outcome resembles that of the deterministic model [1] and fits well the experimental data obtained for a population of cells, data not shown. Here we focus on variability in cell kinetics.

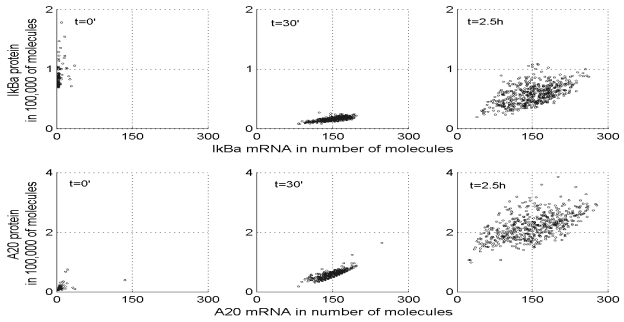


Figure 2: The scatter plots show the abundance of protein versus its mRNA at three time points, for inhibitors IκBα and A20. Prior to TNF signal there is relatively little of IκBα mRNA molecules, while the IκBα protein (which is mostly complexed with NF-κB) is abundant. Then at 30min most of the IκBα protein is degraded, but the number of mRNA molecules is large due to NF-κB induced transcription. For A20, initially there is little of both protein and mRNA, then the growing amount of transcript is followed by the growing amount of protein. The broadening of the distribution in time is caused by the desynchronization of cells due to stochasticity.

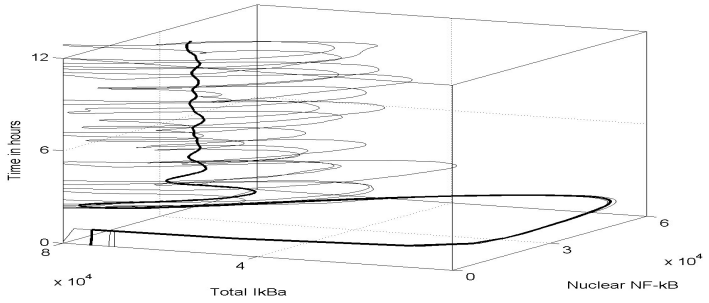


Figure 3: The single cells trajectories (thin lines) keep oscillating despite the equilibrium distribution is reached, and the average trajectory (bold line), a construct resulting from averaging over population, stabilizes.

Patterns of variability indicate that the averaging is accomplished by cancellation of phases of oscillations in individual cells. Fig. 3 shows that none of cells behaves like an average. This outcome suggests experimental testing by following individual cells, which currently is underway.

References

[1] Lipniacki, T., Paszek, P., Brasier, A., Luxon, B. and Kimmel, M. (2004) Mathematical model of NF-kappaB regulatory module. *J. Theor. Biol.,* in press.

E13. Probabilistic Representation of Gene Regulatory Networks

Linyong Mao[1] and **Haluk Resat**[1]

Keywords: gene regulatory network, probabilistic model, Monte Carlo method, gene expression, fluctuation

1 Introduction.

Recent experiments have unambiguously established that biological systems can have significant cell to cell variations in gene expression levels even in isogenic populations. In this paper, we present a new fully probabilistic approach to the modeling of gene regulatory networks that allows for fluctuations in the gene expression levels. By testing the new algorithm on the synthetic gene network library recently bioengineered by Guet et al. [1], we have demonstrated that the new algorithm is robust and very successful in explaining the experimental result.

2 Methods.

In our algorithm, expression levels of the genes in a regulatory network are represented by using multi-state functions. Changes in the expression levels of the genes are described in a probabilistic model. The gene expression levels go up or down with certain probabilities. Transition probabilities, that govern how the system may evolve, depend on the overall state of the network. The new algorithm was implemented using a Monte Carlo approach where a multitude of Markov chains are created using the computed transition probabilities. Within the allowed minimum and maximal values, expression levels of the genes can increase or decrease by one, or stay unchanged between successive steps of the Markov chain according to the employed transition probabilities. Transition probabilities were computed using the following rules: For each pair of interacting genes in the network, we define a weight W_{ik} that indicates the strength of the regulation of gene *k* by gene *i*. The W_{ik} parameters can be negative or positive indicating inhibition or activation, respectively. Total regulating strength for gene *k*, $S_k(t)$, at time t is calculated as $S_k(t) = \sum_i W_{ik} * G_i(t)$, where $G_i(t)$ is the expression level of gene *i* at time t. If $S_k(t)$ is negative, transition probabilities are given as

$$P_{k,t}(\uparrow) = P_0 \tag{1}$$

$$P_{k,t}(\downarrow) = P_0 * \left(1 + \frac{|S_k(t)|^{n_k}}{|S_k(t)|^{n_k} + C_k^{n_k}} \right) \tag{2}$$

$$P_{k,t}(-) = 1 - P_{k,t}(\uparrow) - P_{k,t}(\downarrow) \tag{3}$$

where $P_{k,t}(\downarrow)$ and $P_{k,t}(\uparrow)$ are the probabilities that the expression level of gene *k* will be lower or higher by one unit in the next step, and $P_{k,t}(-)$ is the probability that it stays unchanged. P_0 is the basal transition probability that combines various factors that may give rise to cell-to-cell variations among

[1] Computational Biosciences Group, Pacific Northwest National Laboratory, P.O. Box 999, MS K1-92, Richland WA 99352, E-mail: linyong.mao@pnl.gov and haluk.resat@pnl.gov

the members of a cell culture. Parameters C_k and n_k are constants associated with gene *k*. In the simulation, the model parameters, W_{ik}, C_k and n_k, were determined by trial and error until reasonable agreement with experimental results was obtained. When $S_k(t)$ is positive, $P_{k,f}(\downarrow)$ is assigned with the basal constant, and $P_{k,f}(\uparrow)$ is computed using the sigmoid function (equation 2).

For each simulation, a trajectory of 60 million Monte Carlo steps was run from which information about the gene expression levels and their fluctuations is collected for later analysis. Running average of gene expression levels computed in our simulations can be considered to be either the mean expression value of a certain gene across many cells of a colony or the time averaged mean expression value of a certain gene in a single cell. In these cases the observed fluctuations about the mean correspond to expression value variations among isogenic cell populations or to the dynamical fluctuations in a particular cell.

3 Results.

We used the new algorithm to simulate the synthetic genetic regulatory networks recently engineered by Guet et al. [1]. The synthetic networks were constructed by forming various combinations of four well characterized genes, one of which is the reporter *gfp* gene (Figure 1). Figure 2 compares the predictions of the probabilistic model for the synthetic networks with the experimental results. We have calculated the linear correlation between the results for the GFP expression and found a correlation coefficient of R^2=0.91 and a slope of 0.82, which shows that the predictions of the probabilistic model for the synthetic networks are in good agreement with the experimental results.

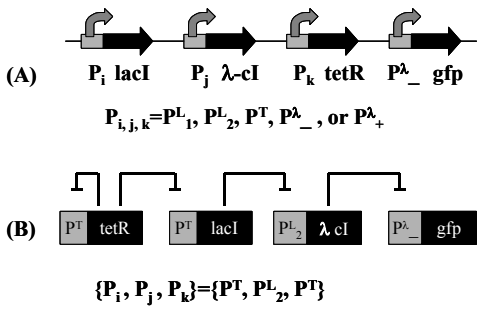


Figure 1. Structure of the studied synthetic gene regulatory networks. (A) General construct, (B) Network obtained with a particular set of promoters.

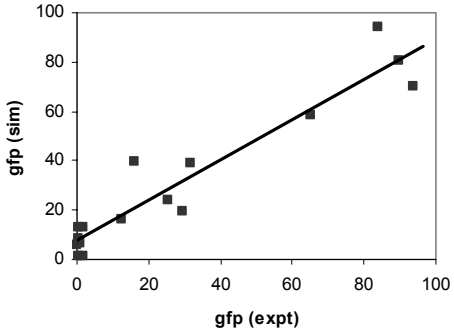


Figure 2. Comparison of simulation results for the mean GFP expression levels with experimental results for the synthetic genetic networks.

4 References.

[1] Guet, C.C., Elowitz, M.B., Hsing, W.H. and Leibler, S. 2002. Combinatorial synthesis of genetic networks. *Science* 296(5572):1466-1470.

E14. Non-exclusive Gene Groupings using SVD: A Critical Approach

Subashini Ramalingam[1], Rajagopalan Srinivasan[1]. Jonnalagadda Sudhakar[1]

Keywords: federation, singular value decomposition, non-exclusive gene groups, SVDMAN

Singular Value Decomposition (SVD) and Principal component analysis (PCA) have been used on microarray datasets to cluster genes and to identify and explain the principal components in a larger effort to reverse engineer genetic networks [2]. However, clustering methods generate mutually exclusive groups unlike cellular networks where genes (or gene products) participate in more than one reaction pathway. SVDMAN [6], addresses this problem, by generating gene groups that are non-exclusive. SVD of a matrix, X, yields components, U, S and V such that $X=U\ S\ V^T$. Columns of U give the coefficients of genes along the principal components defined by the rows of V^T. An SVDMAN group is formed by identifying all genes whose coefficients along a principal component are greater than a defined threshold value, indicating that these genes are significantly influenced by that component. Since the same gene can be highly expressed along many components, genes are found in multiple groups enabling overlaps in the network. The threshold is given by $WN^{-1/2}$, where N is the number of genes and W is a weight factor with a default value of 3.0. The application of this procedure to the publicly available Serum fibroblast dataset [3] and Yeast sporulation datasets used by Tavazoie et al. [5] using the default value revealed that the total number of genes participating in the groups and overlaps between groups were insignificant compared to the total number of genes. For example, out of 2945 genes monitored by the Yeast sporulation dataset, a total of only 124 genes entered the SVDMAN groups and no overlaps were present. The results obtained do not meet the objective, nor are they likely to be representative of real cells. In this paper, we report a modification to better identify non-exclusive gene groups.

The number of gene associations identified by SVDMAN is greatly affected by W. In order to identify a suitable network, we varied W from 5.0 to 0.1 and analyzed the trends in numbers of genes participating in the network and the extent to which these genes repeated among the groups. As W decreased from 5.0 to 0.1, the number of genes included totally and in each individual group increased as shown in Fig. 1. The fraction of genes included increased significantly initially and reached a saturation value equal to N (total number of genes) as W asymptotically approached zero. Repetition of genes among the groups also increased significantly with a decrease in W. A repetition factor for each gene entering the groups was estimated by dividing the total number of connections (number of genes in all groups including multiple occurrences of each gene) by the actual number of genes involved, excluding repeats, for every value of W (Fig. 2). The figure indicates that below a certain value of W, the number of repetitions increases rapidly and each gene appears in almost all groups. For e.g., at W~0.1, each gene in the serum is associated with ten out of all possible groups. In contrast to such high levels of gene associations, cellular networks are believed to be sparse [1, 4] and hence, we assume that each gene participates in 4-8 groups on average. Based on this observation, the range for W for the Serum dataset is [0.4 0.6]. The corresponding value for the Yeast dataset is [0.7 1.1]. This range of W also results in inclusion of 88-95% of the Serum fibroblast genes (and 95-98% of the Yeast genes) and is congruent with the expectation that the vast majority of these genes should participate, since these genes have already been filtered for substantial expression.

[1] Dept. of Chemical Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore. E-mail: eng02122@nus.edu.sg, chergs@nus.edu.sg, g0203685@nus.edu.sg

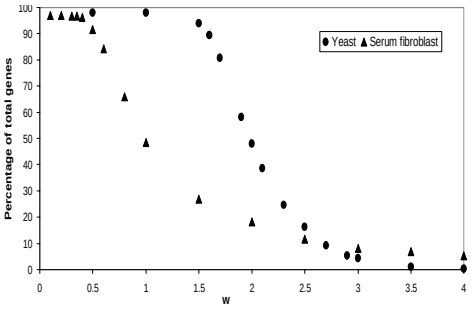


Figure 1: Percentage of genes included in groups vs. W

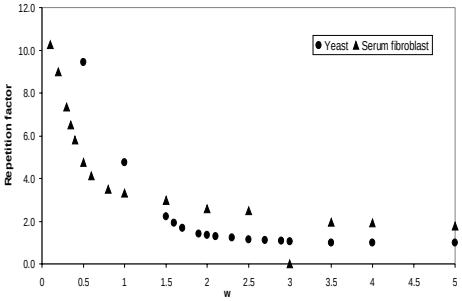


Figure 2: Repetition factor vs. W.

In order to validate the groups obtained using the optimal range of W, we compared the gene associations with clusters obtained using expression profile similarity by Iyer et. al [3] and Tavazoie et al. [5]. At the outset, since SVDMAN groups and Iyer et al. clusters are based on different criteria, associations between the two are not necessary. However, we notice some similarity among the two. Specifically, the clusters and groups have some common units (subsets) with similar expression profiles. Many of these common units participate in multiple groups and thus form the hubs of the network.

These limited observations suggest that SVDMAN can identify closely related genes which are comparable to those from clustering methods. These genes can be compared across groups to identify relations between clusters and the cohesiveness of clusters based on their ability to retain the genes together. Such analyses can be extended to obtain connections between genes and thus, the connectivity matrix of the entire genome.

References

[1] Arnone, M. I. & Davidson, E. H. (1997) *Development* (Cambridge, U.K.) **124,** 1851–1864.

[2] Holter N.S., Mitra M., Maritan A., Cieplak M., Banavar J.R., Fedoroff N.V. Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci USA 2000; 97:8409-14.

[3] Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J., Jr., Boguski, M. S., *et al.* (1999) *Science* **283,** 83–87.

[4] Jeong, H., Mason, S. P., Baraba´si, A.-L. & Oltvai, Z. N. (2001) *Nature (London)* **411,** 41–42.

[5] Tavazoie S. et al. Systematic determination of genetic network architecture. ***Nature genetics***, 22:281–285, 1999.

[6] Wall M.E., Dyck P.A., Brettin T.S. SVDMAN – singular value decomposition analysis of microarray data. Bioinformatics 2001; 17:566-68.

E15. Vector PathBlazer 1.0: A New Pathway Analysis And Visualization Tool

Feodor Tereshchenko[1], Valeriy Reshetnikov[1], Artur Karpov[1], David Pot[1]

Keywords: pathway analysis, pathway prediction, systems biology

1 Introduction.

Biological pathways may be defined as temporal or spatial sets of events leading to state changes in systems. There are several known classes of biological pathways. Most ubiquitous and important for biologists are metabolic, signal transduction, and gene activation pathways. Other examples include developmental pathways and food chain pathways.

Present pathway representations employed by different databases pose several problems for scientists attempting comprehensive computational analysis of this type of information. First, pathway data storage and retrieval is complicated by the very nature of the subject: pathway information is bi-partite: The components (what the parts are) and the topology (how those parts are connected). Many current representations focus primarily on the former, because capturing and storing topology information, is much less straightforward than storing annotation-like information. Second, the problem of pathway analysis is exacerbated by the presence of synonyms, homonyms, content-dependent synonyms, classification hierarchy, etc. Third, there is no universal standard for pathway representation – each database selects its own unique model.

There is a strong demand for software which is able to 1) integrate public data from different sources with proprietary information and 2) analyze and predict both properties of pathway components and pathway topology.

Vector ***PathBlazer*** is a program aimed at the biologist who needs to do complex analyses of different types of biological pathways.

2 Data model and database architecture.

Vector ***PathBlazer*** recognizes three types of objects: compounds, reactions, and pathways. Compounds are the actors in biological reactions. There are multiple classes of compounds: small molecules, proteins, genes, physical factors (UV light, heat), etc. Reactions are defined as spatial or temporary events leading to changes in properties of compounds. Reactions separate participating compounds into educts and products. Pathways store information about participating reactions and the connections between them.

The *Vector* ***PathBlazer*** data model is based on bi-partite directed graphs. Two sets of nodes are compounds and reactions. Directed arcs link educts to a reaction and reaction to products. Both reaction and compounds can be annotated at different levels. Information such as original database, subcellular location, tissue, organism, and disease, can be stored. Arcs carry information about the type of arc, transition probabilities, and stoichiometric constants. Pathways carry their own annotations that can differ from annotations of reactions and compounds. Annotations can be preceded by logical quantifiers such as ONLY_IN {organism|location|tissue}, KNOWN_IN

[1] Invitrogen Life Science Software, 7305 Executive Way, Frederick, MD 21704, USA
E-mail: feodor@informaxinc.com

{organism|location|tissue}, or NOT_IN {organism|location|tissue}. This simplifies storage and retrieval of species-, organism-, and location-specific information.

The population of the *Vector* ***PathBlazer*** database can be achieved by the use of parsers supplied by InforMax. Parsers for KEGG [2], DIP [6], BIND [1] are currently provided. Parsers for TransPath [4] and BioCyc [3] are being developed. Unique or proprietary data can be added to the database manually through a graphical interface or via XML files.

The data imported from different sources are integrated into a single Access database. The integration rules are based on lists of synonyms and annotations. These rules can be fine-tuned to satisfy a user's needs.

3 Pathway discovery, assembly and analysis.

The *Vector* ***PathBlazer*** database can be queried for pre-assembled pathways or novel pathways can be discovered and built using the database query engine and the pathway-building algorithm. A modified Dijkstra algorithm is used to 1) search for the shortest paths between two compounds, and 2) search for the shortest paths and path lengths equal to shortest path plus specified numbers of steps.

These searches can be performed on any reaction subset, Reaction subsets are created manually or by filtering other sets of data. Therefore, subsets can include reactions from one or more databases, making discovery of previously unknown connections across heterogeneous data sources possible. Reactions can be added to resultant pathways in a stepwise fashion. Additionally, Use of Interaction Generality [5] as a restricting parameter can reduce amount of potentially irrelevant interactions.

Expression data can be easily mapped to assembled pathways using customizable templates. The results of a particular experiment can be color coded according to user's specification.

Graphical information can be displayed in different forms using graph layout settings. Editing capabilities allow for the addition and deletion of pathway elements. Annotations can also be accessed and changed directly from the graphical pane.

References

[1] Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B.F., Pawson, T. and Hogue, C.W. 2001. BIND – The Biomolecular Interaction Network Database. *Nucleic Acids Research* 29:242-245.

[2] Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28:29-34.

[3] Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C. and Gama-Castro, S. 2002. The EcoCyc Database. *Nucleic Acids Research* 30: 56-58.

[4] Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A. and Wingender, E. 2003. TRANSPATH: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Research* 31:97-100.

[5] Saito, R., Suzuki, H. and Hayashizaki, Y. 2003. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. *Nucleic Acids Research* 30:1163-1168.

[6] Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M. and Eisenberg, D. 2000. DIP: The Database of Interacting Proteins. *Nucleic Acids Research* 28:289-291.

E16. Simulation mammalian molecular circadian oscillators by dynamic gene network

Yanhong Tong[1], Hava Sieglemann[2].

Keywords: circadian rhythms, gene network

1 Introduction.

Internal biological rhythms that are entrainable to the 24-hr light-dark cycle are driven by endogenous oscillators called circadian clocks. At the molecular level, circadian oscillators are controlled by autoregulatory feedback loops. The transcriptional-translational feedback loops involve transcriptional activation/inhibition and translocation to the nucleus. Several of the genes and proteins involved in the feedback loops have been well studied in different organisms, such as the mouse and drosophila, although questions remain to be answered.

Although approximately 10 genes/proteins (Bmal, Clock, Per1-3, Cry1-2, Rev-erbα in mouse) are involved in the regulation of core clock to generate circadian rhythms, hundreds of genes cycle in different organs such as the suprachiasmatic nuclei and liver. The relationships among the core clock genes/proteins and the cycling genes/proteins in different organs are not yet well understood because of the system's complexity.

Several computational approaches to model biological clocks have been proposed [1]. However, there has not been a model which is basic enough to be of help for molecular biologists in the discovery of new genes or proteins related to circadian rhythms.

This study implements a gene network to model the circadian oscillator at the molecular level. The model is simple and can be easily updated in accordance to new experimental data. While the basic data used was collected from mice, not many modifications would be required to apply it to the human biological clock. This network should prove useful in the discovery of new genes/proteins related to circadian rhythms as well as in the analysis of drug's effects on the biological clock and in understanding of the most effective timing for administering various medications.

2 Methods and Results.

The molecular data for circadian clocks, including the data for gene mutant/deletion mice, are collected from hundreds of studies based on the list of references at http://stke.sciencemag.org/cgi/cm/stkecm;CMP_13296[2].

Our gene network is a directed graph, consisting of three kinds of nodes: a gene node (circle), a protein node (rectangle), or a protein complex (diamond). An edge A → B connects two nodes A and B, and its direction represents the message transmission: A activates/inhibits B via passing

[1] Computer Science Department, University of Massachusetts at Amherst, Computer. E-mail:ytong@cs.umass.edu

[2] Computer Science Department, University of Massachusetts at Amherst, Computer. E-mail:hava@cs.umass.edu

messages of start /stop activation or start/stop inhibition at certain points in time. Each node is associated with an expression function which describes how it is affected by its input; this is based on biological data or published computational models. A node can receive message(s) from its direct upstream node(s), when it reaches a certain threshold, it sends out activation/inhibition message(s) to its direct downstream node(s). The node representing the protein complex is associated with a more complex function that simulates the multi-level regulation according to biological data, such as simulation of light driven, simulation of transcription regulation by protein compound, and simulation of nuclear translocation.

In our model we make the assumption that when a node reaches its peak level, it remains there until it receives a message of stop activation or start inhibition from its direct upstream node(s). Similarly, we assume that when a node reaches its trough level it stays there until a message is received to start activating or stop inhibition. To simplify the model, we incorporate the functions of some other regulate genes/proteins (such as Clock, Cryl and Cry 2) in the core clock system as part of the regulate functions in node M or N.

Figure 1 shows our gene network of the wild type mammalian biological clock. It includes two feedback loops. Solid lines represent positive feedback and dash lines represent negative feedback. M and N are protein complexes, and are used to simplify the simulation. Bmal_M_Rev-erbα feedback loop is non-light driven and cannot be affected by light directly. M_Per_N feedback loop is more important and directly regulated by light. We use this same network to simulate the genes/proteins behaviors in gene mutant/deletion animal, by putting the node representing the dysfunction gene/protein in a "sleep state," rendering it incapable of receiving and sending messages.

In future work we will consider our network of the single cell synchronization as the first step in explaining the synchronization in the different levels of the clock hierarchy. This should also include synchronizations among cells in same tissue/organ, synchronization among tissues, etc. Such a network structure would enable the simulation of the sophisticated relationships among core clock genes/proteins and hundreds of clock controlled genes/proteins, and therefore may be useful in analyzing health problems related to the system level confusion in the circadian rhythms, such as work shift, time zone effect and jet lag.

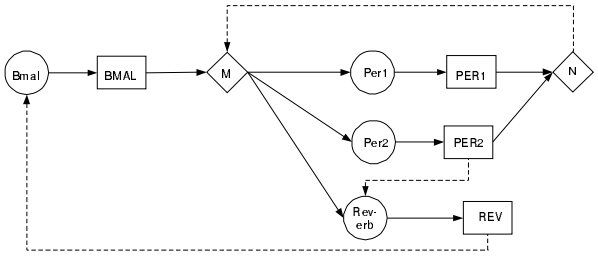


Figure 1: simulation wild type mammalian clock

References

[1] Leloup J. and Goldbeter A. 2003. Toward a detailed computational model for the mammalian circadian clock. PNAS. Vol. 100. pp. 7051-7056.

[2] Van Gelder, R. N. , Herzog, E. D. , Schwartz, W. J. and Taghert P. H. 2003. Circadian rhythms: In the loop at last. Science. Vol .300. pp. 1534-1535.

E18. The EcoTFs Database: *Escherichia Coli* Transcription Factors and Signals

William S. Hlavacek[1], Michael L. Blinov[2], Michael A. Savageau[3], Michael E. Wall[4]

Keywords: autoregulation, database, gene circuits, gene expression, design principles, *Escherichia coli*, signal molecule, transcription factors

1 The EcoTFs database.

The EcoTFs web site (http://EcoTFs.lanl.gov) is dedicated to the assembly and dissemination of information about *Escherichia coli* transcription factors and the signals that control their activity.

This web site was established in November/December of 2003 to provide an online database of information about autoregulation of 50 transcription factors (TFs) in *E. coli* as supplementary material in support of a recently published review of gene circuit design [Wall et al., 2004].

Unique features of this database (currently an html table with links) include annotation of the signal(s) influencing the activities of each TF and classification of each TF based on 1) the response to a stimulus (induction or repression of regulated effector genes), 2) the mode of regulation at the promoters of regulated effector genes (repressor or activator control), and 3) the co-regulation of TF and effector gene products in response to signals. The information catalogued in the database allows the distribution of gene circuit types to be studied (Table 1).

The database was developed to help test theoretical predictions of classifications for elementary gene circuits [Hlavacek and Savageau, 1996; Wall et al., 2003]. A repressor mode of regulation at the TF (*i.e.*, negative autoregulation) is predicted when stability, robustness and responsiveness are important performance criteria, as is expected for many circuits such as those that control metabolic functions. The co-regulation of TF and effector gene products may be classified as direct coupling (TF and effector expression change in the same direction in response to signal), inverse coupling (TF and effector change in opposite directions), or uncoupled (TF expression does not change). Prediction of coupling type is related to the gain of effector gene products with signal (Table 2).

Over 100 TFs in E. coli have been studied experimentally (and there are more putative TFs based on the genome sequence), so the current version of the database is incomplete. We plan to add material until the database is as comprehensive as possible for *E. coli*. We also plan to extend the scope of the database to include information about well-studied TFs in other bacteria.

[1] Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. E-mail: wish@lanl.gov

[2] Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. E-mail: mblinov@lanl.gov

[3] Department of Biomedical Engineering, One Shields Avenue, University of California, Davis, CA 95616, USA. E-mail: masavageau@ucdavis.edu

[4] Computer and Computational Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. E-mail: mewall@lanl.gov

2 Figures and tables.

Effector TU	Repressor mode of control at Regulator TU			Activator mode of control at Regulator TU			TF does not influence transcription of Regulator TU
	I	U	D	I	U	D	N/A
Inducible (+)	4[b]	3[c]	4[d]	0	0	5[e]	4[f]
Inducible (-)	0	0	9[g]	0	0	0	4[h]
Repressible (+)	0	3[i]	0	0	0	0	2[j]
Repressible (-)	0	1[k]	9[l]	0	0	0	1[m]

Table 1: Distribution of system types among 49 surveyed *E. coli* TFs. The following footnotes indicate the sets of transcription factors that correspond to the table entries: [b](AraC, IlvY, MetR, SoxS); [c](CynR, SoxR, TorR); [d](CysB, DsdC, MelR, RhaS); [e](CpxR, IdnR, MarA, RhaR, XylR); [f](MalT, MhpR, Rob, XapR); [g](BetI, CytR, EmrR, GalS, MarR, NagC, PdhR, PutA, UxuR); [h](GalR, GlpR, LacI, RbsR); [i](AsnC, GcvA, PspF); [j](FadR, FruR); [k](TyrR); [l](ArgR, DnaA, Fur, H-NS, IscR, MazEF, MetJ, PurR, TrpR); [m](ModE). D – direct coupling, I – inverse coupling, U – uncoupled, N/A – no TF self-regulation.

Effector TU	Low Gain	Intermediate Gain	High Gain
Inducible (+)	I	U	D
Inducible (-)	D	U	I
Repressible (+)	I	D	D
Repressible (-)	D	U	I

Table 2: Predictions of coupling type for elementary gene circuits. (-) indicates a repressor mode of control, and (+) indicates an activator mode of control. Predictions depend on both the mode of control of effector expression and the magnitude of the steady-state gain of effector gene products with signal. The predictions for inducible and repressible systems are identical except for the case of activator control with intermediate gain. D – direct coupling, I – inverse coupling, U - uncoupled.

3 References and bibliography.

The EcoTFs web site is available at http://EcoTFs.lanl.gov

References

Hlavacek, W. S. and Savageau, M. A. (1996) Rules for coupled expression of regulator and effector genes in inducible circuits. *J. Mol. Biol.* 255:121-139.

Wall, M. E., Hlavacek, W. S. and Savageau, M. A. (2003) Design principles for regulator gene expression in a repressible gene circuit. *J. Mol. Biol.* 332:861-876.

Wall, M. E., Hlavacek, W. S. and Savageau, M. A. (2004) Design of gene circuits: lessons from bacteria. *Nat. Rev. Genet.* 5:34-42.

E19. Discovering Activated Regulatory Networks in the DNA Damage Response Pathways of Yeast.

Chris Workman[1], **Scott McCuine**[1], **Ryan Kelley**[1], **Trey Ideker**[1]

Keywords: regulatory networks, systems biology, DNA damage

1 Abstract

Modeling of the molecular interaction networks induced by DNA damaging agents is likely to reveal rich new insights into the mechanisms of the cellular DNA damage response, a determinant of cancer progression and cellular toxicity. To further elucidate the DNA damage response in eukaryotes, we have developed in-silico network models of regulatory pathways responding to the DNA damaging agent methyl methane sulfonate (MMS). Models were constructed using data derived from an array of classical, genomic and proteomic approaches and were refined using computational systems biology approaches. Regulatory pathways were systematically interrogated to monitor protein-DNA interactions using chromatin immunoprecipitation in conjunction with promoter microarrays (chIP-chip); yeast 2 hybrid based methods to identify protein-protein interactions[1]; DNA microarrays to monitor genome-wide expression patterns; and systematic genetic perturbations via single gene-knockout strains. New computational techniques were used to integrate and model these data and include algorithms for statistical identification of expression-activated network regions[2] and the Cytoscape visualization platform[3] for operating on network models. Using this systems approach, we have generated new hypothesis about cellular factors and mechanisms involved in the complex interactions of DNA damage response.

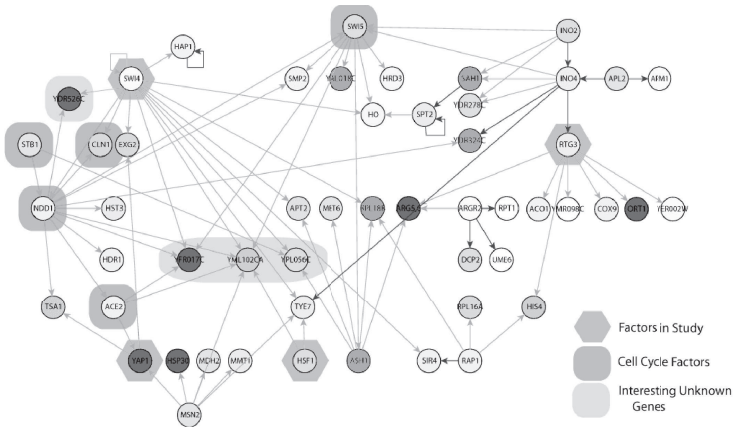


Figure 1: Activated regulatory network found in Cytoscape[3] using the ActiveModules approach[2].

[1]Department of Bioengineering, University of California San Diego, 9500 Gilman Dr. 0412, La Jolla, CA 92093
cworkman@bioeng.ucsd.edu

References

[1] Uetz, P. *et al.* 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403:623-7.

[2] Kelley, B.P. *et al.* Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proc Natl Acad Sci USA* 100, 11394-9.

[3] Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498-504.

E20. Parallel Data Mining of Bayesian Networks from Gene Expression Data

Longde Yin[1], Chun-Hsi Huang[2], Sanguthevar Rajasekaran[3]

Keywords: Gene Regulatory Networks, Genetic algorithm, Parallel Data Mining

DNA microarrays allow monitoring gene expression for tens of thousands of genes in parallel and are already producing huge amounts of valuable Gene Expression Data. Uncovering gene/protein interaction and key biological feature of cellular systems from these data is a major challenge in computational biology. Bayesian network (BN) is a promising method to describe relationships between genes in a genetic regulatory network. However, learning Bayesian network structure is an optimization problem in the space of directed acyclic graphs [1]. The number of such graphs is super-exponential in the number of variables. Therefore, we need to develop high-performance parallel search algorithms.

In the work described here the problem is to find the best Genetic Regulatory Network in a very large solution space of all possible Bayesian networks. Since the problem is NP-hard [3], a heuristic search technique must be used. This leads to the employment of Genetic algorithm (GA), since GA has been shown to be a robust and effective search method requiring very little information about the problem to explore a large search space. The GA works on a population of solutions, which change as the algorithm cycles through a sequence of generations, until a satisfactory solution has been found. Solutions are directed graphs; viable solutions (those which will be scored and allowed to breed in the next generation) are directed acyclic graphs. The scoring function used was the Bayesian scoring metric (BSM) [1], in which the best fit to the experimental data is calculated using Bayesian techniques. High scoring structures have a greater chance of being selected as parents for the next generation [2].

Due to the sheer volume of data involved in data mining, the time required to execute genetic algorithms and the intrinsic parallel nature of genetic algorithms, we decide to parallelize the Genetic algorithm (PGA) and plan to take two different approaches to parallelizing the GA.

[1] Dept. of Computer Science & Engineering, University of Connecticut, USA , E-mail: yin@engr.uconn.edu

[2] Dept. of Computer Science & Engineering, University of Connecticut, USA. E-mail: huang@engr.uconn.edu

[3] Dept. of Computer Science & Engineering, University of Connecticut, USA. E-mail: rajasek@engr.uconn.edu

The first approach is to implement GA in master-slave model Breeding (reproduction, crossover and mutation) was carried out in parallel. In fact the scoring was also implemented in parallel. The selection had to be implemented sequentially and thus remained on the master (the root processor which is the controller, and is connected to the host). This was necessary, as all of the structures from the new generation needed to be re-mixed to form new parents from the gene pool before distribution to the slaves for breeding. The remaining processors are utilized as slaves, which carry out the breeding in parallel and report the new structures and their scores to the master (root processor) [2].

The second approach is to divide the population into subpopulations, run a conventional GA in each subpopulation and allow the periodic communication of information between subpopulations that helps in the search for the solution. The information usually exchanged between subpopulations is a subset of the fittest individuals of each subpopulation. This exchange of individuals is known as migration. This parallel version of GA actually is a distributed GA.

Although this project is still in progress, it is anticipated that the PGA should enhance the efficiency of genetic search and has higher probability to get the optimal solution than the GA. Since the PGA can make use of multiple computing resources at the same time and can divide the large problem into several smaller ones.

In the poster session, presentation will address the problem of parallelization of Genetic Algorithm for the Mining of Bayesian Network based on Gene Expression Data. Two approaches to parallelizing the GA will be presented in detail. The results of the performance study of PGA and PGAs' applications in Data Mining will be presented too.

References

[1] Nir Friedman and et al. 2000. Using Bayesian Networks to Analyze Expression Data, *J Comput Biol.*, 7(3-4):601-20.

[2] Roy Sterritt and et al. 2000. Parallel Data Mining of Bayesian Networks from Telecommunications Network Data, *IPDPS Workshops 2000*: 415-426.

[3] Chickering D.M. and D. Heckerman. 1994. Learning Bayesian network is NP-hard, *Microsoft Research*, MSR-TR-94-17.

E21. Discovery of Gene-Regulation Pathways in Mouse Asbestos Using Background Knowledge

Changwon Yoo[1], Mark Pershouse[2], Elizabeth Putnam[2]

Keywords: Bayesian networks;systems biology; causal discovery;asbestos;gene networks

1 Introduction.

A gene expression study using DNA microarrays usually involves two major steps. The first step typically consists of performing initial experiments to narrow the set of genes to study further in more detail. The experimenter can avoid this first step if he or she already knows the set of genes of interest. For example, the genes involved in galactose metabolism in yeast are relatively well known, so an experimenter could skip the experiments in the first step in the study. After choosing those genes, the experimenter has to produce an experimental design for further study of how those genes regulate each other.

Asbestos fibers small enough to be inhaled, and numerous enough to overcome the normal host defenses can lodge in the lungs, leading to chronic inflammation, pulmonary fibrosis (asbestosis), pleural thickening as well as cancers of the lung and pleura. Asbestos fibers are clearly associated with induction of mesothelioma, a neoplasm derived from the mesothelial lining of the pleural cavity, but the mechanism is unclear [1]. We describe our initial effort in a gene expression study that is designed to learn causal relationships among genes that play an important role in mouse asbestos. In this paper, we concentrate on assessing expert biologist's knowledge of pairwise relationships.

2 Analysis.

Causal networks represent causal relationships using a graphical model. Graphical models hold great promise as representations of molecular biological processes, because they are both expressive and intuitive. In a recent issue of *Science*, the authors of four separate review articles on bioinformatics and related topics described graphical models as one of the most promising methods for representing cellular pathways [2-5]. Three of the articles specifically mention causal Bayesian networks as a promising type of graphical model. Particularly, Kitano [5] refers to one of our causal analysis papers [6] and emphasize the importance of causal discovery in systems biology research. We use Bayesian networks to model interactions of genes in mouse asbestos.

A causal Bayesian network is a directed acyclic graph in which each arc is interpreted as a direct causal influence between a parent node and a child node, relative to the other nodes in the network [7]. One of the challenges in applying Bayesian methods for causal discovery is the assessment of informative priors on possible causal structures and on the parameters of those structures. On the one hand, the ability to represent such prior information is a great strength of the Bayesian approach. With it, we can potentially express prior causal knowledge that comes from many sources other than the observational data. While good progress has been made in facilitating the expression of priors on Bayesian network structures and parameters [8], assessing such prior probabilities (particularly when

[1] Dept. of Computer Science, Univ. of Montana, Missoula, MT. E-mail: cwyoo@cs.umt.edu
[2] School of Pharmacy, Univ. of Montana, Missoula, MT. E-mail: {markp,lizp}@selway.umt.edu

there is a large set of variables) can still be difficult. Note that it will be impractical to assess priors of all of the possible pairwise relationships especially if you are dealing with more than 5,000 genes. We asked expert biologists to assess priors of pairwise relationships that they think are relatively well known, and we assume default prior probabilities for the remaining relationships. We currently use the following assumptions (unless the expert biologist specifies his or her knowledge): $p(E_i^{XY}) = 1/3$ for i=1,2,3. For genes X and Y, E_i^{XY} represents $X \leftarrow Y$, $X \rightarrow Y$, and $X \quad Y$ for i=1,2,3 respectively. We list relatively well known pairwise relationships assessed by expert biologists in Table 1.

Relationship	Prior	Relationship	Prior	Relationship	Prior	Relationship	Prior
SRA→TNFα	0.9	*SRA → IL6*	0.6	*SRA → IL1*	0.6	*IFNγ →TNFα*	0.5
IL1 → IL6	0.8	*IFNγ → IL6*	0.5	*IFNγ → IL1*	0.5		

Table 1: Priors assessed by expert biologists about pairwise relationships among genes that play an important role in mouse asbestos.

3 Future Research.

We will combine the background knowledge shown in Table 1 with microarray experiments with 108 mice that were exposed in different asbestos agents [9]. We are planning to use a pairwise causal algorithm, Implicit Latent Variable Scoring (ILVS) method [6] and its extension Local ILVS Method (LIM) that analyzes local structures with more than pairwise variables [10]. There are many challenges in the planned analysis, e.g., (1) causal discovery with no direct manipulation of genes; (2) global network inference with local networks.

Once we develop a model that combines expert's background knowledge and the results of microarray experiments, we are planning to develop and evaluate a system that recommends experimental design of a gene expression study, e.g., what gene to knock out; how many experimental repetitions to make.

References.

[1] Fung, H., et al., *Patterns of 8-hydroxydeoxyguanosine (8)HdG) formation in DNA and indications of oxidative stress in rat and human pleural mesothelial cells after exposure to crocidolite asbestos.* Carcinogenesis, 1997. **18**: p. 101-108.
[2] Karp, P.D., *Pathway databases: A case study in computational symbolic theories.* Science, 2001. **293**: p. 2040-2044.
[3] Gifford, D.K., *Blazing pathways through genetic mountains.* Science, 2001. **293**(2049-2051).
[4] Mjolsness, E. and D. DeCoste, *Machine learning for Science: State of art and future prospects.* Science, 2001. **293**: p. 2051-2055.
[5] Kitano, H., *Systems Biology: A Brief Overview, , March 1, 2002.* Science, 2002. **295**: p. 1662-1664.
[6] Yoo, C., V. Thorsson, and G.F. Cooper. *Discovery of a gene-regulation pathway from a mixture of experimental and observational DNA microarray data*. in *Pacific Symposium on Biocomputing*. 2002. Maui, Hawaii: World Scientific.
[7] Pearl, J., *Probabilistic Reasoning in Intelligent Systems*. Representation and Reasoning, ed. R.J. Brachman. 1988, San Mateo, CA: Morgan Kaufmann.
[8] Heckerman, D., D. Geiger, and D. Chickering, *Learning Bayesian networks: The combination of knowledge and statistical data.* Machine Learning, 1995. **20**: p. 197-243.
[9] Driscol, K., et al., *Intratracheal instillation as an exposure technique for the evaluation of respiratory tract toxicity: uses and limitations.* Toxicol. Sci., 2000. **55**: p. 24-35.
[10] Yoo, C. and G. Cooper. *Discovery of gene-regulation pathways using local causal search*. in *AMIA*. 2002. San Antonio, Texas.

E22. On Some Choices in Bayesian Network Learning for Reconstructing Regulatory Networks

Xuesong Lu[1], Xing Wang[2], Ying Huang[1], Wei Hu[3], Guang R. Gao[2], Yanda Li[1] and Xuegong Zhang[1,*]

Keywords: Bayesian Network, Parameter Choice, Gene Networks

1 Introduction.

Using high-throughput biological data such as gene expression data to study the gene regulatory networks has become one of the most attractive topics in bioinformatics research. Among the efforts, many researchers are trying to use Bayesian networks (BN) to reconstruct regulatory networks. Different levels of success have been reported on real biological microarray data (such as Friedman et al, [1]), on synthetic data from simulated models (such as Zak et al, [2]), and on both real and simulated data (such as Husmeier et al, [3]). It has been observed that certain choices in BN learning can have big influence on the result, thus it is necessary to take a systemic investigation on the effects of these choices. In this paper we focus on the choices of some basic parameters in the BN learning procedure, such as restrictions on the network structure, input data type (discrete or continuous) and initial structure. We use a well-studied biological network [4] to generate synthetic expression data, use dynamic Bayesian networks (DBN) with different parameters to reconstruct the network and observe the influences of the choices. The observations provide some information useful for reconstructing regulatory networks with BN learning.

2 Data and Methods.

The network model used in this study is a 17-node sub-network (MAPK kinase cascade) from the network in [4]. Starting from certain initial condition of the nodes, the time series of expression of each node can be computed from the set of partial differential equations. We generated five different datasets. Datasets 1, 2 and 3 are composed of the 17 nodes with different sampling intervals and thus result in different sample sizes (200, 500 and 2000). Datasets 4 and 5 include only 9 nodes of the sub-network, the sample sizes are the same with datasets 2 and 3 respectively.

Because the standard Bayesian networks can not solve the problem of loops, we used dynamic Bayesian network or DBN to reconstruct the regulatory network, where each original node (gene) is represented by one node at time t and one node at time $t+\Delta t$. The learning algorithm used in this paper is based on the software package PNL (http://www.intel.com/research/mrl/pnl/). We used the Maximum Likelihood algorithm for discrete data learning and the linear Gaussian algorithm for continuous data learning. Three discretization methods (by difference, by mean or by median) were compared, and we chose the method of binary discretization by difference because the three discretization method does not show big difference on the datasets.

[1] MOE Key Laboratory of Bioinformatics/Dept of Automation, Tsinghua Univ., Beijing 100084, China
[2] Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716, USA
[3] Intel China Research Center, 06-01 Beijing Kerry Centre, Beijing 100020, China
[*] Corresponding author, e-mail: zhangxg@tsinghua.edu.cn. This work is partially supported by NSFC grants 60275007 and 60234020 and an Intel Corporation Grant.

3 Results and Discussion.

The first observation was that adopting proper restrictions to the DBN structure according to prior knowledge can help improving the learning. For example, with the setting of DBN, since the expression of a gene at time $t+\Delta t$ always depends that at time t, such connections in the DBN structure are trivial. However, these signals are sometimes so strong that they prevent the DBN to learn other weaker signals. Experiments showed that forbidding such connections during the learning phase can significantly improve the number of correctly reconstructed links in the network.

There is always the uncertainty about whether expression data should be used in discrete or continuous form for reconstructing gene networks, and what kind of initial structure is better for BN learning. We did a systematic comparison with different choices on our 5 datasets. The observation is, with discrete expression data, using empty initial structure (no links exist at the beginning of learning) usually produces better learning result than using a chain initial structure (where there is a link between any two adjacent nodes at the beginning). However, if continuous expression values are used and when we have a relatively larger sample size (say, 20 times of the number of nodes), starting with a chain structure is better. Generally continuous values show advantage in the learning when we have a large sample size, otherwise they do not show big difference with discrete expression values and thus since BN learning with continuous features are more complicated than with discrete features, we suggest that for limited sample sizes, adopting a proper way to discrete the expression data benefits the learning with regard to both the computation and the performance. The following table summarizes the major suggestions from this study about Bayesian network learning for reconstructing regulatory networks from expression data.

Sample Size	Data type	Initial structure	Prior Restriction
small	Discrete expression values	Empty initial structure	Proper restriction on
large	Continuous expression values	Chain initial structure	BN structure always benefits

It should be noted that with a careful choice of all the learning parameters, the reconstructed network was still far from perfect. For example, in the 9-node model, only 10 of the 16 real links can be correctly recovered, and there are also 7 false positive links. However, this was based on a direct comparison of the edges connecting the nodes. Since the given model is described as network of differential equations and the learned model is a probabilistic model, direct comparison may not be a proper way to evaluate the learning performance. After manually checking those false positive links, it turned out to be that only 1 among them was really a false link, and the probabilistic dependences indicated by the other links actually all exist in the original model. This reveals that the BN is capable of learning regulatory networks from expression data, but a proper way of evaluation needs to be investigated.

References

[1] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. 2000. Using Bayesian networks to analyze expression data. *J. Comp. Bio.,* 7:601–620.

[2] Zak, D. E., Doyle, and Schwaber, J. S., 2001. Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data. In: *Proceedings of the Second International Conference on Systems Biology*, 231-238.

[3] Husmeier D. 2003. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics* 19: 2271-2282.

[4] B. Schoeberl, C. Eichler-Johnsson, E. D. Gilles and G. Mueller. 2002. Computational Modeling of the Dynamics of the MAP Kinase Cascade Activated by Surface and Internalized EGF Receptors. *Nature Biotechnology,* 20:370-375.

E23. PathBLAST : Mining For Conserved Pathways In Genome-Scale Molecular Interaction Networks

Silpa Suthram[1], **Taylor Sittler**[1], **Trey Ideker**[1]

Keywords: interaction networks, PathBLAST, conserved pathways

1 Introduction.

Genome scale protein interaction networks are being mapped and identified at an ever-increasing speed. In order to make sense of this deluge of new data, new tools are needed in order to assemble a coherent picture of these interaction networks and their functional significance. PathBLAST [1]is one such method for identifying conserved pathways between two molecular interaction networks. Pathways are scored and selected based on sequence homology and topological similarity using a dynamic programming algorithm, as previously described [1]. Since interaction networks are obtained from high-throughput methods, these often include many false positives. To help eliminate the false interactions, here we present a new scoring system to assign a confidence value to each interaction. The model considers the co-expression of interactors [2], the number of times an interaction was observed and its topological clustering coefficient [3]. Using the revised PathBLAST, we report new biological findings illustrating conserved pathways between yeast and a variety of other model species for which interaction data is already available (Fig.1). The ability to mine conserved pathways across multiple species provides insight into how these pathways have evolved and can delineate roles for previously uncharacterized proteins through analysis at the level of both sequence and function.

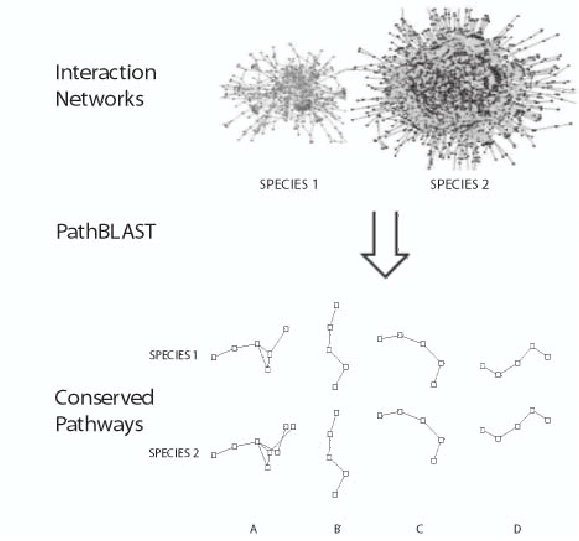


Figure 1: Mining conserved pathways using PathBLAST[1].

[1]Department of Bioengineering, University of California San Diego, 9500 Gilman Dr. 0412, La Jolla, CA 92093
ssuthram@ucsd.edu

2 References and bibliography.

References

[1] Kelley.B.P *et al.* 2003. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *PNAS* 100:11394-11399

[2] Grigoriev.A 2001. A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Research* 29 : 3513-3519

[3] Goldberg.D.S *et.al* 2003. Assessing experimentally derived interactions in a small world. *PNAS* 100 : 4372-4376

E24. Learning kernels from biological networks by maximizing entropy

Koji Tsuda, [1] **William Stafford Noble** [2]

Keywords: function prediction, protein interaction networks, metabolic pathways, support vector machines, diffusion kernels

1 Introduction

When predicting the functions of unannoted proteins based on a protein network, one relies on some notions of "closeness" or "distance" among the nodes. However, inferring closeness among the nodes is an extremely ill-posed problem, because the proximity information provided by the edges is only local. Moreover, it is preferable that the resulting similarity matrix be a valid *kernel matrix* so that function prediction can be done by support vector machines (SVMs) or other high-performance kernel classifiers [2]. Maximum entropy methods have been proven to be effective for solving general ill-posed problems. However, these methods are concerned with the estimation of a probability distribution, not a kernel matrix. In this work, we generalize the maximum entropy framework to estimate a positive definite kernel matrix.

We found that the *diffusion kernel* [1], which has been used successfully for making predictions from biological networks (e.g. [3]), can be derived from this framework. However, one drawback inherent in the diffusion kernel is that, in the feature space, the distances between connected samples have high variance. As a result, some of the samples are *outliers*, which should be avoided for reliable statistical inference. Our new kernel based on local constraints resolves this problem and thereby shows better accuracy in yeast function prediction.

2 Locally Constrained Diffusion Kernels

SVMs work by embedding samples into a vector space called a *feature space*, and searching for a linear discriminant function in such a space [2]. If we have an undirected graph with n nodes and m edges, the n nodes in a graph are mapped to n points in the feature space $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \mathcal{F}$. The embedding is defined implicitly by specifying an inner product via a positive definite kernel matrix $K_{ij} = \boldsymbol{x}_i^\top \boldsymbol{x}_j, i, j = 1, \cdots, n$. Because the discriminant function is solely represented by inner products, we do not need to have an explicit representation of $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$. Once a kernel matrix is determined, the (squared) Euclidean distance between two points can also be computed as $D_{ij} := \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = K_{ii} + K_{jj} - 2K_{ij}$.

We have found that the matrix of the diffusion kernel [1] can be derived as the optimal solution of the following maximum entropy problem:

$$\min_K \operatorname{tr}(K \log K), \quad \operatorname{tr}(K) = 1, \operatorname{tr}(KL) \leq c,$$

where log denotes the matrix logarithm operation, c is a positive constant, and L is the graph Laplacian matrix [1]. Let $\{s_j, t_j\}_{j=1}^m$ denote the node pairs connected by m edges.

[1] MPI for Biological Cybernetics, Spemannstr. 38, 72076 Tübingen, Germany, and AIST CBRC, Tokyo, Japan. E-mail: koji.tsuda@tuebingen.mpg.de

[2] Dept. of Genome Sciences and Dept. of Computer Science, University of Washington, 1705 NE Pacific St., Seattle, WA 98109, USA E-mail: noble@gs.washington.edu

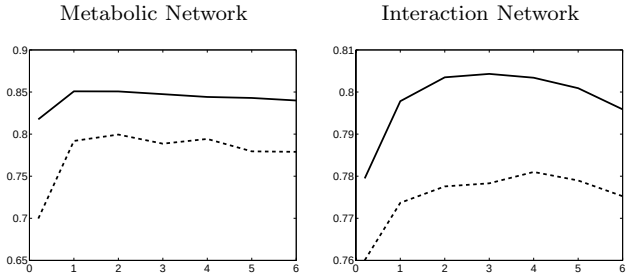


Figure 1: **Mean ROC score as a function of the diffusion parameter.** The plots show the mean ROC scores computed across the set of CYGD categories, using (A) the metabolic network and (B) the protein-protein interaction network. The solid and broken lines correspond to the new and conventional kernels, respectively.

The quantity $\mathrm{tr}(KL)$ equals the sum of Euclidean distances between connected samples: $\mathrm{tr}(KL) = \sum_{j=1}^{m} \|\boldsymbol{x}_{s_j} - \boldsymbol{x}_{t_j}\|^2$. The objective function corresponds to the (negative) von Neuman entropy. In order to impose a more uniform network structure, we consider the following *local constraints*:

$$\min_K \mathrm{tr}(K \log K), \quad \mathrm{tr}(K) = 1, \mathrm{tr}(KV_j) \leq \gamma, \quad j = 1, \cdots, m, \tag{1}$$

where $\mathrm{tr}(KV_j) = \|\boldsymbol{x}_{s_j} - \boldsymbol{x}_{t_j}\|^2$ corresponds to the Euclidean distance between each pair of connected samples.

3 Experiments

We computed kernels from two different types of yeast biological networks. The first network was derived by [3] from the LIGAND database of chemical reactions in biological pathways. The second network was created by [4] from protein-protein interactions. We tested the kernels' utility in the context of an SVM classification task. We used as a gold standard the functional categories of the MIPS Comprehensive Yeast Genome Database. We selected all functional categories containing at least 30 positive examples resulting in 36 categories for the metabolic network and 76 categories for the protein-protein interaction network. Figure 1 compares the classification performance of SVMs. The figure shows that, for both types of network, our new kernel out-performs the conventional diffusion kernel.

References

[1] I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of ICML 2002)*, pages 315–322. Morgan Kaufmann, 2002.

[2] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[3] J.P. Vert and M. Kanehisa. Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. In *Advances in Neural Information Processing Systems 15*, volume 15, pages 1425–1432. MIT Press, 2003.

[4] C. von Mering, *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.

F1. Phylogeny of Tumor Progression from CGH Data

Sven Bilke[1], *Qing-Rong Chen*[1], *Javed Khan*[1]

Keywords: Tumor progression, Comparative Genomic Hybridization, Microarrays

1 Introduction.

Genomic instability, gain or loss of DNA, is frequently observed in tumors [3]. While normal cells contain two copies of each chromosome (with the exception of the gender chromosomes X and Y), tumor cells often loose one or both copies or gain extra copies of specific regions of DNA. The chromosomal location of changes is not random: in many cases instable regions from different patients with the same disease cover the same locations, which are therefore called recurrent regions. It is often argued that, in the competition for nutrition and space, changes at specific locations provide an advantage and are hence more frequent (recurrent) than instabilities at neutral or adverse locations. In this 'micro-evolution' [3, 1] the patterns of genomic instabilities allow to conclude on the inheritance of genomic features and hence the 'phylogeny' of different tumor stages. In this study we use the overlap of the location of genomic instabilities shared between and unique to different phenotypes as indicators for the pathways of tumor progression. To demonstrate the feasibility of this approach we analyze cytogenetic data for three stages of Neuroblastoma the most frequent pediatric childhood tumor, and identify a progression model for this cancer by an exhaustive search in the space of progression models. The data for genomic instabilities used in this study was obtained from 32 neuroblastoma [2] specimens, 12 patients with stage 1, and 20 patients with stage 4 of which 12 were MYCN-amplified (4+) and 8 were MYCN single-copy tumors (4-). Comparative genomic hybridization on a cDNA array with 42000 elements was used to measure the relative DNA copy number. P-values for the presence of gains or losses were estimated by a sliding window method, where the distribution of genome-ordered observations in that window was compared by a t-test to the distribution observed for the full genome.

2 Results and Discussion

Under the assumption that the gains and losses are milestones in the development of the disease, the progression between different stages should be observable in the pattern of genomic instabilities. In order to develop models which can be discriminated by genomic imbalance data, we restrict our analysis to models which follow these biological principles:

1. Unobserved intermediate genotypes are possible but the model with the smallest number of genotypes (observed + unobserved) is utilized
2. All changes found in a parent genotype must be present in the offspring occurring with a similar frequency (the inheritance signature)
3. All tumor stages belonging to the same diagnostic group arise from a common ancestor (i.e. the phylogeny is a rooted tree).

The models compatible with these requirements are discriminated by the patterns of overlap of recurrent genomic instabilities, namely regions *common* to all stages, *specific* to a stage and *shared* between stages. In this feasibility study we analyze a relatively small number of three observed stages, defining seven different subsets which may either be empty or occupied. This allows for $2^7 = 128$ different observations. Not all of those are compatible with the concept of tumor progression in form of a micro-evolution with a specific genetic signature. For example, the case where none of the sets is occupied describes the case where the progression

[1]Oncogenomics Section, Pediatric Oncology Branch, Advanced Technology Center, National Cancer Institute, 8717 Grovemont Circle, Gaithersburg, MD 20877, USA. E-mail: {bilkes,chenqi,khanj}@mail.nih.gov

of neuroblastoma does not manifest itself in a specific signature of genomic instabilities. We identify ten topologically distinct models of tumor progression compatible with the biological assumptions. These models are depicted in the left panel of figure 1. As mentioned in the introduction, no single region is found to be affected in *all* tumors of a specific stage. Also, no single region is affected exclusively in one stage. However, the frequency of their occurance is often significantly different in the different stages of the disease. Therefore the frequency of a genomic instability for the different stages is the primary observable. We define a region to be *specific* to a stage if it is significantly more frequent in one stage as compared to all the other stages. If a region is significantly more frequent in two stages with respect to the third stage, this regions is called *shared*. If a region is frequent in all stages, the region is called *common*. In our analysis of the neuroblastoma-data we have verified that the final result is stable with respect to reasonable changes of the significance thresholds used in defining these regions.

We found recurrent alterations common to all three subgroups, specific for each of the subgroups and common regions of gain for stage 1 and 4- tumors. Interestingly there were no shared alterations of 4+ with 1 or 4- besides the regions common to all. The tumor progression model compatible with these findings is depicted in the right panel of figure 1.

In this study we demonstrate that the pattern of genomic instabilities in neuroblastoma can be used to conclude on the pathway of progression of the disease. The analysis of the frequency of genomic instabilities maps our data onto one of the models compatible with the assumption that 'micro-evolution' governs the progression of cancer. The selected model is in agreement with clinical evidence [4] for neuroblastoma. The identification of the progression pathways for cancer may have an important impact on the strategy for the treatment. In Neuroblastoma our result indicate that the model of a linear progression towards the more aggressive disease (figure 1 (a), sub-diagram I), which is seemingly suggested by the staging system, is not supported by our data. Instead, the best fitting model (right panel of figure 1) suggests that the final outcome of the disease is determined at an very early stage of the cancer development. The stage 4+ disease is fundamentally different from Stages 1 and 4-, also a developed stage 1 tumor does not progress to stage 4-.

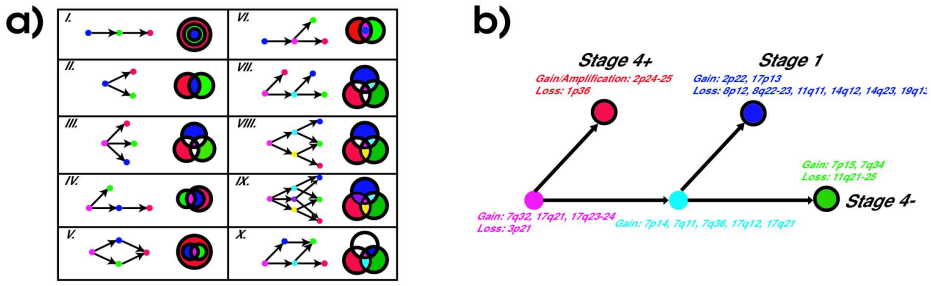


Figure 1: The ten Topologically distinct tumor progression models for three observed stages. Unobserved, intermediate states are drawn in cyan and magenta, observed stages in red, green and blue. b) The selected model summarizing the pattern of recurrent genomic instabilities in our data.

References

[1] For a review see e.g. Bogen K.T. (1989) J. Natl. Cancer Inst. 81, 267-277.

[2] For a review of Neuroblastoma biology see e.g. Brodeur, G. M. (2003) Nat Rev Cancer 3, 203-16.

[3] Rajagopalan, H, et.al. Nat Rev Cancer. 2003 Sep;3(9):695-701.

[4] Westermann, F., Schwab, M. (2002) Cancer Lett 184, 127-47.

F2. Human Transcript Clustering

Ronghua Chen[1], Archie Russell, Guoya Li, Nicholas Tsinoremas, Guy Cavet

Keywords: transcripts, genomic alignment, mRNA, EST clustering

1 Introduction.

Human genome sequencing projects have provided an opportunity for researchers to annotate genes at the whole genome scale. Two analyses of the sequence estimated number of genes ranging from 30,000 to 40,000 [3, 4]. However, a direct comparison of the predicted genes from these two analyses revealed little overlap in the prediction of novel genes [2]. In the absence of any single definitive method for identification of genes, a synthesis of complimentary methods is needed to prioritize a set of model genes and their associated relationships with transcript and genomic sequences. We describe here a method to create a gene model index which is as accurate and comprehensive as possible, folding together sequences from Celera and the public domain.

2 Methods and Results.

We first clustered and aligned all expressed human sequences in GenBank to create EST clusters. We then located the resulting consensus sequences as well as NCBI RefSeq, GenBank mRNA, Ensembl transcripts, UniGene unique representatives, NCBI model RefSeq (XM_) on the Celera human genome assembly by megablast [5]. Detailed alignments were generated with sim4 [1]. The vast majority of transcript sequences from most sources can be aligned to the human genome. Once transcript sequences were aligned to the genome, coordinates of exons, introns and genes with aligned transcripts were loaded into a relational database. Alignments were clustered together into sets which represent genes. Two alignments were considered to represent the same gene only if their exons overlap at the same location of the chromosome. This allows correct resolution of overlapping but distinct genes. In total ~3.5 million transcripts including ESTs were clustered into 130,000 gene models. These models can be prioritized by confidence and an appropriate subset can be selected for any given task. For examples, the gene model index may be used to support microarray design, the calculation of cross-species mappings, and the prediction of splice variants by transcript sequence.

References

[1] Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M. & Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research* 8:967-974.

[2] Hogenesch, J. B. et al. 2001. A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. *Cell* 106:413-415.

[3] Lander, E. S. et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860-921.

[4]Venter, J. C. et al. 2001. The sequence of the human genome. *Science* 291:1304-1351.

[1] Rosetta Inpharmatics, Merck & Co. Inc., 12040 115th Ave NE, Kirkland, Washington 98034, USA. E-mail: ronghua_chen@merck.com

[5] Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *J. Computational Biology* 7:203-214.

F3. Massively Parallel DNA Sequencing using Single Molecule Array Technology

Anthony J. Cox[1]

Keywords: Whole genome resequencing, single molecule array.

1 Introduction

Solexa is developing a unique technology that has the potential to transform the economics of DNA sequencing by allowing the sequence of millions of individual DNA molecules to be rapidly determined in parallel. Our approach obviates the need for sorting, cloning and amplification of genomic DNA samples and so lab preparation and reagent overheads are also drastically reduced. The applications of a re−sequencing technology range from SNP determination to transcriptomics.

As well as outlining the basic ideas behind Solexa's sequencing platform, this presentation will describe the high throughput bioinformatics pipeline that we are developing to process the large volumes of image and sequence data that our platform will generate.

2 Sequencing Technology

Genomic DNA is first purified and sheared and the resulting fragments are then immobilized onto a surface as primed single strands at a density of around 100 million molecules per square centimetre. These molecules are then sequenced with a base−by−base sequencing strategy employing proprietary polymerases and modified fluorescently labelled nucleotides. Sensitive optical methods allow the outcome of sequencing reactions to be observed simultaneously at a resolution of millions of individual DNA molecules, thus enabling their sequences to be determined in a massively parallel fashion. We aim to obtain at least 25 bases of sequence from each molecule imaged.

3 Bioinformatics Pipeline

The primary output of the sequencing process consist of a large number of images, each similar to Figure 1. The first task of our bioinformatics pipeline is to convert these data into a set of DNA sequences. This requires rapid processing of the images in real time. The resulting sequences are then aligned to the reference human genome sequence, allowing for sequencing errors and naturally occurring differences. We have developed software that has made it feasible for the large number of inexact alignments required to be performed on a Linux cluster with a modest number of nodes. Lastly, sequencing errors are filtered out to leave behind the genuine variation between the reference sequence and the genome being sequenced, in a final stage analogous to the consensus generation stage of shotgun sequence assembly.

[1]on behalf of Solexa Limited, Chesterford Research Park, Little Chesterford, Essex CB10 1XL, United Kingdom. anthony.cox@solexa.com

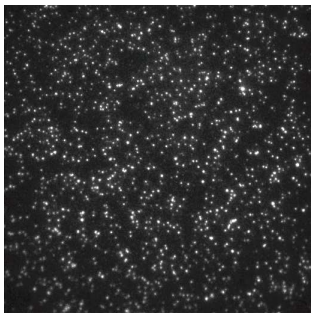


Figure 1: Actual Single Molecule Array image. Each spot is a single DNA molecule with labelled nucleotide attached.

4 Post Sequencing

We have secured UK government funding and set up academic collaborations for the co−development of systems and methods for storage and analysis of genome variation data derived from single molecule array sequencing.

F4. Clann: Software for Phylogenomic investigation and analysis of Horizontal Gene Transfer using supertrees.

Christopher J. Creevey[1] & James O. McInerney[1]

Bioinformatics and Pharmacogenomics Laboratory,
Department of Biology, National University of Ireland Maynooth,
Maynooth, Co. Kildare, Ireland.

Keywords: Whole genome phylogenetics, Supertree construction, Horizontal gene transfer.

Abstract:

Clann is a software tool, developed by Chris Creevey, which uses supertree analyses to investigate the Phylogenomic information content from whole genomes. This in turn allows the identification of Horizontal Gene transfer (HGT) events, and the characterization of the genes that are more or less likely to be involved in such events.

A number of methods of constructing supertrees have been devised (*1-8*) and a variety of supertrees have been constructed using both molecular (*9-11*) and morphological (*12-14*) datasets. However, for the most part, these methods focus on the production of a supertree with the assumption that the input trees are broadly in agreement and have not been used to address the issue of whether or not there really is an underlying phylogeny that can be accurately represented by a tree diagram. Here we use supertree construction as a step in addressing the issue of whether or not there really is an underlying phylogeny that can be accurately represented by a tree diagram. We previously have used this approach to investigate support for a tree-like phylogeny in the prokaryotes (*15*)

The supertree methods that Clann implements are as follows:

1. Source tree compatibility
2. Component /splits compatibility
3. Quartet compatibility
4. Matrix representation with Parsimony (MRP)
5. Most Similar Supertree (*15*)

Exhaustive and heuristic searches of supertree-space are possible using these methods as optimality criteria. Clann includes methods for testing the quality of the input data and the quality of the resulting hypotheses.

Clann is freely available for academic use and can be downloaded at :

http://bioinf.may.ie/software/clann

References:

1. B. R. Baum (1992) Combining Trees as a Way of Combining Data Sets for Phylogenetic Inference, and the Desirability of Combining Gene Trees *Taxon* **41**, 3-10.
2. A. D. Gordon (1986) Consensus Supertrees: The synthesis of rooted trees containing overlapping sets of laballed leaves *Journal of Classification* **3**, 335-348.
3. S. M. Lanyon (1993) Phylogenetic frameworks: towards a firmer foundation for the comparative approach. *Biological Journal of the Linnean Society* **49**, 45-61.
4. A. Purvis (1995) A composite estimate of primate phylgoeny *Philosophical Transactions of the Royal Society of London Series B Biological Sciences* **348**, 405-421.
5. M. A. Ragan (1992) Matrix Representation in Reconstructing Phylogenetic- Relationships among the Eukaryotes *Biosystems* **28**, 47-55.
6. C. Semple, M. Steel (2000) A supertree method for rooted trees *Discrete Applied Mathematics* **105**, 147-158.
7. M. Steel (1992) The complexity of reconstructing trees from qualitative characters and subtrees *Journal of Classification* **9**, 91-116.
8. M. Wilkinson, J. L. Thorley, D. T. J. Littlewood, R. A. Bray, in *Interrelationships of the Platyhelminthes* R. A. Bray, Ed. (Tayloer and Francis, London, 2001) pp. 292-301.
9. N. Salamin, T. R. Hodkinson, V. Savolainen (2002) Building supertrees: an empirical assessment using the grass family (Poaceae) *Systematic Biology* **51**, 136-50.
10. V. Daubin, M. Gouy, G. Perriere (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history *Genome Research* **12**, 1080-90.
11. F.-J. Lapointe, G. Cucumel (1997) The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa *Systematic Biology* **46**, 306-312.
12. O. R. Bininda-Emonds, J. L. Gittleman, A. Purvis (1999) Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia) *Biological Reviews of the Cambridge Philosophical Society* **74**, 143-75.
13. K. E. Jones, A. Purvis, A. MacLarnon, O. R. Bininda-Emonds, N. B. Simmons (2002) A phylogenetic supertree of the bats (Mammalia: Chiroptera) *Biololgical Reviews of the Cambridge Philosophical Society* **77**, 223-59.
14. D. Pisani, A. M. Yates, M. C. Langer, M. J. Benton (2002) A genus-level supertree of the Dinosauria *Proceedings of the Royal Society of London Series B Biological Sciences* **269**, 915-21.
15. C. J. Creevey *et al.* (In Press) Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proceedings of the Royal Society of London* **In Press**.

F5. Monte Carlo Estimation and Graphical Analysis of Likelihood Landscapes (of the Population Structure of Shotgun Libraries)

Ben Felts[1], James Nulton[2]. Joe Mahaffy[3]. Peter Salamon[4]. Forest Rohwer[5]. Mya Breitbart[6]. Beltran Rodriguez Brito[7]. David Bangor[8]

Keywords: Monte Carlo, likelihood, population structure, shotgun libraries, contour map

1 Introduction.

Recently developed techniques are successfully producing "shotgun libraries" composed of genomic fragments from multi-genotype populations. The structure of the underlying population (e.g. richness, evenness, etc.) can be estimated from these shotgun libraries. However, it is cost prohibitive to sequence the entire library, so indirect methods based on sequencing only a very small fraction of the library must be relied upon. These methods typically proceed by extracting some statistics from the sequenced fragments (e.g. the number of fragments that overlap any other fragment, or the size and number of strings of contiguous fragments, etc.) and then searching for the population structure with the maximum likelihood of producing those statistics. This likelihood function is often too complex to be computed directly for the candidate population structures, so various simplifications are employed to estimate it, with an accompanying loss of precision and accuracy. Additionally, the search typically only produces the likelihood values necessary to determine the location of the maximum in the space of candidate population structures and some estimate of error for that location.

We have developed a Monte Carlo simulation method to directly estimate the likelihood of any candidate population structure. This method is applied to a grid of points that spans the range of solutions under consideration. The grid of results is then used to produce a contour map, thus applying contouring algorithms to estimate the likelihood function between grid points. The contour map can be analyzed visually as well as computationally to derive valuable information about the potential solutions (e.g. the relative strength of the maximum likelihood solution, or the size and shape of the regions of nearly maximum likelihood solutions, etc.).

[1] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720. E-mail: bfelts@myth.sdsu.edu
[2] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720. E-mail: jnulton@mail.sdsu.edu
[3] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720. E-mail: mahaffy@sciences.sdsu.edu
[4] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720. E-mail: salamon@math.sdsu.edu
[5] Department of Biology, San Diego State University, San Diego, California, 92182-4614. E-mail: forest@sunstroke.sdsu.edu
[6] Department of Biology, San Diego State University, San Diego, California, 92182-4614. E-mail: mya@sunstroke.sdsu.edu
[7] Department of Computational Science, San Diego State University, San Diego, California, 92182-7720. E-mail: brodrigu@rohan.sdsu.edu
[8] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720. E-mail: heimdalle@yahoo.com

References

[1] Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F. and Rohwer, F. 2002. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences USA 99*:14250-14255.

[2] Breitbart, M., Hewson, I., Felts, B., Mahaffy, J., Nulton, J., Salamon, P. and Rohwer, F. 2003. Metagenomic analyses of an uncultured viral community from human feces. *Journal of Bacteriology 85*:6220-6223.

F6. Using local alignment to discern haplotypes from optical maps

Steve Goldstein[1] [2], **Susan Reslewic**[1], **Scott Kohn**[3], **David C. Schwartz**[1]

Keywords: Optical mapping, haplotyping, local alignment

1 High-throughput haplotype discrimination

A central problem in human genomic studies is the development of appropriate experimental and algorithmic systems for high-throughput, whole genome haplotyping [4]. Single molecule approaches are a necessity for direct haplotyping, which requires unambiguous scoring and phasing of markers on the same homolog [4]. Optical mapping is a system for creating whole genome physical maps from ensembles of single DNA molecule ordered restriction maps [7]. These "single-molecule maps" are assembled into contigs whose consensus maps span an organism's entire genome. Our goal is to develop algorithms that can extract human haplotype information from single-molecule optical mapping data.

As a first step, we have applied our approach to a model system: a diploid form of *Saccharomyces cerevisiae*. We created the diploid yeast by crossing two *S. cerevisiae* strains, S96 and YJM789, and isolating the resulting diploid before meiotic recombination. The strains are markedly divergent from one another on phenotypic, sequence, and karyotypic levels [10]. Here we present a method for resolving the two chromosomal copies, one from S96 and one from YJM789, present in diploid optical mapping data sets.

Recent evidence has highlighted the role played by insertions, deletions, and other genomic rearrangements in human variation [2]. Haplotypes characterized by indels and rearrangements may confound global alignment strategies and instead require local alignment algorithms. Our computational method uses local alignment of single-molecule optical maps against a putative consensus genome to obtain a collection of candidate molecules, which are then clustered and assembled to obtain the two haplotypes. With our model system data set, we have demonstrated the ability to discern the two constituent haplotypes and have developed some general strategies for human haplotyping applications.

2 Algorithms for discerning haplotypes

Our algorithmic approach focuses on genomic loci for which the two haplotypes are globally similar but may or may not have large regions of local variation. Of the five yeast chromosomes analyzed here, there are two instances of significantly large regions of local variation. For example, the restriction maps of the two haplotypes for chromosome III match over a range of approximately 150 kilobases but differ for the remaining 150 kb of the chromosome. Other methods that assume global similarity [3] do not apply because they cannot analyze the regions of large local variation.

In our approach, we assume that we have a reference map that is globally representative of the two haplotypes. This reference may represent one of the two haplotypes or be an averaged representation of both. We assume we have a collection of single-molecule optical

[1]Laboratory for Molecular and Computational Genomics, University of Wisconsin, Madison, WI.
[2]E-mail: steveg@lmcg.wisc.edu
[3]OpGen, Inc. Madison, WI

maps from the diploid genome. Our algorithm will partition the optical maps into sets that represent each haploid genome at a particular locus.

A summary of our algorithm is as follows:

1. For each optical map, find the best scoring local alignment against the reference map.
2. For a particular locus, select maps with statistically significant local alignment scores.
3. Partition the set of maps into two or more clusters and create a consensus map from each partition.
4. If desired, repeat steps 1-3 to increase coverage.

The first step of our approach finds local alignments between the reference map and each optical map. These local alignments are candidate anchors at which each haplotype is very similar to the reference map. The portion of the optical map that does not align with the reference is the portion of the haplotype with the local variation.

Local alignment algorithms have been well-studied in the literature for both sequence comparisons [8] and restriction map comparisons [5]. Our scoring function models the system errors in the optical mapping process [1]. We calculate statistical significance (p-values) for the local alignment scores by fitting scores from randomly generated optical maps to an extreme value distribution [9]. To guard against "false positives" we selected only those maps that do not have statistically significant alignments elsewhere in the genome. We will need a more sophisticated filtering scheme for repetitive genomes. We currently partition the maps at a locus using a variety of ad-hoc approaches but are investigating automatic clustering methods.

We tested our algorithm on the diploid yeast chromosomes I, III, VI, VIII, and IX, using the sequence of the S288C strain to derive the initial reference map. For all chromosomes, the method identified molecules belonging to both homologs in roughly the same proportion. For some of the chromosomes, iteration was not necessary: a single step brought equal representation of the homologs. Others required two or three steps. The clusters were validated by comparison with the consensus maps obtained from the S96 and YJM789 haploid strains.

References

[1] Antoniotti, M., Anantharaman, T., Paxia,S. and Mishra,B. 2001. Genomics via Optical Mapping IV. *Technical Report CIMS-TR-811.* New York: NYU Courant Bioinformatics Group.

[2] Batzer, M. and Deininger, P. 2002. Alu Repeats and Human Genomic Diversity. *Nature Reviews: Genetics,* 3:370-380.

[3] Casey, W. and Mishra, B. 2003. A nearly linear-time general algorithm for bi-allele haplotype phasing. In: Lecture Notes in Computer Science, LNCS 2913. New York: Springer-Verlag.

[4] Gabriel, S., Schaffner, S., Nguyen, H., et al. 2002. The structure of haplotype blocks in the human genome. *Science,* 296: 2225-2229.

[5] Huang, X. and Waterman, M. 1992. Dynamic Programming Algorithms for Restriction Map Comparison. *CABIOS,* 8:511-520.

[6] Mitra, R., Butty, V., Shendure, J., et al. 2003. Digital genotyping and haplotyping with polymerase colonies. *Proceedings of the National Academy of Sciences USA,* 100:5926-5931.

[7] Schwartz, D., Li, X., Hernandez, L. et al. 1993. Ordered Restriction Maps of Saccharomyces cerevisiae Chromosomes Constructed by Optical Mapping. *Science* 262:110-114.

[8] Waterman, M. 1995 *Introduction to Computational Biology: Maps, Sequences, and Genomes.* Boca Raton: Chapman and Hall/CRC.

[9] Waterman, M and Vingron, M. 1994. Sequence Comparison Significance and Poisson Approximation. *Statistical Sciences,* 9:367-381.

[10] Winzeler, E., Richards, D., Conway, A., et al. 1998. Direct Allelic Variation Scanning of the Yeast Genome. *Science,* 821:1194-1197.

F7. Providing an automatically derived high quality immunoglobulin V gene sequence database

Ida Retter[1], Werner Müller[2]

Keywords: sequence alignment, secondary database, automatic generation, immunoglobulins

1 Introduction.

With the exponential growth of the primary nucleotide sequence databases GenBank, DDBJ and EMBL [1] the requirement of automatically generated and annotated secondary sequence databases arises. Our aim is to generate an immunoglobulin nucleotide sequence database derived from the EMBL database in an automatic approach, representing the immunological sequence spectrum present in the germ-line. Within an immunoglobulin gene locus there are different sequence configurations possible: The germ-line configuration, in which multiple gene elements, mainly the so-called V genes, constitute the potential diversity of the antibody molecule, and the rearranged configuration, in which one V gene has been recombined with one or two other gene segments to create a functional antibody coding sequence (for review, see [2]). These rearrangements may or may not include somatic point mutations and deletions. To separate the germ-line encoded from somatically mutated immunoglobulin sequences we use two strategies: On the one hand, we compare V gene sequences from the EMBL database with genomic BAC sequences and regard a 100% match as a germ-line evidence. On the other hand, all rearrangements from the EMBL database are aligned and V genes that are found in at least two independent rearrangements are regarded as germ-line sequences. To maximize data reliability our database does not include information from the annotation part of the EMBL nucleotide entries but provides accurate sequence evaluation by sequence comparison. The program was developed with sequences from the murine immunoglobulin heavy chain locus. However, it can also be applied to the light chain loci, other types of sequences (e.g., T cell receptor genes) and other species.

2 Methods.

Sequence alignment with BLAST. In order to identify all immunoglobulin sequences within the EMBL database we use the BLAST algorithm [3]. Beside the standard subset the High Throughput Genomic Sequence (HTG) and the WGS (Whole Genome Shotgun) databases are included in the search. A number of known immunoglobulin sequences are used as initial query sequences [4]. The BLAST result is subsequently filtered for minimum sequence identity and minimum alignment length.

Sequence alignment with DNAPLOT. DNAPLOT is a sequence alignment program tailored for immunoglobulin nucleotide and protein sequences [5]. The fast alignment algorithm allows sorting the sequences within a multiple alignment and comparing multiple alignments among each other. Furthermore, the DNAPLOT motif recognition functions are used for the automatic V gene annotation.

[1] Department of Experimental Immunology, German Research Centre for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany, Email: ida.retter@gbf.de

[2]Department of Experimental Immunology, German Research Centre for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany, Email: wmueller@gbf.de

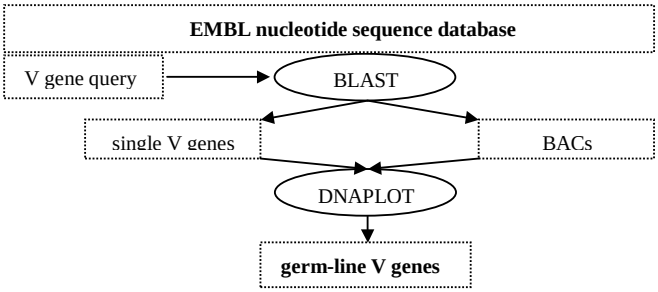


Figure 1: Extraction of V gene sequences from the EMBL database by the BLAST program and germ-line V gene selection by the DNAPLOT program.

3 Results and discussion.

In our database we can classify the V genes into three different quality groups. The first group consists of sequences found in a rearranged form as well as in a non-rearranged configuration. These are germ-line sequences of actively used V gene segments. V gene sequences of the second group are only found as germ-line genes but not recovered in V gene rearrangements. Such V gene segments may represent pseudogenes and the reason for the non-functionality of such sequences might be determined

by a subsequent analysis. The third group of V genes is only found in rearranged sequence list but not in the list of non-rearranged sequences. These V genes represent most likely germ-line genes. However, a little bit of uncertainty remains until in a future generation of the database, a non rearranged counterpart can be recovered from the EMBL nucleotide database.

Due to the automatic generation our sequence data set can be updated any time. It is comprehensive as it takes all published immunological sequences into account. The addition of new entries into the EMBL database will continuously improve the resulting V gene database. In turn, our database provides an important tool for the annotation of genomic sequences of the mouse. The method can be easily adapted to other variable loci, thereby providing the opportunity to analyse species with poor sequence data availability.

4 References and Websites.

[3] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402.

[2] Honjo, T. and Alt, F.[ed.] 1995. *Immunoglobulin Genes.* London: Academic Press.

[5] http://www.dnaplot.org

[1] Kulikova T. et al. 2004. The EMBL nucleotide sequence database. *Nucleic Acids Research* 32:D27-30.

[4] Lefranc, M.P., Giudicelli,V., Ginestoux,C., Bodmer,J., Muller,W., Bontrop,R., Lemaitre,M., Malik,A., Barbie,V. and Chaume,D. 1999. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research* 27:209-212.

F8. A Novel Method to speed up Multiple-Use PCR Primer Design

Yu-Cheng Huang[1], Huai-Kuang Tsai[1], Han-Yu Chuang[1], Chun-Fan Chang[2], and Cheng-Yan Kao[1]

Keywords: multiple-use PCR primer design, partial order graph (POG), polymerase chain reaction (PCR), melting temperature

1 Introduction.

This work proposes a new method to speed up the multiple-use PCR primer design procedure for polymerase chain reaction (PCR). The development of PCR has revolutionized genetic analysis and engineering science. The PCR technique is basically a primer extension reaction for amplifying specific nucleic acids in vitro. Designing proper primers for multiple sequences is a knotty task [2]. The proposed method reduces the computing time of multiple-use PCR primer design by representing multiple DNA sequences in a new data structure, named partial order graph (POG). POG preserves all information of sequences and reduces computing time. The melting temperature of primer candidates is determined by nearest-neighbor thermodynamic approximation. To determine the melting temperature of primer candidates, a lookup table is built to accelerate the calculating process by recording pairs of neighbors appeared in the sequences. The proposed method is applied on four datasets which is selected from the UniGene database with different species. It efficiently reduces the number of necessary computed pairs of neighbor to only 0.62% of the original size. The efficient multiple-use PCR primer design algorithm is useful in many aspects of genomic researches including unique primer design, minimal primer set design, probe design, and so forth.

2 System and method.

The proposed method works as follows. Assume we would like to find the primers of n DNA sequences. These sequences will be transferred into a compact partial order graph (POG). Fig.1(a) shows an example of merging three sequences into a POG. In the POG, letters are represented as nodes and directed edges drawn between consecutive letters in each sequence. Identical letters of sequences are merged if they appear at the same position. Primer candidates are then generated by shifting the windows whose size is predefined. A primer candidate is considered to be good if the melting temperature (T_m) is high enough to a particular region in a sequence but low in all the other sequences. We calculate T_m using a thermodynamic data [3] with the following formula [1]:

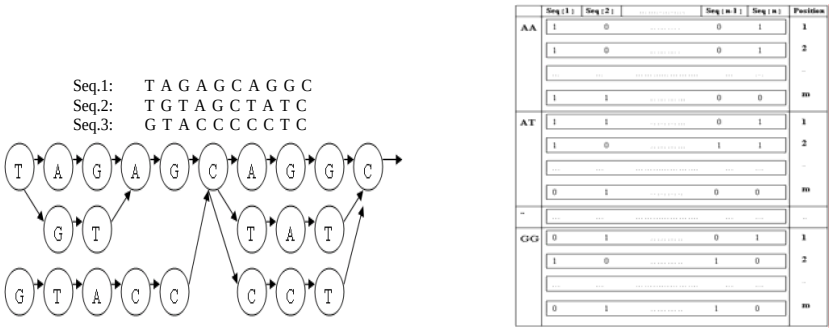
$$T_m = -273.15\frac{\Delta H}{\Delta S + R\ln(Ct/4)} + 16.6\log[Na+],$$

where ΔH is the sum of the nearest-neighbor enthalpy changes for hybrids; ΔS is the sum of the nearest-neighbor entropy changes for hybrids; R is the Gas Constant (1.987 cal/°C/mol); C_t is the molar concentration of strands; and [Na+] is the monovalent ion concentration.

However the costs of calculating pair of neighbors are time consuming. To compute T_m efficiently, a lookup table of all pair of neighbors in those sequences is created. If an edge (i.e., a pair of neighbor) appears in the sequence, the corresponding entry will be set to 1. Fig. 1(b) shows the lookup table.

[1] Bioinfo Lab., Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan. E-mail: {r91021, d7526010, r90002, cykao}@csie.ntu.edu.tw

[2] Chinese Culture University, Taiwan. E-mail: chunfan@faculty.pccu.edu.tw



(a) (b)

Fig. 1: (a) shows an example of how three sequences are merged into a POG; (b) shows a lookup table of all pair of neighbors in those sequences. The table is used to accelerate the process of calculating t_m for primer candidates.

3 Result.

The proposed method is applied on four genome datasets to verify the robustness. Four genome datasets, including *Homo sapiens* (*Hs*), *Mus musculus* (*Mm*), *Rattus norvegicus* (*Rn*), and *Hordeum vulgare* (*Hv*), are selected from the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene/). We cut the long sequence of species into smaller segments (3,000 bp) and merge them into a more compact partial order graph. Table 1 shows the results on the proposed methods tested on these datasets with a PC (800MHz *Intel Celeron* and 512 Mbytes of RAM). Among these experiments, the proposed method can significantly reduce the number of neighbors which are necessary in calculating T_m. The reduce rates are at least 99% in these tests. Although the building time is somewhat long, once the POG is built, we can easily add more target sequences in a short time without destroying the established POG.

Table 1: Experimental results of the proposed methods applying on four different datasets.

	Hordeum vulgare	*Rattus norvegicus*	*Mus musculus*	*Homo sapiens*
size of UniGene database	11.1 MBytes	64.7 MBytes	117 MBytes	137 MBytes
original number of sequences	11491	53354	88185	108094
total edges should be calculated	8,406,053	46,518,085	90,434,462	102,295,866
building time(sec.)	208.8	1120.2	2150.7	2489.8
edges should be calculated in POG	43,786	183,201	403,385	634,711
ratio of calculating edges	0.52 %	0.39 %	0.45 %	0.62 %

References

[1] Breslauer, K. J., Frank, R., Blöcker, H., and Marky, L. A. 1986. Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences USA*, vol. **83**, pp. 3746-3750.

[2] Hsieh, M.-H., Hsu, W.-C., Chiu, S.-K., and Tzeng, C,-M. 2003. An efficient algorithm for minimal primer set selection, *Bioinformatics*, vol. **19**, no 2, pp. 285-286.

[3] Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamuta, H., Ohmichi, T., Yoneyama, M., and Sasaki, M. 1995. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*, vol. **34**, pp. 11211-11216.

F9. Building a Laboratory Information Management System for FP-TDI Genotyping Research

Daniel C. Koboldt[1], Pui-Yan Kwok[2], and Raymond D. Miller[1]

Keywords: SNP, FP-TDI, Laboratory Information Management System, LIMS, HapMap Project, open source, XML, MySQL, Perl, assay design, quality control

1 Introduction

Single-nucleotide polymorphisms (SNPs) can change the function or the regulation of a protein. They also are useful as genetic markers that can be used to find the actual DNA sequence variants that cause differences in gene function or regulation; many such differences directly contribute to disease processes.[1] The ability to genotype SNPs rapidly and cost-effectively is essential for many applications such as mutation detection and linkage mapping.[2]

Our laboratory uses a genotyping method for SNPs that combines the specificity of nucleotide incorporation by DNA polymerase and the sensitivity of fluorescence polarization, known as Template-directed Dye Terminator Incorporation assay with Fluorescent Polarization detection (FP-TDI). Following PCR amplification of the target region, a SNP-specific primer anneals immediately upstream of the polymorphic site in the target DNA. Appropriate dye-labeled terminators extend the SNP-specific primer by one base; by determining which terminator is incorporated, the allele present in the target DNA can be inferred. The highly specific and sensitive nature of FP-TDI allows us to perform high-throughput SNP genotyping in a 384-well plate format, generating about 10,000 genotypes per day.[3]

The development of a laboratory information system became especially critical with our participation in the International HapMap Project, the goal of which is to develop a haplotype map of the human genome. The freely-available information produced by the Project is expected to be an important resource for researchers who want to find genes related to health, drug response, and disease. As a genotyping center for the HapMap project, our laboratory must have the capability to receive large incremental SNP allocations from a data control center (DCC) over a thousand miles away. Our FP-TDI machines output data in a raw format which has to be processed, organized, put through quality control, and then submitted back to the center.

2 Data Requirements

Participation in the HapMap project requires a data pipeline that begins at SNP allocation. On a regular basis the data control center sends us massive compressed XML files that contain the sequence and alleles for each SNP that has been found in our region, chromosome 7p. We must parse, sort, and store all of that information. With each new allocation, we must determine the SNPs for which we can successfully design assays, a computationally-intensive process that takes advantage of freely available databases including dbSNP and RepBase[4]. We select from among assayable SNPs those which meet our criteria for genotyping. When technicians run the assays, a

[1] Washington University School of Medicine, Campus Box 8123, St. Louis, Missouri 63110 USA. E-mail: dkoboldt@psts.wustl.edu

[2] University of California, San Francisco, 505 Parnassus Ave, Long 1332A, Box 0130, San Francisco, CA 94143-0130 USA.
E-mail: kwok@cvrimail.ucsf.edu

raw text file is the only output. Those raw files must be processed so that our quality control team can view the results in PerkinElmer's SNPScorer software. An interface is necessary for the team to view call rates and update QC assessments for each assay without making direct changes to the database records. In preparation for submission, called data from our laboratory must be combined with that from collaborators located in San Francisco. Finally, the data must be exported into a precise XML format and delivered to the data control center.

3 LIMS Solution

We felt that an open-source platform offered the flexibility and capabilities that best suited the vast quantities of data in our pipeline. A customized UNIX environment provides the backbone for our data processing. Three relational MySQL databases store our laboratory data in every stage of the data pipeline; they also provide the storage and organization capabilities required for complex informatics tasks including assay design, SNP choosing, primer orders, data storage, analysis, quality control, and data submission. Once DCC files are downloaded and decompressed, Perl scripts parse out the requisite SNP information and update our database. In the ensuing assay design pipeline, the most ideal PCR and SNP-specific primers are selected and ranked for every SNP possible. A mapping program written in Perl selects the SNPs in our region that have assays, have not been ordered, and meet the distribution requirements set by the HapMap Project Steering Committee. Additional software generates the order files for mapped SNPs and parses the raw output files into our database once the assays have been run. Our intranet provides web interfaces built in Perl CGI that allow our staff to analyze, call, and assess the quality of each assay. Additional web interfaces chart our progress on 7p, help staff members to update plate ordering information, and display calculated assay errors (such as Mendelian failures). Data is exported into XML format for transmission between laboratories and submission to the DCC.

4 SUMMARY

We used open-source technology to build a Laboratory Information Management System (LIMS) to store, manipulate, and share the data produced by our FP-TDI technology.

References

[1] Brooks, Lisa D. 2003. SNPs: Why Do We Care?. *Methods in Molecular Biology, vol. 212: Single Nucleotide Polymorphisms: Methods and Protocols.*, Humana Press Inc., Totowa, NJ, USA 1-14.

[2] Gibson, Greg and Muse, Spencer V. 2002. *A Primer of Genome Science.* Sunderland, Mass.: Sinauer Associates, Inc.

[3] Hsu, Tony M. and Kwok, P.Y. 2003. Homogeneous Primer Extension Assay With Fluorescence Polarization Detection. *Methods in Molecular Biology, vol. 212: Single Nucleotide Polymorphisms: Methods and Protocols.*, Humana Press Inc., Totowa, NJ, USA 177-187.

[4] Vieux, E.F., Kwok, P.Y., and Miller, R.D. 2002. Primer Design for PCR and Sequencing in High-Throughput Analysis of SNPs. *Biotechniques* , USA Suppl:28-30, 32.

F10. Combinatorial chemistry discriminating analysis of complex microbial systems with restricted site tags (RST)

Alexey Kutsenko[1], Veronika Zabarovska[1], Lev Petrenko[1], Tore Midtvedt[2], Ingemar Ernberg[1], Eugene R. Zabarovsky[1]

Keywords: biodiversity, gemone, microorganism, human microflora, combinatorial chemistry

1 Introduction.

The current strategies for mapping and sequencing are clearly not meeting the challenge of high-throughput comparative genomics. Alternative and complementary strategies need to be developed, and it is imperative now to find cost-effective and convenient methods that allow comparative genomics projects to be undertaken by a wide range of laboratories.

We develop a new robust and high efficient technique for large scale scanning of genomes in complex multiorganisms mixture on a quantitative and qualitative basis. This comparative genomic technique are currently effectively applying [1] in the area of sequencing of related bacterial strains and species, in order to identify the genomic basis for differences in their biological properties, particularly pathogenicity. Our approach allows analysis of complex microbial mixtures such as in human gut and identification with high accuracy of a particular bacterial strain on a quantitative and qualitative basis.

2 Method and results.

To achieve the aim of large scale scanning of microbial genomes we propose to create restricted site tags (RST) set of a organism: databases containing specie's short sequences surrounding restriction sites - tags, of a particular rare cutting restriction enzymes. We introduced an information value of a rare cutting restrictases (with restriction site of eight base pairs in lengths and more, Tab. 1, Fig. 1) and analyzed RST passports of all selected restrictases. Thus, a comparison of 1 312 tags from available sequenced *E. coli* genomes, generated with the *Not*I, *Pme*I and *Sbf*I restriction enzymes, revealed only 219 tags that were not unique. None of these tags matched human or rodent sequences.

Restriction enzyme	***Pme*I**	***Sbf*I**	*Pac*I	*Fsp*A	***Not*I**	*Sgf*I	*Sgr*AI	*Srf*I	*Sse*2321	*Asc*I	*Fse*I	*Swa*I
Score	**62.5**	**49.5**	37	35.5	**22.5**	22.5	21.5	16.5	16	14	13	-21

Table 1. Information value of recognition sites for rare cutting restriction enzymes in selected 70 microbial genomes.

Thus, RST set for a particular organism represents in fact a unique genomic fingerprint of this specie or strain that is easy to generate. This distinctive feature of the method allows a

[1] Microbiology and Tumor Biology Center, Karolinska Institute, Stockholm, Sweden. E-mail: Alexey.Kutsenko@mtc.ki.se

[2] Department of Cell and Molecular Biology, Karolinska Institute, Stockholm, Sweden.

discriminating different species and even strains from mixture of genomes. From the total DNA fraction of a complex microbial system a total RST set is produced and sequenced. Then an information about the original genomes mixture is extracting by comparison of the total set with the fingerprints of known bacteria (with known genomes) using combinatorial chemistry. At the same time it was shown that short sequences randomly taken from any bacterial genomes fall in clusters when principal component analysis (PCA) is applied [2]. Due to this fact we can suggest the potential of our method to discriminate even between unknown (unsequenced) species.

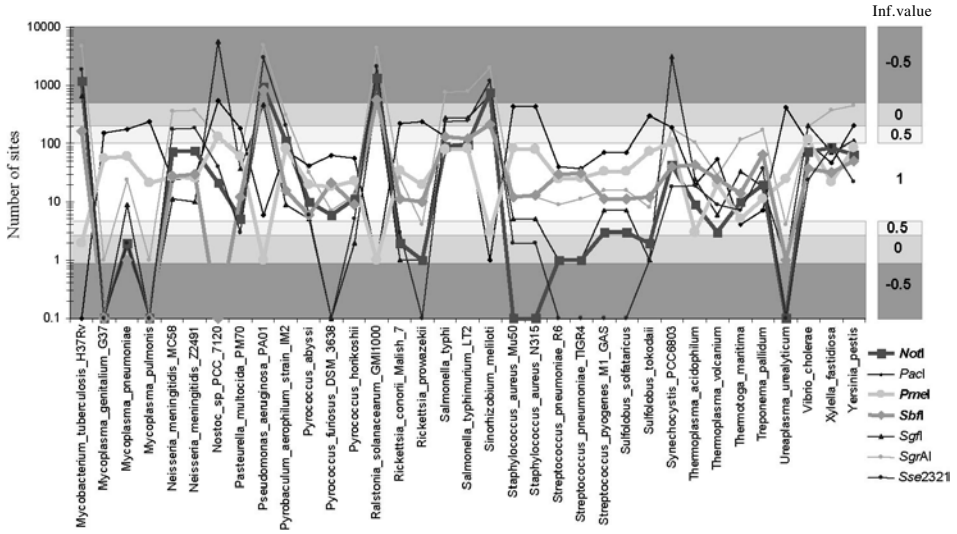


Figure 1. Schematic distribution of recognition sites for rare cutting restriction enzymes in selected completely sequenced bacterial genomes.

The procedure for generating tagged sequences is easily realized experimentally and can be adapted to any rare cutting restriction enzyme. We demonstrated experimentally that the *Not*I tags comprising 19 bp of sequence information could be successfully generated using DNA isolated from intestinal samples. Such *Not*I passports allow the discrimination between closely related bacterial species and even strains. Therefore the approach allows analysis of complex microbial mixtures such as in human gut and identification with high accuracy of a particular bacterial strain on a quantitative and qualitative basis.

The remarkable advantage of our method verified on microbial systems is the ability to identify even strain composition. This gives the opportunity of identifying faint differences between relative organisms, e.g. pathogenic islands.

References.

[1] Zabarovska, V., Kutsenko, A.S. *et al.* 2003 *Not*I passporting to identify species composition of complex microbial systems. *Nucleic Acids Res.*, 31:e5.

[2] Sandberg, R., Winberg, G., Branden, C.-I., Kaske, A., Ernberg, I. and Coster, J. 2001 Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. *Genome Res.*, 11:1404-1409.

F11. Aligning Optical Maps

Yu-Chi Liu [12], Michael S. Waterman [1], Anton Valouev [1], Lei Li [1], Yu Zhang [1], Yi Yang [1], Jong-Hyun Kim [1], David C. Schwartz[3]

Keywords: optical mapping, restriction maps alignment, dynamic programming

1 Introduction.

Optical Mapping is a single molecule system for the construction of ordered DNA restriction maps [4]. It uses light microscopy to directly image individual DNA molecules bound to charged surfaces, cleaved by restriction endonucleases. Cut sites are flagged by small, visible gaps. Notably, restriction fragments retain their original order, and the resulting string of restriction fragment masses is termed an optical map. Shotgun Optical Mapping is an approach that constructs whole-genome DNA restriction maps by overlapping optical maps derived from randomly sheared genomic DNA molecules. Each molecule is mapped and aligned to form contigs consisting of $10-50$x coverage of a given locus; as such, these maps have served as scaffolds for sequence assembly and validation [3].

In the map assembly problem, sensitive detection of overlaps between optical maps plays a crucial role, and experimental errors make the analysis of this problem complex. By modifying the dynamic programming algorithm in [2] for restriction maps alignment with suitable scoring function and parameters, we are able to find accurate matchings between two overlapping optical maps under the limits of known experimental errors.

2 Data.

Due to experimental condition, the measured maps have the following features:

I. *missing cuts*,
which occur when molecules are not cleaved at all restriction sites.
In one data set, the probability of missing cut at a cut site is about 0.2.

II. *false cuts*,
which occur when there are cleavages detected which are not at a restriction site.
We have about 5 false cuts per Mb in one data set.

III. *sizing errors*,
which are imprecisions in measuring the length of restriction fragments.
Suppose l_r is the true fragment length and l_m is what we observed in the data, then $|l_m - l_r|$ is modeled to be proportional to $\sqrt{l_r}$.

3 Methods.

Let a *segment* $[i,k]$ of a map consists sites i through k. Let a *matching pair* between two maps be denoted by $(i,j;k,l)$.

[1]Program of Molecular and Computational Biology, University of Southern California, Los Angeles, California, 90089.
[2]E-mail: ycliu@usc.edu
[3]Laboratory for Molecular and Computational Genomics, Department of Chemistry and Laboratory of Genetics, University of Wisconsin-Madison, Madison, Wisconsin 53706. E-mail: dcschwartz@facstaff.wisc.edu

Suppose we have a global alignment Π between a map A with m sites and the other map B with n sites is a sequence of ordered matching pairs $(i_1, j_1; k_1, l_1)$ $(i_2, j_2; k_2, l_2)$ $\ldots$ $(i_d, j_d; k_d, l_d)$, where $k_t < i_{t+1}$ and $l_t < j_{t+1}$ for each $t < d$. Let q_x denote the position of site x on map A, and r_y the position of site y on map B. With $\lambda \geq 0$ and $\nu \geq 0$, the score of Π is defined as:

$$score(\Pi) = \sum_{t=1}^{d} \sigma(i_t, j_t; k_t, l_t) + l(q_{i_1}, r_{i_1}) + \sum_{t=2}^{d} l((q_{i_t} - q_{k_{t-1}}), (r_{j_t} - r_{l_{t-1}}))$$
$$+ l((q_m - q_{k_d}), (r_n - r_{l_d})) - \lambda \left[m + n - \sum_{t=1}^{d} (k_t - i_t + 1) - \sum_{t=1}^{d} (l_t - j_t + 1) \right],$$

where

$$\sigma(i_t, j_t; k_t, l_t) = \nu \cdot (\#\text{ of matching sites pair in segment } [i_t, j_t; k_t, l_t])$$
$$+ l((q_{k_t} - q_{i_t}), (r_{l_t} - r_{j_t})) - \lambda((k_t - i_t) + (l_t - j_t)).$$

Each pair of matching sites in segment is rewarded by ν in $\sigma(i_t, j_t; k_t, l_t)$. It also takes care of random permutation of restriction sites position. Each site not matched with any others (due to a false cut or a missing cut) is penalized by λ. $l(a, b)$ is the scoring function of length similarity for two segments, one with length a and the other with length b. So the function $l(a, b)$ takes care of the distance discrepancy between a matching pair.

To obtain an appropriate function for $l(a, b)$, we first, using a central limit theorem, model the measurement distribution (conditional on the true fragment length). Then we derive a log likelihood function to use as $l(a, b)$.

This approach, which is for global map alignment, can be modified to yield local alignments and overlap alignment. To obtain an algorithm for overlap alignment, we modify the recursion in [2] as follows.

Initialize both the first row and column as 0 when calculating the maximum score matrix of alignments (which means both the open and end gaps are not penalized [1]). We are able to find the best overlap alignment (with highest score) between two optical maps under the scoring scheme we mentioned above.

References

[1] Huang, X. 1992. A contig assembly program based on sensitive detection of fragment overlaps. *Genomics*, 14:18-25

[2] Huang, X. and Waterman, M.S. 1992. Dynamic programming algorithms for restriction map comparison. *Comp. Appl. Bio. Sci.*, 8:511-520.

[3] Lim, A., Dimalanta, E.T., Potamousis, K.D., Yen, G., Apodoca, J., Tao, C., Lin, J., Qi, R., Skiadas, J., Ramanathan, A., Perna, N.T., Plunkett, G. 3rd., Burland, V., Mau, B., Hackett, J., Blattner, F.R., Anantharaman, T.S., Mishra, B., Schwartz, D.C. 2001. Shotgun optical maps of the whole Escherichia coli O157:H7 genome. *Genome Res.*, 11(9):1584-93.

[4] Schwartz, D.C., Li, X., Hernandez, L.I., Ramnarain, S.P., Huff, E.J., Wang, Y.K. 1993. Ordered restriction maps of Saccharomyces cerevisiae chromosomes constructed by optical mapping. *Science*, 262:110-114.

F12. Unsupervised Learning of Biological Sequences and Its Applications in Genomic DNA Sequence Annotation

Jing Liu, [1] **L. Ridgway Scott,** [2] **John Goldsmith** [3]

Keywords: unsupervised learning, annotation, lexicon

Abstract

We describe an algorithm that can automatically annotate segments of genomic DNA sequences, identifying coding regions, promoter regions, introns and so forth. We do this by using unsupervised learning techniques developed in computational linguistics. The key new element we borrows from unsupervised learning is the concept of a lexicon, a list of building blocks of the words of a language. There are distinct lexicons for different genomic regions, determined by a learning algorithm automatically from training data.

In our algorithm, context features of different domain of the training genomic DNA sequences are extracted into corresponding lexicons, which serve as probabilistic models to detect featured domains from the un-annotated test genomic DNA sequences. A general probabilistic model of genomic sequence structure is built up to combine all the featured domains together and automatically find out the best annotation of the test genomic DNA sequence.

To achieve this task, we first revised Carl de Marcken's unsupervised learning algorithm and implemented the Viterbi training algorithm to process the large warehouse of the genomic DNA sequences. When we applied it to English corpus parse. it was observed that our Viterbi training showed comparable performance and higher efficiency than Carl de Marcken's unsupervised learning algorithm.

Then we tested the Viterbi training algorithm on a real warehouse of protein sequence database like PIR-NREF, containing 892580 entries and up to $2.38 * 10^8$ characters. The result protein lexicon was composed of $2,423,358$ words. The maximum length of word is 360 characters. The time complexity of this experiment was $O(10^{12})$ floating point computation. Our algorithm is efficient in handling large database.

Afterwards, we integrated the Viterbi training algorithms to human promoter region identification from un-annotated human genome sequences. We used the promoter, intron and CDs training dataset from a public Representative Benchmark Data Sets of Human DNA Sequences provided by Berkeley Drosophila Genome Project web site and trained promoter, intron and CDs lexicons respectively. Two classifiers were built based on those three lexicons to identify the promoter regions. The result was comparable with other promoter region identification methods based on the review dataset by Fickett and Hatzigeiogiou paper.

[1]Dept. of Computer Science, University of Chicago. E-mail: jliu@cs.uchicago.edu

[2]Dept. of Computer Science and Mathematics, University of Chicago. E-mail: ridg@cs.uchicago.edu

[3]Dept. of Computer Science and Linguistics, University of Chicago. E-mail: ja-goldsmith@uchicago.edu

Currently, we are building a simplified model of sequence structure, consisting of intergenic region, introns, exons, acceptor sites and donor sites. We are training the intron and exon lexicons and position specific weight matrixes of acceptor and donor sites from human genome training data. Given an unannotated genomic DNA sequence, the intergenic regions, introns, exons, acceptor sites and donor sites are annotated automatically.

In future work, we will extend current simplified model to annotate the complete genomic DNA sequence more specifically. And we will refine current signal and content scoring parts and other parameters to improve the annotation accuracy.

F13. Characterization of Retroid Agents in the Human Genome: An Automated Approach

Marcella A. McClure, Rochelle A. Clinton, Hugh S. Richardson, Vijay A. Raghavan, Crystal M. Hepp, Brad A. Crowther, Angela K. Olsen, Eric F. Donaldson[1] and Aaron R. Juntunen.

Keywords: Retroid agents, retroviruses, retrotransposons, Genome Parsing Suite, human genome

1 Introduction.

Retroid agents are genomes that replicate by reverse transcription of an RNA intermediate. Although once considered to be "junk" DNA, some Retroid agents are implicated in disease via insertional mutagenesis while others have been found to encode proteins essential to mammalian reproduction or to provide regulatory sequences for host cell processes. We have developed new software, the Genome Parsing Suite (**GPS**), to identify and characterize reverse transcriptase (RT) signals in the human genome database (HGD), and to annotate the Retroid Agents that encode them. The **GPS** approach is quite different in concept from **RepeatMasker** [1], a program designed to identify and mask out Retroid Agents in the human genome with consensus DNA for repetitive elements. Recent work conducted on the December, 2001 freeze of the human genome identified 90 L1 LINEs with intact open reading frames (ORF) [2]. The prototype **GPS** provides precise information about the Retroid Agent, including genes present, condition of genes, agent boundaries, location of the agent in the genome etc. The **GPS** utilizes protein rather than nucleotide sequences to screen for the presence of Retroid Agents, thereby providing a deeper query into a genome. Using the **GPS** to analyze 257,278 RT hits retrieved by **BLAST** from the July 2003 freeze of the HGD a total of 95,537 unique RT signals have been identified, and 111 signals are not unambiguously classified to date. As expected, a human LINE sequence pulls out 93% of the unique RT hits. This is less than the reported 500,000 LINEs found by **RepeatMasker**. The estimated number of LINEs in the HGD using **RepeatMasker** does not account for two possibilities regarding small sequence length hits; 1) some may be random, and 2) small hits that are close in proximity actually belonging to the same highly divergent LINE genome. In addition the **RepeatMasker** results include small remnants of untranslated regions (UTRs), while the **GPS** screens for potentially functional Retroid Agents. We have identified 156 LINEs that contain the coding capacity to be potentially functional. All chromosomes except 19 and 21 have full length LINEs without stop codons or frame shifts (table 1). The **GPS** is designed to not only identify highly conserved Retroid agents, but also very distant ancestors. A complete analysis of Retroid information content per genome will also allow the correlation of the position of these agents with higher order genomic feature, e.g., regions of increased gene expression or silencing within CpG islands. The development of this research tool provides the ability to quickly evaluate all Retroid information of a given genome and generate hypotheses regarding the nature of the Retroid Agent landscape in genomes from all three domains of life. By populating the **GPS** with protein sequences representing all known RT genes not only is a complete analysis of the major Retroid Agents present in the any genome feasible, but low frequency RT genes and in some cases “cryptic” retroviral genomes may be discovered.

2 Software and files.

The **GPS** can be populated with any set of phylogenetically distributed protein sequences and the corresponding ordered-series-of-motifs (OSM) representing functional or structurally important amino acids to search, annotate and assess probable function of new members of a protein family in any organismal genome database. Tests were performed to determine which of three external search methods provide the greatest amount of raw data to analyze. **BLAST** [3] with the PAM70 matrix, retrieved more significant RT hits than **WU-BLAST** [4] using the same matrix, or **RepeatMasker** driven by **Cross_Match** [1] using our in-house RT gene libraries. Twenty-one representative RT sequences were used to generate the data presented in table 1. The GPS is designed to initially evulate the RT retrieved hits and then search the surrounding genomic sequences for other genes encoded by Retroid agents. Raw **BLAST** hits are evaluated for the presence of the RT OSM. Unique hits are determined by comparison of: 1) hits in the same reading frame by multiple probes to the same location, and 2) compound hits at the same location from multiple reading frames. In stage one, a hit with a blast score that is at least 10% greater than all others is determined to be the unique hit. All hits meeting this 10% criterion are reevaluated in the next procedure. In stage two, if more than one probe retrieves a hit to the same location, the unique hit is

All at: Montana State University-Bozeman, Dept. of Microbiology and the Center for Computational Biology, 109 Lewis Hall, Bozeman MT 59717. Email: mars@parvati,msu.montana.edu

[1] University of North Carolina, Chapel Hill, NC 27514. Email: eric_donaldson@med.unc.edu

determined by a score of at least 50% greater than all others. Ambiguity arises when the range of scores are all within 50% of one another. The unique hit will then be determined from among these ambiguities at the Retroid agent gene component analysis stage of the GPS. This procedure removes redundancy of retrieved hits due to cross coverage by multiple queries and solves the problem of multiple small hits retrieved by a given query which represents a distantly related RT gene. Failure to account for these small hits results in each hit being scored as a unique hit, thereby overestimating the number of potential RT genes and Retroid genomes within a host genome. Based on the OSM scores for potential RTs, corresponding Retroid agent genome sequences are extracted from the organismal database. The **GPS** then analyzes these potential Retroid agents for the expected genes given the RT classification based on our gene component libraries. In addition, each potential Retroid agent sequence will be evaluated by all component libraries to screen for recombination events. Future extensions to the **GPS** will include the ability to populate the system with nucleotide sequences using our in-house Retroid nucleotide libraries and a graphical display of all results.

3 Figures and tables.

Chr	Unique	Unclass	Full	Perfect
1	6722	4	127	17
2	7793	4	167	12
3	6858	5	133	11
4	7136	14	154	10
5	6361	7	139	12
6	5793	8	122	10
7	4944	3	86	9
8	4817	2	109	7
9	3842	3	71	4
10	4029	2	74	6
11	4486	9	95	9
12	3890	6	83	6
13	3167	2	42	3
14	2850	3	51	3
15	2520	5	35	5
16	1624	2	26	7
17	1458	1	19	1
18	2448	1	46	5
19	1222	6	18	0
20	1374	3	26	3
21	986	1	8	0
22	745	3	7	2
x	9091	4	187	13
y	1381	13	23	1
Totals	95537	111	1848	156

Table 1. Chr is chromosome number. Unique indicates all unique RT signals, Unclass includes all RTs that could not be unambiguously classified by query RTs. Full indicates full length LINEs and Perfect indicates LINEs that are full length and contain no frame-shifts or stop codons.

4 References.

[1] Smit, AFA & Green, P RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html
[2] Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., Farley, A. H., Moran, J. V. & Kazazian, H. H., Jr. (2003)*Proc Natl Acad Sci U S A* 100, 5280-5.
[3] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997)*Nucleic Acids Res* 25, 3389-402.
[4] Gish, W. 1996-2003 http://blast.wustl.edu

F14. MGAW : a Microbial Genome Annotation Workbench under Web-based Analysis Interface

Hwajung Seo[1], Hyeweon Nam[1], Daesang Lee[1,2], Hongseok Tae[1], Kiejung Park[1]

Keywords: annotation, genome, ontology, COG, viewer, browser

1 Introduction

As the genome projects have produced tremendous bioinformatic data, annotation is considered as an essential part of genome sequencing projects to elucidate the value of the sequence[1]. Through annotation systems, molecular biologists could deal with genome data easily and all related information could be accessed, edited, or updated without additional efforts. To satisfy these needs, we have developed MGAW(Microbial Genome Annotation Workbench), a web-based microbial genome annotation system, which provides web-based analysis interface for gene prediction, homology search, promoter analysis, motif analysis and gene ontology analysis. The annotated information can be retrieved with database searching and browsed with a genome map browser and a gene classification viewer. Public microbial genome databases are imported and can be searched and browsed through the same interface.

2 Methods

MGAW database contains from contig data to functional analysis data of the final gene set. The interface of each annotation tool was implemented not only for running each tool and viewing the result but also for monitoring the progress. Analysis results are saved in the database with primary keys indicating the relationship between data.

Gene prediction in a genome project is the first step of annotation. General methods for gene prediction are applied in this system and a few analysis options are provided. For promoter analysis, we implemented a general promoter pattern search and a two-component analysis search against all the predicted genes. For motif analysis, the Prosite DB patterns are searched against all protein sequences which are translated from the gene prediction. Fast algorithms were developed to accomplish fast searching for motif patterns of regular expression. The progress/status of promoter and motif analysis can be monitored through web interface. For homology analysis of all the predicted genes, we implemented the interface for NCBI BLAST and both COG(Clusters of Orthologous Groups) and GO(Gene Ontology) databases were used to classify the homology search result[4].

Database searching module was implemented to query for the annotation results of an in-progress or finished genome project and a linear map browser was implemented to visualize the whole genome map and detailed annotation information for each selected gene by further clicking. A gene classification viewer was implemented to show gene ontology analysis result with COG for a whole genome. A circular map is generated after retrieving gene ontology information of all the

[1] Information and Technology Institute, SmallSoft Co., Ltd. Junmin-Dong 461-71, Yusung-Gu, Daejeon, 305-811, South Korea. E-mail: {hjseo, hwnam, dslee, hstae, kjpark}@smallsoft.co.kr

[2] Dept. of Biological Science, KAIST, Kusung-dong, Yusung-gu, Daejeon, 305-701, South Korea. E-mail: dslee@bioneer.kaist.ac.kr

genes of a genome and calculating a few features for the whole genome area. A few options were implemented to select a specified category, a region and a drawing mode

For public microbial genome data, input programs were implemented to parse the genome data of GenBank format and import into MGAW databases. Each imported genome can be searched and browsed as an annotated genome can be.

3 Results and Discussions

MGAW has stepwise and intuitive interface as shown in Figure 1. We have tested and improved through a few microbial genome projects.

Figure 1: Overview of MGAW features : annotation interface, genome browsing, gene ontology viewing

More features will be added in the near future, including gene ontology viewing for GO, genome alignment, and a lot of comparative genome analysis modules. MGAW will be very helpful not only for analysis of public microbial genome annotation but also as a practical information system of genomics/comparative genomics for real genome projects.

4 References

[1] Lee, D., Seo, H., Tae, H., Kim, Nam, H. and Park, K. 2003. Development of a web-based genome annotation system. *Research in Computational Molecular Biology (RECOMB 2003).*.

[2] Seo, H., Kim, K.-B., Tae, H., Park, W., and Park, K. 2002, Development of gene ontology analysis and classification tools for microbial genome annotation. *Research in Computational Molecular Biology (RECOMB 2002).*

[3] Tae, H., Seo, H., Nam, H., Lee, D. and Park, K. 2003. Development of a web-based genome annotation system and two analysis tools *Intelligence system for Molecular Biology (ISMB 2003).*

[4] Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V. 2001. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research.* 29:22-28.

F15. A Bioinformatics Approach Toward Identification of Genes involved in Hematopoiesis and Leukemia

Twyla T. Pohar[1], Hao Sun[2], Sandya Liyanarachchi[3], S. James S. Stapleton[4] and Ramana V. Davuluri[5]

Keywords: database, data-mining, EST, hematopoiesis, leukemia, statistical analysis, gene expression.

1 Introduction.

Hematopoiesis describes the process of the normal formation and development of blood cells, involving both proliferation and differentiation from stem cells. Abnormalities in this developmental program yield blood cell diseases, such as leukemia. This complex, biological process is under stringent control, with various extra- and intra-cellular stimuli that result in the activation of downstream signaling cascades. Ultimately, these signaling cascades converge at the level of gene expression where positive and negative modulators of transcription delineate the pattern of gene expression [1].

In order to elucidate the mechanisms involved in such regulation, it is important to identify the genes, which play integral roles in the hematopoietic process. Our in-house developed Hematopoiesis Promoter Database, HemoPDB, includes experimentally defined genes, which were manually curated from published literature [2]. Characterization of the gene expression profiles of these genes may allow us to efficiently identify previously unannotated genes via a combination of statistical analysis and data-mining of expressed sequence databases. We have designed a data-mining pipeline in order to manage the data obtained from dbEST [3] to determine the UniGene [4] clusters likely representing genes expressed preferentially in hematopoietic tissues and organs.

2 Results.

We downloaded all of the ESTs via dbEST: ftp://ncbi.nlm.nih.gov/repository/dbEST to create a standardized gene expression database for each sequence according to organ/tissue. Our automated pipeline then determines the genomic coordinates of each sequence by its alignment to the genome by BLAT [5]. These coordinates are then associated with the corresponding UniGene cluster. We then perform a genome-wide statistical analysis by applying a binomial test to compare the proportion of organ/tissue-specific ESTs expressed in each cluster vs. the entire EST population. The imposed criterion of a p-value, which conforms to a specific level of statistical significance, allows us to determine the clusters that are hematopoietic-related. We intend to substantiate these data obtained

[1] Div. of Human Cancer Genetics, Dept. of Mol. Virology, Immunol. and Med. Genetics, 420 West 12th Ave. Room 570A, Columbus, Ohio, USA. E-mail: pohar-2@medctr.osu.edu

[2] Div. of Human Cancer Genetics, Dept. of Mol. Virology, Immunol. and Med. Genetics, 420 West 12th Ave. Room 570B, Columbus, Ohio, USA. E-mail: sun.143@osu.edu

[3] Div. of Human Cancer Genetics, Dept. of Mol. Virology, Immunol. and Med. Genetics, 420 West 12th Ave. Room 570B, Columbus, Ohio, USA. E-mail: liyanarachchi-1@medctr.osu.edu

[4] Div. of Human Cancer Genetics, Dept. of Mol. Virology, Immunol. and Med. Genetics, Dept. of Physics, 420 West 12th Ave. Room 570, Columbus, Ohio, USA. E-mail: stapleton.41@osu.edu

[5]Div. of Human Cancer Genetics, Dept. of Mol. Virology, Immunol. and Med. Genetics, 420 West 12th Ave. Room 524, Columbus, Ohio, USA. E-mail: davuluri-1@medctr.osu.edu

by comparing them to the gene expression profiles of experimentally characterized hematopoietic genes.

We will present the development of our automated pipeline, in addition to results obtained from this genome-scale analysis. We are hopeful that this study will allow the identification of key genes in the normal and malignant hematopoietic environments.

3 Figures.

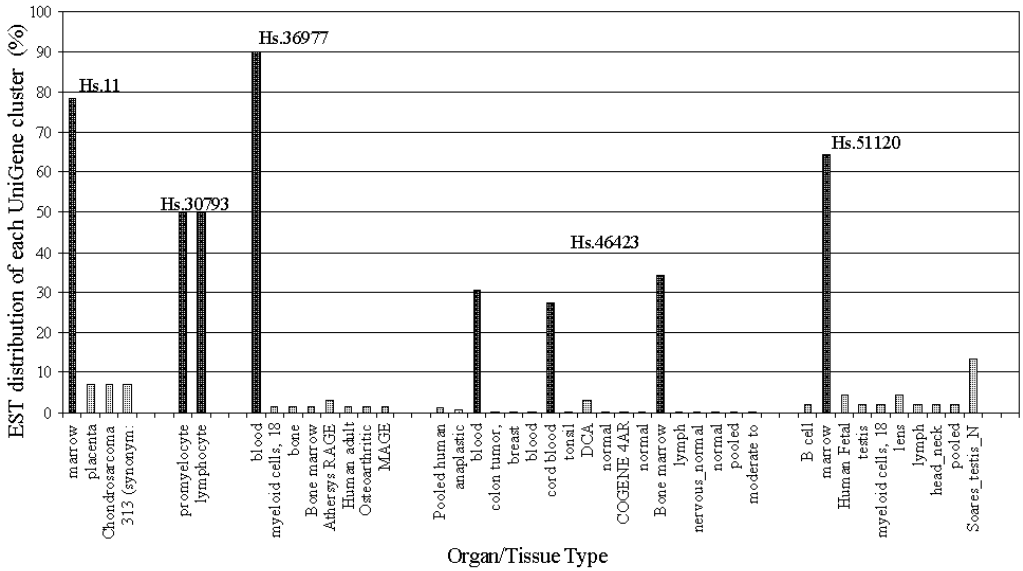


Figure 1. Sample EST percentage distribution of 5 UniGene clusters obtained via our automated method, which identifies hematopoietic-related genes. The highest percentages, conveyed by black bars, are consistent with hematopoietic organ/tissue types.

4 References.

[1] Barreda, D.R. and Belosevic M. 2001. Transcriptional regulation of hemopoiesis. *Developmental and Comparative Immunology* 25: 763-789.

[3] Boguski MS, Lowe TM, Tolstoshev CM. dbEST--database for "expressed sequence tags". 1993. *Nat Genet.* 4(4): 332-3.

[5] Kent, W.J. and Brumbaugh, H. 2002. BLAT--the BLAST-like alignment tool. *Genome Res.* 12: 656-664.

[2] Pohar, TT, Sun, H. and Davuluri, RV. 2004. HemoPDB: An information resource of transcriptional regulation in blood cell development. *Nuceleic Acids Research* 32: D86-D90.

[4] Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. and Wagner, L. 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31: 28-33.

F16. The complete genome sequence of *Rickettsia typhi* and comparison with other rickettsial genomes

Xiang Qin[1], Michael P. McLeod[1,5], Sandor E. Karpathy[1,5], Jason Gioia[2] , Sarah K. Highlander[2], George E. Fox[4], Thomas Z. McNeill[1,4], Huaiyang Jiang[1], Donna Muzny[1], Leni S. Jacob[1], Alicia C. Hawes[1], Erica Sodergren[1], Anita G. Amin[1], Rachel Gill[1], Jennifer Hume[1], Maggie Morgan[1], Guangwei Fan[1], Richard A. Gibbs[1], Chao Hong[3], Xue-jie Yu[3], David H. Walker[3], George M. Weinstock[1,*]

Keywords: genome comparison, annotation, sequencing, rickettsial, *Rickettsia typhi*, pseudogenes

The complete genome sequence (1,111,496 bp) of *Rickettsia typhi*, an obligate intracellular pathogen and the causative agent of murine typhus, was determined and compared with the two published rickettsial genome sequences (*R. prowazekii* and *R. conorii*). A total of 877 genes were identified in *R. typhi* genome, including 3 rRNAs, 33 tRNAs, 3 ncRNAs, 838 proteins, 3 of which are frameshifts, and over 40 pseudogenes, which include the entire cytochrome c oxidase system. *R. typhi*, *R. prowazekii* and *R. conorii* share 775 genes: 23 are found only in *R. prowazekii* and *R. typhi*; 15 are found only in *R. conorii* and *R. typhi*; and 23 are unique to *R. typhi*. While most of the genes are collinear among these three genomes, there is a 124 kb inversion in *R. typhi,* when compared to *R. prowazekii* and *R. conorii*. In addition, a 35 kb inversion close to the replication terminus was also found in *R. typhi* and *R. prowazekii* when compared to *R. conorii*. Inversions in this region are also seen in the unpublished genome sequences of *R. sibirica* and *R. rickettsii* indicating this region is a hot spot for rearrangements. Genome comparisons also revealed a 12 kb insertion in the *R. prowazekii* genome, relative to *R. typhi* and *R. conorii*, which appears to have occurred after the typhus (*R. prowazekii* & *R. typhi*) and spotted fever (*R. conorii*) groups diverged. Repeat sequence analyses of *R. typhi* genome revealed that, like *R. prowazekii* and *R. conorii*, this genome contains very few repetitive sequences, which may explain the similarity in gene orders among the rickettsial genomes. The three-way comparison allowed further *in-silico* analysis of the SpoT split genes leading us to propose that the stringent response system is still functional in these rickettsiae.

[1] Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, 77030
[2] Baylor College of Medicine, Dept. of Molecular Virology and Microbiology, Houston, TX, 77030
[3] University of Texas Medical Branch, Dept. of Pathology, Galveston, Texas 77555
[4] University of Houston, Dept. of Biology and Biochemistry, Houston, TX 77204-5001
[5] University of Texas Health Science Center at Houston, Houston, TX 77030

F17. Providing an automatically derived high quality immunoglobulin V gene sequence database

Ida Retter[1], Werner Müller[2]

Keywords: sequence alignment, secondary database, automatic generation, immunoglobulins

1 Introduction.

With the exponential growth of the primary nucleotide sequence databases GenBank, DDBJ and EMBL [1] the requirement of automatically generated and annotated secondary sequence databases arises. Our aim is to generate an immunoglobulin nucleotide sequence database derived from the EMBL database in an automatic approach, representing the immunological sequence spectrum present in the germ-line. Within an immunoglobulin gene locus there are different sequence configurations possible: The germ-line configuration, in which multiple gene elements, mainly the so-called V genes, constitute the potential diversity of the antibody molecule, and the rearranged configuration, in which one V gene has been recombined with one or two other gene segments to create a functional antibody coding sequence (for review, see [2]). These rearrangements may or may not include somatic point mutations and deletions. To separate the germ-line encoded from somatically mutated immunoglobulin sequences we use two strategies: On the one hand, we compare V gene sequences from the EMBL database with genomic BAC sequences and regard a 100% match as a germ-line evidence. On the other hand, all rearrangements from the EMBL database are aligned and V genes that are found in at least two independent rearrangements are regarded as germ-line sequences. To maximize data reliability our database does not include information from the annotation part of the EMBL nucleotide entries but provides accurate sequence evaluation by sequence comparison. The program was developed with sequences from the murine immunoglobulin heavy chain locus. However, it can also be applied to the light chain loci, other types of sequences (e.g., T cell receptor genes) and other species.

2 Methods.

Sequence alignment with BLAST. In order to identify all immunoglobulin sequences within the EMBL database we use the BLAST algorithm [3]. Beside the standard subset the High Throughput Genomic Sequence (HTG) and the WGS (Whole Genome Shotgun) databases are included in the search. A number of known immunoglobulin sequences are used as initial query sequences [4]. The BLAST result is subsequently filtered for minimum sequence identity and minimum alignment length.

Sequence alignment with DNAPLOT. DNAPLOT is a sequence alignment program tailored for immunoglobulin nucleotide and protein sequences [5]. The fast alignment algorithm allows sorting the sequences within a multiple alignment and comparing multiple alignments among each other. Furthermore, the DNAPLOT motif recognition functions are used for the automatic V gene annotation.

[1] Department of Experimental Immunology, German Research Centre for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany, Email: ida.retter@gbf.de

[2]Department of Experimental Immunology, German Research Centre for Biotechnology, Mascheroder Weg 1, 38124 Braunschweig, Germany, Email: wmueller@gbf.de

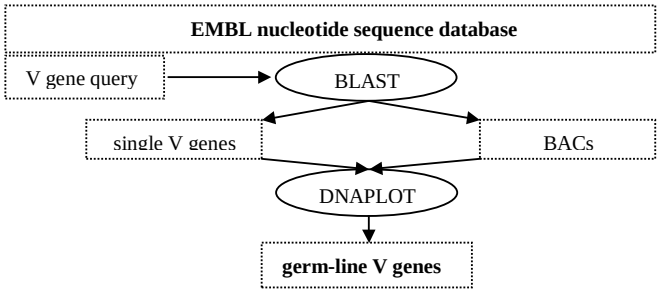


Figure 1: Extraction of V gene sequences from the EMBL database by the BLAST program and germ-line V gene selection by the DNAPLOT program.

3 Results and discussion.

In our database we can classify the V genes into three different quality groups. The first group consists of sequences found in a rearranged form as well as in a non-rearranged configuration. These are germ-line sequences of actively used V gene segments. V gene sequences of the second group are only found as germ-line genes but not recovered in V gene rearrangements. Such V gene segments may represent pseudogenes and the reason for the non-functionality of such sequences might be determined

by a subsequent analysis. The third group of V genes is only found in rearranged sequence list but not in the list of non-rearranged sequences. These V genes represent most likely germ-line genes. However, a little bit of uncertainty remains until in a future generation of the database, a non rearranged counterpart can be recovered from the EMBL nucleotide database.

Due to the automatic generation our sequence data set can be updated any time. It is comprehensive as it takes all published immunological sequences into account. The addition of new entries into the EMBL database will continuously improve the resulting V gene database. In turn, our database provides an important tool for the annotation of genomic sequences of the mouse. The method can be easily adapted to other variable loci, thereby providing the opportunity to analyse species with poor sequence data availability.

4 References and Websites.

[3] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402.

[2] Honjo, T. and Alt, F.[ed.] 1995. *Immunoglobulin Genes.* London: Academic Press.

[5] http://www.dnaplot.org

[1] Kulikova T. et al. 2004. The EMBL nucleotide sequence database. *Nucleic Acids Research* 32:D27-30.

[4] Lefranc, M.P., Giudicelli,V., Ginestoux,C., Bodmer,J., Muller,W., Bontrop,R., Lemaitre,M., Malik,A., Barbie,V. and Chaume,D. 1999. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research* 27:209-212.

F18. The role of pre-mRNA secondary structure in splicing of *Saccharomyces cerevisiae*

Sanja Rogic [1], Holger H. Hoos [1,2],B.F. Francis Ouellette [2], Alan K. Mackworth [1]

Keywords: gene splicing, RNA secondary structure

1 Introduction.

Splicing of pre-mRNA, which involves the excision of introns from a primary RNA transcript and ligation of exons into mature mRNA, is one of the essential cellular processes in eukaryotic organisms. Although this process has been extensively studied since the discovery of splicing almost three decades ago, the highly reliable identification of splice sites by the splicing machinery is still not a very well understood phenomenon. It is currently believed that the identification is accomplished through the base-pairing interactions between the conserved sequences at the exon/intron boundaries as well as within introns and the small spliceosomal RNAs. However, these sequences are short and often hard to distinguish from the numerous unutilized sequences throughout the genome.

The inadequacy of the primary sequence to unambiguously specify splice sites prompted scientists to speculate about the effects of higher order RNA structure. There are a number of research articles discussing various effects of RNA secondary structure on splicing, but more general role has not been established. In our current research we are attempting to characterize the types and role of RNA secondary structure in the context of gene splicing.

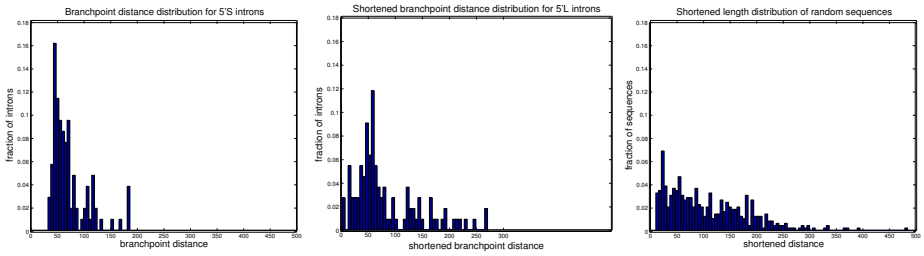
2 Methods.

S. cerevisiae, which was chosen to be the model organism for this research, has a characteristic bimodal intron length distribution [5], with shorter introns ranging in length from 50-200 nt and longer ones from 200-1000 nt. Since the intron length and the distance between 5' splice site and the branchpoint sequence are tightly correlated in yeast, the distribution of the latter is also bimodal. It is believed that shorter branchpoint distance (35-200 nt) is optimal for splicing since it facilitates direct interactions between the small nuclear RNAs involved in splicing. Analysis of the few yeast introns with a longer branchpoint distance, also called 5'L introns, revealed the existence of a stem/loop structure that effectively shortens the distance between the 5' splice site and the branchpoint to the optimal value [1, 4, 2].

To investigate if the formation of the stem/loop structure between the 5' splice site and the branchpoint is common for all yeast introns with long branchpoint distance we used two approaches. In the first one we computationally predicted secondary structures of 111 *S. cerevisiae* 5'L introns and searched for 'zipper' stems that satisfied certain thermodynamic criteria. A 'zipper' stem is a stem located between the 5' splice site and the branchpoint which maximally shortens the distance between these two sequences. The distribution of branchpoint distances shortened by 'zipper' stems was then compared to the branchpoint

[1]Department of Computer Science, The University of British Columbia, Vancouver, Canada. E-mail: {rogic,hoos,mack}@cs.ubc.ca

[2]UBC Bioinformatics centre, The University of British Columbia, Vancouver, Canada. E-mail: francis@bioinformatics.ubc.ca



(a) (b) (c)

Figure 1: Comparing distributions for (a) branchpoint distance for 5'S introns, (b) branchpoint distance for 'zipped' 5'L introns , (c) length of 'zipped' random sequences.

distance distribution of introns with short branchpoint distance (5'S introns) using the standard Kolmogorov-Smirnov test. The distributions were found not to be significantly different. To further test the significance of these results we performed the same 'zipper' stem analysis for random and exon sequences and obtained distributions that were significantly different from the original branchpoint distance distribution for 5'S introns (Figure 1).

We also used a comparative genomics approach to test for the existence of 'zipper' stems in yeast introns. In this context, we used homologous intronic sequences from three species closely related to *S. cerevisiae*: *S. paradoxus*, *S. mikatae*, and *S. bayanus* [3]. Multiple sequence alignments of these sequences were visually searched for conserved structural elements that could function as 'zipper' stems. The alignments were also processed by several programs that predict use conserved secondary structure elements based on compensatory mutations. Potential 'zipper' stems were found for each of the alignments and the resulting distribution of shortened branchpoint distances was not significantly different from the original distribution for 5'S introns.

In addition to the 'zipper' stem analysis we also looked at the secondary structure of the branchpoint region of the yeast introns. We compared the predicted secondary structures of the branchpoint sequences with those of other intronic sequences of the same length by counting the number of unpaired bases. The results we obtained indicate that the branchpoint sequences have significantly less base-pairing interactions than an average intronic sequence of the same length.

References

[1] Goguel, V. and Rosbash, M. 1993. Splice site choice and splicing efficiency are positively influenced by pre-mRNA intramolecular base pairing in yeast. *Cell.* 72:893–901.

[2] Howe, K. J. and Ares, M. Jr. 1997. Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. *Proc Natl Acad Sci USA.* 94:12467–12472.

[3] Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature.* 423:241–254.

[4] Libri, D., Stutz, F., McCarthy, T., and Rosbash, M. 1995. RNA structural patterns and splicing: molecular basis for an RNA-based enhancer. *RNA.* 1:425–436.

[5] Parker, R. and Patterson, B. 1987. Architecture of fungal introns: implications for spliceosome assembly. In Inouye M. and Dudock B. S., editors, *Molecular biology of RNA: new perspectives*, Academic Press, Inc., San Diego, CA, USA. 133–149.

F19. e2g - A Web-Based Tool for Efficiently Aligning Genomic Sequence to EST and cDNA data

Alexander Sczyrba, Jan Krüger, Robert Giegerich [1]

Keywords: EST alignment, gene structure prediction, suffix array

1 Introduction.

High throughput cDNA and EST sequencing projects have generated a vast amount of data representing the transcribed portion of the organisms in study. As soon as (parts of) the sequence of the associated genome becomes available, gene structures can be determined by mapping the cDNA data to the genomic sequence. This allows the detection of genes missed by gene prediction tools and the determination of splice variants of already known genes.

While several tools for mapping ESTs and cDNAs to genomic sequence already exist [1, 2, 3], they can hardly be used in an interactive web-based application because of the huge amount of data to be searched against. `e2g` is a web-based tool which efficiently aligns genomic sequence to indexed cDNA and EST databases. This allows users to rapidly detect the exon-intron structure of genes, including variants, in the genomic region of interest.

`e2g` is online available on the Bielefeld University Bioinformatics Server at:

`http://bibiserv.techfak.uni-bielefeld.de/e2g/`

2 Method.

The web interface accepts either (i) genomic sequence or (ii) genomic sequence and cDNA/EST data as input. In the first case the sequence will be matched against a database of cDNAs and ESTs. (Currently, databases for human and mouse are available.) In the second case, the user provides the cDNA data to be matched against the genomic sequence. In both cases, repeats and low-complexity regions will be masked before matching.

As good efficiency is critical for the approach, an enhanced suffix array is built on the server as an persistent index of the EST sequences using `mkvtree` [4, 5]. The index efficiently represents all substrings of the database sequences and allows to solve matching tasks in time independent of the size of the index, done using the string matching algorithm `vmatch` [4, 5]. Matches can be computed either exactly, or approximately by extending the exact seeds using the X-drop strategy [6]. Matching the 16.5kb genomic sequence of the example in figure 1 against all mouse EST data (approx. 2.5 GB), takes 50 seconds on a SUN UltraSparc III (800 Mhz) with 64 GB RAM. Allowing mismatches using the X-drop strategy (99% identity, seedlength 15) increases the running time to 73 seconds.

3 Web interface.

Figure 1 shows a screenshot of the `e2g` web interface. Matches of the ESTs reveal the exon-intron structure of the gene. On the top the annotation of the submitted genomic region is shown. This can be overlayed by dragging a transparent image over the lower part of the window, allowing the user to easily compare the annotated gene structure to the matches

[1] Technische Fakultät, Bielefeld University, D-33594 Bielefeld, Germany.
E-mail: {asczyrba,jkrueger,robert}@TechFak.Uni-Bielefeld.DE

found. Information about the matches of the circled exons is shown on top. The alignment of each match can be calculated on the fly, reusing the existing index (see popup window in the lower left).

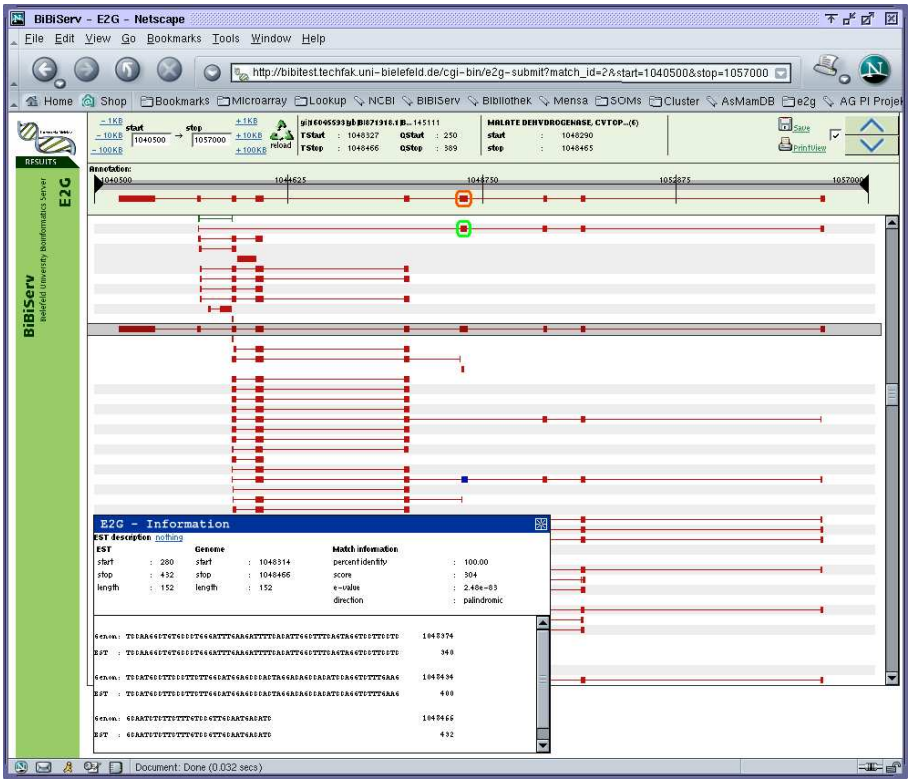


Figure 1: Screenshot of the `e2g` web interface.

References

[1] R. Mott. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *CABIOS*, 13(4):477–8, 1997.

[2] W.J. Kent. BLAT-The BLAST-Like Alignment Tool. *Genome Res.*, 12(4):656–664, 2002.

[3] C. Del Val, K.H. Glatting, and S. Suhai. cDNA2Genome: A tool for mapping and annotating cDNAs. *BMC Bioinformatics*, 4(1):39, 2003.

[4] M.I. Abouelhoda, S. Kurtz, and E. Ohlebusch. The Enhanced Suffix Array and its Applications to Genome Analysis. In *Proceedings of the Second Workshop on Algorithms in Bioinformatics*. Springer-Verlag, Lecture Notes in Computer Science, 2002.

[5] M.I. Abouelhoda, E. Ohlebusch, and S. Kurtz. Optimal Exact String Matching Based on Suffix Arrays. In *Proceedings of the Ninth International Symposium on String Processing and Information Retrieval*, pages 31–43. Springer-Verlag, Lecture Notes in Computer Science 2476, 2002.

[6] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *J Comput Biol*, 7:203–14, 2000.

F20. LinkageView: a powerful graphical tool for integrating statistical data with Ensembl

Judith E. Stenger[1,2,3], Hong Xu[1,2], Carol Haynes[1,2], Elizabeth R. Hauser[1,2], Margaret A. Pericak-Vance[1,2], Pascal J. Goldschmidt-Clermont[2] and Jeffery M. Vance[1,2]

Keywords: bioinformatics, linkage analysis, genome annotation, database integration, complex disease susceptibility genes.

1 Introduction.

To facilitate efficient prioritization of candidate disease susceptibility genes for association analysis, increasingly comprehensive tools are essential to smoothly integrate vast quantities of disparate private and public data generated by genomic screens, association studies, whole genome sequencing and annotation data and ancillary biological databases. For this purpose, we have developed "LinkageView", a tool that enables statistical data to be viewed in the context of Ensembl's [1] Contig- and Cyto- View pages. Drawn under the ideogram, LinkageView illustrates lod score values for markers from a particular study plotted against the position that is determined in base pairs by mapping the primers to the chromosome using electronic PCR (e-PCR)[2]. The graphs are transparently integrated into the Contig- and Cyto-View pages and feature a sliding selection box to display a region of interest enlarged in the Overview and the Detailed View. Genomic features mapping to regions with evidence of linkage are accentuated when LinkageView is used with the Distributed Annotation System (DAS) [3] to display supplemental laboratory information as tracks containing information such as differentially expressed genes pertaining to the disease under investigation. LinkageView is a powerful visual feature that enhances the use of Ensembl as a resource for integrative genomic-based approaches for identifying candidate disease susceptibility genes. We are currently developing a panel to display association data in a region of interest.

2 Software and files.

Linkage data is traditionally displayed as a graph in which lod scores are plotted along the ordinate against the genetic location (in cM) of the markers used in the genetic screen and any subsequent analysis. To integrate such data with the physical genome sequence, it is imperative that the abscissa be expressed in base pairs so that the position of markers along the abscissa correctly align with, and strictly correspond to, the horizontal illustration of the ideogram that is displayed immediately above the linkage study graph in the Ensembl Contig View (Figure 1, second panel).

First we downloaded the human genome assembly (NCBI build 34) from the UCSC genome site [4] and the most recent version of the Ensembl annotation system. Then, we mapped the markers used as probes to the actual sequence of the human genome. We used the NCBI UniSTS database (ftp://ftp.ncbi.nlm.nih.gov/repository/UniSTS/) as the source of marker probe sequence information to convert the map units into base pair coordinates by locating the position of the STSs in human

[1] Center for Human Genetics, Duke University Medical Center, P.O. Box 3445, Durham, North Carolina 27710, USA. E-mail: center@chg.duhs.duke.edu

[2] Department of Medicine, Duke University Medical Center, P.O.Box 3703, Durham, North Carolina 27710, USA. E-mail: pascal.goldschmidt@duke.edu.

[3] Presenter and corresponding author, E-mail: jstenger@chg.duhs.duke.edu

genome build 34 using e-PCR [2] and finally populated the statistical results table in the database.

We developed four BioPerl modules to draw the linkage plot in Contig View using the data uploaded in the DAS server for a particular study. The Linkage.pm module is a constructor for the linkage object, essentially a table encapsulating the record for a single linkage point. This allows linkage data to be added to the database. Next, LinkageAdaptor.pm, provides the functionality for accessing linkage data from the DAS database. A "slice" object is created which defines the region from which linkage data will be retrieved and returns the records from the linkage objects that are mapped to the chromosome between the coordinates of the slice. The GlyphSet::lodplot.pm module contains all the information for drawing the graph of the statistical data. The WebUserConfig::chrplot.pm module is then added to configure the lod plot so that it is displayed in the LinkageView panel in Contig- and Cyto- View. Finally the plot image was scored into the Ensembl Contig View and Cyto View by modifying the cytocide code so that the linkage plot can displayed into both Contig- and Cyto- View.

3 Figure(s).

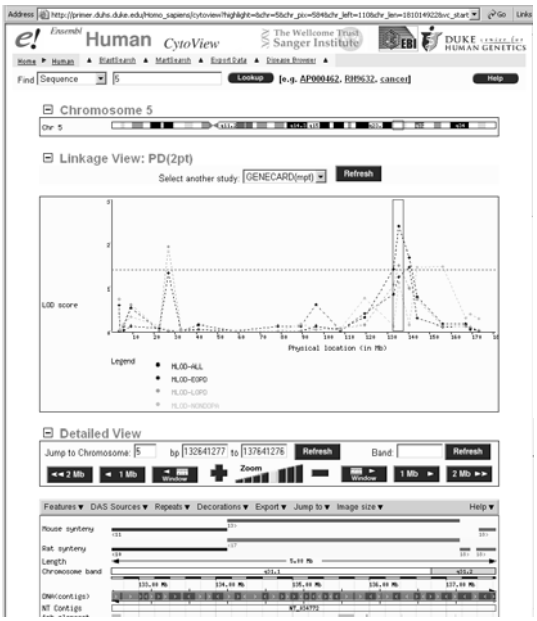


Figure 1: A screen shot of linkage data displayed in Contig View on a local server

3 References.

[3] Dowell R.D., Jokerst R.M., Day A., Eddy S.R., and Stein L. 2001. The Distributed Annotation System. *BMC Bioinformatics.* 2:7.

[1] Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* 30:38-41.

[4] Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., and Haussler D. 2002. The human genome browser at UCSC. *Genome Res.* 12:996-1006.

[2] Schuler GD. 1997. Sequence mapping by electronic PCR. *Genome Res.* 7:541-50.

F21. HMM-based System for Identification of Related Gene/Protein Names

L. Yeganova[1], L. Smith[2], W. J. Wilbur[3]

Keywords: gene name variation, hidden Markov model, information extraction.

Abstract

Gene and protein names follow few, if any, true naming conventions and are subject to great variation in different occurrences of the same name. This gives rise to two important problems in searching biological literature: first, can one locate the names of genes or proteins in free text, and second, can one determine when two names denote the same gene or protein?

Several researchers have looked at the first problem of identifying gene or protein names in molecular biology texts. A large number of different methods have been applied to this problem, including part-of-speech tagging, hidden Markov models (HMM), decision trees, Bayesian methods, rule based systems, regular expressions, and a variety of knowledge based resources.

The problem of determining when two differing strings represent the same gene or protein seems to have received much less attention. Few methods for handling such term variations have been developed. A BLAST[4]-based system presented by Krauthammer et al. [4] uses approximate string matching techniques and dictionaries to recognize spelling variations in gene or protein names. They have encoded gene names and text in terms of the nucleotide alphabet and have used BLAST to look for 'homologies' between a query gene name and the text. A similar problem was addressed by Cohen et al. [3] who studied contrast and variability in gene names to develop heuristics to distinguish between gene or protein names with different meaning from names that are synonyms. They found that capitalization could be ignored, parenthesized material and hyphens were optional, and vowels tended to be interchangeable. However, they did not implement a system based on these observations.

Here we describe a system which, given a query gene or protein name, identifies related gene or protein names in a large list. The system is based on a dynamic programming algorithm for sequence alignment in which the mutation matrix is allowed to vary under the control of a fully trainable hidden Markov model [5], [6]. We have used this method to provide an access portal to GenBank [2] in which one may enter a putative gene name and retrieve the names (with accompanying GenBank id) that appear to be the closest. From these the user may select and access those GenBank records judged by him to be useful matches to his query. The website is available at http://web.ncbi.nlm.nih.gov/IRET/Gene_Prot_Match.

References

[1] Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., 1990. Basic local alignment search tool. Journal of Molecular Biology, 215, 403-410.

[2] Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L., 2003. GenBank. Nucleic Acids Res, 31(1), 23-27.

[1] Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. Tel:301-402-0776; E-mail:yeganova@ncbi.nlm.nih.gov
[2] Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. Tel:301-594-2845; E-mail:lsmith@ncbi.nlm.nih.gov
[3] Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. Tel:301-435-5926; E-mail:wilbur@ncbi.nlm.nih.gov
[4] BLAST - Basic Local Alignment Search Tool [1]

[3] Cohen, K. B., Acquaah-Mensah, G. K., Dolbey, A. E., Hunter, L., 2002, July 11. Contrast and variability in gene names. Paper presented at Natural Language Processing in the Biomedical Domain, University of Pennsylvania.

[4] Krauthammer, M., Rzhetsky, A., Morozov, P., Friedman, C., 2000. Using BLAST for identifying gene and protein names in journal articles. Gene, 259, 245-252.

[5] Smith, L., Yeganova, L., Wilbur, W. J., 2003. Hidden Markov models and optimized sequence alignments. Computational Biology and Chemistry, 27, 77-84.

[6] Yeganova, L., Smith, L., Wilbur, W. J., 2003. Identification of related Gene/Protein names based on an HMM of name variations. Computational Biology and Chemistry, in press.

F23. A New Tool for Enumerative Combinatorics?

James Nulton[1], Peter Salamon[2], Mya Breitbart[3], Joe Mahaffy[4], Ben Felts[5], Beltran Rodriguez Brito[6], David Bangor[7], Forest Rohwer[8]

Keywords: shotgun sequencing, contigs, bernoulli polynomials, convolution

1 Introduction.

In 1988 Lander and Waterman [1] presented a mathematical framework for approximating various statistics related to shotgun sequencing, a tool used in planning projects related to the physical mapping of the genome. More recently [2], that framework has been used as part of a scheme for inferring population structural features of a community of marine bacteriophage from a statistical analysis of a shotgun library assembled from that community. In a shotgun sample for a single genome a maximal assembly of q sequences that covers a contiguous portion of the genome map is called a *q-contig*. Good estimates for the expected number of q-contigs is critical for this recent application. Lander/Waterman does very well for lower values of q, but Monte Carlo simulations show that, as q increases, the quality of the required estimates deteriorate.

The result showcased below was developed in an effort to analyze the shotgun sampling experiment by exact combinatorial counts. A special class of basis events was identified in terms of which all other events of interest (including contig events) could be expressed by standard combinatorial methods. The method of counting the events in the special class uses a construct that appears to be new to the field of combinatorics. It is a type of convolution in the ring of polynomials.

2 Sampling, Basis Events, and the B-convolution.

A sample (with replacement) of K numbers, called *points* is selected with uniform probability from the set $[G] = \{1, \ldots, G\}$. The sample space consists of G^K equally likely outcomes and can be identified with the set of maps $\{f : [K] \to [G]\}$. We are interested in clustering patterns among the points of an outcome. Let p and p' be a pair of sample points with no sample points between them. If $|p - p'| \geq C$, the pair forms a *gap*, otherwise it forms a *link*. The number C is called the *gap threshold*. In the context of a shotgun experiment, $[G]$ is the genome map, the sample points mark the starts (on the map) of the sequenced fragments, and C is the effective fragment length.

List the K points of an outcome (with possible repetitions) in non-decreasing order and consider consecutive pairs in the list. Each outcome has an ordered pattern of gaps and

[1]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720. E-mail: jnulton@mail.sdsu.edu

[2]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720. E-mail: salamon@saturn.sdsu.edu

[3]Department of Biology, San Diego State University, San Diego, California, 92182-4614. E-mail: mya@sunstroke.sdsu.edu

[4]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720.

[5]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720.

[6]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720.

[7]Department of Mathematics, San Diego State University, San Diego, California, 92182-7720.

[8]Department of Biology, San Diego State University, San Diego, California, 92182-4614. E-mail: forest@sunstroke.sdsu.edu

links that can be represented as a sequence of $K-1$ binary digits, each of which records the status of a consecutive pair, "0" for gap, "1" for link. The jth digit records the status of the jth consecutive pair in the sorted list. Conversely, such a binary string denotes the event consisting of all outcomes with such a pattern of gaps and links. For example, 1001011 is an event ($K=8$) in which the first consecutive pair in the sorted list of points is a link, the second and third are gaps, etc. More generally, a symbol such as $XX0110X$ represents an event in which we are indifferent to the status of pairs in positions 1, 2, and 7, but the status of the others are specified. Incidentally, this event has a 3-contig (2 links) starting at position 4. The problem is to determine the cardinality of such a set. Standard counting methods (principle of inclusion/exclusion) allow us to reduce the problem to a simpler class of *basis events* whose binary symbols use only "X" and "0". Our method of counting such sets uses the following construct on the ring of polynomials.

Let g and h be polynomials of degrees m and n. The equation below defines a polynomial of degree $m+n+1$.

$$f(y)=\sum_{x=0}^{y} g(y-x)h(x) \tag{1}$$

We call f the *B-convolution* of g and h and denote it by $g\#h$. "B" is for Bernoulli, whose polynomials are required for the reduction of the right side to polynomial form.

We now illustrate with an example how this convolution is used to count the number of outcomes in a basis event. To that end, for each positive integer j, define the special polynomials: $b_j(x)=x^j$ and $a_j(x)=(x+1)^j-x^j$. Consider the basis event $E=X0XXX00XX$ (for $K=10$). Introduce 10 asterisks to set off the 9 symbols: $*X*0*X*X*X*0*0*X*X*$. Now remove the X's: $**0****0*0***$. The 10 asterisks are partitioned into $2+4+1+3$. Construct the polynomial $\pi_E=b_2\#a_4\#a_1\#a_3$. Count the number of terms, M, in the partition: $M=4$. The cardinality of the event E is $\pi_E(G-(M-1)C)$. It is a theorem that π_E is independent of the partition order, despite the apparent broken symmetry with a "b" as the first factor in the convolution.

3 The Expected Number of q-Contigs.

Finally, we summarize the result for the number, x_q, of q-contigs in a sample of size K. For every q there is a set of polynomials, $\{\,p_q^{(j)}\,|\,j=0,\ldots,q+1\}$, all of degree $q-1$, and *depending only on* q, in terms of which the expected number of q-contigs is given by

$$E[x_q]=G^{-K}\sum_{j=1}^{q+1}(2p_q^{(j)}\#b_{K-q}+p_q^{(j-1)}\#d_{K-q})(G-jC), \tag{2}$$

where $d_i(x)=x(x+1)a_{i-1}(x)$ are polynomials of degree i. All of the functions represented here as polynomials must, nevertheless, be taken to vanish on negative arguments.

We remark that the p's are integral linear combinations of π's associated with basis events, and there are recursion relations among them. We emphasize that the p's are computed independently of the parameters K, G and C.

References

[1] Lander E. S. and Waterman M. S. 1988. Genomic mapping by fingerprinting random clones. *Genomics* 2:231–239.

[2] Breitbart, M. et al. 2002. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences USA* 99:14250–14255.

G1. A Latent Process Decomposition Model for Interpreting cDNA Microarray Datasets

Simon Rogers, [1] **Mark Girolami,** [2] **Colin Campbell,** [3]

Keywords: cDNA Microarray, Latent Variable Modelling, Cluster Analysis

1 Introduction

We outline a hierarchical Bayesian model which enables the probabilistic analysis of cDNA microarray data. The parameters in this model are estimated using an efficient variational technique in which each sample is represented as a combinatorial mixture over a finite set of latent processes. For determining informative substructure in such datasets the proposed model has several important advantages over the standard use of dendrograms, such as:

- the ability to objectively assess the optimal number of sample clusters.
- the ability to represent samples and gene expression levels using a common set of latent variables (dendrograms cluster samples and gene expression values separately which amounts to two distinct reduced space representations).
- observations are not assigned to a single cluster and thus a gene expression level is modelled via a combination of latent processes.

In analogy to Latent Dirichlet Allocation [1] we define a Dirichlet probability distribution for the distribution of $\mathcal{K}$ possible processes labelled by a parameter $\boldsymbol{\alpha}$. For a sample c, a distribution $\boldsymbol{\theta}$ over a set of mixture components indexed by the discrete variable k is drawn from the Dirichlet distribution. Then, for each of the $\mathcal{G}$ features (generally uniquely mapping to genes) g, we draw a process index k from the distribution $\boldsymbol{\theta}$ with probability θ_k which selects a Gaussian defined by the parameters μ_{gk} and σ_{gk}. The level of expression e_{gc} for gene g for the sample c is then drawn from the k'th Gaussian denoted as $\mathcal{N}(e_{gc}|k, \mu_{gk}, \sigma_{gk})$. This is then repeated for each of $\mathcal{C}$ tissue samples. This is a biologically more realistic representation than the mixture model which underlies many forms of clustering and the shared latent space facilitates easy extraction of genes differentially expressed across different processes.

2 Parameter Inference

The likelihood of the c^{th} sample under this model is

$$p(c|\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\alpha}) = \int_{\Delta} \left\{ \prod_{g=1}^{\mathcal{G}} \sum_{k=1}^{\mathcal{K}} \mathcal{N}(e_{gc}|k, \mu_{gk}, \sigma_{gk}) \theta_k \right\} p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta}. \quad (1)$$

Variational inference can be employed to estimate the parameters of this model by introducing a set of sample specific Dirichlet priors $p(\boldsymbol{\theta}|\boldsymbol{\gamma}_c)$ and using Jensen's inequality to lower bound the log likelihood of all $\mathcal{C}$ training samples

[1]Dept. of Engineering Mathematics, Queen's Building, University of Bristol, Bristol BS8 1TR, United Kingdom. E-mail: simon.rogers@bristol.ac.uk

[2]Bioinformatics Research Centre, Dept. of Computing Science, University of Glasgow, Glasgow G12 8RZ, United Kingdom. E-mail: girolami@dcs.gla.ac.uk

[3]Dept. of Engineering Mathematics, Queen's Building, University of Bristol, Bristol BS8 1TR, United Kingdom. E-mail: c.campbell@bristol.ac.uk

$$\sum_{c=1}^{\mathcal{C}} \log p(c|\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\alpha}) \geq \sum_{c=1}^{\mathcal{C}} \mathbf{E}_{p(\boldsymbol{\theta}|\boldsymbol{\gamma}_c)} \left[\log \left\{ p(c|\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\theta}) \frac{p(\boldsymbol{\theta}|\boldsymbol{\alpha})}{p(\boldsymbol{\theta}|\boldsymbol{\gamma}_c)} \right\} \right] \tag{2}$$

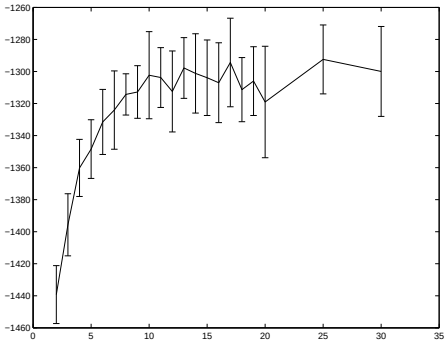
and by introducing the variational variable Q_{kgc} such that $\sum_{k=1}^{\mathcal{K}} Q_{kgc} = 1$ we can obtain the following bound of the likelihood conditioned on $\boldsymbol{\theta}$ rather than $\boldsymbol{\alpha}$

$$\sum_{c=1}^{\mathcal{C}} \log p(c|\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{\theta}) \geq \sum_{k=1}^{\mathcal{K}} \sum_{c=1}^{\mathcal{C}} \sum_{g=1}^{\mathcal{G}} Q_{kgc} \log \left\{ \frac{\mathcal{N}(e_{gc}|k,\mu_{gk},\sigma_{gk})\theta_k}{Q_{kgc}} \right\} \tag{3}$$

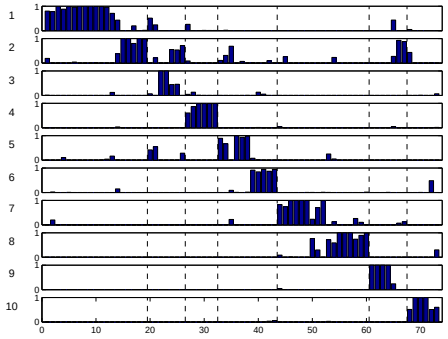
Employing these bounds together, we can define iterative EM-like updates for the model parameters Q_{kgc}, γ_{kc}, μ_{gk}, σ^2_{gk} and α_k. On termination, normalised γ_{kc} represent the confidence in allocating the c'th sample to process k. We can also derive an estimate of the likelihood enabling comparison between models and an estimate of the best value of $\mathcal{K}$.

3 Numerical Experiments

To illustrate the performance of this model and the resulting algorithm we will briefly illustrate its performance on a microarray dataset for lung cancer. This dataset [2] consists of 73 gene expression profiles from normal and tumour samples with the tumours labelled as squamous, large cell, small cell and adenocarcinoma. 10-fold cross validation was performed for values of $\mathcal{K}$ between 2 and 30 and the likelihood of the held out samples is shown in figure (a). Figure (b) shows the assignment of samples (each column represents one sample) to processes (rows). The dotted lines represent the clustering defined in [2]. Clustering by dendrogram is accurately reproduced by Latent Process Decomposition (LPD). Indeed two adenocarcinoma samples (66,67) wrongly placed by the dendrogram with small cell carcinomas (61-67) and correctly placed by LPD. For other cancer datasets the likelihood peaks at a distinct value of $\mathcal{K}$ which gives a clear indication of the number of latent processes required to accurately fit the data. Evaluation on further datasets and more details about the method will be given elsewhere.



(a)



(b)

References

[1] Blei, D.M and Ng, A.Y. and Jordan, M.I. 2003 .Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3:993–1022.

[2] Garber, E et al 2001. Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of the national academy of sciences USA* 98:12784–13789.

G3. Discovering Statistically Significant Clusters by Using Genetic Algorithms in Gene Expression Data

Hwa-Sheng Chiu[1], Han-Yu Chuang[1], Huai-Kuang Tsai[1], Tao-Wei Huang[1], Cheng-Yan Kao[1]

Keywords: genetic algorithms, clustering, statistically significant, gene expression data, microarray

1 Introduction.

This study presents an iterative GA (genetic algorithm) approach to find significant clusters in gene expression data. Clustering genes of similar expression patterns is useful for predicting gene functions and pathways. Many heuristic algorithms have developed to cluster genes on microarray data [1]. However, the results of most heuristics may be dominated by some predefined criteria, such as the number of clusters and initial mean points of clusters. Moreover, the present methods assign each one of genes in the dataset into some cluster while some of them are irrelevant to the interested conditions. To automatically cluster informative genes together, we regard clustering as an optimization problem of maximizing correlativity among genes in a set. Thus, an iterative GA approach is proposed to find the tightest clusters subject to our statistical definition of a significant cluster. The proposed method is applied on a yeast cell cycle dataset of 614 genes and 77 conditions [2]. It efficiently finds 21 statistically significant clusters and discards 105 genes. The experiment results show that genes of the same cluster are highly related to their mean pattern, and genes not belonging to any clusters are far away from each mean pattern. We proceed to apply the proposed method to biclustering.

2 System and method.

The proposed method works as follows. Our GA finds the cluster of the maximal fitness in the interested dataset, and removes the corresponding genes in the duster from the dataset. Then the GA starts the whole searching process again in the remaining dataset. Until there are no any significant clusters at one run, the iterative process is terminated. In our GA, each chromosome is represented as a bit string with g bits, where g is the number of genes in current target dataset. A chromosome corresponds to a cluster of genes. If a bit of a chromosome has value of 1, its corresponding gene is included in the gene set corresponding to the chromosome. N sets of genes, subject to the constraints of a cluster, are randomly generated as the initial population. After evaluating the fitness, the sequential steps of family competition [3] and stochastic universal sampling (SUS) are run with a k-point probabilistic-elitism crossover and a bitwise uniform mutation. The GA is terminated when the following criterion is satisfied: the improvement ratio between all of the children generated and their respective family parents is less than 0.001 in five continuous generations.

■ *Significant Clusters*

A subset of genes can be regarded as a cluster in our method, if the average pairwise Pearson correlation (aPPC) of genes in the subset is more than the one of genes in the whole set excluding the subset. Moreover, the cluster is statistically significant if its aPPC is more than the sum of the expected pairwise Pearson correlation in the dataset and $\alpha\,\sigma$, where σ is the expected S.T.D. and α is a threshold to decide the significance.

[1] Bioinfo Lab., Dept. of Computer Science and Information Engineering, National Taiwan University, Taiwan. E-mail: {r91031, r90002, d7526010, d90016, cykao}@csie.ntu.edu.tw

■ *Fitness Function*

Each chromosome S_i in the population is evaluated by using the sum of all pairwise Pearson correlation of its corresponding genes, instead of using the average one. If only using the average Pearson correlation to choose a cluster, the trivial cluster of two genes will be derived. Our GA prefers chromosomes of larger sum subject to the constraints of the significant cluster. Besides, we introduce the aPPC as the weight to the fitness function, in order to get the set of quite correlative genes.

■ k-point Probabilistic-elitism Crossover

The crossover is modified from the multipoint crossover. The crossover works by randomly selecting *k* points in each parent and splitting them into *k* segments. Then it glues segments of parents to obtain offspring by the probability proportional to the fitness of the segment. If the fitness of the segment of one parent is better than the corresponding fitness of the other, the offspring would inherit the better segment more likely.

3 Result.

The proposed method is applied on a yeast cell cycle dataset with 614 genes and 77 conditions. 21 statically significant clusters are derived and 105 genes are discarded. The top-6 statically significant clusters are list in Fig. 1. The thick line is the mean pattern of this cluster and the thin line above or below to the mean pattern is 1 standard deviation from the mean pattern. As shown, the different clusters have the different expression patterns and the standard deviation in all clusters is small. We also found that the pairwise Pearson correlations between 105 discarded genes and 21 cluster mean patterns are small. The experimental results indicate that the proposed approaches can find tight and significant clusters efficiently and filter out irrelative genes effectively.

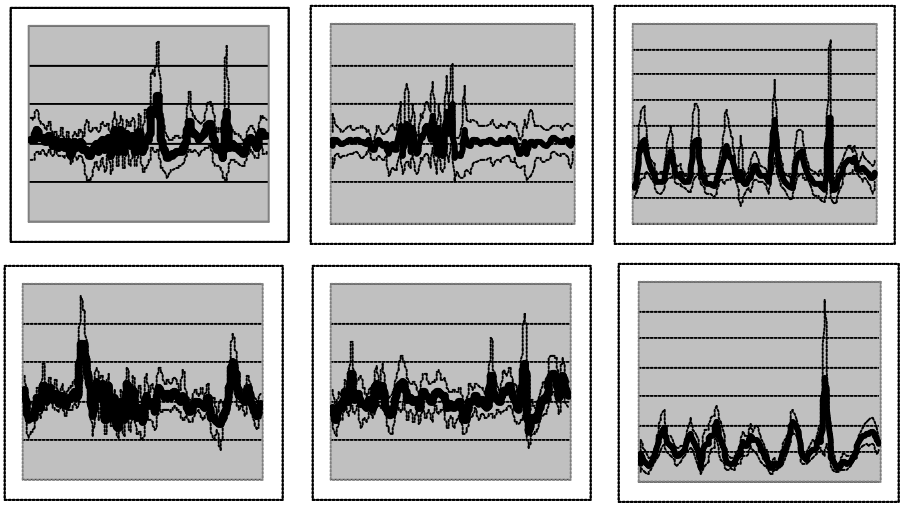


Fig. 1: The top-6 statically significant clusters. The x-axis represents experiments and the y-axis represents the expression levels. Each cluster has different expression patterns from others.

References

[1] Tou, J. T. and Gonzalez, R.C.. 1974. *Pattern Recognition Principles*, Addison-Wesley, Reading.

[2] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B.1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces erevisiae by microarray hybridization. *Molecular Biology of Cell*, vol. 9, no. 12, pp. 3273–3297

[3] Yang, J. M. 2001. A Family Competition Evolutionary Approach of Global Optimization in Neural Networks, Optical Thin-film Design, and Structure-based Drug Design. *Ph. D. thesis, National Taiwan University, Taiwan*.

G4. J-Express - an integrated tool for processing and analyzing microarray gene expression data

Bjarte Dysvik[1], Kjell Petersen[2,3], Trond Hellem Bø[1], Kristin Sandereid[4], Inge Jonassen[1,2]

Keywords: Gene expression analysis, software, microarray data analysis, J-Express, MAGE.

1 Introduction.

A new version of J-Express Pro [1] will soon be released. In addition to upgrading most of the existing framework, we have also added functionality for low-level data processing and quality control. The MAGE [2] compatibility has also been upgraded, and we are now able not only to read data in MAGE format but also create new objects for export. We give a short overview of the system, its functionality, and we briefly explain basic principles and special features of the system.

2 Software description.

Microarray gene expression experiments generate relatively large amounts of data that need to be processed (normalized, filtered, etc) and analyzed using for example data reduction, clustering, and classification approaches. The particular processing and analysis that is appropriate depends on the data set; the experimental design, the experimental protocols used, and the nature of the questions one wants to address.
Data, intermediate and final results need to be visualized to facilitate interpretation and understanding.

The J-Express tool is developed to offer an integrated tool that at the same time offers a wide range of processing and analysis functionality while keeping the user interface simple and intuitive. Special features of J-Express include (1) automatic recording of information on all processing and analysis steps used to arrive at a result (meta data integrated in a project management system), (2) produce high quality graphical results, (3) facilitate easy extension of the system by including plug-ins or use of an internal scripting language, (4) allow integration with external tools and databases through MAGE-ML files or MAGE objects. The implemented MAGE functionality is developed in collaboration with the EBI ArrayExpress team and is compatible with the latest adapted policies for submission to ArrayExpress. The J-Express tool includes a flexible quality-control, filtering, and normalization system, clustering (hierarchical, K-means, and self-organizing maps), projection methods (principal component analysis and multi-dimensional scaling), supervised methods (feature selection), and tools for integration of external data, e.g., pathway and genome data.

[1] Department of Informatics, University of Bergen
[2] Computational Biology Unit, Bergen Centre for Computational Sciences, University of Bergen
[3] DESPRAD – Development and Establishment of Standards and Prototype Repository for DNA-Array Data. (EU funded project)
[4] Forinnova AS, Bergen

3 Availability.

Software available at http://www.molmine.com.

Other relevant web addresses:
Microarray Gene Expression Data Society: http://www.mged.org
The Norwegian Microarray Consortium: http://www.mikromatrise.no
The Norwegian Functional Genomics program: http://www.bioinfo.no

References

[1] Bjarte Dysvik and Inge Jonassen. J-Express: exploring gene expression data using Java *Bioinformatics*, 17, 369-370 (2001)
[2] Paul T Spellman, Michael Miller, Jason Stewart, Charles Troup, Ugis Sarkans, Steve Chervitz, Derek Bernhart, Gavin Sherlock, Catherine Ball, Marc Lepage, Marcin Swiatek, WL Marks, Jason Goncalves, Scott Markel, Daniel Iordan, Mohammadreza Shojatalab, Angel Pizarro, Joe White, Robert Hubley, Eric Deutsch, Martin Senger, Bruce J Aronow, Alan Robinson, Doug Bassett, Christian J Stoeckert Jr, Alvis Brazma. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, 3(9), (2002), research0046.1-0046.9

G5. Sorting Points Into Neighborhoods ($SPIN$): a novel data organization and visualization tool

Ilan Tsafrir [1], Liat Ein-Dor[1], Dafna Tsafrir[1], Or Zuk[1], Eytan Domany[1]

Keywords: Microarray, data analysis, data visualization, data organization, sorting

Exploratory data analysis is critical in a broad range of research areas, where large collections of data need to be meaningfully arranged and presented in a human-oriented manner. We present $SPIN$, a novel method for the organization and visualization of data, implemented in a simple tool. $SPIN$ utilizes traits of distance matrices to sort objects in a natural ordering that highlights the underlying structure of the original, multidimensional data. As an unsupervised analysis tool, $SPIN$ does not rely on any external labels, but rather explores the inherent characteristics of the data. In the analysis of high-throughput biological experiments, discretely - labelled data, such as clinical labels of 'sick' versus 'healthy', is traditionally organized by various clustering approaches. However, when the objects are characterized by one or more continuous variables, e.g. survival intervals for patients, any sharp separation into distinct clusters will be rather arbitrary. Thus, a different organization approach, one which emphasizes ordering rather than grouping, could be more relevant.

This work focuses on finding a one-dimensional ordering of a set composed of n data points, and to present as output the matching (2-dimensional) n by n distance matrix D. An element D_{ij} of D represents the dissimilarity between objects i and j. Our aim is to find a permutation of the data points, such that the correspondingly reordered distance matrix reveals the underlying structure of the data.

$SPIN$ is especially suitable for analyzing high-throughput biological experiments, such as gene array experiments, where results are typically summarized in an expression matrix, in which each element denotes the expression level of a particular gene in a specific sample [1]. In this context two types of distance matrices can be produced: the distances between all pairs of samples can be calculated based on their expression levels over the measured genes, and the distance between all pairs of genes can be measured in the sample dimensions [2]. The sorted distance matrix generated by $SPIN$ is particularly useful in time-series experiments, where an elongated cluster represents the temporal evolution of a particular biological module, such as cell-cycle progression (see figure 1c). Another example where the shape revealed by $SPIN$ has a clear biological interpretation comes from cancer research where samples are often composed of mixtures of cells: for instance, colon tissue samples isolated from liver metastases arrayed into an elongated, ellipsoid cluster [3]. The genes that induced the elongation were characteristic of liver, suggesting that this pattern reflects a mixture of the metastasis samples with cells originating from the liver. Furthermore, the degree of cell mixture in each sample may be deduced from its position in the ordering. Once the genes associated with the contamination are identified in this manner, they can be distinguished from the cancer related genes that the experiment aims at discovering.

[1]Department of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: Ilan.tsafrir@weizmann.ac.il

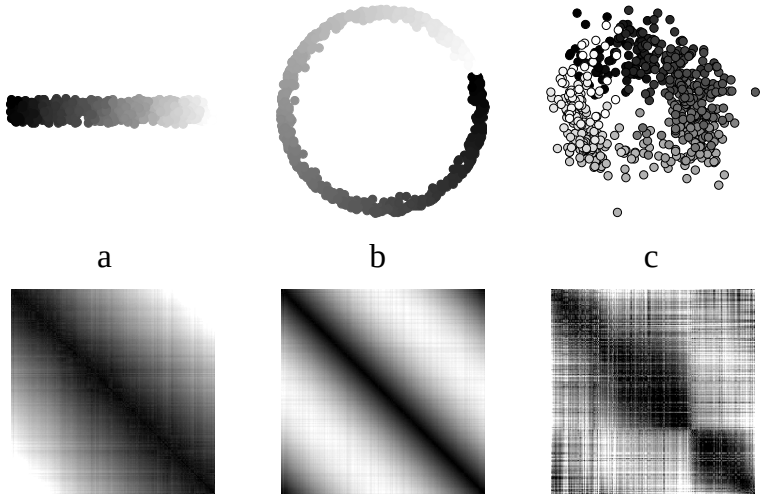


Figure 1: A properly ordered distance matrix is indicative of the shape of a set of points. The color of element D_{ij} reflects the relative distance between points i and j , where black (white) denotes small (large) distances. The top row depicts the placement of the points in two dimensions. Below each object is the corresponding sorted distance matrix. The sorted distance matrix allows a human observer to deduce structural information. The colors of the objects in the top row represent the linear ordering of the points, where the first point is dark ranging to the last point in white. This is the same order that $SPIN$ imposed on the distance matrix, i.e. the first row and first column contain the distances from the first point (black in the PCA image) to all other points. (a) An elongated shape, in this case a cylinder, displays a clear gradient of distances that increase as one moves away from the main diagonal. (b) A ring of points is characterized by a cyclic pattern, with small distances (black) at the corners. (d) $SPIN's$ results for yeast cell-cycle expression data taken from [4], where the cyclic nature is visualized in a ring conformation. Assigning such periodic nature to genes associated with cell-cycle is in accordance with known biological dynamics and functions.

References

[1] M. Eisen, P.Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[2] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745–6750, 1999.

[3] D. Tsafrir, W. Liu, Y. Yamaguchi, I. Tsafrir, Y. Wen, W. Gerald, R. Stengel, F. Barany, P. Paty, E. Domany, and D. Notterman. A novel mathematical approach to analyzing gene expression data: results from an international colon cancer consortium. In *AACR*, 2004.

[4] P.T. Spellman, G. Sherlock, MQ. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297, 1998.

G6. Reproducibility, Variance Stabilization, and Normalization in CodeLink^{TM} Data with Application to Cancer in Rats

Sue Geller[1], David M. Rocke[2], Danh Nguyen[3], Raymond Carroll[4].

Keywords: microarray, variance, transformation, CodeLink^{TM}

Although the experimental aspects of microarray methods are maturing rapidly, the analysis of the array data is still a difficult exercise with many issues that have not been fully resolved in the literature. One of the issues is that, in order to use parametric model-based analysis and even most non-parametric methods, the variance of the replicates of the data needs to be constant across expressions. It was determined by Rocke and Durbin ([4]), that the variance was not constant for spotted microarray data and instead conformed to a model, which had been used in the analytical chemistry and the environmental science literature. This is a two error component model, $y = \alpha + \mu e^{\eta} + \epsilon$, in which there is an additive error that dominates when μ is small and the proportional error that dominates when μ is large. In a subsequent manuscript ([1]), they developed a transformation to stabilize the variance. The applicability of the two component model to Affymetrix data as well as an algorithm for simultaneously finding the constant of the transformation and normalizing the data was given in Geller, Gregg, Hagerman, and Rocke ([2]). In all these (and other) cases, a data set was used that consisted of technical replicates, so the effect of the transformation on variance stabilization is unstudied in sets with multiple technical replicates.

The effect, of lack thereof, of diet on cancer has been a long-standing question. In addition to the usual usual observational trials, the effect of diet in the presence of carcinogens was studied on the genetic level in rats using the CodeLink^{TM} microarray platform ([3]). Three diets (corn oil, fish oil, olive oil) were given with either saline or the carcinogen AOM, and the rats killed at 12 hours or 10 weeks. Twenty two of the 59 rats had two technical replicates, one three technical replicates, and three had four technical replicates, making this data set a useful one for studies of reproducibility and variance stabilization. While many questions remain unanswered, the following are some of the conclusions of our analysis to date.

- The CodeLink^{TM} data from these studies appear to conform in broad terms with the two error component model, $y = \alpha + \mu e^{\eta} + \epsilon$.
- CodeLink^{TM} data has a great deal of variability among technical replicates.

[1]correspondence to Sue Geller, Department of Mathematics – MS 3368, Texas A&M University, College Station, TX 77843-3368 or email to `geller@math.tamu.edu`.

[2]CIPIC, 2343 Academic Surge, University of California, One Shields Ave, Davis, CA 95616

[3]Department of Epidemiology and Preventive Medicine, University of California, Davis, CA 95616

[4]Department of Statistics – MS 3143, Texas A&M University, College Station, TX 77843-3143

- It is possible to transform the gene expressions so that the variance within genes across arrays is approximately the same regardless of the level of expression of the genes. This transformation resembles the logarithm at high expression levels and resembles a linear transformation at low expression levels.
- After transformation, the data for highly expressed genes appear approximately normally distributed, but for low expression levels there are frequent outliers. These may be caused by dust or scratches that have only a small effect on highly expressed genes, but larger proportional effects on genes with low expression.
- The choice of constant for the transformation can be in a broad range and still produce transformed technical replicates with constant variance and symmetric errors. Thus, the same constant can be used for all the data to produce a data set with approximately constant variance regardless of the level of expression of the genes for each set of technical replicates.
- Normalization in conjunction with stabilization of variance is more effective than normalization and then stabilization of variance, i.e., normalization may be incomplete or inaccurate if it is done before transformation.

References

[1] Durbin, B. P., Hardin, J. S., Hawkins, D. M., and Rocke, D. M. 2002. A Variance Stabilizing Transformation for Gene Expression Microarray Data, *Bioinformatics* 18:S105-S110.

[2] Geller, S. C., Gregg, J. P, Hagerman, P., and Rocke, D. M. 2003. Transformation and Normalization of Oligonucleotide Microarray Data, *Bioinformatics* 19:1817-1823.

[3] Ramakrishnan, R., Dorris, D., Lublinsky, A., Nguyen, A., Domanus, M., Prokhorova, A., Gieser, L., Touma, E., Lockner, R., Tata, M., Zhu, X., Patterson, M., Shippy, R., Sendera, R. J., and Mazumder, A. 2002. An Assessment of Motorola CodeLink Microarray Performance for Gene Expreseesion Profiling Applications. *Nucleic Acids Res.* 30 (7):e30.

[4] Rocke, D. M. and Durbin, B. P. 2001. A Model for Measurement Error for Gene Expression Arrays. *J. Comp. Bio* 8:557-569.

G7. ChipQC: Microarray Artifact Visualization Tool

Peter A. Henning[1#], Paul K. Tan[2#], Tung Yu Chu[3], David A. Stiles[3], David Wheeler[4], Pushkar Mukewar[5], Margaret C. Cam[3]*, May D. Wang[1]*

Keywords: Microarray, Quality Control, Error Detection, Web Server

1 Background

Recent advances in biotechnology have led to the development of several high-throughput procedures, such as DNA microarrays, have dramatically increased the amount of information obtained in a relatively short time span[1]. Although experiments using these new techniques are highly automated, downstream analytical methods still lag behind in terms of automation and efficiency[2]. Microarray studies typically investigate gene expression profiling[1, 3], which can be greatly skewed by the slightest amount of contamination. The fabrication or hybridization processes would easily introduce areas of high local variability, which can only be accurately detected when several replicates are analyzed together[4].

ChipQC is a novel web-based software tool that is joined researched and developed by MIBLab at Biomedical Engineering Department of Georgia Tech and Emory University and the Microarray Core Facility at National Institute of Diabetes and Digestive and Kidney Diseases. It is designed to perform standard error analysis and statistical techniques on multiple array sets (technical or biological replicates). Although chip quality control is often overlooked in microarray experiments, it the reproducibility and reduction of systematic error afforded by extensive quality control that will allow microarray systems to migrate from an experimental research tool to a clinical diagnostic device.

2 System Development

ChipQC has evolved from a simple Unix program used to investigating small sections of Affymetrix GeneChips® to a web-based comprehensive microarray analysis tool that can employed on any of six supported chip types, both commercial and "homemade". ChipQC include calculating the coefficient of variation, standard deviation, mean intensity, fold change, and statistical significance. ChipQC, using a heat map scheme, graphically represents the error analysis and various metrics of each gene at its proper chip coordinates, thus mimicking the analyzed chip's configuration. Use of this visualization tool revealed localized areas of high variability pattern in some arrays consistent with an edge effect, which persisted even after lowess or linear normalization had been applied, demonstrating the need to select and flag specific spots which would potentially yield falsely

[1] Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, 313 Ferst Dr., Atlanta, Georgia, USA. Email: maywang@bme.gatech.edu

[2] John A. Burns School of Medicine, University of Hawaii, Manoa, Hawaii, USA. Work done while at NIDDK Microarray Core Facility.

[3] Microarray Core Facility, National Institute of Diabetes Digestive and Kidney Disorders (NIDDK), National Institutes of Health, Building 8, Bethesda, Maryland, USA. Email: MaggieC@intra.niddk.nih.gov

[4] National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Building 38, Bethesda, Maryland, USA.

[5] School of Biology, Georgia Institute of Technology, Atlanta, Georgia, USA.

[#] These authors contributed equally.

[*] Authors whom correspondence should be addressed.

positive or negative gene expression results. The latest version of the software still contains the orginial Perl core but has been enhanced by C and Java programs, which put common normalization schemes and the ability to create histrograms just a click away. The heatmap image generation is made possible by the Boutell GD Perl module (www.boutell.com). There are six key steps contained in the ChipQC workflow (Figure 1).

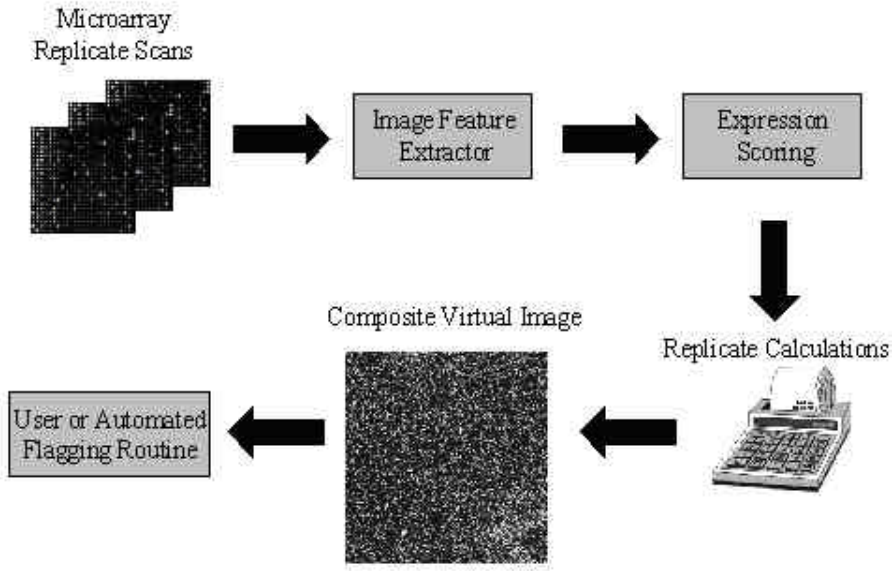


Figure 1 ChipQC System Diagram

3 Future Work

The composite image has been used to identify systematic errors presented such as edge effect. While we are in the process of reporting such discovery in full detail to the technical community, we have been working to add other features such as more normalization options, statistical options, and more supported chip types to the existing software. The hope is to also pursue more substantial improvements by implementing the latest image processing methods to enhance image feature extraction, expression scoring, and possibly the automated flagging routine.

References

1. Schena, M., et al., *Quantitative Monitoring of Gene-Expression Patterns with a Complementary-DNA Microarray*. Science, 1995. **270**(5235): p. 467-470.
2. Holloway, A.J., et al., *Options available - from start to finish - for obtaining data from DNA microarrays.* Nature Genetics, 2002. **32**: p. 481-489.
3. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nature Biotechnology, 1996. **14**(13): p. 1675-1680.
4. Slonim, D.K., *From patterns to pathways: gene expression data analysis comes of age.* Nature Genetics, 2002. **32**: p. 502-508.

G8. A Database Aiding Probe Design System for Virus Identification

Feng-Mao Lin[1], Pak-Leong Chan[2], Yu-Chung Chang[3], Hsien-Da Huang[4], Jorng-Tzong Horng[5]

Keywords: probe design, microarray, database system, virus

1 Introduction.

To identify viruses causing disease becomes more and more important today. If viruses had been identified early, virologists would have had sufficient time to treat them. However, a few systems have been developed previously to design probes to identify virus sequences [1-5]. User can select all the virus-specific oligonucleotide probes under the group of viruses. However, user cannot select viruses of different groups. A system for identifying different host categories of viruses in reasonable time is crucial. Probe design for virus sequence identification is very time-consuming, especially in avoiding the cross-hybridization of designed probes against the non-target sequences. We propose a faster algorithm to implement the program of LIS [6]. We apply LIS algorithm to calculate the similarity between each pair of probe and non-target sequences. Our algorithm is faster than the method of using BLAST only. In addition, we make use the database technology aiding for designing an oligonucleotide microarray that can identify virus sequences selected by users.

2 Materials and methods.

We downloaded two kinds of data form different databases. One is virus taxonomy data from the universal virus database of the international committee on taxonomy of virus (ICTVDB) [7]. Another is the virus sequence from NCBI GenBank database. Finally, we integrated virus taxonomy data and virus sequence data to our local database.

The process of probe design contains two sections. They are probe candidates generation and probe selection. A virus sequence was divided into many fragments by sliding a window with 5 nucleotides a time along the whole virus sequence. The size of window is ranged at 20 to 60 nucleotides. Sequence fragments will be stored in probe candidate table in our local database, if and only if the sequence fragment satisfies all the following criteria [8]:

1. Number of any single base (As, Cs, Ts or Gs) does not exceed half of the fragment length.
2. The length of any contiguous As, Cs, Ts, or Gs does not exceed a quarter of the fragment length.
3. GC-content of sequence fragment is in the range of 40% to 60%.
4. No self-complementary is within the sequence fragment.

[1] Department of Computer Science and Information Engineering, National Central University, Taiwan. E-Mail: meta@db.csie.ncu.edu.tw

[2] Department of Computer Science and Information Engineering, National Central University, Taiwan. E-Mail: leong@db.csie.ncu.edu.tw

[3] Department of Biotechnology, Ming Chuan University, Taiwan. E-mail: d80106@mcu.edu.tw

[4] Department of Biological Science and Technology & Institute of Bioinformatics, National Chiao-Tung University, Hsin-Chu, Taiwan. E-Mail: bryan@mail.nctu.edu.tw

[5] Department of Computer Science and Information Engineering and Department of Life Science, National Central University, Taiwan. E-Mail: horng@db.csie.ncu.edu.tw

3 Results.

The system was built under Linux Red Hat 9.0 and MySQL 3.x. The LIS program was implemented in c programming language. It was compiled with a Gnu-compiler and ran on Linux workstation. The web interface was implemented in PHP. Virus sequences can be selected in a web page. As the user selects the virus sequences, inputs the temperature and length of probe, the system will select all the probes candidates belonging to the virus sequences selected by the user from the probe candidate database.

To verify probe with low LIS-identity is specific, we selected 7 species virus under a genus (coronavirus) from virus sequence database. The length of probe was set to 50mer and the temperature threshold was set between 75^{o}C and 85^{o}C. 9,097 probes candidates were selected from probe candidate database. The LIS-identity of each probe was calculated against its non-target sequence. We defined cross-hybridization as a probe that is large than 70% similarity to its non-target sequence. Table 1 shows the number of probes and the percentage of cross-hybridization of every 5 LIS-identity. The lower the LIS-identity is, the less possibility the probe has cross-hybridization to its non-target sequences. Although the LIS-identity is approximately proportional to the similarity of probe to its non-target sequence, there are still some cases that the LIS algorithm cannot calculate the similarity of probe to its non-target sequence accurately.

LIS-identity	Number of data	Identity > 70%	Possibility of being cross-hybridization
10~15	31,420	134	0.42%
15~20	28,653	627	2.1%
20~25	4,031	1,133	28%
25~30	1,932	1,267	65%
30~35	947	879	92%
35~40	453	453	100%
40~45	182	182	100%
45~50	41	41	100%

Table 1. Show the data distribution of each scope of LIS-identity and the correlation between LIS-identity and cross-hybridization.

References

[1] Chang, P. C. and Peck, K. 2003. Design and assessment of a fast algorithm for identifying specific probes for human and mouse genes. *Bioinformatics* 19:11 1311-7.
[2] Kaderali, L. and Schliep, A. 2002. Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics* 18:10 1340-9.
[3] Kane, M. D., Jatkoe, T. A., Stumpf, C. R., Lu, J., Thomas, J. D., and Madore, S. J. 2000. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res* 28:22 4552-7.
[4] Rouillard, J. M., Herbert, C. J., and Zuker, M. 2002. OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* 18:3 486-7.
[5] Wang, X. and Seed, B. 2003. Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics* 19:7 796-802.
[6] Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., and Salzberg, S. L. 1999. Alignment of whole genomes. *Nucleic Acids Res* 27:11 2369-76.
[7] Buechen-Osmond, C. and Dallwitz, M. 1996. Towards a universal virus database - progress in the ICTVdB. *Arch Virol* 141:2 392-9.
[8] Li, F. and Stormo, G. D. 2001. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* 17:11 1067-76.

G9. Quickly Choosing Choice SNPs for Chips

Earl Hubbell, Teresa Webster, Hajime Matsuzaki [1]

Keywords: SNPs, greedy, optimization

1 Introduction.

Oligonucleotide microarrays are currently being used for genotyping of SNPs (Liu, et al, 2003). Given the millions of SNPs estimated to exist (Kruglyak and Nickerson, 2001) and the large subset already in databases (Thorisson and Stein, 2003), we wish to prune this number down to a number that will fit on a few microarrays at current feature sizes. After bioinformatic and biochemical constraints are applied, we had an empirical pool of approximately 150,000 SNPs with known allele frequencies, call rates, and genomic locations, from which to select 100,000 SNPs to cover the genome.

We constructed an entropy-based measure of the value of each SNP, which was then combined with the expected decay of information with distance along the genome to approximate the information provided by each SNP for locations in the genome. Inspired by the set cover problem, a simple greedy algorithm was used to choose the SNPs that covered the genome with the most information, combining spacing and allele frequency.

2 Problem formulation

SNP Selection Problem: Given N SNPs, each of which has a genomic location, allele frequency, and call rate (proportion of time SNP information can be detected from a sample), choose the K SNPs that provide the most uniform information across the genome.

Multiple Enzyme SNP Selection Problem: Given N SNPs as above, where each SNP is assigned to one of L enzyme classes, choose the $K_1, K_2, \ldots, K_L$ best SNPs providing the most uniform information across the genome.

How do we define the information provided by a SNP about a location? There are several options, and we choose to measure the information in bits, using a channel capacity formulation. SNPs occur with a given allele frequency f, are linked with a nearby location with a recombination fraction r, are called with a frequency q, and have accuracy a. Using the Haldane recombination formula $r = (1-exp(-2*d/s))/2$, we can convert genetic distance d to recombination fraction, with a scale for decay of information set by s. From this, and standard measures of mutual information, we obtain that a SNP provides approximately

$$I = q * [((1-re) * log(1-re) + re * log(re) - ae * log(ae) - (1-ae) * log(1-ae))] \quad (1)$$

bits of information about a location d away, where $re = (1-r) * (1-a) + r * a$ and $ae = f * (1-re) + (1-f) * re$.

A nice feature of the information measure is that by changing the genomic scale over which information decays we can adjust the relative weighting of allele frequency and spacing between SNPs.

[1] Affymetrix, Santa Clara CA. E-mail: Earl_Hubbell@affymetrix.com

3 Solving the Problem

We use a natural greedy selection criterion for a SNP. Pick the SNP that provides the least redundant information about the genome, where redundant information is defined by $o - I$, where o is the information provided by any already selected SNPs about the genomic location of the chosen SNP.

With this criterion using a priority queue, we choose SNPs from a pool to add to one or more chips. In the multiple enzyme case, we set up one priority queue for each chip and cycle through the chips, choosing one SNP per chip to ensure that no chip gets only the lowest priority SNPs.

In practice, SNPs are very much more clumped than items drawn from a random distribution, and there is a large fraction of low allele frequency SNPs, which implies that even with good optimization it is difficult to produce an evenly spaced distribution of high minor allele frequency SNPs.

4 References and bibliography.

Thanks to the whole genotyping team, including Giulia Kennedy, Simon Cawley, Rui Mei, Jing Huang, Guoying Liu, Geoffrey Yang, Xiaojun Di, and Keith Jones.

References

[1] Brookes, A.J. 1999. The essence of SNPs. *Gene* 234:177-186.

[2] Liu, W.-m., X. Di, G. Yang, H. Matsuzaki, J. Huang, R. Mei, T.B. Ryder, T. A. Webster, S. Dong, G. Liu, K.W. Jones, G.C. Kennedy, and D. Kulp. 2003. Algorithms for Large Scale Genotyping Arrays *Bioinformatics* 19:2397-2403.

[3] Thorisson, G.A. and L.D. Stein. 2003. The SNP Consortium Website: past, present and future. *Nucleic Acids Res* 31:124-127.

G10. Genome-wide statistical analysis of gene coexpression: application to GATA transcription factors in *Arabidopsis thaliana*

Chih-hung Jen[1] and David Robert Westhead[2]

Keywords: *Arabidopsis thaliana*, GATA factor, microarray, coexpression analysis

1 Introduction

Identifying the coexpressed genes from the microarray data can be used to assign potential functions to new genes and help the discovery of transcriptional regulation networks [1], [2]. Currently, the coexpressed genes are usually analysed by many sophisticated clustering algorithms e.g. SOM, hierarchical clustering, k-mean clustering. However, these clustering approaches usually depend on the distance cut-off value or arbitrary k value to group the genes, and these criteria do not really indicate the significance of the similarity within the clusters. Besides, they assign particular genes to only one cluster that may cause loss the information where genes may have multiple biological roles or respond to different transcription factors.

In order to identify the *in vivo* potential targets of *Arabidopsis* GATA family transcription factors [3] using microarray data and avoid the drawbacks of clustering algorithms, we propose a novel robust approach of assessing the significance of relationships in expression. This approach explores the probability distribution of correlation scores that occur by chance between the GATA gene and unrelated gene expression profiles. Based on the stable correlation score distributions, the correlation scale (p-values) associated with the corresponding correlation distance values can be estimated. Furthermore, for different numbers of probe sets on the array, the e-value can also be calculated. The p-value and e-value of a correlation value can help us discover potential GATA gene co-expressed genes with confidence and will be useful indicators for the further researches and applications.

2 Methods

The microarray datasets we used consist of 59 Affymetrix arrays (ATH1) for 14 different experimental purposes and were obtained from the NASC [4]. The estimation of the correlation distribution for the GATA gene is achieved through the three iterative steps.

Step1: Random sampling. A gene expression profile is randomly sampled from the actual expression profiles.

Step2: Randomisation. Shuffled the order of the arrays, and added pseudo-Gaussian distributed random noise.

Step3: Correlation calculation. The r-value (Pearson correlation coefficient) between the GATA gene and randomised profiles is calculated.

The procedure is executed $5*10^5$ times to simulate the correlation distribution of unrelated profiles which can cover all currently possible size of arrays. Each simulation was repeated three times and the stability of the empirical distribution was assessed by Kolomogorov confidence bands [5]. All possible factors that could affect the correlation distribution were also examined, such as using

[1]Bioinformatics Research group, School of Biochemistry and Molecular Biology, University of Leeds, U.K. E-mail: bmbcj@bmb.leeds.ac.uk

[2]Bioinformatics Research group, School of Biochemistry and Molecular Biology, University of Leeds, U.K. E-mail: D.R.Westhead@leeds.ac.uk

different number of arrays, different size of the arrays, difference experimental datasets, and different GATA genes.

3 Results

The correlation distributions of all repeated simulations were within Kolomogorov confidence bands and imply that the distribution simulation process is stable and robust. The shape of the correlation distribution is only affected by the number of arrays (Figure 1). Based on the p-value derived from the correlation distribution simulation, we can reveal the GATA coexpressed genes with confidence. An example result of the top ten genes coexpressed with GATA-2 is shown in Table 1.

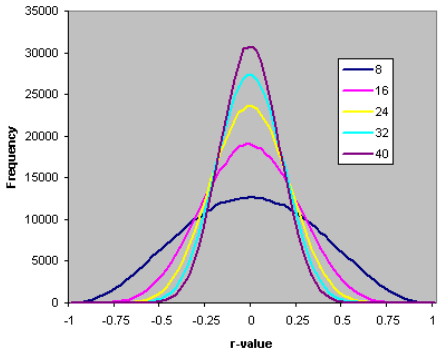


Figure 1: Distribution of the correlation coefficient in random, unrelated profiles using different numbers of arrays.

Probe ID	Gene ID	Pearson CC	P value	E value	Annotation (from Affymetrix)
266125_at	AT2G45050	1.0000	0	0	GATA zinc finger protein
254100_at	AT4G25020	0.9463	0.00011908	2.70859368	expressed protein
266310_at	AT2G26990	0.9445	0.0001342	3.0525132	COP9 complex subunit CSN2, putative
260000_at	AT1G68060	0.9429	0.00014764	3.35821944	expressed protein
246085_at	AT5G20540	0.9370	0.0001972	4.4855112	expressed protein
266407_at	AT2G38560	0.9361	0.00020476	4.65747096	elongation factor -related
265789_at	AT2G01210	0.9193	0.00039832	9.06018672	leucine-rich repeat transmembrane protein kinase, putative
262519_at	AT1G17160	0.9144	0.00048456	11.02180176	pfkB type carbohydrate kinase protein family
246995_at	AT5G67470	0.9136	0.00049864	11.34206544	formin homology 2 (FH2) domain-containing protein
247461_at	AT5G62100	0.9061	0.00063064	14.34453744	BAG domain containing protein
253204_at	AT4G34460	0.9034	0.00067816	15.42542736	transducin / G-protein beta-subunit (AGB1)

Table1: Top ten genes are correlated to *Arabidopsis* GATA-2 gene.

References

[1] D'haeseleer, P., Liang, S., and Somogyi, R. 2000. Genetic Network Inference: From Co-Expression Clustering to Reverse Engineering. *Bioinformatics*, 16, 707-726.

[2] Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders, R., Brazma, A., Holstege, F. C. 2002. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell*, 9, 1133-1143.

[3] Teakle, G.R., Manfield, I.W., Graham, J.F., and Gilmartin, P.M. 2002. Arabidopsis thaliana GATA factors: organisation, expression and DNA-binding characteristics. *Plant Mol Biol.*, 50, 43-57.

[4] Craigon, D.J., James, N., Okyere, J., Higgins, J., Jotham, J., May S. 2004. NASCArrays: A repository for Microarray Data generated by NASC's Transcriptomics Service. *Nucleic Acids Res.*, 32, D575-D577.

[5] Hayter, A. J. 1996. *Probability and statistics for engineers and scientists*. London : PWS Pub. Co.798-806.

G11. Using the Human Genome as a Framework for Sequence Clustering and Microarray Design

Barbara Lin[1], Timothy Burcham[1]

Keywords: human genome, BLAT, sequence, alignment, microarray design

1 Introduction.

At diaDexus, Inc., we utilize genomic and bioinformatic tools to identify and validate cancer-associated molecular targets for diagnostic and therapeutic applications. These gene discovery efforts generate large amounts of sequence data on a daily basis, which poses significant challenges to organize and eliminate redundancy between sequences. One approach commonly used to group similar sequences is to BLAST the sequences against one another. However, BLAST results can be hard to interpret and difficult to reproduce, and may be inconclusive with large sequence sets. With the completion of the human genome and its sequence information publicly available, a reliable, stable framework now exists and can be utilized to classify and structure large sequence sets. We have used the genome as a means of organizing sequence data, and have developed an algorithm to quickly isolate non-overlapping clusters of sequences. These clusters allow us to quickly find the most prevalent forms of sequences that are of interest to disease target discovery, and also enable a systematic approach for oligonucleotide microarray design.

2 Software and Method.

Using the human genome (Build 30-34) as a template, we developed a process to identify a unique set of sequences from a highly redundant set of sequences. This process required an ability to identify the highly similar sequences, but not identical, sequences in an efficient and reproducible manner. We used the resulting data to identify a high quality, non-redundant set of oligonucleotides to print on microarrays for gene expression profiling.

First, approximately 263,000 sequences of specific interest to target discovery, each having different quality and length, and generated from different sources and methods, were aligned to the human genome using BLAT[2]. Out of the 263,000 sequences, about 16,000 sequences were from the Human Refseq mRNA database, 23,000 were Ensembl genes, 129,000 were from UniGene, and 95,000 were either generated as a result of internal sequencing efforts or from other proprietary sources. Sequences were allowed to map to more than one location on the genome and the coordinates of each mapping were stored in a relational database for downstream processing.

After the BLAT alignment, we developed and implemented an algorithm to cluster sequences based on their exact coordinates and locations on the genome. In the first step in the algorithm, "chromosome walking", we iterate through every chromosome and walk each chromosome from one end to another, in both orientations, defining and separating stretches of overlapping sequences into individual "super-clusters". Sequences are grouped together if their beginning and ending chromosomal coordinates overlap one another. After chromosome walking, we were able to very rapidly reduce the complexity of the set of the 263,000 sequences, and further group the sequences into approximately 80,000 bins or super-clusters.

[1] diaDexus, Inc, 343 Oyster Point Blvd, South San Francisco, CA 94080. E-mail: tburcham@diadexus.com

After the first step, many super-clusters contained sub-clusters which had no overlap with the parent clusters. For example, a small sub-cluster of sequences would be intronic to a super-cluster whose chromosomal coordinates encompassed the sub-cluster. The second step of the algorithm, "exon walking", was used to identify all sets of truly overlapping sequences from the results generated by chromosome walking, based on the coordinates of the exons within each super-cluster.

Using exon walking, sequences with non-overlapping exons are separated out of the chromosome clusters and given their own bin or cluster. This is accomplished by taking all the sequences in a super-cluster, and for each sequence, creating a virtual sequence in which every nucleotide is characterized as a bit. The nucleotides that are within an exon are represented as 1's, and those that fall outside exons are represented as 0's. By imposing a logical bitwise AND relationship between the virtual bit sequences, we were able to quickly and efficiently identify the sequences with overlapping regions. Approximately 90,000 bins or clusters resulted after exon walking, a number significantly less than the 263,000 we started out with.

We were then able to use the resulting clusters to identify a non-redundant, high quality set of oligonucleotides to print on microarrays. About 38,000 pre-designed oligonucleotides[1] from proprietary sequences were melded into the clusters obtained from the method described above. We then prioritized the clusters them based on a combination of factors. The criteria that we looked for in characterizing these clusters were:

1. Clusters with a public annotated gene, transcript, or EST from RefSeq, Ensembl, or UniGene.
2. Clusters with multiple exons.
3. Clusters with multiple sequences (non-singleton clusters).
4. Clusters containing oligos with measured, disease-specific, expression in our in-house expression profiling studies.

Using the criteria above, we identified 19,000 "high priority" cluster bins. After the high priority clusters were selected, we selected the best, most representative, oligo from each cluster for array design. This algorithm will be discussed in more detail in the poster, but generally we printed the most 3′ oligo that most generally represented the sequences in the cluster, and for which we had good previous expression results. These oligos were then printed onto several custom oligonucleotide microarrays and are currently in use for cancer-target discovery and validation.

In conclusion, we were able to use the genome a powerful framework to organize and efficiently cluster a large sequence set into individual sequence clusters. Using these clusters, we were able to design high quality microarrays with minimally redundant oligonucleotide sequences.

3 References.

[1] Hughes, T.R., *et al.*, 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* 19: 342-347.

[2] Kent, W. J. 2002. BLAT —The BLAST-Like Alignment Tool. *Genome Res.* 12: 656-664.

G12. Programs for the Inference and Analysis of Gene Influence Networks

Gary Livingston,[1] Liwu Hao,[2] Guangyi Li,[3] and Xiao Li[4]

Keywords: gene expression analysis, microarray data analysis, pathway inference, pathway analysis

The induction and analysis of gene networks generated from microarray gene expression data may greatly aid the understanding and cure of gene-based diseases such as cancer. We have devised a novel set of programs for (1) inducing gene networks from gene expression data that are based upon rule induction and (2) analyzing the induced networks. The first program, InduceNet, induces networks from gene expression data; CompareNet performs an edgewise comparison of two networks; and ComparePop uses a given network to compare the "fit" of the network to two populations. We demonstrate the utility of our programs by using them to analyze a lung cancer dataset formed by merging data from Beer et al. [1] and Bhattacharjee et al. [2].

1 Programs

InduceNet uses a discovery program called HAMB [4] to generate rules predicting the expression levels of each gene by using the expression levels of other genes in the dataset. Rule sets using one gene to predict another with an accuracy (proportion of correct predictions) exceeding a given threshold are used to form the edges of the induced network.

CompareNet creates a "difference network" from two networks by comparing the edges in the two networks, thus creating a new network comprised of the edges that differ between the networks as follows: edges that appear in the first network but not in the second, edges that appear in the second network but not in the first, and edges that appear in both networks but with different correlations (e.g., a positive correlation in the first network and a negative correlation in the second).

ComparePop compares the "fit" of each edge of a reference network to each set of data by using a standard Z-test to compare the proportions of cases from each set of data that are correctly predicted by the rule family used to infer the edge. The resulting Z-score may then be used to compute the p-value of the "fit difference." If the p-value is sufficiently small, one would reject the possibility that the fit difference could have occurred by random variation in the sampling process and therefore accept that the edge probably represents a difference in the populations from which the data were obtained. Thus, ComparePop can be used to identify which gene interactions are likely to be particular to cancer by identifying the edges in the network that differ significantly between cancer cases and non-cancer cases.

CompareNet and ComparePop are not limited to analyzing networks generated by InduceNet; they may be used to analyze gene networks generated in any fashion. For example, CompareNet could be used to compare an induced network to a manually derived network, such as a known gene network. We have devised a simple program that can accept a manually created network and

[1] Dept. of Computer Science, Univ. of Massachusetts—Lowell, Lowell, MA. Email: gary@cs.uml.edu
[2] Dept. of Computer Science, Univ. of Massachusetts—Lowell, Lowell, MA. Email: lhao@cs.uml.edu
[3] Dept. of Computer Science, Univ. of Massachusetts—Lowell, Lowell MA. Email: lucyligy@yahoo.com
[4] Dept. of Computer Science, Univ. of Massachusetts—Lowell, Lowell MA. Email: xiaolee88@yahoo.com

generate the rule families needed for ComparePop, which allows it to accept and evaluate manually created networks representing hypotheses devised by biologists.

2 Results

We used InduceNet to induce a gene influence network using cases from our combined gene expression dataset that were from stage 2, 3, or 4 tumors. We then used ComparePop to identify the edges in this network having associated rule families with proportions of correct predictions that varied significantly between non-tumor cases from the combined dataset and cases from patients with stage 2, 3, or 4 tumors.

When we examined the gene interactions identified in our inferred network, many genes were found to belong to lists of oncogenes obtained from the World Wide Web: GAS6 and LAMB1 are in the AKT group of genes, FGF7 belongs to the WNT group of genes, PF4 is a member of the TGF-B group of genes, and IL11 is a member of the MEK group of genes.

We also found several structures contained in the inferred network that are similar (structurally and ontologically) to some of the structures of the cancer pathway presented in [3]:
PF4 ← FGF7 → CD36; FABP4 ←LMO2 → FOXF1 ←COL9A3; GAS1 and LAMB1 → SH3GL2 ← ABCC2; FBLN2 → CDH11 ← CTSK? →SPARC ← LUM →COL9A3 ← GAS6; and GATA → IL11 → SPP2 → TNNT.

We found several genes not known to be oncogenes that are in one or more edges of our inferred difference network, suggesting that these genes may be biologically related to lung cancer. Each additional edge in the network that a gene belongs to dramatically increases the probability that the gene is biologically related to lung cancer. The gene FBN1 occurred in 7 of the edges; COL5A2 and KCNAB1 each occurred in 5 of the edges; THY1 occurred in 4 of the edges; HSU15552 and COL3A1 each occurred in 3 of the edges. Eleven more genes occurred in 1 edge each.

3 References

[1] Beer, D., Kardia, S., Huang, C., Giordano, T., Levin, A., Misek, D., Lin, L., Chen, G., Gharib, T., Thomas, D., Lizyness, M., Kuick, R., Hayasaka, S., Taylor, J., Iannettoni, M., Orringer, M., and Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nature* 8(8): 816–824.

[2] Bhattacharjee, A., Richards, W. G., Staunton J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, [3] D. J., and Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, *Proceedings of the National Academy of Science USA* 98(24): 13790-5.

[3] Hahn, W., and Weinberg, R. A. (2002). A subway map of cancer pathways. http://www.nature.com/nrc/journal/v2/n5/weinberg_poster/.

[4] Livingston, G., Rosenberg, J. and Buchanan, B. (2003). An Agenda- and Justification-Based Framework for Discovery Systems, *Journal of Knowledge and Information Systems* 5(2), in press.

G13. Analysis of Microarray Time Course Data

Tanya Logvinenko[1], **David Schoenfeld**[2],**Douglas Hayden**[3]

Keywords: Microarray Data, Time Course Data, Analysis of Variance, Two-Sample Comparison, Cluster Analysis, Fold Change.

1 Introduction

With the development of microarray technology, it became possible to study biological processes over time. For example, of interest can be changes in gene expression profile over time. In case of a disease, discovering which genes are induced or repressed at different stages can help in the understanding of the disease process, as well as in the development of the treatment procedures. We will address a problem of analyzing time-course microarray data. The data that will be used is part of the Inflammation and the Host Response to Trauma collaborative project (`www.gleugrant.org`). It was obtained from eight healthy human volunteers who were randomized into two groups of four individuals each. One of the groups was exposed to *in-vivo* LPS (lipo-polysacharide stimulation), whereas the other group was administered a placebo treatment. Six microarrays per individual were obtained over the course of one day. We will study the efficiency of different methods of time-course microarray data analysis applied to this data set.

2 Methods

A variety of methods for analysing time-course microarray data were developed recently. To detect differentially expressed over time genes, one can apply

[1]Division of Biostatistics, Massachusetts General Hospital, 50 Staniford St, Suite 560, Boston, MA 02114. E-mail: `tlogvinenko@partners.org`

[2]Division of Biostatistics, Massachusetts General Hospital, 50 Staniford St, Suite 560, Boston, MA 02114. E-mail: `dschoenfeld@partners.org`

[3]Division of Biostatistics, Massachusetts General Hospital, 50 Staniford St, Suite 560, Boston, MA 02114. E-mail: `doug@gcrc.mgh.harvard.edu`

standard statistical models like analysis of variance [3] to expression profiles of each genes, or examine fold changes of gene expressions over time. A statistical technique, *SAM*, for discovering genes differentially expressed between different samples was developed [5]. Modified to account for the higher degree of similarity between observations obtained from the same subject, rather than from different subjects at the same time point, it can be applied to time-course microarray data. A whole range of clustering procedures can be used for discovering genes with the same expression profiles over time [1, 2, 4]. In cases when differential expression of genes is caused by a known factor (such as a disease or a trauma), a control sample can be examined. In such situations, we propose directly comparing distributions of the gene expressions from the two groups. We will examine a few techniques and models applied to time-course microarray data, and present their advantages and disadvantages.

References

[1] Bar-Joseph, Z., Gerber, G., Gifford, D., Jaakkola, T., and Simon, I. (2003). A new approach to analyzing gene expression time series data. *Proc. Natl. Acad. Sci. USA*, 100:10146–10151.

[2] Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868.

[3] Kerr, M., Martin, M., and Churchill, G. (2000). Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, 7:819–837.

[4] Tseng, G. and Wong, W. (2003). Tight clustering: A resampling-based approach for identifying stable and tight patterns in data.

[5] Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, 98:5116–5121.

G14. GOArray: Interpreting microarrays with GODB

Michael V. Osier[1], David Tuck[2]. Kevin P. White[3], Christopher E. Mason[3], Hongyu Zhao[4], Kei-Hoi Cheung[1]

Keywords: microarray interpretation, Gene Ontology Database, inference, permutation

1 Introduction.

Once a researcher has performed a microarray experiment and determined which genes are differentially expressed, the process of interpretation begins. For the rare case where only a few genes appear relevant, this can be an easy task. When dozens or hundreds of genes are differentially transcribed, however, it might be best to allow computers to reduce the effort of elucidation. Several analysis tools have been implemented to do this in the context of the terms in the Gene Ontology Database (GODB, http://www.godatabase.org/dev/database/). GODB is a rooted directed acyclic graph (rooted-DAG) of terms which represent a biological concept. Some of the tools that interpret arrays in the context of GODB score by "strict association" in which a GO term is scored purely by which genes are directly associated with it. Other tools score by "inclusive association" in which a term is scored by which genes are directly associated with it or one of its child terms.

A problem faced by both scoring methods is how to correct for the large number of statistical tests that must be performed in the analysis. The Bonferoni correction, multiplying the p-value by the number of tests performed, is overly-conservative to the point of being an impediment since few, if any, terms will ever be significant. Untangling the many interdependencies of the rooted-DAG of terms in GO to find an appropriate correction, however, is impractical.

As a way of assessing confidence in the results of the analysis from our tool GOArray, formerly named GOMine [1], which uses an inclusive association analysis of microarray data in the context of GODB, we have implemented two tests based on a permutation of the microarray data: a False Detection Rate to estimate how many significant terms we would expect on average, and a Confidence Test of how frequently we observe permuted arrays to have at least as many significant terms as the observed data.

2 Methods.

GOArray uses an inclusive association algorithm to identify genes associated with a term in GODB. The statistical analysis for each term tests if, among the genes associated with this term or one of its descendants, there is an overrepresentation of genes considered of interest (Genes Of Interest, GOI) relative to none GOI (NGOI). Note that determination of GOI and NGOI is made by an outside source, allowing the researcher to answer the question of "what is a gene that is differentially expressed?" in a manner appropriate for their experiment. A z-score and a p-value are

[1] Yale Center for Medical Informatics, Yale University School of Medicine, New Haven, Connecticut. E-mail: michael.osier@yale.edu
[2] Dept of Pathology, Yale University School of Medicine, New Haven, Connecticut
[3] Dept of Genetics, Yale University, New Haven, Connecticut
[4] Dept of Epidemiology and Public Health, Yale University, New Haven, Connecticut

calculated for each term based on the overall frequency of GOI [1]. Terms with a p-value less-than-or-equal-to a user-defined cutoff value are reported.

Permutations of the GOI are then generated, and statistics calculated for each permutation. For each permutation, the number of terms with a p-value less than or equal to the cutoff are counted (T_f). For the real dataset, the same statistic is calculated (T_r). The False Discovery Rate (FDR) is determined as the mean T_f divided by the T_r. The Confidence Test (CT) is the number of permutations where the T_f is greater than or equal to the T_r divided by the number of permutations. Together, these two tests give a feel for how confident one can be in the results of the analysis.

In the output of GOArray, the list of significant terms, FDR, and CT are reported as are two tree representations of the significant terms and an archive of how all statistics are calculated. An HTML format is used so that a) results are viewable in the future, minimizing the chance of data loss when application software is retired, and b) GOArray can be easily harnessed by a web interface in the future. GOArray is available from the web site "http://ycmi.med.yale.edu/gomine/".

3 Results and Discussion.

We have used GOArray to analyze a Drosophila expression dataset using a cutoff p-value of 0.001 and 1000 permutations. The data was taken from the Arbeitman et al. [2] stage 0-1 hours results, available from the NCBI database GEO (http://www.ncbi.nlm.nih.gov/geo/) with the accession GSM3612. All genes with at least a five-fold increase in expression were marked as GOI (1374 spots), and all other genes as NGOI (7425 spots). On a 2.4 GHz Xeon processor, the analysis took ~10 min with permutations and ~0.5 sec without. The FDR was high (51.4%) and the confidence test was marginal (10.7%). In contrast to what one would expect based purely on the FDR and CT results, however, the terms determined to be significant appear to be biologically appropriate. Multiple terms were related to rapid cell replication (e.g. "DNA replication and chromosome cycle", "S phase of mitotic cell cycle", "pre-replicative complex", etc.). The only unexpected term was "leucyl aminopeptidase activity" (p=0.00007). The biological role of this metalloexopeptidase activity is uncertain. It may be involved in the modification of signaling peptides, or it could be a false positive. A closer examination of the specific GOI that resulted in this high z-score would be interesting. Given the high percentage of significant terms that seemed biologically appropriate, it may be that the FDR and CT are themselves overly conservative measures. Alternatively, it may be that other, more statistically robust, means of separating GOI from NGOI may result in a higher confidence in the results. Indeed, we are examining methods to maximize the power of GOArray. The HTML formatted results of this analysis are available from the web site "http://microarray.yale.edu/ymd_public/white_science_ratio5.html".

References

[1] Osier, M.V., Tuck, D., White, K.P., Mason, C.E., Zhao, H., and Cheung, K.H. GOMine – a model for microarray interpretation. *Submitted*.
[2] Arbeitman, M.N., Furlong, E.E., Imam, F., Johnson, E., Null, B.H., Baker, B.S., Krasnow, M.A., Scott, M.P., Davis, R.W., White, K.P. 2002. Gene expression during the life cycle of Drosophila melanogaster. *Science* 297:2270-2275.

G15. SVM Model Selection for Microarray Classification

David A. Peterson[1]

Keywords: support vector machine, model selection, microarray, lymphoma

1 Introduction.

Support vector machines (SVMs) [4] are gaining broad acceptance as state-of-the-art classifiers for microarray data analysis [3]. However, most studies that use SVMs to predict sample class consider only a small subset of SVM kernels and parameters. The effect of the kernel type and parameter values is usually not studied in microarray classification. The choice of kernel and classifier parameters is a form of model selection. Although the machine learning community has extensively considered model selection with SVMs [2], optimal model parameters are generally domain-specific. The present study evaluates the impact of kernel type and parameter values on the accuracy with which a SVM can classify microarray data. We hypothesized that classification accuracy would vary with the kernel type and parameter values, and that the optimal parameter values would vary with the kernel type.

2 Methods.

We evaluated linear and non-linear SVMs. Two non-linear SVM kernels were tested: second order polynomial and radial basis function (RBF). In each case, the regularization parameter C was varied over [1 10 100 1000 10000]. For the polynomial kernel, gamma serves as inner product coefficient in the polynomial. In the case of the RBF kernel, gamma determines the RBF width. In both cases, gamma was varied over [0.00001 0.0001 0.001 0.01 0.1 1]. Classification accuracy was measured as the mean accuracy on held-out test data from 30 iterations of stratified 10-fold cross-validation. We used the lymphoma gene expression data reported in [1].

3 Results.

For the linear SVM, variations in C did not produce statistically significant difference in performance. For the polynomial kernel, only a conjunction of low gamma and low C values produced substantially degraded performance. Either high values of gamma or high values of C produced optimal classification accuracy. For the RBF kernel, values of gamma < 0.001 increased classification accuracy by over 20% relative to higher values of gamma for the three highest values of C.

4 Discussion.

We demonstrate that the ability with which support vector machines can classify microarray data is highly dependent upon both kernel type and parameter settings. As in many other domains, model selection is an important issue for microarray data analysis. In the present study, the regularization parameter C exhibited no significant influence on the linear SVM but was an important factor in the performance of the non-linear SVM classifiers. Gamma, playing different roles in the two non-

[1] Department of Computer Science and Program in Molecular, Cellular, and Integrative Neuroscience, Colorado State University, Fort Collins, Colorado. E-mail: petersod@cs.colostate.edu

linear classifiers, also influenced their performance. For both non-linear classifiers, the influence of gamma depended on the regularization parameter C, and vice-versa. This is an important result, because it suggests that the two parameters should be optimized jointly, which is much more computationally expensive than optimizing them separately. Such computational demands are particularly important when SVMs are used for feature selection [5]. Feature selection approaches that go beyond per-feature evaluations and allow for feature interaction commonly involve extensive search of the feature subset space and, consequently, many iterations of training and testing predictors such as SVM classifiers. Thus, further research to optimize SVM model selection will be particularly important for feature selection in microarray data analysis.

References

[1] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.G., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L.M., Marti, G.E., Moore, T., Hudson, J., Lu, L.S., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403 (2000) 503-511.

[2] Duan, K., Keerthi, S.S. and Poo, A.N., Evaluation of simple performance measures for tuning SVM hyperparameters, *Neurocomputing*, 51 (2003) 41-59.

[3] Guyon, I., Weston, J., Barnhill, S. and Vapnik, V., Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, 46 (2002) 389-422.

[4] Vapnik, V.N., *Statistical learning theory*, Wiley, New York, 1998, xxiv, 736 pp.

[5] Weston, J., Elisseeff, A., Scholkopf, B. and Tipping, M., Use of the Zero-Norm with Linear Models and Kernel Methods, *Journal of Machine Learning Research*, 3 (2003) 1439-1461.

G16. Stability Analysis of Time Series Gene Expression Data

J. Gebert, M. Lätsch, S. W. Pickl, N. Radde [1], G. - W. Weber [2], Röbbe Wünschiers [3], Bob Veroff [4]

Keywords: Gene Expression Analysis, Stability Analysis

Abstract

The analysis of gene expression data is a challenging field of computational biology. We represented the behavior of time series gene expression patterns by a system of differential equations, which we analytically and algorithmically investigate. Therefore we use an algorithm which computes regions of stability and instability. It takes advantage of extremal points of polyhedra, which grow step by step, up to a possible stopping. The overall goal is to correlate the stability pattern with metabolic states. New numerical results will be presented.

1 Mathematical Modelling of Gene Expression Data

Several approaches to analyze time-series gene expression data have been presented. Chen *et al.* [3] proposed to use the differential equation $\dot{E} = ME$ to model gene expression data. M represents a matrix with constant entries and $E(t)$ the vector of mRNA and protein concentrations at time t. The vector $\dot{E}$ represents the change of concentrations in time. The modelling approach suggested by De Hoon *et al.* [5] is based on Chen *et al.*. However, only mRNA measurements and different preconditions are used. In 2001, Sakamoto and Iba [6] proposed a more flexible approach $\dot{E}_i = f_i(E_1, ..., E_n) \, \forall i = 1, ..., n$ with n being the number of genes and f_i being a function in $E_1, ..., E_n$. Genetic programming is used to construct the model which has also been proposed by Cao *et al.* To be more effective Sakamoto and Iba also used the least mean square method along with the genetic programming. The method worked very well for small samples. For a large-sized network they will search for methods of pre-processing to reduce the given networks to small-sized sub-networks. Our approach is a further extension, since we firstly make M dependent on E and secondly analyze the system for stability.

2 Data Sets

Our initial experiments were based on artificial data sets. Thereby, we could for example control the innode degrees. Currently, we are working on the time-series gene expression data set from a diauxic shift of yeast. This data-set and the experimental background can be found in [4].

[1]Institute of Mathematics, Center for Applied Computer Science, University of Cologne, Weyertal 80, 50931 Cologne, Germany. E-mail: gebert@zpr.uni-koeln.de, laetsch@zpr.uni-koeln.de, pickl@zpr.uni-koeln.de, radde@zpr.uni-koeln.de

[2]Institute of Applied Mathematics, Middle East Technical University, 06531 Ankara, Turkey. E-mail: gweber@metu.edu.tr

[3]Institute of Genetics, University of Cologne, Weyertal 121, 50931 Cologne, Germany. E-mail: robbe.wunschiers@uni-koeln.de

[4]Compputer Science Department, University of Mexico, 153 Farris Engineering Center Albuquerque, NM 87131. E-mail: veroff@cs.unm.edu

3 Mathematical Modelling

Step 1 Based on the data of microarray experiments we want to develop a system of differential equations

$$(\mathcal{CE}) \qquad \dot{E} \;=\; M(E)E\,,$$

which describes the cell's process. The expression levels of the genes at time t are described by the vector $E(t) = E_1(t), ..., E_n(t)$, n being the number of genes in the microarray experiment, and $M(E)$ being a matrix depending on E.
How can $M(E)$ be calculated? For the time being we will cluster the data and use least squares to solve this problem.

Step 2 In the second step we want to analyze the system of differential equations for stability using an algorithm of Brayton and Tong [2, 1].
i) We apply Euler's discretization on the time-continuous system of differential equations $(\mathcal{CE})$ by which we describe the metabolic process. Based on our good differentiable approximation of the biochemical processes, the time-discretization by Euler's method supports and simplifies the analysis and approximate resolution of the continuous process $(\mathcal{CE})$. This method is not a very refined approximation but convenient and a strong tool, here. We get a set of matrices $\mathcal{M} = \{M_0, ..., M_{m-1}\}$ which we analyze for stability.
ii)We apply the algorithm of Brayton and Tong on the set $\mathcal{M}$ of matrices. This set is **stable**, if for every neighborhood of the origin $U \subset \mathbb{C}^n$ there exists another neighborhood of the origin $\tilde{U}$ such that, for each $M \in \mathcal{M}'$ it lasts $M\tilde{U} \subseteq U$. Brayton and Tong proved that $\mathcal{M}$ is stable, iff $\breve{B}$ is bounded, where

$$\breve{B} := \bigcup_{k=0}^{\infty} B_k\,, \text{ with } B_k := \mathcal{H}\left(\bigcup_{i=0}^{\infty} M_{k'}^{i} B_{k-1}\right) \quad \text{and} \quad k' = (k-1) \textit{ modulo } m\,,$$

for $k \in \mathbb{N}$ and $B_0 \in \mathbb{C}^n$ a bounded neighborhood of the origin.
The algorithm constructs polytopes B_i consecutively and checks whether $\breve{B}$ is bounded and therefore whether the set $\mathcal{M}$ is stable.

References

[1] R. Brayton and C. Tong. Constructive stability and asymptotic stability of dynamical systems. *IEEE Transactions on Circuits and Systems*, 27(11):1121–1130, 1980.

[2] R. K. Brayton and C. H. Tong. Stability of dynamical systems: A constructive approach. *IEEE Transactions on Circuits and Systems*, 26(4):224–234, 1979.

[3] T. Chen, H. L. He, and G. M. Church. Modeling gene expression with differential equations. *Pac Symp Biocomput*, pages 29–40, 1999.

[4] I. V. B. P. DeRisi, J.L. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–6, 1997.

[5] M. D. Hoon, S. Imoto, and S. Miyano. Inferring gene regulatory networks from time-ordered gene expression data using differential equations. In C. H. S. Steffen Lange, Ken Satoh, editor, *Lecture Notes in Computer Science*, volume 2534, pages 267–274. Springer-Verlag, 2002.

[6] E. Sakamoto and H. Iba. Inferring a system of differential equations for a gene regulatory network by using genetic programming. *Proc. Congress on Evolutionary Computation*, pages 720–726, 2001.

G17. Non-Unique Probe Selection by Matrix Condition Optimization

Sven Rahmann,[1] **Tobias Müller,**[2] **Martin Vingron**[3]

Keywords: microarray, DNA chip, design, probe selection, matrix, condition, optimization

1 Introduction: Non-Unique Probe Selection

We are interested in selecting oligonucleotide probes for DNA arrays [1]. In large transcript families, such as alternative splice variants of a gene, or in a large family of closely homologous genes (e.g., human heat shock proteins), it is often impossible to find enough unique 25-mer probes that can be taken as a signature for a specific variant. Therefore we consider *non-unique probes* [2].

For this study, we assume that we have many potential probe candidates and the task is to select an appropriate subset of them for use on the chip. We assume that we know (an approximation to) the probe-transcript *affinity matrix* A that relates transcript expression level to observed signal. We have $y = A \cdot x + c$, where $y \in \mathbb{R}^m$ are the observed probe signals, $x \in \mathbb{R}^n$ contains the transcript expression values, $A \in \mathbb{R}^{m \times n}$ contains the affinity coefficients between probes and transcripts, and c models additional noise or unspecific hybridization. A probe i that matches a transcript j leads to a high affinity value $A_{ij} \approx 0.1$ to 1, say. The target set of probe i is denoted by $T(i)$. A probe i that is unrelated to transcript j leads to a low affinity value of less than 10^{-4}, say.

From the $m \gg n$ probe candidates whose affinity values form the m rows of the affinity matrix A, we would like to select at most $\mu \leq m$ rows. We write H for the *hybridization matrix* defined by $H_{ij} := 1$ if $j \in T(i)$, and $H_{ij} := 0$ otherwise. We denote the index set of the chosen rows by D for *design*. We have $D \subset \{1, 2, \ldots, m\}$ and desire $|D| \leq \mu$. Let A^D and H^D denote the matrices obtained from A resp. H by removing all rows whose index is not in D. The requirements on D are that the equation $y = A^D \cdot x$ must be stably and robustly solvable for the n expression levels x, given the $|D|$ probe signals y.

2 Condition Optimization

Let A be an $m \times n$ matrix of full rank $n \leq m$, and let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n > 0$ be the singular values of A. Then the condition of A is defined as $\mathrm{cond}(A) := \sigma_1/\sigma_n$. If A does not have full rank n, then $\sigma_n = 0$, and we set $\mathrm{cond}(A) := \infty$. The condition measures how changes in the measurement y influence the solution x of the minimization problem $\|y - A \cdot x\| \to \min$: We have $\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}(A) \cdot \frac{\|\Delta y\|}{\|y_A\|}$, where y_A is the projection of y on the range of A.

We assume that $\mathrm{cond}(A) < \infty$, i.e., that the affinity matrix that consists of all candidates has full rank n, and that the associated hybridization matrix H satisfies the minimum and average *coverage constraints* $\min_j \sum_i H_{ij} \geq \mathcal{M}$ and $\sum_{i,j} H_{ij} \geq n\mathcal{A}$. We let

[1]Max Planck Institute for Molecular Genetics, Computational Molecular Biology, Ihnestr. 73, D-14195 Berlin, Germany; and Dept. of Mathematics and Computer Science, Free University of Berlin, Germany. E-mail: Sven.Rahmann@molgen.mpg.de

[2]Department of Bioinformatics, Biozentrum, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany. E-mail: Tobias.Mueller@biozentrum.uni-wuerzburg.de

[3]Max Planck Institute for Molecular Genetics, Computational Molecular Biology, Ihnestr. 73, D-14195 Berlin, Germany. E-mail: Martin.Vingron@molgen.mpg.de

$\mathcal{D} := \{D \subset \{1,2,\ldots,m\} : |D| \le \mu,\ \mathrm{cond}(A^D) < \infty,\ \min_j \sum_i H^D_{ij} \ge \mathcal{M},\ \sum_{i,j} H^D_{ij} \ge n\mathcal{A}\}$ denote the set of admissible designs under the side conditions. It is assumed that $\mathcal{D}$ is not empty; otherwise we have to increase μ or decrease $\mathcal{M}$ or $\mathcal{A}$. The combinatorial problem is to minimize $\mathrm{cond}(A^D)$ among all $D \in \mathcal{D}$.

As far as we are aware, the *condition optimization problem* has not been posed before in the mathematical literature. It appears to be a difficult problem because the singular values of two matrices A^D and A^{D-i}, where $D - i := D \setminus \{i\}$, are not related in an obvious way, and also because the landscape of admissible designs potentially has many local minima. In spite of these difficulties, we propose a greedy heuristic to obtain a good admissible design.

Greedy Condition-based Design

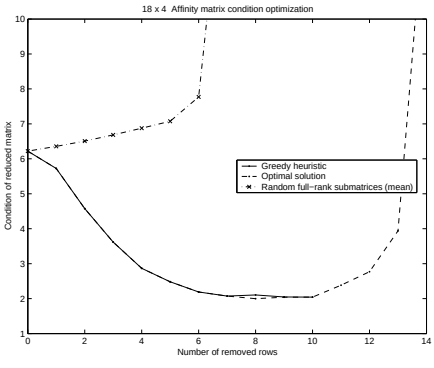
Input: An $m \times n$ affinity matrix A and hybridization matrix H

1. $D \leftarrow \{1,2,\ldots,m\}$
2. $B \leftarrow \emptyset, \quad C \leftarrow +\infty$
3. **while** $(|D| > n)$
4. $\quad c \leftarrow +\infty, \quad i^* \leftarrow 0$
5. $\quad$ **for each** $i \in D$
6. $\quad\quad$ **if** $(\min_j \sum_{i'} H^{D-i}_{i'j} \ge \mathcal{M})$ **and** $(\sum_{i',j} H^{D-i}_{i'j} \ge n\mathcal{A})$ **and** $(\mathrm{cond}(A^{D-i}) < c)$ **then** $c \leftarrow \mathrm{cond}(A^{D-i}),\ i^* \leftarrow i$
7. $\quad$ **if** $i^* = 0$ **then break**
8. $\quad D \leftarrow D - i^*$
9. $\quad$ **if** $(|D| \le \mu)$ **and** $(c < C)$ **then** $B \leftarrow D,\ C \leftarrow c$
10. **if** $(B = \emptyset)$ **then** $B \leftarrow D,\ C \leftarrow c$

Output: Design B with condition $C = \mathrm{cond}(A^B)$

The procedure starts with a full design and iteratively removes a single row to locally minimize the condition while still satisfying the coverage constraints(lines 5–6). If the resulting design is admissible it is compared against the current best admissible design B (line 9). This is repeated until the design size equals the number of targets (line 3) or no smaller design satisfying the coverage constraints can be found (line 7).

We evaluated the greedy heuristic against the optimal selection for small artificial matrices with 18 probe candidates and 4 targets. It was attempted to reduce the number of probes as far as possible with a minimum and average coverage requirement of 3 probes per target. Although the greedy heuristic does not always find the optimal solution, its performance is reasonably close to the optimal design (found by exhaustive search) and much better than choosing random subsets. The typical behavior is shown to the right: Removing a few probes improves the condition; removing too many will eventually worsen the condition again.



References

[1] S. Rahmann. Fast large scale oligonucleotide selection using the longest common factor approach. *Journal of Bioinformatics and Computational Biology*, 1(2):343–361, 2003.

[2] A. Schliep, D. C. Torney, and S. Rahmann. Group testing with DNA chips: Generating designs and decoding experiments. In *Proceedings of the 2nd IEEE Computer Society Bioinformatics Conference (CSB 2003)*, pages 84–93. IEEE, 2003.

G18. Incorporation of Target RNA Secondary Structure Parameter into Synthetic Oligomer Probe Design

Vladyslava G. Ratushna[1], Jennifer W. Weller[2], Cynthia J. Gibas

Keywords: microarray design, target RNA structure

1 Introduction.

Lack of reliable bioinformatic tools, which would incorporate the phisico-chemical characteristics of the nucleic acids as part of a synthetic oligomer probe design criterion, is a major challenge of modern microarray research. One of these phisico-chemical parameters, which reduces the binding effectiveness of microarray probes, is the propensity of long target RNA molecules to form stable secondary structures. Different 3-D foldings such as hairpins and stacked regions have the potential to pre-empt target nucleotides, thus blocking regions of the long RNA molecules from hybridizing to their intended probes. Our computational analysis indicates that, although raising the probe-target hybridization temperature and implementation of shearing techniques remove some of the target secondary structure problems, they cannot eliminate all of the stable short helices across the RNA molecule. These temperature resistant structures create potential problems for the on-chip probe annealing that is not predicted by consideration of probe structure alone. Identification and masking of the target RNA stable secondary structure sequences and the blocked regions of the target molecule should be included as one of the algorithms used for the design of synthetic oligo probes.

2 Experiment and Methods.

Thermodynamic analysis of the secondary structures was performed for a set of ten out of fifteen different urease genes present in *Brucella suis* genome using the Mfold 3.1 software [1]. This program package utilizes the nearest neighbor energy approach, which assigns free energies to loops rather than to base pairs [2]. Our calculations were performed for the RNA secondary structure predictions at four different temperatures, chosen as being those most often utilized in laboratories as the hybridization temperatures for microarrays: 37˚C, 42˚C, 52˚C and 65˚C for full length RNA molecules at buffer conditions of 1.0 M sodium concentration and no magnesium ion. The free energy increment for computing the suboptimal foldings, $\Delta\Delta G$, was set to 5% of the computed minimum free energy ΔG. The default values of the window parameters, controlling the number of foldings automatically computed by Mfold 3.1 were picked based on the sequence length.

The computations revealed that for approximately 300 bp long sequences at 37˚C about 7 % of the double stranded helices within the suboptimal ΔG range are at least 6 bp long with the minimum free energy of the overall folding containing these structures ranging from -930 to -980 10^{th} of the kcal/mol. For the same sequence at 42˚C also about 7 % of the double stranded helices within the suboptimal ΔG range are at least 6 bp long with the minimum free energy of the overall folding

[1] Department of Biology, Virginia Polytechnic Institute and State University, 4083 Derring Hall, Blacksburg, Virginia, U.S.A. E-mail: vratushn@vt.edu

[2] School of Computational Science, George Mason University, PW1 Rm. 441, Manassas, Virginia, U.S.A. E-mail: jweller@gmu.edu

containing these structures ranging from -860 to -820 10^{th} of the kcal/mol. At 52˚C about 9 % of the double stranded helices within the suboptimal ΔG range are at least 6 bp long with the minimum free energy of the overall folding containing these structures ranging from -640 to -600 10^{th} of the kcal/mol. At 65˚C about 12 % of the double stranded helices within the suboptimal ΔG range are at least 6 bp long with the minimum free energy of the overall folding containing these structures ranging from -380 to -350 10^{th} of the kcal/mol. These data suggest that the temperature raise up to 65˚C although removes a substantial part of the secondary structure but doesn't eliminate competing structure or cause complete unwinding of all molecules. Below is an illustration of an optimal and suboptimal folding for the ureA1 mRNA at 37˚C and 65˚C [3]:

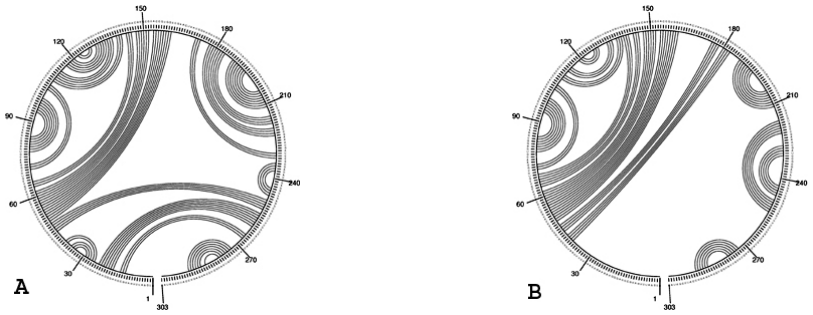


Figure 1: **A** 37˚C, ΔG=-111.33 kcal/mol **B** 65˚C, ΔG=-37.5 kcal/mol

A computational shearing simulation was performed 37˚C, 42˚C, 52˚C and 65˚C with the 37˚C, 42˚C, 52˚C and 65˚C 10 nucleotide window and the sheared fragment size of 200, 100 and 50 nucleotides. The obtained results indicate that although shearing has a major effect on the elimination of the secondary structure formation, up to 12 % of the double stranded helices within the suboptimal ΔG range are at least 6 bp long with the minimum free energy of the overall folding containing these structures ranging from 30 to -110 10^{th} of the kcal/mol.

3 Conclusions.

Computational analysis of the secondary structures at a range of temperatures of the full length mRNA sequences as well as the sheared fragments revealed a set of stable secondary structures, which do not disappear with either raise of the hybridization temperature, shearing or both, and may cause problems during the probe annealing stage of the microarray experiment. Therefore, we suggest to develop a new microarray probe design criteria based on the presence of the stable secondary structures in the target mRNA sequence.

References

[2] Mathews, D.H., Sabina, J., Zuker , M. and Turner, D.H. 1999. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *Journal of Molecular Biology* 288: 911-940.

[3] Zuker , M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research.* 31 (13): 3406-15.

[1] Zuker, A.M., Mathews, B.D.H. and Turner C.D.H. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide.

G19. Identification of Transcribed Differentiating Genes in *Brucella abortus, B.melitensis and B.suis*

Vladyslava Ratushna[1], David Sturgill, Sheela Ramamoorthy, Sherry Poff, Nammalwar Sriranganathan, Stephen Boyle[2], Cynthia Gibas

Keywords: comparative, functional, genomics, differential, regions, gene, expression, *Brucella*

1 Introduction.

Design of diagnostics based on microarray technology inherently involves a multiple genome comparison. Unique regions must be unequivocally identified; this process can not rely solely on comparison of annotated gene sequences, especially when annotation is tentative or incomplete. A three-way comparison of published and annotated genomes of *Brucella. melitensis* and *B. suis*, and a draft sequence of the *B. abortus* genome identified a group of unique known and hypothetical gene coding sequences in these species [1]. The comparison was performed using a prototype of the GenoMosaic toolkit [2]. Although only *B. suis* was found to have a significant number of unique genes, patterns of genes that exist in only two out of three genomes were also identified. These patterns will discriminate unambiguously between *B. suis*, B. *melitensis*, and B. *abortus*, and are important for explaining the differences in virulence and host specificity of the three *Brucella* spp. The existence and *in vitro* transcription behavior of the differentiating genes were confirmed by PCR.

Brucella is a facultative intracellular pathogen with approximately 3 Mb genome, split between the two chromosomes the size of 1.85 Mb and 1.35 Mb. Human brucellosis is quite common but often not diagnosed. There are six recognized *Brucella* species that differ in their host preference. *B. abortus* preferentially infects cattle, *B. melitensis* infects sheep and goats, and *B. suis* infects pigs. All three of these species and *B. canis* can infect humans, although *B. melitensis* is associated with the most serious human infections. The *Brucellae* are grouped with the α-proteobacteria and are related to other cell-associated parasites of plants and animals [3]. The true pattern of *Brucella* intracellular survival and proliferation, and the reasons for the different virulence patterns among the species are not conclusively known.

2 Experiment and Methods.

Reverse transcriptase (RT-PCR) analyses were performed for identified unique and differential regions of three *Brucella* spp.**: *B. abortus, B.*** *melitensis* and ***B. suis,*** to determine whether the identified differentiating genes are transcribed *in vivo*. To compensate for the absence of completed annotation in *B.abortus*, the identified differentiating open reading frames from *B. melitensis* and *B. suis* were used to develop the RT-PCR primers. The same primers were used to interrogate each of the three genomes, testing for the existence and expression of these genes in *B. Abortus*. Patterns of identified unique and differentiating genes that exist in only two out of three genomes were

[1] Department of Biology, Virginia Polytechnic Institute and State University, 4083 Derring Hall, Blacksburg, Virginia, U.S.A. E-mail: vratushn@vt.edu

[2] Virginia-Maryland Regional College of Veterinary Medicine, Virginia Polytechnic Institute and State University, CMMID, 1410 Prices Fork Road, Blacksburg, Virginia, U.S.A. E-mail: smboyle@vt.edu

confirmed by PCR. Both PCR and RT-PCR reactions were performed for 104 unique genes or partial differential genes detected in the three genomes by whole genome sequence comparison[2]. 23 unique genes, and 79 differential genes common between the two of the three species were tested for transcription. Table 1 summarizes transcripts detected for genes in each differentiating sequence island.

Genomic	*Brucella suis*			*Brucella melitensis*			*Brucella abortus*		
location	Predicted	Observed	NB	Predicted	Observed	NB	Predicted	Observed	NB
S1	4	4	0	1	1	0	0	0	0
S2	18	17	1	0	0	0	0	0	0
M1	0	0	0	1	1	0	0	0	0
A	0	0	0	0	0	0	1	1	0
SM1	1	1	0	2	2	0	0	0	0
SM2	25	16	9	24	6	18	0	0	0
SA1	11	3	8	0	0	0	9	7	2
SA2	11	7	4	0	0	0	11	6	5
MA1	0	0	0	30	21	9	26	23	3

Table 1: Summary of RT-PCR results for differential ORFs.
A – *B.abortus*, M – *B.melitensis*, S – *B.suis*, 1 – chromosome I and 2 – chromosome 2 and NB – no band.

Several groups of genes with potential significance for virulence were detected as differentials and shown to be transcribed. Some of these genes were likely of phage or plasmid origin, suggesting possible mechanisms for their appearance as differentials.

3 Conclusions.

In three very closely related pathogen genomes, systematic genome-wide comparison was used to identify targets for a diagnostic microarray. RT-PCR tests confirmed computational predictions. PCR amplification of the genomic DNA confirmed the presence of the predicted amplicon even where transcription was not detected, and transcription was not detected in any case where we had predicted that the target would be absent. However, in some cases, predicted differentiating genes did not appear to be transcribed under the conditions of the experiment, as the predicted amplicon was not observed even though the gene was detected by PCR in the genomic DNA. All of the identified differentials have been used in development of a Brucella microarray containing probes for both common and differentiating targets.

References

[2] Gibas C.J., Sturgill D.M., Weller, J.W. GenoMosaic: On-Demand Multiple Genome Comparison and Comparative Annotation. 2003. In *Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering*, IEEE Press: 158-167.

[3] Paulsen, I.T., Seshadri, R., Nelson, K.E., Eisen, J.A., Heidelberg, J.F., Read, T.D., et al. 2002. The *Brucella suis* genome reveals fundamental similarities between animal and plant pathogens and symbionts. *Proceedings of the National Academy of Sciences U S A* 99: 13148-13153.

[1] Sturgill, D.M., Ratushna, V.G., Ramamoorthy, S., Poff, S.A., He, Y., Lathigra, R., Sriranganathan, N., Halling, S., Boyle, S.M., Gibas, C.J. 2004. Comparative expression analysis of differentiating regions in *Brucella melitensis, B. abortus* and *B. suis. Journal of Bacteriology*. (submitted).

G20. MeSH Key Terms for Validation and Annotation of Gene Expression Clusters

Andreas Rechtsteiner and Luis M Rocha [1]

Keywords: gene expression analysis, validation, information retrieval, automated functional annotation

Integration of different sources of information is a great challenge for the analysis of gene expression data, and for the field of Functional Genomics in general. As the availability of numerical data from high-throughput methods increases, so does the need for technologies that assist in the validation and evaluation of the biological significance of results extracted from these data. In mRNA assaying with microarrays, for example, numerical analysis often attempts to identify clusters of co-expressed genes. The important task to find the biological significance of the results and validate them has so far mostly fallen to the biological expert who had to perform this task manually.

One of the most promising avenues to develop automated and integrative technology for such tasks lies in the application of modern Information Retrieval (IR) and Knowledge Management (KM) algorithms to databases with biomedical publications and data. Examples of databases available for the field are bibliographic databases containing scientific publications (e.g. MEDLINE/PUBMED), databases containing sequence data (e.g. GenBank) and databases of semantic annotations (e.g. the Gene Ontology Consortium and Medical Subject Headings (MeSH)).

We present here an approach that uses the MeSH terms and their concept hierarchies to validate and obtain functional information for gene expression clusters[2]. The controlled and hierarchical MeSH vocabulary is used by the National Library of Medicine (NLM) to index all the articles cited in MEDLINE. Such indexing with a controlled vocabulary eliminates some of the ambiguity due to polysemy (terms that have multiple meanings) and synonymy (multiple terms have similar meaning) that would be encountered if terms would be extracted directly from the articles due to differing article contexts or author preferences and background. Further, the hierarchical organization of the MeSH terms can illustrate the conceptual/functional relationships of genes associated with MeSH terms. MeSH terms can be associated with genes through co-occurrence of these in MEDLINE citations, i.e. the genes occur in titles or abstracts and the MeSH terms are assigned by experts.

To identify MeSH terms associated with a group of genes we used the tool MESHGENE developed at the Information Dynamics Lab at HP Labs (http://www-idl.hpl.hp.com/meshgene/). When presented with a list of human genes, MESHGENE uses some sophisticated techniques to search for these gene symbols in the titles and abstracts of all MEDLINE citations. MeSH terms and the number of co-occurrences can be retrieved[3]. Gene symbols that are aliases of each other are pooled from several databases. This addresses the problem of synonymy, the fact that several symbols can refer to the same gene. MESHGENE employs some sophisticated algorithms that disregards symbols that are likely to be acronyms for other concepts than a gene. This addresses the problem of polysemy, i.e. possible multiple meanings of a gene symbol[4].

We applied our approach to gene expression data from herpes virus infected human fibroblast cells [1]. The data contains 12 time-points, between 1/2 hrs and 48 hrs after infection. Singular Value Decomposition was used to identify the dominant modes of expression. 75% of the variance in the expression data was captured by the first two modes, the first exhibiting a monotonly increasing expression pattern and the second a more transient pattern. Projection of the gene expression vectors onto this

[1] CCS-3, Los Alamos National Laboratory. E-mail: andreas@lanl.gov

[2] Masys et al. [2] presented a proof-of-concept utility related to this approach.

[3] MESHGENE also returns 'scores', a statistical measure of association that attempts to take into account how often one would find a certain gene-MeSH term co-occurrence by chance.

[4] Both of these techniques improve on some of the short-comings of the approach by Masys et al. [2]

first two modes identified 3 statistically significant clusters of co-expressed genes[5]. 500 genes from cluster 1 and 300 genes from clusters 2 and 3 each were uploaded to MESHGENE and the MeSH terms and co-occurrence values were retrieved. MeSH terms were also obtained for 5 groups of randomly selected genes with similar numbers of genes. The log was taken of the co-occurrence values and for each MeSH term these log co-occurrence values were summed for each group over the genes in that group. A matrix with 8 columns for the 8 groups of genes and with 14,000 rows with the MeSH terms was obtained. To analyze this association matrix we used a Latent Semantic Analysis (LSA) approach. We applied SVD to this gene-group vs. MeSH term association matrix. The first 2 modes that capture most of the variation (and therefore most times also information) in the association matrix were highly associated with MeSH terms that occurred uniquely or disproportionally in the 3 gene clusters. MeSH terms highly associated with the 5 groups of randomly selected genes were associated with the lower modes. These modes seem to just capture 'noise' in the association matrix. This result by itself is of great interest for gene expression analysis. We were able to show that the 3 clusters of genes not only separated in 'expression space' but also in the MeSH term space with which they are associated through the literature. Further, our results indicate that the genes within the 3 clusters are not only similar in expression space (i.e. co-expressed) but are also more coherent in MeSH space (and therefore probably also functionally coherent) than the randomly grouped genes. These two observations support, and in some way validate, the clustering results obtained from the expression data.

We also inspected the MeSH terms most associated with the 3 clusters for functional information and compared the results to a manual annotation of clusters 1 and 2 that was done by biological experts[6]. Many MeSH terms highly associated with cluster 1 were related to viruses, indicating that many genes in cluster 1 must have been reported in the literature in connection with these. A disproportionate number of MeSH terms for cluster 1 were also related to oncogenesis, transcription regulation, antigene processing, antibody formation, and Major Histocompatibility Complexes I and II. The manual annotation by the biological experts identified a disproportionate high number of genes for cluster 1 related to all of the above biological processes or cellular components suggested by the MeSH terms. MeSH terms highly associated with cluster 2 were related to immune response, T cells, macrophages, infections, inflammation, cytokines and cytokine receptors. Manual annotation supported the above findings by identifying a disproportionate number of genes in cluster 2 related to these concepts. For cluster 3 highly associated MeSH terms were related to connective tissue and connective tissue diseases (note the herpes infected cells were human fibroblast cells), enzymes involved in apoptosis, immune response and oncogene proteins. (No manual annotation of cluster 3 genes was performed.) The results indicate that the MeSH terms associated with the genes in the 3 clusters are informative about functions of many of the genes in these clusters, and would at least be a good guide for the biological expert trying to investigate the functions of these in more detail.

We are working on further exploration and improvements on the above methodology. We are exploring, for example, different measures of association between MeSH terms and genes. We also explore the usefulness of our approach for other areas of Functional Genomics, like protein function inference.

References

[1] E.P. Browne, B. Wing, D. Coleman, and T. Shenk. Altered cellular mRNA levels in human cytomegalovirus-infected fibroblasts: Viral block to the accumulation of antiviral mrnas. *Journal of Virology*, 75(24):12319–30, 2001.

[2] D.R. Masys, J.B. Welsh, J. Lynn Fink, M. Gribskov, I. Klacansky, and J. Corbeil. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17(4):319–26, 2001.

[5] A manuscript describing this SVD based algorithm is in preparation.

[6] A manuscript of this study is in preparation.

G21. Evaluation of Statistical Methods for cDNA Microarray Differential Expression Analysis

Wei Sha[1], Keying Ye[2], Pedro Mendes[3]

Keywords: microarray data analysis, differential expression, t-test, SAM, ANOVA, mixed model

1 Introduction.

There are many statistical methods to identify differentially expressed genes in cDNA microarray experiments. A straightforward method is to use the traditional two sample t-test [1]. Tusher et al. [2] proposed the Significance Analysis of Microarray (SAM) method. Kerr et al. [3] and Kerr and Churchill [4] were the first to propose the study of gene expression data using analysis of variance (ANOVA) models. These models perform both normalization and identification of differentially expressed genes. However, none of these methods have yet gained wide acceptance, and it has perhaps been difficult to decide which method performs better than the others. Thus, the evaluation of these methods for differential expression is an important issue. The performance of four statistical methods for differential expression was compared in this simulation study. Gene expression data was simulated, and then analyzed by a) Welch t-test b) SAM c) ANOVA gene model d) mixed ANOVA model. The relative performance of these methods in terms of accuracy in detecting differentially expressed genes was evaluated.

2 Simulating microarray data.

Artificial gene networks were generated following one of three topologies, Erdös-Renyi random networks, Watts-Strogatz small-world networks [5], and Albert-Barabási scale-free networks [6]. Details of artificial gene networks can be found in our recent paper [7]. These networks were used to generate gene expression data using the program Gepasi [8]. This program allows us to generate different type of experiments, such as null mutants, or time courses following treatments. Noise arising from array, spot, and other sources was simulated and added by PERL script at three different levels following the normal distribution. In this study, we simulated three cDNA microarray experiments in a reference design by using three different gene networks. In each simulated microarray experiment, there were 100 genes, six treatments and three replicates.

3 Data analysis and results.

Simulated data were normalized by median center before analyzed by t-test, SAM and ANOVA gene model. Simulated data were not normalized before analyzed by mixed ANOVA model, since this method itself does normalization. Bonferroni adjustment was used for t-test, ANOVA gene model and mixed ANOVA model. ANOVA gene model and mixed ANOVA model are shown as follows.

[1] Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, 1880 Pratt Dr., Blacksburg, VA 24061, USA. E-mail: wsha@vt.edu

[2] Dept. of statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA. E-mail: keying@vt.edu

[3] Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, 1880 Pratt Dr., Blacksburg, VA 24061, USA. E-mail: mendes@vt.edu

$$y_{jk} = \mu_{jk} + A_j + T_k + \varepsilon_{jk} \quad \text{-------} \quad \text{ANOVA gene model}$$

$$y_{ijk} = \mu_{ijk} + G_i + A_j + T_k + (AT)_{jk} + (AG)_{ji} + (TG)_{ki} + \varepsilon_{ijk} \quad \text{-------} \quad \text{mixed ANOVA model}$$

The false positive rates and false negative rates were calculated for each method. Results are shown in Table 1. We found that all these methods have good performance when noise level is low. However, as noise increases, the mixed model performs better than others.

	Low noise		*Medium noise*		*High noise*	
	False positive	False negative	False positive	False negative	False positive	False negative
t-test	1.33	0	3.70	44.44	5.67	52.00
SAM	1.33	0	3.70	22.22	5.67	48.33
ANOVA gene model	0	0	0	18.52	0	48.33
Mixed model	0	0	0	18.52	0	35.33

Table 1: False positive rate (%) and false negative rate (%).

Although we have studied only a limited number of data sets, out findings may provide some guidance on the selection of differential expression methods. To get more convincing results, we have automated the data analysis process in PERL and SAS. Hundreds of data sets are now being analyzed in these programs. The results from these data sets will also be shown on the poster.

References

[1] Devore, J. and Peck, R. 1997. *Statistics: the Exploration and Analysis of Data*, 3rd edn, Duxbury Press, Pacific Grove, CA.

[2] Tusher, V.G., Tibshirani, R. and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, 98, 5116-5121.

[3] Kerr, M.K., Martin, M. and Churchill, G.A. 2000. Analysis of variance for gene expression microarray data. *J. Comput. Biol.,* 7, 819-837.

[4] Kerr, M. and Churchill, G. 2001. Experimental design for gene expression microarrays. *Biostatistics* 2, 183-201.

[5] Watts, D. J. & Strogatz, S. H. 1998. Collective dynamics of 'small-world' networks, *Nature.* 393, 440-2.

[6] Barabási, A. L. & Albert, R. 1999. Emergence of scaling in random networks, *Science*. 286, 509-12.

[7] Mendes, P., Sha, W. and Ye, K. 2003. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics.* 19, ii122-ii129.

[8] Mendes, P. (1997) Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3., *Trends Biochem. Sci.* 22, 361-363.

G22. How Noisy are DNA Microarray Data?

Suman Sundaresh[1§], She-pin Hung[1*], G. Wesley Hatfield[*], and Pierre Baldi[§]

Keywords: DNA microarrays, differential gene expression, correlation, replicated experiments

1 Abstract

What factors are important in improving consistency in replicated measurements? Every stage of a microarray experiment has the potential to introduce "noise" [3,6]. This work analyses variability in highly replicated measurements (up to 32x) of DNA microarray data (using wild type *E. coli* as the model organism [1,2,4,6]) and quantifies the noise attributed by differences in the biological, experimental and technological setup of the experiments. Replicability is assessed quantitatively using correlation analysis as a global measure and differential expression analysis at the level of individual genes. The factors introducing variability that are considered in our study are differences in cDNA targets (complementary to independently prepared mRNA preparations), filters, imaging technology (Affymetrix MAS4.0 or MAS5.0, dChip[7]), DNA microarray formats (nylon filters and Affymetrix GeneChips™) and improvements in labeling. Our findings show that the major source of variance comes from biological factors such as using different cDNA targets. In nylon filter experiments, higher correlation (0.95) is observed when biological factors are kept consistent as opposed to experimental conditions (filters) (0.92). Changes in both filters and cDNA targets result in a significant drop in correlation (0.86). GeneChip measurements show an average correlation in the 0.81-0.85 range when a given image processing software is used. However, when two different software are compared (dChip [7] versus MAS 4.0 or 5.0), these values drop to around 0.78. We also find that signal intensities obtained from different microarray platforms do not correlate well with one another (not higher than 0.4) possibly due to probe effects. Differential gene expression analysis on the same data supports the findings from the correlation analyses. We observe that changing the biological factors alone introduces 2-3 times more false positives than experimental differences (using standard t-test on log transformed data [8] with $p<0.005$). As the differences are compounded, these numbers sharply increase up to 10 times the false positive levels. We present ways in which biological variations can be minimized such as harvesting the RNA as quickly as possible to prevent the effects of temperature shifts or osmotic stress and adopting standard cell-specific media for the growth of cells. Low correlation between the different microarray platforms suggests that the sets of replicates may not be combined. However, studies [5] have shown that signal ratios and overall differential expression profiles maybe comparable.

2 Results

Results obtained from pre-synthesized nylon filter experiments	Average Correlation	Average Number of False Positive Genes (t-test with p<0.005)
Duplicate measurements from each filter (same target)	0.974	0
Same targets hybridized to different filters	0.951	25.88
Different targets hybridized to the same filters	0.917	68.38
Different targets hybridized to different filters	0.859	124.03

Table 1. Comparing the effects of biological and environmental variables: Average correlation and false positive count of 32 measurements (of genes from wild type Escherichia coli) obtained using different cDNA targets and filters (excluding timeframe and labeling).

[1] These authors contributed equally to this work. E-mail: {suman, shung, gwhatfie, pfbaldi}@uci.edu
[*] The Department of Microbiology and Molecular Genetics, College of Medicine, University of California, Irvine
[§] The School of Information and Computer Science, University of California, Irvine
[§*] The Institute for Genomics and Bioinformatics, University of California, Irvine

Correlation		E1	E2	E3	E4	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
Filter	**E1**	1.00															
	E2	0.90	1.00														
	E3	0.60	0.75	1.00													
	E4	0.73	0.86	0.92	1.00												
MAS 4.0	**C1**	0.34	0.29	-0.01	0.13	1.00											
	C2	0.35	0.30	-0.03	0.12	0.87	1.00										
	C3	0.34	0.30	-0.01	0.14	0.83	0.84	1.00									
	C4	0.35	0.31	0.01	0.15	0.80	0.84	0.79	1.00								
dChip	**C1**	0.35	0.30	0.00	0.15	0.80	0.84	0.79	0.77	1.00							
	C2	0.34	0.29	0.01	0.15	0.77	0.80	0.76	0.84	0.87	1.00						
	C3	0.34	0.30	0.02	0.15	0.86	0.82	0.79	0.76	0.91	0.83	1.00					
	C4	0.34	0.30	0.01	0.15	0.80	0.81	0.84	0.76	0.88	0.84	0.88	1.00				
MAS 5.0	**C1**	0.36	0.30	0.00	0.14	0.92	0.86	0.83	0.80	0.81	0.78	0.86	0.81	1.00			
	C2	0.37	0.31	-0.02	0.14	0.85	0.92	0.83	0.83	0.84	0.80	0.82	0.80	0.90	1.00		
	C3	0.37	0.33	0.01	0.16	0.84	0.85	0.92	0.81	0.81	0.78	0.81	0.85	0.87	0.87	1.00	
	C4	0.35	0.31	0.02	0.16	0.78	0.81	0.77	0.86	0.76	0.82	0.76	0.76	0.82	0.84	0.81	1.00

False Positives Count	Filter	MAS4.0	dChip	MAS5.0
Filter	0			
MAS4.0	674	0		
dChip	662	102	0	
MAS5.0	678	9	101	0

Table 2. Comparison of different microarray platforms: (a) Correlation between measurements from pre-synthesized nylon filter and Affymetrix GeneChip™ experiments (processed with dChip, MAS 4.0 and MAS 5.0) (b) Matrix showing number of false positive genes identified when comparing filter and Affymetrix GeneChip™ data (t-test with $p<0.005$).

3 References and acknowledgments[2]

[1] Arfin, S. M., Long, A. D., Ito, E. T., Tolleri, L., Riehle, M. M., Paegle, E. S., and Hatfield, G. W. (2000) Global gene expression profiling in Escherichia coli K12. The effects of integration host factor. *J. Biol. Chem.* 275, 29672-29684

[2] Baldi, P., and Hatfield, G. W. (2002) DNA microarrays and gene expression: From experiments to data analysis and modeling, *Cambridge University Press*, Cambridge, UK

[3] Coombes, K.R., Highsmith, W.E., Krogmann, T.A., Baggerly, K.A., Stivers, D.N., Abruzzo, L.V. (2002) Identifying and quantifying sources of variation in microarray data using high-density cDNA membrane arrays. *Journal of Computational Biology* 9, 655-669

[4] Hatfield, G. W., Hung, S.-P., and Baldi, P. (2003) Differential analysis of DNA microarray gene expression data. *Mol. Microbiol.* 47, 871-877

[5] Hung, S.-P., Baldi, P., and Hatfield, G. W. (2002) Global gene expression profiling in Escherichia coli K12: The effects of leucine-responsive regulatory protein. *J. Biol. Chem*. 277, 40309-40323

[6] Hung, S.-P., Hatfield, G. W., Sundaresh, S., and Baldi, P. (2003) Understanding DNA Microarrays: Sources and Magnitudes of Variances in DNA Microarray Data Sets. Genomics, Proteomics, and Vaccines. *G. Grandi Editor, John Wiley and Sons*

[7] Li, C., and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U. S. A.* 98, 31-36

[8] Speed, T. (2002) Always log spot intensities and ratios, Speed Group Microarray Page, http://www.stat.berkeley.edu/users/terry/zarray/html/log.html

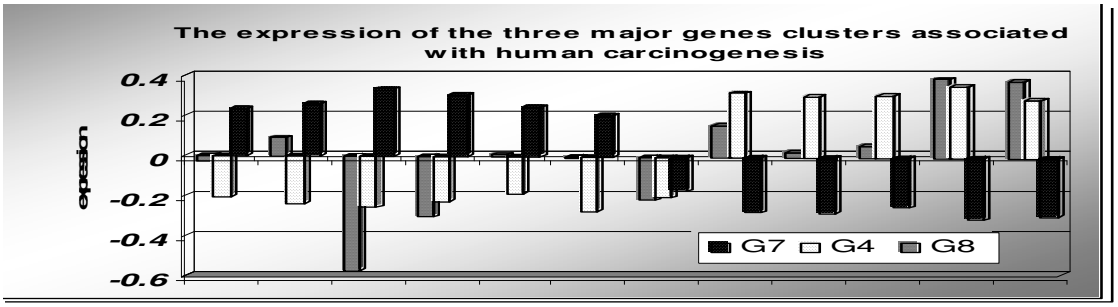
[2] This work was supported in part by the UCI Institute of Genomics and Bioinformatics, by grants from the NIH (GM-055073 and GM068903) to GWH, by a Laurel Wilkening Faculty Innovation Award to PB, and by a Sun Microsystems Award to PB. SH was supported by a postdoctoral training grant fellowship from the University of California Biotechnology Research and Education Program. We are grateful to Cambridge University Press for permission to reproduce materials from a book by PB and GWH titled "DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling" ISBN: 0521800226.

G23. Identification of transcriptional programs along defined stages of human carcinogenesis

Tabach Yuval[1,2], Milyavsky Michael[1], Zuk Or[2], Pilpel Yitzhak[3], Domany Eytan[2], Rotter Varda[1]

Keywords: microarray, Gene expression, data analysis, SPC, cancer, p53, H-Ras, *INK4A locous*,

Global profiling of human transcriptome during distinct stages of carcinogenesis presents a challenge, which requires novel experimental models and advanced analytical tools. We developed such a model of the stepwise transformation in vitro, which is based on human primary fibroblasts[1]. Upon hTERT expression spontaneous variants with inactivated INK4A locus arose, allowing further transformation by H-Ras. In the present study we showed that cells with concomitant expression of H-Ras and dominant negative p53 were endowed with more transformed features, which were sufficient for tumorigenic potential in vivo. Distinct stages along the multistage carcinogenesis were selected for microarray profiling.
Unsupervised analysis of the stepwise transformation, using SuperParamagnetic Clustering (SPC) [2], identified specific genetic signatures that differentiate samples according to the INK4A locus status, presence of functional p53 and H-Ras overexpression. Several important conclusions were drawn from our analysis. Dramatic downregulation of growth inhibitory molecules, putative tumor suppressors and proapoptotic factors were evident at the earliest stages of the neoplastic process. In contrast, robust induction of protein translation machinery, antiapoptosis genes and multiple cancer specific antigens characterize progression from a slow to a fast growing stage. Inactivation of wild type p53 in the absence of INK4A resulted in upregulation of multiple genes required for cell cycle progression, mitosis regulation and rapid proliferation resembling "proliferation signature" found in many aggressive human tumors. Lastly, genes involved in chemotactic attraction of endothelial cells as well as in inflammatory response and metastasis were induced in a synergistic manner in cells expressing both H-Ras and dominant negative p53, adding novel facets to H-Ras transforming activities. The detailed understanding of the alterations in the transcriptional programs occurring during human carcinogenesis will ultimately lead to the identifications of the novel anticancer targets.



Days in culture	21	30	40	40	40	193	319	480	600	480	480	480
INK4A status	+	+	+	+	+	+	-	-	-	-	-	-
hTERT					Up	Up	Up	Up	Up	Up	Up	Up
p53		Down*	Up**	Down*		Down*					Down*	Down*
H-Ras V12										Up		Up

* GSE insertion: cause inactivation of p53
** p53 activity is increased in fibroblasts during senescence [3]

[1] Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel. .
E-mail: yuval.tabach@weizmann.ac.il
[2] Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel
[3]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 76100, Israel.

Figure 1: The mean expression of the 3 clusters that are correlated with the malignance process. The clusters were identified by unsupervised analysis using the SPC procedure. The table represent the status of 3 manipulated genes: p53 that was repressed using GSE, hTERT, inserted to the cells in day 40, and Ras, inserted into two samples: with and without GSE. Cluster G7: 397 down regulated genes, that correlated with INK4A locus expression and anti-correlated with cell proliferation rate. Cluster G4: 250 up regulated genes, that anti-correlated with INK4A locus expression and with Cluster G7; correlated with cell proliferation rate. Cluster G8: A cluster of 168 genes, which showed associated with p53 activity, cell proliferation and tumor aggressiveness.

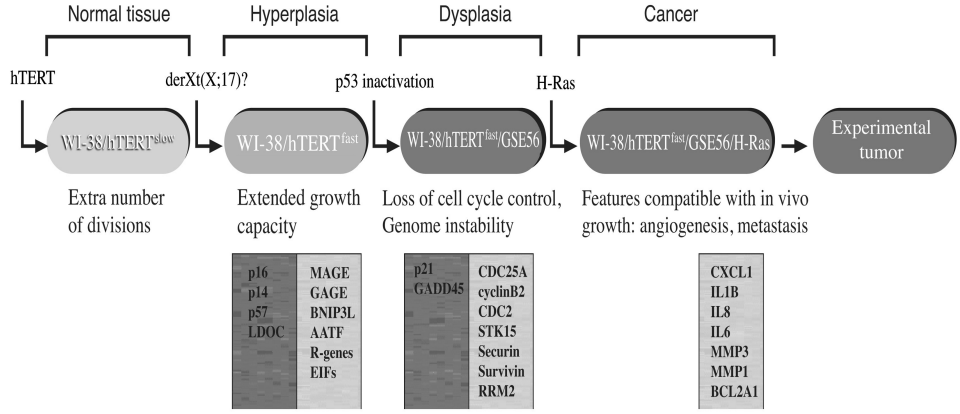


Figure 2: stepwise malignant transformation model of the human diploid fibroblast and the underlying transcriptional changes. Microarray profiling revealed specific genetic signatures, associated with the particular stages in the in vitro transformation model (selected genes in the boxes are colored according to their expression level: bright for high expression and dark for lower expression). These alteration reflect the biological features acquired by cells spontaneously (derX, t(X;17) or induced by engineered mutations (GSE56 and H-Ras) along the process. We hypothesize that genetic signatures identified by our study provide a conceptual framework of similar transcriptional alterations associated with transition from normal tissue to hyperplasia, dysplasia and then to cancer.

Data Analysis

Gene expression analysis was done on 12 data points with two replicates for each one of the data points. For probe-level data analysis, Affymetrix® Microarray Suite software 5.0 (MAS 5.0.) was used as follows. First, the abundance of each transcript represented on the array was estimated and was labeled as either present, absent, or marginal. Then absent /present and variance filters were applied; only those genes that were present in at least one data point in both repeats and had variance > 2 were kept. For each gene (row), the $\log_2$ of the expression was mean-centered (subtracting the average) and normalized to generate the final gene expression matrix. SPC [2], ANOVA, T-test and fold change were used to analyze the data.

References

1. Milyavsky, M., et al., Prolonged culture of telomerase-immortalized human fibroblasts leads to a premalignant phenotype. Cancer Res, 2003. 63(21): p. 7147-57.
2. Blatt, M., S. Wiseman, and E. Domany, Superparamagnetic clustering of data. Physical Review Letters, 1996. 76(18): p. 3251-3254.
3. Atadja, P., et al., Increased activity of p53 in senescing fibroblasts. Proc. Natl. Acad. Sci. USA, 1995. 92: p. 8348-8352.

G24. Tight Clustering: a method for extracting stable and tight patterns in expression profiles

George C. Tseng[1] and Wing H. Wong[2]

Keywords: microarray, cluster analysis, scattered genes, data mining

Abstract

We propose a method for clustering that produces tight and stable clusters without forcing all points into clusters. Many existing clustering algorithms have been applied in microarray data to search for gene clusters with similar expression patterns. However, none has provided a way to deal with an essential feature of array data: many genes are scattered randomly and do not belong to any of the significant biological functions (clusters) of interest. In fact, most current algorithms have to assign all genes into clusters. For many biological studies, however, we are mainly interested in the most informative, tight and stable clusters with sizes of, say, 20-60 genes for further investigation. Tight Clustering has been developed specifically to address this problem. The tightest and most stable clusters are identified in a sequential manner through an analysis of the tendency of genes to be grouped together under repeated resampling. We validated this method in the expression profiles of the Drosophila life cycle and mouse embryonic development. The result is shown to better serve biological needs in microarray analysis.

1. Methods

1.1 Algorithm A

The following algorithm is used to select candidates of tight clusters when the number of clusters *k* in the *K*-means algorithm is pre-specified. The subsampling procedure is used to create variabilities so that a pair of points stably clustered together can be distinguished from those clustered by chance.

(a) Take a random subsample X' from the original data X, say with 70% of the original sample size. Apply *K*-means with the pre-specified *k* on X' to obtain cluster centers $C(X', k)=(C_1, C_2,\ldots,C_k)$.

(b) Use the clustering result C(X',k) as a classifier to cluster the original data X according to the distances from each point to the cluster centers. Following the convention of Tibshirani et al. [1], the resulting clustering is represented by a co-membership matrix D[C(X',k), X] where $D[C(X',k), X]_{ij}$, the element of the matrix in row *i* and column *j*, takes value 1 if point *i* and *j* are in the same cluster and 0 otherwise.

(c) Repeat independent random subsampling *B* times to obtain subsamples $X^{(1)}, X^{(2)},\ldots,X^{(B)}$. The average co-membership matrix is defined as $\overline{D}$ =mean($D[C(X^{(1)},k), X],\ldots, D[C(X^{(B)},k), X]$).

(d) Search for a set of points $V=\{v_1,\ldots,v_m\}\in\{1,\ldots,n\}$ such that $\overline{D}_{v_i v_j} > 1-\alpha\ \forall i,j$, where α is a constant close to 0. Order sets with this property by size to obtain $V_{k1}, V_{k2},\ldots$. These V sets are candidates of tight clusters.

1.2 Sequential identification of tight clusters

[1] Department of Biostatistics and Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, 15260. Email: ctseng@pitt.edu

[2] Department of Statistics and Department of Biostatistics, Harvard University, Cambridge, MA, 02138. Email: wwong@hsph.harvard.edu

The following algorithm is used to identify a tight cluster that is stably chosen by consecutive *k*. We first define a similarity measure of two sets V_i and V_j to be $s(V_i, V_j)=|V_i \cap V_j|/|V_i \cup V_j|$ where $|V|$ is the size of set V.

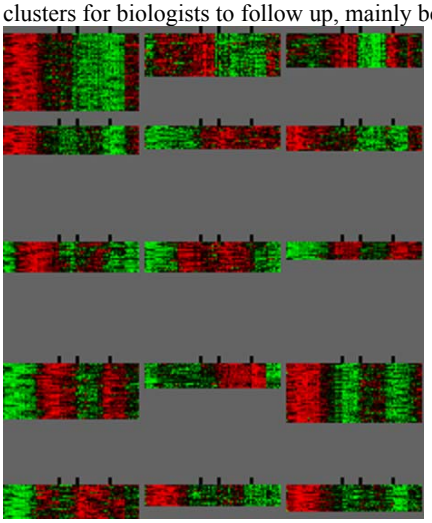
(a) Start with a suitable k_0. Apply Algorithm A on consecutive *k* starting from k_0. Choose the top 3 tightest clusters for each k, namely $\{V_{k0,1}, V_{k0,2}, V_{k0,3}\}$, $\{V_{k0+1,1}, V_{k0+1,2}, V_{k0+1,3}\}$,…

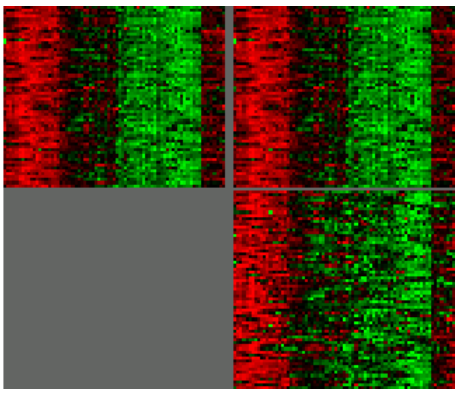
(b) Stop when $s(V_{k',l}, V_{(k'+1),m})>\beta$. Here β is a constant close to 1, $k' \geq k_0$ and $l,m \in \{1,2,3\}$. Identify $V_{(k'+1)m}$ as the tightest and most stable cluster. Remove it from the whole data.

(c) Decrease k_0 by 1 and repeat step (a) and (b) to identify the next tightest cluster.

2. Result

We applied our algorithm to a cDNA microarray data [2]. In Figure 1., the heat map [3] of 15 tight clusters when α =0.1, β =0.6, B=10 and k_0=25 is presented. The four life cycle periods are separated by black marks above the heat map. Figure 2. gives a side-by-side comparison of Tight Clustering and *K*-means algorithm. The left cluster is the first cluster identified by Tight Clustering in Figure 1. The right cluster is the corresponding cluster in *K*-means clustering when *k*=15. The two clusters have 61 common genes that were ordered and shown in the upper region. *K*-means, however, includes additional 67 genes with more variable patterns in the cluster and is likely to introduce many more false-positives. This figure shows the ability of Tight Clustering to produce tight and informative clusters for biologists to follow up, mainly because it does not require assigning all genes into clusters.





The method is further applied to a set of mouse embryonic development expression profile (data not yet published). Tight Clustering identifies a cluster of 26 genes containing seven mini chromosome maintenance (MCM) deficient genes. When using *K*-means with *k*=30, 50, 70, the resulting clusters containing these MCM genes are much larger (96, 60, 77 respectively). For *k*=100, MCM genes were distributed in two different clusters (size 31 and 15), making it harder to detect the co-regulation of the MCM genes.

3. References

[1] R. Tibshirani, G. Walther, D. Bostein, and P. O. Brown (2001). "Cluster validation by prediction strength.", Technical report, Department of Statistics, Stanford University.

[2] M. Arbeitman, E. Furlong, F. Imam, E. Johnson, B. Baker, R. Davis. and K. White (2002). "Gene expression during the life cycle of Drosophila melanogaster." *Science* **297**, 2270-2275.

[3] M. B. Eisen (2000). "TreeView (version 1.5)." Software download: http://rana.lbl.gov/

G25. A Method for 3D Visualization of Microarray Data

L. G. Volkert,[1] M. Tamboli,[1], P. Siddula,[1] and J. I. Maletic[1]

Keywords: microarray analysis, microarray visualization, data visualization, 3D visualization

1 Introduction.

Microarray based technologies allow researchers to simultaneously measure the expression levels of thousands, and even tens of thousands of genes, providing a new and powerful means of discovering, characterizing, and analyzing genes and their expression patterns. The need for new types of analysis tools is warranted by the exponential increase in the number of gene expression experiments being conducted, and thus concomitantly by the amount of data generated. Existing visualization tools do not provide the analyst a simple means for viewing the complex relationships present in the data. We present a novel visualization method that aims to provide such a mechanism in a way flexible enough to allow the user to choose a combination of visual metaphors that provide the most explanatory view of the data for the question at hand. Our approach utilizes a method for 3D visualization originally developed for aiding in the analysis of large software systems [1]. An initial implementation of this concept currently supports the use of up to nine visual metaphors such as (x,y)-placement, height, and color. The large scale and scope of microarray data make it an ideal domain for utilization of this type of data visualization tool.

2 Visual metaphors.

Our tool renders a three dimensional (3D) visualization to support the simultaneous display of multiple attributes on partitioned information spaces. The tool is decoupled from the data source so that it may be utilized as a standalone generic data visualization tool approriate for a wide variety of data domains. This is possible because the input data is an XML file containing the mapping information for rendering the supplied data attributes with user selected visual metaphors. Data attribute values can be mapped to various visual metaphors including a poly-cylinder, a container (group of poly-cylinders), the height or depth of a cylinder, and the cylinder color, shape, texture and relative position within a container. Certain visual metaphors are better for presenting enumerated data versus continuous values and we have developed guidelines for proper mapping for the user.

We present a simple example using a subset of the first 400 probes listed after the housekeeping probes from three Affymetrix GeneChip Mouse Expression Array 430A chips. In the following discussion we use the terminology developed by the Minimum Information About Microarray Experiments (MIAME) standards [2]. Each of the 400 reporters are aranged in identical relationship to each other as poly-cylinders in three seperate contaners, each container representing data from a different hybridization. The poly-cylinders oriented in the downward direction represent reporters with an *Absent* detection value as obtained from the unfiltered .txt file produced by the Affymetrix image anaylsis process. The polycylinders oriented in the upward direction represent reporters with either a *Present* (dark) or a *Marginal* (light) detection value.

[1]Department of Computer Science, Kent State University, Kent, OH, 44242. E-mails: volkert@cs.kent.edu, mtamboli@cs.kent.edu, psiddula@cs.kent.edu, jmaletic@cs.kent.edu

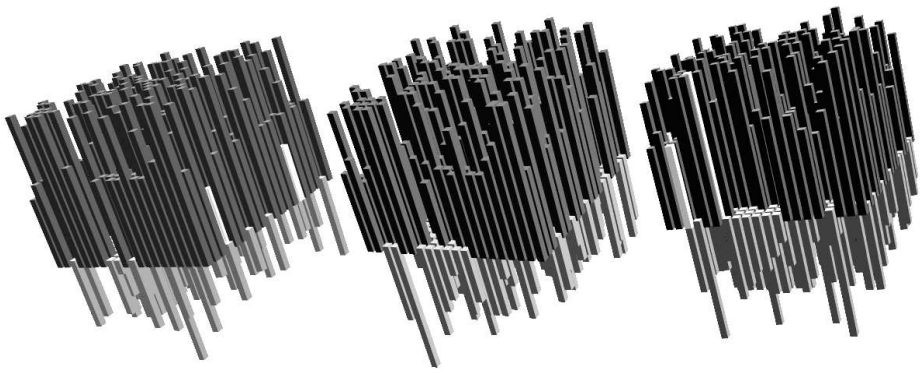


Figure 1: An example of a 3D data visualization of 400 reporters from three hybridizations of an Affymetrix 430a Mouse genechip. The poly-cylinders oriented in the upward direction represent the signal expression levels for*Present* (dark) or *Marginal* (light) reporters and those oriented in the downward direction represent signal expression levels values for *Absent* reporters

3 Discussions and directions of future work.

Complex fuctional relationships also exist among the genes represented on a chip. A set of genes may be related in terms of belonging to the same regulatory pathway or the same protein family. Thus a potentially interesting view would be to cluster data into functional profiles as per the molecular function, cellular component, and biological process. Genes with similar molecular function could be grouped together in containers representing different functions, which would potentially aid in recognizing similarities and differences in the expression profiles of groups of genes. The possibilities for useful mappings are only limited by the imagination of the user.

Our tool, while still a prototype, is quite robust and efficient. It allows for rendering of over 100,000 poly-cylinders while still maintaining quick user interaction. Users can select individual objects in the visual presentation space and manipulate them separately. This is a distict advantage over full-space-only manipulation tools. Additionally, the tool supports filtering via transparency control on selected attribute values and the ability to drill down to the underlying input data. We are currently in the process of developing a collection of data preperation scripts to automatically construct a variety of XML input files from standard Affymetrix chips. These scripts will then be embedded into a GUI that will enable easy access to this exciting new visualization method.

References

[1] Marcus, A., Feng, L., Maletic, J. I. 2003 "3D Representations for Software Visualization" In: *Proceedings of the ACM Symposium on Software Visualization (SoftVis 2003),* San Diego, CA. pp. 27-36

[2] A Brazma, P Hingamp, J Quackenbush, G Sherlock, P Spellman, C Stoeckert, J Aach, W Ansorge, C A Ball, H C Causton, T Gaasterland, P Glenisson, F C P Holstege, I F Kim, V Markowitz, J C Matese, H Parkinson, A Robinson, U Sarkans, S Schulze-Kremer, J Stewart, R Taylor, J Vilo & M Vingron, 2001 "Minimum information about a microarray experiment (MIAME)toward standards for microarray data" *Nature Genetics,* 29:365-371.

G26. A Robust, Noise-Insensitive Variable Selection Algorithm for Molecular Profiling Data

Michael Wagner[1] **and Zhongming Yang** [1]

Keywords: variable selection, L_1-SVM, molecular expression profiling.

1 Introduction

Variable Selection is a standard, well-known combinatorial problem in machine learning with wide applicability and a number of successful published algorithms (see, e.g., [3] for a recent overview). Given a number of descriptors (e.g., gene or protein expression measurements) for samples and a classification of the samples, the variable selection problem consists of determining which combination of descriptors (genes) constitutes the "best" set of predictors of class membership of the samples. Variable Selection has an obvious and important application for biomedical research. Molecular profiling of tissues and blood samples for mRNA expression levels (using microarray technology) and, more recently, for protein expression (using chromatographic affinity or antibody chips and mass spectrometry, see, e.g., [4]) have enabled the simultaneous measurement of thousands of molecular expression levels. One obvious use of this kind of data is to determine the combination of genes or proteins which can be used to diagnose disease.

Besides the general difficulties with the variable selection problem due to its combinatorial nature, biological data presents additional complications. It is inherently very noisy, both due to technical variability in the measurements and naturally occurring biological variability. Secondly, sample sizes are typically small, and the numbers of descriptors (genes, proteins) very large, making it especially difficult to obtain statistically significant results.

2 L_1-based Variable Selection

We present a new algorithm based on a variant of the well-known and widely popular linear support vector machine. Using the L_1-norm to measure the size of the weight vector instead of the standard Euclidean norm is well-known to yield sparse discriminating hyperplanes[1]. As an illustration of this, the left panel of Figure 1 shows the final values of the feature weights w_i for an implementation of the standard 2-norm SVM (libsvm[2]) as well as the L_1-SVM on the Wisconsin breast cancer testproblem[1], where 30 clinical features were used to describe 568 patients. One immediate scheme to perform dimensionality reduction on a given dataset now is as follows: Solve the L_1-SVM problem once on the entire dataset, filter out all features with small $|w_i|$, then solve again, repeat if appropriate. However, precautions must be taken in order to make this a reasonable idea: the data must be scaled appropriately in order to ensure that two w_i's are indeed comparable, and, secondly, we should like to account for the fact that the data will in general be noisy, and thus we want to ensure that we are not dealing with artifacts created by noise. Our algorithm now deliberately adds a controlled amount of noise to the data and ranks features based on their sensitivity to this additional noise, as manifested in the distribution of the resulting weight components w_i. In short, we propose ranking the features according to the z-scores of their weights from several L_1-SVM runs on perturbed data.

[1]Cincinnati Children's Hospital Research Foundation, Cincinnati, OH. E-mail: [mwagner, zhongming.yang]@cchmc.org

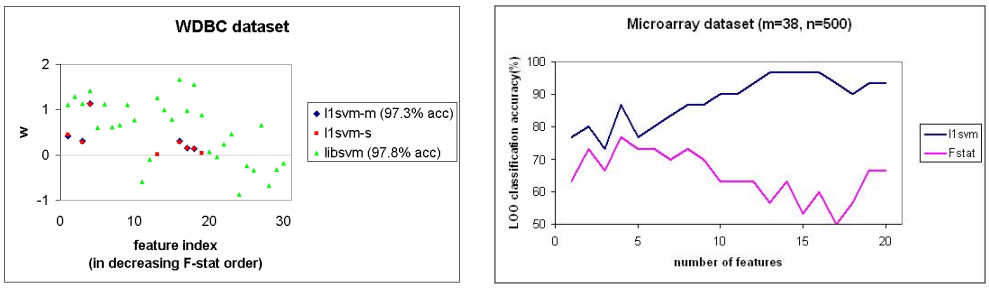


Figure 1: Left Panel: Resulting weights of L_1-SVM and a standard SVM on a testproblem. Only nonzero $w_i's$ are represented. Classification accuracies are comparable, while the L_1-SVM (l1svm-s) naturally selects a small feature subset. Points labeled with l1svm-m represent weights used as a result of running the new algorithm described here. Right panel: Comparison of variable selection strategies on a JRA microarray dataset.

3 Results

To illustrate the potential of our algorithm we present results on one dataset that motivated us to work in this direction in the first place. As part of a program project study at Cincinnati Children's Hospital, 38 patients with various forms of juvenile rheumatoid arthritis were profiled using microarrays on peripheral blood samples. The question is whether we can distinguish polyarticular JRA versus all others based on mRNA expression levels, and if yes, which combination of genes can be used to predict the classification accurately. We compare 2 variable selection methods: the commonly used F-statistic, which ranks genes based on the ratio of between group variance and within group variance, and our L_1-SVM algorithm briefly outlined above. The right panel of Figure 1 shows the cross-validated leave-one-out classification accuracies (obtained with a simple linear SVM) for both methods. As can be seen from the non-monotone behavior, the noise in this small dataset is significant, as one would typically expect monotonically increasing accuracy for increasing numbers of features. We also see that the L_1-SVM variable selection method is significantly more robust and achieves much better classification results. Experiments on several other testproblems have shown that our L_1-SVM based variable selection technique is at least as good as various competing techniques, and sometimes significantly superior.

References

[1] K. Bennett and O. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–34, 1992.

[2] C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[3] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[4] M. Wagner, D. Naik, and A. Pothen. Protocols for disease classification from mass spectrometry data. *Proteomics*, 3:1692–1698, 2003.

G27. Finding biclusters in gene expression data by random projection

Stefano Lonardi[1] **Wojciech Szpankowski**[2] **Qiaofeng Yang**[3]

Keywords: microarray, biclustering, random projection

1 Introduction and Problem Definition

Recent research has been focused on the problem of finding hidden sub-structures in large matrices composed by thousands of high dimensional vectors (see, e.g., [1, 4, 3]). This problem is known as *biclustering*. In biclustering, one is interested in determining the similarity among the vectors in a *subset* of the dimensions (subset that has to be determined as well).

Given a matrix X composed of symbols, a bicluster is a submatrix obtained by removing some of the rows and some of the columns of X in such a way that each row of what is left reads the same string. Here we are concerned with the problem of finding the bicluster with the largest area in a large matrix X. The problem is proven to be **NP**-complete. We present a fast and efficient randomized algorithm that discovers the largest bicluster by random projections.

2 Methods

Assume that we are given a large matrix $X \in \Sigma^{n \times m}$ with $|\Sigma| = a$ in which a submatrix $X_{(R^*,C^*)}$ is implanted. Assume also that the submatrix $X_{(R^*,C^*)}$ is maximal. To simplify the notation, let $r^* \equiv |R^*|$ and $c^* \equiv |C^*|$. Given a selection of rows R, we say that a column j, $1 \leq j \leq m$, is *clean* with respect to R if the symbols in the j-th column of X restricted to the rows R, are identical.

The algorithm works as follows. Select a random subset S of size k uniformly from the set of columns $\{1, 2, \ldots, m\}$. Assume for the time being that $S \cap C^* \neq \emptyset$. If we knew $S \cap C^*$, then (R^*, C^*) could be determined by the following three steps (1) select the string w that appears exactly r^* times in the rows of $X_{[1:n,S \cap C^*]}$, (2) set R^* to be the set of rows in which w appears and (3) set C^* to be the set of clean columns corresponding to R^*. Since we do not know the set $S \cap C^*$ and r^*, we try every subset of S as possible candidate for $S \cap C^*$. For every subset of S, we decide the row selection R by picking the row set in which the strings w_i induced by $X_{[1:n,S']}$ appear at least $\hat{r}$ times, where $\hat{r}$ is a user defined parameter. Thereafter the clean columns restricted to the row selection R would be our column selection C. In order to restrict the number of solutions we also allow the user to define the column threshold $\hat{c}$. The algorithm performs a number of t independent trials. In each such trial, it chooses a random subset S of the columns and for each subset of S it performs a row selection and a column selection as explained above. Finally, the algorithm returns all the candidate submatrices which satisfy user defined row and column threshold. Among all the candidate submatrices, the largest one is chosen as the final solution (R^*, C^*).

[1]Dept. of Computer Science and Engineering, University of California, Riverside, CA 92521. E-mail: stelo@cs.ucr.edu

[2]Dept. of Computer Sciences, Purdue University, West Lafayette, IN 47907. E-mail: spa@cs.purdue.edu

[3]Dept. of Computer Science and Engineering, University of California, Riverside, CA 92521. E-mail: qyang@cs.ucr.edu

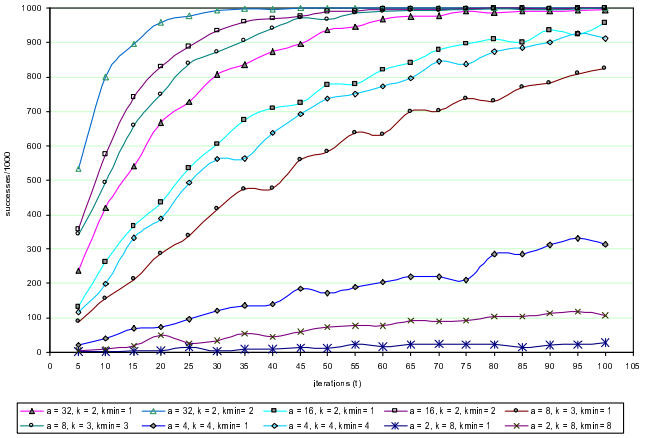


Figure 1: Comparing the performance of the randomized algorithm LARGEST_BICLUSTER when $k_{\min} = k$ versus $k_{\min} = 1$, for different choices of the alphabet size a. The projection size is $k = \log_a m$

3 Simulations

In order to evaluate the performance of the algorithms, we designed several simulation experiments. In these experiments we randomly generated one thousand 256×256 matrices of symbols drawn from an symmetric i.i.d. distribution over an alphabet of cardinality a. Then, in each we embedded a random 64×64 submatrix at random columns and random rows. We ran the algorithms for a few tens of iterations ($t = 5, \ldots, 100$), and for each choice of t we measured the number of successes out of 1,000 matrices. Figure 1 summarizes the performance of LARGEST_BICLUSTER, for several choices of alphabet size a and projection size k, and minimum subset size $k_{\min}$. In order to make a fair comparison between $k_{\min} = k$ and $k_{\min} = 1$, the number of iterations for the case $k_{\min} = k$ was multiplied by $2^k - 1$. Note that by doing so, we are assuming that one projection for $k_{\min} = 1$ takes about the same time as one projection for $k_{\min} = k$, which is not necessarily very accurate. Under this assumption, however, $k_{\min} = k$ outperforms $k_{\min} = 1$ (see Figure 1).

References

[1] CHENG, Y., AND CHURCH, G. M. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular (ISMB-00)* (Menlo Park, CA, Aug. 16–23 2000), AAAI Press, pp. 93–103.

[2] HARTIGAN, J. A. Direct clustering of a data matrix. *Journal of the American Statistical Association 67*, 337 (1972), 123–129.

[3] MISHRA, N., RON, D., AND SWAMINATHAN, R. On finding large conjunctive clusters. In *Proc. of the ACM Conference on Computational Learning Theory (COLT'03)* (2003), p. to appear.

[4] SHENG, Q., MOREAU, Y., AND MOOR, B. D. Biclustering microarray data by gibbs sampling. In *Proceedings of European Conference on Computational Biology (ECCB'03)* (2003), p. to appear.

G28. gMap: extracting and interactively visualizing nonlinear relationships of genes from expression

Chaolin Zhang[1], Yanda Li[1], Xuegong Zhang[1,*]

Keywords: gMap, microarray, interactive visualization, nonlinear projection, Isomap

1 Introduction.

DNA micro-array technologies have made it possible to monitor the expression of thousands of genes across a set of conditions or cases simultaneously, enabling us to study the functional or regulatory relationship among genes. Assuming that genes with similar expression profiles have similar functions or are in the same regulatory pathway, various methods of gene clustering and dimension reduction by projection are commonly used to extract and represent the underlying knowledge. Two critical problems of these methods are measuring the relations between genes and visualizing these discovered relations in a way that is easy for biologists' analysis.

Due to the high-complexity of gene systems, distance metrics such as correlation and Euclidean distance may not always capture the relationships among genes in the event of time shifts or the different timing rates of biological processes, etc [1].

In these cases, nonlinear manifold might exist in the high dimensional space of gene expression data. By measuring the distance of genes with geodesic distance, the nonlinear manifold can be linearized and preserved more faithfully when the data is projected into the low dimensional space. Zhou, *et al* [3] demonstrated the effectiveness of "shortest path analysis" (which is actually an equivalence of geodesic distance) in transitive functional annotation. In our study, we employed a more systematic framework of applying geodesic distance to extract the relationship among genes. An easy-to-use tool named gMap (*g* means both *geodesic* and *gene*) was developed for this purpose. Another important advantage of gMap is that it includes a number of interactive features to facilitate the scrutiny of biologists upon the expression data.

2 Methods and interactive features.

At the center of gMap is the Isomap algorithm proposed by Tenenbaum, *et al* [2]. Instead of measuring the similarity/dissimilarity of genes with correlation or Euclidean distance (or with any other distance metric) directly, we use their geodesic distance estimated by their shortest path. The authors of [2] proved that when sufficient data (in our context, number of genes) are given, the estimated geodesic distance can be arbitrarily accurate.

After obtaining the geodesic distance matrix of all gene pairs, we project the relationship of genes into the low dimensional embedding. Particularly for micro-array data analysis, gMap can display genes in a 2D or 3D scatter-plot composed of 2 or 3 major components. A number of flexible interactive features are developed based on the scatter-plots, such as capture, query, sort, cluster, etc, which facilitate the interpretation of results. A part of them is list as follows: (1)Capture. To

[1] MOE Key Lab of Bioinformatics / Dept of Automation, Tsinghua University, Beijing 100084, China.
[*] Corresponding author, email: zhangxg@tsinghua.edu.cn. This work is partially supported by NSFC under grant numbers 60275007 and 60234020.

interpret the biological meaning of a component or the local pattern of a set of genes in the embedding, users can select any gene(s). These selected/captured genes are marked out in the scatter-plot and more information of them, such as gene name, descriptions and expression profiles can be given and visualized. (2) Query. When users have a set of interested genes in advance, they can examine/query these genes to investigate their distribution in the embedding and in turn their relationship. (3)Sort. This feature is extremely useful for cell-cycle data. It allows users to re-order all (cell-cycle related) genes according to their phases in the life period.

3 Applications and results.

To demonstrate the utility of gMap, we applied our tool to a simulated dataset and three typical real micro-array datasets. The data sets are: CAKE model (a simulated genetic network), colon cancer, yeast cycle and response of human fibroblasts to serum. References of data source are not listed here due to the limit of space.

We found that gMap can preserve more information in the low dimensional space because of the linearization effect of geodesic distance, which makes it possible to identify weak relationships or patterns in terms of correlation or Euclidean distance measure. Figure 1 shows the comparison of gMap and a previous classical method MDS in the residual variance of projection, which reflects ratios of information preserved after projection. The residual when the order is 2 or 3 by gMap is less than the counterpart by MDS, especially when array number is not so small. Larger residual at the tail is due to the approximation of geodesic distance. We also observe clusters of genes are more compact in the low dimensional embeddings by gMap compared with MDS.

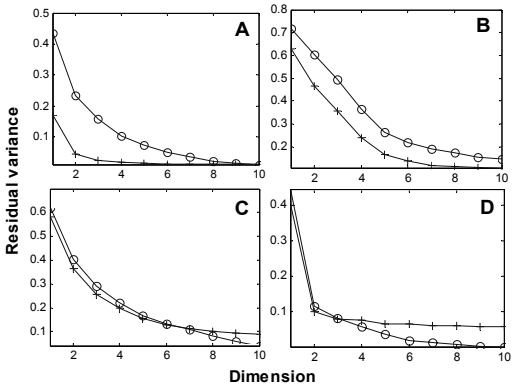


Figure 1: Comparison of gMap and MDS in the residual variance of projection. (+) gMap and (o) MDS. A. CAKE model, B. colon cancer, C. yeast cell cycle, D. response of human fibroblasts to serum

References

[1] Bar-Joseph,Z., Gerber,G., et al. (2003). Continuous Representations of Time Series Gene Expression Data. *J. Comput. Biol.* 3-4, 341-356

[2] Tenenbaum, J. B., Silva,V. D. et al. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*. 290, 2319-2323.

[3] Zhou, X., Kao, M. J. and Wong, W. H. 2002, Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA*. 99, 12783-12788

G29. Efficient Selection of Unique and Popular Oligos for Large EST Databases

Jie Zheng [1], **Timothy. J. Close** [2], **Tao Jiang** [3], **Stefano Lonardi** [4]

Keywords: oligo design, EST analysis, pooled oligo probes, string algorithms

1 Introduction.

EST databases have grown exponentially in recent years and now represent the largest collection of genetic sequences. An important application of these databases is that they contain information useful for the design of gene-specific oligonucleotides (or simply, oligos) that can be used in PCR primer design, microarray experiments, and genomic library screening.

We study two complementary problems concerning the selection of short oligos, *e.g.*, 20–50 bases, from a large database of tens of thousands of EST sequences: (i) selection of oligos each of which appears (exactly) in one EST sequence but does not appear (exactly or approximately) in any other EST sequence and (ii) selection of oligos that appear (exactly or approximately) in many ESTs. The first problem is called the *unique oligo* problem and has applications in PCR primer and microarray probe designs. The second is called the *popular oligo* problem and is useful in screening genomic libraries (such as BAC libraries) for gene-rich regions.

A large family of pattern discovery algorithms has been proposed and implemented, for instance, PROJECTION [4, 3], VERBUMCULUS [1, 2], among others. Although these tools perform well on small and medium-size datasets, they cannot handle large datasets such as the barley EST dataset.

2 Methods and Results

We present an efficient algorithm to identify all unique oligos in the ESTs and an efficient heuristic algorithm to enumerate the most popular oligos. Our strategy for unique oligos is to eliminate all the those l-mers that cannot be unique oligos. First, we cluster all the possible seeds from the ESTs into groups such that within each group, a seed has no more than one mismatch with the other seeds. Then, we check whether extending the flanking regions of a seed would result in a d-match with the corresponding extension of any other seed in the same group. If so, the l-mer given by this extension is not a unique oligo. For popular oligos, we cluster the l-mers in the EST sequences into groups by their cores, enumerate candidate l-mers by comparing the members of each cluster in a hierarchical way, then filter out candidates with unsuitable GC content, melting temperatures, secondary structures, etc, finally select and output a list of oligos. An outline of the algorithm is illustrated in Figure 1.

[1]Department of Computer Science & Engineering, University of California, Riverside, CA 92521, USA. E-mail: zjie@cs.ucr.edu

[2]Department of Botany & Plant Sciences, University of California, Riverside, CA 92521, USA. E-mail: timothy.close@ucr.edu

[3]Department of Computer Science & Engineering, University of California, Riverside, CA 92521, USA. E-mail: jiang@cs.ucr.edu

[4]Department of Computer Science & Engineering, University of California, Riverside, CA 92521, USA. E-mail: stelo@cs.ucr.edu

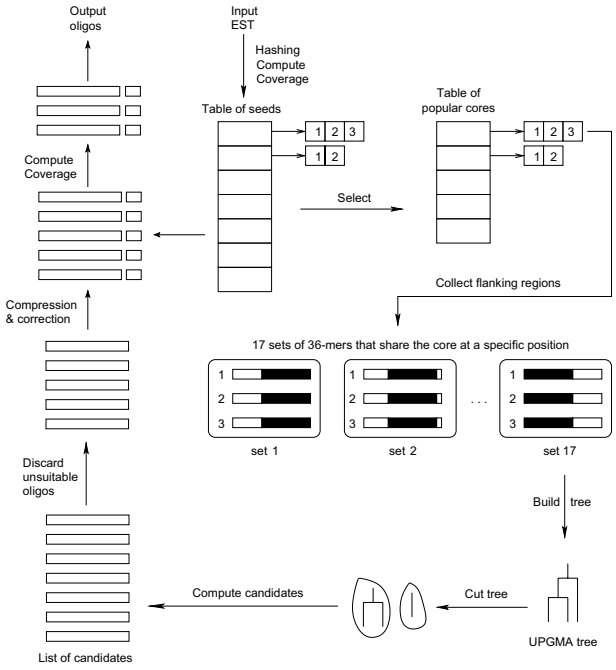


Figure 1: An overview of the algorithm for selecting popular oligos. For convenience of illustration, the length of the oligos is assumed to be $l = 36$, and the length of the cores is assumed to be $c = 20$.

3 Results

By taking into account the distribution of the frequencies of the words in the EST database, the algorithms have been carefully engineered to achieve remarkable running times on regular PCs. Each of the algorithms takes only a couple of hours (on a 1.2 GHz CPU, 1 GB RAM machine) to run on a dataset 28 Mbases of barley ESTs from the HarvEST database. Our results on synthetic data and the barley EST database show that the average relative error for popular oligos is below 2%.

References

[1] Apostolico, A., Bock, M. E., and Lonardi, S. Monotony of surprise and large-scale quest for unusual words (extended abstract). In *Proc. of Research in Computational Molecular Biology (RECOMB)* (Washington, DC, April 2002), G. Myers, S. Hannenhalli, S. Istrail, P. Pevzner, and M. Waterman, Eds. Also in *J. Comput. Bio.*, 10:3-4, (July 2003), 283–311.

[2] Apostolico, A., Gong, F., and Lonardi, S. Verbumculus and the discovery of unusual words. *Journal of Computer Science and Technology*, 1 (2004).

[3] Buhler, J., and Tompa, M. Finding motifs using random projections. *J. Comput. Bio. 9*, 2 (2002), 225–242.

[4] Tompa, M., and Buhler, J. Finding motifs using random projections. In *Annual International Conference on Computational Molecular Biology* (Montreal, Canada, Apr. 2001), pp. 67–74.

G30. Gene Co-regulation vs. Co-expression

Ya Zhang[1], Hongyuan Zha[2], James Z. Wang[3], Chao-Hsien Chu[4]
The Pennsylvania State University, University Park, PA 16802

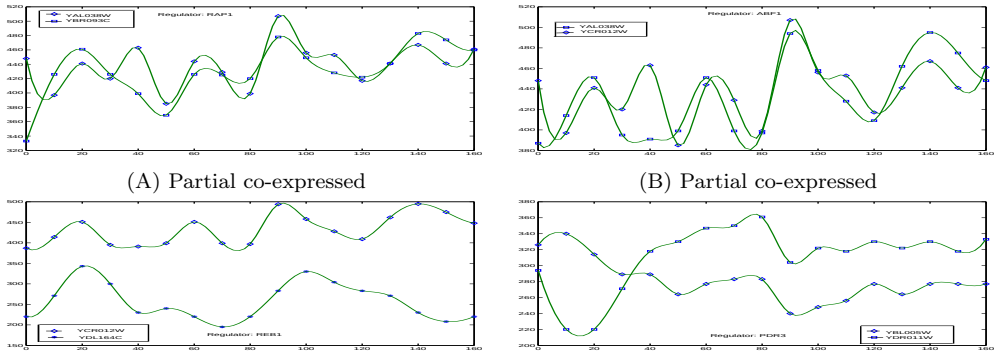
Keywords: Microarray, gene expression, co-regulation, regulon, regulator

One important goal of analyzing gene expression data is to discover co-regulated genes. For a relatively long time, it has been assumed that similar patterns in gene expression profiles usually suggest relationships between the genes. According to [3], genes targeted by the same transcription factors tend to show similar expression patterns along time. Analyzing expression profiles of genes targeted by the same transcription factors revealed complex relationships between co-regulated gene pairs, including co-expression, time shifted, inverted, and inverted and time-shifted relationships. To investigate how co-regulation corresponds to co-expression, we retrieved regulator-regulon pairs from the Yeast Promoter Database [1] and examined the expression profiles of the regulons with the same regulator. We plotted expression profiles of target genes regulated by several regulators, respectively (Fig. 2). The gene expression data were generated by Cho et al[2]. They sampled 17 time points at 10 minutes time interval, covering nearly two full cell cycles of yeast *Saccharomyces cerevisiae*. As shown from the plots, the relationships among co-regulated genes are very complex and beyond the description of the four relationships identified by Yu et al [3]. We further identified the partial co-expression relationship between genes: gene profiles may simultaneously rise and fall in a sub-range of the time course rather than the overall time course. For example, among the set of genes regulated by $ABF1$, genes $CDC19$ ($YAL038W$) and $PGK1$ ($YCR012W$) show similar expression profile in the second half of the time course (Figure 1(B)). Partial time-shift relationship is exemplified by genes $PGK1$ ($YCR012W$) and $CDC9$ ($YDL164C$) (Figure 1(C)). Moreover, we also observed partial inverted and partial co-expression relationship between genes $PDR3$ ($YBL005W$) and $SNQ2$ ($YDR011W$) (Figure 1(D)). The above observation suggests that a regulator may only function in some particular stage of cell development, and therefore genes may be co-regulated in part of the time course, which corresponds to certain phase of cell cycle or cell development.

A gene may be regulated by multiple regulators. When genes sharing multiple regulators, their expression profiles are more likely to be similar. For example, genes HO ($YDL227C$), $CLN1$ ($YMR199W$), and $CLN2$ ($YPL256C$), who share regulator $CCBF$, SCB, and $SWI6$, exhibit close expression patterns (Figure 2). However, for some genes regulated by several regulators, its expression profile may only reflect the effect of a dominate regulator. For example, regulators of gene HO ($YDL227C$) include $BAS1$, $CCBF$, $MATalpha2$, and SCB. However, according to the plots of gene expression profiles, HO ($YDL227C$) is highly correlated genes regulated by $CCBF$ and SCB (Figure 2). Similar phenomena is shown by gene $PGK1$. The regulator of $PGK1$ ($YCR012W$) includes $GCR1$, $CPF1$, and $ABF1$. In this case, $PGK1$ ($YCR012W$) showed similar expression profile to some genes regulated by $GCR1$ and $ABF1$ in the second half of the time course (Figure 2).

Current analysis of microarray gene expression focuses on relationships based on overall correlation in expression profile, identifying clusters of genes whose expression levels simultaneously rise and fall throughout a time course. However, genes may be regulated by different regulators over a time course. Co-regulating in part of the time course does not guarantee

[1]School of Information Sciences and Technology. E-mail: yzhang@ist.psu.edu
[2]Department of Computer Science and Engineering. E-mail: zha@cse.psu.edu
[3]School of Information Sciences and Technology. E-mail: jwang@ist.psu.edu
[4]School of Information Sciences and Technology. E-mail: chu@ist.psu.edu



(A) Partial co-expressed (B) Partial co-expressed

(C) Partial time shifted (D) Partial inverted and co-expressed

Figure 1: Expression profiles for co-regulated genes. The time-series gene expression data are from [2].

a global similarity in gene profiles. Therefore, new clustering algorithms are needed to address this issue. Several biclustering algorithms have been proposed to discover sets of genes that co-regulated in only part of the experiments conditions under study. However, these algorithms are not applicable to clustering gene expression time-series data because they ignored the internal relationship between time points. When analyzing the time-course gene expression data, it is necessary to consider the internal connection between time points and preserve the time locality in time-course gene expression data.

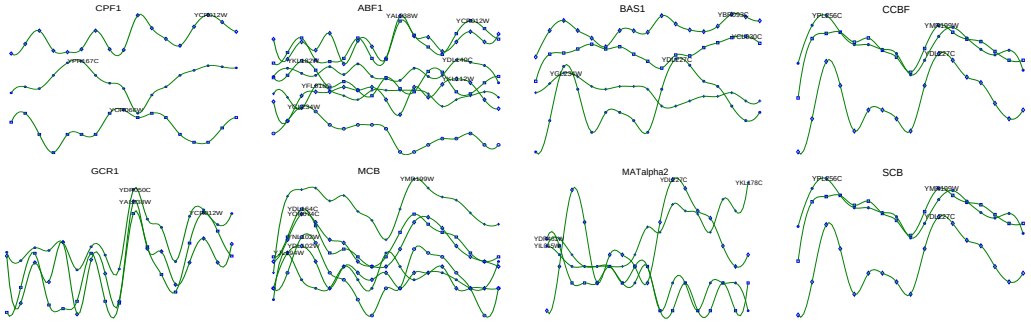


Figure 2: Expression profiles of co-regulated gene groups. Each curve represents the expression profile of a gene. Each sub-plot represents gene expression profiles of a co-regulated gene group. The gene expression time-series data are obtained from [2]. The time range is from 0 min to 160 min.

References

[1] http://cgsigma.cshl.org/jian/

[2] Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L. et al. (1998) A genome-wide transcriptional analysis of the cell cycle. *Mol. Cell,* 2, pp.65-73.

[3] Yu,H., Luscombe,N., Qian,J., and Gerstein,M. (2003) *Trends Genet* 19(8):422-7.

G31. Deconvolution of cDNA Microarray Images and Significance Testing for Gene Expression Levels

Hye Young Kim[1], Min Jung Kim[1], Yong Sung Lee[2], Young Seek Lee[3], Tae Sung Park[4], Ki Woong Kim[4], and Kim Jin Hyuk[1]

Keywords: microarray, deconvolution, significance testing

1 Introduction.

The functional characterization and optimization of the microarray scanner are essential for the accurate quantitation of microarray data [1]. The most commonly used fluorescent dyes, Cy3 and Cy5, are relatively unstable, have different incorporation and quantum efficiencies, and are detected by the scanner with different efficiencies [2]. For these reasons, most microarray images are produced with different scanner settings, such as laser power and photomultiplier tube (PMT) gain between chips and between fluorescence dyes. However, there is neither a criterion for adjusting the scanner settings nor a proved relationship between the intensities of pixels in the microarray image and the 2-D fluorophore concentration (the quantity of fluorophores in unit area). Therefore, this manual adjustment must be checked for the accurate conversions of gene expression levels into the pixel intensities.

2 Software and files.

Several research papers indicate that the microarray scanner has dynamic range and that normalization between fluorescence labels is non-linear and depends on slides [3]. The relationship between fluorophore quantities and intensity reported by a scanner is linear only within a certain range of intensities, being dominated by noise below and subject to saturation above that range. There are many differences between the photochemical characteristics of both fluorescent tags. Therefore, the identical scanner settings for the Cy3 and Cy5 images do not convert the 2-D fluorophore concentration representing the gene expression levels into the images under the same condition.

Deconvolution is an image processing techniques, which is utilized for improving the contrast and resolution of digital images captured in the microscope. The synchronous deconvolution of both Cy3 and Cy5 images makes possible the reduction of the errors originated from scan process as well as the enhancement of the images. In a point of view that most gene expressions are *not* significantly up- nor down-regulated in cDNA microarray experiment, the spot intensity distributions of Cy3 and Cy5 images should be overlapped in a wide range. Thus, the synchronous deconvolution can make the equal prevalence of the spot intensities of Cy3 and Cy5 images as much as possible and increase the resolution of image intensity.

[1] Department of Physiology, College of Medicine, Hanyang University, Seoul, 133-791, Korea. E-mail: jhkim1@hanyang.ac.kr
[2] Department of Biochemistry, College of Medicine, Hanyang University, Seoul, 133-791, Korea. E-mail: yongsung@hanyang.ac.kr
[3] Department of Biochemistry and Molecular Biology, College of Science and Technology, Hanyang University, Ansan, 425-791, Korea. E-mail: yslee@hanyang.ac.kr
[4] Department of Statistics, College of Natural Sciences, Seoul National University, Seoul, 151-741, Korea. E-mail: tspark@stats.snu.ac.kr

3 References.

[3] Dudley, A.M., Aach, J., Steffen M.A. and Church, G.M. 2002. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proceedings of the National Academy of Sciences USA* 99:7554-7559.

[2] Tseng, G.C., Oh, M.K., Rohlin, L., Liao, J.C. and Wong, W.H. 2001. Issues in cDNA microarray analysis: quality filtering, Channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* 29:2549-2557.

[1] Worley, J., Bechtol, K., Penn, S., Roach, D., Hanzel, D., Trounstine, M. and Barker, D. 2000. A systems approach to fabricating and analyzing in DNA microarrays. In: *Microarray biochip technology* (Schena, M., ed.), Sunnyvale: Eaton Publishing. pp. 65-85.

G32. A New HMM-based Clustering Technique for the Analysis of Gene Expression Time Series Data

Yujing Zeng[1], Javier Garcia-Frias[1]

Keywords: gene expression time-course data, clustering, temporal dependences, hidden Markov model (HMM)

1 Introduction.

This paper proposes a novel Hidden Markov Model (HMM)-based clustering technique for the analysis of gene expression time series data. The proposed model, called the profile-HMM, is specifically designed to explicitly take into account the dynamic nature (time dependences) of temporal gene expression profiles, which is ignored by standard clustering methods. In this model, gene expression dynamics are represented by a special set of paths, with each path characterizing a stochastic pattern. The profile-HMM is trained to represent the most likely set of stochastic patterns given the dynamic microarray data, and the clustering result is obtained by grouping together the time series that are most likely to be related to the same pattern. The novelty of the method is that the behavior of all expression data is modeled by a single HMM and all the clusters are implicitly and simultaneously defined in the model during the training procedure. An attractive property of the proposed clustering algorithm is its ability to automatically identify the number of clusters. As demonstrated on real dynamic gene expression data collected for the study on the transcriptional program of sporulation in budding yeast, the proposed method outperforms standard clustering techniques such as K-means, single-linkage and average-linkage.

2 Simulation.

We utilized the experimental data from [1], which reported the result of a study on the transcriptional program of sporulation in budding yeast. The study used DNA microarrays, which contained 97% of the known or predicted genes of Saccharomyces cerevisiae, to explore the temporal program of gene expression during meiosis and spore formation. Changes in the concentrations of mRNA transcripts from each gene were measured at 7 uneven time intervals. In their original paper, the authors identified 477 genes grouped in 7 clusters, with each cluster associated with a different temporal pattern of induced transcription according to visual inspection and prior studies.

When applied over the 477 genes described above, the proposed clustering method is able to determine the number of clusters automatically. Moreover, the number of clusters is relatively stable with respect to the model complexity, as long as the complexity stays in a reasonable range: when the total number of states in the profile-HMM ranges between 49 and 350, the resulting number of clusters is always between 16 and 21, and all the resulting clustering structures are very similar. In the rest of this section, we consider the result of the simplest model (49 states, which results in 16 clusters) for further analysis.

In order to validate the proposed clustering method, 3 widely used algorithms, K-means, single-linkage, and average-linkage hierarchical clustering, were also implemented on the same dataset. The number of clusters was always fixed to 16, so that we could compare the performance of these 3

[1] Department of Electrical and Computer Engineering, University of Delaware, Delaware, USA. E-mail: {zeng,jgarcia}@ee.udel.edu

methods with that of our proposed approach. Several validation indexes were calculated for all the clustering results (including the original model in [1]), and their values are shown in Table 1.

As shown in Table 1, the clustering result produced by the proposed algorithm has much better homogeneity than the single-linkage result, but worse separation. The incorporation of both properties in the validation process, provided by the DB index, seems to show a better performance for the single-linkage approach. However, by checking the details of the clustering results, we found that 14 out of 16 clusters in the single-linkage approach are singletons, and more than 96% of genes (461 out of 477 genes) are clustered into one big group. Intuitively, it is clear that such a clustering structure cannot describe the real patterns in the dataset. These results obtained by visual inspection are confirmed by the Silhouette index, which shows that the single-linkage result is the worst of all the results considered in Table 1. The average-linkage method provides a similar result with fewer singletons. However, most clusters in the average-linkage approach still have a very small number of members, and one of the clusters contains the profiles of more than 2/3 of the input genes. Similar to the single-linkage result, when compared with the profile-HMM result the average-linkage result has a worse evaluation for the homogeneity and the Silhouette indexes.

Criterion	Profile-HMM	K-means	Single-link.	Average-link.	Model in [1]
Homogeneity	0. 3222	0.2590	0.5428	0.3732	0.3240
Separation	0.9941	0.7881	1.1290	1.2279	0.8193
DB index	0.8605	1.1439	0.4201	0.4513	1.2278
Silhouette index	0.2952	0.2668	-0.1353	0.2410	0.2820

Table 1: Validation indexes for different clustering methods applied over the proposed dataset

A similar analysis was performed to compare the profile-HMM with K-means. In this case, most indexes, especially the DB and Silhouette indexes that consider the tradeoff between separation and homogeneity, show that the proposed algorithm outperforms the K-means approach. We also compared our result with the model described in the original paper [1]. The most evident difference between the two clustering approaches is in the number of clusters (16 in our method versus 7 in [1]). Usually, increasing the number of clusters leads to a better homogeneity within clusters, but may degrade the separation between different clusters. However, we can see in Table 1 that this does not occur here. In fact, our result provides more compact clusters (.3222 homogeneity coefficient versus .3249 in the original model) with a better separation index (0.994 versus 0.8193). This shows that the profile-HMM method succeeded in capturing different dynamics existing in the data while keeping each cluster homogeneous. Its advantage over the original method is also proved by all the other indexes.

In addition to the inspection of the characteristics of the clustering results as a whole, shown in Table 1, we also performed a cluster-by-cluster comparison between the results obtained by the proposed method and those of the original model. Although the details cannot be shown here due to space constraints, we found that the profile-HMM identified different patterns that get mixed in the same subgroup by the original model, and assigned them into different clusters, providing an improved description on the global patterns of the data.

References

[1] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brow, P.O., and Herskowitz, I. *The Transcriptional Program of Sporulation in Budding Yeast.* Science, 282 (23): 699-705, 1998.

G33. A Novel HMM-based Cluster Validity Index for Gene Expression Time-Course Data

Yujing Zeng[1], Javier Garcia-Frias[1]

Keywords: gene expression time-course data, clustering validation, temporal dependence, hidden Markov model (HMM)

1 Introduction.

Gene expression time-course data is usually obtained by performing microarray experiments at consecutive time points. The analysis of these data is helpful to reveal the mechanisms regulating different cellular processes. Different from other microarray data, the pattern represented by each profile is decided not only by the observations at different time-points, but also by the order of these observations. This sequential dependence is crucial for clustering processing and validation.

Our research addresses the problem of clustering validation for time-course data by proposing a novel hidden Markov model (HMM)-based clustering validity index. In the index definition, we use a specially designed HMM to model the data distribution under the constraints of the clustering result. The evaluation is calculated based on the likelihood of each time series given this HMM. The main novelty of the proposed index is its ability to take account of the temporal dependences in the sequential data. Contrary to other validity indexes, in the proposed model the observations at different time-points are not considered independently, and the dependences in each time-interval are modeled and used to evaluate the clusters quality explicitly. In other words, if these dependences change because of permutations among time-points, the validation result is able to reflect such a change accurately. The simulation discussed in next subsection was designed to test this ability.

2 Simulation.

In this experiment, we generated two datasets which have the same observations but in different order. Any object in both datasets has 9 attributes, which can be divided into two categories. The first category includes 4 attributes, whose values are i.i.d. samples from zero-mean Gaussian distributions. Since there is no useful information for clustering in these 4 features, we called them noninformative features. The rest 5 attributes, called informative features, are generated from three different basic patterns, as shown in Figure 1(a). To generate similar but distinct sequences, random variables with different distributions are added to each component of the basic patterns, which is shown in Figure 1(b, c). As shown in Figure 1(d), all the informative features are observed at the first 5 consecutive time-points in dataset I, which shows three distinctive patterns. Dataset II is generated by permuting the order of the attributes in dataset I so that each informative feature is separated by noninformative features, as shown in Figure 1(e). Since all the noninformative features are i.i.d. and unrelated with the basic patterns, the dependence between a noninformative and an informative feature is very weak, and, therefore, interleaving the two kinds of features reduces the sequential dependences in the whole data.

In order to corroborate the effectiveness of the proposed HMM-based index, the partitions with different numbers of clusters, from 2 to 10, were obtained using one of the most popular and simplest methods, K-means. As shown in Table 1, although the datasets share the same observations, the

[1] Department of Electrical and Computer Engineering, University of Delaware, Delaware, USA. E-mail: {zeng,jgarcia}@ee.udel.edu

proposed algorithm produces different evaluations for the two datasets, and suggests different optimal numbers of clusters for them.

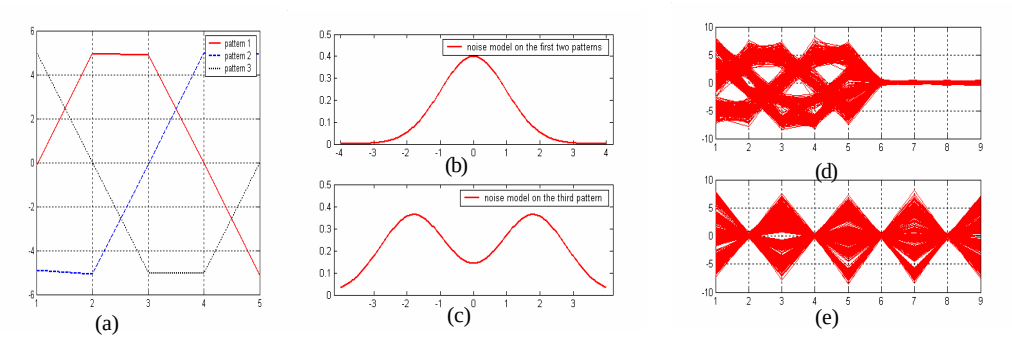


Figure1: Illustration of the generation of the simulation data. (a) three basic patterns for the informative features; (b) density function of the noise model added on the first two patterns; (c) density function of the noise model added on the third pattern; (d) simulation dataset I; (e) simulation dataset II.

According to the results of the proposed index, the optimum number of clusters for dataset I is estimated as 3, which agrees with the original model generating the data. Notice that because of the bimodal noise added to the third basic pattern, this cluster shows a more dispersive distribution than the other two and tends to be split into several groups. This is reflected in the curve of the proposed index, which displays a sub-optimum for the partition with 6 clusters. In dataset II, the sequential dependences are reduced by permutation, so that the noise model becomes overwhelming and gets emphasized in the clustering evaluation. As shown in Table 1, although a sub-optimum is shown for the partition with 3 clusters, the proposed index suggests that the optimal number of clusters for dataset II is 6.

Comparison has been performed with the results of the Silhouette [1] and Davis-Bouldin indexes [2]. Both indexes ignore the sequential dependences in the data and give the same evaluation on both datasets, which, as shown in Table 1, is similar to the result of the proposed index for dataset II.

Index	2	3	4	5	6	7	8	9	10
HMM-based index for Dataset I	0.5000	**0.7042**	0.4845	0.3590	***0.6542***	0.4793	0.3987	0.2058	0.2101
HMM-based index for Dataset II	0.5000	0.6115	0.4196	0.4960	***0.6156***	0.4473	0.3741	0.1915	0.1959
Davis-Bouldin index	0.6607	0.4407	0.8404	0.6252	***0.4306***	2.2809	2.2748	2.2252	2.2057
Silhouette index	0.5388	0.7134	0.5593	0.6688	***0.7260***	0.6536	0.4945	0.4133	0.3422

Table 1: Values of different indexes for the two simulation datasets

3 Conclusion.

We have illustrated the ability of the proposed index to capture the sequential dependence of time series data. This ability is reflected in the higher robustness of the HMM-based index to deal with data corrupted by noise models more general than the Gaussian one, which is a very attractive property in the context of gene expression data analysis.

References

[1] Davies, D.L and Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.1, pp. 224-227, 1979.

[2] Gordon, A. *Classification*. Chapman and Hall, 2nd edition, 1999.

G34. Clustering of Time-Course Gene Expression Data

Ya Zhang[1], Hongyuan Zha[2], James Z. Wang[3], Chao-Hsien Chu[4]
The Pennsylvania State University, University Park, PA 16802

Keywords: Microarray, gene expression, time series, clustering

1 Introduction.

Microarray experiments have been used to measure genes' expression levels under different cellular conditions or along certain time course. Initial attempts to interpret these data begin with grouping genes according to similarity in their expression profiles. The widely adopted clustering techniques for gene expression data include hierarchical clustering, self-organizing maps, and K-means clustering. Bayesian networks and neural networks have also been applied to gene clustering. Sharan & Shamir [3] provided a survey on this topic. Clustering techniques typically discover the inherent structure of the genes expression profiles based on some similarity measures. The clustering results largely depend on how the similarity measure corresponds to the biological correlation between genes. Before reliable conclusion about biological functions can be drawn from the data, the gene clusters obtained from microarray analysis must be investigated with respect to known biological roles of those clusters.

The current analysis of whole-genome expression focuses on relationships based on global correlation over a whole time-course, identifying clusters of genes whose expression levels simultaneously rise and fall. However, genes may be regulated by different regulators in a long time course. Co-regulating in part of the long time course does not guarantee a global similarity in gene profiles.

Biclustering of microarray gene expression data has recently been introduced by Chen & Church [1] as a means to discover sets of genes that co-expressed in only part of the experiment conditions under study. Essentially, overlapping in gene clusters is allowed, and many subtle gene clusters are revealed. Since then, several other algorithms have been developed to bicluster gene expression data [4]. However, existing biclustering algorithms do not consider the differences between time-series gene expression data and multi-condition gene expression data. The relations between time points are ignored, and the time points are clustered independently. It is marginally biologically meaningful if two genes show similar expression pattern in non-consecutive time points. It is therefore necessary to preserve the time locality in time-course gene expression data.

2 Method.

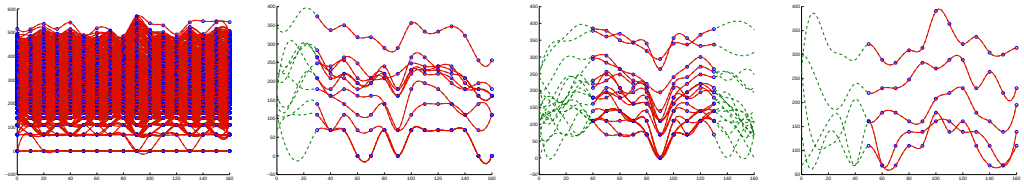
We present our time series biclustering algorithm to cluster time course microarray data. The aim of this clustering is to discover genes that are co-regulated in an interim of the time course but do not show highly correlated gene expression over the whole time course. The mean square residue score H is used as a measurement for the biclustering. While enforcing H to be smaller than a user-selected parameter δ, we try to simultaneously maximize H,

[1]School of Information Sciences and Technology. E-mail: yzhang@ist.psu.edu
[2]Department of Computer Science and Engineering. E-mail: zha@cse.psu.edu
[3]School of Information Sciences and Technology. E-mail: jwang@ist.psu.edu
[4]School of Information Sciences and Technology. E-mail: chu@ist.psu.edu

the number of genes, and the length of the time course in the cluster. The bicluster is first initialized as the entire data matrix. We adopt a deletion-based method to eliminate genes with expression profiles deviating from those of the majority. A row (gene) is removed from the bicluster if the ratio of the mean square residue score of the row to that of the bicluster larger than a user-defined threshold. Similarly, time points are removed. To ensure that the time points in a bicluster are always consecutive, in time point deletion, we only allow the deletion to be exerted on the border time points – the first and the last column in the bicluster. The deletion has demonstrated the capability to reduce the mean square residue score of the resulting bicluter [1]. The deletion stops when the mean square residue score of the resulting bicluter is less than δ. Some previously deleted genes may have strong correlation with genes in the bicluster in terms of similarity in the interim of expression profile in the bicluster. These genes are then add into the bicluster, and it is guaranteed that the addition will reduce the mean square residue score. Similarly, the time points adjacent to the border of the bicluster may be considered for addition into the bicluster.

3 Results and Discussion.

We test our algorithm on the yeast cell cycle data provided by [2]. Figure 1 presents some clusters obtained by our time series biclustering algorithm. The solid lines represent the interims of gene profiles that are in the biclusters, and the dash lines represent the deleted time points. Clearly, the variability of expression profile in the biclustered range is smaller than the that in the range of deleted time points. By deleting some time points in the two ends, we are able to discover some subtle genes clusters. However, further investigation of the gene clusters with respect to known biological roles of cluster members is desired. Further experimental confirmation may be required to reveal the true 'biological relationships' among genes.



cluster 1 cluster 35 cluster 59 cluster 83

Figure 1: Expression profile of gene clusters.

References

[1] Cheng,Y., and Church,G. 2000. Biclustering of expression data. In *Proceedings of 8th International Conference on Intelligent System for Molecular Biology (ISMB)*, pp.93-103.

[2] Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L. et al. (1998) A genome-wide transcriptional analysis of the cell cycle. *Mol. Cell,* 2, pp.65-73.

[3] Sharan,R. and Shamir,R. 2002. Algorithmic approaches to clustering gene expression data. In T. Jiang *et al.* (eds), *Current Topics in Computational Molecular Biology.* The MIT Press, pp. 269-300.

[4] Yang,J., Wang,H., Wang,W., and Yu,P. 2003. Enhanced biclustering on expression data. In *Proceedings of the 3rd IEEE Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 321-327.

G35. Improving temporal gene expression profiles with probabilistic models

Marta Milo[1], M.C. Holley[1], M. Rattray[2], M. Niranjan[3] and N. D. Lawrence[3]

Keywords: temporal gene expression profiles, high-density short oligonucleotide microarrays, probabilistic models, gamma distribution.

1 Introduction.

High-density short oligonucleotide microarrays are a widely used tool for measuring gene expression on large scale [3]. A key issue for short oligonucleotide probes, such as the one used by Affymetrix, Inc., is the way of selecting probe sequences with high sensitivity and specificity. The use of multiple probe pairs (11-20 pairs), referred as *probe set*, to target a single gene is the current approach to this problem. One of each pair exactly matched the DNA fragment of the gene (*PM probe*) and the other contains a single mismatch base in the middle (*MM probe*). The sensitivity of the gene expression signal is given by the PM values and the MM values, offering a measure of non-specific binding, improve the specificity. However, around 30% of the pairs consistently have a negative signal, indicating that the MM values are unreliable as pure measure of non-specific binding [5, 7]. Moreover the variation of the order of magnitude of the probe signals within a probe set suggests that not all the probes have optimal sensitivity [2]. Given the above limitations, it becomes crucial to design a model for the extraction of gene expression signals from oligonucleotide arrays. Many recent studies have focused on statistical methods [1, 2] to summarise these expression levels, choosing not to use the MM probes for the unreliability mentioned above. In this work we chose to use a probabilistic model to describe both the sensitivity and the specificity of the probe set including in the model the MM observed signals. The model, gamma Model for Oligonucleotide Signal (gMOS) [4] proved to perform well both on publicly available data set and on a real case study. To improve the specificity of the model we take in account the correlation between MM probes and PM probes, using a different approach that allows us to learn the parameter of the joint distribution of PM and MM from the observed data. The *modified gMOS* (mgMOS) is tested on a temporal profile of a cell line, UB/OC-1, [6] derived from epithelial cells in the cochlear duct at embryonic day 13.5 (E13.5). The extracted profiles are compared with a real time RT-PCR profile and with profiles obtained with both the Affymetrix MAS v.5 and with different statistical methods.

2 Material and Methods.

For each probe set we model the PM signal (y) and MM (m) using a gamma probability functions. The gene expression signal (s) is then derived from the joint probability of y and m. In the basic gMOS we assume independency in the probe set. The joint probability of y and m is defined as:

$$p(y_i, m_i) = p(y_i \mid a, \alpha, b)\, p(m_i \mid a, b)$$

and $i=1,...N$. N is the number of probe pairs in a probe set. The parameters a, α, b are estimated using Maximun Likelihood. The parameter b does not vary within the probe set, modeling the genetic affinity of corresponding short oligonucleotide with the gene sequence as constant within

[1] Institute of Molecular Physiology, Department of Biomedical Science, Addison Building, Western Bank, Sheffield, S10 2TN, UK E-mail: M.Milo@sheffield.ac.uk
[2] Department of Computer Science, University of Manchester, UK. E-mail: magnus@cs.man.ac.uk
[3] Department of Computer Science, Regent Court, 211 Portobello Street, Sheffield S1 4DP,UK. E-mail: niranjan@dcs.shef.ac.uk, neil@dcs.shef.ac.uk

the probe set. The mgMOS is a latent variable model that defines this genetic affinity as varying within the probe set. This difference is defined by modeling the correlation between *y* and *m* that is also detectable from the observed *y* and *m* signals. In the modified model the parameter *b* now varies within the probe set. The joint probability distribution of *y* and *m* becomes:

$$p(y_i, m_i) = \int p(y_i \mid a, \alpha, b_i)\, p(m_i, a, b_i)\, p(b_i)\, db_i$$

The equation is tractable for Gamma distributions. Given $p(b_i) = \frac{d^c}{\Gamma(c)} b_i^{c-1} \exp(-db_i)$ it is possible to calculate the joint distribution of *y* and *m* and therefore obtain the estimate of the signal *s* with the above equation. The parameters are again estimated using Maximum Likelihood.

3 Results and conclusions.

The two methods are tested on benchmark data and on a temporal profile of the transcription factor *gata3* from an inner ear cell line. The temporal profile consists in 12 time points sample in 14 days of development after differentiation. The method gives very promising results both for the application to real dataset (Figure 1) and for further theoretical exploitation.

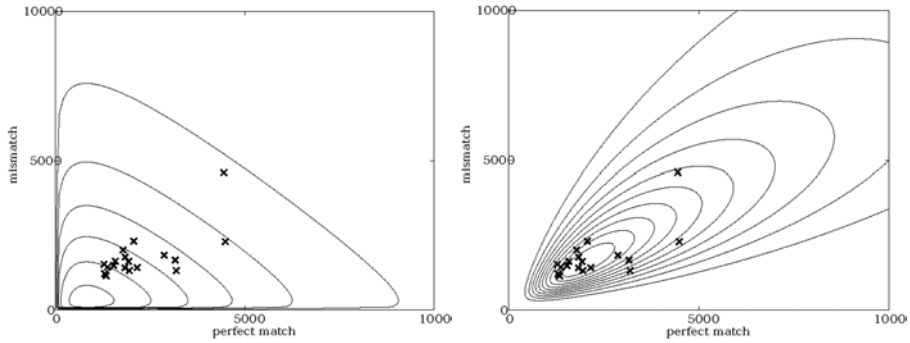


Figure1: Correlation of PM and MM for one time point of the gata3 profile. The plot the left shows the contours of the distributions obtained with gMOS; the plot on the right shows contours obtained with mgMOS. The plot clearly shows that mgMOS better fits the observed data.

References

[1] Irizarry R., B. Bolstad, F. Collin, L. Cope, B. Hobbs and T. Speed. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acid Research*, **31**, No. 4 e15.

[2] Li, C. and Wong, W.H. 2001. Model-based analysis of oligonucleotide arrays: expression index, computation and outlier detection. *Proc. Natl Acad. Sci.*USA, **98**, 31-36

[3] Lockhart, D.J. and Winzeler E.A. 2000. Gene expression and DNA arrays. *Nature*, 405, 827-836.

[4] Milo, M., Fazeli, A., Niranjan M. and N. D. Lawrence 2003. A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochemical Transations*, **31**, 6.

[5] Naef, F. Hacker, C.R., Patil N., and Magnasco M. 2002. Characterization of the expression ratio noise structure in high-density oligonucleotide arrays. *Genome Biology*, **3**, research0018

[6] Rivolta, M. N.,Hasall,A., Johnson, C.M., Tones, M.A., Holley, M.C. 2002. Transcript profiling of functionally related Groups of Genes During Conditional Differentiation of Mammalian Cochlear Hair Cell Line. *Genome Research,* **12**, 7, pp. 1091-1099.

[7] Zhou, Y. and Abagyan, R. 2003. Match-Only Integral Distribution (MOID) Algorithm for high density oligonucleotide array analysis. *BMC Bionformatics*, **3**, 3.

H1. Approximating Geodesic Tree Distance

Nina Amenta, [1] **Matthew Godwin,** [2] **Katherine St. John** [3]

Keywords: phylogeny, tree metric.

1 Introduction.

What is the distance between two possible evolutionary trees? Biologists use a variety of distance metrics [2], like RF, TBR, SPR, and NNI. The mathematicians Billera, Holmes and Vogtmann have proposed a new metric [1], which resembles the usual Euclidean metric used in geometry and statistics in a way that the other tree metrics do not: there is a unique shortest path between any two trees. Their hope is that it will help to adapt statistical techniques from Euclidean geometry, to compute better trees or to understand relationships between possible trees. Unfortunately the metric, which we shall call *geodesic distance*, is not obviously easy to compute. In this poster we observe that it is, however, easy to approximate: we give simple upper and lower bounds which differ by a multiplicative factor of at most $\sqrt{2}$.

2 Geodesic distance and the bounds

When two fully-resolved trees $T1, T2$ have the same topology, we treat the two trees as points in the Euclidean space whose coordinates are the edge weights; see Figure 1. The geodesic distance is the Euclidean distance $\left(\sum_e \delta(e)\right)^{1/2}$, where $\delta(e) = (w_1(e) - w_2(e))^2$ and $w_i(e)$ is the weight of edge e in tree Ti.

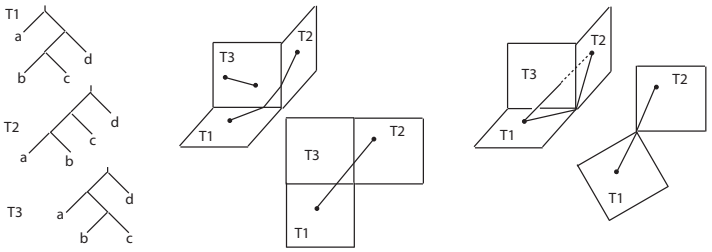


Figure 1: On the left, three tree topologies; the lengths of the two non-terminal edges in each tree form the coordinates of a planar part of tree-space. In the middle, the shortest path between two trees with topology T3 is a line segment, while between two trees with topologies T1 and T2 the shortest path includes some topology changes. The relevant parts of tree-space can be unfolded to straighten the path. On the right, the lower bound path is not constrained to lie in the tree-space. The upper bound path is constrained to go through the parts corresponding to T1 and T2 and the parts that they share. Again, the relevant parts of tree-space can be unfolded to straighten the path.

[1]Computer Science Department, University of California at Davis. E-mail: amenta@cs.ucdavis.edu. Supported by NSF ITR DEB-0121651.

[2]Department of Computer Science Department, University of California at Davis. E-mail: mjgodwin@ucdavis.edu. Supported by NSF ITR DEB-0121651.

[3]Lehman College, City University of New York. E-mail: stjohn@lehman.cuny.edu. Supported by NSF ITR DEB-0121682.

Every fully-resolved topology gives a Euclidean space. Trees with polytomies are shared by the topologies produced by different resolutions. These shared trees connect the corresponding spaces, giving the overall *tree-space*. The geodesic distance between two trees is the length of the shortest path between them in tree-space, which generally includes topology changes. If there are many possible ways to order the topology changes, finding the shortest path may be difficult.

A lower bound on the geodesic distance is $D_{lo}(T1,T2) = \left(\sum_e \delta(e)\right)^{1/2}$ where now, if e is not an edge of $T1$, $w_1(e) = 0$, and visa versa. This does not correspond to a path in tree-space, except when $T1, T2$ have the same topology.

The upper bound is given by the length of a particular path in tree-space. The path goes directly from $T1$ to a strict consensus tree of $T1$ and $T2$, and then to $T2$. By unfolding the space to straighten the path (see Figure 1) we can find the edge weights on the consensus tree which give the best bound. We get the formula $(D_{hi}(T1,T2))^2 = \left[\left(\sum_{e\in(T1-T2)} \delta(e)\right)^{1/2} + \left(\sum_{e\in(T2-T1)} \delta(e)\right)^{1/2}\right]^2 + \sum_{e\in(T1\cap T2)} \delta(e)$. The ratio D_{hi}/D_{lo} is maximized when the common edges have the same weight, and the edges in $T1-T2$ contribute the same weight a as the edges in $T2-T1$. In this case $D_{hi} = 2\sqrt{a}$ and $D_{lo} = \sqrt{2}\sqrt{a}$.

3 Unweighted trees

When all the edges have weight one, the upper bound path is always the shortest path, and the geodesic distance is the square root of the RF distance. This is easy to compute.

Tree metrics are useful for understanding distributions of evolutionary trees [3], such as those produced by parsimony or maximum likelihood search methods. Here, we use geodesic distance to get a visual sense of the distribution of a set of equally-optimal tree topologies on different species of the genus Caesalpinia.

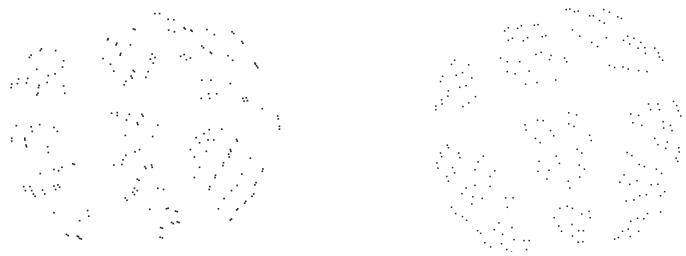


Figure 2: On the left, a layout of a distribution of unweighted phylogenies, reflecting RF distance. On the right, the same trees laid out to reflect geodesic distance gives a similar picture. The layout heuristic (MDS) places points so that their distances reflect distances between corresponding trees.

References

[1] Billera, L.J., Holmes, S.P., and Vogtmann, K. 2001. Geometry of the Space of Phylogenetic Trees. *Advances in Applied Mathematics* 27:733–767.

[2] Hillis, D., Moritz, C. and Mable, B. *Molecular Systematics*, Sinauer Pub., Boston, 1996.

[3] Klingner, J. and Amenta, N. 2002. Case Study: Visualization of Evolutionary Trees. In *Proceedings of IEEE Information Visualization*, pp. 71–74.

H2. Identification of interacting sites in protein families

Vijayalakshmi Chelliah[1], Simon Lovell[2]. Tom L Blundell[3]

Keywords: functional sites, conservation, evolutionary restraints, substitution table

1 Introduction.

Although structural studies of proteins have been traditionally carried out on systems that have been functionally well-characterized, structural genomic projects are reversing this tendency. Structural genomics projects are producing many three-dimensional structures of proteins that have been identified only from their gene sequences. As a consequence, the structures of many proteins that are poorly characterized in terms of function and biochemistry are now being determined. It is therefore important to develop computational methods that will predict sites involved in productive intermolecular interactions that might give clues about functions.
The conservation of amino acid residues has been shown to be strongly dependent on the environment in which they occur in the folded protein and amino acid substitution tables that give the likely substitutions of amino acids in particular local environments have been derived from the structural alignment of the database HOMSTRAD (HOMologous STRructure Alignment Database),[1,2,3]. We have developed a method to distinguish those evolutionary restraints placed on protein structure from additional restraints due to particular functions mediated by interactions with other molecules using these environment-specific substitution tables. Though there are several techniques available to predict the functional site, majority of the techniques do not use all the available structural and sequence information, and are not able to distinguish between evolutionary restraints that arise from the need to maintain structure and those that arise from function. The method presented here takes advantage of all the available sources i.e sequence, structure and evolutionary information.

2 Method.

The method compares two input distributions (observed substitution pattern – calculated from the homologous protein sequences, expected substitution pattern – derived from the environment specific substitution table) and assigns a score for each alignment position to assess their statistical similarity.

Given two distributions P and Q, the commonly used measure of statistical similarity between two arbitrary probability distributions, is the Kullback-Leibler (D^{KL} [4],) divergence defined as:

$$D^{KL}[P||Q] = \Sigma_i \, Px_i \log(Px_i/Qx_i); \;\; D^{KL}[P||Q] \geq 0$$

[1] Department of Biochemistry, 80 Tennis Court Road, University of Cambridge, Cambridge, CB2 1GA, United Kingdom. E-mail: viji@cryst.bioc.cam.ac.uk
[2] School of Biological Sciences, University of Manchester, 2.205 Stopford Building, Oxford Rd, Manchester, M13 9PT. United Kingdom. E-mail: lovell@bioinf.man.ac.uk
[3] Department of Biochemistry, 80 Tennis Court Road, University of Cambridge, Cambridge, CB2 1GA, United Kingdom. E-mail: tom@cryst.bioc.cam.ac.uk

As this measure has the disadvantages of being asymmetric and unbounded, a better measure of statistical similarity as defined by Jensen-Shannon (D^{JS} [5], as suggested by Yona and Levitt [6].) was used to find the divergence score between the two distributions P ("Observed substitution pattern") and Q ("Predicted substitution pattern"). Given two (empirical) probability distributions P and Q, for every $0 \leq \lambda \leq 1$, the λ-JS divergence is defined as:

$$D^{JS}_{\lambda}[P||Q] = \lambda D^{KL}[P||R] + (1 - \lambda)\, D^{KL}[Q||R];\ 0 \leq D^{JS}_{\lambda}[P||Q] \leq 1 \text{ where: } R = \lambda P + (1 - \lambda)Q$$

R is the common source distribution of both distributions P and Q, with λ as a prior weight. This measure is symmetric and ranges between 0 and 1, where the divergence for identical distributions is 0. The divergence score is calculated for each position *t* of the multiple sequence alignment. If the score at an alignment position is high then two distributions at that position is different. The positions where environment-specific substitution tables make poor predictions of the overall amino acid substitution pattern are thus identified.

3 Results.

We find that the clusters of high scoring residues apparently subjected to these additional restraints in evolution correlate well with the functional sites in protein defined by experimental methods. The definition of protein functional sites is surprisingly difficult. There will be strong evolutionary pressure for the residues involved in protein interactions that mediate the function to remain unchanged. This will be true for cofactor binding, enzyme-substrate interactions, protein interactions in multiprotein complexes, allosteric effectors and so on. We define all these sites as functional sites. Functional residue clusters include all residues that contribute to the maintenance of a functional interface. This broader interface definition includes residues that participate directly in protein-ligand contacts and others beneath the site that are crucial for the maintenance of the interface. The method predicts all these residues as functional.

Alignment accuracy is important for the prediction to be accurate. The method is applied to a set of well-characterised protein families and is able to identify functional sites. The technique is fast, automatic and predicts functional sites with a high degree of accuracy.

4 References.

[1] Overington, J., D. Donnelly, et al. (1992). "Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds." Protein Sci **1**(2): 216-26.

[2] Overington, J., M. S. Johnson, et al. (1990). "Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction." Proc R Soc Lond B Biol Sci **241**(1301): 132-45.

[3] Mizuguchi, K., C. M. Deane, et al. (1998). "HOMSTRAD: a database of protein structure alignments for homologous families." Protein Sci **7**(11): 2469-71.

[4] Kullback, S. (1959) Information theory and statistics, John Wiley and Sons, New York.

[5] Lin, J. (1991). "Divergence measures based on the Shannon entropy." IEEE Trans Info theory **37**: 145-151.

[6] Yona, G. a. L., M (2002). "Within the twighlight zone: a sensitive profile-profile comparison tool based on information theory." J Mol Biol **315**(5): 1257-1275.

H3. Streamlining the Conserved Domain Database: A Taxonomic Approach

Praveen F. Cherukuri[1,2], Aron Marchler-Bauer[2], Lewis Y. Geer[2], and Stephen H. Bryant[2]

Keywords: protein domain, taxonomy

1 Introduction

The rapid growth of sequence databases has elevated the need for computational annotation of protein models. Not surprisingly, a variety of protein annotation resources have emerged, some examples include Pfam [1], SMART [3] and COG [5]. The Conserved Domain Database (CDD) [4] imports these and other publicly available alignment model collections, and adds curated domain definitions. The incorporation of several different source databases has created redundancy, ranging from duplication to hierarchical parent-child relationships, which may be caused by differing levels of representation in source databases. In addition, a fairly large subset of domains in CDD describes lineage-specific protein families with narrow taxonomic coverage. We describe a taxonomic-filter approach which is exclusively directed to the removal of such domain models, as we intend to offer a resource aimed at annotating "ancient" conserved domains.

2 Estimating a Domain's Age

We define a set of nodes in the taxonomic tree of life which cover all branches represented by a significant amount of sequence data. We pick these taxonomic nodes so that the presence of a protein or domain family in more than one node indicates a certain minimum age (unless caused by horizontal gene transfer). Focusing on cellular organisms only, the final list has 66 taxonomic nodes (ex: mammalia, alphaproteobacteria, etc). The number of taxonomic nodes covered by a domain family gives a rough indication of that domain's age.

3 Taxonomy Filter

We count the number of preferred taxonomic nodes a conserved domain detects in the NCBI NR (Non-Redundant) protein database, using pre-calculated RPS-BLAST results stored in the CDART database [2]. We recognized 1319 out of 13436 protein domains (~10%) in CDD version 1.63 to be specific to only one of the 66 taxonomic nodes. The majority of these protein domains originate from Pfam (8.5%), followed by COG (1.0%) and other databases (~0.5%). We investigate whether the presence of low-complexity regions or low sequence diversity in the domain model alignments correlates with narrow taxonomic distribution, and whether apparent narrow taxonomic distribution may be caused by bad performance of the search model.

4 Figures and tables.

Table 1: Analysis of protein domains with low taxonomic coverage

[1]Bioinformatics Program, Boston University, 44 Cummington St., Boston, MA 02215, USA. E-mail: cherukur@ncbi.nlm.nih.gov

[2]NCBI, National Institutes of Health, Bldg. 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA. E-mail: bryant@ncbi.nlm.nih.gov

Source Database	Total number of protein domains	Protein Domains with Low Taxonomic Coverage	
		Number	%
Pfam	5426	1146	21.12
COG	4099	129	3.15
SMART	642	28	4.36
Cd	347	13	3.75
LOAD	53	0	0.00
KOG	2869	3	0.10
CDD [V1.63]	13436	1319	9.82

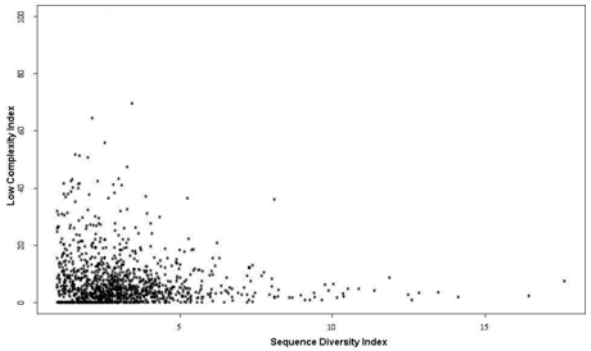


Figure 1: Sequence Diversity Index vs. Low complexity Index of protein domains with only one preferred taxonomic node hit

References

[[1] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C. and Eddy, S. R. 2004. The pfam protein families database. Nucleic Acids Res 32:D138-41.

[2] Geer, L. Y., Domrachev, M., Lipman, D. J. and Bryant, S. H. 2002. Cdart: Protein homology by domain architecture. Genome Res 12:1619-23.

[3] Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R. R., Ponting, C. P. and Bork, P. 2002. Recent improvements to the smart domain-based sequence annotation resource. Nucleic Acids Res 30:242-4.

[4] Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S., Hurwitz, D. I., Jackson, J. D., Jacobs, A. R., Lanczycki, C. J., Liebert, C. A., Liu, C., Madej, T., Marchler, G. H., Mazumder, R., Nikolskaya, A. N., Panchenko, A. R., Rao, B. S., Shoemaker, B. A., Simonyan, V., Song, J. S., Thiessen, P. A., Vasudevan, S., Wang, Y., Yamashita, R. A., Yin, J. J. and Bryant, S. H. 2003. Cdd: A curated entrez database of conserved domain alignments. Nucleic Acids Res 31:383-7.

[5] Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J. and Natale, D. A. 2003. The cog database: An updated version includes eukaryotes. BMC Bioinformatics 4:41.

H4. Heterogeneous Maximum Likelihood Methods for the Detection of Adaptive Evolution.

Jennifer M. Commins[1], Dr Peter Foster[2], Dr James O. McInerney[3]

Keywords: methods, heterogeneous maximum likelihood, adaptive evolution.

1 Introduction.

Proteins evolve by means of random genetic drift (Kimura 1983) and adaptive evolution. Adaptive events result in an organism being more fit than previous members of the same group (Hughes 1999). Our aim is to identify where these events have occurred using a Maximum Likelihood approach. Maximum Likelihood (ML) is a complex statistical method (Edwards 1972). Current implementations of this type of analysis appear to contain certain weaknesses and limitations (Suzuki and Nei 2001). As such, ML and Bayesian methods have come under increasing criticism regarding their reliability; empirical evidence suggests that they may not be using realistic models of sequence evolution (**Yu** *et al*. 2003).

2 Aims and Objectives.

The aim of this work is to design new methods for robustly inferring the evolutionary history of extant sequences and for precisely identifying signatures of adaptive evolution events by using rigorous computational optimisation procedures. Our approach requires the minimum of user input in order to find adaptive evolution events. A phylogenetic tree is constructed from a sequence alignment and assumed to be correct. For each internal node of the tree, we evaluate silent (Ds) and replacement (Dn) substitutions between it and adjoining nodes, both ancestral and descendent. Heterogeneity is achieved by dividing the data into categories according to rules describing the substitution process (Foster, 2003). Unlike other approaches, this method performs analysis across entire lineages of a phylogenetic tree as apposed to a single branch (Fig 1). Additionally, we examine whether or not the replacement substitutions were subsequently changed. We perform a similar analysis for silent substitutions. In this way, a path through the tree is maximised, identifying where the number silent and replacement substitutions differ significantly from each other, indicating evolutionary events.

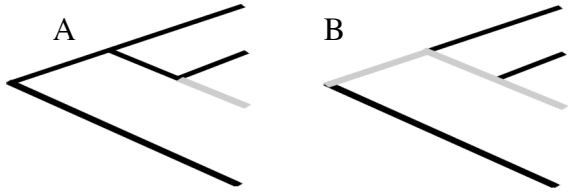


Figure 1: A) Alternative approach B) Lineage analysis implemented in the method described here.

[1] National University of Ireland, Maynooth, Co. Kildare, Ireland. ✉ jennifer.commins@may.ie.

[2] Natural History Museum, London, UK. ✉ p.foster@nhm.ac.uk.

[3] National University of Ireland, Maynooth, Co. Kildare, Ireland. ✉ james.o.mcinerney@may.ie

3 Discussion and Further Work.

This approach directly addresses claims that current ML methods are sensitive to violations of the assumptions regarding which model of evolution to use. The software will be tested on both real and simulated data sets, where the likelihood can be verified using the existing Adaptive Evolution Database. The output of this work will be a software product that is capable of performing analyses on multiple sequence alignments, using methods that are closer to biological reality and more user-friendly than existing methods.

4 References and bibliography.

[3] Edwards, A. W. F. 1972. Likelihood. Cambridge University Press, Cambridge.

[6] Foster, P. 2003. Personal Correspondence.

[2] Hughes, A. L. 1999. Adaptive Evolution of Genes and Genomes. Oxford University Press, New York, Oxford.

[1] Kimura, M. 1983. The neutral theory of molecular evolution. Evolution of Genes and Proteins. Cambridge University, New York, London.

[4] Suzuki, Y., and M. Nei. 2001. Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. 18:2179-2185.

[5] Yu, Y.-K., J. C. Wootton, and S. F. Altschul. 2003. The compositional adjustment of amino acid substitution matrices. PNAS 100:15688-15693.

H5. Homogeneous Phylogenetic Models: Invariants and Parametric Inference

Nicholas Eriksson [1]

Keywords: Phylogenetic trees, parametric inference, phylogenetic invariants.

We study the model on phylogenetic trees in which each node is a binary, observable random variable Y_i and the transition probabilities are given by the same matrix $A = \left(\begin{smallmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{smallmatrix}\right)$ on each edge. We write the joint probabilities as $p_{\sigma_1\sigma_2\ldots\sigma_n} := P(\mathbf{Y} = \sigma)$. Let $\rho(i)$ denote the parent of node i in the tree.

The homogeneous Markov model is a fundamental model for statistics. We believe that an understanding of the homogeneous model will lead to more general results. For example, by adding nodes to subdivide edges, we can think of the homogeneous model as a discrete approximation to a continuous Markov model. We are interested in two questions from [4] that are of fundamental importance to the use of statistical models in biology.

1. Given observations $\sigma = (\sigma_1, \ldots, \sigma_n)$, describe the set of parameters a_{ij} such that p_σ is maximal among the coordinates of p.
2. Which (parameter independent) relations on the probabilities $p(\mathbf{Y} = \sigma)$ does the model imply?

Problem 1 has been studied in [3, 6] and shown to be of importance for sequence alignment problems. To solve Problem 1, first we write the joint probabilities in terms of the parameters $a_{00}, a_{01}, a_{10}, a_{11}$

$$p_{\sigma_1\sigma_2\ldots\sigma_n} = a_{\sigma_{\rho(2)}\sigma_2} a_{\sigma_{\rho(3)}\sigma_3} \cdots a_{\sigma_{\rho(n)}\sigma_n}.$$

That is, the probability of observing σ is the product of the a_{ij} that correspond to the transitions on the edges of the tree. Next, transform to logarithmic coordinates $b_{ij} = -\log(a_{ij})$. The condition that $p_{\sigma_1\ldots\sigma_n}$ is maximal among the coordinates of p becomes the linear system of inequalities

$$b_{\sigma_{\rho(2)}\sigma_2} + \cdots + b_{\sigma_{\rho(n)}\sigma_n} \geq b_{l_{\rho(2)}l_2} + \cdots + b_{l_{\rho(n)}l_n} \qquad \text{for all } (l_1, \ldots l_n) \in \{0,1\}^n.$$

The set of solutions to these inequalities forms a polyhedral cone. If the cone is full dimensional, then $\sigma_1, \ldots, \sigma_n$ is the most likely observation for some choice of the parameters. Such a sequence is called a *Viterbi sequence*. The collection of the cones of all Viterbi sequences is the normal fan of the *Viterbi polytope*, P_T. This polytope is three-dimensional and has one vertex for every Viterbi sequence. Given this polytope, we can quickly solve Problem 1 for any $\sigma_1, \ldots, \sigma_n$. Our main result is an explicit description of the polytope for a class of binary trees. For an example, see Figure 1.

Theorem 1 *If T is a binary tree with $n > 3$ nodes in which all leaves are an odd distance from the root, then there are exactly 8 Viterbi sequences. Furthermore, the polytope P_T has the same combinatorial structure for all such trees.*

Problem 2 is solved by computing the ideal of *polynomial invariants* among the probabilities $p_{\sigma_1\ldots\sigma_n}$. The invariants vanish for a given distribution $(p_{i_1\ldots i_n})$ essentially when that distribution comes from our model. Therefore, invariants have been used in phylogenetics to identify good trees for aligned sequences, see [1, 2]. In the observed, homogeneous Markov

[1]Department of Mathematics, University of California, Berkeley, CA 94720-3840 E-mail: eriksson@math.berkeley.edu

model, the invariants can be computed using the theory of Gröbner bases of toric ideals (see [5]). We are able to calculate the ideal of invariants for trees with 11 nodes. These are computations in 2048 indeterminants, which we believe to be the largest number of indeterminants ever in a Gröbner basis calculation. We conjecture that binary trees require only linear and quadratic generators for the ideal of invariants.

Example 1 Let T be the path with 4 nodes. Then the ideal of invariants has 32 minimal generators. Four generators are linear (e.g., $p_{0100} - p_{0010}$), twenty-four are quadratic (e.g., $p_{0001} \cdot p_{0010} - p_{0000} \cdot p_{0101}$), and four are cubic (e.g., $p_{1000} \cdot p_{1100} \cdot p_{1111} - p_{0000} \cdot p_{1110}^2$). We believe that the ideal of invariants of a path is always generated by these three types of relations.

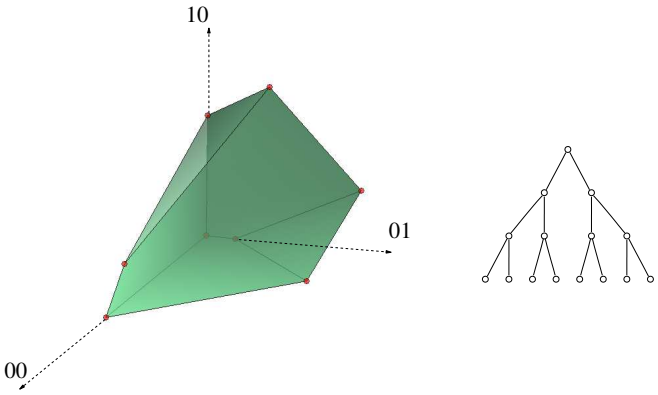


Figure 1: Let T be the pictured tree with 15 nodes. By Theorem 1, since all leaves are at an odd distance from the root, exactly 8 of the $2^{15} = 32768$ possible observations are Viterbi sequences and therefore the polytope P_T has 8 vertices. The polytope is displayed with the x-coordinate counting $0 \to 0$ transitions, the y-coordinate counting $0 \to 1$ transitions and the z-coordinate counting $1 \to 0$ transitions. For example, the front left vertex $(14, 0, 0)$ corresponds to the all zero observation, which is Viterbi sequence with the parameters $a_{00} = 1$, $a_{01} = 0$, $a_{10} = 1$, $a_{11} = 0$.

References

[1] Allman, E. and Rhodes, J. 2003. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, 1–33.

[2] Cavender, J. and Felsenstein, J. 1987. Invariants of phylogenies in a simple case with discrete states. *Journal of Classification*, 4:57–71.

[3] Gusfield, D., Balasubramanian, K., and Naor, D. 1994. Parametric optimization of sequence alignment, *Algorithmica* 12, 312–326.

[4] Pachter, L. and Sturmfels, B. 2003. The geometry of statistical models for biological sequences, eprint: `arXiv q-bio.QM/0311009`

[5] Sturmfels, B. 1996. *Gröbner bases and convex polytopes*, volume 8 of *University Lecture Series*. American Mathematical Society, Providence, RI.

[6] Waterman, M., Eggert, M. and Lander, E. 1992. Parametric sequence comparisons, *Proc. Natl. Acad. Sci. USA* 89:6090–6093.

H6. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency

P. Clote[1], F. Ferré[2], E. Kranakis[3] [4], D. Krizanc[5]

Keywords: tRNA, folding energy, secondary structure, random RNA, structural RNA.

1 Introduction

It is well-known that there is a compositional bias in nucleotide usage of various classes of RNA, depending on function. Seffens and Digby [4] and Workman and Krogh [5] investigated whether the folding energy of mRNA is lower than that of random RNA of the same mononucleotide [4] resp. dinucleotide [5] frequency, reaching opposite conclusion. In [4], Seffens and Digby reported that the minimal free energy (mfe) of the optimal secondary structure, or *native state*, of mRNA is significantly lower than that of random RNA having the same mononucleotide frequency (i.e. random RNA generated by a 0-th order Markov chain), while Workman and Krogh [5], arguing that random RNA must be generated with the same dinucleotide frequency, found that mRNA does not have any statistically significant lower mfe than random RNA. This is coherent with the idea of mRNA existing in an ensemble of low energy states, lacking any functional native state for which the structure has a functional relevance. To help solving this controversy, we present evidence that for several RNAs for which the native state (minimum free energy structure) is functionally important (as tRNAs, type III hammerhead ribozymes, SECIS elements, signal recognition particle RNAs, small nucleolar spliceosomal RNAs) the folding energy is significantly lower than that of random RNA of the same dinucleotide frequency, while there is no such distinction between mRNAs and random RNAs of the same dinucleotide frequency.

2 Methods and Results

Using the mono- and dinucleotide frequencies computed for each RNA class, we generated for each RNA 1000 random RNAs of the same *expected* dinucleotide frequency using a first-order Markov chain algorithm and 1000 random RNAs of the same dinucleotide frequency using an implementation of the Altschul-Erikson algorithm [1]. Then we computed the Z-score of its minimum free energy (mfe) using version 1.5 of Vienna RNA Package `RNAfold` [2] with respect to the mfe of the corresponding 1000 random RNAs, obtained with the two different methods. We followed the same procedure for each class of RNA we investigated, with the exception of mRNA, when to avoid unduly long computational times, we generated only 100 random RNA per mRNA. We were able to demonstrate that the considered classes of RNA sequences with known important secondary structure have, on average, a lower folding free energy than dinucleotide-frequency-matching control sequences; moreover, we validate the conclusion of Workman and Krogh [5] concerning mRNA. These findings suggest that selective pressure for structural RNA sequences has driven them to find low free energy conformations. Table 1 shows that using the Altschul-Erikson algorithm all classes of structurally important RNA investigated show a significantly lower folding energy than random RNAs of the same dinucleotide frequency (a similar trend is found using the first-order Markov chain algorithm). On the contrary, this is not true for mRNA; moreover,

[1]Dept. of Biology, Boston College, `clote@bc.edu`
[2]Dept. of Biology, Boston College, `ferref@bc.edu`
[4]School of Computer Science, Carleton University, `kranakis@scs.carleton.ca`
[5]Dept. of Computer Science, Wesleyan University, `dkrizanc@mail.wesleyan.edu`

RNA type	Number of sequences	Mean	Stdev	Max	Min
tRNA	530	-1.592244	0.885706	0.689074	-3.995676
Hammerhead III	114	-3.209455	0.917812	-1.188390	-5.701443
SECIS	5	-4.656427	1.171537	-3.303051	-6.962124
srpRNA	94	-3.350788	1.887907	0.301066	-8.929866
U1	53	-1.914551	1.216397	0.378252	-7.878986
U2	62	-4.606619	1.739807	-1.787787	-10.733675
mRNA whole lenght	41	-0.178294	1.439988	2.441546	-3.869000
mRNA 3'UTR	41	-0.250845	1.412864	1.986367	-5.796952
mRNA 5'UTR	41	0.284761	1.330637	4.330794	-3.970213
mRNA cds	41	-0.214827	1.446859	3.058202	-4.259037

Table 1: Z-score statistics for structural RNA compared to random RNA of the *same* dinucleotide frequency using Altschul-Erikson Algorithm.

also the 3′ and 5′ untranslated regions, that usually contain structured elements recognized by specific translation factors have the same folding energy as that of random RNA.

Figure 1 present superposed histograms of Z-scores for the RNAs analyzed. The general trend is the shifting towards negative values of the curves related to structural RNAs; Z-score curves obtained using the two algorithms are similar. Similar curves to those of Figure 1 were obtained by Rivas and Eddy [3] (eg. Figure 14 of [3]) when comparing certain RNA genes with random RNA of the same mononucleotide (not dinucleotide) frequency.

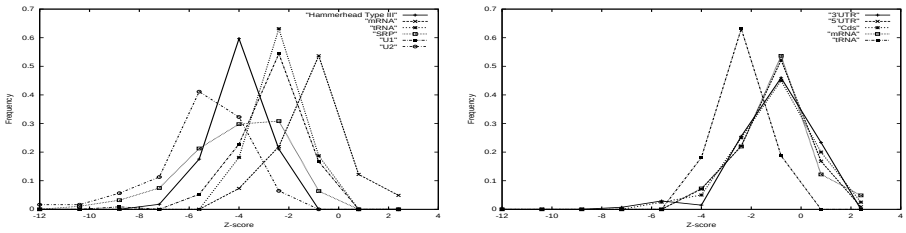


Figure 1: Histograms of Z-scores of minimum free energy (mfe) of RNA classes versus 1000 random RNAs of the same dinucleotide frequency using Altschul-Erikson Algorithm. Left panel: structural important RNAs curves are shifted toward negative values with respect to the whole length mRNAs curve. Right panel: whole lenght mRNA, coding sequences (cds), 3' untranslated region (UTR) and 5' UTR compared to tRNA.

References

[1] S.F. Altschul and B.W. Erikson. Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol*, 2(6):526–538, 1985.

[2] I. Hofacker et al. Vienna RNA Package. `http://www.tbi.univie.ac.at/~ivo/RNA/`

[3] E. Rivas and S. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNA. Bioinformatics, 16: 573-585, 2000

[4] W. Seffens and D. Digby. mRNAs have greater negative folding free energies than shuffled or codon choice ransomized sequences. *Nucl. Acids. Res.*, 27:1578, 1999.

[5] C. Workman and A. Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucl. Acids. Res.*, 27:4816–4822, 1999.

H7. Convergent evolution of domain architectures

Julian Gough, [1]

Keywords: evolution, domain, structure, superfamily, architecture

1 Introduction.

A domain is the smallest unit of evolution in the definition from the SCOP database of known protein structures. Small proteins consist of a single domain, and some larger proteins consist of more than one domain. A part of a protein is only considered a domain in its own right if it is observed elsewhere in nature on its own or in combination with different partner domains. In SCOP, domains which share a common evolutionary ancestor belong to the same superfamily. The domain architecture of a protein is described by the order of the domains and the superfamilies to which they belong. The repertoire of architectures present in the genomes has arisen by the duplication and recombination of the ancestral superfamily domains.

The question which is addressed here, is to what extent the architectures observed in the genomes are due to functional necessity or due to evolutionary descent. Convergent evolution is defined here as more than one independent evolutionary event (recombination) leading to the same domain architecture. If the shuffling of domains is functionally driven then we expect to find a great deal of evidence of convergent evolution, whereas a failure to detect convergent evolution points to evolutionary descent.

2 Analysis

From the SUPERFAMILY database 70 genomes were chosen to represent a wide and even spread across all kingdoms of life. The subset of 4075 multi-domain architectures in these 70 genomes whose sequences are completely covered by domain assignments was used in this work.

The proteins with each domain architecture were first clustered into phylogenetic groups. If an architecture is observed in every genome belonging to a particular branch of the phylogenetic tree, then the most likely explanation is that the architecture was inherited from the root node of that branch.

Any architecture which forms more than one distinct phylogenetic group is a candidate for convergent evolution. The two explanations alternative to convergent evolution are: horizontal gene transfer and gene loss. Both of these cause the true evolutionary group descending from a common ancestral architecture to be split into more than one observed phylogenetic group. To ascertain whether, for any given architecture, phylogenetic groups should be joined to form a single evolutionary group, two different principles were used:

(1) If two proteins come from the same complete-architecture evolutionary ancestor, then they will have a closer homology than the components of non-related proteins sharing the same architecture. Furthermore they will have homology across the entire length. BLAST

[1] RIKEN Genome Sciences Centre, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan. E-mail: gough@supfam.org

groups	phylogeny	(1) multi-d	(1) mono-d	method (2)	(1) or (2)
1	2891	3517	428	3665	3911
2	460	399	210	270	149
3	191	98	109	66	12
4	69	22	69	21	0
5	64	12	33	15	0
>5	400	27	86	38	3
>1	29%	14%	54%	10%	4%

Table 1: The number of architectures for different methods.

was used with local scoring, to link groups sharing a protein with an E-value < 0.00001 and a local alignment across n/(n+1) of the residues of the longer sequence (up to 0.9), 'n' being the number of domains.

(2) The individual domains belonging to two proteins sharing the same complete-architecture ancestor will have diverged for the same length of time, and in the same environment as each other. In two proteins which have convergently evolved the same architecture, the component domains will have a different evolutionary history from each other. Sequence identity from SUPERFAMILY alignments of the individual domains (from a pair of sequences) was used to link groups sharing a protein with no domains more than 10% different in identity from all of the other domains.

Table 1 shows four different results. The first column corresponds to the phylogenetic groups, before any method is applied. The second and third columns relate to method (1). As well as being applied to the set of multi-domain architectures (first column) it was also applied to the set of mono-domains (second column). Since the mono-domain proteins are by definition descended from a common evolutionary ancestor, they should all be listed as having a single evolutionary group. This (control) clearly highlights the weakness of method (1); some evolutionarily related proteins have diverged beyond the point of recognition by BLAST. This is expected to be much worse for the mono-domains since they are more ancient.

The weakness of method (2) is that different superfamilies have a different propensity to diverge and change their sequence identities at different rates within the same protein. It is expected that both methods overestimate the extent of convergent evolution, and the architectures forming more than one evolutionary group should only be considered as candidates for further investigation.

3 Conclusion

The vast majority of proteins are descended from a single ancestor with a unique domain architecture, and convergent evolution of domain architectures is relatively rare. Although further work is required to produce an accurate estimate of the extent, we can conclude from this work that at least 95% of all architectures (including mono-domain) observed in the genomes are due to evolutionary descent rather than functional necessity (see introduction).

H8. Extracting the phylogenetic signal from Mutual Information estimates

Rodrigo Gouveia-Oliveira[1], Anders Gorm Pedersen[1]

Keywords: mutual information, phylogeny, parametric bootstrap

Predicting interaction between protein sites (in the same or different proteins) is an interesting problem, which has been tackled in different ways. One common approach to the problem is by measuring the Mutual Information between sites in an alignment, as:

$$I(A;B) = \sum_i \sum_j P(a_i, b_j) \log \left[\frac{P(a_i, b_j)}{P(a_i) P(b_j)} \right]$$

where *a* and *b* are columns of the alignment, and *i* and *j* run through all categories (aminoacids, or nucleotides).
High Mutual Information between two sites means these sites tend to co-vary, and thus the knowledge of one improves our guesses on the other. If biological sequences were independent this would pinpoint interacting sites. However, biological sequences share a common ancestry, and that also gives rise to joint variation, as illustrated in the figure:

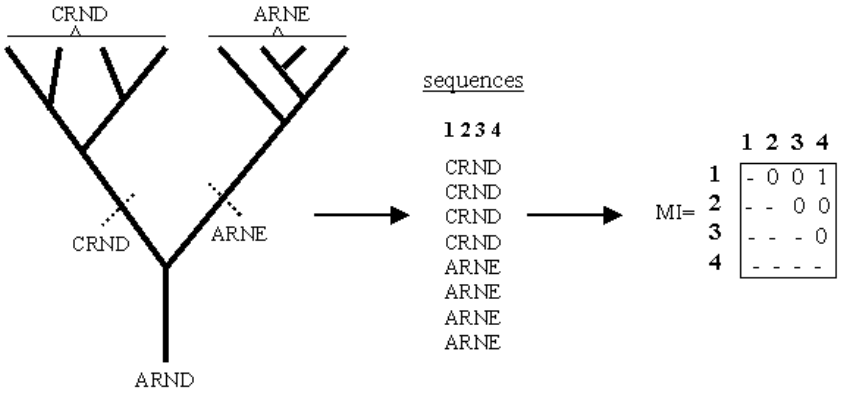


Figure 1: A hypothetic phylogeny, the extant sequences and the Mutual Information matrix between the 4 sites of the sequence. Two independent events result in a set of sequences with very high mutual information between sites 1 and 4.

[1] Center for Biological Sequence Analysis, Danish Technical University, Building 208, DK-2800 Lyngby, Denmark
E-mail:gorm@cbs.dtu.dk (corresponding author)
E-mail: rodrigo@cbs.dtu.dk

Several authors have noticed this problem and tried to extract the phylogenetic signal from the Mutual Information estimates (Felsenstein 1985; Gulko and Haussler 1996;Akmaev et al. 1999,2000; Wollenberg and Atchley 2000). Specifically, (Wollenberg and Atchley 2000) used parametric bootstrapping to simulate the density function of Mutual Information in a sequence due to phylogeny and chance. They then used it as a null-distribution for testing the Mutual Information values observed in their sequences.
In this work we extend their analysis by investigating the use of null models that explicitly include parameterizations of site-specific features. This allows us to obtain null-distributions that are specific to individual pairs of sites (or class of pairs).

References

Akmaev V.R., Kelly S.T., Stormo G.D. ,2000. Phylogenetically enhanced statistical tools for RNA structure prediction. In:*Bioinformatics* 6:501-12.

Akmaev V.R., Kelly S.T., Stormo G.D. ,1999. A phylogenetic approach to RNA structure prediction. In:*Proc Int Conf Intell Syst Mol Biol.* 1999;:10-7.

Felsenstein J, 1985 . Phylogenies and the comparative method, In: *The American Naturalist* 125:1-15

Gulko,B , Haussler D. 1996. Using multiple alignments and phylogenetic trees to detect RNA secondary structure. In: Pac Symp Biocomput. 1996;:350-67.

Wollenberg K.R. and Atchley W.R.,2000. Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. In: *Proceedings of the Natural Academy of Sciences USA*. 97(7):3288-91

H9. Optimal, Efficient Reconstruction of Root-Unknown Phylogenetic Networks with Constrained Recombination

Dan Gusfield [1,2]

Keywords: Molecular Evolution, Phylogenetic Networks, Ancestral Recombination Graph, Recombination, SNP

1 Introduction

With the growth of genomic data, much of which does not fit ideal evolutionary-tree models, and the increasing appreciation of the genomic role of such phenomena as recombination, recurrent and back mutation, horizontal gene transfer, gene conversion, and mobile genetic elements, there is greater need to understand the algorithmics and combinatorics of phylogenetic networks on which extant sequences were derived. Phylogenetic networks are models of sequence evolution that go beyond trees, allowing biological operations that are not tree-like. One of the most important biological operations is recombination between two sequences.

Hein et al. introduced and studied the *phylogenetic network problem (with recombination)*: Construct a phylogenetic network that derives a given set of binary sequences M, minimizing the number of recombinations used [3, 4, 8, 7, 6, 5]. The minimization criteria is motivated by the general utility of parsimony in biological problems, and because most evolutionary histories are thought to contain a small number of observable recombinations. The assumption that the sequences are binary is motivated today by the importance of SNP data, where each site can take on at most two states (alleles).

No efficient, general algorithm is known for the phylogenetic network problem. Wang et al. [9] introduced a restricted version of the problem: Construct a phylogenetic network when the network is constrained to be a "galled-tree", and the ancestral sequence for the galled-tree is specified in advance. A galled-tree is a phylogenetic network where all cycles (created by recombinations), must be disjoint from each other. Simulations have shown that galled-trees are common when the recombination rate is moderate. The problem of determining whether a set of sequences can be derived on a galled-tree, with a specified ancestral sequence, has an efficient solution [1, 2]. Moreover, when there is a galled-tree for the input, with the specified ancestral sequence, the algorithm produces one that minimizes the number of recombinations over all possible phylogenetic networks with that ancestral sequence. However, the more biologically realistic case is that *no* ancestral sequence is known in advance, and the only previous algorithmic solution for that case takes exponential time.

[1] Department of Computer Science, University of California, Davis, E-mail: gusfield@cs.ucdavis.edu Dan Gusfield

[2] Research partially supported by NSF Grant EIA-0220154. Thanks to C. Langley and S. Eddhu for helpful conversations on this topic.

2 Solution to the Root-Unknown Galled-Tree Problem

We have now developed an efficient solution to the root-unknown galled-tree problem, i.e., in the case when no ancestral sequence is known in advance. For input consisting of n sequences, each of length m, the algorithm runs in $O(nm + n^3)$ time. We show that when there is a galled-tree for the input, the algorithm finds one that minimizes the number of recombinations over all possible phylogenetic networks and over all possible ancestral sequences. This result holds even if multiple-crossover recombinations are allowed.

The main tools that we use to solve the root-unknown galled-tree problem are two graphs representing "incompatibilities" and "conflicts" in M. The conflict graph was used to solve the galled-tree problem when an ancestral sequence is specified in advance [1, 2]. The incompatibility graph is the analogous graph when no ancestral sequence is known. The conflict graphs can be different for different ancestral sequences, so the main difficulty in extending the previous solution to the root-unknown case is that we do not know which conflict graph to use, and there may be an exponential number of them. The main new structural result is that there is always an ancestral sequence A such that its conflict graph is the same as the incompatibility graph. The algorithmic consequence is that even without knowing A, we can determine its conflict graph, and build the network based on that graph, along with other needed modifications to other parts of the previous solution that depend on knowing the ancestral sequence.

References

[1] D. Gusfield, S. Eddhu, and C. Langley. Efficient reconstruction of phylogenetic networks (of SNPs) with constrained recombination. In *Proceedings of 2'nd CSB Bioinformatics Conference*. IEEE Press, 2003.

[2] D. Gusfield, S. Eddhu, and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination, To appear in J. Bioinformatics and Computational Biology. Technical report, UC Davis, Department of Computer Science, 2003.

[3] J. Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci*, 98:185–200, 1990.

[4] J. Hein. A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, 36:396–405, 1993.

[5] R. Hudson and N. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164, 1985.

[6] S. R. Myers and R. C. Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163:375–394, 2003.

[7] Y. Song and J. Hein. On the minimum number of recombination events in the evolutionary history of DNA sequences. *Journal of Mathematical Biology (to appear)*, 2003.

[8] Y. Song and J. Hein. Parsimonious reconstruction of sequence evolution and haplotype blocks: Finding the minmimum number of recombination events. In *Proc. of 2003 Workshop on Algorithms in Bioinformatics*. Springer-Verlag LNCS, 2003.

[9] L. Wang, K. Zhang, and L. Zhang. Perfect phylogenetic networks with recombination. *Journal of Computational Biology*, 8:69–78, 2001.

H10. Alu Clustering in the Human Genome: Origins and Consequences

Michael Hackenberg[1] and Jose L. Oliver[1]

Keywords: Alu, Isochores, Recombination, Evolution, Negative Selection, Alu Clustering

1 Introduction

Approximately 10% of the human genome is made up of Alu elements. Although the Alu repeats have been the focus of investigation over the last few years, many aspects of their genomic role still remain unsolved. For example, they appeared in the genome some 50-80 MYA and had their major spread approx. 30 MYA, which is about two orders of magnitude higher than the present amplification rate [1]. Other topics which are still controversial deal with the Alu distribution over the human genome. In general, Alu density is positively correlated with the GC content of the genomic region (isochore) in which they reside [2]. However, recently inserted Alus present a different pattern. As they depend on the LINE-1 transposition machinery, they show an initial density maximum in GC-poor regions. In [3], we analysed the Alu distribution as a function of evolutionary age and isochore membership, as well as the densities of possible recombination outcomes, inferring that recombination is probably the most important mechanism driving the density shift from GC-poor (L) to GC-rich (H) isochores. Here, we present an analysis of the Alus as a function of isochore membership and physical distance to the next Alu repeat (**D**istance to the **N**earest **N**eighbour - DNN), which reinforces our argument and sheds light on this controversy from a different side.

We analysed the human reference sequence (April 2003 freeze, UCSC version hg15), based on NCBI Build 34 and produced by the International Human Genome Sequencing Consortium (IHGSC), downloaded from (ftp://genome.ucsc.edu/goldenpath/10april2003/bigZips/chromFa.zip).

The partition of the chromosomes into isochores was performed using the IsoFinder segmentation algorithm (see [4] and references therein). The Alu densities indicated have been calculated as the number of elements per 10 kb.

2 Results and Discussion

As a measure for Alu clustering, we used the distance to the next Alu (DNN). The Alu density ratios (densities in H isochores divided by densities in L isochores) in the different isochores as a function of DNN are shown in Figure 1. For short distances, the densities are up to 8 times higher in H than in L isochores, which corresponds to an extremely higher clustering in the H isochores. For example, in the GC-richest isochore H4, almost 40% of all Alus are closer than 20 bp to each other, whereas in the GC-poorest isochore L1 this fraction declines to 5%. With growing distances to the closest Alu, however, the density maximum shifts gradually towards the L isochores; thus, Alus with distances greater than 2000 bp to the next Alu show a density maximum in L isochores. This means that interactions with other Alus are crucial for the density shift, as single Alus with long distances to the next Alu remain in L isochores. There are two possible explanations: Alu/Alu recombination and preferential insertion in or near preexisting Alus, both leading to pronounced Alu clustering. It is known that the poly-A tail of an Alu may constitute a good insertion target for

[1] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, E-18071 Granada, Spain. E-mail: genmol@ugr.es, oliver@ugr.es

new insertions [5]. Analysing the densities of Alu/Alu insertions (i.e., insertions in the AT-rich linker, unpublished), we found that the insertion probability is positively correlated with the isochore GC content. We believe, nevertheless, that the preferential targeting on preexisting Alus in H isochores can only partially contribute to form the pronounced overall Alu density maximum in H2/H3 isochores, but that the density shift and the formation of the biased distribution over isochores is driven mainly by recombination [3].

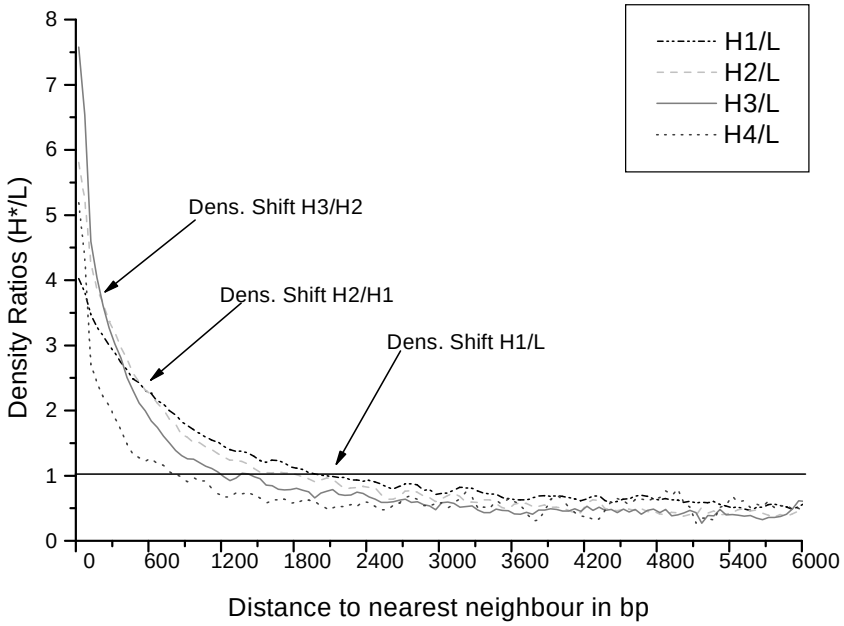


Figure 1: Alu density ratios as a function of DNN (Distance to Nearest Neighbour) and isochore membership are shown. It can be seen that very closely spaced Alus have a pronounced density maximum in the H3 isochore. With growing distance, the density maximum shifts gradually towards L isochores.

3 References and bibliography

[1] Batzer M.A., Deininger P.L. 2002. Alu repeats and human genomic diversity. Nature Reviews Genetics 3:1-10.

[2] Bernardi G. 2001. Misunderstandings about isochores. Part 1. Gene 276:3-13.

[3] Hackenberg M, and Oliver J.L. 2004. .The biased distribution of Alus in the human isochores might be driven by recombination. Submitted.

[4] Oliver JL, Carpena P, Hackenberg M, Bernaola-Galván P. IsoFinder: computational prediction of isochores in genome sequences, in preparation

[5] Stenger, J.E. et al. 2001. Biased Distribution of Inverted and Direct *Alus* in the Human Genome: Implications for Insertion, Exclusion, and Genome Stability. Genome Research 11:12-27.

H11. Pruned PDM Method for Detecting Recombination

Dirk Husmeier [1]

Keywords: phylogenetics, interspecific recombination, sliding window methods, Markov chain Monte Carlo, probabilistic divergence measure.

1 Introduction

The underlying assumption of most phylogenetic tree reconstruction methods is that there is one set of hierarchical relationships among the taxa. While this is a reasonable approach when applied to most DNA sequence alignments, it can be violated in certain bacteria and viruses due to interspecific recombination. The resulting transfer or exchange of DNA subsequences can lead to a change of the branching order (topology) in the affected region, which results in conflicting phylogenetic information from different regions of the alignment. If undetected, the presence of these so-called mosaic sequences can lead to systematic errors in phylogenetic tree estimation.

The idea of a recently proposed method for detecting evidence of recombination in DNA sequence alignments is illustrated in the left panel of Figure 1. Consider a given alignment of DNA sequences, $\mathcal{D}$, from which we select a consecutive subset $\mathcal{D}_t$ of predefined width W, centred on the tth site of the alignment. Let S be an integer label for tree topologies, and consider the marginal posterior probability of tree topologies S conditional on the 'window' $\mathcal{D}_t$, $P(S|\mathcal{D}_t)$, which is numerically computed with Markov chain Monte Carlo by marginalizing over the branch lengths of the phylogenetic tree and the parameters of the nucleotide substitution model. The basic idea of the probabilistic divergence method (PDM) for detecting recombinant regions is to move the window $\mathcal{D}_t$ along the alignment and to monitor the distribution $P(S|\mathcal{D}_t)$. We would then, obviously, expect a substantial change in the shape of this distribution as we move the window into a recombinant region. To quantify the degree of change, a probabilistic divergence measure is computed, as discussed in (2).

A shortcoming of the PDM method is that its performance deteriorates as the number of taxa increases. This is because for an increased number of taxa the posterior distribution over tree topologies, $P(S|\mathcal{D}_t)$, becomes more diffuse unless the size of the data set $\mathcal{D}_t$ is increased. An increased amount of data $\mathcal{D}_t$, however, corresponds to an increased length of the sliding window, which compromises the spatial resolution of the detection and is not an option for short alignments.

2 Method

A possible remedy to this problem is to reduce the vagueness of the posterior distribution by reducing the cardinality of the support of $\langle P(S|\mathcal{D})\rangle$, where $\langle . \rangle$ denotes an average over all window positions. This can be effected with a pruning scheme based on the Robinson-Foulds (RF) distance (3). First, identify a set of principal tree topologies, for instance, those that maximize $\langle P(S|\mathcal{D})\rangle$. Next, assign each non-principal tree topology to the principal topology with the minimum RF-distance. Finally, renormalize the posterior distributions $P(S|\mathcal{D}_t)$ and recompute the PDM signal. An illustration is given in Figure 1. The reduction in the vagueness of $P(S|\mathcal{D}_t)$ is likely to reduce the noise in the PDM signal. Note that similar pruning methods are used in machine learning to improve the generalization performance of a predictor. Also note that the pruning of the support of $\langle P(S|\mathcal{D})\rangle$ can be justified in a Bayesian way as bringing the data-based prediction in line with our prior assumptions about the expected frequency of recombination events.

[1] Biomathematics & Statistics Scotland, JCMB, The King's Buildings. Edinburgh EH9 3JZ, UK. E-mail: dirk@bioss.ac.uk

3 Results

The method was applied to a DNA sequence alignment of ten strains of Hepatitis-B virus with the following Genbank accession numbers: D00329, X68292, V00866, M57663, D00330, M54923, X01587, D00630, M32138 and L27106. Without pruning, the probabilistic divergence signal, shown in the left panel of Figure 2, contains erratic oscillations that obscure any breakpoint patterns. This is dramatically improved with the pruning method (middle and left panels of Figure 2). Note that three clear breakpoints occur, which were also found in an independent earlier study (1). Also note that for sufficiently small values of the cutoff threshold K, the results are rather independent of K.

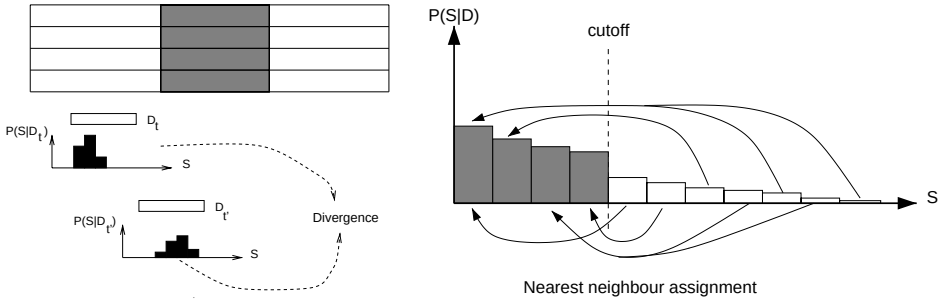


Figure 1: *Left:* PDM method. The figure shows the posterior distribution $P(S|D_t)$ of tree topologies S conditional on two subsets D_t and $D_{t'}$ selected by a moving window. When the window is moved into a recombinant region, the posterior distribution $P(S|D_t)$ can be expected to change significantly, which leads to a large probabilistic divergence score. *Right:* On the average posterior distribution of tree topologies, averaged over all sliding window positions, a cutoff threshold is defined. Tree topologies above this threshold are kept as "principal topologies". Tree topologies below the threshold are assigned to the principal topology with the minimal RF distance. After this re-assignment, the posterior distributions $P(S|D_t)$ are re-normalized.

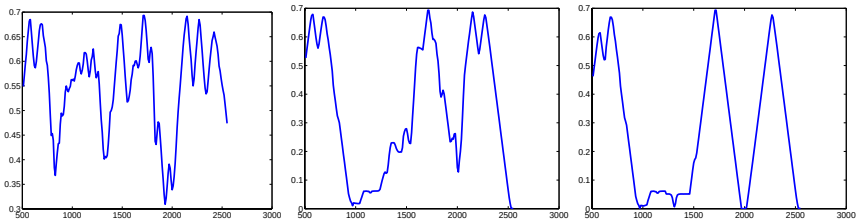


Figure 2: Detection of recombination in a DNA sequence alignment of ten strains of Hepatitis B virus. The graphs show the probabilistic divergence signals obtained with a window size of 500. From left to right: No pruning, pruning down to 7 and 3 tree topologies.

References

[1] Bollyky, P. L., Rambaut, A., Harvey, P. H., and Holmes, E. C. (1996). *Journal of Molecular Evolution*, 42:97–102.
[2] Husmeier, D. and Wright, F. (2001). *Bioinformatics*, 17(Suppl.1):S123–S131.
[3] Robinson, D. and Foulds, L. (1981). *Mathematical Biosciences*, 53:131–147.

H12. Phylogenetic analyses detect site-specific perturbations in asymmetric mutation gradients

Neeraja M. Krishnan[1], Hervé Seligmann[1], Sameer Z. Raina[1], and David D. Pollock[1]

Keywords: deamination gradient, Hidden Markov Model, single-strandedness, Monte Carlo Markov Chain, mRNA secondary structure, protection against mutations

1. Introduction

During replication in mitochondria portions of the heavy strand remain single-stranded and exposed to mutations for varying periods of time. Thus deaminations from cytosine to thymine (C→T) and adenine to hypoxanthine (A→H, which leads to A→ G substitutions) accumulate [1, 2] on that strand. Previous likelihood-based analyses in vertebrates indicated that while C→T substitutions accumulate rapidly and then saturate with increased time spent single-stranded, A→G substitutions accumulate proportionally to single-strandedness [3]. Here, we present two new phylogeny based Bayesian methods for modeling this substitution response profile at the single site level, and apply them to a set of complete primate mitochondrial genomes.

2. Methods

Our first approach assumes a "base" model with a symmetric substitution probabilities matrix; asymmetric increase in a specific substitution type (according to model of [4])is added as a linear function of single-strandedness, measured by D_{ssH}, with the slope and intercept of the response model as unknown parameters. The second approach implemented a Hidden Markov Model (HMM) with a correlation in rates between sites as a function of distance between them, to determine and compare details of the response curve for different substitutions, and to detect consistent regional deviations from the expectations of our model. To build complex models at each site with relatively small computational burden, complex models were evaluated based on the distribution of likely ancestral sequences obtained using a general-time reversible (GTR) model. Equilibrium frequency ratios were evaluated based on the corresponding substitution probabilities and reversibility constraints. For e.g. $\pi_G/\pi_A = \lambda_{GA}/\lambda_{AG}$. In particular, C→T deaminations appear to increase linearly for only a very short genome section, reaching a plateau that is best explained by some form of protection or repair mechanism.

3. Results and Discussion

In all primate groups[2], A→G substitution probabilities increased linearly with D_{ssH}, although there were large differences between the two primate sub-groups ('high' and 'low'), as predicted by previous analysis (Raina et al, unpublished data). We confirm previous speculation that C→T substitutions increase rapidly at low single-strandedness values, and then remain approximately saturated for the rest of the genome [2]. Variations in the frequency ratios (G/A and C/T) in the form of dips and rises in specific genes exist (Figures 1 and 2). These correlate significantly with site-specific mRNA secondary structure, suggesting that secondary structure provides at least some protection from mutations. There are also considerable differences in the details of the C→T response between the 'high' and 'low' groups. The relative steepness of the early increase in C→T response corresponds to the relatives steepnesses of the A→G responses, but the plateau is lower for the group with the steeper slope (Figure 1). The larger, steep-sloped group also has unexplained drops in the COI and cytb regions, while the smaller group has an unexplained hump at $D_{ssH} = 0.8$ before reaching a slightly slower and constant level.. These analyses suggest that complex yet identifiable changes in mutation processes have occurred during the evolution of primate mitochondria, and the methods presented here should be useful in elucidating the mechanistic bases and historical patterns that have emerged.

[1] Biological Computation and Visualization Center, Dept of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA.

[2] The groups termed 'high' and 'low' consist of species: *Cercopithecus*, *Cynocephalus, Gorilla, Homo, Hylobates, Macaca, Pan, Papio, Pongo*; and *Cebus, Colobus, Lemur, Nycticebus, Tarsius, Trachypithecus, Tupaia*, respectively and 'combined' consists of all the eighteen species. These were chosen based on previous analyses (Raina et al., unpublished data).

4. Figures

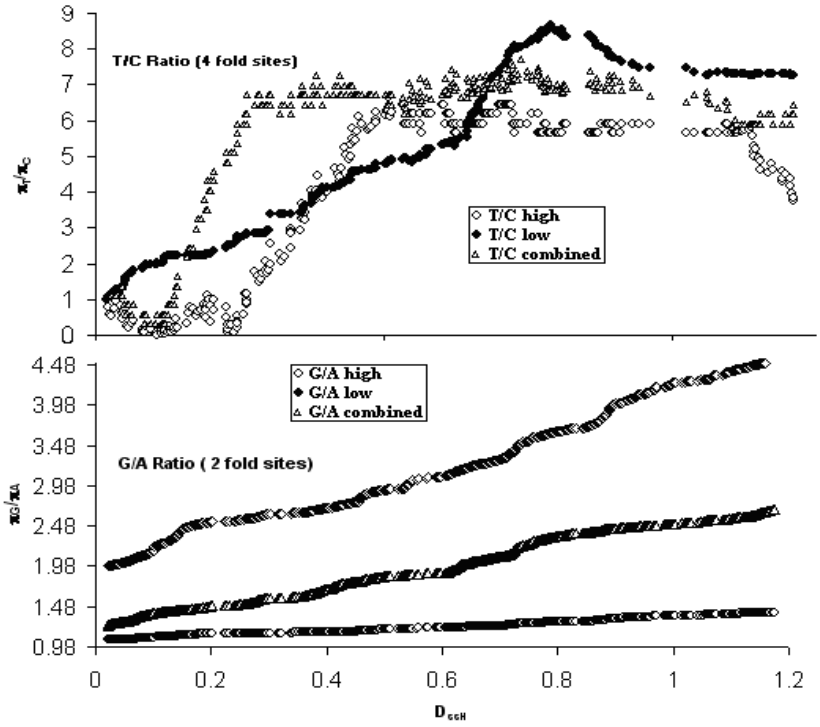


Fig. 1a,b. C→T and A→G response profile for the high, low, combined primate datasets.

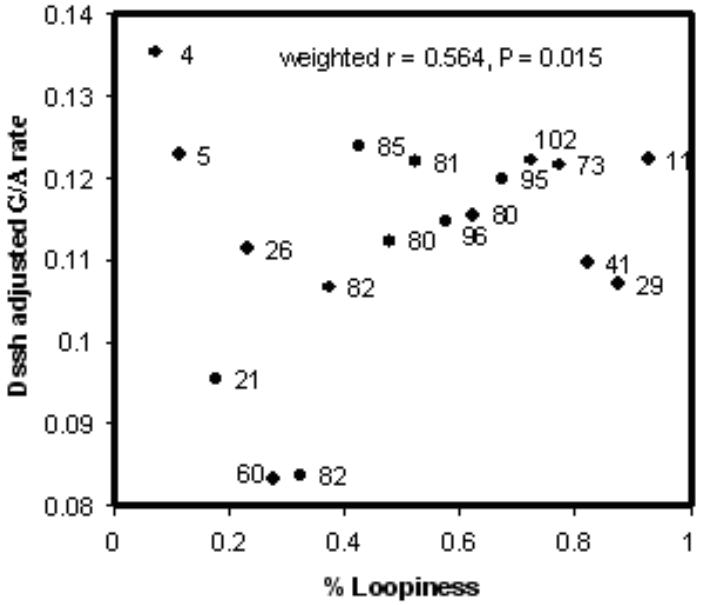


Fig. 2. Site specific G/A ratio adjusted for D_{ssH} as a function of % of alternative structures where that site is in a loop. Numbers of sites pooled into categories according to % loopiness are indicated.

5. References

[4] Bielawski, J.P. and Gold, J.R., 2002 Mutation patterns of H- and L- strand DNA in closely related Cyprinid fishes. Genetics 161: 1589-97
[1] Clayton, D. A., 2000 Transcription and replication of mitochondrial DNA. Hum Reprod 15: 11-7
[3] Faith, J. J. and Pollock D. D., 2003 Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Genetics 165: 735-45
[2] Graziewicz, M.A., Day, B.J., and Copeland W.C., 2002 The mitochondrial DNA polymerase as a target of oxidative damage. Nucleic Acids Research 30: 2817-24
[5] Tanaka, M., and Ozawa, T. 1994 Strand asymmetry in human mitochondrial DNA mutations. Genomics 22:327-35

H13. Length distributions of exons and introns imply the evolutionary constraints for exon/intron length

Yiyu Jia,[1] Yan Zhang[1], Chee Keong Kwoh[1], Vivek Gopalan[2]

Keywords: Intron, Exon, SNPs, Skewness, Jackknife

1 INTRODUCTION

In the postgenomic era, a valuable and interesting undertaking is to investigate the distribution of the number of exon and intron over their lengths. We made experiments on several large intron/exon databases. We observed that both of distributions of exon length from GenBank and Homo Sapiens genome database had single peaks. However, as the most developmental complex organism, the distribution of Homo sapiens database was more symmetric than GenBank. This maybe implies an evolutionary constraints for exon length. Since these distributions do not be any parameterized distributions when tested by χ^2 GOF test, we used resampling methods, including bootstrap and jackknife, to analyze the distribution of exon and intron lengths. Unlike exons, we found that distribution of intron length had several peaks. Coupled with the fact that SNPs in intron sequences account for a large proportion in the total quantity of SNPs, we conclude that introns are under looser evolutionary constraint than exons and introns are functionally more complex than exons.

2 MATERIALS AND METHODS

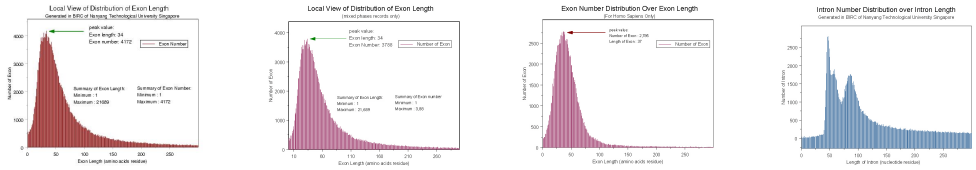
We used two databases, which refer to GenBank release 132 and Homo Sapiens genome database. To make sure that the sequences are low homologous and independent, we purged original databases to be 40% similarity because pairwise sequence alignment methods often fail to correctly align protein pairs with 20-30% pairwise sequence identity and significant sequence identity. Several statistical methods, including χ^2 GOF testing, Skewness testing, Bootstrap, and Jackknife, were used for the analysis. Skewness is a signed measure that describes the degree of symmetry in a distribution; Jackknife after bootstrap could be a technique for providing information on the influence of each observation on the functionals.

3 RESULTS AND DISCUSSION

Our first discovery from purged ExInt database is that 40.677% of the total nucleotides are located in exon region. We got one "mixed" ExInt by deleting sequences, which only contain pure phase0, phase1 or phase2 introns. We also investigated the Homo Sapiens genome database. Their distributions were plotted in Figure 1. According to our results, the exon distribution peaks at lengths of 34 and 37 amino acids residues. After making χ^2 GOF testing, we found that those distribution were neither exponential distribution nor

[1]Bioinformatics Research Centre, Nanyang Technological University Singapore. Yiyu Jia's E-mail: yyjia@pmail.ntu.edu.sg; Yan Zhang's E-mail: yzBirc@pmail.ntu.edu.sg; C.K. Kwoh's E-mail: asckkwoh@ntu.edu.sg

[2]Department of Biochemistry, National University of Singapore. Vivek Gopalan's E-mail: vivek@bic.nus.edu.sg



(a) (b) (c) (d)

Figure 1: Exon and Intron length distribution

Table 1: The shape description for exon length distribution

Methods	ExInt	mixed phases ExInt	Homo sapiens
Skewness value	4.961943	4.651485	3.889752

Poisson distribution. Meanwhile, we investigated the intron length distribution, which is shown in figure 1 (d). Compared with exon, intron length distribution has total three peaks, which make it seems more "nonregular" than exon length distribution. Hence, how and why these peaks formed could be interesting questions. After doing the Jackknife after bootstrap analysis on standard deviation methods, we saw that the points, which are particularly influential to the standard variance for distribution of intron, are much more than exons'. From table 1, it can be seen that exon length distribution of homo sapiens is the most symmetric, the "mixed" one is the second to it, and the ExInt, which could be considered as whole population including homo sapiens and other organisms, is the most asymmetric. Does this phenomena imply evolution constraints for exon length? Because evolution requires only function instead of biochemical structure, we may hypothesize that the amino acids residues with the length $30 - 40$ most likely form functional segments of protein.

4 CONCLUSION

In this article, we show that the distribution of exon length is amazingly regular than that of intron length. By comparing the distributions of the length of exon and intron, we conclude that there is much looser evolution constraints for introns than for exons. These observations fit the Neutral Theory, which predicts that gene regions of functional "importance" should evolve more slowly than less important regions. Further more, introns are much more functionally complex than exons. In fact, it is observed that the numbers of protein-coding genes do not increase exponentially in complex organisms and hence cannot provide large-scale cellular connectivity, which does increase exponentially. Our study on SNPs supports this point. In the level of phenotype, the polymorphism of genes will lead to the different features of an individual. According to dbSNP, there is 67.93% SNPs located in intron regions. Hence, we believe that it is much more challenging to study the function of intron than exons. This is probably one way to solve the questions about the origin and evolution of introns. Treating introns as regulation part of multi-tasked gene networks could be a prospective direction. Investigating other organisms than homo sapiens for clear inside is our future work as well.

H14. Evaluating Indels as Phylogenetic Markers for the Prokaryotes

Timothy G. Lilburn[1] Yufeng Wang[2]

Keywords: indel, gap, alignment, phylogeny, classification

1 Introduction

Currently, the standard for inferring molecular phylogenies of the Prokaryotes uses SSU rRNA sequence comparison. This approach is not perfect, as it cannot satisfactorily shed light on the oldest pages in the evolutionary history of the Prokaryotes and, furthermore, some of the "higher level" taxa established using these sequences contain groups that are similar in sequence but that stand out as dissimilar in phenotype. The use of rare genomic changes (RGCs) to infer phylogenies has been proposed [7] as a way of supplementing or even supplanting other data. Numerous studies using RGCs to resolve difficulties in phylogeny have been published. R. S. Gupta has pioneered the use of indels to examine the phylogeny of the Prokaryotes [2]. In a recent review [3] Gupta defined a set of 25 indels in 21 proteins that were used to establish the root of the Prokaryotic tree and the evolutionary history of this domain.

As for many of the RGCs proposed as characters for inferring phylogenies, the statistics of indels are not well understood [6]. Furthermore, since an indel is defined only in relation to other sequences, the alignment parameters used in the multiple sequence alignment could affect the size and placement of an indel and, therefore, the results of the phylogenetic inference. However, as the mechanisms giving rise to insertion and deletion events are relatively complex, the probability that an indel would appear in the same place in a protein through two separate events is thought to be very low [4]. This means that shared indels are most likely the result of shared descent and have a high phylogenetic information content. The problems of lateral gene transfer, recombination and paralogy can still obscure this information and some researchers have shown that indels can be misleading [1].

2 Methods and Data

In this poster we attempt to assess the robustness of indels as phylogenetic characters. Firstly, we assessed the effects of different alignment algorithms and parameters on the appearance of Gupta's 25 indels. We screened 148 Prokaryotic genomes for homologues to the 21 protein set, using the *E. coli* proteins as the query sequences in a blastp search. The protein sets were aligned using ClustalW and T-Coffee and screened for all indels, including the 25 defined by Gupta. Column statistics for the alignment were generated and only indels occurring in conserved regions were considered. With the sequences in the alignment arranged in order of their place in the Bergey's Taxonomic Outline of the Procaryotes (DOI 10.1007/bergeysoutline200310), indels that corresponded to taxonomic groups were easy to see. The sequences were also arranged according to Gupta's phylogeny and indels supporting or contradicting each classification were scored.

[1] Bioinformatics Department, American Type Culture Collection, 10801 University Boulevard, Manassas, VA USA. E-mail: tlilburn@atcc.org

[2] Department of Biology, University of Texas at San Antonio, 6900 North Loop, 1604 West, San Antonio, TX USA. E-mail: ywang@utsa.edu

Finally, we attempted a phylogenetic analysis that characterizes the indels more rigorously than previously. Gupta scored the indels on a presence or absence, plus/ minus basis, regardless of their length. Simmons and Ochoterena have developed a method called “simple indel coding” [8] that allows indels to be scored in a more sophisticated way as multistate characters. The method has been automated in the GapCoder algorithm [10] and we used the algorithm to score all the indels in the alignments. As the number of indels in each alignment was relatively small, the GapCoder scores were simply examined to see if using the indel as a multistate, rather than binary character contradicted, supported or supported and enhanced the classifications.

3 References

[1] Bapteste, E. and Philippe, H. 2002. The potential value of indels as phylogenetic markers: position of Trichomonads as a case study. *Molecular Biology and Evolution* 19:972-977.

[2] Gupta, R.S. 1998. Protein phylogenies and signature sequences: a reappraisal of evolutionary relationships among Archaebacteria, Eubacteria and Eukaryotes. *Microbiology and Molecular Biology Reviews* 62:1435-1491.

[3] Gupta, R.S. 2003. Evolutionary relationships among photosynthetic bacteria. *Photosynthesis Research* 76:173-183.

[4] Lloyd, D.G. and Calder, V.L. 1991. Multi-residue gaps, a class of molecular characters with exceptional reliability for phylogenetic analyses. *Journal of Evolutionary Biology* 64:9-21.

[5] Notredame, C., Higgins, D.G. and Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* 302:205-217.

[6] Philippe, H. and Laurent, J. 1998. How good are deep phylogenetic trees? *Current Opinion in Genetics and Development* 8:616-623.

[7] Rokas, A. and Holland, P.W.H. 2000. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution* 15:454-459.

[8] Simmons, M.P. and Ochoterena, H. 2000. Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology* 49:369-381.

[9] Thompson, J., Higgins, D. and Gibson, T. 1994. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22:4673-4680.

[10] Young, N.D. and Healy, J. 2003. GapCoder automates the use of indel characters in phylogenetic analysis. *BMC Bioinformatics* 4:6.

H15. A triplet approach to approximations of evolutionary trees

Jesper Jansson, Andrzej Lingas, Eva-Marta Lundell [1]

Keywords: evolutionary trees, consensus methods, approximation algorithms, polynomial time approximation scheme

1 Introduction.

Recently, methods for evolutionary tree construction using partial topological relationships between the species have been proposed and analyzed. Among these are quartet methods [3] which compute unrooted topologies for subsets of cardinality four and combine them to form an unrooted evolutionary tree, and rooted consensus methods. In this paper, we study two variants of the rooted consensus approach: *the maximum agreement subtree problem* (MAST) and *the maximum triplet consistency problem* (MTC).

Let T be a rooted, unordered tree distinctly leaf-labeled by a set of labels S. For any subset S' of S, $T|S'$ denotes the tree obtained by first deleting from T all leaves which are not in S' and all internal nodes without any descendants in S' along with their incident edges, and then contracting every node having just one child. In *the maximum agreement subtree problem* (MAST) the input is a set $\mathcal{T} = \{T_1, T_2, ..., T_k\}$ of rooted, unordered trees, where each $T_i \in \mathcal{T}$ is distinctly leaf-labeled by S, no $T_i \in \mathcal{T}$ has a node of degree 1, and the goal is to find a subset S' of S of maximum cardinality such that $T_1|S' = T_2|S' = ... = T_k|S'$. In the complementary problem of MAST, called co-MAST, the goal is to find a subset $\tilde{S}$ of S of minimum cardinality such that $T_1|(S \setminus \tilde{S}) = T_2|(S \setminus \tilde{S}) = ... = T_k|(S \setminus \tilde{S})$.

Next, a *rooted triplet on* S is a constraint of the form $(\{i, j\}, k)$, where $i, j, k \in S$, which specifies that the lowest common ancestor of i and j is a proper descendant of the lowest common ancestor of i and k. The *maximum triplet consistency problem* (MTC) is the following: given S and a set R of rooted triplets on S, find a rooted, unordered tree with leaves distinctly labeled by S which satisfies as many of the triplets in R as possible.

It is known that that MAST restricted to three trees is already hard to approximate. It is also known that another special case of MAST consisting of instances containing only trees of height 2 is hard to approximate. However, Amir and Keselman [1] gave a factor 4 algorithm with $O(kn^5)$ running time for co-UMAST[2]. In this paper, we obtain a factor $(3 - \frac{6 \log \log n}{\log n})$ approximation algorithm for co-MAST, i.e., the complement of MAST.

Let m be the maximum number of leaves in an agreement subtree of $\mathcal{T}$. Removing our approximate solution to co-MAST from S yields an approximate solution to MAST of size $\geq n - 3(n - m) = 3m - 2n$. This gives a constant approximation factor for MAST whenever $m > 0.67n$ and a good approximation factor for MAST if m is large (which seems to be the case in practice); e.g., if $m \geq 0.95n$ then $3m - 2n \geq 0.85n$.

It is known how to construct a tree which is consistent with all of the rooted triplets in a given set in polynomial time, if such a tree exists. One knows that MTC is NP-hard; also approximation algorithms for MTC are known.

In the consensus approach for the unrooted case, typically input constraints specifying topologies of quadruples of leaves (quartets) are considered. Unfortunately, the consistency problem for quartets is MAX SNP-hard [3]. Despite this, it was shown that the corresponding maximum quartet consistency problem admits a PTAS when the input instance is complete,

[1]Department of Computer Science, Lund University, Box 118, 221 00 Lund, Sweden

[2]UMAST is defined like MAST except that all trees are unrooted and $T|S'$ now denotes the tree obtained by first deleting from T all nodes which are not on any path between two leaves in S' and their incident edges, and then contracting every node with degree 2. co-UMAST is the complement of UMAST.

i.e., contains a constraint for each possible quadruple of leaves. In this paper, we make progress on the open approximation status of MTC by presenting a PTAS for MTC in the dense case where the number of different input triplets is $\Omega(n^3)$.

2 A $(3 - \frac{6 \log\log n}{\log n})$-approximation of co-MAST

Let $H = (V, \mathcal{F})$ be a hypergraph with vertex set V and edge set $\mathcal{F} \subseteq 2^V$. A *vertex cover of* H is a subset C of V such that for every $F \in \mathcal{F}$, there exists a vertex $v \in C$ with $v \in F$. A *minimum vertex cover* is a vertex cover with as few vertices as possible. The size of a minimum vertex cover of H is denoted by $\tau(H)$.

Fact 1 ([2]). *Let $H = (V, \mathcal{F})$ be a 3-uniform hypergraph on n vertices. A vertex cover C of H of size at most $(3 - \frac{6 \log\log n}{\log n}) \cdot \tau(H)$ can be computed in polynomial time.*

Fact 1 can be applied to co-MAST as follows: Construct a 3-uniform hypergraph $H = (S, \mathcal{F})$, where for each cardinality 3 subset S' of S, S' belongs to the set of hyperedges $\mathcal{F}$ if and only if $T_i|S' \neq T_j|S'$ for some $\{i, j\} \subseteq \{1, 2, ..., k\}$. By testing all cardinality 3 subsets of S one at a time, H can be constructed in $O(n^4 k)$ time.

Lemma 2.1 *Let C be a vertex cover of H. Then $T_1|(S \backslash C) = T_2|(S \backslash C) = ... = T_k|(S \backslash C)$.*

Hence, to obtain a factor $(3 - \frac{6 \log\log n}{\log n})$ approximation for co-MAST, compute an approximative minimum vertex cover of H as described in the proof of Lemma 2 and output this as the solution.

3 A PTAS for dense MTC

An instance of MTC is complete if the set R of triplet topologies contains one topology for every one of the $\binom{n}{3}$ possible three-leaf subsets over S. Here we study the more general problem of *dense* instances of MTC, i.e., instances where R contains $\Omega(n^3)$ resolved triplets. Jiang *et al.* [3] have studied the complete problem in the setting of quartets, and given a PTAS for complete MQC. We adopt the method from [3] to construct a PTAS also for dense MTC.

T_{opt} is decomposed into a tree consisting of kernel with the leaves grouped into bins of finite sizes. For each possible such decomposition an approximate optimal assignment of leaves to bins can be found in polynomial time. Hence, the dense MTC can be approximated in the following way:

Theorem 3.1 *There exists a constant c such that for each $\epsilon > 0$, there is a polynomial time algorithm which, for each instance R of dense MTC, produces a tree T_k that approximates T_{opt} in such a way that $|R_{T_k} \cap R| \geq |R_{T_{opt}} \cap R| - (c/k + \epsilon) \cdot n^3 \geq (1 - c/(c'k) - \epsilon/c')|R_{T_{opt}} \cap R|$ where c' is a constant satisfying $|R_{T_{opt}} \cap R| \geq c' n^3$.*

References

[1] A. Amir and D. Keselman. 1997. Maximum agreement subtree in a set of evolutionary trees: Metrics and efficient algorithms. *SIAM Journal on Computing*, Vol. 26, No. 6, pp. 1656–1669.

[2] E. Halperin. 2000. Improved approximation algorithms for the vertex cover problem in graphs and hypergraphs.
In *Proc. 11th ACM-SIAM Symp. on Discrete Algorithms*, pp. 329–337.

[3] T. Jiang, P.Kearney, M. Li. 2001. A polynomial time approximation scheme for inferring evolutionary trees from quartet topologies and its application. *SIAM Journal on Computing* 30(6), pp. 1942–1961.

H16. Dual Multiple Change Point Model Leads to More Accurate Recombination Detection

Vladimir N. Minin, [1] **Karin S. Dorman,** [2] **Marc A. Suchard** [3]

Keywords: Recombination, HIV evolution, MCMC, phylogenetics

Recombination is an important force in the evolution of HIV, promoting the developing of multiple drug resistant strains and deterring the construction of therapeutic vaccines. Consequently, research interests expand in recombination detection methods using multiple sequence alignments that accurately identify the recombination sites. We extend a computational method to detect homologous recombination [3] to improve its recombination detection resolution. In the original model an alignment is partitioned into K segments, where K is an unknown parameter. Each partition $k \in 1, \ldots, K$ has a vector of phylogenetic parameters (Θ_k, τ_k) associated with it, where Θ_k is a vector of parameters describing the nucleotide substitution process and τ_k is a bifurcating tree topology describing evolutionary relationships between sequences. End points of the partitions ξ_k are called change-points. Recombination is inferred if there is at least one change-point ξ_k such that $\tau_{k-1} \neq \tau_k$. This model was successfully applied to test recombination hypotheses in HIV strains by Suchard et al. [2]. Modeling spatial variation of all parameters with a single change-point process results in prior correlation between sites where substitution parameters vary and sites where topologies change. This can lead to loss of accuracy of recombination site identification, when recombination occurs near the boundary of regions with varying evolutionary pressures. Here, we develop a dual Multiple Change Point (MCP) model that decouples substitution parameters change-points from topology break-points by introducing two *a priori* independent change-point processes to describe the variation. We use reversible jump MCMC sampling to approximate the posterior distribution of model parameters [3].

To demonstrate improved accuracy, we start with a previously used test example involving mtDNA sequences from 4 primates and generate a series of datasets with simulated recombination events near a site where evolutionary pressures change greatly. In Figure 1 we plot inferred most probable recombination sites against simulated recombination sites for the single and dual MCP models. The single MCP model shows strong attraction between inferred recombination sites and the evolutionary pressure change. The dual MCP model yields more accurate inference with small variation about the diagonal caused by randomly distributed informative sites. For better interpretation of this variation, we divided all informative sites into two classes: supportive or contradictory of the recombinant structure, denoted in Figure 1 as light and dark gray circles respectively. As expected, the greatest inaccuracies in detection occur when the simulated recombination event is located in an uninformative region, especially those bordered by contradictory sites.

We also apply the dual MCP model to the *gag* gene sequenced from HIV-1 isolate VI557 and 9 different subtype consensus sequences. VI557 is a reported recombinant, with ambiguous support [1]. Figure 2 depicts the marginal posterior probabilities of the different possible topologies for each site in the alignment. We see one region near the 5' end in the alignment with large uncertainty in the topology, but no change in the most probable topology. This suggests little support for recombination in this alignment.

[1]Department of Biomathematics, UCLA, Los Angeles, CA. E-mail: vminin@ucla.edu
[2]Department of Statistics, Iowa State University, Ames, IA. E-mail: kdorman@iastate.edu
[3]Department of Biomathematics, UCLA, Los Angeles, CA. E-mail: msuchard@ucla.edu

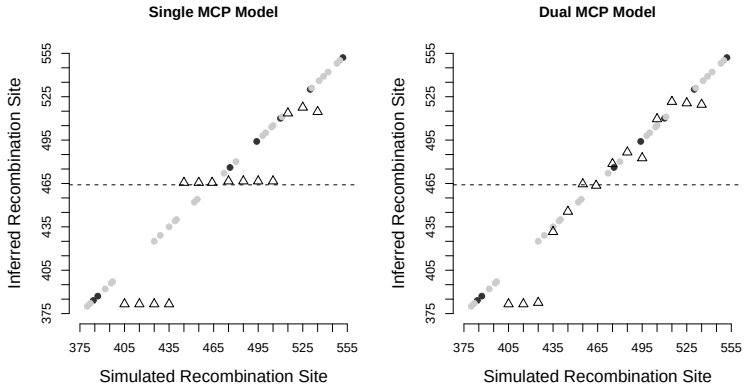


Figure 1: Simulation results. Inferred most probable recombination sites are plotted against simulated recombination sites near a substantial change in evolutionary pressures (fixed at site 464, dashed line). Circles on the diagonal denote informative sites that support (light gray) or contradict (dark gray) the recombinant structure.

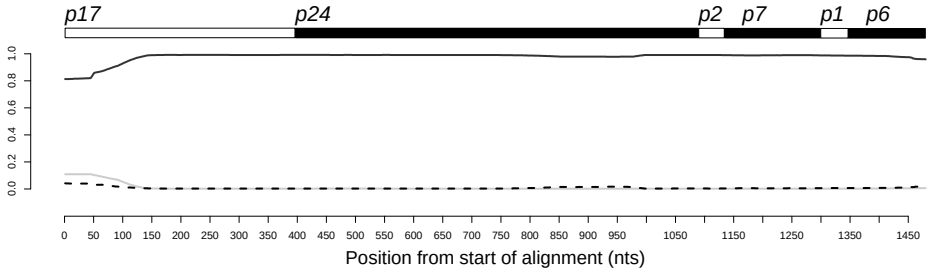


Figure 2: HIV results. Plot shows the locations of the gene products within the *gag* gene and the marginal posterior probabilities of the two most likely tree topologies (light gray, dark gray) and the sum of the marginal posterior probabilities of all other topologies for each site in the alignment(dashed line).

The dual MCP model inherits a major strength of the original MCP model in its realistic modeling of spatial phylogenetic variation using a parsimonious number of parameters. Using two change-point processes results in better sampling of topologies during MCMC simulations (not shown) and increases the accuracy of recombination site identification.

References

[1] Siepel, A. and Korber, B. 1995. Scanning the database for recombinant HIV-1 genomes. Pages III 35-60 in Human retroviruses and AIDS compendium.

[2] Suchard, M. A., Weiss, R. E., Dorman K. S. and Sinsheimer, J. S. 2002. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. *Systematic Biology* 51(5):715–728.

[3] Suchard, M. A., Weiss, R. E., Dorman K. S. and Sinsheimer, J. S. 2003. Inferring spatial phylogenetic variation along nucleotide sequences: A multiple change-point model. *Journal of the American Statistical Association* 98:427–437.

H17. Internal Gene Duplication Patterns in Transmembrane Protein Evolution

Hironori Mitsuke[1], Keisuke Noto[1], Masafumi Arai[1,2], Toshio Shimizu[1]

Keywords: internal gene duplication, transmembrane protein, transmembrane topology

1 Introduction.

It has been revealed that transmenbrane (TM) topology (the number of TM segments (TMSs), the TMS position and the orientation of TMS to the membrane lipid bilayer) evolved by internal gene duplication [1-3]. A large number of TM proteins, such as Bacteriorhodopsin, Na^{+}/Ca^{2+} exchanger and the major facilitator superfamily, have been recognized to have internal repeats that presumably have arisen from the duplication event [2-4]. In this study, we have searched comprehensively for sequences with internal repeat by partial sequences comparison within a single sequence from prokaryotic genomes, and investigated evolutionary pathways of TM topologies by analyzing in detail duplication patterns observed in the sequences.

2 Materials and Methods.

We extracted putative 52,686 TM protein sequences out of 239,359 ORFs of 87 prokaryotic (72 bacterial and 15 archaean) genomes in GenBank by using DetecSig [5], SOSUI [6] and their TM topologies were predicted by ConPred [7]. Next, partial sequences (PSs) containing one or consecutive 2~6 TMSs (1~6-tms) and flanking loop regions were extracted from all the predicted TM proteins. To judge whether or not two PSs were homologous in each TM protein sequence, we defined the threshold values of sequence identity for 1~6-tms PSs (detailed procedure not described here). The threshold identity values thus obtained are 48.4, 33.8, 29.8, 28.2, 27.1 and 26.2% for 1-tms, 2-tms, 3-tms, 4-tms, 5-tms and 6-tms PSs, respectively. The PSs of the same size were compared one another within a single TM protein sequence to find internal repeats which satisfy the defined conditions, i.e., larger sequence identity between them than the thresholds determined above.

3 Results and Discussion.

Finally, we found 377 TM protein sequences in total appeared to have evolved by internal duplication, which were distributed over 70 out of 87 genomes (Figure. 1). Various duplication patterns are observed in the detected sequences. One duplication pattern is “diploid-type”, which occupies 99.4, 8.3, 34.4, 54.3 and 21.8% among the detected TM protein sequences for 4, 6, 8, 10 and 12-tms TM proteins, respectively. If “quasi-diploid-type” is classified as “diploid-type”, these fractions increase to as much as 99.4, 41.6, 75.8, 64.9 and 70.1%. Other duplication patterns than "diploid-type" are prevailing in the TM protein sequences with the odd numbers of TMSs. The case of 7-tms TM protein

[1] Department of Electronic and Information System Engineering, Faculty of Science and Technology, Hirosaki University, Hirosaki 036-8561, Japan. E-mail: mitsuke@si.hirosaki-u.ac.jp (H. Mitsuke), gs0214@si.hirosaki-u.ac.jp (K. Noto), slsimi@si.hirosaki-u.ac.jp (T. Shimizu)

[2] Department of Developmental Biology and Neuroscinece, Graduate School of Life Science,Tohoku University, Sendai 980-8577, Japan. E-mail: d01603@si.hirosaki-u.ac.jp

is interesting in particular. There are mainly three duplication patterns recognized: from primordial 4-tms TM protein with the duplication of 3-tms element, from 3-tms TM protein with the duplication of 3-tms element and the generation of a new TMS, and from 5-tms to 7-tms, as illustrated in Figure 2. The function of the sequences with the first pattern is assigned to "rhodopsin pump" [7], and the second is reported as "zinc transporter". A similar evolutionary pathway to the second one was reported earlier for "lysosomal cystine transporter" [8]. The function of the third one is not known yet at this stage, being remained for future research.

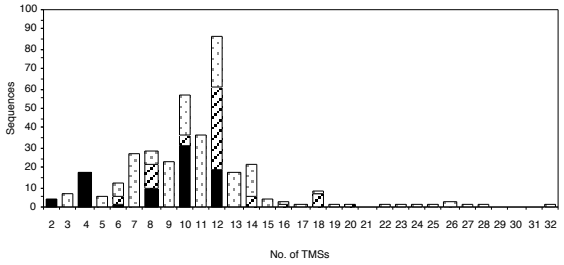


Figure 1: The distribution of the 377 detected TM protein sequences with internal repeat over the number of TMSs, black bars, "diploid-type"; hatched bars, "quasi-diploid-type"; dotted bars, others.

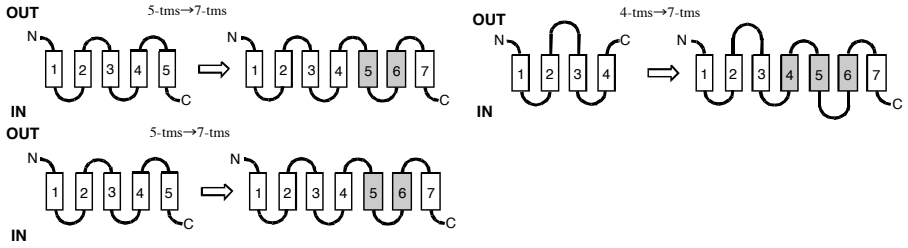


Figure 2: Schematic presentation of three duplication patterns for the evolution of 7-tms TM proteins.

References

[1] Arai, M., Ikeda, M. and Shimizu, T. 2003. Comprehensive analysis of transmembrane topologies in prokaryotic genomes. *Gene* 304:77-86.

[2] Saier, M.H.Jr. 2003. Tracing pathways of transport protein evolution. *Mol. Microbiol.* 48:1145-1156.

[3] Taylor, E.W. and Agarwal, A. 1993. Sequence homology between bacteriorhodopsin and G-protein coupled receptors: exon shuffling or evolution by duplication? *FEBS Lett.* 325:161-166.

[4] Sääf, A., Baars, L. and von Heijne, G. 2001. The internal repeats in the Na^+/Ca^{2+} exchanger-related *Escherichia coli* protein YrbG have opposite membrane topologies. *J. Biol. Chem.* 276:18905-18907.

[5] Lao, D.M. and Shimizu, T. 2001. Methods for detecting the signal peptide in transmembrane and globular proteins. *Genome Informatics* 12:340-342.

[6] Hirokawa, T., Boon-Chieng, S. and Mitaku, S. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14:378-379.

[7] Ikeda, M., Arai, M., Lao, D.M. and Shimizu, T. 2002. Transmembrane topology prediction method: a re-assessment and improvement by consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico Biol.*, 2:19-33.

[8] Zhai, Y., Heijne, W.H.M., Smith, D.W. and Saier, M.H.Jr. 2001. Homologues of archaeal rhodopsins in plants, animals and fungi: structural and functional predictions for putative fungal chaperone protein. *Biochem. Biophys. Acta* 1151:206-223.

H18. A liberal Supertree approach to test the Ecdysozoa hypothesis

Gayle K. Philip[1], Christopher J. Creevey[1], James O. McInerney[1]

Keywords: Coelomata, Ecdysozoa, supertree

1 Introduction.

Researchers have often been divided on the relationship of nematodes to arthropods and vertebrates. Traditionally, vertebrates and arthropods have been grouped together with nematodes occupying a basal position. This classic hypothesis, named "Coelomata" argues that vertebrates and arthropods are more closely related as they have a true body cavity (coelem), which nematodes lack. However, a recent hypothesis now joins the nematodes with the arthropods in a molting clade, the Ecdysozoa. Since the publication of the Ecdysozoa hypothesis, evidence has appeared both for and against it [1,2]. It was our aim to test these hypotheses through the construction of a supertree [3] from all the single gene families identified from all the available eukaryotic genomes.

2 Materials and Methods.

Our approach was to identify single gene, orthologous families containing a minimum of four members, from ten fully characterised eukaryotic genomes. The ten genomes consist of three vertebrates, two arthropods, one nematode, two yeasts, one apicomplexan and one plant species. Hypotheses of relationships for each gene family were re-constructed. The supertree software program Clann (http://bioinf.may.ie/software/clann) was then used to find the supertree that best described the relationships from all the source trees.

3 Discussion.

The tree that best described the relationships in the 780 source trees is shown in Figure 1. We recover this tree using a variety of methods and we have not been able to find a method that does not recover this tree.

In conclusion, it can be seen that our supertree supports the traditional Coelomata hypothesis with the vertebrates more closely related to the arthropods.

[1] National University of Ireland, Maynooth, Co. Kildare, Ireland. E-mail: gayle.k.philip@may.ie

4 Figures.

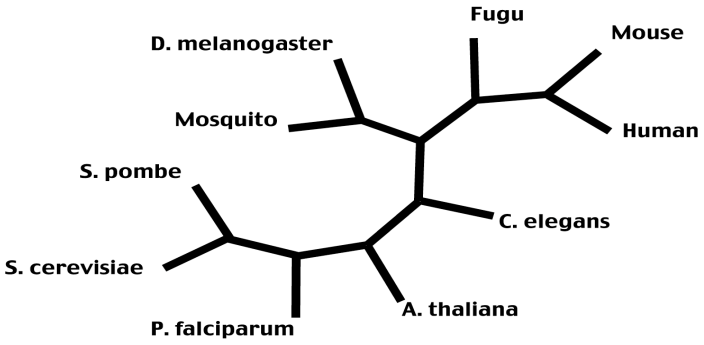


Figure 1. The supertree best describing the relationships in the 780 source trees.

5 References and bibliography

[1] Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A. & Lake, J. A. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387: 489-493.

[3] Bininda-Edmonds, O.R.P., Gittleman, J.L. and Steel, M.A. 2002. The (Super))tree of life: Procedures, Problems, and Prospects. *Annu. Rev. Ecol. Syst.* 33:265-289.

[2] Blair, J.E., Ikeo, K., Gojobori, T., and Hedges, S.B. 2002. The evolutionary position of nematodes. *BMC Evolutionary Biology*. 2:7-13.

H19. Joint Bayesian Estimation of Alignment and Phylogeny

Benjamin D. Redelings, [1] **Marc A. Suchard** [2]

Keywords: Phylogeny, Alignment uncertainty, Bayesian, MCMC

We describe a model and algorithm for simultaneously estimating multiple alignments for biological sequences and the phylogenetic trees that relate the sequences. Unlike current techniques that base phylogeny estimates on a single best estimate of the alignment, we take into consideration the myriads of near-optimal alignments. We also avoid the trap of conditioning on an inaccurate external guide tree in constructing the alignment by estimating the alignment and phylogeny simultaneously. This eliminates the bias towards the guide tree that is inherent in phylogenies based on alignments constructed with progressive alignment [3]. The availability of the phylogeny during alignment construction also allows for more accurate models of both substitution and insertion/deletion that do not over-count single indels and substitutions that are shared between multiple taxa by common descent. Furthermore, this allows us to use shared indels as evidence in clustering taxa on the tree. We note that improved substitution models, such as those allowing invariant sites and rate variation between sites, may improve alignments in the joint estimation framework, whereas currently these models are only available in constructing phylogenies.

We use a continuous-time Markov chain process to describe the substitution process, with extensions for varying rates between sites. While current models implicitly condition on the alignment, we introduce an alignment prior, which allows us to treat the alignment as a parameter to be estimated. Our multiple alignments are built up from pairwise alignments along each branch of the tree. The alignment model is constructed from a hidden Markov model (HMM). We use a HMM with affine gap penalties, which avoids treating long indels as several unit-length indels. In addition to modeling alignments more accurately, this extension is important when using indels to group taxa as it does not exaggerate the number of rare events shared between taxa.

We take a Bayesian approach that allows us to estimate probable phylogenies and alignments, as well as measures of their support, by using Markov chain Monte Carlo (MCMC) techniques to sample from the joint posterior distribution for the phylogeny, alignment, and model parameters. We construct our Markov chain from straightforward Metropolis-Hastings steps for updating branch lengths and substitution parameters and several unique steps for updating the alignment and the topology that rely on dynamic programming. To update the alignment, we use modified versions of the two MCMC steps (branch alignment re-sampling, internal node re-sampling) proposed by [2]. In addition, we introduce a novel MCMC proposal to improve mixing that re-samples both a branch alignment and the internal node at one end of the branch. This proposal decreases burn-in substantially because it allows portions of the alignment to be aligned or unaligned without going through an unfavorable intermediate. We also introduce a new proposal to update the topology based on nearest-neighbor-interchange proposals, with some modifications to deal with internal nodes that lose definition when the topology is changed.

One problem that has intrigued molecular biologists is the question of whether the Archaea form a monophyletic group. To date, some analyses have supported monophyly of

[1]Department of Biomathematics, UCLA, Los Angeles, CA E-mail: bredelin@ucla.edu
[2]Department of Biomathematics, UCLA, Los Angeles, CA E-mail: msuchard@ucla.edu

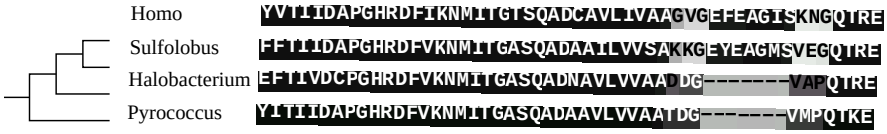


Figure 1: Topology and alignment for a part of EF-Tu. Darker regions represent residues or gaps which are well resolved. Homo (a Eukaryote) and Sulfolobus (an eocyte) share an indel which is not present in the other Archaea, supporting paraphyletic Archaea.

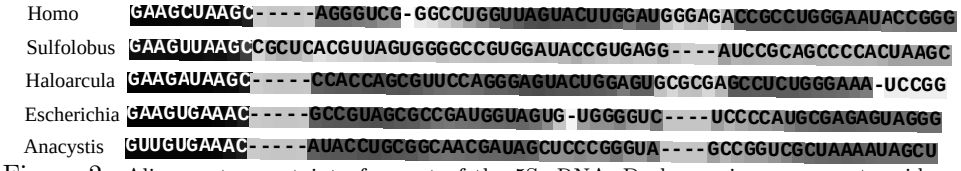


Figure 2: Alignment uncertainty for part of the 5S rRNA. Darker regions represent residues or gaps which are well resolved. The latter half of the alignment is ambiguous (shown), especially in regard to Sulfolobus. This makes the position of Sulfolobus difficult to resolve on the tree.

the Archaea, and some have placed the eocyte Archaea as sister taxa to Eukaryotes. This lack of resolution results partly from the fact that the inference depends on distantly related sequences that are difficult to align. Joint Bayesian estimation of alignment and phylogeny is an ideal method with which to approach this problem; joint estimation can deal with alignment ambiguity, avoids problems of bias in ambiguous alignments, and makes use of more information in the data than current phylogenetic reconstruction methods. To address the issue of the Archaea monophyly, we analyze both the 5S rRNA, and the EF-Tu/EF-1α gene. For each gene, we analyze data sets consisting of 5 taxa and 12 taxa.

The 5S rRNA left the location of the eocyte Archaea unresolved on the tree. However, based on EF-Tu/EF-1α, we find strong evidence against monophyly in that the eocytes are placed as sister taxa to the Eukaryotes (see Figure 1). Furthermore, we find strong support that the remaining Archaea are also paraphyletic. Our strong support for this topology stems from our methodology's use of evidence from common indels shared by eocytes and Eukaryotes. According to [1], about 75% of residues in EF-Tu have very well resolved homology; joint estimation can make use of the information from the 25% of residues with less resolved homology without being overconfident in their alignment. Our methodology can show alignment uncertainty in addition to uncertainty on trees. Figure 2 shows one important reason for the inability of the 5S rRNA to resolve the topology near the root; while some of the alignment is well resolved, approximately half of it is unusable for phylogenetic reconstruction because it is too ambiguous.

References

[1] Baldauf, S. L., Palmer, J. D., and Doolittle, W. F. 1996. The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proceedings of the National Academy of Sciences USA* vol. 93, pp. 7749–7754

[2] Holmes, I. and Bruno, W. J. 2001. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics* vol. 17 no. 9 pp. 802–820

[3] Lake, J. A. 1991. The order of sequence alignment can bias the selection of tree topology. *Molecular Biology and Evolution* vol. 8 pp. 378-385

H20. Conservation Patterns of Human Phosphorylation Sites

Keith Robison[1]

Keywords: signal transduction, molecular evolution, pathways, phosphorylation, protein kinases, comparative genomics

1 Introduction.

The reversible phosphorylation of eukaryotic cellular protein by protein kinases is an important mode of cellular signal transduction, modulating the activity, localization, stability, and protein-protein interaction potential of many proteins. Protein kinases have also become the topic of intense interest in the pharmaceutical industry, with multiple approved drugs and many more in clinical testing. However, the experimental determination of phosphorylated amino acids and their cognate kinases and phosphatases remains experimentally challenging. I have assembled a large (>1900 sites; >700 proteins) catalog of human phosphorylation sites as a basis for examining the conservation of phosphorylation sites. Pairwise comparisons of all human proteins to the complete proteomes of two yeasts (*S.pombe* and *S.cerevisiae*), fly (*Drosophila*), nematode (*C.elegans*), sea squirt (*Ciona*), two fish (*Fugu* and *Danio*) and mouse identify patterns of conservation which shed light on the evolution of phosphosignalling pathways and protein kinase specificities and may enable the prediction of substrates for particular kinases.

2 Methodology.

The phosphosite catalog was built by extracting information from PhosphoBase [1], and the Ingenuity LifeSciences[2] database, and the primary literature. The existence of the correct amino acid at the specified position was verified, and corrections applied where consistent shifts in numbering were identified (such as numbering the first amino acid after the signal peptide as +1).

Human RefSeq proteins were downloaded and the top pairwise match (WuBLAST 2.0, with Smith-Waterman option) used for further analysis. Hits were not filtered for reciprocal best similarity; this allows identifying all cases of a conserved amino acid in a recently expanded protein family, but may overcount conservation for the same reason. Unassembled genomes were searched with TBLASTN, all hit traces assembled with phrap, and the candidate protein translated using a GeneWise alignment of the query protein versus concatenated contigs from phrap.

3 Evolutionary history of amino acids phosphorylated in human.

Comparison of the degree of conservation of phosphorylated amino acids versus the background conservation of Ser, Thr and Tyr shows that phosphorylated positions are only slightly more likely to be conserved across multiple eukaryotes than a randomly drawn amino acid of the same type. Only for positions conserved across all the proteomes looked at are enrichments of 2-fold or more observed, and this effect may be due to a few protein families.

[1] Computational Sciences, Millennium Pharmaceuticals Inc. 640 Memorial Drive, Cambridge MA 02139. E-mail: Robison@mpi.com

The language 'serine-threonine' kinases might engender the naïve assumption that Ser and Thr are equivalent, an equivalence which is not present in the data. Threonine is a much smaller fraction of the catalog than serine, and serine and threonine show no greater exchange at phosphorylated sites than non-phosphorylated ones. Threonine positions are more likely to be strongly conserved, though this may be due to a few specific protein families.

The conservation of phosphosites in homologs from specific species was also examined. Protein phosphorylation is much less extensive in bacteria such as *E.coli* [3,4], but many important metabolic enzymes are regulated via phosphorylation[4,5]. Phosphosites conserved with *E.coli* homologs are rare but do occur. Few phosphosites found in a *Saccharomyces* proteomics study[6] are conserved with human, though analysis of protein cellular abundance estimates for yeast[7] underscores the difficulty of experimental phosphoproteomics.

The Anaphase Promoting Complex (APC) is a multiprotein assemblage which plays a critical role in eukaryotic mitosis. Extensive phosphorylation of APC proteins has recently been mapped [8]. Analysis of these sites reveals that none are conserved across all of the eukaryotes examined, and few are conserved even in all metazoans or deuterostomes. Closer examination of the draft genomes for fragmented proteins and analysis of additional, unannotated draft genomes (chicken, frog, sea urchin) enables a more detailed estimate as to which positions are present broadly and which appear to represent sites unique to the chordate lineage.

4 References.

[3] Cortay, J.C., Rieul, C., Duclos, B., Cozzone, A.J. 1986. Characterization of the phosphoproteins of *Escherichia coli* by electrophoretic analysis. Eur J Biochem 159: 227-237.

[6] Ficarro, S. B. *et al.* 2002. Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat Biotechnol* 20: 301-305.

[2] Ficenec, D. *et al.* 2003. Computational knowledge integration in biopharmaceutical research. *Brief Bioinform* 4(3): 260-278.

[7] Ghaemmaghami, S. *et al.* 2003. Global analysis of protein expression in yeast. *Nature* 425: 737-741.

[5] Hurley, J.H. et al. 1990. Regulation of an enzyme by phosphorylation at the active site. *Science* 249: 1012-1016.

[8] Kraft, C. *et al.* 2003. Mitotic regulation of the human anaphase-promoting complex by phosphorylation. *EMBO J* 22: 6598-6609.

[1] Kreegipuu, A., Blom, N., and Brunak, S. 1999. PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res* 27(1): 237-9.

[4] Link, A.J., Robison, K. and Church, G.M. 1997. Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K12. Electrophoresis 18: 1259-1313

H21. Protein Structure and Evolutionary History Determine Sequence Space Topology

Boris E. Shakhnovich[1], Eric Deeds[2], Charles Delisi[1] and Eugene Shakhnovich[3]
[1]*Bioinformatics Program, Boston University, 44 Cummington Street, Boston MA, 02215*
[2]*Department of Molecular and Cellular Biology, Harvard University*
[3]*Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street, Cambridge MA, 02138*
Keywords: protein domains, evolution, gene family size, power law.

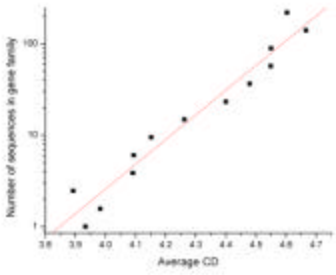
1 Introduction.

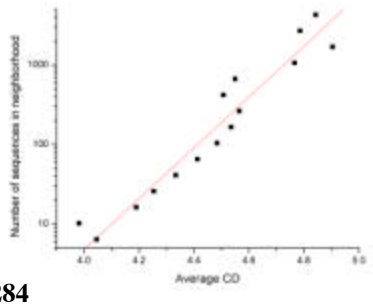
Understanding the observed variability in the number of homologs of a gene is a very important, unsolved problem that has broad implications for research into co-evolution of structure and function, gene duplication, pseudogene formation and possibly for emerging diseases. Here we attempt to define and elucidate the reasons behind this observed unevenness in sequence space. We present evidence that sequence variability and functional diversity of a gene or fold family is influenced by certain quantitative characteristics of the protein structure that reflect potential for sequence plasticity i.e. the ability to accept mutation without losing thermodynamic stability. We identify a structural feature of a protein domain – contact density - that serves as a structural determinant of entropy in sequence space, i.e. ability of a protein to accept mutations without destroying the fold (also known as fold designability). We show that the (log) of the average gene family size exhibits statistical correlation (R^2>0.9.) with the contact density of its three-dimensional structure. We present evidence that the sizes of individual gene families are influenced also by their evolutionary history e.g. the amount of time the gene family was in existence. We further show that our observed statistical correlation between gene family size and designability of the structure is valid on many levels of evolutionary divergence i.e. not only for closely related gene but also for less related fold families.

2 Number of Sequences correlate With Structure

From a biological perspective, gene family size is at least in part influenced by functional constraints related to the number of different but perhaps related functions needed by the cell. For example, some functions such as kinase activity have varied specificities within a relatively small number of sequence mutations[1] while others such as globins have much less functional flexibility despite, in some cases, substantial sequence divergence[2]. From a physical perspective, the potential of a gene to obtain new function upon duplication may depend on its ability to accept mutations without destroying the three-dimensional structure of a protein domain that it encodes.

Our first task is to determine what, if any, physical factors are responsible for the variability in gene family size. To this end we define an inherent *structural* characteristic related to the number of sequences that a structure can accommodate without loss of thermodynamic stability i.e. we employ a structural determinant of designability. This feature has been previously hypothesized to be one of the key influences responsible for over-representation of some folds over others. Recent analysis suggested that structures with greater values of traces of powers of their contact matrices (CM) (i.e. Tr[CM], Tr[CM] etc) are predicted to be more designable[3]. Sequence space Monte Carlo calculations for simple lattice models show that this characteristic of a structure does indeed correlate strongly with its designability that we define as logarithm of the number of sequences that are stable in the structure. Remarkably, we observe that there is a marked positive correlation between a domain's designability calculated via CD and the average gene family size of that domain



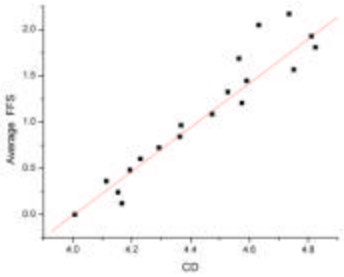


We perform a similar analysis on distantly related gene families as defined through the structural comparisons within the PDUG. To this end we take the structural neighbourhood of a given domain to be all those domains that are connected to it by an edge on the PDUG[4]. Physically this means that all domains that are structurally but not sequentially similar to a given domain (beyond some threshold Z-score value) are included in this structural neighbourhood. We then look at the correlation between the sizes of families of gene sequences that

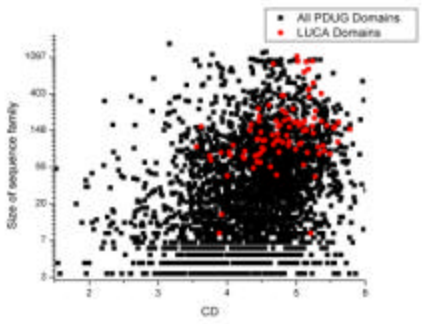
fold into structures belonging to the same structural neighbourhood on the PDUG and the average CD for that neighbourhood. This shows that average the CD, which serves as a proxy for average designability of a structural neighbourhood, itself correlates with the (log) of the gene sequence family size of that neighborhood

3 Structural and functional flexibility.

Next, we determine how gene family size is related to the diversity of functions that family performs. We define the *functional* determinant of a gene family as entropy in function space. When we calculate this measure in the context of PDUG, we utilize Gene Ontology (GO)[5] to define the functional variability (functional flexibility score or FFS) of a set of genes. Perhaps not surprisingly, FFS statistically correlates with CD (Fig 4). This is not surprising because FFS statistically correlates with the total number of sequences in a gene family (data not shown). However, this analysis serves two purposes. First the correlation of FFS and CD shows that designability directly affects the underlying biology of the domain. Domains with low CD have a much lower chance of performing many different functions. Secondly, this serves as a corroboration of the previous result using a different database, annotation method, and a completely different measure of entropy.



4 Evolution matters



We present a scatter plot of gene family size versus CD that shows all domains in the PDUG. The scatter is very significant and it is clear that CD is hardly a predictor of gene family size for an each domain. This is perhaps not surprising given that other factors may have influenced gene family sizes. Any domain that exists in every proteome within a given set can be placed in the last universal common ancestor (LUCA)[6] of that set representing domains that were the predecessors of all others. We may thus highlight the LUCA domains on the scatter plot. Two observations are immediately apparent. First, LUCA domains clearly feature greater CD's, suggesting that ''first'' domains were more designable. (difference of means .48, t-test P-value is <1e-14) Secondly, even at equal CD (designability) with their younger counterparts, LUCA domains feature greater family sizes, on average 116 more members (red points are markedly shifted towards higher family size in Fig.5, P-value < 1e-14). This observation provides evidence that, as simulations on simple lattice models suggest, designability is only the *potential* for larger family size that has to be coupled with other mitigating factors for a full understanding of the evolutionary history of that domain. For two domains with the same CD but differing times of divergence, the domain with the longer divergence time will most likely have more sequence members.

5 References

1. Manning, G., Plowman, G. D., Hunter, T. & Sudarsanam, S. (2002). Evolution of protein kinase signaling from yeast to man. *Trends Biochem Sci* 27, 514-20.
2. Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold. Unique features of the globin amino acid sequences. *J Mol Biol* 196, 199-216.
3. England, J. L. & Shakhnovich, E. I. (2003). Structural determinant of protein designability. *Physical Review Letters* in press.
4. Dokholyan, N. V., Shakhnovich, B. & Shakhnovich, E. I. (2002). Expanding protein universe and its origin from the biological Big Bang. *Proc Natl Acad Sci U S A* 99, 14132-6.
5. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-9.
6. Mirkin, B. G., Fenner, T. I., Galperin, M. Y. & Koonin, E. V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol* 3, 2.

H22. Minimal Convex Recoloring of Phylogenetic Trees

Shlomo Moran, [1] **Sagi Snir** [2]

Keywords: Phylogenetics, Perfect Phylogeny, Graph Algorithms, Fixed Parameter Tractability, Approximation Algorithms

1 Introduction.

Given a set of taxa (a group of related biological species), the goal of phylogenetic reconstruction is to build a tree which best represents the course of evolution for this set over time. The leaves of the tree are labeled with the given, extant taxa. Internal vertices correspond to hypothesized, extinct taxa. Phylogeny reconstruction methods are broadly divided into *character-based* and *distance-based* methods.
Distance based methods construct trees with weighted edges whose pairwise tree distances approximate numerical "evolutionary distances". In contrast, character based methods work directly on character data, which describe biological attributes of the species under consideration that are used to reconstruct phylogenetic trees. A natural biological constraint is that the reconstructed phylogeny has the property that each of the characters could have evolved without reverse or convergent transitions: In a reverse transition some species regains a character state of some old ancestor whilst its direct ancestor has lost this state. A convergent transition occurs if two species possess the same character state, while their least common ancestor possesses a different state. The concept behind this constraint is of "innovation". That is, each time the character state changes, it acquires a new state. A character exhibits neither reverse nor convergent transitions, is denoted *homoplasy free* character. In nature, homoplasy does occur, however these events are considered relatively rare. The acquisition of teeth by the birds ancestor, the Archaeopterix, and their subsequent loss is an example of reverse transition, and the convergence of evolution in placental and Australian mammals is an example of convergent transition.

In graph theoretic terms, a character in a phylogenetic tree is homoplasy free if it is convex, that is: for each state of this character, all species (extant and/or extinct) possessing that state induce a subtree. Thus, the above discussion implies that in a phylogenetic tree, each character is likely to be convex or "almost convex". This makes convexity a fundamental property in the context of phylogenetic trees.

In the Perfect Phylogeny (PP) problem, the tree is unknown and the task is to find a tree on which each of the input characters is homoplasy free. PP was shown to be NP-Complete by [1] and independently by [4].

In this work we deal with the following aspect of PP. For some set of species, the evolutionary tree, T, is believed to be known (e.g. the primates tree shown in Figure 1). Given a character relating to a subset of all the species (extant and extinct) in T, we want to know how much this character "agrees" with T. That is, how far this character is from PP on T. In [2] this distance, denoted the penalty of the character on T, was defined as the number of violation to convexity. In this paper we study another natural measure for this distance: the minimal number of species whose states should be changed to make the given character convex. Indeed, the evolution of a character which can be made convex by removing a small number of exceptions, can be explained by searching biological reasoning for few exceptional

[1]Computer Science dept., Technion, Haifa 32000, Israel. moran@cs.technion.ac.il
[2]Computer Science dept., Technion, Haifa 32000, Israel. ssagi@cs.technion.ac.il

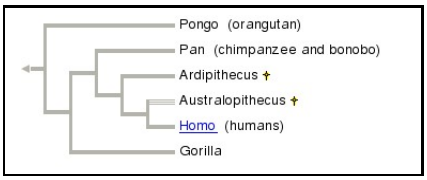


Figure 1: The primates tree, taken from "The Tree of Life" project.

phenomena, as was done for the two above mentioned cases (see e.g. [3]). If however a very large number of state changes is needed to make the character convex, a biological explanation becomes less probable. Therefore, the reliability of the given phylogeny for describing the evolution of this character diminishes accordingly.

First we show that the above problem is NP-hard even for a simple tree - the string, and for the case where character states are given only at the leaves (so that changes on extant species are not counted); we also consider a variant of the problem, in which an atomic operation is a block-recoloring, which changes the color of all the vertices in a given block of the input to a different color. We show that finding the minimum number of block-recolorings needed to obtain convexity is NP-Hard as well. This last result is a bit surprising, since it can be shown that this implies that computing the number of violations that must be removed for making the coloring convex is NP-hard, while trivially, computing the number of violations to convexity can be done efficiently (eg, by finding the parsimony score of the given coloring).

On the positive side, we present an efficient algorithm when the number of states (colors) at the character is fixed. Then we show that for strings and trees of bounded degree, the (unweighted version of the) problem can be solved by a fixed parameter tractable algorithm. Finally, we present polynomial time 2-approximation algorithm for the string version and a 3-approximation algorithm for tree version.

References

[1] H. Bodlaender, M. Fellows, and T. Warnow. Two strikes against perfect phylogeny. In *Proceedings of the 19th International Colloquium on Automata, Languages, and Programming*, Lecture Notes in Computer Science, pages 273–283. Springer Verlag, 1992.

[2] D. Fernndez-Baca and J. Lagergren. A polynomial-time algorithm for near-perfect phylogeny. *SIAM Journal on Computing*, 32(5):1115–1127, 2003.

[3] E. J. Kollar and C. Fisher. Tooth induction in chick epithelium: Expression of quiescent genes for enamel synthesis. *Science*, 207:993–995, 1980.

[4] M.A. Steel. The complexity of reconstructing trees from qualitative characters and subtress. *Journal of Classification*, 9(1):91–116, 1992.

H23. Evolutionary Study of Amino Acid Substitution Patterns Associated with Accelerated Evolution in Endosymbiotic Bacteria

Jun-ichi Takeda[1], Takeshi Itoh[1], Tadashi Imanishi[1], Takashi Gojobori[1]

Keywords: endocellular symbiont, evolutionary rate, amino acid substitution, genome evolution

1 Introduction.

It was previously reported that the evolutionary rate of *Buchnera* spp., endocellular symbionts of aphids is about two times faster than that of *Escherichia coli*, their free-living cousin, and it was suggested that a major factor of this evolutionary rate acceleration is the enhanced mutation rate rather than fixation of slightly deleterious mutations [1]. On the other hand, other studies about the amino acid replacement indicates that the higher evolutionary rate in *Buchnera* is due to the slightly deleterious or beneficial mutation [2, 3]. If the neutral evolution of the enhanced mutation rate is a major factor, the amino acid substitution patterns in *Buchnera* may not change very much compared with those in *E. coli*, while accumulation of slightly deleterious or beneficial mutations would result in a drastic change of the substitution patterns (e.g. increase of radical substitutions). Here we examined the substitution patterns by using the Grantham matrix, which can be used to estimate differences of the amino acid substitution properties [4].

2 Materials and Methods.

Date set
We selected *Buchnera aphidicola* as an endocellular symbiont and *Escherichia coli* as a free-living bacterium. We used seven complete sequences of the following prokaryotes: *Buchnera aphidicola* str. APS (*Acyrhosiphon pisum*), *Buchnera aphidicola* (*Schizaphis graminum*), *Buchnera aphidicola* (*Baizongia pistaciae*), *Escherichia coli* K12 MG1655, *Escherichia coli* O157:H7 RIMD 0509952, *Haemophilus influenzae* Rd KW20, and *Salmonella enterica* subsp. *enterica serovar* Typhi CT18.

Comparison of amino acid substitutions
We used 85 genes, of which orthology was confirmed manually [1]. We made amino acid alignments of orthologs among *Buchnera aphidicola* str. APS, *Escherichia coli* K12 and *Haemophilus influenzae* by using ClustalW. Amino acid substitutions in each lineage were calculated by using the maximum parsimony method. Then, we examined the difference of amino acid property during the substitution by using the Grantham matrix, classifying them into four categories: very radical, radical, moderate, and conservative [6]. Likewise, we carried out the same analysis for *Buchnera aphidicola* str. APS, *Buchnera aphidicola* (*Schizaphis graminum*) and *Buchnera aphidicola* (*Baizongia pistaciae*) (outgroup), and for *Escherichia coli* K12, *Escherichia coli* O157 and *Salmonella enterica* (outgroup).

Motif search

[1] Biological Information Research Center, National Institute of Advanced Industrial Science and Technology, Time24 Bldg. 10F 2-45 Aomi, Koto-ku, Tokyo 135-0064, Japan. E-mail: jtakeda@jbirc.aist.go.jp

We searched for functional motifs in *Escherichia coli* K12 by using InterProScan, and examined the positions of the amino acid substitutions in terms of the locations of the motifs. The motif locations in *Buchnera aphidicola* str. APS were determined by using the motifs found in *Escherichia coli* K12.

Estimation of d_N/d_S
We estimated the nonsynonymous-synonymous ratios (d_N/d_S) of 85 genes between *Buchnera* spp. and between *Escherichia coli* and *Salmonella enterica* by using the Nei-Gojobori method.

3 Results.

If two-fold accelerated evolution in *Buchnera* is because of fixation of slightly deleterious or beneficial mutations, it was expected that the radical and very radical amino acid substitutions increase. However, the amino acid substitution patterns in *Buchnera* were not largely different from those in *Escherichia coli*. In addition, the amino acid substitute patterns were not very different between the inside and outside of functional motifs in the two species. Although d_N/d_S of *Buchnera* was larger than that of *Escherichia coli*, d_N/d_S of *Buchnera* is not different from other species except *Escherichia coli* [5]. These observations implies that the acceleration of the evolutionary rate is due to the enhanced mutation rate rather than the slightly deleterious or beneficial mutations.

References

[1] Itoh, T., Martin, W. and Nei, M. 2002. Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proceedings of the National Academy of Sciences USA* 99:12944-12948.

[2] Shigenobu, S., Watanabe, H., Sakaki, Y., and Ishikawa, H. 2001. Accumulation of species-specific amino acid replacements that cause loss of particular protein functions in *Buchnera*, an endocellular bacterial symbiont. *Journal of Molecular Evolution* 53:377-386.

[3] Moran, A. N. 1996. Accelerated evolution and Muller's racher in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences USA* 93:2873-2878.

[4] Grantham, R. 1973. Amino acid differences formula to help explain protein evolution. *Science* 185:862-864.

[5] Fares, M.-A., Barrio, E., Sabater-Munoz, B. and Moya, A. 2002. The evolution of the heat-shock protein GroEL from *Buchnera*, the primary endosymbiont of aphids, is governed by positive selection. *Molecular Biology and Evolution* 19:1162-1170.

[6] Li, W.-H., Wu, C.-I., and Luo, C.-C. 1985. A new method for Estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* 2:150-174.

H24. Genome comparison allowing complex rearrangements

Mariel Vazquez[1], Dan Levy[2], Rainer K. Sachs[3]

Keywords: genome comparison, rearrangement multigraph, radiation cytogenetics, evolution

Modern studies of comparative genetics in evolution and of tumor cytogenetics analyze large-scale genome rearrangements between species or between normal and cancer cells respectively ([2], [3]). When studying evolution of genomes the rearrangements (apart from fusions and fissions) are usually reduced to a sequence of reactions involving just two DNA breaks (i.e. reversals within one chromosome and simple translocations between two chromosomes); sometimes some events involving three breaks (e.g. transpositions as in figure 1) are considered ([4]).

Radiation cytogenetics is concerned with rearrangements of the genome caused by ionizing-radiation, especially with chromosome aberrations observed at the first mitosis after irradiation. Radiation cytogenetics has strong similarities to comparative genomics at the coarse-grained, large-scale level of Zoo-FISH or synteny-bloc studies.

We have developed a mathematical framework, related to the well-explored theory of cubic (i.e. 3-regular) multigraphs, for characterizing genome rearrangements, including *complex* ones involving 3, 4 or more breaks in reactions not reducible to a sequence of simpler reactions each one of which involves fewer breaks [6]. A genome "rearrangement multigraph" specifies not only breakpoints and misrejoinings but also the way in which one or more chromosomes are involved (figure 1). It defines a unique cycle decomposition. For example, reversals (*i.e.* inversions) and simple translocations are 2-break cycles, c2. *Complex* rearrangements such as musical chair rearrangements are higher order cycles (*i.e.* cN for N>2) or are sequences of such cycles (e.g. c4+c3+c2+c2+c1 as in figure1C). In radiation data, higher order cycles are biomarkers of densely ionizing radiation (such as neutrons or alpha particles) [5]. In contrast to evolutionary comparative genomics these higher order cycles are usually not regarded as successions of simpler events; the evidence speaks against the idea that there are enough cryptic breaks or reused breakpoints to account for the data in terms of multiple 2-break cycles c2. Rather, higher order cycles are considered to characterize the complexity of the DNA interactions (radiation-induced breaks followed by enzymatic misrejoinings).

We are conducting searches for higher order cycles (cN for N>2) in the comparative genomics literature and databases. When comparing mouse and human genomes small (<1Mb) inversions, duplications, transpositions and deletions are common [1]. By lowering the resolution of the magnifying glass one can consider each genome as a short sequence of blocks (synteny blocks) and analyze the rearrangements taking one sequence of blocks into the other. In [2], Pevzner and Tesler compare human and mouse genomes by sorting by reversals. We here explore the possibility that large-scale insertions and transpositions (c3) and higher order cycles cN occur.

[1] Mathematics Department, 970 Evans Hall, University of California Berkeley, Berkeley CA 94720-3840, USA. E-mail: mariel@math.berkeley.edu

[2] Mathematics Department, 970 Evans Hall, University of California Berkeley, Berkeley CA 94720-3840, USA. E-mail: levyd@math.berkeley.edu

[3] Mathematics Department, 970 Evans Hall, University of California Berkeley, Berkeley CA 94720-3840, USA. E-mail: sachs@math.berkeley.edu

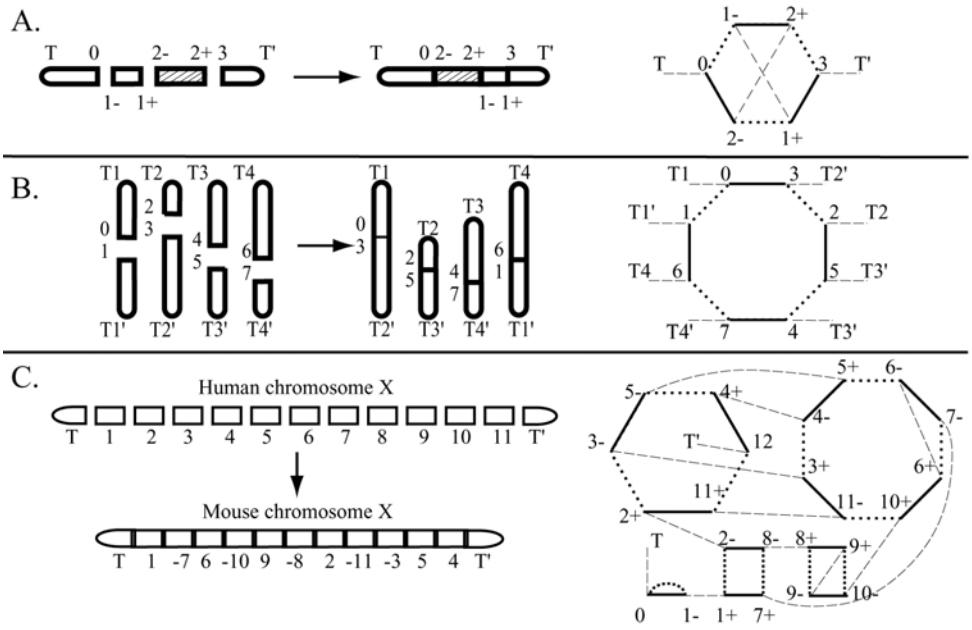


Figure 1: **A.** A transposition involving three breaks in one chromosome, and the corresponding rearrangement multigraph. Telomeres are represented by a vertex each and each breakpoint gives rise to two vertices in the graph. The edges represent chromatin segments (dashed lines), initial partnership between two free ends of one break (dotted lines), and misrejoinings (solid lines), The cycle structure is obtained by considering just the initial and misrejoining edges. In this case this procedure gives the cyclic graph with 6 vertices, i.e. a 3-break cycle c3. **B.** 4-chromosome musical-chairs rearrangement and its rearrangement multigraph, c4. The multigraph uniquely specifies the rearrangement, even with four chromosomes involved. C. Human chromosome X and its rearrangement into mouse chromosome X as presented in [3]; c4+c3+c2+c2+c1. Assuming no cryptic breaks or breakpoint reusage, the multigraph shown in C results. Likewise, the multigraph can be used to characterize c2 interaction by adding cryptic breaks.

References

[1] Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. and Haussler, D. 2003 Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences USA* 100(20):11484-11489.

[3] Pevzner, P.A. and Tesler, G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences USA* 100(13):7672-7677.

[4] Raphael, B.J., Volik, S., Collins, C. and Pevzner, P.A. 2003. Reconstructing tumor genome architectures. *Bioinformatics*,19 Suppl 2: II162-II171.

[5] El-Mabrouk, N. 2001 Sorting signed permutations by reversals and deletions/insertions of contiguous segments. *Journal of Discrete Algorithms,* 1(1): 105-122.

[6] Hlatky, L., Sachs, R.K., Vazquez, M., Cornforth, M.N. 2002 Radiation-induced chromosome aberrations: insights gained from biophysical modeling. *Bioessays,* 24(8):714-23.

[7] Sachs, R.K., Arsuaga, J, Vazquez, M., Hlatky, L., Hahnfeldt, P. 2002 Using graph theory to describe and model chromosome aberrations. *Radiation Research,* 158(5):556-567.

H25. A Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy

Qiong Wang[1], George M. Garrity[1], James M. Tiedje[1], James R. Cole[1]

Keywords: classification, naïve Bayesian classifier, ribosomal RNA, bacterial taxonomy

1 Introduction.

Starting in the mid '80s, Carl Woese revolutionized the field of microbiology with his ribosomal RNA-based phylogenetic comparisons delineating the three main branches of life [5]. Today, rRNA based analysis remains a central method in microbiology, used not only to explore microbial diversity, but as a day-to-day method for bacterial identification. rRNA identification (classification) methods, as opposed to phylogenetic (clustering) methods have been hindered due to the lack of a consistent higher-level bacterial classification structure (taxonomy). This situation changed recently, when in 2002, Bergey's Trust published a revised higher-order taxonomy attempting to reconcile bacterial taxonomy with rRNA based phylogeny [2].

We have developed a naïve Bayesian classifier for classifying bacterial rRNA sequences into the new Bergey's bacterial taxonomy. This classifier is fast, does not require sequence alignment and works well with partial sequences. (The vast majority of rRNA sequences in the public databases are partial.) This classifier is currently being used internally by the Ribosomal Database Project [1] (RDP; http://rdp.cme.msu.edu) to organize its publicly available sequence library.

2 Data and Methods.

Small subunit ribosomal RNA sequences from approximately 4400 bacterial species type strains in 900 genera were obtained from Bergey's Trust, along with associated taxonomic assignment information [2]. The sequences averaged 1459 bases in length with a range of 1200 - 1775 bases. These training sequences were each labeled with a set of taxa, from domain to genus. The classifier uses a feature space consisting of all possible eight-base subsequences (words) in the query molecule. Word-specific priors were calculated from their frequency in the entire training set. As with text-based Bayesian classifiers, only those words occurring in the query contribute to the score [3]. A similar word-based classification scheme has been used to search for horizontal gene transfer events in whole-genome sequences [4].

To classify a query, the joint probability of observing the words in the query was calculated separately for each genus from the training set probability values. For bootstrap analysis, the collection of all overlapping unique words in the query was first calculated. Then a subset of these words was randomly chosen (with replacement) and the words in this subset were then used to calculate the joint probability. (Since overlapping words are highly dependent, we conservatively chose only one-eighth of the words for each trial.) The number of times a genus was selected out of 100 bootstrap trials was used as an estimate of confidence in the assignment to that genus. For higher-rank assignments, we sum the results for all genera under each taxon.

[1] Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824, E-mail: {wangqion, garrity, tiedjej, colej}@msu.edu
This research was supported by DOE-OBER grant DE-FG02-99ER62848 & NSF grant DBI-0328255.

3 Results.

We tested the classifier by exhaustive leave-one-out testing. For each test, we reserved a single training set sequence as query and re-trained the classifier on the remaining sequences. The process was repeated for all sequences in the training set. In addition to the near-full-length sequences, we also tested the classifier on small contiguous regions of 400 and 200 bases chosen at random from the test sequences (Table 1). For the near-full-length and 400 base partial rRNA sequences, the classifier was highly accurate down to the genus level, while with 200 base partial sequences the classifier was accurate at the phylum and class levels. The bootstrap provided a good estimate of classification reliability (Table 2). Overall, 90.4% of taxon assignments matched in 95 or more of the 100 bootstrap trials and these assignments were correct 98.7% of the time.

This classifier is fast enough to handle large sample volumes. On a 1Ghz Apple G4 processor, it can classify approximately 5 sequences per second (with 100 bootstrap samples of each). The new taxonomy is still evolving as species are reevaluated and discrepancies are resolved. As these changes occur, it has proved relatively simple to re-train the classifier and update the assignments of the greater than 86,000 sequences in the RDP library.

length	phylum (%)	class (%)	order (%)	family (%)	genus (%)
1459	99.4	98.7	97.2	94.2	91.0
400	99.2	98.4	96.6	93.0	87.7
200	91.6	87.4	77.3	60.4	46.5

Table 1: Classifier accuracy at different taxonomic ranks for varying query lengths.

	Number of bootstrap assignments out of 100 trials					
rank	100-95	94-90	89-80	79-70	69-60	59-50
phylum	4332/4340†	51/51	19/19	10/10	6/6	6/13
class	4186/4214	40/50	41/42	18/20	15/18	4/6
order	4005/4058	79/81	55/61	31/37	27/30	18/26
family	3705/3786	110/117	98/126	54/72	47/68	37/59
genus	3025/3113	192/207	157/191	97/126	60/95	69/108
overall	98.7%	93.2%	84.3%	79.2%	71.4%	63.2%

† Number of correct assignments over total number of assignments.

Table 2: Classifier accuracy versus bootstrap confidence estimate.

References

[1] Cole, J.R., Chai, B., Marsh, T.L., Farris, R.J., Wang, Q., Kulam, S.A., Chandra, S., McGarrell, D.M., Schmidt, T.M., Garrity, G.M. and Tiedje, J.M. 2003. The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic Acids Res*, 31(1):442-3.

[2] Garrity, G.M., Bell, J.A. and Lilburn, T.G. 2003. Taxonomic outline of the prokaryotes. *Bergey's Maunal of Systematic Bacteriology*, Second Edition. Release 4.0. New York: Springer-Verlag. DOI: 10.1007/bergeysmanual.

[3] Li, Y.H. and Jain, A.K. 1998. Classification of text documents. *The Computer Journal*, 41(8):537-546.

[4] Sandberg, R., Winberg, G., Branden, C.I., Kaske, A., Ernberg, I. and Coster, J. 2001. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Research*, 11(8):1404-9.

[5] Woese, C.R., Kandler, O. and Wheelis, M.L. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci*, 87(12):4576-9.

H26. Amino Acid Coevolution in COI

Zhengyuan Wang[1], David Pollock[1]

Keywords: coevolution, residues, phylogeny, C-alpha distance, domain, sequence, amino acid

1 Introduction

Coevolution in proteins results from structural and functional constraints, and the study of coevolution can therefore provide valuable information about protein structure and function. Noise resulting from phylogeny is a major problem of coevolutionary analysis [1] [2], and to overcome this problem a large amount of data and efficient algorithms are required.

The present study analyzed cytochrome c oxidase subunit I (COI) from 253 vertebrata species using maximum likelihood analysis [2]. Our results show that coevolution is related to pairwise residue distance, domain structure, secondary structure, and protein function. The detected differential coevolutionary patterns in transmembrane helices and other structures suggests that coevolutionary analysis might be a good method to predict helical structures based on sequences.

2 Methods and Materials

Vertebrata COI sequences were downloaded from GenBank and aligned using Clustalx. A phylogenetic tree was constructed using all 13 mitochondrial encoded proteins by NEIGHBOR of PHYLIP [3]. Branch lengths were added using PROML of PHYLIP [3].

Following the method of Pollock et al. [1], aligned COI sequences were first segregated according to residual physico-chemical characters (burial preference, polarity, or side chain volume) and models of independent and dependent were analyzed. To decrease noise, only sites with non-zero and less than average substitution rates were considered. For each pair of sites considered, the primary statistic was the ratio of maximum likelihoods (MLs). Pairs exhibiting significant likelihood ratios (greater than the upper 0.2% of likelihood ratio values from parametric bootstrapping) were extracted as hypothetical coevolved pairs. These pairs were further examined based on C-alpha distances, domain locations, and secondary structures.

3 Results

Stronger coevolution was detected in transmembrane helices, indicating that the degree of coevolution is domain dependent (Table 1). Pairs with short distances exhibited stronger coevolution, suggesting that residual distance is an important factor for coevolution (Figure 1). Functional properties are also related to coevolution because a large cluster of sites (20 sites) that showed correlated coevolution was found to co-localize with supposed H^+ and water release channels. In the central region of transmembrane helices (the first loop on the two ends of each helix were excluded), pairs separated by less than 11 residues along the chain had very weak coevolution (0 coevolved pairs detected) with side chain volume segregation (Table 2). This weak

[1] Department of Biological Sciences, Louisiana State University, Baton Rouge, LA70803, USA. E-mail: zwang3@lsu.edu, dpollock@lsu.edu

coevolution could be one character of the evolution of the transmembrane helices and may be used for secondary structure prediction.

Clustering	Locations	Pairs observed	Percentage of coevolved pairs
Polarity segregation	All locations	6293	2.78
	Transmembrane domain	2926	3.73
	Surface domains	345	2.32
	Across domains	3022	1.92

Table 1: Differential coevolution in different domain structures. Results segregated according to residual polarity are shown. Results of other two segregations are similar.

	Burial preference segregation		Side chain volume segregation	
	TotalPairs	CoevolvedPairs	TotalPairs	CoevolvedPairs
CRH	66	5	67	0
Non-CRH	133	11	148	13

Table 2: Comparing coevolutions in central regions of helices (CRH) to other regions (Non-CRH). Pairs are those with two sites 10 or fewer residues apart. Results of polarity segregation are similar to those of burial preference segregation

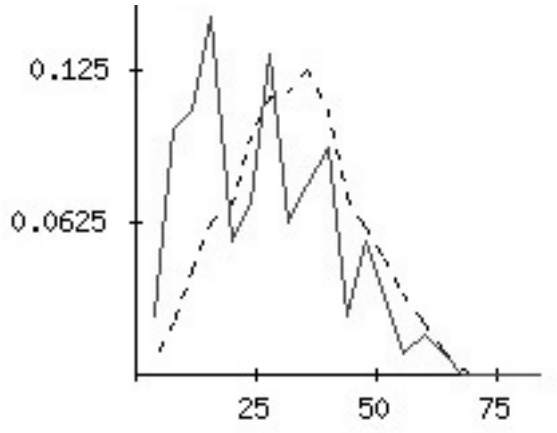


Figure 1: C-alpha pairwise distance distribution showing that pairs having short distances are more likely to coevolve. Coevolved pairs, solid line; all observed pairs, broken line. Vertical axis, frequency; horizontal axis, pairwise C-alpha distance in A°. Results shown are from burial preference segregation (results from other two segregations are similar).

4 References

[3] Felsenstein, J. 1989. PHYLIP—Phylogeny inference package (version 3.2). *Cladistics* 5:164-166.

[1] Fukami_Kobayashi, K., Schreiber, D.R., and Benner, S.A. 2002. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *Journal of Molecular Biology* 319:729-743.

[2] Pollock, D.D., Taylor, W.R., and Goldman, N. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of Molecular Biology* 287:187-198.

H27. A Markovian model of genome evolution: distribution of paralogs

Jerzy Tiuryn[1], **Damian Wójtowicz**[1]

Keywords: genome evolution, homologous genes, paralogs, Markov chain

1 Introduction.

Several sequences of genomes have been recently established. This creates an unprecedented chance of understanding a genomes' evolution and organization. The data allows us to study a genome not only as simply a set of genes, but also as a dynamic collection of genes which changes in time. Various biochemical and evolutionary processes constantly act on genomes and drive them to evolve dynamically.

Two genes are called *homologous* if they are of the same evolutionary origin. When two homologous genes belong to the same genome they are called *paralogs*. The homology means common evolutionary origin and as such, when applied to two particular genes, can be stated only as a hypothesis. Homologous genes should show some level of similarity as sequences. It is clearly conceivable that in case of distant homology (i.e. when divergance has occurred very long time ago) the similarity level may be too low to distinguish it from the random event. In our model the event of mutation represents an accumulated evolutionary divergence beyond which it is impossible to tell two genes being homologous (even though, as it follows from our definition of homology, they never cease being homologous).

In recent 5 years, many publications concentrate on quantitative comparative analysis of the frequency distributions of genes. Initially, these papers concentrate mainly on empirical analysis of the available genomes [4, 5, 1, 2]. Their authors claim different distributions of gene families sizes. Recently, Koonin's group has developed a simple model that based on a well known a *birth-and-death model* with an innovation process [3]. Unfortunately, one of the drawbacks of this model is that it sets apriori a bound on the family size.

In this note we propose a new model of genome evolution without the above mentioned limitation.

2 Model of genome evolution.

In order to express the concept of gene homology we will assume that all genes we are working with are colored. The convention will be that genes with the same color are *homologous* and genes of different colors are not homologous. We will assume that we have an unlimited supply of colors. A *genome* is a finite set of colored genes. A *gene family* in a genome is a set of all genes of that genome which have the same color. We group families according to their size. For any $i > 0$, let $\mathcal{C}_i$ denote the class of all i element families of the genome.

Evolution of genomes will be modeled by a Markov chain. States of the Markov chain are genomes. The transition from a genome $\mathcal{G}$ to $\mathcal{G}'$ is based on the following process of *evolution* which is performed independently for each gene of $\mathcal{G}$. Let p_R, p_M, p_D, p_U be non negative reals such that $p_R + p_M + p_D + p_U = 1$. A gene, which is subject to the process of evolution

[1]Institute of Informatics, Warsaw University, ul. Banacha 2, 02–097 Warsaw, Poland. E-mail: {tiuryn,dami}@mimuw.edu.pl This work was partly supported by the Polish KBN grant 7 T11F 016 21.

is: *removed* from the genome, *mutated* [2], *duplicated* [3] or *unchanged* with probabilities p_R, p_M, p_D and p_U, respectively. Since the processes of removal, mutation and duplication are independent of each other as the evolutionary events, it follows that we can assume that p_R, p_M, p_D are pairwise different. Moreover, it is natural to assume that $p_U >> p_R+p_M+p_D$.

3 Main results.

Let $C_i^{(n)}$ be the expected value at time n of the cardinality of the class $\mathcal{C}_i$ in the described Markov chain with the initial state being a one element genome. It follows that the long term behavior of this chain falls into one of the following two categories: eventually all genes disappear (when $p_R \geq p_D$) with probability 1, or the total population of genes expotentially increases to infinity (when $p_R < p_D$). Nevertheless, we are interested in studying the relative proportions between $C_i^{(n)}$ and $C_{i+1}^{(n)}$ with n turning to infinity. This coreresponds to the possibility of observing the distribution of paralogs in a genome which has evolved for a sufficiently long time.

Model with mutation. Using computer simulations, we were able to discover elementary properties of our model. These calculations show that our results are in accord with real biological data published in the mentioned papers.

Model without mutation. In this case, we have proved the following

Theorem *Let $p_M = 0$ and $p_R \neq p_D$. Then the sequence $(C_1^{(n)}, C_2^{(n)}, \ldots)$ leads neither to geometric nor to logarithmic distribution.*

However, when probabilities p_R and p_D are relatively small, computer simulations show that the observed distribution prefectly matches the geometric one. On the contrary, when these probabilities increase, then the geometric distribution does not fit the data at all.
An additional assumption $p_D < p_R$ also allows us to prove that this distribution does not lead to power law.

Remark. Our computer simulations are based on computing a generating function of the sequence $(C_1^{(n)}, C_2^{(n)}, \ldots)$, rather than on randomized simulations of the Markov chain. There is an analytic formula for this function.

References

[1] Huynen, M.A., van Nimwegen, E. ,,The Frequency Distribution of Gene Family Size in Complete Genomes." *Molecular Biology Evolution* 15(5), pp. 583–589, 1998.

[2] Jordan, K., Makarova, K.S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. ,,Lineage-Specific Gene Expansions in Bacterial and Archeal Genomes." *Genome Research* 11, pp. 555–565, 2001.

[3] Karev, G.P., Wolf, Y.I., Rzhetsky, A.Y., Berezovskaya, F.S., Koonin, E.V., "Mathematical Modeling of the Evolution of Domain Composition of Proteomes: A Birth-and-Death Process with Innovation.", In *Frontiers in Computational Genomics*, Caister Academic Press, 2003.

[4] Slonimski, P.P., Mosse, M.O., Golik, P., Henaût, A., Diaz, Y., Risler, J.L., Comet, J.P., Aude, J.C., Wozniak, A., Glemet, E., Codani, J.J. "The first laws of genomics." *Microbial and Comparative Genomics* 3:46, 1998.

[5] Slonimski, P.P., "Comparision of complete genomes: Organization and evolution." *Proceedings of the Third Annual Conference on Computational Molecular Biology (RECOMB'99)*, Stanislaw Ulam Memorial Lecture, pp. 310, ACM Press, 1999.

[2] It changes its color to a new color which is not present in the genome.
[3] An new gene is created in the genome and the new gene inherits the color of the older one.

H28. Retention of functionality in protein recombinants

Yanlong O. Xu[1], Randall W. Hall[2], Richard A. Goldstein[3], David D. Pollock[4]

Keywords: protein, lattice models, divergence, recombination, evolution

1 Introduction.

Recombination of divergent proteins is an important means of generating diversity and developing functional innovation in both natural evolution and protein engineering. As proteins diverge, there is an inherent tradeoff between the probability of creating a useful innovation and the probability that the recombinant protein no longer functions properly, but the dynamics of this tradeoff are largely unknown. Here, we use model proteins to explore how divergence alters function in recombinants protein in detail.

2 Model and Methods.

We modeled proteins as compact structures on a two-dimensional lattice, with folding energies obtained from Miyazawa-Jernigan amino acid potentials. Sequences were of length 25 and they could fold into 1081 compact 5x5 structures. For a given sequence the potential energy of folding to a specific structure (or fold), can be given by the summation of potentials from all interaction pairs; thus, the equilibrium probability that a protein sequence will fold to confirmation can be given by Boltzmann statistic. We model evolution in constant-size haploid populations of 1000 individuals with mutation rates of 0.05 mutations per proteins per generation. Individual fitness is based primarily on the probability of folding into the "native" or functional structures, f_N, which are usually pre-specified in our simulation. The populations were initialized with ten random sequences at equal frequency and were allowed to evolve to equilibrium. At equilibrium, the populations were duplicated and each population was then allowed to evolve independently under identical conditions. Samples were taken every 500 steps (sampling intervals). At each sampling point, the most frequent sequences in each population were recombined at all 24 possible sites (thus 48 recombinants produced) and probabilities of folding into the native structures and all alternative structures were evaluated at each recombination site. We classify structures based on the designability (Ds) which is defined as the fraction of random sequences that will fold to the structure with a probability of at 98%. We sorted structures into high(C_H), medium(C_M) and low(C_L) designable groups. There are 10 structures in C_H group with Ds≥1%. The C_M group consists of 24 randomly sampled structures with 0.1%≤Ds<1% , and C_L group consists of 32 randomly sampled structures with Ds<0.1%. The structural occupancy in evolution is defined as the average probability of protein fold to that structure over recombinants at each sampling point and over evolution time.

3 Results and Figures.

Although highly designable structures have, by definition, many more sequences that will fold into their structure, it is not designability but the sharing of contact pairs that determines the degree to

[1] Dept. of Chemistry, Louisiana State University, Baton Rouge, LA, USA. Email:yxu3@lsu.edu
[2] Dept. of Chemistry, Louisiana State University, Baton Rouge, LA, USA. Email:rhall@lsu.edu
[3] National Institute for Medical Research, Mill Hill, London, rgoldst@nimr.mrc.ac.uk
[4] Dept. of Biology, Department of Biological Sciences, and Biological Computation and Visualization Center, Email: dpollock@lsu.edu

which mutants will partially fold into an alternative structure in recombinants (figure1). The occupancy of alternative structures in the recombinants is consistently about seven times higher in the low-designable structures than in the medium- and highly designable structures. The retention of the structure in the recombinants is much worse for the low designable structure than for the high or medium designable structure in the evolution (figure2). The location of the least foldable recombinants is very well predicted by the number of interactions that are disrupted in the recombination, As the crossover position gets closer to the most disruptive point, the recombinants have lower and lower fitness, For the high and medium designable structures this trend is slight, but for the low designable structures the trend is steeper, and the recombinants become dramatically less fit when they are very close to the most disruptive point (figure3). Our results, based on a simple but effective model, show that the designability of the native structure and crossover site have a dramatic effect on the fitness of recombinants in the evolution; The contact pairs overlap determines which alternative structure will arise in the mutants.

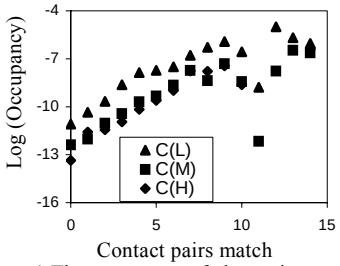


Figure1.The occupancy of alternative structure is correlated with its contact pairs overlap with native structure

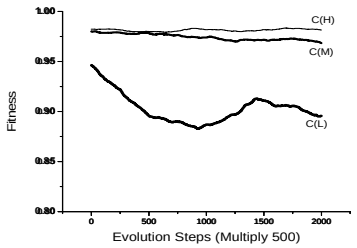


Figure2. Fitness in the recombinants keeps high in high designable structure and get much lower in the low designable structures along the evolution.

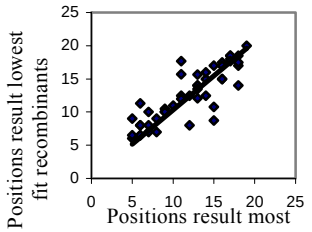


Figure 3(a)

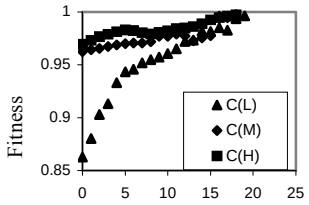


Figure 3(b)

Figure3. (a) Lowest fit recombinants correspond to the crossover positions that disrupt the contact pairs most. (b) Fitness decrease when crossover position gets close to the position that disrupts contact pairs most, the trend is clear especially for low designable structures..

4 References and bibliography.

[1]Y. Cui, 2002, Recombinatoric exploration of novel folded structures, *Proceedings of the National Academy of Sceicnes*, *USA*, 99: 809-14

[2]C A.Voigt, 2002, Protein building blocks preserved by recombination, *Nature Structural Biology*, 9(7): 553-8.

[3] P D. Williams, D D. Pollock, R A. Goldstein, 2001, Evolution in Functionality in lattice proteins, Journal of Molecular Graphic and Modeling, 19(1):157-167.

H29. Fitting nonreversible substitution processes to multiple alignments

Von Bing Yap [1]

Keywords: DNA base substitution, maximum likelihood, EM algorithm

Consider a DNA base substitution model on a rooted phylogenetic tree with initial distribution x and rate matrix Q. This process is stationary if x is the equilibrium distribution of Q, i.e., $xQ = 0$, and it is reversible if in addition, the detailed balance condition is satisfied:

$$XQ = Q'X,$$

where Q' denotes the transpose of Q and X is the diagonal matrix with x as its diagonal. Thus, this model allows base composition to evolve and is more general than the reversible models used in routine phylogenetic analysis. Given a multiple alignment related by a rooted tree, the EM algorithm introduced by Holmes and Rubin for fitting reversible models [2] turns out to be an efficient tool for fitting the present nonreversible model. This algorithm is applied to two datasets described by Yang [3]: six primate $\psi\eta$ globin pseudogenes, and mitochondrial DNA from nine primates. The estimate of Q from the pseudogenes is similar to previous estimates by Yang and by Arvestad and Bruno [1]. However, on the mtDNA, the new estimate is very different from Yang's. Generally, the new estimates fit the data better in the sense that the predicted base compositions are closer to the observed base compositions.

If it is known that the root lies on a particular branch, then its distance from a reference node can also be estimated jointly with x and Q, by maximum likelihood. On the pseudogene dataset, the most likely root position is at orangutan, followed very closely by spider monkey, the latter being intuitively more sensible. On the mtDNA dataset, the most likely root position is at the node connecting orangutan to the others. Another local maximum is somewhere on the branch conneting the ancestors of gibbon and crab-eating macaque.

References

[1] Arvestad, L. and Bruno, W. J. 1997. Estimation of reversible substitution matrices from multiple pairs of sequences. In: *Journal of Molecular Evolution* 45:696–703.

[2] Holmes, I. and Rubin, G. H. 2002. An expectation maximization algorithm for training hidden substitution models. In: *Journal of Molecular Biology* 317:753–764.

[3] Yang, Z. 1994. Estimating the pattern of nucleotide substitution. In: *Journal of Molecular Evolution* 39:105–111.

[1] Department of Mathematics, University of California. E-mail: vonbing@math.berkeley.edu

H30. A Web of Prokaryotic Life

Stefan R. Henz,[1] Daniel H. Huson,[2] Alexander F. Auch,[2] Vincent Moulton, [3] and Stephan C. Schuster[1]

Keywords: molecular evolution, tree of life, web of life, prokaryotes, consensus methods

1 Introduction.

The evolution of species is generally believed to be a *branching process* [3] and, thus, the ultimate goal of phylogenetic analysis is to compute the "Tree of Life". However, there are a number of evolutionary mechanisms such as hybridization, recombination or swapping of genes, that may imply that evolutionary history is perhaps more accurately described by a "Web of Life" [2], that is, by a network that generalizes the concept of a phylogenetic tree [1, 7, 6].

Traditionally, in molecular phylogeny, trees are built from singular phylogenetic markers such as 16S rRNA [8]. In Figure 1(a) we show such a tree for 91 prokaryotes. More recently, given the rapidly rising abundance of whole genome sequences, it has become possible to base such phylogenies on many different molecular markers, such as different shared genes, gene-order or gene-content, for example. In Figure 1(b) we show a phylogenetic "consensus network" [5] for the same 91 prokaryotes based on their 30 most conserved genes [4]. This graph shows every split that occurs in two or more of the 30 trees. By increasing this threshold, one obtains increasingly less cluttered graphs (not shown here). This phylogenetic network gives a good indication of which parts of the phylogeny shown in Figure 1(a) are supported by the 30 genes, and which parts potentially have multiple evolutionary histories.

References

[1] H.-J. Bandelt and A.W.M. Dress. A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92:47–105, 1992.

[2] F. W. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2128, 1999.

[3] Ernst Haeckel. Monophyletischer Stammbaum der Organismen, 1866.

[4] S. R. Henz, D. H. Huson, A.F. Auch, and S. C. Schuster. Computing the web of prokarayotic life. In preparation, 2004.

[5] B. Holland and V. Moulton. Consensus networks: A method for visualizing incompatibilities in collections of trees. In G. Benson and R. Page, editors, *Proceedings of "Workshop on Algorithms in Bioinformatics"*, volume 2812 of *LNBI*, pages 165–176. Springer, 2003.

[6] D. H. Huson and D. Bryant. jSplits - a framework for phylogenetic trees and networks. `http://www-ab.informatik.uni-tuebingen.de/software/jsplits/`, 2004.

[7] D.H. Huson. SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics*, 14(10):68–73, 1998.

[8] C. R. Woese. Bacterial evolution. *Microbiol. Rev.*, 51:221–272, 1987.

[1]Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany.
E-mail: `stefan.henz@tuebingen.mpg.de`, `stephan.schuster@tuebingen.mpg.de`
[2]Center for Bioinformatics, Tübingen University, Sand 14, 72076 Tübingen, Germany.
E-mail: `huson@informatik.uni-tuebingen.de`, `auch@informatik.uni-tuebingen.de`
[3]The Linnaeus Centre for Bioinformatics, Uppsala University, Sweden.
E-mail: `vincent.moulton@lcb.uu.se`

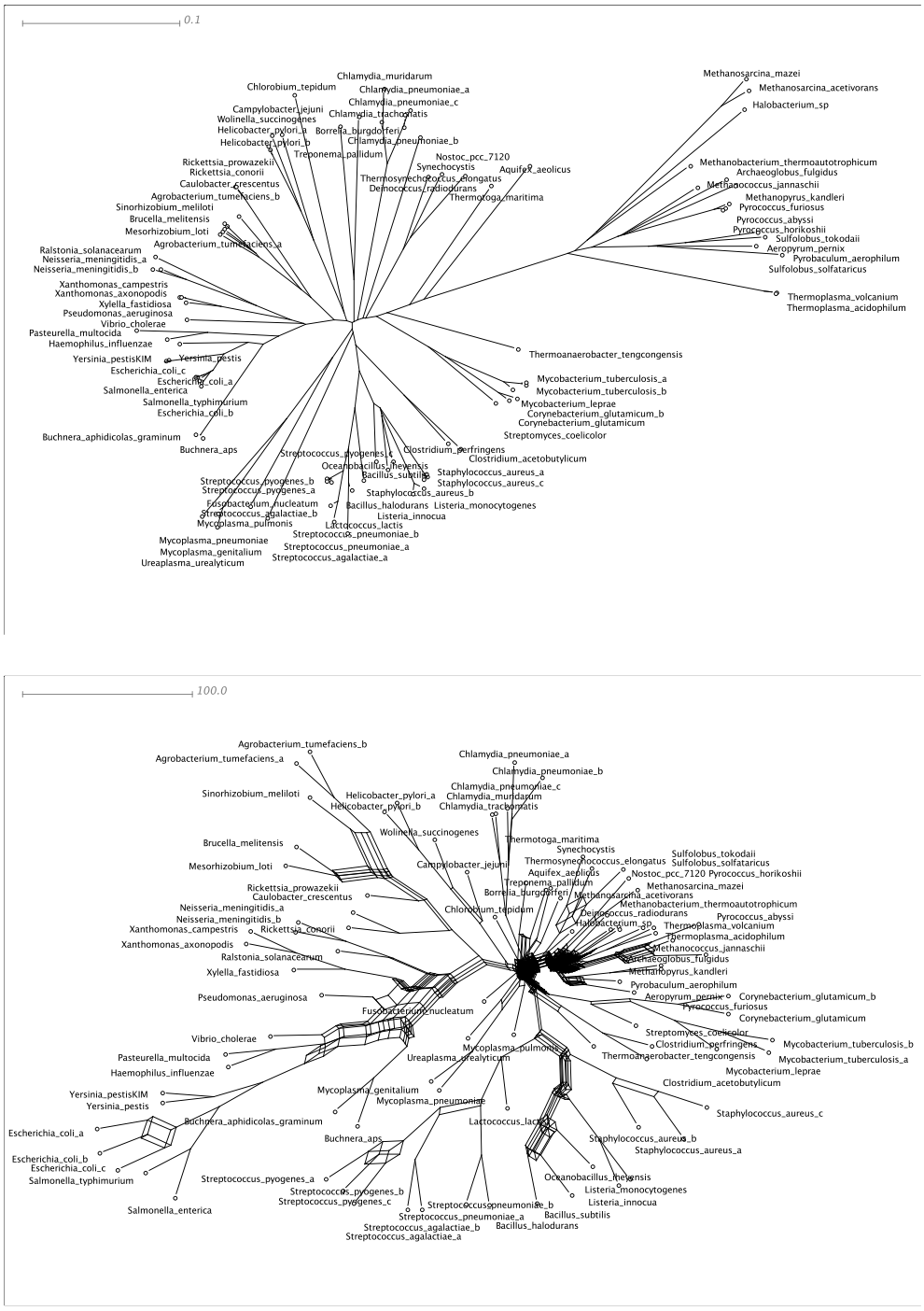


Figure 1: (a) A tree of prokaryotic life based on 16S rRNA. (b) A "web" of prokaryotic life based on the 30 most conserved genes.

I1. Computational Target Discovery: Using HMMs to Identify the Druggable Proteome

Joanne I. Adamkewicz[1], R. Glenn Hammonds[2]

Keywords: Target discovery, HMM, Pfam, protein classification, enzymes, drugs

1 Abstract.

We present a method for the classification of proteins for the purpose of target discovery. Using the functional hierarchy of the Enzyme Commission (EC) classification, with extensions where required, we classified over 9000 publicly available Hidden Markov Models (HMMs) representing evolutionarily conserved protein domains, noting catalytic activity and other properties. We use 4 fungal genomes with differing levels of annotation to illustrate how the method can rapidly identify all potentially "druggable" genes in a genome of interest, using only publicly available software and data. The model collection currently covers 88% of the 52475 Swiss-Prot protein entries with EC assignments. Those enzymatic proteins not hit by the collection will direct building of custom HMMs.

2 Introduction.

Modern target discovery has become a process of sifting wheat from chaff. With full genome sequence available for many organisms, the number of possible target genes is in the thousands or tens of thousands for each species of interest. Validating potential targets via biological experiments is labor-intensive and time-consuming, yet if the product of a gene is not amenable to modulation by drug therapy, prior biological validation work is wasted. As a first step in the drug discovery process, then, it is highly cost-effective to restrict your focus to the fraction of an organism's genes that are considered druggable according to criteria defined for each specific project. Ideally, the filtering process would be rapid and automated; applicable to annotated, unannotated, and even partial genomes; flexible (so each user can define their own criteria for selecting desired targets); use only publicly available software and data; and be applicable to any species of interest.

We present here a curation of publicly available models (or alignments that can be made into models), which has proven useful in target selection and practical to implement using largely off the shelf software. The essence of our curation is a classification of the models into categories of interest to our customers: biologists at biotech or pharmaceutical companies. This classification allows biologists to quickly select candidate targets for a wide variety of projects based on user-defined criteria.

[1] Exelixis, Inc., South San Francisco, CA, USA. E-mail: jadamkew@exelixis.com
[2] Exelixis, Inc., South San Francisco, CA, USA. E-mail: rghammonds@earthlink.net

3 Methods and Results.

Models or alignments were obtained by download from Pfam[1], Interpro[2] for SMART [3] models, TIGR [4], NCBI for COG [5] alignments, and Affymetrix for GPCR models [6]. In addition, custom alignments were built as desired to cover additional protein families. Alignments were converted to models where necessary using either hmmer or SAM, with later conversion of all models to hmmer format. The resulting model collection was curated in collaboration with biologists via automated text classification followed by manual inspection.

We estimate the fraction of known enzymes covered by the current model collection from the fraction of SwissProt sequences annotated with an EC number retrieved by any enzymatic model in the collection: 88% (E value <= 1).

Given the curated collection, the models are then run against a proteome of interest using the hmmpfam or hmmsearch algorithms. The raw output files are processed and stored in a database for easier searching and downstream analysis. Each sequence hit by one or more domains then inherits the classification of those domain hits, according to priorities as set by the user. For newly-sequenced genomes where no protein predictions are available, gene prediction tools such as Orfinder can be used in combination with HMMs calibrated to find partial matches to a domain.

To illustrate the utility of the curated model collection, we analyzed 4 fungal genomes in various states of sequencing and gene annotation: Saccharomyces cerevisae, Schizosaccharomyces pombe, Ustilago maydis, and Phanerochaete chrysosporium. We present comparative results for classes of particular interest, including kinases, proteases, and glycosyltransferases, and show that the method can be used successfully with an unannotated genome by adding an ORF-finding step.

References

[1] Bateman, A. *et al*. 2004 The Pfam protein families database. *Nucleic Acids Research* 32: D138-D141

[4] Haft DH, Selengut JD, White O. 2003 The TIGRFAMs database of protein families. *Nucleic Acids Research* 31:371-3.

[3] Letunic, I., Copley, R.R., Schmidt, S., *et al*. 2004 SMART 4.0: towards genomic data integration. *Nucleic Acids Research* 32: D142-D144

[2] Mulder N.J., Apweiler R., Attwood T.K., *et al*. 2003 The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research* 31:315-318.

[6] Shigeta R., Cline M., Liu G., and Siani-Rose MA. 2003 GPCR-GRAPA-LIB-a refined library of hidden Markov Models for annotating GPCRs *Bioinformatics* 19:667-668.

[5] Tatusov RL, Natale DA, Garkavtsev IV, *et al*. 2001 The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research* 29: 22-8.

I2. Phylogenetic Relationship of Sessile Barnacles Based on Mitochondrial DNA

Rowshan Ara Begum[1], Toshiyuki Yamaguch[2]. Shugo Watabe[1]

Keywords: 12S rRNA, 16S rRNA, sessile barnacle, *Chthamalus, Megabalanus*, *Tetraclita*, molecular phylogeny

1 Introduction.

Barnacle, a sessile organism abundantly distributed in the intertidal area, occupies an important phylogenetical position in Crustacea due to diverged speciation. Previous attempts by several groups could not unequivocally establish a phylogenic relationship within closely related barnacles because of inconsistency in phylogenetic data derived from nuclear DNA [1] in comparison with those from morphology [2]. To resolve such dispute over the phylogeny of barnacle, especially for sessile barnacle, the relationship was reinvestigated based on mitochondrial DNA (mtDNA).

2 Materials and Methods.

Three species of sessile barnacle, *Chthamalus challengeri*, *Tetraclita japonica* and *Megabalanus volcano*, and one goose barnacle, *Capitulum mitella*, were collected from the coastal region of the Pacific ocean near Miura Peninsula, Kanagawa Prefecture, Japan and the muscle tissues were subjected to DNA extraction. PCR was employed to amplify the partial nucleotide sequence of the 12S and 16S rRNA genes of mtDNA.

3 Results and Discussion.

Amplified DNA fragments encoding partial lengths of the 12S rRNA and 16S rRNA genes contained 350 and 450 bp, respectively (Figure 1). Both neighbor joining and maximum parsimony analyses based on 12S rRNA and 16S rRNA separately and on combined data produced a monophyly of the three sessile barnacle species from different genera while the goose barnacle formed a paraphylic group with the three sessile barnacles. However, based on the12S rRNA gene alone, the two sessile barnacle species, *M. volcano* and *C. challengeri*, formed a different group than *T. japonica*. On the other hand, based on the 16S rRNA gene and combined data of the 12S and 16S rRNA genes, however, *M. volcano* and *T. japonica* appeared to be the closest relatives among the three sessile barnacle species (Figure 2). These results were congruent with those based on larval characters [3] and the nuclear 18S rRNA gene [1] reported previously, but differed from morphological classification [2]. Thus, the 16S rRNA gene may be considered to be more reliable than the 12S rRNA gene to investigate the phylogenetic relationship at the generic level within sessile barnacle species as far as the present three genera are concerned.

[1] Department of Aquatic Bioscience, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo, Tokyo 113-8657, Japan. E-mail: awatabe@mail.ecc.u-tokyo.ac.jp

[2] Marine Biosystems Research Center, Chiba University, Inage, Chiba 263-8522, Japan. E-mail: tyamaguc@msi.biglobe.ne.jp

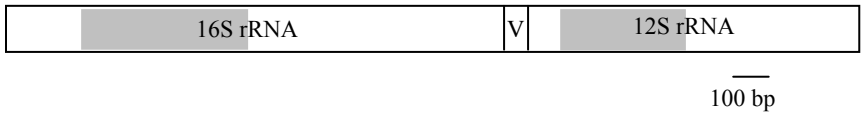


Figure 1: A schematic diagram of a part of the mitochondrial DNA showing the amplified regions. Shaded areas represent the amplified region and 'V' indicates $tRNA^{Val}$.

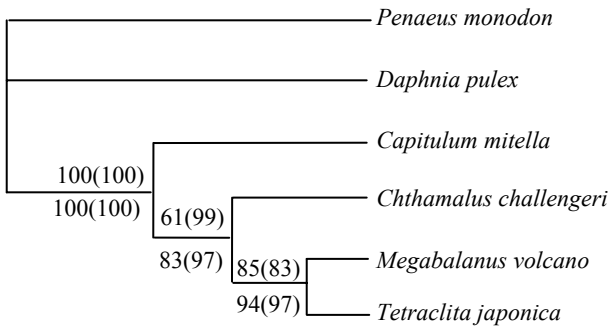


Figure 2: Molecular phylogenic trees for *Capitulum mitella*, *Chthamalus challengeri*, *Megabalanus volcano* and *Tetraclita japonica* according to the neighbor joining (NJ) and maximum parsimony (MP) methods based on the partial nucleotide sequences of the 16S rRNA gene and based on the combined data of the 12S rRNA and 16S rRNA genes. Identical phylogeny was obtained for the 16S rRNA alone and combined data. Numbers above the branches indicate bootstrap values from 1000 replicates using the NJ method based on the 16S rRNA gene only, whereas the numbers of the parentheses indicate the bootstrap values according to the MP method. The bootstrap values of the NJ and MP methods of combined data are shown below the branches. The nucleotide sequences of *Daphnia pulex* and *Penaeus monodon* were cited from the GenBank database with accession numbers NC000844 and NC002184, respectively, in order to root the molecular phylogenic tree.

References

[1] Spears, T., Abele, G. L., and Applegate, M. A. 1994. Phylogenetic study of ciripedes and selected relatives (Thecostraca) based on 18S rDNA sequence analysis. *Journal of Crustacean Biology* 14:641-656.

[2] Nishimura, S. 1995. *Guide to Seashore Animals of Japan with Color Pictures and Keys, II*. Hoikusha Publishing Co. Ltd. pp. 42-133.

[3] Newman, A. W. and Ross, A. 2001. Prospectus on larval cirriped setation formulae, revisited. *Journal of Crustacean Biology* 21:56-77.

I4. Resequencing the Human Genome using Short Sequence Fragments

Anthony J. Cox[1], Lisa J. Davies[2], Clive G. Brown[3]

Keywords: Whole genome resequencing, single molecule array, sequence alignment.

1 Introduction

Now that the sequence of the human genome is essentially complete, attention has turned towards the challenge of characterizing the variation in the genome across the human population, as exemplified by initiatives such as the HapMap project [3].

Solexa's Single Molecule Array platform will allow the massively parallel sequencing of millions of short DNA fragments and other groups are seeking to provide similar data by different means. It is therefore a pertinent time to consider what information can be gleaned about an individual human genome from short (20−30 bp) fragments of its sequence. We describe a method of predicting whether sequence fragments of a given length can be used to detect single base variation at any given point in the human genome.

2 Discussion

Sequence fragments of say 20−30 bases in length are not likely to permit a *de novo* assembly of the human genome (although some assembly of contigs may be possible), but it is anyway more sensible to take an approach that makes use of the high quality reference sequence that is already available. Instead, then, one can take each fragment and try to find its most likely origin in the genome, a task that is possible only if the fragment has a unique "best match" in the genome. Uniqueness is also a prerequisite for the successful design of oligonucleotide probes for microarrays, a consideration that motivated a recent paper by Healy *et al* [2], which described a method of annotating the human genome using counts of exact occurrences of fragments, together with an efficient algorithm for the substantial amount of computation involved.

A similar annotation method would suffice for us to decide whether a fragment drawn from a given position in the reference genome could be uniquely mapped back to that spot. However, in practice both mutations in the genome being sequenced and errors caused by the sequencing process will manifest themselves as differences between the fragment and its originating position in the genome. We would like to know how these differences will affect our ability to assign the fragment to its correct position. For this, we need an annotation method that incorporates information about *inexact* copies of fragments within the genome.

[1] Solexa Limited, Chesterford Research Park, Little Chesterford, Essex CB10 1XL, United Kingdom. anthony.cox@solexa.com

[2] lisa.davies@solexa.com

[3] clive.brown@solexa.com

We can think of two fragments as being "neighbours" if they differ only by a few bases and so any fragment will have a "neighbourhood" of similar fragments scattered throughout the genome. If its neighbourhood is sparsely populated, then errors or mutations are less likely to cause a sequenced version of the fragment to be erroneously mapped to one of its neighbours instead.

We show [1] that the neighbourhoods of the set of fragments that overlap a given base position in the genome can provide an accurate forecast of our ability to detect a single base variation at that base position using short sequence fragments of a given length. A new algorithm has enabled us to surmount the computational challenge of computing the neighbourhood of every fragment in the human genome in a reasonable time on modest hardware. This data has been used to annotate every base in the human genome with our prediction of our ability to detect a single base variation at that point using fragments of 20, 25 and 30 bases in length. We discuss our ability to detect variation in loci and other regions of particular interest within the human genome.

References

[1] Cox, A.J. ,Davies, L.J. and Brown, C.G. Resequencing the human genome using short sequence fragments. In preparation.
[2] Healy, J., Thomas, E.T., Schwartz, J.T. and Wigler, M 2003. Annotating large genomes with exact word matches. *Genome Research* 13:2306−2315.
[3] International HapMap Consortium 2003. The International HapMap Project. *Nature* 426:789−796

I5. Detecting correlated amino acid substitutions using Bayesian phylogenetic techniques

Matthew W. Dimmic, [1] **Melissa J. Todd,** [2] **Carlos D. Bustamante,** [3] **Rasmus Nielsen** [4]

Keywords: protein coevolution, mutational mapping, MCMC, likelihood models

1 Introduction

The evolution of protein sequences is constrained by complex interactions between amino acid residues. Because harmful substitutions may be compensated by other substitutions elsewhere in the protein, residues can co-evolve. Identification of these coevolving sites can aid in the prediction of binding sites between protein domains, as well as providing insight regarding the process of protein evolution and adaptation.

We describe a Bayesian approach and MCMC computational techniques to detect correlated substitutions in protein families. This method minimizes errors due to phylogenetic correlations by exploiting the information in the evolutionary tree, yet it does not require prior knowledge of the true tree topology. Variance in the estimates of branch lengths and mutation rate are also accounted for, and posterior predictive distributions are used to determine the significance of each putative correlation. The sensitivity of several different test statistics are assessed under varying evolutionary conditions using simulated datasets.

2 Methods

The method is based upon the notion that coevolving sites will be more likely to substitute along the same branch of the evolutionary tree, and that these co-substitutions can be inferred using ancestral reconstruction (e.g. [1]). In some cases, however, the true tree may be difficult to determine, and variation in evolutionary rate and branch length can bias the results [5] and mislead the parsimony reconstruction.

To account for these issues, we use the MCMC approach in MrBayes [2] to draw from the posterior distribution of tree topologies, evolutionary distances, and GTR model parameters. Customized software has been written to assign rates to each site for each posterior draw, and a Bayesian mutational mapping technique [3] (BMM) is then used to assign mutations to branches of the tree for each site pair of interest. Once these mutational maps have been calculated, a set of test statistics are used to detect whether coevolution is occurring at each site pair. To assess the significance of each test, posterior predictive P-values are calculated using MCMC simulation [4].

To assess the power and sensitivity of the BMM method under various conditions, it was evaluated on codon sequences simulated under a model of correlated evolution. The substitution rate for an i, j amino acid substitution at site A given amino acid k at site B is

$$q_{ij} = \mu\omega_A g_{ij} e^{\rho_{AB}(\psi_{jk} - \psi_{ik})}$$

[1]Biological Statistics and Computational Biology Department, Cornell University, Ithaca, New York. E-mail: mdimmic@umich.edu

[2]E-mail: melissa@mail.bscb.cornell.edu

[3]E-mail: cdb28@cornell.edu

[4]E-mail: rn28@cornell.edu

The first portion is similar to previous codon models: μ is the nucleotide substitution rate, ω_A is the selection parameter for making any nonsynonymous change at site A, and g_{ij} is the rate of nucleotide substitution for the nucleotide change required to go from amino acid i to j. The exponential term introduces the correlation: ρ_{AB} is the strength of correlation between amino acid sites A and B, and ψ_{ik} and ψ_{jk} is the favorability or unfavorability of interaction between the amino acid pairs i, k and j, k respectively. In the case where $\rho_{AB} = 0$, the sites become independent.

3 Results

With the correct choice of test statistic, the Bayesian mutational mapping (BMM) approach can provide significant improvement over methods which do not account for uncertainty in topology or parameter estimates. For example, on a 32-sequence tree with moderate levels of simulated correlation and evolutionary rate, the best test statistic (Z) detected 70% of the truly correlated sites, with no false positives (see Table 1). As the strength of the correlation is decreased the sensitivity also decreases, but the number of false positives remains low. The BMM method can reliably detect coevolving sites in difficult phylogenetic situations, such as when branch lengths are long (i.e., the sequences are distantly related with few close homologs) or when there is extreme rate variation. Preliminary results on real datasets indicate that the BMM method can detect weak levels of correlated evolution at protein domain interfaces.

	Z		LR		c_+		S_-		T_{corr}	
μ	.05	.01	.05	.01	.05	.01	.05	.01	.05	.01
0.10	39%	5%	16%	5%	32%	2%	38%	2%	10%	2%
0.25	70%	25%	45%	19%	46%	24%	40%	18%	17%	4%
0.50	88%	71%	83%	59%	61%	36%	32%	14%	33%	20%
0.75	100%	90%	94%	80%	91%	61%	36%	21%	42%	21%

Table 1: Sensitivity (TP/(TP+FN)) of some test statistics used in the BMM method at the $\alpha = 0.05$ and $\alpha = 0.01$ cutoffs for the test statistic's P-value. Datasets were simulated at $\rho = 1$ with different evolutionary rates μ, which is the number of expected nucleotide changes per codon site per branch under no correlation. The first three test statistics use a likelihood model, while the last two are model-independent. False positive rates for uncorrelated sites are not shown, but in all cases are 2% or less.

References

[1] Fukami-Kobayashi, K, Schreiber, D.R., and Benner, S.A. 2002. Detecting compensatory covariation signals in protein evolution using reconstructed ancestral sequences. *J. Mol. Biol.* 319(3):729-743.

[2] Ronquist, F. and Huelsenbeck, J. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19(12):1572-1574.

[3] Nielsen, R. 2002. Mapping mutations on phylogenies. *Systematic Biology* 51(5):729-239.

[4] Nielsen, R. and Huelsenbeck, J. 2002. Detecting positively selected amino acid sites using posterior predictive P-values. *Pacific Symposium on Biocomputing* 2002:576-588.

[5] Tufféry, P. and Darlu, P. 2000. Exploring a phylogenetic approach for the detection of correlated substitutions in proteins. *Mol Biol Evol* 17(11):1753-1759.

I6. Scientific Workflows for High Resolution Genetic Sequence Analysis

Luke Ulrich[1], Elizabeth Marland Glass[1], Mark D'Souza[1], Praveen Chandramohan[1], Natalia Maltsev[1]

Keywords: phylogeny, scientific workflow, HMM, functional motifs, classification

Modification of proteins in the course of evolution leads to the emergence of new specificities, altered recognition properties, or changes in their biological functions. Extreme proliferation of some families within an organism, perhaps at the expense of other families, may correspond to functional innovations during evolution [1]. Phylogenetic analysis for inferring relationships among genes and reconstructing evolutionary events provide a powerful way to interpret an increasing body of sequence data and enables investigation into the mechanisms that have led to the establishment of particular biological functions (e.g., convergent evolution, evolution by enzyme recruitment) [2]. However, the level of resolution of existing genetic sequence analysis tools sometimes is not sufficient for high-resolution genetic sequence analysis.

To assist such analyses we have developed the following tools for analysis of protein families: 1. PhyloBlocks (http://compbio.mcs.anl.gov/ulrich/phyloblock) is an interactive tool that allows an expert to analyze protein functions in a framework of phylogenetic information and to develop high-resolution HMM profiles for particular implementation of protein function interactively and 2. SVMmer (http://compbio.mcs.anl.gov/svmmer) (in collaboration with N. Samatova, ORNL), is a tool for classification of protein families using a support vector machines (SVM) algorithm.

Analysis of protein families using PhyloBlocks involves several steps. First, sets of protein sequences corresponding to a particular family (e.g. COGS [3], HobaGen [4], WIT3 [5], or a set of homologs from Blast) are aligned by using CLUSTALW [6]. Then the resulting CLUSTALW tree is presented to the user. The user can choose particular subsets of sequences (for example, sequences corresponding to particular protein function or particular version of enzyme) and submit them to Blockmaker [7] for the development of specific hidden Markov models-based (HMM-based) Blocks profiles. Resulting profiles can be saved in PhyloBlocks database and later used for characterization and annotation of proteins. PhyloBlocks also allow users to compare HMM profiles for different families and to identify features that are common or variable between the families. PhyloBlocks allows for increased efficiency and precision of high-throughput automated genetic sequence analysis. Such increased resolution and accuracy of predictions is essential for the development of metabolic reconstructions and flux analysis of metabolic networks. It also allows for the identification and characterization of different implementations of functions characteristic of a species or taxonomic group or of a specific environmental niche (i.e. hypothermophilic, psychrophylic, high salinity environments, etc). Resultant specific HMM profiles may be used for the development of the SVM models in SVMer for automated classification of proteins in newly sequenced genomes as well as for the development of relevant PCR primers for experimental identification and characterization of genes of interest. The poster will present the results of analysis of evolution of the aminotransferases II family using PhyloBlocks [8] and SVMMER [9].

[1] Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Ave., Argonne, Illionis, US. E-mail: ulrich@mcs.anl.gov, marland@mcs.anl.gov, dsouza@mcs.anl.gov, mohan@mcs.anl.gov, maltsev@mcs.anl.gov

References

[1] Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., and Hood, L. 1997. Gene families: the taxonomy of proten paralogs and chimeras. Science, 278: 609-614.

[2] Thornton, J.W. and DeSalle, R. 2000. Gene family evolution and homology: genomics meets phylogenetics. *Annual Rev Genomics Hum Genet. 1*:41-73.

[3] Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J., Natale, D.A., McLysaght, A.., Seoighe, C. and Wolfe, K.H. 2000. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* 4:41.

[4] Perriere, G., Duret, L. and Gouy, M. 2000. HOBACGEN: database system for comparative genomics in bacteria. *Genome Res*., 10:379-385.

[5] http://compbio.mcs.anl.gov/wit3

[6] Higgins, D., Thompson, J., Gibson, T., Thompson, J.D., Higgins, D.G., Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., 22:4673-4680.

[7] http://blocks.fhcrc.org/blocks/make_blocks.html

[8] http://compbio.mcs.anl.gov/ulrich/phyloblocks

[9] http://compbio.mcs.anl.gov/svmmer

I7. Anchors: Pre-Classification and its Effects on Hidden Markov Models

Jeremy Fisher[1] and Alan Sprague[2]

Keywords: Hidden Markov Models, Data Mining

1 Introduction.

Hidden Markov Models have been prominent in the categorization of biological data in the past decade. Hidden Markov Models (HMM) make predictions about an observable sequence from a finite alphabet.

Genemark and Genscan, two popular tools used to identify regions in DNA sequence, incorporate Hidden Markov Models into their design. A large interest in these programs is to classify sequences of DNA that are composed of the small alphabet of four nucleotides. Part of the classification of these sequences is to locate introns and exons. While performing admirably, there seems to be a cap of the performance of these programs [1].

Verbumculus is a program that analyzes substrings within a sequence. It can be used on DNA sequences to try to find under-represented and over-represented substrings within that sequence. It calculates the expected value and variances of all substrings of length n in O (n^2) worst case and O(n log n) expected time. From this you can give a score to the substrings [2].

A previous experiment constructed a HMM to classify the languages of English and Spanish based on consonant and vowel rhythms, the rationale being that insights might be applied to the classification of introns and exons, and eventually other parts of the gene such as promoter regions. The experiment yielded promising results at low language transition probabilities [3].

We apply the concept of under-represented and over-represented substrings to find unusual substrings to pre-classify a small portion of an observable sequence at a high accuracy. Once pre-classified, the sequence is run through the HMM and the pre-classified parts are used as 'anchors' that alter the probabilities of the HMM to suggest a specific (pre-determined) classification. We revisited the previous experiment classifying English and Spanish with this pre-classification to investigate the results.

The addition of the pre-classified 'anchors' in a sequence increased the accuracy of the HMM. Figure 1 shows the comparison of the unaided classification with that of the pre-classified sequence. Languages were run through the HMM at both frequent (10%) and infrequent (as low as 0.1%) language transition probabilities. The sequence was pre-classified at 0.3%, 0.6%, 1.2% and 3.0 % of total sequence.

These promising results indicate that pre-classification might increase accuracy of existing classification models which incorporate HMM.

[1] University of Alabama at Birmingham. E-mail: fisherje@cis.uab.edu
[2] University of Alabama at Birmingham. E-mail: sprague@cis.uab.edu

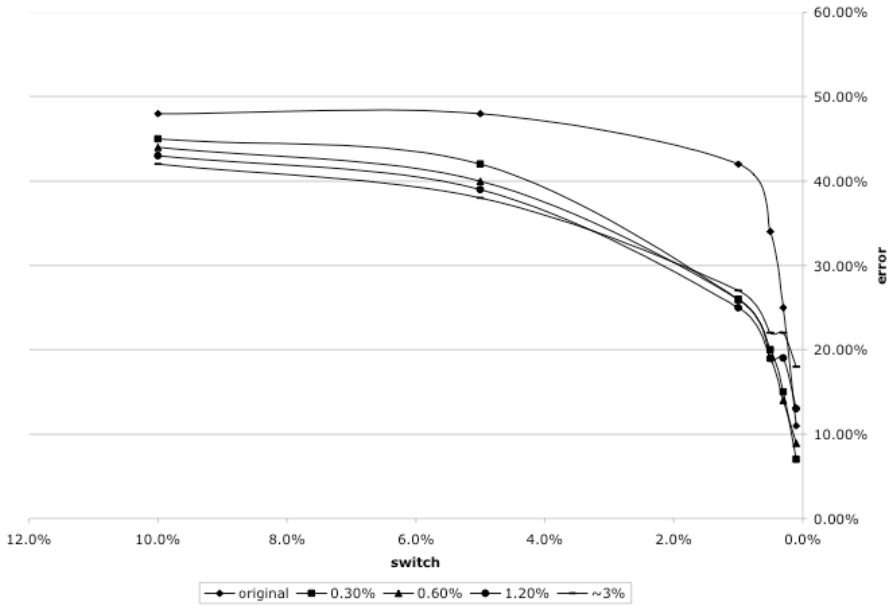


Figure 1: Alignment error of HMM given at 0%, 0.3%, 0.6%, 1.2% and 3.0% pre-classfication. Horizontal axis depicts the language transition probabilities. Vertical axis denotes the alignment error.

References

[2] Apostolico, et al. "Efficient Detection of Unusual Words" *Journal of Computational Biology*, Volume 7, Number 1/2, 2000. P 71-94.

[3] J. Fisher, F. Hernandez, A. Sprague. "Language Patterns: Comparison and Prediction Using Hidden Markov Models". *Proceedings of the ACMSE'03- ACM Southeastern Conference*

[1] Rogic et al. "Evaluation of Gene-Finding Programs on Mammalian Sequences" *Genome Research*, 2001. May 11 (5): 817-832.

I8. Genome Organization Analysis Tool

Aaron Kaluszka, Cynthia Gibas[1]

Keywords: genome comparison, visualization, web applications

1 Introduction.

Traditionally, point mutations of genes have been the basis of comparative sequence analyses, most often used as a tool for phylogenetic and evolutionary studies. However, point mutations are of limited utility when comparing either slowly evolving genes and genomes or rapidly evolving genes and genomes such as some pathogenic viral genomes. Even when the nucleotide sequences of genes are highly conserved, their position within the genome may change. In these cases, a feature-based genome comparison is a useful complement to genome sequence comparison.

Genome Organization Analysis Tool (GOAT) uses the visualization concept introduced in GeneOrder[1] and incorporates a variety of additional features. These include a variety of input types, a flexible scheme of match filtering criteria, the caching of results to improve speed, and the ability to store data under a username so that one can return at a later date to refilter and redisplay results. Interactive gene order plots allow users to navigate the genome by selecting specific features, to customize plots by zooming and filtering, to navigate to the relevant sequence information and BLAST pairwise alignments represented by each point in the plot, and to view multiple alignments for equivalent genes.

2 Experiment and Methods.

GOAT compares one or more query genome annotations against a reference genome annotation and outputs a two-dimensional graphical plot of gene matches relative to the reference genome, where each gene match is a probable protein homologue. GOAT allows the use of four different cutoff criteria in any combination: bit score, expect value, percentage identity, and percentage of the query length represented in the match segment. The density and linearity of this plot represents the genomes' similarity to each other. Tabular output is also available, if desired. GOAT compares each gene in each query annotation to each gene in the subject genome using BLAST[2]. In the current implementation, we assume that the input will be a pre-annotated genome sequence, and low-complexity filtering is turned off so matches will not be missed. The BLAST output files are saved locally in a pre-parsed format, and are used as an index for quick information reference and to provide information necessary to generate dot plots of the query genome(s) against the subject genome. GOAT is written in Perl and may be run as a standalone application, but is intended to be used with an accompanying web-based front-end. An administration interface allows the user to install GOAT locally with a specific set of genome data, to restrict or enable user-initiated comparisons, and to restrict or enable multi-processor access, in effect allowing GOAT to be used to maintain a focused online genome comparison resource for specific species. GOAT is designed to be a user-friendly application for performing comparative analyses of gene order in small genomes or chromosomes, while being extensible for integration with other data mining

[1] Department of Biology, College of Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, U.S.A. E-mail: cgibas@vt.edu

applications. GOAT is accessible from any of the more common GUI-based web browser applications, such as Internet Explorer, Netscape, Mozilla, and Safari.

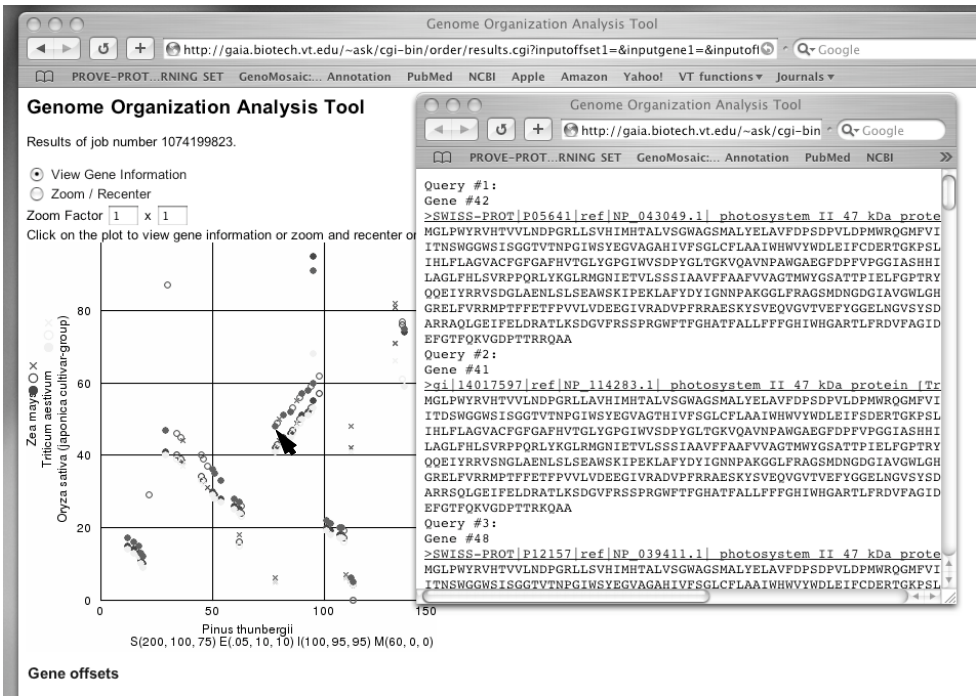


Figure 1: Screen shot from Genome Organization Analysis Tool. The plot shows chloroplast gene order in grasses relative to *p. thunbergii*. Each dot in the plot links to an informational screen showing pairwise and multiple alignments of genes that match the query and fall within the user-selected cutoff ranges.

3 Conclusions.

The two-dimensional gene order dot plot is a convenient format for visualization of structural changes in small genomes, and for visual identification of regions of colinearity and large-scale inversions. We have developed a web application which adds several features to the basic gene order dot plot, including comparison of multiple queries to the reference genome, interactive rescaling, GenBank link-outs, and point-and-click access to pairwise and multiple alignments of related genes in the query and reference. A GOAT demonstration website featuring publicly available chloroplast genome sequences is available at http://gaia.biotech.vt.edu/goat/. The software is freely available for noncommercial use under the Gnu Public License (GPL).

References

[2] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang,J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* 25: 3389-3402.

[1] Mazumder, R., Kolaskar, A., and Seto, D. 2000. GeneOrder: comparing the order of genes in small genomes. *Bioinformatics* 17: 162-166.

I9. Computing the Global Similarity of two Strings, with a Vector Algorithm

Sylvie Hamel, [1]

Keywords: Global similarity, vector algorithms, automata, proteins

1 Introduction.

Global similarity between proteins was first discussed by Needlemann and Wunsch in 1970 [6]. The classic solution computes an $n \times m$ table, where n and m are the length of the two proteins being compare, in $\mathcal{O}(nm)$ time.

This poster presents new efficient vector algorithms for the *global string similarity problem*. The key idea is to first reduce the computation to a Moore automaton, as done in [1] for approximate string matching problem using edit distance, and then to obtain the output of this automaton in a bounded number of steps, regardless of the length of the input.

This yields a very efficient algorithm since the operations are bit-wise operations widely available in processors, and the number of operations is independant of the length of the column in the classic table.

2 Global String Similarity Problem.

The *global string similarity problem* is to find a *maximum score* alignment between two strings, given a similarity matrix $\mathcal{S}$ on the input alphabet (Habitually PAM [2] or BLOSUM [4] matrix, when dealing with proteins). Let $x = x_1x_2 \dots x_m$ and $y = y_1y_2 \dots y_n$ be two strings on an input alphabet Σ. An *alignment* A of x and y produces two strings x', y' of length $l \geq \max(m,n)$, on the alphabet $\Sigma' = \Sigma \cup \{-\}$, where "$-$" correspond to the insertion or deletion of a letter in one of the strings. The *score* of the alignment A is then defined as $\sum_{i=1}^{l} \mathcal{S}(x'[i], y'[j])$.

The classic solution, referred as the Needleman-Wunsch algorithm [6] and requiring $\mathcal{O}(nm)$ time, is obtained by computing the matrix $V[0..m, 0..n]$ with the recurrence relation:

$$V[i,j] = \max \begin{cases} V[i-1, j-1] + \mathcal{S}(x[i], y[j]) \\ V[i, j-1] - c \\ V[i-1, j] - c \end{cases} \quad (1)$$

where c is the cost associated to an insertion or a deletion. The value $\mathcal{S}(a,b)$ is the score obtained by aligning letters a and b and, the initial conditions are $V[i,0] = -ci$ and $V[0,j] = -cj$. The entry $V[m,n]$ of the matrix then give the maximal score of an alignement between x and y.

Classically the computation of each column of V takes $\mathcal{O}(m)$ steps. In order to recast the computation as an automata computation, one needs to restrict the possible outputs to a range that is independant of m. This is done by noting that the differences between two adjacent horizontal or vertical cells in the table V are bounded. More precisely, define $\Delta v_{i,j} = V[i,j] - V[i-1,j] + c$ and $\Delta h_{i,j} = V[i,j-1] - V[i,j] + (c + M)$, where M is the maximal element of the given similarity matrix. We claim the following:

[1]DIRO, Université de Montréal, CP 6168 succ. Centre-Ville, Montréal, Québec, Canada, H3C 3J7. E-mail: sylvie.hamel@umontreal.ca, with the support of NSERC

(1) Both $\Delta v_{i,j}$ and $\Delta h_{i,j}$ are in the interval $[0, 2c + M]$.

(2) $V[m, j]$, can easily be computed as $V[m, j-1] - \Delta h_{m,j} + c + M$.

(3) Given $\boldsymbol{\Delta h_{i-1}} = (\Delta h_{i-1,1}, \ldots, \Delta h_{i-1,n})$ and $\boldsymbol{\mathcal{S}(x_i, y)} = \mathcal{S}(x_i, y_1), \ldots, \mathcal{S}(x_i, y_n))$, we can computed the value of $\boldsymbol{\Delta v_i} = (\Delta v_{i,1}, \ldots, \Delta v_{i,n})$ with a Moore automaton having states $\{0, \ldots, 2c + M\}$, with initial state 0 and transition fonction F given by $F(s, (\Delta h, \mathcal{S})) = max\{\Delta h - M + \mathcal{S}, \Delta h - (2c + M) + s, 0\}$.

(4) $\boldsymbol{\Delta h_i}$ can then be computed with the vector equation $\boldsymbol{\Delta h_i} = \uparrow_0 \boldsymbol{\Delta v_i} - \boldsymbol{\Delta v_i} + \boldsymbol{\Delta h_{i-1}}$, where $\uparrow_0 \boldsymbol{\Delta v_i}$ stands for a right shift of the vector $\boldsymbol{\Delta v_i}$ with the value 0 filled in in the first position.

Claims (1) to (4) thus reduce the computation of $V[m, j]$ to a few vector operations and one automata computation, which still takes $\mathcal{O}(m)$ steps.

3 From Automata to Vector Algorithms.

Given an input $\boldsymbol{e} = e_1 \ldots e_m$, a Moore automaton yields the output $\boldsymbol{r} = r_1 \ldots r_m$ of the visited states on input $\boldsymbol{e}$. For a particular class of automata, called *solvable* automata in [1], there is a general way to produce a vector algorithm that computes the output -regardless of its length- in $\mathcal{O}(q)$ steps, where q is the number of states of the automaton. Fortunately, the automaton of claim (3) is shellable.
The key step of the algorithm is the following. Each value r_i in the output $\boldsymbol{r} = r_1 \ldots r_m$ belongs to the interval $[0, 2c + M]$. Suppose that, for a given k, all the values r_i such that $r_i \geq k + 1$ are known. Then one can decide whether $r_i = k$, or not, in a bounded number of steps with the following procedure:

a) If $r_{i-1} < k$ and the input is $(\Delta h, \mathcal{S})$, where $\Delta h + \mathcal{S} = k + M$ then $r_i = k$.

b) If $r_{i-1} = k$ and the input is $(\Delta h, \mathcal{S})$, where $\Delta h = 2c + M$ or $\Delta h + \mathcal{S} = k + M$ then $r_i = k$.

c) If $r_{i-1} > k$ and the input is $(\Delta h, \mathcal{S})$, where $\max\{\Delta h - M + \mathcal{S}, \Delta h - (2c + M) + r_{i-1}\} = k$ then $r_i = k$

Cases *a*) and *c*) are rather straigtforward to implement. Only case *b*) requires knowledge of a bounded "past" history. This case is solved by a clever use of binary addition with carry propagation, first developed in [5] and generalized in [1].

Using the above procedure, we were able to implement the algorithm using $\mathcal{O}(c + M)$ steps, where c is the cost associated to an insertion or a deletion [3].

References

[1] Bergeron, A. and Hamel, S. 2002. Vector Algorithms for Approximate String Matching, *International Journal of Foundations of Computer Science*, 13-1: 53-66.

[2] Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. 1978. A model of evolutionary change in proteins, *Atlas of Protein Sequence ans Structure*, 5:345-352.

[3] Hamel, S. Finding global similarities in proteins using vector algorithms, in preparation.

[4] Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks, *Proc. Natl. Academy of Science*, 89: 10, 915-19.

[5] Myers, G. 1999. A Fast Bit-Vector Algorithm for Approximate String Matching Based on Dynamic Programming. *J. ACM* 46-3:395–415. pp. 75-83.

[6] Needleman, S.B. and Wunsch, C.B. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48: 443–453.

I12. Exact Algorithms for Matrix Based Motif Extraction

Paul Horton and Wataru Fujibuchi [1]

Keywords: motif extraction, branch and bound algorithm, transcription binding sites

1 Introduction.

Motif discovery is the problem of finding local patterns (*motifs*) from a set of unlabeled sequences. One common representation of a motif is a position specific score matrix. The problem of discovering motifs from unlabeled sequences has been extensively studied and a large number of heuristic algorithms have been proposed to solve it (reviewed by Stormo [1]). Heuristic algorithms have been used because obvious exact algorithms suffer from exponential time complexity, and indeed a general formulation of this problem has been shown to be APX-hard [2]. Thus exact matrix based motif discovery is generally considered to be an intractable problem.

However the APX-hard result only applies to problem instances with unrealistically long motif widths ($\Omega(n^{25})$ for n sequences), and indeed some serious attempts have been made to create exact branch and bound algorithms for this problem [3], [4]. In particular the TsukubaBB program [4] can solve large problems for very short motif widths such as five. In this poster we present results using an improved version of TsukubaBB which can solve more realistic problems. In particular we demonstrate its usefulness for finding transcription binding sites.

We also present a web server (seq.cbrc.jp) which provides quick heuristic algorithms for motif extraction, as well as other tools begin developed such as protein localization prediction from amino acid sequence and cell type classification from microarray data.

2 Results

Recently we have theoretically validated the TsukubaBB approach to motif extraction with a non-trivial upper bound on the running time [5]. In this poster we show empirical validation of the method when applied to characterizing transcription binding sites. For example we ran TsukubaBB on a problem instance of 14 sequences (and their reverse complements) taken from the yeast promoter database SCPD [6] for REB1 transcription factor binding sites. The sequences were of length 250 to 900 with a harmonic mean of approximately 467. The consensus given in SCPD was of length 7 with 20 sites known. Using pseudocounts of one for each base as a prior and the base composition of the input sequences (with their reverse complements added, so really just g+c content) as the background model TsukubaBB found the matrix (including pseudocounts) shown in figure 1 in a sequence logo [7] type representation. We have confirmed that most of the known sites given in SCPD match this matrix very well. The running time on the same 2.8GHz PC was 574 minutes. (Note that the TsukubaBB program does not currently take advantage of the input symmetry caused by adding the reverse complement, which could be used to speed things up).

[1]Computational Biology Research Center AIST, Japan. E-mail: horton-p@aist.go.jp, fujibuchi-wataru@aist.go.jp

Figure 1: An optimal scoring motif found in the REB1 dataset is shown

3 References and bibliography.

References

[1] Stormo, G. DNA binding sites: representation and discovery. *Bioinformatics*, 16:16–23, 2000.

[2] Akutsu, T., Arimura, H. and Shimozono S. 2000. On approximation algorithms for local multiple alignment. In *Proceedings of the fourth annual international conference on computational molecular biology (RECOMB2000)*, pages 1–7. ACM Press, 2000.

[3] Horton, P. A branch and bound algorithm for local multiple alignment. In *Pacific Symposium on Biocomputing '96*, pages 368–383, 1996.

[4] Horton, P. Tsukuba BB: A branch and bound algorithm for local multiple alignment of DNA and protein sequences. *Journal of Computational Biology*, 8(3):249–282, 2001.

[5] Horton P. and Fujibuchi W. 2004. In preparation.

[6] Zhu, J., Zhang, M. SCPD: a promoter database of the yeast saccharomyces cerevisiae. *Bioinformatics*, 15:607–611, 1999.

[7] Schneider, T., Stephens, R. Sequence logos: A new way to display consensus sequences. *Nucl. Acids Res.*, 18:6097–6100, 1990.

I13. An improved method of finding over-represented sequence motifs in sets of DNA sequences.

Tadashi Imanishi[1], Hiroki Hokari[2], Motohiko Tanino[3], Jun-ichi Takeda[4], Taichiro Sugisaki[5] and Shin Nurimoto[6]

Keywords: sequence motifs, OSMO-finder, computer simulation

1 Introduction.

Aside from gene coding sequences, eukaryotic genomes contain important functional sequences such as regulatory elements of gene expression and chromosomal replication. However, many of them remain uncharacterized. Finding functional sequence motifs in genomic sequences is thus a very important and urgent issue in bioinformatics. We thus developed a method for finding sequence motifs with the aim of discovering unknown functional sequence motifs in genomes, and we implemented the method in a computer program called OSMO-finder (Over-represented Sequence MOtif finder)[1]. This is a heuristic method of finding sequence motifs that appear frequently in sets of DNA sequences. We have extensively improved the algorithm of OSMO-finder and examined its efficiency of finding sequence motifs by using computer simulations. In this paper, we describe the outline of the improved algorithm and the results of computer simulations.

2 An improved algorithm of OSMO-finder.

The purpose of the method is to find a consensus DNA sequence (motif) without gaps that are over-represented in a set of unaligned sequences. Motifs can be evaluated based on the properties including the length, the number of occurrences (L) in the data set, and the level of sequence conservation (the number of nucleotide mismatches allowed). In OSMO-finder, we evaluate the identified motifs by the exact probability (P-value) that the motifs appear L times or more in random sequences of the same size as the data set. OSMO-finder first finds "seeds" of motifs of a given length W that appear frequently, allowing some mismatches, in a set of DNA sequences. It then searches for the optimum motif by calculating P-values for variously modified motifs by extending the seeds and changing the levels of sequence conservation.

[1] Biological Information Research Center, National Institute of Advanced Industrial Science and Technology. Time24 Bldg. 10F, Aomi 2-45, Koto-ku, Tokyo 135-0064, Japan. E-mail: imanishi@jbirc.aist.go.jp
[2] Mitsui Knowledge Industry Co., Ltd. Harmony Tower 21F, Honcho 1-32-2, Nakano-ku, Tokyo 164-8721, Japan. E-mail: hhokari@jbirc.aist.go.jp
[3] Japan Biological Information Research Center, Japan Biological Informatics Consortium. Time24 Bldg. 10F, Aomi 2-45, Koto-ku, Tokyo 135-0064, Japan. E-mail: mtanino@jbirc.aist.go.jp
[4] Biological Information Research Center, National Institute of Advanced Industrial Science and Technology. Time24 Bldg. 10F, Aomi 2-45, Koto-ku, Tokyo 135-0064, Japan. E-mail: jtakeda@jbirc.aist.go.jp
[5] Mitsui Knowledge Industry Co., Ltd. Harmony Tower 21F, Honcho 1-32-2, Nakano-ku, Tokyo 164-8721, Japan. E-mail: sugisaki@hydra.mki.co.jp
[6] Mitsui Knowledge Industry Co., Ltd. Harmony Tower 21F, Honcho 1-32-2, Nakano-ku, Tokyo 164-8721, Japan. E-mail: nurimoto@hydra.mki.co.jp

Finally, the motif with the smallest *P*-value is chosen as the optimum one. The program is written in JAVA and can be run under various operating systems.

3 Computer simulation.

To examine how efficiently OSMO-finder can discover sequence motifs in sets of DNA sequences, we conducted computer simulation using artificial sequences with hiddensequence motifs. We generated 10 artificial DNA sequences of 600 base pairs (bps) and inserted 10 or 20 sequence motifs of 15 bps. We then run the OSMO-finder (version December 2003) and measured the efficiency of finding hidden motifs. The test was repeated 5 times. The efficiency was measured by the level of overlap of correct and predicted motifs. As a result, it appeared that OSMO-finder can discover hidden motifs efficiently (Table 1). For example, 81% of the hidden motifs were successfully identified when there are 20 motifs with 10% sequence diversity from a consensus sequence. We also compared the efficiency of OSMO-finder with that of MEME (version 3.0.4) with "Two-Component Mixture" (TCM) option[2]. These two programs showed comparable efficiency. In particular, OSMO-finder showed better performance when the level of sequence conservation of motifs is low. These results clearly demonstrate the usefulness of OSMO-finder in identifying functional sequence motifs from genomic sequences.

number of hidden motifs in the data	diversity of sequence motifs	OSMO-finder	MEME (TCM)
10	0%	0.718	0.940
10	10%	0.731	0.819
10	20%	0.242	0.017
20	0%	1.000	1.000
20	10%	0.811	0.920
20	20%	0.346	0.030

Table 1. Efficiency of finding hidden sequence motifs by OSMO-finder and MEME (TCM).

References

[1] Imanishi T., Shikanai T., Nurimoto S. and Sugisaki T. 2003. A new method of finding functional sequence motifs and its application to human GC-rich genomic sequences. In Spang R., Beziat P. and Vingron M. editors, Currents in Computational Molecular Biology 2003 (RECOMB 2003, Berlin), pp. 61-62.

[2] Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36.

I14. Exact algorithm for discovery of consensus sequences among multiple sequences to design degenerate primers

Chul Hyun Joo, Hwisun Lee, Jinho Lee, Heuiran Lee, Yoo Kyum Kim [1]

Keywords: degenerate primer, trie, multiple sequences

1 Introduction.

The identification of consensus sequences among multiple DNA fragments has assumed importance following the development of the polymerase chain reaction technique. A primer is called degenerate if some of its positions have several possible bases [3]. Degenerate primers have proven useful for studying gene families [4]. Many algorithms and programs have been developed to solve the problem by heuristic approaches. We describe an exact algorithm for finding all consensus sequences that can complementarily hybridize with degenerate primers of given length, number of degenerate positions, and coverage.

2 Algorithm.

We divide the problem into multiple exact matching problems. First, all possible degenerate patterns are generated for a given length and number of degenerate positions. Then all suffixes of a given set of sequences are transformed into same sized substrings with omission of the degenerate positions. Finally, strings common to the transformed substrings above the coverage are archived by using a trie [2] that was modified from the generalized suffix tree [1]. Example of a generalized *pm* trie with a given proper mask $pm(101101)$ and sequences set $\{s_1, s_2\}$ is shown in Figure 1.

The worst case time complexity of the algorithm to find (l, d, k) degenerate probe among a set $\{s_1, \ldots, s_i\}$ is $O(\binom{l-1}{d}\Sigma|s_i|)$, where l is the length of primer, d is the degeneracy, and k is the coverage. We made two major improvements by premature termination of each trie construction step using the k value.

3 Experimental Results.

The algorithm was tested using synthetic data sets with various parameters typical of experimental settings, and showed acceptable computation times. We used this algorithm to find four degenerate primer binding sites to design nested primer sets, among 63 full genomic sequences of enteroviruses, where the length of each sequence was approximately 7400 bases. We searched for primer of length 18-20 with minimum degeneracy among the set. The four consensus sequences covering the 63 serotypes were successfully selected from the region of the 5' untranslated terminal repeat lesion of the viral gene. The (l, d, k) parameters for the algorithm were (17,2,63), (23,1,63), (21,1,63), and (20,0,63). We can found each primer binding site within a few seconds when run on a Pentium 4 2.4GHz, 1G RAM machine.

[1]Department of Microbiology, University of Ulsan College of Medicine, 388-1 Pungnap-dong Songpa-gu, Seoul 138-736, Korea. E-mail: ykkim@amc.seoul.kr

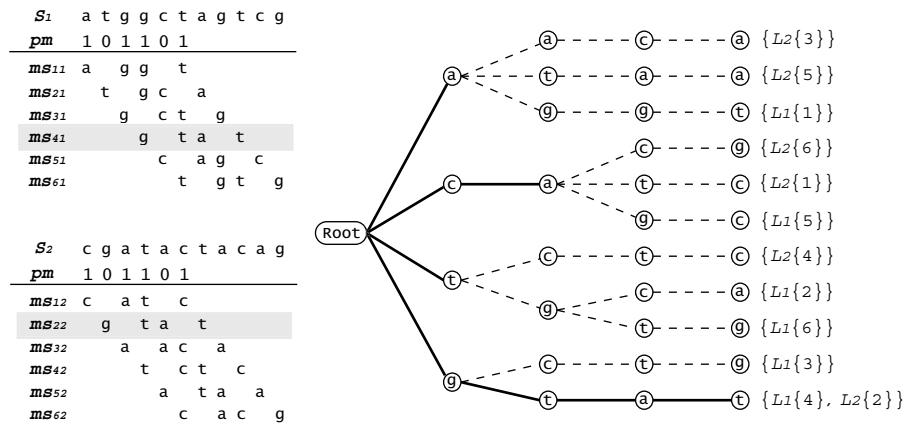


Figure 1: The consensus sequence is the path from the root node to leaf nodes that has a leaf cardinality of 2.

4 Discussion.

We implemented the algorithm in the program called Prober that can be downloaded at no cost from "http://prober.biocodes.org". The Prober is a Windows based consensus sequence archive program that can be run on a PC. The number of patterns, number of probes found, elapsed time and remaining time are displayed at each step to assist the user in deciding whether to quit the thread without waiting for termination. The output file contains information such as positions, sequences, degenerate description of probes using ambiguity codes. One merit of the Prober is that the time required to find a good probe is much less than that required to a bad probe (i.e. a probe that covers less genes or has more degenerate positions).

References

[1] Gusfield, D. 1997. Linear time construction fo suffix trees. In Gusfield, D. editors, *Algorithms on strings, trees, and sequences*, New York: Cambridge. vol. 3: pp. 94-119.

[2] Knuth, D. E. 1998. Digital searching. In Knuth, D. E. editors, *The Art of Computer Programming 2nd ed.*, Reading, MA: Addison-Wesley. vol. 3: pp. 492-507.

[3] Kwok, S., Chang, S. Y., Sninsky, J.J., and Wang, A. 1994. A guide to the design and use of mismatched and degenerate primers. *PCR Methods and Appl* 3:S39-47.

[4] Rose, T.M., Schultz, E.R., Henikoff, J.G., Pietrokovski, S., McCallum, C.M., and Henikoff, S. 1998 Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res* 26:1628-35.

I15. GAME: Genome Alignment by Match Extension

Jeong-Hyeon Choi[1] Hwan-Gue Cho[2] Sun Kim[3]

Keywords: genome sequence alignment, anchor filtering, maximal exact match

1 Introduction

As the number of completely sequenced genomes is increasing rapidly, there is an urgent need for efficient and effective computational methods for comparing multiple genomes [4]. One important computational method is to align whole genomic sequences. Aligning two genomic sequences requires detecting numerous local alignments – homologous regions such as genes common to two genomes. Due to the large size of genomes, almost all genome sequence alignment algorithms first find anchors and extend or combine them to generate sequence alignments. There are two common approaches used in detecting anchors, one based on exact matches and the other based on approximate patterns.

In this paper, we present a new adaptive genome sequence alignment based on maximal exact match (MEM) anchor. The major problem with the use of MEM anchor is that the number of hits in non-homologous regions increases exponentially when shorter MEM anchors are used to detect more homologous regions. To deal with this problem, we have developed a fast and accurate anchor filtering scheme based on simple match extension. Due to its simplicity and accuracy, all MEM anchors in a pair of genomes can be exhaustively tested and filtered. As a result, our genome alignment algorithm outperforms existing algorithms and can align large genomes, e.g., *A. thaliana*, without the typical large memory requirement problem.

2 Extended MEM (EMEM): Anchors by Gap-free Extension

The key idea is simply to extend MEMs without computing alignment scores until the percent identity becomes lower than the ones used in existing genome alignment algorithms. The EMEM anchors are generated in two steps:

1. All MEMs are detected using suffix array [3]. Use of suffix tree suffers the intrinsic large memory space problem. However, the space problem can be significantly reduced by using suffix array.
2. Each MEM anchor is tested with a filter, called *the MEM filter*. More specifically, given a maximal exact match (MEM) m of length l_m and a threshold T_{pi} for the minimum percent identity, the *gap-free match extension* of m is defined as the extent up to which m can be extended without gaps before its percent identity drops below T_{pi}. Let *gfe(m)* denote the grap-free extension of m and |gfe(m)| denote its length. Given a threshold T_e for the minimum extension length, *the MEM filter* is defined to filter (discard) any m that fails to extend beyond T_e , i.e., $|gfe(m)| < T_e$.

Due to its simplicity, the filtering scheme can further utilize a well developed string pattern matching technique using bit parallelism; multiple characters can be compared simultaneously using the technique, e.g., [2]. Our filtering scheme is accurate [1] and runs

[1]Department of Computer Science, Pusan National University, Korea. jeochoi@indiana.edu
[2]Department of Computer Science, Pusan National University, Korea. hgcho@pusan.ac.kr
[3]School of Informatics, Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN, USA, sunkim@bio.informatics.indiana.edu.

exceptionally fast, performing over one million filtering tests per second on a 2.0 Ghz Pentium machine, which makes it possible to test *all* initial anchors exhaustively in a pair of genomes. As a result, our alignment algorithm can align genomes fast and accurately.

3 Experiments

Two tests were performed on a Pentium 2.0 GHz machine with 4GB memory running RedHat Linux 9.0, all pairwise comparisons of 10 bacterial genomes from close to distant in terms of phylogeny (*Mycoplasma genitalium, Mycoplasma pneumoniae, Bacillus halodurance, Bacillus subtilis, Staphylococcus aureus, Mycobacterium tuberculosis, Helicobacter pyroli, Escherichia coli, Haemophilus influenza,* and *Methanococcus jannaschii*) and all pairwise comparisons of 5 chromosomes of *A. thaliana*. The performance in terms of the number of COG families detected for 45 genome pair comparison are shown in Figure 1. The memory usage of GAME for all pairs ranged only from 25MB to 125MB for the actual numbers). GAME was able to perform *all* pairwise comparisons of the five chromosomes of *A. thaliana*, the shortest being 17.5 Mb and the longest being 29.6 Mb, without the typical memory problem. Experimental data will be available shortly at [1].

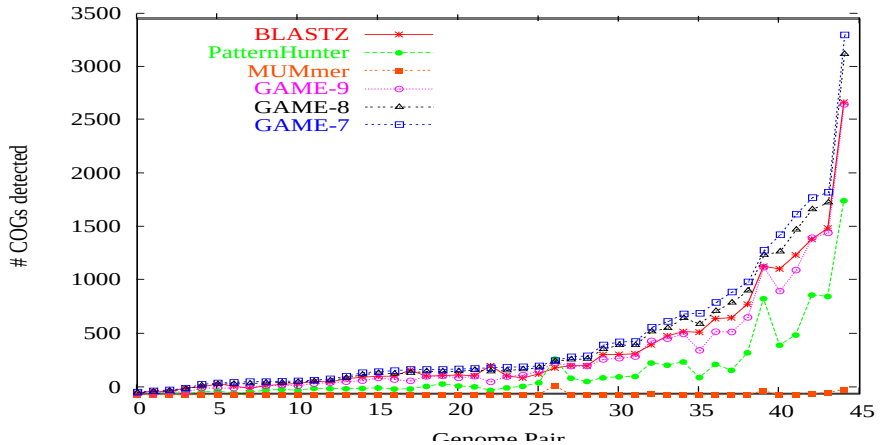


Figure 1: MUMmer ran with the MUM option. GAME-7, GAME-8, and GAME-9 are with MEMs of length 7, 8, and 9 respectively. Blastz and GAME-9 were competitive. GAME-7 and GAME-8 outpeformed others in all cases except 1 and 4 genome pairs respectively.

References

[1] http://bio.informatics.indiana.edu/projects/GAME/.

[2] S. Kim and Y. Kim. A fast multiple string-pattern matching algorithm. In *Proc. of The 17th AoM/IAoM Interantional Conference on Computer Science*, pages 44–49, 1999.

[3] U. Manber and G. Myers. Suffix arrays: A new method for on-line string searches. *SIAM J. Comput.*, 22(5):935–948, 1993.

[4] W. Miller. Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics*, 17(5):391–397, 2001.

I16. Cumulative Local Cross-Correlation – an Algorithm for the Decomposition of Sequence Patterns

Simon Kogan[1]

Keywords: cross-correlation, pattern, repeat, sequence

1 Introduction.

Nucleotide sequences (DNA), a store of biological inheritance information, contain multiple codes (messages) responsible for an organism's functioning and structure [8]. A nucleic sequence can be seen as a signal and investigated by formal methods of signal processing theory: Fourier [6] and wavelet [1] transforms, cross-correlation analysis [2, 7], etc. Cross-correlation is especially suited for the indication of common (frequently encountered) sequence parts (i.e., repeats). A dispersed repeat (frequent combination of oligonucleotides following each other at some specific distance with an arbitrary sequence in between) will be captured as well.

Examining cross-correlation between AA and TT dinucleotides in the human genome reveals a peak in lag '12' that suggests an existence of sequence repeat(s) containing AA dinucleotide followed by TT dinucleotide at a distance of 12 bases. One can find all positions of such a motif (AA-12-TT) in a given sequence, and then, calculate a nucleotide local distribution in the vicinity of the motif. If there is only one unique repeat containing this motif, the distribution is actually a pattern representing the repeat. If there are several different repeats containing the motif, the distribution is an overlapping of all of the patterns. To reconstruct one of the patterns or all of them, one has to decompose the distribution picture. This is exactly the objective of the algorithm developed in this work.

2 The algorithm.

In order to calculate cross-correlation functions and also use the Cumulative Local Cross-Correlation (i.e., CLCC) algorithm, one needs to represent an investigated nucleic sequence by 4 discrete signals: the first one for 'A' nucleotide positions ('1' – where it is present, '0' – where it is absent); the second one for 'C' nucleotide positions, etc.

The algorithm is as follows:

For a given nucleic sequence, calculate nucleotide local distribution in the vicinity of the initial motif (AA-12-TT for example).

For each occurrence of the motif in the sequence, calculate the cross-correlation (including auto-correlation) between the signals in each pair of positions inside a window enclosing the motif. The result of this operation is a symmetric matrix (local cross-correlation instance).

Calculate cumulative local cross-correlation matrix as a sum of local cross-correlation instances. The number of CLCC matrices is equal to the number of cross-correlation functions (16 in our nucleotide example). Each element of CLCC matrix characterizes correlation between a pair of positions in the local distribution.

Now, in order to know, whether two specific peaks in the local distribution are related to the same pattern or not, one just needs to check a value of CLCC matrix element associated with them (larger value for correlated peaks than for uncorrelated ones). Theoretically, by checking CLCC matrix elements associated with all pairs of peaks in a local distribution and collecting mutually correlated peak groups, all patterns can be separated from each other.

[1] *Genome Diversity Center, Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel. E-mail:* skogan@research.haifa.ac.il

3 Application to human genomic sequences.

The following example illustrates the CLCC application to one contig (7MB) of human chromosome **1**. The development of this algorithm took place during the investigation of a nucleosome pattern based on dinucleotides [5]. There are 16 local distributions and 256 cross-correlation functions (and CLCC matrices as well) in dinucleotide case (against 4 of the former and 16 of the latter in nucleotide case).

AA-12-TT was chosen as an initial motif and CLCC matrices (and local dinucleotide distributions) in its vicinity were obtained as described in the previous section. Analysis of CLCC matrix elements associated with several pairs of peaks in the local distribution permitted an isolation of mutually correlated group of peaks. The new, more detailed motif was constructed from these peaks and dinucleotide local distributions in the vicinity of the new motif were calculated.

Knowing, that the new distribution is a non overlapped one, we converted it to a nucleic sequence (the non-overlapping is a necessary condition for the conversion). The resulting sequence was found to be very similar to the well-known Alu tandem repeat widely present in the human genome [3]. Therefore, we were able to align the sequence with Alu consensus [4] to validate the CLCC technique (see Fig. 1 for the alignment result).

```
AAAAAAAAAAAATTCAGGCgcGGgGCGGgGGggCcCcCgccTAAaCCCcacCcTTGGGgGGggGgGGgG
                GGCcgGGcGCGGtGGctCaCgCctgTAAtCCCagcaCtTTGGGaGGccGaGGcG

GGgGGgaacCCcGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACcTGGTGAAACCCCGgCTCTACTAAAA
GGcGGatcaCCtGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACaTGGTGAAACCCCGtCTCTACTAAAA

ATACAAAAATTAGCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGA
ATACAAAAATTAGCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGA

ATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGAC
ATCGCTTGAACCCGGGAGGCGGAGGTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGAC

AGAGCGgGACcCCGTCcCAAAAAAAAAAAAAAAAAAAAAAA
AGAGCGaGACtCCGTCtCAAAAAAAA
```

Fig. 1: Sequence alignment: upper (underlined) rows – the extracted nucleic sequence; lower rows – Alu consensus.

References

[1] Arneodo, A., E. Bacry, P.V. Graves, and J.F. Muzy. 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. Physical Review Letters. 74(16): pp. 3293-3296.

[2] Herzel, H., E.N. Trifonov, O. Weiss, and I. Grosse. 1998. Interpreting correlations in biosequences. Physica A. 249(1-4): pp. 449-459.

[3] Hwu, H.R., J.W. Roberts, E.H. Davidson, and R.J. Britten. 1986. Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution. Proc Natl Acad Sci U S A. 83(11): pp. 3875-9.

[4] Jurka, J. and T. Smith. 1988. A fundamental division in the Alu family of repeated sequences. Proc Natl Acad Sci U S A. 85(13): pp. 4775-8.

[5] Kato, M., Y. Onishi, Y. Wada-Kiyama, T. Abe, T. Ikemura, S. Kogan, A. Bolshoy, E.N. Trifonov, and R. Kiyama. 2003. Dinucleosome DNA of human K562 cells: experimental and computational characterizations. J Mol Biol. 332(1): pp. 111-25.

[6] Peng, C.K., S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H.E. Stanley. 1992. Long-range correlations in nucleotide sequences. Nature. 356(6365): pp. 168-70.

[7] Trifonov, E.N. and J.L. Sussman. 1980. The pitch of chromatin DNA is reflected in its nucleotide sequence. Proc Natl Acad Sci U S A. 77(7): pp. 3816-20.

[8] Trifonov, E.N. 1989. The multiple codes of nucleotide sequences. Bull Math Biol. 51(4): pp. 417-32.

I17. Predicting the modular domain architecture of a protein

Gülriz Aytekin-Kurban [1]

Keywords: protein modular domain architecture, domain-specific function prediction

Structural biology defines a domain as an independently folding, compact and stable fragment of a protein. Such useful structural properties are likely to have transformed the domains into evolutionary units, observed as recurring modules of distinct structure and possibly a related function throughout the protein universe [4]. In computational or experimental analysis of a protein, advance knowledge of its **modular domain architecture**(MDA), i.e., the number and relative positions of the domains in the amino acid sequence, is usually a prerequisite. For instance, in automated function annotation by sequence similarity, consensus function should be assigned only to the domain that shares similarity [1]. The automation of domain-specific function prediction can aid large scale genome annotation projects [2]. In experimental structure determination, if the domain positions are known, a large multi-domain protein can be partitioned into small domains that are more suitable to analysis via nuclear magnetic resonance or crystallization.

This poster presents a method that brings together domain boundary prediction to aid structural analysis and domain-specific function annotation into a single framework. The method builds a probabilistic model for the MDA of a multi-domain protein using the sequence alignments returned from a similarity search of a large dataset of non-redundant proteins. Domain boundary prediction and domain-specific function annotation using sequence alignments have been investigated before [1, 3]. Accomplishing both together increases the likelihood of identifying novel domains in little explored areas of protein space.

In local sequence alignments, we assume that each of the domain segments is aligned to a subset of instances of the same domain family in matching proteins. Consecutive domains are covered by a single alignment. A protein composed of domains, belonging to distinct families and recurring in different combinations, should have a pattern of alignments such that the sets of aligned proteins for any two distinct domain positions are distinguishable.

The MDA model is a probabilistic latent class model, i.e., mixture model. Each latent class has a conditional probability mass distribution over all matching proteins, the distribution models the homolog occurrences of a domain on matching proteins. Class assignment probabilities for positions partitions amino acid positions into segments, multiple segments can belong to a single domain.

The dot plot in Figure 1(a) displays a pattern of matching protein alignments for a multi-domain protein, a fructose specific enzyme. Each line represents one or more aligned segments of the protein to a database protein obtained by using PSI-BLAST to search **nr**[2]. There are two PFAM functional domains on the protein. The PTS_EIIA domain, positions from 2 to 142, contains the primary phosphorylation site. The PTS_HPr domain, from position 285 to 371, has an acceptor for the phosphoryl group of EI(eiia). Figure 1(b) shows a probabilistic partitioning of protein positions into 3 domains by the class assignment probabilities of a computed MDA model. If we commit each sequence position to a class with the highest assignment probability, we get domains: C1:[149-242], C2:[246-376], and C3:[1-148]. Domain C3 corresponds to PTS_EIIA domain and C2 corresponds to PTS_HPr

[1]Dept. of Computer Science, The University of Chicago, USA E-mail: gulriz@cs.uchicago.edu
[2]Detected fragmented sequences are removed.

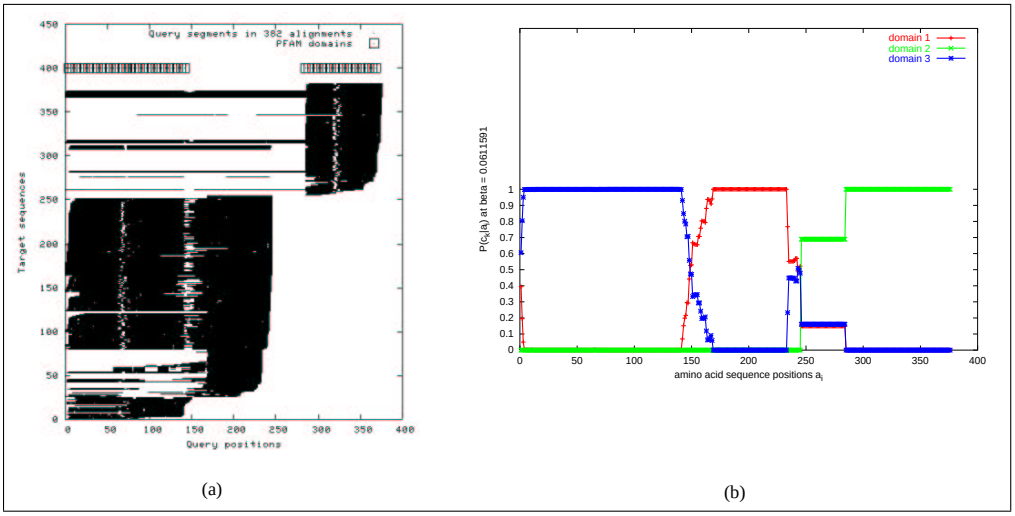


Figure 1: (a) PTFA_ECOLI sequence segments in local alignments, (b) predicted domain positions.

domain while C1 could be a structural domain. Class conditional probability distributions are used to process annotations of matching proteins to obtain phrases that best describe each domain as shown in Table 1.

Domain	Annotation phrases ranked by class conditional probabilities
1	emb probable phosphocarrier protein ***hpr***‖ emb similar to transcriptional antiterminator ‖ nitrogen regulatory ***iia*** like protein
2 hpr	gb aaa86048 ***hpr***‖ 1 putative phosphotransferase system enzyme ‖ np_312377 1 putative phosphotransferase system ‖ putative frv operon regulatory protein ‖ fructose specific pts transport system ‖ putative bglb family transcriptional antiterminator ‖ phosphocarrier protein ***hpr*** streptococcus suis
3 eiia	emb similar to transcriptional antiterminator ‖ probable phosphotransferase system enzyme i‖ phosphoenolpyruvate dependent sugar phosphotransferase system ‖ gb aal00742 conserved hypothetical protein ‖ putative nitrogen regulatory ***iia*** transcription ‖ a42374 phosphotransferase system phosphohistidine containing ‖ regulatory ***iia*** transcription regulator protein

Table 1: The domain-specific annotation of PTFA_ECOLI.

A further improvement planned is to integrate other types of information, such as the conservation covariance of pairs of positions, to the model. Moreover, initial heuristic processing of matching protein annotations will be improved to obtain better descriptions for domains.

References

[1] M A. Andrade. Position-specific annotation of protein function based on multiple homologs. In *Proceedings of Intelligent Systems in Molecular Biology (ISMB)*, 1999.

[2] T. Gaasterland and C.W. Sensen. Magpie: Automated genome interpretation. *Trends in Genetics*, 12:96–98, 1996.

[3] R. A. George and J. Heringa. Protein domain identification and improved sequence similarity searching using psi-blast. *PROTEINS: Structure, Function and Genetics*, 48:672–681, 2002.

[4] E. V. Koonin, Y. I. Wolf, and G. P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420:218–223, 2002.

I18. A web-based tool for the identification of conformationally flexible segments in protein sequences.

Igor B. Kuznetsov[1], Byron Gerlach[2], S. Rackovsky[3]

Keywords: intrinsic propensity, amino acid, profile, structural flexibility, probability, prediction

1 Introduction.

It has long been axiomatic that an amino acid sequence defines a unique three-dimensional native structure of a protein. An increasing amount of data from NMR spectroscopy and other sources shows that not all residues in the native structure have the same microscopic stability. This means that at least certain segments of the native structure are very flexible and adopt a set of different dynamically inter-converting conformations. It is nearly impossible to characterize the structure of proteins that contain long flexible segments by means of X-ray crystallography or NMR spectroscopy. The success of large-scale structural genomics projects therefore crucially depends on our ability to identify such proteins and exclude them from the list of potential targets selected for experimental structure determination. Identification of conformationally flexible segments is also important for the understanding of the mechanism of conformational changes observed in misfolding diseases. One of the best-known examples of misfolding diseases is the group of neurodegenerative disorders (CJD, mad cow disease, etc.) caused by the prion protein, which contains a long highly flexible N-terminal domain. This protein represents an interesting subject for theoretical research designed to study conformational flexibility, since theoretical conclusions can be directly compared to the wealth of existing experimental data.

2 Methods and software.

We developed a novel entropic index, Generalized Local Propensity (GLP) [1], which provides a quantitative measure of conformational flexibility of a sequence fragment, and a propensity-based method that utilizes the GLP to identify segments with high and low conformational flexibility. A software program, Conformational Flexibility Profile (CFP) Tool, that implements this method has been developed. A unique feature of this program is that it not only identifies flexible segments, but also provides, for each given segment, a probability of observing segments of the same or greater length by chance. The program has a user-friendly web interface and is available at http://jay.bioinformatics.ku.edu/~gerlach/propensity.html. The application of the software to study conformational flexibility of the prion protein (PrP) has shown that it is able to detect a long conformationally flexible segment PrP(52-97) of size 46 located in the flexible N-terminal domain of the prion protein (Fig.1). The estimated probability of observing flexible fragment of this or greater length in random sequences with the same amino acid composition as that of human PrP and in the non-redundant protein sequence database is only about 10^{-4}. The N-terminal domain of PrP has been identified as highly flexible by NMR spectroscopy. This agreement with experimental results indicates that the software is capable of correctly identifying unusually long conformationally flexible segments.

[1] Bioinformatics Group, University of Kansas, 1002 Haworth Hall, Lawrence, KS 66045 E-mail: igor@ku.edu
[2] EECS Department, University of Kansas, Lawrence, KS 66045, E-mail: gerlach@mail.ku.edu
[3] Department of Biomathematical Sciences, Mount Sinai School of Medicine, One Gustave Levy Pl., New York, NY 10029, E-mail: shelly@camelot.mssm.edu

3 Figures and tables.

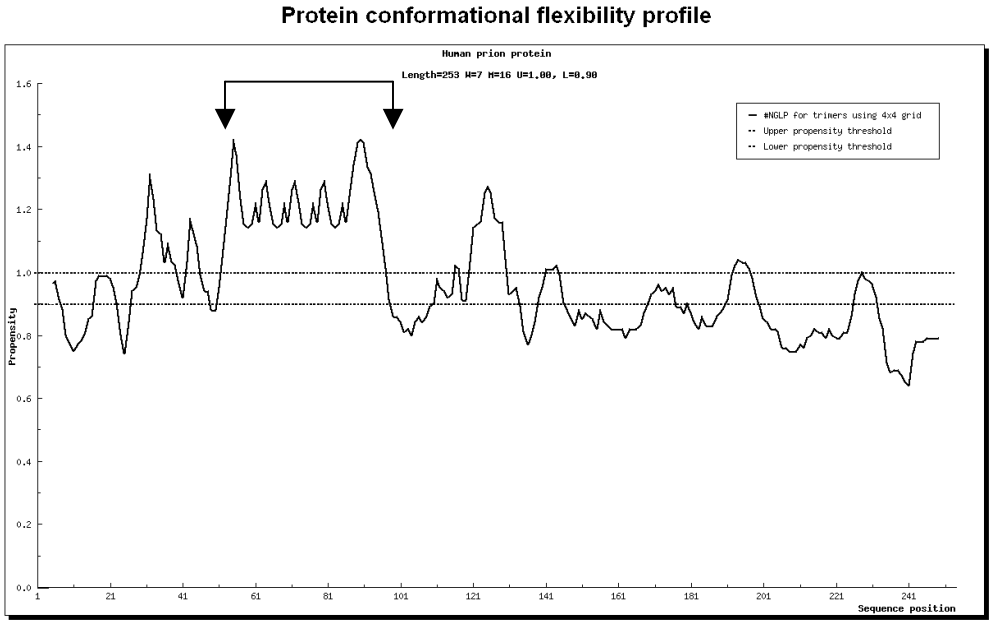


Figure 1: Flexibility profile for human prion protein (smoothed using a window of size 7) computed using the CFP tool. A long flexible fragment (GLP above 1.0) comprising residues 52 through 97 is shown by two arrows.

# of the fragment	Start	End	Length	Raw propensity	Smoothed propensity	Random probability	Database probability	Sequence
1	30	39	10	1.08	1.12	9.835000e-01	N/A	GWNTGGSRYP
2	42	45	4	1.03	1.10	1.000000e+00	N/A	GSPG
3	52	97	46	1.24	1.22	3.000000e-04	N/A	QGGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQPHGGGWGQGGGTHS
4	116	117	2	0.70	1.02	1.000000e+00	N/A	AA
5	120	130	11	1.11	1.16	9.505000e-01	N/A	AVVGGLGGYML
6	141	144	4	1.14	1.01	1.000000e+00	N/A	FGSD

Table 1: A table of all flexible fragments identified in human prion protein by the CFP tool. The fragment comprising residues 52 through 97 has a low probability of being observed by chance (p-value$\approx 10^{-4}$).

4 Acknowledgements.

This work was supported by NSF EPSCoR and grant number 1R01 LM06789 from the National Library of Medicine of the National Institutes of Health.

5 References.

[1] Kuznetsov, I.B. and Rackovsky, S. 2003. On the properties and sequence context of structurally ambivalent fragments in proteins. *Protein Science* 12:2420-2433.

I19. *In silico* Analysis of LASS1 (LAG1 Longevity Assurance homology 1) and Related Orthologs Using Target Identification Software Tools

Darryl LeÛn[1] and Scott Markel[2]

Keywords: bioinformatics, alignment, UOG, validation, prediction, patent

The LAG1 gene was first cloned from yeast, and it was found to extend the life span of the yeast up to 48% [1]. In 1998, the human LAG1 longevity assurance homolog 1 (LASS1) gene was cloned, sequenced, and characterized [2]. It has been reported that LASS1is associated with regulating C18-ceramide (N-stearoyl-sphinganine) synthase activity where ceramides have been associated with cell differentiation, growth, regulation, and death. [3]. With the aim of illustrating the value of LIONís target identification software and open-source bioinformatics tools, we selected the LASS1 gene as a potential drug target for analysis. LION technology confirmed that the LASS1 gene is located at 19p12 in humans, while the mouse and rat orthologs were also confirmed to be located on chromosomes 8 and 10 and chromosome 16p14, respectively. A BLAST search against genomic databases revealed other putative LASS1 sequences identified from *plasmodium falciparum*, chimpanzee, puffer fish, zebra fish, mosquito, and fly. A multiple sequence alignment of the LASS1 protein ortholog sequences was performed, and we identified several conserved regions in the encoded protein sequence. An expansion to a previous phylogenetic tree [1] was constructed to show the sequence relatedness among the various organisms. A secondary structure prediction showed there are five conserved putative transmembrane regions and three significant predicted helical regions. Three of the predicted transmembrane regions overlap helical regions and two of the predicted transmembrane regions do not. The number of predicted hydrophobic regions ranged from three to six depending on the complexity of the organism. There is a linear relationship between the number of hydrophobic regions per 100 amino acids and the organism type. It is proposed that simple organisms require a higher number of hydrophobic regions per 100 amino acids in LASS1 than more complex organisms. Further analyses included sequence motif searching, profile building, and patent searching, and results are presented. The prediction analysis of LASS1 demonstrates the value of an integrated *in silico* analysis for uncharacterized novel drug targets. These results collectively demonstrate that target identification tools provide insight for providing scientific direction when designing drug discovery experiments.

References

[1] Dímello, N.P., Childress, A.M., Franklin, D.S., Kale, S.P., Pinswasdi and Jazwinski, S.M. 1994. Cloning and Characterization of *LAG1*, a Longevity-assurance Gene in Yeast. *J Biol Chem*., 269:22, 15451-15459.

[2] Jiang, J.C., Kirchman, P.A., Zagulski, M., Hunt, J. and Jazwinski, S.M. 1998. Homologs of the Yeast Longevity Gene *LAG1* in *Caenorhabditis elegans* and Human. *Genome Research*, 8,1259-1272.

[3] Venkataraman, K., Riebeling, C., Bodennec, J., Riezman, H., Allegood, J.C., Sullards, M.C., Merrill, A.H. Jr. and Futerman, A.H. 2002. Upstream of Growth and Differentiation Factor 1 (*uog1*), a Mammalian Homolog of the Yeast Longevity Assurance Gene 1 (*LAG1*), Regulates N-Stearoyl-sphinganine (C18- (Dihydro)ceramide) Synthesis in a Fumonisin B_1-independent Manner in Mammalian Cells. *J Biol Chem*., 277:38, 35642-35649.

[1] LION bioscience Inc. 6126 Nancy Ridge Drive, Suite 118 San Diego, California, USA. E-mail: darryl.leon@lionbioscience.com

[2] LION bioscience Inc. 6126 Nancy Ridge Drive, Suite 118 San Diego, California, USA. E-mail: scott.markel@lionbioscience.com

I20. Novel Gene Discovery with Sequence Profile Comparison

Weizhong Li[1], Lukasz Jaroszewski[2], Adam Godzik[3]

Keywords: genome annotation, sequence profiles, sequence homology

1 Introduction.

Despite a human genome draft being available for over two years, the number of protein coding genes is still a matter of debate and novel genes are continuously being found. Homology based methods, which predict new genes by comparing an entire genome with known genes, have played an important role [1]. These methods can find real genes that may be systematically missed by other methods such as those that employ *ab initio* gene prediction. However, homology based methods also give rise to huge amount of false positive and pseudo-genes.

In the past years, we have been developing sensitive algorithms to efficiently detect more remote homologies and that implement comprehensive approaches to validate predictions [2-5]. Our key techniques are sequence and profile alignment tools and sequence clustering methods. Here, we present a novel gene discovery system that incorporates with these protein sequence and genome analysis tools. This system has been applied in identifying several novel proteins families, some of which represent potential new drug targets.

2 Methods.

In general, homology based approaches use known genes as queries to search against a genome to identify fragments that may encode genes. Typically, a very large number of fragments with statistically significant similarities to known genes could be found, which in itself is an interesting and not completely understood phenomenon. These fragments need to be ranked in order to produce a reasonable number of fragments that can advance to more detailed analysis such as gene assembly.

Our gene discovery system incorporates several processes: a) protein sampling, b) protein clustering and analysis, c) profile building, d) profile-sequence and profile-profile search against genome, e) fragment characterization, f) backward profile-sequence and profile-profile search, g) fragment filtering, prioritizing and validating, h) full gene assembly, and i) experimental validation. A list of algorithms, programs, and resources employed in this system are listed in table 1.

Given one or more known proteins the system a) first searches the **NR+** database with an intermediate sequence or profile search to retrieve all the close and remote homologues of this protein family; b) selects representative sequences of the family (depending upon diversity within the family) by applying an appropriate clustering tool such as **PSI-Clstr.** These representative

[1] Quorex Pharmaceuticals, Carlsbad, California, E-mail: wli@quorex.com
[2] San Diego Supercomputer Center, La Jolla, California, E-mail: lukasz@sdsc.edu
[3] The Burnham Institute, La Jolla, California, E-mail: adam@burnham.org

sequences are then used to retrieve a conserved sequence pattern from the multiple sequence alignment; c) then builds sequence profile for each representative protein from **rep-NR+**; d) performs profile-sequence and profile-profile search against **PCF** or **PCF-Profile** to collect genomic fragments; e) calculates a broad array of properties of fragments, such as sequence complexity, exon probability, relative positions to known genes and **EST**s, match score of specific sequence pattern. This calculation incorporates publicly available genome annotation databases; f) annotates the fragments by comparing them against known protein profile databases such as **PDB-Profile** and **Pfam-Profile**; g) filters out random hits and hits corresponding to known genes, and it also ranks fragments according to all available calculated results. The more detailed analyses are performed to the top ranking fragments; h) full length genes are assembled; i) fragments and assembled gene are tested experimentally.

CD-HIT	Protein sequence clustering algorithm with very high speed to handle medium to high homology on very huge database
BLAST-Clstr	BLAST-based clustering algorithm which can handle medium to low homology
PSI-Clstr	PSI-BLAST-based clustering algorithm which can handle very remote homology
Saturated-BLAST	Intermediate sequence and profile search algorithm, which can effectively explore diverse protein family. It also offers multiple sequence alignment and clustering.
FFAS	Sensitive profile-profile alignment algorithm
NR	Public protein database from NCBI
NR+	NR plus other proteins predicted or annotated from some specific genomes
REP-NR+	Representative protein families from NR+ prepared with CD-HIT
Genome	Complete human genome sequence
Annotation	Publicly available genome annotations such as NCBI genome, Ensembl and Goldenpath
PCF	Putative coding fragments, translated peptides from human genome
PCF-Profile	Sequence profiles for PCF
EST	Public EST database
PDB-Profile	Sequence profiles for PDB sequences
Pfam-Profile	Sequence profiles for Pfam database
Other	Sequence profiles calculated from other public databases

Table 1: A list of algorithms and resources.

3 Applications.

We have used this system to identify several novel protein families such as putative apoptotic proteins and kinases.

References

[1] Li, W. and Godzik A. 2002. Discovering new genes with advanced homology detection. Trends in Biotechnology. 20:315-316.

[2] Li, W., Jaroszewski, L. and Godzik A. 2002. Database clustering strategies improve PSI-BLAST remote homology recognition while cutting down on search time. Protein Engineering, 15:643.

[3] Li, W., Jaroszewski, L. and Godzik A. 2002. Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics, 18:77-82

[4] Li, W., Pio, F., Pawlowski, K and Godzik A. 2000. Saturated BLAST: An automated multiple intermediate sequence search used to detect distant homology. Bioinformatics, 16:1105-1110

[5] Rychlewski, L., Jaroszewski, L., Li, W. and Godzik A. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Science, 9:232-241

I21. Clustering protein sequences with a novel metric transformed from sequence similarity scores and sequence alignments with neural networks

Qicheng Ma, N. R. Nirmala, Gung-wei Chirn, Richard Cai [1]

Keywords: comparative genomics, neural networks, protein sequence clustering

1 Introduction.

Clustering of protein sequences from different organisms has been used to identify orthologous and paralogous protein sequences, to find protein sequences unique to an organism, and to derive the phylogenetic profile for a cluster of protein sequences. These are some of the essential components of a comparative genomics study of protein sequences across several genomes.

Algorithms used to cluster protein sequences can be either domain-based or family-based. All the clustering methods start with an all-against-all pairwise protein sequence similarity searches. The domain-based clustering methods organize the protein sequence universe into domain clusters where domains are the structural units of proteins, e.g., COG [1]. Family-based clustering methods group protein sequences into families, which contain a group of evolutionarily related proteins that share similar domain architecture, e.g., PROTONET [2].

We propose a novel family-based clustering method to address two problems: how to detect whether two aligned sequences have similar domain structures; and how to quantify transitive homologies through intermediate sequences to detect remote homologies at the superfamily level. These two problems are simultaneously solved by a new metric for clustering of protein sequences.

2 Method

From the all against all pairwise sequence similarity searches, we extract four sets of features to represent the homology between a pair of sequences. The first two sets of input features detect the homology of two aligned sequences, the last two sets of input features test whether two aligned sequences have similar domain structures. We use neural networks to map these input features to a new metric, a probability value which scales from 0 to 1. This metric is interpreted as the likelihood that two sequences are of the same homologous superfamily.

The first input feature is the log scale of the pairwise E-value. To model the correlation between two consecutive positions in the alignment, we use the 2-gram encoding [3] of the aligned regions as the second input. Intuitively, if two aligned sequences have similar domain structures, the alignment will divide the two aligned sequences in similar proportions. Thus the third input measures how similarly the two aligned sequences are cut by their aligned regions. Furthermore, the more similar domain structure two aligned sequences have, the more similar neighbour sets two aligned sequences have. The last input feature is to measure the overlap of the two neighbor sets of two aligned sequences.

[1]Life Science Informatics, Functional Genomics Area, Novartis Institute of Biomedical Research Inc, 100 Technology Squere, Cambridge, MA 02139, USA. E-mail: {qicheng.ma,nanguneri.nirmala,gung-wei.chirn, richard.cai}@pharma.novartis.com

After we represent the sequence homology between a pair of sequences by a set of input features, we can train the neural network. Each homologous pair of sequences in the training dataset is labelled as 1 if it belongs to the same Interpro [4] superfamily or the same domain if they are single domain proteins, and 0 otherwise. The neural network we use is fully connected feed-forward back propagation neural network and has one hidden layer with sigmoid activation functions. The network is trained with the scaled conjugate gradient algorithm implemented in MATLAB, and has a specificity of 94.18% and a sensitivity of 91.81%.

To take advantage of the transitive homology between sequence A and C through the third intermediate sequence B, we calculate the product of the metric score for A & B, $P(A,B)$ and the metric score for B & C, $P(B,C)$. If the metric score between A and C is smaller than $P(A,B)P(B,C)$, it is replaced by $P(A,B)P(B,C)$. Then the hierarchical average linkage clustering method is applied to clustering of the protein sequences in the new metric space using the geometric mean of the metric value as the merging rule.

3 Results.

The benchmark data set consists of all Swissprot sequences which satisfied the following criteria. One criterion is that the Interpro annotation for the sequence is consistent, e.g., the same superfamily or domain assignment in at least two member databases which include PROSITE, Pfam, SMART, TIGRFAM, and PRINTS. In addition, the alignment of the sequence with respect to either a hidden Markov model or a profile is at least 30 amino acids long. 41480 Swissprot sequences satisfied these criteria.

We evaluated the performance measure at different threshold values. At the metric score threshold of 0.5, 2073 clusters are generated with specificity 98.7%, sensitivity 99.1%, goodness is 72.6%. There were 59 orphan clusters which can not be mapped to the corresponding Interpro family or domain. Details of the clustering results with regard to specific families together with an analysis of false positives and false negatives will shed light on the strength and weakness of our clustering algorithm, and will be presented in the poster.

4 Conclusion.

This poster describes a novel clustering method of protein sequences into families based on the new metric derived from the prediction by neural networks and further utilizing the metric to model the transitive sequence homologue to detect the remote homologue. Good performance with respect to the Interpro protein sequence database has been achieved on the benchmarking dataset.

References

[1] Tatusov R. L. , Koonin E. V. and Lipman D. J. 1997. A genomic perspective on protein families. *Science* 278(5338), 631-637.

[2] Sasson O., Linial N. and Linial M. 2002. The metric space of proteins-comparative study of clustering algorithms. *Bioinformatics, Suppl 1.* S14-21.

[3] Wang T.J., Ma Q., Shasha D. and Wu C. 2001. New techniques for extracting features from protein sequences. *IBM Systems Journal, Special Issue on Deep Computing for the Life Sciences.* 40(2), 426-441.

[4] Zdobnov E.M. and Apweiler R. 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics.* 17(9), 847-848.

I23. A Tool for Constructing EST Splice Graphs and Consensus Sequence Assembly

Ketil Malde[1], Eivind Coward[2], and Inge Jonassen[3]

Keywords: EST clustering, splice graphs, consensus sequence assembly

1 Introduction

ESTs provide an abundant and quickly growing source of genetic data, and devising efficient algorithms and tools for analysing EST data remains an important challenge for the field of Bioinformatics. We present a tool for constructing *splice graphs* from EST clusters, both for a visual rendering of the structure of cluster, and for fast assembly of high-quality consensus sequences representing the different splice variants of the gene.

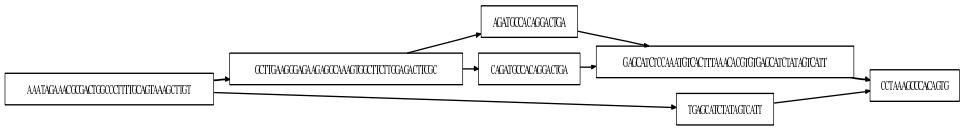


Figure 1: Visualization of a simple splice graph.

A *splice graph* [1] is an innovative way to present an EST cluster (see Figure 1 for an example). Originally based on ideas developed for "sequencing by hybridization", the splice graph is a graph where, ideally, each node in the graph represents an exon, and each edge in the graph represents a possible concantenation of exons in one or more splice variants from the gene.

It is also possible to efficiently construct consensus sequences from splice graphs. There already exist graph based methods for DNA assembly [5], but due to the different natures of mRNA and DNA, we need a slightly different approach.

2 Algorithm and Implementation

Since we do not have any a priori knowledge of the exons in the gene, we start by letting every $n-1$-word in the data set represent a potential exon. To simplify the graph, we then collapse edges where the source node has outdegree one and the destination has indegree one.

The graph is stored as dictionary of the words constituting the edges, with the nodes implicitly defined from the edges. "Lightweight" edges (with little support from the data) can then be filtered out, before visualization is done using GraphViz [6].

For consensus sequence generation, the graph is traversed using a greedy heuristic that tries at each branching point to follow the branch being supported by the most sequences consistent with previous branches.

[1]Department of Informatics, University of Bergen, Norway. E-mail: ketil@ii.uib.no
[2]Department of Informatics, University of Bergen, Norway. E-mail: coward@ii.uib.no
[3]Computational Biology Unit, BCCS, University of Bergen, Norway. E-mail: inge@ii.uib.no

3 Results

Both graph visualization and consensus sequence construction is very fast, on an 800MHz Sun Fire 880, a relatively large cluster consisting of 1180 sequences was assembled in less than three minutes. For comparison, Phrap, which is usually considered a fast assembler, used thirteen minutes on the same data set.

When comparing the results to six mRNAs from the same gene, our tool consistently produced contigs that more closely resembled the mRNAs, with BLAST alignment scores ranging from 2500 to 3600 bits, while Phrap's contigs were in the range 1500 to 1900.

The software is freely available [8].

4 Acknowledgements

This work was funded by the Norwegian Salmon Genome Project [7], a Norwegian Research Council program.

5 References and bibliography.

References

[1] Heber, S., et al. 2002. Splicing graphs and EST assembly problem *Bioinformatics* pp. S181-S188

[2] Malde, K. et al. 2003. Fast sequence clustering using a suffix array algorithm *Bioinformatics* vol. 19 no. 10, pp. 1221–1226.

[3] Mironov, A. et al. 1999. Frequent alternative splicing of human genes. *Genome Research* 9:1288–1293.

[4] Modrek, B. and Lee, C. 2001. A Genomic View of Alternative Splicing. *Nature Genetics* 30:13–19.

[5] Pevzner, P. A., Tang, H., and Waterman M. A. 2001. An Eulerian path approach to DNA fragment assembly *PNAS*

[6] The GraphViz web site. `http://www.research.att.com/sw/tools/graphviz/`

[7] The Salmon Genome Project. `http://www.salmongenome.no/`

[8] Further information and program downloads. `http://www.ii.uib.no/~ketil/bioinformatics`

I24. Evolutionary Analysis of Enzymatic Functions

Elizabeth Marland Glass[1], Tanuja Bompada[1], Jason C. Ting[2], Barnett Glickfeld[2], Natalia Maltsev[1]

Keywords: enzymes, knowledge base, phylogeny, database, evolutionary mechanisms

Evolutionary analysis of diverse sets of organisms is essential for understanding mechanisms of their adaptation to environments. Common ancestry of eukaryotes, prokaryotes and archaeobacteria leads to similarity of many molecular functions. However, differences in organisms' structural complexity, physiology, and lifestyle result in divergent evolution and emergence of variants of molecular function, metabolic organization, and phenotypic features. Recent progress in genomics, bioinformatics and physiological studies now allows for systematic exploration of adaptive mechanisms that led to diversification of biological systems. Such adaptive changes usually are not limited to one component of the system, on the contrary, in the process of adaptation, organisms undergo co-adaptive changes, such as the complementary changes of protein sequences to accommodate changes in an enzyme's active site or co-evolution of properties of different steps in metabolic pathways. Therefore, the development of a scientific framework that will allow studying evolution of the functional processes (e.g. metabolism, signal transduction) and variations in phenotypic properties is essential for interpretation of observations accumulated by molecular evolution studies.

We have developed an *enzymatic knowledge base* (EKB) that contains information regarding genetic variations for each enzymatic function, including variations characteristic for particular phylogenetic neighborhoods, phenotypic versions, etc. When possible we suggest evolutionary mechanisms involved in emergence of such variations (e.g. divergent evolution, convergent, horizontal transfer). EKB also provides an environment for interactive evolutionary analysis of the enzymes and relevant protein families by the high resolution bioinformatics tools (e.g. PhyloBlocks [1], SVMMER [2]) developed by us. A poster will present EKB and demonstrate its usefulness for evolutionary study of metabolism and high-throughput genetic sequence analysis on an example of analysis of Alcohol dehydrogenases and evolution of Alcohol fermentation pathway.

References

[1] http://compbio.mcs.anl.gov/ulrich/phyloblock

[2] http://compbio.mcs.anl.gov/svmmer

[1] Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Ave., Argonne, Illionis, US. E-mail: marland@mcs.anl.gov, tanuja@mcs.anl.gov, maltsev@mcs.anl.gov

[2] Department of Computer Science, Northern Illinois University, DeKalb, Illionis, US. E-mail: jting@mcs.anl.gov, glickfel@cs.niu.edu

I25. Molecular characterisation of a versatile peroxidase from a novel *Bjerkandera* strain.

Patrícia R. Moreira [1,2*], C. Duez [3], A. Antunes [4], E. Almeida-Vara[1], F. Xavier Malcata[2], & J. Cardoso Duarte[1]

Keywords: Ligninolytic peroxidases, White-rot fungi, Versatile peroxidase, Bjerkandera, Molecular biology

1 Introduction.

A novel class of ligninolytic peroxidases, with high affinity for manganese and dyes, has recently been described; these enzymes can also oxidise 2,6-dimethoxyphenol (DMP) and veratryl alcohol (VA) in a manganese-independent reaction. Until now, they have only been isolated from *Pleurotus ostreatus, Pleurotus eryngii, Bjerkandera adusta* and *Bjerkandera* sp. strain BOS55 [1-3]. Purified MnP1 and MnP2 from *B. adusta*, as well as MnPL1 and MnPL2 from *P. eryngii* are similar with respect to their ability to decolourise several dyes in manganese-independent reactions, as well as in manganese-independent oxidations of DMP and veratryl alcohol.

The recently sequenced enzymes MnPL1 and MnPL2 from cultures of *P. eryngii* exhibit high sequence and structural homologies with LiP and with MnP from *Phanerochaete chrysosporium*, but their molecular models show a putative manganese interaction site near the internal propionate of heme that accounts for their ability to oxidise low concentrations of this cation [2-3].

A novel fungal strain was identified in our laboratory as belonging to the *Bjerkandera* genus, and tentatively named *Bjerkandera* sp. strain B33/3. The objectives of the present study were to characterise the versatile peroxidase from this novel strain, both at sequence and structural levels.

2 Materials and Methods.

With the propose of cloning the gene corresponding to the cDNA obtained that codifies for the RBP versatile peroxidase, the oligonucleotides F-Peroxi (forward) and a primer designed from 3' terminus of the cDNA sequence obtained previously Cterm6 (reverse) 5'-CACAATTCCTACGACGACGCCTTATTCCCTCC-3' (Eurogentec) where used. Since the primer F-Peroxi encodes seven amino acids beginning at the third amino acid at the N-terminus of the mature protein, the amplified fragment was devoid of the 5'end of the gene and in particular of the leader sequence.

* E:mail: pmoreira@mail.esb.ucp.pt
[1] Departamento de Biotecnologia, Instituto Nacional de Engenharia e Tecnologia Industrial (INETI), Est. Paço do Lumiar, 22, P-1649-038 Lisboa, Portugal.
[2] Escola Superior de Biotecnologia, Universidade Católica Portuguesa, R. Dr. António Bernardino de Almeida, P-4200-072 Porto, Portugal.
[3] Centre d'Ingénierie des Protéines, Université de Liège, Institut de Chimie, B6, Sart Tilman, B-4000 Liège, Belgique.
[4] Département de Chimie Générale et Physique. Laboratoire de Spectrométrie de Masse, Université De Liège, B6c, Sart-Tilman, B-4000 Liège, Belgique.

In order to obtain it an oligonucleotide Nterm2 5'-GCACTTCTTCGCCGCACTCGCCGCCGTC-3' (Eurogentec), directed towards the 5'end of the gene, located near the N terminal of the RBP protein coding region, was combined with an hexamer (random priming) in a PCR reaction with the total genomic DNA. An oligonucleotide PS1 5'-ATGGCCTTCAAGCAACTCCTCACTG-3' (Eurogenetec), coding for the first eight residues of a sequence giving good homology with signal peptides of ligninolytic peroxidases, was therefore synthesized and used in PCR reactions. The primers 6c 5'-CGGCGCCTGCGAATTGAATACTGTTTGTGTG-3' (Eurogenetec) or 6g 5'-GATCATGATAGACCCGTCGGCACC-3' located within the non-coding region of gene sequence previously obtained where used as downstream primers. Using the genomic DNA extracted as a template and Ps1, 6c or 6g oligonucleotides as primers, several DNA fragments were amplified by PCR.

To get a better specificity of PCR reactions, a final oligonucleotide 6h 5'-CGTTGTTGGCGTGGAAGTTGGGCTCGATGTCGTC-3' was designed from the nucleic acid coding sequence for DDIEPNFHANN, a relatively conserved region among ligninolytic peroxidases where the proline residue was unique and characteristic of the cDNA of the RBP previously obtained. The pair of oligonucleotides 6h (complementary to the sequence encoding the above peptide) and PS1 allowed amplifying the expected signal peptide.

3 Results and Discussion.

A 1625 bp fragment was amplified and cloned. Six different templates were completely sequenced on both strands, using the internal specific oligonucleotides 6a to 6f as primers. The Bestfit program (GCG software) clearly identified the correspondance between the amplified fragment and the cDNA and allowed the accurate localization of the introns interrupting the ORF encoding the mature protein.

Sequence and structure function studies of the cloned gene where made, revelling high sequence and structural homologies with both LiP and MnP from different white rot fungal strains.

4 References.

[1] Giardina P, Palmieri G, Fontanella B, Rivieccio V, Sannia G 2000. Manganese peroxidase isoenzymes produced by Pleurotus ostreatus grown on wood sawdust. *Arch Biochem Biophys*; 376(1):171-79.

[2] Martinez AT. 2002 Molecular biology and structure-function of lignin- degrading heme peroxidases. *Enzyme and Microbial Technology*; 30(4):425-44.

[3] Mester T, Field JA 1998. Characterization of a novel manganese peroxidase-lignin peroxidase hybrid isozyme produced by Bjerkandera species strain BOS55 in the absence of manganese. *J Biol Chem*; 273(25):15412-17.

I27. Primer Designer for Site-Directed Mutagenesis

Alexey Novoradovsky[1], Vivian Zhang[2], Madhushree Ghosh[2], Holly Hogrefe[2], William Detrich[2], Joseph A. Sorge[2], Terry Gaasterland[1].

Engineering mutations within cloned DNA fragments involves the design of primer-template duplexes containing the target mutation. These primers have single or multiple base mismatches, or in the case of a deletion or an insertion, single-stranded DNA loops. Such DNA duplexes have a higher free energy than a mismatch-free perfect duplex. Our program suggests the most energy-saving, and thus the most effective, base substitutions to generate single or multiple amino acid changes, frame-shifts, deletions, or insertions within the target DNA molecules. Preference in codon replacement is given to the codon changes that involve fewer nucleotide substitutions and are more energetically favorable. The energy "c ost" is calculated as a difference between the summary stacking and nearest neighbor energies of the perfect primer-template duplex compared to the duplex with mismatches. In addition to free energy conservation principles, the software incorporates several empirical rules for nucleotide changes, which were established through mutagenesis experiments using degenerate primers. The program is written in PHP and is accessible as a free web service in the Stratagene web site: http://labtools.stratagene.com.

[1]Scripps Institution of Oceanography Center for Marine Genomics, 9500 Gilman Dr., La Jolla, CA 92093
[2]Stratagene, 11011 North Torrey Pines Rd., La Jolla, CA 92037

I28. PLOC: Analysis of features for protein's subcellular localization prediction

Keun-Joon Park and Paul Horton[1]

Keywords: subcellular localization, amino acid composition, SVMs, sequence analysis

1 Introduction.

To understand the functions of proteins especially in genome sequencing projects, it is desirable to obtain a protein's subcellular locations automatically from its protein sequence. Park and Kanehisa have developed a prediction method PLOC (Protein LOCalization prediction) using the compositions of amino acid and (gapped) amino acid pairs by support vector machines [1]. PLOC is available at http://www.genome.ad.jp/SIT/ploc.html. In this research, we analyzed the relationship between prediction rate and feature subgroup selection in PLOC. We also have investigated some new features, such as the biological features from PSORT2 [2] (http://psort.ims.u-tokyo.ac.jp/).

2 Method.

We considered 12 subcellular locations in eukaryotic cells (maximum case): chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracellular medium, Golgi apparatus, lysosome, mitochondrion, nucleus, peroxisome, plasma membrane, and vacuole. For the construction of our dataset, protein sequences were collected from the SWISS-PROT database. In the SWISS-PROT database, we checked keyword information about subcellular locations in the CC field, and also checked the OC field to remove prokaryotic proteins from the dataset. From the protein sequence data set, a set of Support Vector Machines (SVMs) for each subcellular location was trained based on its amino acid, amino acid pair, and from one to three gapped amino acid pair compositions. The case of 12 subcellular locations, 12 SVMs were prepared using five different kinds of composition data. The feature vector contains 20 elements for the amino acid composition, and 400 coordinates for the four kinds of amino acid pair compositions. The prediction methods based on these five different compositions information were then combined using a voting scheme.

The prediction performance was examined by the five-fold cross-validation test, in which the data set was divided into five subsets of approximately equal size (Table 1). In order to assess the accuracy of prediction methods we use two measures, the total accuracy (TA) and the location accuracy (LA) defined by:

$$TA = \frac{\sum_{i=1}^{k} T_i}{N}, \qquad LA = \frac{\sum_{i=1}^{k} P_i}{k},$$

where:

$$P_i = \frac{T_i}{n_i}.$$

[1] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Aomi Frontier Bldg. 17F, 2-43 Aomi, Koto-ku, Tokyo 135-0064, Japan, E-mail: park-kj@aist.go.jp, horton-p@aist.go.jp

Here N is the total number of proteins in the data set, k is the number of subcellular locations, n_i is the number of proteins in each location i, and T_i is the number of correctly predicted proteins in each location i. In this method, the kernel of SVMs is RBF (Radial Basis Function). We also tested for more realistic repertoires of subcellular locations in different cell types, 11 subcellular locations excluding lysosome for a plant cell, 10 locations excluding chloroplast and lysosome for a fungal cell, and 10 locations excluding chloroplast and vacuole for an animal cell. Note that vacuoles in fungi or plants are thought to correspond to lysosomes in animals.

To investigate the importance of each amino acid pair compositions, forward feature selection analysis was done with the PLOC dataset. The forward selection procedure starts from the evaluation of each individual amino acid pair feature along with the base set of 20 amino acid composition features. The pair feature used in the best combination is then added to the base set and the procedure is repeated until no further improvement is obtained.

3 Results and Discussions.

Table 1 shows the first 10 amino acid pair features selected and each prediction rate from the forward feature selection analysis.

TA	LA	Amino acid pair features
0.727	0.569	GxxxP
0.728	0.572	PxxY
0.728	0.575	NF
0.729	0.577	PxxH
0.729	0.579	YF
0.730	0.580	GxQ
0.730	0.581	RxxY
0.732	0.581	SxxxF
0.733	0.581	LxxxP
0.734	0.582	HxP

Table 1: The first 10 informative amino acid pair compositions selected by forward subset feature selection in PLOC method.

We also acquired some informative feature set from PSORT2. PSORT2 contains 31 biological meaningful features. These features would be added as the new features to the new version of PLOC2 for yeast proteins. Further practical prediction method may be constructed by adding new subcellular locations or defining finer classifications, for example mitochondrial inner, outer membrane or matrix protein groups. And some researchers could also want prediction system for some specific locations only. We have to gather and consider these various needs from the users. Perhaps improvements of prediction rate can be obtained using additional new feature vectors for training of SVMs.

References

[1] Park, K. -J. and Kanehisa, M. 2003. Prediction of proteins subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics* 19:1656-1663.

[2] Nakai, K. and Horton, P. 1999. PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24:34-35.

I29. An Evolutionary Computation Approach for Detecting Repetitions in Biosequences

Adam Adamopoulos, [1] **Katerina Perdikuri,** [2]

Keywords: biological sequences, repetitions, evolutionary programming, genetic algorithms

1 Introduction.

One of the most important goals in computational molecular biology is allocating repeated patterns in nucleic or protein sequences, and identifying structural or functional motifs that are common to a set of such sequences. In this paper we introduce a new approach to detect the repetitions of fixed length in Biosequences using an Evolutionaty Computation approach. Our approach involves evolving a population of patterns in an evolutionary manner and gradually improving the fitness of the population as measured by an objective function, which measures the approximate repetitions of the patterns in the given sequence. The general attraction of the approach is the ability to detect repeated schemas, thus inferring motifs of fixed length from biosequences. Genetic Algorithms and Evolutionary Computation have been succesfully applied so far in the Multiple Molecular Sequence Alignment problem in order to identify similarities among sequences [1].

2 Methodology.

Biosequences, such as DNA and Protein sequences, can be seen as long texts over specific alphabets, encoding the genetic code of living beings. Searching for repeated sub-sequences of any length over those texts could be modeled as searching for a set of given patterns in a "text".

In our approach we consider the population of a Genetic Algorithm as the population of p words of length l. For each particular run, the population size p (i.e. the number of individuals) of each generation, as well as the length l of each one word of the population are kept constant. After establishing a population of words the population is randomly initialized. When the initialization procedure is completed all words of the population are random strings drawn from the Σ_{DNA} alphabet. The fitness f of each particular word is evaluated considering as fitness (or evaluation) function the number of approximate occurrences (repetitions) of the word in the input sequence.

The overall structure of the method is shown in Figure 1. To go from one generation to the next, children are derived from parents that are chosen by some kind of natural selection. To create a child, an operator is selected that can be a crossover (mixing the contents of the two parents) or a mutation (modifying a single parent). Each operator has a probability of being chosen. Thus the algorithm is divided in two stages. The first one is the evolutionary phase where the new population of individuals/words is generated and the searching phase where each individual is evaluated by counting its number and exact positions of occurrences.

Compared to other techniques ([2], [3]) our algorithm is linear to the length of the input sequence and has the advantage of allowing the user to specify the exact length of

[1]Laboratory of Medical Physics, Department of Medicine, Democritus University of Thrace, Alexandroupolis, Greece. E-mail: adam@med.duth.gr

[2]Research Academic Computer Technology Institute, 61 Riga Feraiou Street, 26221 Patras, Greece. E-mail: perdikur@ceid.upatras.gr

the repetitions the biologist looks for. Taking into consideration the easy parallelisation of Genetic Algorithms we believe that our method can be used in many practical applications. Moreover Genetic Algorithms could be successfully used as a practical way to solve many computationally difficult problems in the areas of Sequence Search and Alignment. They are intellectually satisfying in their simplicity and the way they attempt to mimic biological evolution.

FIND REPETITIONS(X,l, p, n, p_m, p_c, $elitism$)
Initialize population of words
WHILE $n \geq 1$, **DO**
  Evaluate-Fitness: compute the repetitions of each word of the population;
  Produce Next Generation: compute the next generation;
    If $elitism \neq 0$, perform elitism;
    If $p_m \leq const$, perform mutation;
    If $p_c \leq const$, perform uniform crossover;
  Report individuals in descending order
END FIND REPETITIONS

Figure 1: Schematic View of the Genetic Algorithm Methodology

3 Conclusions and Future Work.

Our method efficiently computes the repetitions inside a biosequence by evolving a population of repeated patterns in an evolutionary manner (mutation and crossover) and finally reporting those with high fitness function. Our future work is three fold. The first one concerns the modification of the algorithm by assigning a credit to the operators of mutation and crossover. Thus, each time a new individual is generated, if it yields some improvement over its parents, the operator that was directly responsible for its creation gets the largest part of the credit and so in the new generation we can dynamically change the probability of the mutation or crossover operator. This can reduce the time complexity needed to compute the mutation and crossover operation for the population in each generation. The second research direction concerns the addition of one operator responsible for inserting gaps inside repeated patterns thus giving the possibility of inferring structured patterns from the input biosequence. Finally an interesting problem arises from having "don't care symbols" in the input sequence [4]. A "don't care" symbol has the property of matching any symbol of a given alphabet. We believe that our approach can efficiently compute the repetitions even in biosequences with "don't cares".

References

[1] Zhang, C., Wong, A.K. 1997. A genetic algorithm for multiple molecular sequence alignment. *Comput. Appl. Biosci.*, Vol. 13. (1997) pp. 565-581.

[2] Kurtz, S., Schleiermacher, C. 1999. REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, Vol. 15, (1999) pp. 426-427.

[3] Tsunoda, T., Fukagawa, M. Takagi, M.T. 1999. Time and memory efficient algorithm for extracting palindromic and repetitive subsequences in nucleic acid sequences. In *Proceedings of the Pacific Symposium on Biocomputing*, Vol. 4, (1999) pp. 202-213.

[4] Iliopoulos, C., Mohamed, M., Mouchard, L., Perdikuri, K., Smyth, W.F., Tsakalidis, A. 2003. String Regularities with Don't Cares. *Nordic Journal of Computing*, Vol. 10 (2003) pp. 40-51.

I31. Euclidean Distance Measure of Markov Models for Genome Comparison Without Alignment

Tuan Pham, [1] **James O'Connell,** [2] **Johannes Zuegg,** [3]

Keywords: Sequence comparison, alignment-free, Markov models

1 Introduction.

Comparison between sequences is a key step in bioinformatics [2] when analysing similarities of functions and properties of different sequences. We discuss herewith a stochastic method for modeling biological sequences and its Euclidean distance measure for comparison without the need of alignment.

2 Euclideance distance of Markov models.

Let $A = [a_{ij}]$ denote the state transition probability matrix of a discrete, first-order Markov process. Each state transition probability a_{ij} is defined as

$$a_{ij} = P[q_t = S_j | q_{t-1} = S_i],\ 1 \leq i, j \leq N \tag{1}$$

where q_t stands for the actual state at time t, S_j a state j of a set of N distinct states.

We make use of the 20 amino acids and the stop-codon group to specifiy N=441 states that come from the different combinations of pair-wise sets of the amino acids and the stop codons ($21^2 = 441$). To construct the transition matrix A, we apply circular shifting of every single nucleotide in the sequence and count the occurances of the transitions.

Let $\pi = \{\pi_i\}$ be the initial state distribution where

$$\pi_i = P(q_1 = S_i),\ 1 \leq i \leq N \tag{2}$$

Let $\lambda_1 = (A_1, \pi_1)$, and $\lambda_2 = (A_2, \pi_2)$ be two Markov models of the two sequences, and assuming equal initial probability distribution, we can measure the similarity between two sequences by the Euclidean distance, denoted by $d_E(\lambda_1, \lambda_2)$, and defined as follows:

$$d_E(\lambda_1, \lambda_2) = \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} (a_{ij}^1 - a_{ij}^2)^2 \right)^{\frac{1}{2}} \tag{3}$$

where a_{ij}^1 and a_{ij}^2 are the elements of A_1 and A_2, respectively.

[1]School of Computing and Information Technology, Griffith University, Nathan Campus, QLD 4111, Australia. E-mail: t.pham@griffith.edu.au

[2]School of Computing and Information Technology, Griffith University, Nathan Campus, QLD 4111, Australia. E-mail: james.oconnell@student.gu.edu.au

[3]Research Computing Services, Griffith University, Nathan Campus, QLD 4111, Australia. E-mail: j.zuegg@griffith.edu.au

3 Results.

We test our proposed method with 33 complete mamalian mtDNA sequences available in a public database (http://megasun.bch.umontreal.ca/ogmp/projects/other/mt_list.html). The distance matrix derived from these 33 mtDNA genomes is used to study the relationship between the species and their groups. Figure 1 shows the phylogenetic tree of these genomes using the PHYLIP package [1]. Overall, the method has shown reasonable relationships between these species and their groups.

4 Concluding Remarks.

We have applied a simple distance measure for comparison between Markov models of 33 complete mtDNA genomes and obtained reasonable result. Investigation into more robust distance measures for stochastic models will be useful in this study.

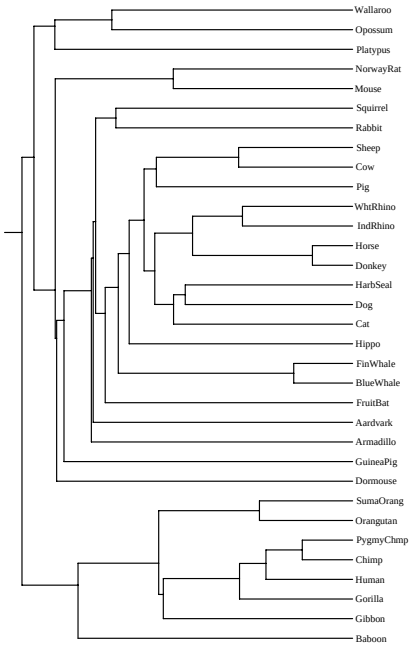


Figure 1. Phylogenetic tree constructed from complete mtDNA for 33 species

References

[1] Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

[2] Miller, W. 2001. Comparison of genomic DNA sequences: solved and unsolved problems, *Bioinformatics*, **17**, 391-397.

I32. HMMERHEAD - Accelerating HMM Searches On Large Databases

Elon Portugaly[1] **Matan Ninio**[1]

Keywords: HMM, acceleration, sequence, search, Pfam, SWISS-PROT, BLAST

1 Introduction

HMMs have been proven useful in protein sequence analysis [1]. However, a full search of a sequence database using an HMM is a computationally expensive process - running all the Pfam [3] HMMs on the SWISS-PROT database [4] takes almost three months of computer time. The two-hit method used by Altschul et al [2] allows BLAST to accelerate both sequence vs. sequence searches and profile vs. sequence searches. In this work we build a framework that uses a similar method for HMM searches.

We provide *HMMER Hashing Enabled Acceleration Device* (HMMERHEAD) - a software package that filters out sequences for hmmsearch. Our experiments show that we typically achieve a 15-fold acceleration of running time, while retaining 99% of the results.

2 The Two-Hit Method

The two-hit method was introduced in [2]. Following is a short description of the method.

Preprocessing: In a preprocessing step, a database of k-mers (i.e. words of size k over the alphabet used) is compiled from the sequence database. For each possible k-mer this database provides quick access to the list of all occurrences of the k-mer in the sequence database. This database is fixed for all queries, and need only be computed once for each sequence database.

Queries: When presented with a query, the algorithm finds all k-mers that locally match any part of the query with a local score above some given threshold. Next, the k-mer database is queried for the occurrences of each of the above k-mers in the target sequences. Each such occurrence resides within a *diagonal* path in the alignment graph of the sequence and the query. If two such occurrences share a diagonal, and are within a fixed distance from each other, the target sequence is reported as a candidate for dynamic programming search.

3 HMMER and HMMERHEAD

We have implemented the two-hit-method as a filter stage for HMMER [1]. HMMER is a software package that implements HMMs for families of protein and DNA sequences. Given a query HMM and a sequence database, HMMERHEAD filters the sequence database using the two-hit-method, and pipes the sequences that passed the filter to HMMER's hmmsearch program for the final search. HMMERHEAD accepts HMMER HMMs files as input. The HMMERHEAD package has been tested on Linux and is provided under the GNU license at `http://www.cs.huji.ac.il/labs/compbio/hmmerhead`.

[1]School of Computer Science & Engineering, Hebrew University, Jerusalem 91904, Israel.
E-Mail: {elonp,ninio}@cs.huji.ac.il

Filtering[1] (%)	90	95	97.5	99	99.5
Average recall[2]	99.4	99.3	99.0	90.6	88.3
Speedup factor[3]	5.5	8.4	15.2	34.4	52.5
Recall[4]					
Above 99%	93.6	92.0	89.0	82.8	77.2
99%-95%	4.3	5.4	5.6	4.0	4.0
95%-90%	1.6	2.1	3.2	4.0	5.4
Below 90%	0.5	0.5	2.1	9.1	13.4

[1]percent of sequences that are filtered out
[2]percent of total required matches that survive filtration
[3]decrease factor in total cpu time (user+system)
[4]Percent of HMMs for which recall is within range

Table 1. Recall and speedup by filtering

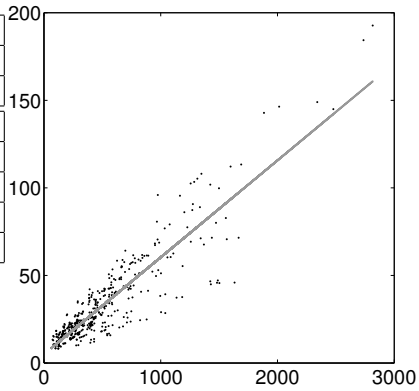


Figure 1. Running times

4 Results

HMMERHEAD performance is evaluated by two measures: *recall* - the percentage of true positives that survive the filtering, and *filtering* - the percent of sequences that pass the filter. We also measure machine running time directly. All runs where performed with k-mers of size 4 and a two-hit window of size 25.

We tested HMMERHEAD performance by searching for Pfam family members in the SWISS-PROT (rel. 41.21) database [4].

Each Pfam family is defined by two HMMs and two corresponding cutoff scores. We randomly chose about 5% of the Pfam families, and collected all 476 related HMMs. We searched the SWISS-PROT database with each of the HMMs, first with HMMER, and then with HMMERHEAD using several different levels of filtering. The results are shown in Table 1. Note that when filtering 97.5% of the sequences, we achieve a speedup factor of more then 15, and loose only 1% of the required matches. The bottom part of the table shows the performance over the different HMM profiles - here recall is computed separately for each family. Figure 1 shows the running times, in cpu seconds, of the different HMMs. X-axis – search without HMMERHEAD; Y-axis – HMMERHEAD filtration with 97.5%. A least squares linear fit sets $f = 5.028 + 0.0553 \times u$ for f filtered time and u unfiltered time.

References

[1] Eddy, S. R. 2001. HMMER: Profile hidden Markov models for biological sequence analysis (`http://hmmer.wustl.edu/`)

[2] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. in *Nucleic Acids Res* 25, 3389-3402.

[3] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy S. R., Griffiths-Jones S, Howe K. L., Marshall M, and Sonnhammer E. L. 2002. The Pfam protein families database. In *Nucleic Acids Research* 30, 276-280

[4] Boeckmann B., Bairoch A., Apweiler R., Blatter M. -C., Estreicher A., Gasteiger E., Martin M. J., Michoud K., O'Donovan C., Phan I., and Pilbout S., Schneider M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL. In *Nucleic Acids Res.* 31, 365-370.

I33. Understanding RNA Pseudoknotted Structures

Anne Condon, Beth Davy, Baharak Rastegari, Shelly Zhao, [1] Finbarr Tarrant [2]

Keywords: RNA secondary structure prediction, RNA pseudoknots

1 Introduction

RNA molecules play diverse roles in the cell. The structure of RNA molecules is often the key to their function and predicting the secondary structure of RNA molecules is thus an important problem. Since the general problem of predicting pseudoknotted secondary structures is NP-hard, several algorithms have been proposed that find the minimum free energy (mfe) secondary structure from a restricted class of secondary structures: Rivas and Eddy (R&E) [7], Dirks and Pierce (D&P) [5], Lyngso and Pederson (L&P) [6] and Akutsu and Uemura (A&U) [1, 8] (two algorithms with the same class). In this work, we order the classes by the generality of the structures that they handle and the result is as follows (PKF stands for pseudoknotted free structures): PKF $\subseteq$ L&P $\subset$ D&P $\subset$ A&U $\subset$ R&E. The classes of structures that some of these algorithms can handle is defined implicitly by the lengthy recurrences in the algorithm. We provide simple, concise characterizations of the classes of structures handled by R&E, D&P and A&U algorithms as well as linear time methods to test whether a given secondary structure is in each class. Using our methods, we provide data on the percentage of real structures that can be handled by each algorithm.

2 Characterization of R&E class

RNA secondary structure is usually defined as a set R of base pairs $i.j$ such that each base is paired at most once. Here, we will also use an alternative *pattern* representation of secondary structures. To define patterns precisely, we introduce some notation. We use ϵ to denote the empty string. Let $\mathbb{N}_n$ denote the natural numbers between 1 and n (inclusive). A string p (of even length) over some alphabet Σ is a *secondary structure pattern*, or simply a *pattern*, if every symbol of Σ occurs either exactly twice, or not at all, in p. We say that p *is a pattern for secondary structure* R of a strand of length n if there exists a mapping $m : \mathbb{N}_n \rightarrow \Sigma \cup \{\epsilon\}$ with the following properties: (i) if $i \cdot j \in R$ then $m(i) \in \Sigma$ and $m(i) = m(j)$, (ii) if $i \cdot j \notin R$ for all $j \in \mathbb{N}_n$, then $m(i) = \epsilon$, and (iii) $p = m(1)m(2)\ldots m(n)$.

We have developed formal characterizations of the R&E, D&P and A&U classes. Here, we describe the characterizations of the R&E class. (The others are defined in a similar way)

The patterns in the R&E structure class can be defined as follows: Let p be a string over alphabet Σ. Symbol σ is *directly adjacent* to symbol τ in p if and only if either $\sigma\tau\sigma$ is a substring of p or there are two disjoint substrings x, y of p, both of length 2, such that τ and σ are both in x and τ and σ are both in y. If σ is directly adjacent to some symbol in pattern p, we say σ is directly adjacent in p. (The direct adjacency relation is not necessarily symmetric.) Let p be a pattern. We say that $p \xrightarrow[\text{R\&E}]{} p'$ if $p' = p \downarrow \sigma$ for some σ that is either self-adjacent or directly adjacent in p. ($p \downarrow \sigma$ denotes the string p with

[1]Department of Computer Science, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada. Contacting E-mail: condon@cs.ubc.ca

[2]The Kane Building, Department of Computer Science, University College Cork,College Road, Ireland

all occurrences of σ removed.) Also, $p \xrightarrow[\text{R\&E}]{*} p'$ if $p = p'$ or $\exists$ patterns $p_1, \ldots, p_k$ such that $p \xrightarrow[\text{R\&E}]{} p_1 \xrightarrow[\text{R\&E}]{} \ldots p_k \xrightarrow[\text{R\&E}]{} p'$. Pattern p is an R&E pattern if $p \xrightarrow[\text{R\&E}]{*} \epsilon$.

3 Tests and Results

We have developed linear time tests for membership in each of the R&E, D&P, and the L&P classes from PseudoBase (PBase) [2], the Nucleic Acids Database (NDB) [3], 16S and 23S ribosomal RNA and Group I and Group II Introns from the Gutell Database [4]. Our results,presented in Table 1, shows that the R&E structure class is indeed very general, containing all of the secondary structures except for three (long) Group II Intron sequences. Although the D&P class does not contain most of the 23S rRNA structures, it compares well with the R&E class. The L&P class additionally misses almost all of the 16S rRNA structures, yet still contains almost all of the structures in PseudoBase.

	PBase	16S	23S	Gp I Intron	Gp II Intron	NDB
# Strs	240	152	69	10	3	12
Avg. #Bps	14.1	455.6	733	126.1	207	268
PKF	0	0	21	0	0	6
L&P	231	12	21	10	0	6
D&P	232	152	21	10	0	11
R&E	240	152	69	10	0	12

Table 1: Structure classification. Columns 2-7 present data for each RNA data set. For each data set (column), the entry in the first row lists the number of structures in the data set. The second row lists the average number of base pairs in the structures. The remaining rows list the number of structures of the data set that are in the PKF, L&P, D&P, and R&E classes, respectively.

References

[1] Akutsu, T. 2000. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics* 104:45–62.

[2] Batenburg, F. H. D. van, Gultyaev, A. P., Pleij,C. W. A., Ng, J. and Oliehoek, J. 2000. Pseudobase: a database with RNA pseudoknots. *Nucl. Acids Res.* 28(1):201–204.

[3] Berman, H. M. et al. 1992. The Nucleic Acid Database: A Comprehensive Relational Database of Three-Dimensional Structures of Nucleic Acids. *Biophys. J.*, 63:751–759.

[4] Cannone J. J. et al. 2002. The Comparative RNA Web (CRW) Site: An Online Database of Comparative Sequence and Structure Information for Ribosomal, Intron, and other RNAs. BioMed Central Bioinformatics, 3:2. [Correction: BioMed Central Bioinformatics. 3:15.]

[5] Dirks, R. M. and Pierce, N. A. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.* 24(13):1664-1677.

[6] Lyngsø R. B. and Pedersen, C. N. 2000. RNA pseudoknot prediction in energy-based models. *J. Computational Biology* 7(3):409–427.

[7] Rivas E. and Eddy, S. R. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Molecular Biology* 285:2053–2068.

[8] Uemura, Y., Hasegawa, A., Kobayashi, S. and Yokomori, T. 1999. Tree adjoining grammars for RNA structure prediction. *Theoretical Computer Science* 210:277–303.

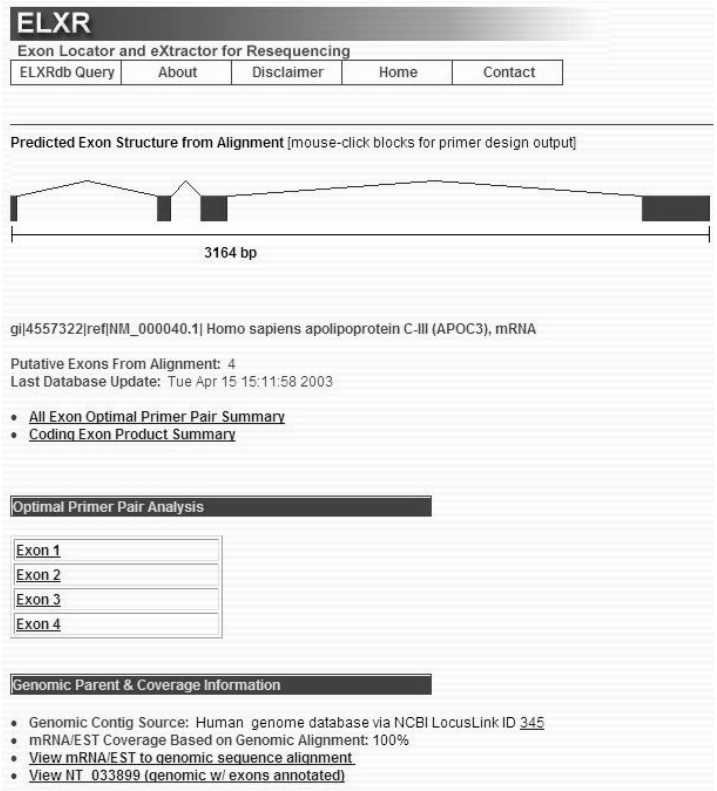
I35. ELXR: A Resource For Rapid Exon-Directed Sequence Analysis

Jeoffrey J. Schageman[1] and Alexander Pertsemlidis[2]

Keywords: Polymorphism, exon, sequence analysis

Abstract

DNA resequencing of exons followed by comparative sequence analysis is the most common method used for detecting sequence variation. Typically, this process requires determination of intron/exon boundaries, design of PCR and DNA sequencing primers, PCR amplification from genomic DNA, and subsequent sequencing of the resulting PCR product. Here we describe a bioinformatics resource called ELXR (Exon Locator and Extractor for Resequencing) that automates this tedious process. We have pre-computed ELXR primer sets for all exons identified from the entire NCBI-curated human, mouse, and rat mRNA reference sequence (RefSeq) public database. The resulting 360,000 exon-flanking PCR primer pairs with accompanying genomic and exon-specific annotations have been compiled into a queryable system called ELXRdb, which may be searched by keyword, gene name or RefSeq accession number. An example of an ELXRdb entry is depicted in figure 1. ELXR and ELXRdb available on the World Wide Web at http://elxr.swmed.edu/.



[1] Howard Hughes Medical Institute and Eugene McDermott Center for Human Growth and Development, Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, Texas 75390. E-mail: Jeff.Schageman@UTSouthwestern.edu

[2] Eugene McDermott Center for Human Growth and Development, Department of Internal Medicine, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, Texas 75390. E-mail: Alexander.Pertsemlidis @UTSouthwestern.edu

References

[1] Pruitt, K.D. and Maglott D.R. 2001. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Research* **29** (1):137-40.

[2] Rozen, S. and Skaletsky H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* **132**:365-86.

[3] Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M. and Miller, W. 1998. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research* **8**(9): 967-74.

I37. Effect of alternative splicing on structure and function of mouse transcription factors.

Bahar Taneri[1], Ben Snyder[2], Terry Gaasterland[3].

Keywords: alternative splicing, transcription factor, proteome

Abstract

Analyzing proteins in the context of all available genome and transcript sequence data has the potential to reveal functional properties not accessible through protein sequence analysis alone. We are studying the effect of alternative splicing on mouse proteins and specifically on transcription factors. We hypothesize that by creating a protein structure alteration, alternative splicing could be a critical determining factor for the specific DNA binding sites and cofactors that interact with a given transcription factor. We further hypothesize that expression data will indicate tissue specific control of changes in protein structure and function due to alternative splicing. We use MouSDB3, a database of splice variants in the mouse transcriptome (http://genomes.rockefeller.edu/MouSDB3) to study these phenomena.[1] Protein domain organizations for different splice forms are determined by using SMART (Simple Modular Architecture Research Tool- http://smart.embl-heidelberg.de).[2] Initial analyses revealed that 62% of the transcription factor loci in MouSDB3 have variant exons, compared to 29% of all loci. These variant transcription factor loci contain a total of 325 facultative exons, which are excluded in some transcripts and included in others. 24% of these facultative exons are in-frame, i.e their nucleotide number is in multiples of three, they do not introduce a stop codon when skipped and the exon starts at the first base of a codon. When excluded, in-frame facultative exons alter the domain architecture of the protein 81% of the time, as computed by SMART. 67% of these in-frame facultative exons are either fully or partially within the coding regions for motifs that are important in transcription factor function. These include helix-loop-helix, leucine zipper and homeobox domains. Our integrated genomic and proteomic approach addresses the general question of how alternative splicing affects the proteome, and gives insight into control of transcription.

[1] Laboratory of Computational Genomics, The Rockefeller University, 1230 York Avenue, New Yok, New York, 10021. E-mail: bahar@genomes.rockefeller.edu

[2] Laboratory of Computational Genomics, The Rockefeller University, 1230 York Avenue, New Yok, New York, 10021. E-mail: ben@genomes.rockefeller.edu

[3] Laboratory of Computational Genomics, The Rockefeller University, 1230 York Avenue. New Yok, New York 10021. E-mail: gaasterl@genomes.rockefeller.edu

References

[1] Zavolan, M., van Nimwegen, E. and Gaasterland, T. 2002. Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Research* 9:1377-1385.

[2] Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P. and Bork, P. 2002. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Research* 30(1):242-4.

I38. A Novel Approach for Efficient Query of Single Nucleotide Variations in DNA Databases

Hsiao Ping Lee [1], Yin Te Tsai [2], Ching Hua Shih [3], Tzu Fang Sheu [4], Chuan Yi Tang [5]

Keywords: Single Nucleotide Variations, RNA Editing, Filtration

1 Introduction.

RNA editing [2] plays a crucial role in altering the functions of the resulting protein products. *Single nucleotide variations (SNV)* among DNA sequences can serve as the potential RNA editing targets. Given a genomic database D, a query sequence Q and a segment width w, the problem of identifying SNVs is to discover all segments $b \in D$ and $q \in Q$ such that b and q have the same width w, the difference between b and q is in only one position, and the difference is a mutation of A↔G or C↔T. We refer to such b as a w-SNV of q. In the literature, there are several efficient database search algorithms have been proposed, for example [1, 3, 4]. The sorting-based method [3] is an efficient approach for all-against-all SNV detection. Nevertheless, the issue of ad hoc query is not well considered in the sorting approach. This insufficiency limits the feasibility to serve as an online web tool. In this poster, we propose a novel approach for efficient SNV identification based on some simple properties. The ad hoc query is also fully supported in the approach. By the evaluation experiments, we demonstrate the efficiency of our new approach.

2 SNVF Algorithm.

Let X be a DNA sequences of length w over Σ={A,C,G,T}. $X[i,j]$ denote the substring of X from positions i to j. An m-seed is a segment, a short substring, of length m. We call $X[1, \lfloor w/2 \rfloor]$ and $X[\lceil w/2 \rceil + 1, w]$ the *head* and *tail seeds* of X, respectively. For each seed e of X, the substring $X - e$ is called its *flanking counterpart*. Let F_X^c denote the number of occurrences of a $c \in \Sigma$ in X. Let $V_X^\Sigma = \{ F_X^{\mathtt{A}}, F_X^{\mathtt{C}}, F_X^{\mathtt{G}}, F_X^{\mathtt{T}} \}$. Let $\mathrm{DP}(V_X^\Sigma) = F_X^{\mathtt{A}} + 2^v F_X^{\mathtt{C}} + 2^{2v} F_X^{\mathtt{G}} + 2^{3v} F_X^{\mathtt{T}}$, where $v = \lceil \log_2 w \rceil$. Let $d = 2^{2v} - 1$. We call $C_\Sigma^w = (2^v + 1)d$ and $R_\Sigma^w = (2^v - 1)d$ as the *center* and *radius* of w-SNV over Σ respectively.

For a given query sequence Q and the user-defined segment width w, the *SNV finder (SNVF)* is designed to efficiently extract all w-SNVs from the genomic databases. Our strategy is first to filter out all the substrings that can not be SNV counterparts. Let S and T be two DNA sequences of length w. The filtration is based on the following observations.

Observation 1. If T is a w-SNV of S then one of the following two conditions holds: (1) the head seeds of S and T are identical and the flanking counterpart of head seed of T is a $(w - \lfloor w/2 \rfloor)$-SNV of S, or (2) the tail seeds of S and T are identical and the flanking counterpart of tail seed of T is a $(w - \lfloor w/2 \rfloor)$-SNV of S.

Observation 2. Let $d_1 = |\mathrm{DP}(V_{S[1,\lfloor w/2 \rfloor]}^\Sigma) - \mathrm{DP}(V_{T[1,\lfloor w/2 \rfloor]}^\Sigma)|$ and $d_2 = |\mathrm{DP}(V_{S[\lfloor w/2 \rfloor + 1, w]}^\Sigma)$

[1]Department of Computer Science, National Tsing-Hua University, Taiwan, ROC. E-mail: shopping@cs.nthu.edu.tw

[2]Department of Computer Science and Information Management, Providence University, Taiwan, ROC. E-mail: yttsai@pu.edu.tw

[3]Department of Life Science, National Tsing-Hua University, Taiwan, ROC. E-mail: stewardshih@yahoo.com.tw

[4]Institute of Communication Engineering, National Tsing-Hua University, Taiwan, ROC. E-mail: sunnie@totoro.cs.nthu.edu.tw

[5]Department of Computer Science, National Tsing-Hua University, Taiwan, ROC. E-mail: cytang@cs.nthu.edu.tw

$- \mathrm{DP}(V^{\Sigma}_{T[\lfloor w/2 \rfloor+1,w]})|$. If T is a w-SNV of S, $d_1 \times d_2 = 0$ and $|C^w_\Sigma - 2(d_1 + d_2)| = R^w_\Sigma$ hold.

Based on the observations, we design an efficient three-phase algorithm to find w-SNVs. In the first phase, we file all seeds of length $l = \lfloor w/2 \rfloor$ from the genomic database into a dictionary with 4^l entries. Using the dictionary, we can immediately locate identical l-segments in the database. This work can be done in the preprocess beforehand. The second phase is to extract the possible candidates of SNVs. For each w-segment Q_w in the query sequence, we generate the set G_h of all the w-segments having the head seeds same as Q_w's from the database. In addition, the set G_t of all the w-segments having the tail seeds identical to Q_w's is also established. We examine the flanking counterpart of head seed of each item within G_h by the criterion in Observation 2. The test is also performed on the flanking counterpart of tail seed of each segment in G_t. The satisfied segments are selected as SNV candidates, and passed to the next phase. We perform the SNV verification in the last phase by the comparisons between the target and candidate segments.

Our evaluation experiments are performed on a PC running Red Hat Linux release 9, equipped with one Intel P4 2.8Ghz CPU, 1GB DDR memory, and 80GB hard disk. In each test, we randomly select the specified size of query sequences from the chromosome 1 ESTs, and set the segment width to 20. The CPU time spent for each query is measured and presented in Table 1, where D_1, D_2 and D_3 are chromosome Y ESTs ($\approx$ 1.4M mers), chromosome 21 ESTs ($\approx$ 21M mers) and chromosome 18 ESTs ($\approx$ 31M mers), respectively.

Query(mers)	D_1	D_2	D_3
0.3K	<0.01	0.04	0.15
0.6K	<0.01	0.09	0.14
1.2K	<0.01	0.11	0.17
10K	0.04	0.97	4.00

Figure 1: The CPU time (seconds) in the experiments.

3 Conclusion.

To massively study genes is a hallmark of the transition from 'structural' to 'functional' genomics. With huge DNA sequence data, one of the biologists' desires is to understand the secrets of DNA sequences, the language of life. For biological applications, our method can apply to locate gene positions, SNP, alternative splicing events and cross species genomic comparison to look for orthologous genes.

References

[1] S. Burkhardt, A. Crauser, P. Ferragina, H. P. Lenhof, E. Rivals, and M. Vingron. q-Gram Based Database Searching Using a Suffix Array (QUASAR). In the proceedings of *RECOMB 1999*, pages 77–83.

[2] T. R. Cech. RNA Editing: World's Smallest Introns?. In *Cell*, Volume 64, pages 667–669, 1991.

[3] C. N. Chen, C. H. Peng, C. T. Chang, W. Y. Chow and C. Y. Tang. Identify Single Nucleotide Variation in Whole Genome Sequences by External Sorting. In the proceedings of RECOMB 2003, poster 184.

[4] H. P. Lee, Y. T. Tsai, C. Y. Tang, C. H. Shih and T. F. Sheu. A Seriate Coverage Filtration Approach for Homology Search. To appear in the proceedings of SAC 2004.

I39. Predicting Regulatory Motif based on Multiple Genome Sequences

Ting Wang, Gary D. Stormo[1]

Keywords: comparative genomics, motif discovery, ALLR statistic

1 Introduction.

Discovery of regulatory motifs is one of the fundamental problems in computational biology. Identification of all TF binding sites will provide the information necessary to eventually construct models for global networks of transcriptional regulation. In recent years, the rapid accumulation of complete genome sequences and the advance of high-throughput expression profiling technology are changing the ways that we look at genomic sequences and redefining the type of problems a motif discovery algorithm can tackle. For example, due to statistical limitations, current motif discovery tools are only applicable to a small group of sequences that are also limited in length. These sequences are often promoter sequences of some co-regulated genes identified through expression assay or chromatin immunoprecipitation assay, or promoters of orthologous genes from several genomes. However, few tools can take the advantage of both type of data and predict regulatory motifs from promoter sequences of multiple genes from several related genomes.

2 Methods and Results.

We recently developed a new algorithm called PhyloCon (Phylogenetic Consensus) that takes into account both conservation among orthologous genes and co-regulation of genes within a species. PhyloCon first aligns conserved regions of orthologous sequences into multiple sequence alignments, or profiles, then compares profiles representing non-orthologous sequences using a newly developed statistic we named "ALLR (average log likelihood ratio)". In this study, we present ALLR statistic and show its properties for sequence alignment and model comparison. We also present the basic PhyloCon algorithm and its derivatives for motif discovery based on comparative genomic data and motif comparison. We show by example that PhyloCon can accurately and efficiently identify biological meaningful motifs from relatively large sequence dataset.

[1] Department of Genetics, Washington University Medical School. 4566 Scott Avenue, Campus Box 8232, St. Louis MO 63110. E-mail: stormo@ural.wustl.edu

I41. Predictive Classification

Jialu Zhang, [1] **Françoise Seillier-Moiseiwitsch** [2]

Keywords: prequential tests, goodness-of-fit test, cell tropism, HIV

1 Introduction.

Much effort has been devoted to deriving statistical tests to assess model fitness. The usual goodness-of-fit tests are often too optimistic in estimating model prediction error. Indeed, the same set of data is used to estimate parameters and to evaluate the model.

Seillier-Moiseiwitsch and collaborators [1] [2] proposed prequential tests as a means to evaluate models as data come in sequentially. In prequential testing, data are first divided into a training sample $\{y_1, y_2, \cdots, y_k\}$ and an evaluation sample $\{y_{k+1}, y_{k+2}, \cdots, y_n\}$. The training sample is then extended by one observation at a time to update parameter estimates and make a prediction for the next outcome until the last event is predicted. The test statistics are based on the prediction errors.

The uniqueness of prequential tests is that by extending the training sample and updating model parameters, the tests evaluate models on the basis of the accuracy of their probabilistic predictions for future events. Also, prequential tests achieve maximum usage of data in evaluating models. Specifically, in this research, we applied prequential tests to HIV envelope sequences in order to determine the key loci in the sequences that are associated with cell tropism.

2 Prequential tests in HIV sequences.

Cell tropism is categorized by whether HIV is CCR5-tropic or CXCR4-tropic. It is therefore considered as a binary variable. There are a total of 795 HIV envelope sequences considered. Each sequence is composed of 35 amino acid residuals spanning the V3 loop. Aligned sequences form a block with 795 rows and 35 columns. Columns are scanned one at a time.

For each column, a logistic regression model is applied with cell tropism as the dependent variable and amino acid residuals in each column as covariates. Prequential test is performed to determine whether any amino acid residual plays significant role in cell tropism. Seillier-Moiseiwitsch proved that under correct model specification, prequential test statistics follow a standard normal distribution. In this case, we include one covariate at a time into the logistic model. If the amino acid residual is related to cell tropism, the distributions of the prequential test statistics obtained by including the covariate and not including it into the model should be significantly different. Empirical distributions of test statistics are computed from the data as an approximation for the true reference distributions.

Using this method, we identified several important positions in the HIV envelope sequences including columns 11 and 25. Columns 11 and 25 of the V3 loop were already recognized as positions related to HIV cell tropism. Besides these, columns 9, 13, 19, 20 and 24 also have some amino acid residuals showing a statistically siginificant impact on viral tropism. This result indicates the potential biological relationship of these positions to the viral-host membrane fusion process.

[1]Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland, USA. E-mail: jzhang8@math.umbc.edu

[2]Director, Bioinformatics Research Center, University of Maryland Baltimore County, Baltimore, Maryland, USA. E-mail: seillier@math.umbc.edu

References

[1] F. Seillier-Moiseiwitsch, T. J. Sweeting and A. P. Dawid. 1992. Prequential Tests of Model Fit. *Scandinavian Journal of Statistics*, Volume 19: 45–60.

[2] F. Seillier-Moiseiwitsch and A. P. Dawid. 1993. On Testing the Validity of Sequential Probability Forecasts. *Journal of the American Statistical Association*, Volume 88, No. 421.

[3] F. Seillier-Moiseiwitsch. 1996. Predictive Diagnostics for Logistic Models. *Statistics in Medicine*, Volume 15: 2149–2160.

I42. An Eulerian Path Approach to Local Multiple Alignment of DNA Sequences

Yu Zhang[1], **Michael S. Waterman**[2]

Keywords: local multiple alignment, de Bruijn graph, repeat finding

1 Introduction.

Many available local alignment programs are limited by running time and accuracy when aligning large sequence sets. We present an Eulerian path approach to local multiple alignment for DNA sequences. The computational time and memory usage of this approach is almost linear to the total size of sequence set. By constructing a de Bruijn graph, most of the conserved segments are amplified as heavy paths in the graph, and the original patterns distributed in sequences are recovered even if they do not exist in any single sequence. This approach can detect both short (< 20bps) and long (> 300bps) conserved segments, as well as degenerate patterns (70% pairwise similarity). In addition, a Poisson heuristic [1] is applied to estimate the significance of local multiple alignments.

We demonstrate the performance of our method by an application in Alu repeat finding in the human genome. Five genomic sequences were tested, ranging from 22Kb to 1Mb in length. We compared the result to Alus marked by RepeatMasker[5], which uses detailed information about Alu repeats. Two programs are in good agreement.

2 Method.

The initial motivation for the method arises from the algorithm for DNA fragment assembly using the Eulerian superpath approach [2][4]. We first construct a de Bruijn graph using overlapping k-tuples from the given sequence set. Each k-tuple is represented by a directed edge in the graph, and two edges are joint by a node if their k-tuples overlap at (k-1) letters in any sequence. Identical k-tuples are represented by same edges. Under this construction, each sequence is mapped to a path traversing the graph. If a k-tuple appears in multiple sequences, the corresponding sequence paths will intersect at the edge representing the k-tuple. Based on this property, we define the multiplicity of an edge to be the number of sequence paths visiting the edge, i.e., the number of sequences containing the represented k-tuple. Using a probabilistic analysis, the larger the multiplicity is, the more likely the edge represents a conserved k-tuple. And vise versa, the conserved segments tend to be amplified in the graph by edges of large multiplicities. Therefore, we can extract a consensus of conserved regions by traversing the graph even if the original pattern do not exist in any sequence. We then apply constrained local alignment methods to find all segments similar to the consensus and output the result as a local multiple alignment. We use a Poisson heuristic to control the false positive rate and to estimate the significance of a local alignment as well.

[1]Department of Mathematics, University of Southern California, Los Angeles, CA 90089-1113. E-mail: yuzhang@usc.edu

[2]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-1340. E-mail: msw@usc.edu

3 Results.

We generated 3 different patterns of various lengths ($| S |= 63$, $| M |= 42$, $| W |= 21$), and inserted mutated copies of each pattern into N random sequences of length L. The number of copies inserted in each sequence follows the Poisson distribution. Table 1 shows the result.

Sequences			Patterns Inserted				Patterns Found					
N	L	identity	Total	S	M	W	Total	S	M	W	FP	FN
10	2K	90%	27	10	10	7	26	10	10	6	0	1
30	2K	90%	77	33	19	25	76	33	19	24	0	1
10	20K	90%	33	10	9	14	29	10	9	10	0	4
30	20K	90%	82	26	24	32	80	26	24	30	0	2
10	2K	80%	33	8	8	17	31	8	8	15	0	2
30	2K	80%	92	29	28	35	87	29	28	30	0	5
10	20K	80%	22	6	5	11	19	5	5	9	0	3
30	20K	80%	81	38	20	23	76	38	20	18	0	5

Table 1: Simulation results.

To test on real data, we used our program to find Alu repeats in the human genome. Five sequences of various lengths were randomly selected from NCBI database. Our results agrees well with Alus found both by the authors who submitted the sequence and by RepeatMasker, a program which knows Alu consensuses in prior. Result is shown in Table 2. As another comparison, a repeat finding program REPuter [3] fails to find many Alu repeats presented and runs slower than our method does.

ID	L	RepM	Family	Len(avg:std)	Div(avg:std)	FP	FN	Time/sec
AF435921	22Kb	29	10	261 : 69	15.0% : 6.4%	0	0	11
Z15025	38Kb	53	13	245 : 85	15.7% : 5.7%	3	2	15
AC034110	167Kb	89	18	261 : 72	12.2% : 5.9%	0	4	65
AC010145	199Kb	120	13	277 : 55	15.0% : 5.6%	1	3	134
Chr22	1Mb	717	32	252 : 79	15.2% : 6.1%	5	107*	1095

Table 2: Alu finding in the human genome. 'RepM': number of Alus marked by RepeatMasker; 'Family': number of Alu subfamilies; 'Len': Alu lengths; 'Div': percentage divergence to Alu consensuses. * A second run of our program reduces the number of false negatives to 30.

References

[1] Aldous, D. 1989. *Probability approximations via the Poisson clumping heuristic.* Springer, New York.

[2] Idury, R. and Waterman, M.S. 1995. A new algorithm for DNA sequence assembly. *J. Comp. Biol.* **2**:291-306.

[3] Kurtz, S. and Schleiermacher, C. 1999. REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*, **15**:426-427.

[4] Pevzner, P.A., Tang, H., and Waterman, M.S. 2001. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci.* **98**:9748-9753.

[5] Smit, A.F.A. and Green, P. Unpublished results.

I43. FastR: Fast database search tool for structured RNA sequences[1]

Vineet Bafna[2], Shaojie Zhang[2]

Keywords: non-coding genes, RNA, database search, filtration, dynamic programming

1 Introduction.

The discovery of novel non-coding RNAs has been among the most exciting recent developments in Biology. Yet, many more remain undiscovered. It has been hypothesized that there is in fact an *abundance* of functional non-coding RNA (ncRNA) with various catalytic and regulatory functions [1]. Computational methods tailored specifically for ncRNA are being actively developed. As the inherent signal for ncRNA is weaker than that for protein coding genes, comparative methods offer the most promising approach, and are the subject of our research.

We consider the following problem: Given an RNA sequence with a known secondary structure, efficiently compute all structural homologs (computed as a function of sequence and structural similarity) in a genomic database. Our approach, based on structural filters that eliminate a large portion of the database, while retaining the true homologs allows us to search a typical bacterial database in minutes on a standard PC, with high sensitivity and specificity. This is two orders of magnitude better than current available software for the problem.

2 Methods and Results

Recently, Klein and Eddy [3] developed a tool, RSEARCH, for searching a database with a query RNA molecule. The method depends upon existing algorithms for computing alignments between an RNA sequence and substrings of a database, where the alignment score is a function of sequence and structural similarity. Known algorithms for computing such alignments are computationally intensive (approximately $O(mw^2n)$, where m is the length of the query sequence, n is the length of the database sequence, and w is the maximum length of a database substring that is aligned to the query). Not surprisingly, RSEARCH is slow to use. For a test run on an Intel/linux PC with 2.8GHz, 1Gb memory, a microbial database of size 1.67M, and a query 5S rRNA sequence, the program took over 6.5 hrs. to run. This makes it impractical when either the query or the database is large.

We propose *FastR* , an efficient database search tool for ncRNA. An analogy can be drawn from fast search tools (BLAST/FASTA) for DNA and Protein sequences that has made database searching practical. The speed and effectiveness of BLAST in particular has contributed in large measure to the exponential growth of sequence databases, and the use of database search as an accepted method for finding novel DNA/protein homologs. By proposing FastR, which includes a novel idea for RNA structure filtering, and a novel & simple RNA alignment algorithm, we hope to do the same for ncRNA. As an example, FastR reduces the compute time of the previously mentioned query to 103s.

[1]The full-paper version of this work has been submitted to ISMB/ECCB 2004.

[2]Department of Computer Science and Engineering, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0114. Email: vbafna@cs.ucsd.edu; shzhang@cs.ucsd.edu

	Query	Hits (TP/Tot)	Time
RSEARCH	Asn-tRNA(AE001087.1/4936-5008)	85/93	3411s
FastR	”	71/87, 72/97	52s
RSEARCH	5S rRNA (AE016770.1/210436-210555)	97/97	14939s
FastR	”	79/99	44s
RSEARCH	Purine-Rs (AE010606.1/4680-4581)	33/39	9215s
FastR	”	27/35	30s
RSEARCH	Hammerhead (M83545.1/56-3)	50/58	2741s
FastR	”	44/51, 45/61	34s

Table 1: Comparison of FastR and RSEARCH.

Query	Genome	FastR (hits/TP/FN)	RSEARCH (E-val $\leq$ 10)	FastR time	RSEARCH time
Asn tRNA	*A. pernix*	25/24/9	57/31/2	2m57s	146m22s
5S rRNA	*A. pernix*	9/1/1	2/1/1	1m43s	390m7s

Table 2: Comparison of RSEARCH and FastR results on querying the 1.67Mb *A. pernix* genome (NC_000854.1). The true positives are obtained from known annotations. For False Negatives, we do not consider tRNAs with introns.

To test our algorithms, we worked with arbitrary ncRNA subfamilies of known/predicted structure from the RFAM [2] and the 5S Ribosomal RNA database [4]. Four sub-families are considered here, tRNA, 5S rRNA, a Purine Riboswitch, and the Hammerhead Ribozyme. For every sub-family, we chose some members arbitrarily, and inserted them in a random database of 1Mb, and tested our algorithms on the composite sequence. Table 1 summarizes the results of our search. As can be seen, FastR is close to two orders of magnitude faster than RSEARCH while maintaining comparable sensitivity.

We have also tested FastR on real genomes, where it is difficult to distinguish true hits. As shown in Table 2, querying the 1.67 Mb *A. pernix* genome yielded comparable results. FastR could not detect the 14 intron containing tRNAs, but detected 24 out of the remaining 33. For 5S rRNA, the single known annotation was the top hit, but there were other alignments of similar quality, indicative of novel 5S rRNAs. In the other two cases (Hammerhead and Purine-Rioboswitch), RSEARCH did not return any significant hit, and no annotations were available, hence no comparison could be made. FastR dominates again in speed.

References

[1] Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nature Reviews in Genetics*, 2: 919–929.

[2] Griffiths-Jones, S., Bateman, A., Marshall, M, Khanna, M. and S.R. Eddy. 2003. Rfam: an RNA family database. *NAR*, 31(1): 439–441.

[3] Klein, R.J. and Eddy, S.R. 2003. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics* 4(1): 44.

[4] Szymanski, M., Barciszewska, M.Z., Erdmann, V.A. and Barciszewski, J. 2000. 5S ribosomal RNA database. *NAR*, 28(1): 166–167.

I44. Transcription unit organization in Prokaryotes

Gabriel Moreno-Hagelsieb[1] and Warren F. Lamboy[2]

Keywords: genome organization, operons, inter-genic distances

1 Abstract.

Here we show that a lot can be learned of the transcription unit (TU) organization of prokaryotes by analyzing the number of genes of directons, stretches of genes in the same strand with no intervening gene in the opposite strand, and of the inter-genic distances. We give a detailed study of *Escherichia coli* K12, compared with data on experimentally determined TUs available thought RegulonDB [1]. We use *E. coli* K12 as a basis to compare the tendencies in TU organization of other prokaryotes.

References

[1] Salgado, H., et al., *RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12.* 2004. *Nucleic Acids Res* **32 Database issue**: p. D303-6.

[1] Program of Computational Genomics, CIFN-UNAM, Apdo Postal 565-A, Cuernavaca, Morelos, 62100 Mexico.
[2] Center for Agricultural Bioinformatics, USDA-ARS, Cornell Theory Center, Cornell University, Ithaca, NY 14853.

I45. Performance Comparison of Multiple Sequence Alignment Programs Using Nonparametric Statistics

Conrad Shyu, [1] **Luke Sheneman,** [2] **James A. Foster** [3]

Keywords: multiple sequence alignment, progressive alignment, nonparametric statistics, Clustal W

1 Introduction.

We present an approach that utilizes the nonparametric statistical procedures to compare the performance of two multiple sequence alignment programs, Clustal W and EVALYN. In particular we employ the sign and permutation tests to compare the score difference. We establish the hypothesis that EVALYN performs better than Clustal W and show that the score difference is statistically significant. In addition we employ the bootstrap methods to find the estimate of the MSE (mean square error), and construct the confidence intervals of the mean and variance [2]. Our study has shown that nonparametrics provides an attractive alternative for the comparison of sequence alignment programs. It gives a more statistically sound interpretation on comparison of the alignment scores, and does not rely on any benchmark sequence database [3].

2 EVALYN.

EVALYN [1] employs a genetic algorithm (GA) to evolve guide trees, which is a fundamental component of progressive alignment algorithms. The population in the GA consists of viable guide trees that are represented in an efficient, coalescing binary tree structure. Variability operators such as crossover and mutation are constructed such that the viability of an individual tree is never compromised. Rank-based selection is implemented via a random number generator that follows the beta distribution. This non-uniform random selection allows for strongly biased rank-based selection wherein highly-fit individuals are more likely to be selected for crossover and produce offspring. The low-fit individuals are replaced with the newly created offspring that has higher fitness value. Elitism is implemented to preserve the fittest individual in the population. The fitness value is objectively computed by performing the progressive alignment in the pairwise order specified by the individual guide tree.

3 Discussion and Conclusion.

We conducted the experiments on synthesized data in order to better control the similarities and lengths of the sequences. Over 30 tests were run, with each test working against a different input sequence file, which contained different number of sequences and/or different sequence lengths. The input sequences for all of the experimental runs were constructed

[1]Initiative for Bioinformatics and Evolutionary Studies (IBEST), Department of Bioinformatics and Computational Biology (BCB), University of Idaho, Moscow, Idaho 83844-1010, USA. E-mail: shyu4751@uidaho.edu

[2]Email: shen0614@uidaho.edu

[3]Email: foster@cs.uidaho.edu

using the Jukes-Cantor model. Both alignment programs were parameterized identically. The best scores of each run from GA were recorded. For this study, we used the minimum scores from EVALYN for the analysis. We first applied the sign and permutation tests to obtain the sample means and variances. The p-value is about 0.043, which is statistically significant at the 95 percent confidence interval. We then obtain 5000 bootstrap samples for the estimates of the confidence intervals of the p-value, means and variances. Approximately 60.58 percent of the bootstrap p-values are less than 0.05 and the largest one is 0.99. The frequency distribution of the bootstrap sample means clearly follows the normal distribution and centers around 10968 or 9.30 in the log scale. The distribution of the bootstrap variances centered around 287198676 or 19.41 in the log scale.

Our study has shown that nonparametric statistics provides a better alternative for the performance comparison of alignment programs. Comparisons that rely on a specific benchmark database can potentially introduce bias, which might favor certain types of algorithms. Furthermore one can certainly fine-tune an algorithm and exclusively target the benchmark database. On the contrary, nonparametric statistics is not susceptible to such a problem. The comparison can be conducted more objectively and solely focuses on the quality of alignments.

4 Acknowledgements.

This research used equipment funded in part by NIH NCRR 1P20 RR16448, and NIH NCRR 1P20 RR16454. Shyu was partially funded by a grant from Proctor and Gamble, and Sheneman was partially funded by NIH NCRR 1P20 RR16448. Foster was partially funded for this research by NIH NCRR 1P20 RR16448.

References

[1] Shyu, C., Sheneman, L., and Foster, J.A. 2004. Multiple sequence alignment with evolutionary computation. *Genetic Programming and Evolvable Machines, Special Issue* vol. 5(2).

[2] Higgins, J.J. 2003. *Introduction to Modern Nonparametric Statistics.* New York: Thomson.

[3] Thompson, J.D., Plewnaik, F., and Poch, O. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Research* 27(13).

I46. Searching Bioinformatic Sequence Databases using UM-BLAST—A Wrapper for High-Performance BLASTs

Xue Wu, Chau-Wen Tseng [1]

Keywords: Bioinformatics, sequence comparison, parallel processing, cluster computing

1 Introduction

BLAST [1] is the most widely used search tool for screening large bioinformatic sequence databases, and accounts for a large portion of the computation performed in bioinformatics. As a result, researchers have developed many versions of BLAST to improve its performance. Among them, NCBI BLAST [2] and WU BLAST [3] support parallel processing on a variety of SMP computer architecture. mpiBLAST [4] is the most recent version of parallel BLAST for distributed memory architecture. While BLAST++ [5] tried to improve the throughput of BLAST by batching the query sequences and exploiting sharing of results on common subsequences of queries. In addition, several organizations (TurboWorx Inc., Paracel Inc. and SGI Inc.) and researchers [6, 7, 8] have also developed their own version of high performance BLASTs.

However, our experiments show that no single version of BLAST is able to achieve the best performance given variations in sequence database size, query batch size, and query sequence length. We find mpiBLAST is best at exploiting database partitioning over multiple nodes to keep large databases in memory. BLAST++ is best at amortizing search costs for multiple batched queries. Threaded BLAST is best at reducing search costs for very long single queries. Based on our evaluation, we design UM-BLAST, a wrapper capable of selecting the proper combination of threaded BLAST, BLAST++, and mpiBLAST to achieve good performance over a range of search parameters.

2 UM-BLAST Wrapper

UM-BLAST is designed to select and invoke the most efficient version of BLAST for the database size, batch size, and query length selected. The basic algorithm for UM-BLAST is as follows:

1. Pre-partition large sequence databases for mpiBLAST so that each partition fits in memory for a single node
2. For each batched sequence search query
   (a) If database is too large to fit in memory on a single node, use mpiBLAST
   (b) Else
      i. If (batch size $> B$) and (query length $< L$) use replicated BLAST++
      ii. Else use replicated threaded-BLAST
      iii. Combine outputs for batch from replicated BLASTs

[1] Computer Science Department, University of Maryland at College Park. E-mail: {wu,tseng}@cs.umd.edu

3 Performance Measurement and Results

We measured the performance of threaded BLAST, mpiBLAST, BLAST++ and UM-BLAST on a Linux PC cluster with 10 worker nodes (each with dual AMD Athlon 1.6 GHz processors and 1G bytes of memory). Detailed results are shown in Figure 1, 2 and Table 1.

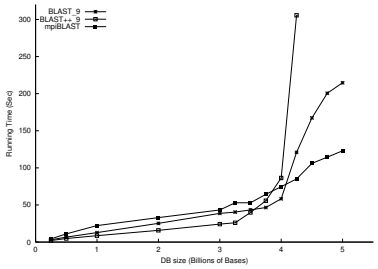


Figure 1: Impact of DB Size on BLASTs' Parallel Performance

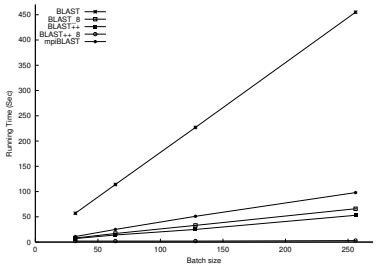


Figure 2: Impact of Batch Size on BLASTs' Performance

DB Type	Example DB	Batch Size	Query Length	UM-BLAST's Choice	Time(s)	Speedup
Large	4.5G	any	any	mpiBLAST	85.1	20 - 30
Medium	1.0G	256	3250	replicated threaded BLAST	438	1.4 - 100+
	1.0G	1	16384	mpiBLAST	6	5 - 6
	250M	256	64	replicated BLAST++	3	17.7 - 151.7
Small	13M	1024	64	BLAST++	2	46 - 123.5

Table 1: UM-BLAST's Performance

4 Conclusions

The results show that our UM-BLAST, a wrapper for high-performance BLASTs, can automatically select the appropriate version and configuration of BLASTs based on the target database size, query batch size and query sequence length. We feel UM-BLAST will be very useful to bioinformatics researchers seeting up their own BLAST servers.

References

[1] Altschul, SF, Gish, Miller, W, Myers, EW and DJ Lipman, A basic local alignment search tool, *Journal of Molecular Biology (1990)* 215:403-410

[2] NCBI BLAST, *http://www.ncbi.nih.gov/BLAST/*

[3] WU-BLAST, *http://blast.wustl.edu/blast/README.html*

[4] Aaron Darling, Lucas Carey, Wu-chun Feng, The Design, Implementation, and Evaluation of mpiBLAST, *ClusterWorld 2003*

[5] H. Wang, T.H. Ong, B.C. Ooi, K.L. Tan, BLAST++: A Tool for BLASTing Queries in Batches, *Proceedings of the 1st Asia-Pacific Bioinformatics Conference*, Feburary, 2003

[6] J. D. Grant, R. L. Dunbrack, F. J. Manion, and M. F. Ochs, BeoBLAST: distributed BLAST and PSI-BLAST on a Beowulf cluster, *Bioinformatics* Vol 18, pp. 765-766

[7] Michel Dumontier1,2 and Christopher W. V. Hogue, NBLAST: a cluster variant of BLAST for NxN comparisons, *BMC Bioinformatics*, 2002; 3(1):13

[8] Akira Naruse, Naoki Nishinomiya, Hi-per BLAST: High Performance BLAST on PC Cluster System, *Genome Informatics*, 2002; Vol 13, pp 254-255

FastGroup II: A web-based bioinformatics platform for analyses of large 16S rDNA libraries

Yanan Yu[1] , Pat McNairnie[1] and Forest Rohwer[1,2]

Key words: prokaryote, bacteria, 16S rDNA, biodiversity, richness, rarefaction

1 Introduction

Prokaryotes are the most abundant and diverse components of the biosphere. Most of these microbes cannot be cultured in the lab and are therefore studied by sequencing their ribosomal RNA genes (16S rDNA) [1]. We have been using high-throughput 16S rDNA sequencing to study the diversity of *Bacteria* and *Archaea* associated with reef-building corals. This has created a data glut because there are few bioinformatic tools for the analyses of large numbers of 16S rDNA sequences. To address this problem, FastGroup II has been developed.

2 Method

FastGroup II is a web-based software tool to de-replicate large 16S rDNA libraries. After cloning and sequencing, the 16S rDNA sequences are imported into FastGroup II as individual text files or as one FASTA file. The sequences are then trimmed to commonly used conserved sites [2], or to sites specified by the user. Poor-quality sequence, as determined by user-defined criteria is also eliminated. Then, using one of several different algorithms, the 16S rDNA sequences are grouped together. This grouping de-replicates the library, displays one representative sequence for all sequences that are >97% identical, and produces a table noting the number of sequences that fell into each group. The user has the ability to change both the algorithm and the grouping criteria used.

When the grouping is completed, FastGroup II automatically calculates standard diversity and richness indices, including the Shannon index, Chao1, and rarefaction of each 16S rDNA library. The result is also visualized in the rank-abundance curve using web-based graphical views. The data from each library is then stored in a MySQL database for later retrieval and comparison against other 16S rDNA libraries.

3 Results

FastGroup II is user-friendly and is currently being used in our laboratory for several studies of coral-related *Bacteria* and *Archaea*. This software package is built using Objected-Oriented Perl and new methods and functions are continuously being added. This client-server program will soon be available for public use.

[1]San Diego State University, Department of Biology, LS316, 5500 Campanile Drive San Diego, California 92182-4614 USA
[2]Center for Microbial Sciences, San Diego State University, 5500 Campanile Drive San Diego, California 92818-4614 USA

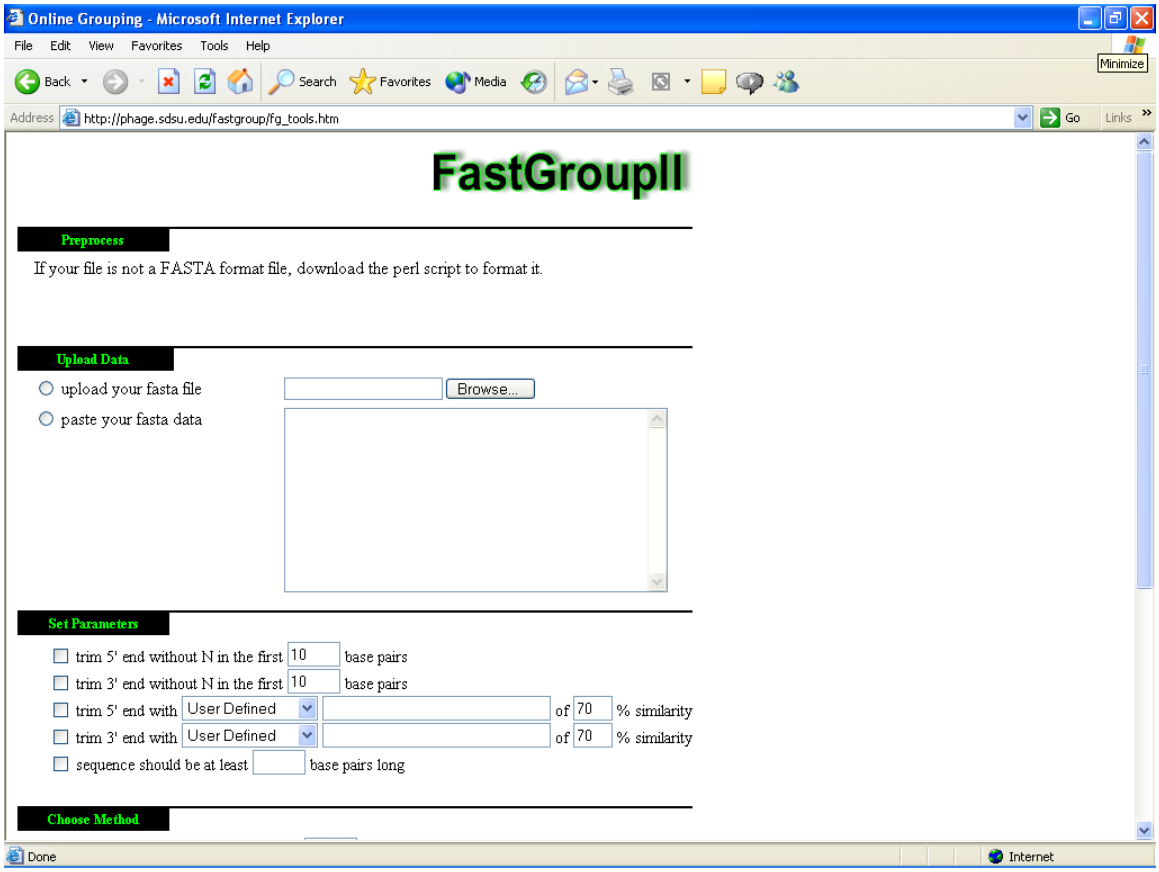


Figure 1. Part of FastGroupII user interface.

4 References

[1] Hugenholtz P, Goebel BM, Pace NR: Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of Bacteriology 1998, 180: 4765-4774*

[2] Seguritan, V, Rohwer, F: FastGroup: A program to dereplicate libraries of 16S rDNA. *BMC Bioinformatics 2001, 2:9*

I48. Paradigms for Computational Nucleic Acid Design

Robert M. Dirks[1], Milo Lin[2], Erik Winfree[3] and Niles A. Pierce[4]

Keywords: DNA, RNA, secondary structure, positive design, negative design, affinity, specificity, partition function

Abstract

The design of DNA and RNA sequences is critical for many endeavors, from DNA nanotechnology, to PCR-based applications, to DNA hybridization arrays. Results in the literature rely on a wide variety of design criteria adapted to the particular requirements of each application. Using an extensively-studied thermodynamic model, we perform a detailed study of several criteria for designing sequences intended to adopt a target secondary structure [1]. We conclude that superior design methods should explicitly implement both a positive design paradigm (optimize affinity for the target structure) and a negative design paradigm (optimize specificity for the target structure). The commonly used approaches of sequence symmetry minimization and minimum free energy satisfaction primarily implement negative design and can be strengthened by introducing a positive design component. Surprisingly, our findings hold for a wide range of secondary structures and are robust to modest perturbation of the thermodynamic parameters used for evaluating sequence quality, suggesting the feasibility and ongoing utility of a unified approach to nucleic acid design as parameter sets are further refined. Finally, we observe that designing for thermodynamic stability does not determine folding kinetics, emphasizing the opportunity for extending design criteria to target kinetic features of the energy landscape.

Introduction

A fundamental design problem consists of selecting the sequence of a nucleic acid strand that will adopt a target secondary structure [7]. As depicted in Figure 1a, this is the inverse of the more famous folding problem of determining the structure (and folding mechanism) for a given sequence. To attempt the rational design of novel nucleic acid structures, we require both an approximate empirical physical model [6, 4] and a search algorithm for selecting promising sequences based on this model. Experimental feedback on the quality of the design and the performance of the design algorithm can then be obtained by folding the molecule *in vitro*. Alternatively, if this feedback loop can be closed computationally by folding the molecule *in silico*, the quality of sequence designs could be rapidly assessed and improved before attempting laboratory validation.

In designing nucleic acid sequences, we consider the two principal paradigms illustrated in Figure 1b. *Positive design* methods attempt to select for a desired outcome by optimizing sequence affinity for the target structure. *Negative design* methods attempt to select against unwanted outcomes by optimizing sequence specificity for the target structure. A successful design must exhibit both high affinity and high specificity [8], so useful design algorithms must satisfy the objectives of both paradigms, even if they explicitly implement only one.

[1]Chemistry, Caltech. E-mail: dirks@caltech.edu
[2]Undergraduate, Caltech. E-mail: miloiq@its.caltech.edu
[3]Computer Science and Computation & Neural Systems, Caltech. E-mail: winfree@caltech.edu
[4]Applied & Computational Mathematics, Bioengineering, Caltech. E-mail: niles@caltech.edu

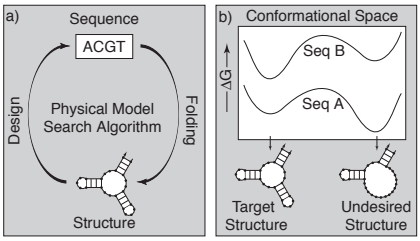


Figure 1: a) Feedback loop for evaluating nucleic acid sequence designs and methodologies. b) Positive and negative design paradigms. Two sequences are evaluated using an empirical potential on both the desired target structure and an undesired structure. Using a positive design paradigm, sequence A would be selected since it exhibits a stronger affinity than sequence B for the target structure (i.e. lower ΔG). Using a negative design paradigm, sequence B would be selected since it exhibits specificity for the target structure while sequence A exhibits specificity for the undesired structure.

For some applications, it may be desirable to supplement these thermodynamic design considerations with additional kinetic requirements. For example, in designing molecular machines [10], it may be crucial to select sequences that fold or assemble quickly. Alternatively, it may be important to design interactions with intentionally frustrated folding kinetics in order to control fuel delivery during the work cycle [9].

The present study uses efficient partition function algorithms [5, 2] and stochastic kinetics simulations [3] to examine the thermodynamic and kinetic properties of sequences designed using seven methods that capture aspects of the positive and negative design paradigms. Although several of these design criteria have been widely used, we are not aware of any previous attempt to assess their relative performance. Evaluated based on thermodynamic considerations, we consistently observe that sequence selection methods that implement both positive and negative design paradigms outperform methods that implement either paradigm alone. This trend appears to be robust to changes in both the target secondary structure and the parameters in the physical model, and to the choice of either RNA or DNA as the design material. The trend does not hold when the design criteria are judged based on kinetic considerations, as favorable thermodynamic properties do not ensure fast folding.

References

[1] R.M. Dirks, M. Lin, E. Winfree, and N. A. Pierce. Paradigms for computational nucleic acid design. *Nucleic Acids Res.*, 2004. In press.

[2] R.M. Dirks and N. A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, 24:1664–1677, 2003.

[3] C. Flamm, W. Fontana, I.L. Hofacker, and P. Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 2000.

[4] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[5] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[6] J. SantaLucia, Jr. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, 35:3555–3562, 1996.

[7] N. C. Seeman. Nucleic acid junctions and lattices. *J. Theor. Biol.*, 99:237–247, 1982.

[8] N. C. Seeman and R.K. Kallenbach. Design of immobile nucleic acid junctions. *Biophys. J.*, 44:201–209, 1983.

[9] A. J. Turberfield, J.C. Mitchell, B. Yurke, Jr. Mills, A. P., M.I. Blakey, and F.C. Simmel. DNA fuel for free-running nanomachines. *Phys. Rev. Lett.*, 90(11):118102, 2003.

[10] B. Yurke, A.J. Turberfield, Jr. Mills, A.P., F.C. Simmel, and J.L. Neumann. A DNA-fuelled molecular machine made of DNA. *Nature*, 406:605–608, 2000.

I49. Modeling Phage Species Abundance

David Bangor[1], Beltran Rodriguez Brito[2], Peter Salamon[3], James Nulton[4], Ben Felts[5], Joe Mahaffy[6], Mya Breitbart[7], Forest Rohwer[8]

Keywords: Lander/Waterman, viruses, phage, biodiversity, metagenomes

Bacteriophage (phage) are viruses that infect bacteria. Once inside their host bacteria, phage rapidly multiply and eventually kill the host by causing it to explode and release the new phage particles. These free phage particles are the most abundant biological entities in the biosphere, with and estimated 10^{31} phage particles on the planet. The total number of phage species, however, is essentially unknown. Recently we have started to shotgun sequence the DNA from uncultured phage communities. From this data, we obtained overlapping fragments, which means that the same phage genome has been resampled. This observation allows us to mathematically model the phage community.

Previous studies of phage diversity have been performed using culture-based methods [1]. This involves first culturing a bacterial host from the environment, and then isolating phage that can infect that bacterial host. Unfortunately, most bacterial hosts (>99%) cannot be cultured and not all phage produce identifiable plaques. To circumvent these limitations, we developed a method to shotgun sequence DNA from uncultured phage communities. To do this, total genomic DNA was isolated from natural phage communities containing approximately 10^{12} particles. The genomic DNA was physically sheared into 1-2 kilobase long fragments. Then the fragmented DNA sequences were cloned and sequenced. Finally, the DNA sequences were analyzed using Sequencher [2] to identify sequences that overlapped with 98% identity over 20 bp. An overlap between sequences means that the same genome has been re-sampled. A contig spectrum was created from the Sequencher analysis, where a 1-contig means the sequence had no overlaps, a 2-contig means two sequences overlapped, a 3-contig means three sequences overlapped, etc... The contig spectra from 4 environmental samples were then used to predict the population structure of the phage communities using a modified Lander/Waterman algorithm [3, 4, 5, 6].

In the current analysis a number of different distributions were compared. The first two models are the Broken Stick and the Niche Preemption [7], which are ecological models based upon the division of resources into niches. The other models were the common empirical functional forms - Power Law, Exponential Law, Logarithmic, and Lognormal. To make the comparisons, the contig spectrua obtained from the samples were used to model the populations assuming one of the 6 functional forms. The error between the predicted and the actual contig spectra were determined.

[1]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: heimdalle@yahoo.com

[2]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: brodrigu@rohan.sdsu.edu

[3]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: salamon@saturn.sdsu.edu

[4]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: jnulton@mail.sdsu.edu

[5]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: bfelts@myth.sdsu.edu

[6]Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: mahaffy@math.sdsu.edu

[7]Department of Biology, San Diego State University, San Diego, California, 92182-4614.E-mail: mya@sunstroke.sdsu.edu

[8]Department of Biology, San Diego State University, San Diego, California, 92182-4614.E-mail: forest@sunstroke.sdsu.edu

Table 1 shows that the Power Law and the Lognormal best described the observed contig spectra. In contrast, the ecological models did a very poor job of explaining the observed data.

	Scripps Pier	Mission Bay	Fecal Data	Mission Bay Sediment
Power Law	1.81	2.11	9.20	0.0104
Exponential Law	12.1	16.2	59.9	0.0126
Logarithmic	2.51	2.81	10.3	0.0104
Broken Stick	10.7	14.6	51.6	0.0156
Niche Preemption	29.5	38.1	145	ND
Lognormal	1.89	2.31	9.66	0.0104

Table 1: Errors for the given species abundance model and the environmental sample above. ND = not determined

In the environmental samples the phage community structure is best described by a Power Law or Lognormal distribution. This does not mean that all phage community structures will be a Power Law or Lognormal distribution, but it does give some idea of what the actual mathematical distribution may look like. The distinction between the phage community falling under actual Power Law distribution or a Lognormal distribution needs to be determined in the future because the absolute number of species calculated by the Lognormal is approximately 3X as much as that predicted by the Power Law. The shape of the community distribution is also important for modeling how phage and their bacterial host interact. Currently, we are performing higher coverage sequencing to differentiate between these two functions.

References

[4] Breitbart, M., B. Felts, et al. (2004). Diversity and population structure of a nearshore marine sediment viral community. Proceedings of the Royal Society B, in press.

[5] Breitbart, M., I. Hewson, et al. (2003). Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 85(20): 6220-6223

[6] Breitbart, M., P. Salamon, et al. (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 99(22): 14250-14255.

[3] Lander, E. S. & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2, 231-239.

[7] MacArthur, R. H. (1957). On the relative abundance of bird species. Proc. *Nat. Acad. Sci*. Wash. 43: 293-295.

[2] Sequencher http://www.genecodes.com/

[1] Wommack, K. and R. Colwell (2000). "Virioplankton: Viruses in aquatic ecosystems." *Microbiol Mol Biol Reviews* 64(1): 69-114.

I51. Cyber Infrastructure for Phylogenetic Research

Fran Berman[1], Bernard Moret[2], Satish Rao[3], David Swofford[4], Tandy Warnow[5]

Keywords: phylogenetics, phyloinformatics, algorithms, biological dataset analysis

1 Introduction.

CIPRes (Cyber Infrastructure for Phylogenetic Research) is an NSF-funded community effort to design and build an integrated environment for large-scale phylogenetic analysis.

The environment will integrate high-performance computing platforms, large databases, biological datasets and their analyses, benchmark datasets, optimization software, and a flexible user interface. It will serve both biologists carrying out analyses of biological data and algorithm designers developing and testing new phylogenetic reconstruction methods.

The collaboration involves directly 13 universities and museums and 33 researchers in North America and indirectly many more institutions and individuals worldwide. CIPRes researchers coordinate closely with national and international initiatives for the reconstruction Tree of Life -- an ambitious project to reconstruct the evolutionary history of all living species (in the tens of millions).

2 Project Overview

CIPRes will be composed of a large computational platform, a collection of interoperable high-performance software for phylogenetic analysis, and a large database of datasets (both real and simulated) and their analyses; it will be accessible through any web browser by developers, researchers, and educators. The software, freely available in source form, will be usable on scales varying from laptops to high-performance, Grid-enabled, compute engines such as our platform, and will be packaged to be compatible with current popular tools. In order to build this resource, CIPRes will support research programs in phyloinformatics (databases to store multilevel data with detailed annotations and to support complex, tree-oriented queries), in optimization algorithms, Bayesian inference, and symbolic manipulation for phylogeny reconstruction, and in simulation of branching evolution at the genomic level, all within the context of a virtual collaborative center.

Biology, and phylogeny in particular, has been almost completely redefined by modern information technology, both in terms of data acquisition (new genomic data accumulates at a rate exceeding

[1] Department of Computer Science/San Diego Supercomputer Center, University of California at San Diego. La Jolla, California, USA. E-mail: berman@cs.ucsd.edu

[2] Department of Computer Science, University of New Mexico, Albuquerque, New Mexico, USA. E-mail: moret@cs.unm.edu

[3] Computer Science Division, University of California at Berkeley, Berkeley, California, USA. E-mail: satishr@cs.berkeley.edu

[4] Department of Biological Science, Florida State University, Tallahassee, Florida, USA. E-mail: swofford@csit.fsu.edu

[5] Department of Computer Sciences, University of Texas at Austin, Austin, Texas, USA. E-mail: tandy@cs.utexas.edu

Moore's law) and in terms of analysis (the literature shows over 10,000 citations to the top three phylogenetic software packages). Phylogeneticists have formulated specific models and questions that can now be addressed using recent advances in database technology and optimization algorithms. The time is thus exactly right for a close collaboration of biologists and computer scientists to address the IT issues in phylogenetics, many of which call for novel approaches, due to a combination of combinatorial difficulty and overall scale. CIPRes includes computer scientists working in databases, algorithm design, algorithm engineering, and high-performance computing, evolutionary biologists and systematists, bioinformaticians, and biostatisticians, with a history of successful collaboration and a record of fundamental contributions, to provide the required breadth and depth.

CIPRes brings together researchers from many areas and foster new types of collaborations and new styles of research in computational biology; moreover, the interaction of algorithms, databases, modeling, and biology will give new impetus and new directions in each area. CIPRes will help create the computational infrastructure that the research community will use over the next decades, as more whole genomes are sequenced and enough data is collected to attempt the inference of the Tree of Life. It will help evolutionary biologists understand the mechanisms of evolution, the relationship between evolution, structure, and function of biomolecules, and a host of other research problems in biology, eventually leading to major progress in ecology, pharmaceutics, forensics, and security (including computer security). The project will publicize evolution, genomics, and bioinformatics through informal education programs at our museum partners and will motivate high school students and college undergraduates to pursue careers in bioinformatics. CIPRes provides an extraordinary opportunity to train students, both undergraduate and graduate, as well as postdoctoral researchers, in one of the most exciting interdisciplinary areas in science; our institutions serve a large number of underrepresented groups and are committed to increase their participation in research.

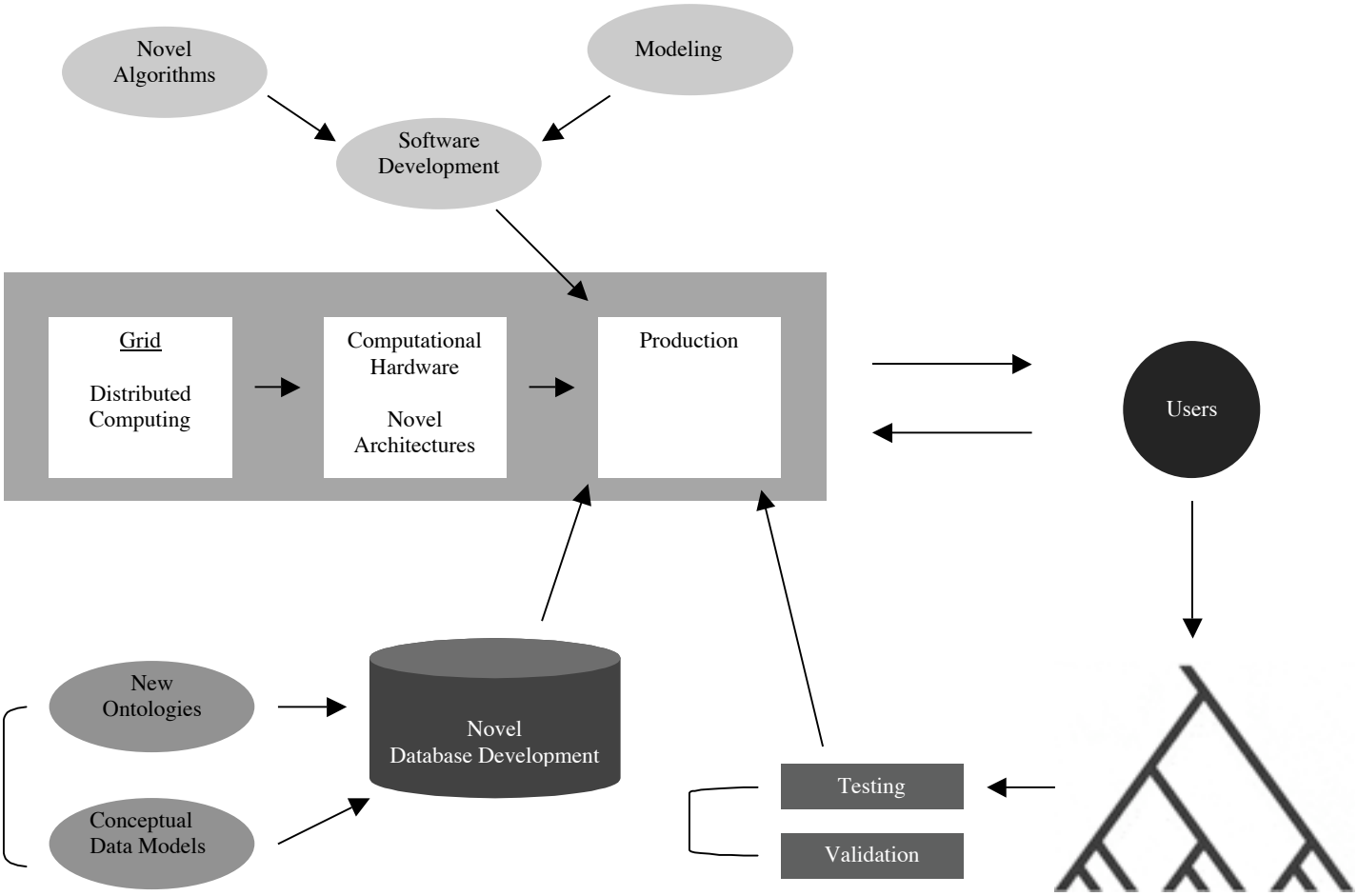


Figure 1: Overview of the proposed schematic model of CIPRes.

J1. Cross-Link Analysis and Experiment Planning for Elucidation of Protein Structure

Xiaoduan Ye,[1] **Janusz M. Bujnicki,**[2]
Alan M. Friedman,[3] **Chris Bailey-Kellogg**[4]

Keywords: Protein structure prediction, protein-protein complexes, experiment design, cross-linking mass spectrometry, disulfide trapping, structural genomics.

Emerging high-throughput experimental techniques for the characterization of protein and complex structure yield noisy data with sparse information content, placing a significant burden on computation to predict, optimize, and interpret the information provided. One such experiment employs residue-specific chemical cross-linkers to confirm or select among proposed structural models by testing consistency of cross-linking data with respect to model geometry (Fig. 1). Recent applications include those by Young et al., using high-resolution mass spectroscopy alone to correctly discriminate threading models of fibroblast growth factor [7]; Scaloni et al., analyzing the binding mode of the calmodulin-mellitin complex [4]; and Sorgen et al., determining the arrangement of transmembrane helices in lac permease [6].

We have developed a mechanism for analyzing cross-linking information with respect to a set of models, predicting the ability of experiments to discriminate among those models, and optimizing a set of experiments accordingly. A probabilistic framework selects models based on consistency with data (Fig. 1(c)), mediated by the geometric feasibilities of the cross-links for the models [3], represented by "cross-link maps" (Fig. 1(b)), and the experimental conditions, represented with noise and capture rates. We formalize model discriminability in terms of differences in cross-link maps, and formulate experiment planning problems to select sets of experiments that maximize such differences, accounting for the key factors of discriminability, coverage, balance, ambiguity, and cost. We have developed a greedy algorithm, generalizing those for related SetCover problems [1], that effectively navigates the design space defined by these terms.

We are applying this mechanism in a study of the bacteriophage lambda Tfa chaperone protein, and have planned dicysteine mutations for model discrimination by disulfide formation. Fig. 2 summarizes our planning results on 103 Tfa models, including three high-quality threading models from our fold recognition meta-server [2] and 100 decoys from the *ab initio* folding program Rosetta [5]. We are currently carrying out a minimized, balanced, and least ambiguous set of six experiments to discriminate the three threading models with discriminability two.

Our mechanism is very general, and we plan to study additional applications not tested here, for example in the discrimination of protein-protein complexes, incorporating different types of experimental data, and planning combinatorial possibilities of cross-linkers and mutations. Our methods provide the experimenter with a valuable tool for understanding and optimizing cross-linking experiments.

[1]Department of Computer Sciences, Purdue University. ye@cs.purdue.edu
[2]International Institute of Molecular and Cell Biology, Warsaw, Poland. iamb@genesilico.pl
[3]Department of Biological Sciences, Purdue University. afried@bilbo.bio.purdue.edu
[4]Corresponding author. Department of Computer Sciences, Purdue University. 250 N. Univ. St., West Lafayette, IN 47907, USA. cbk@cs.purdue.edu

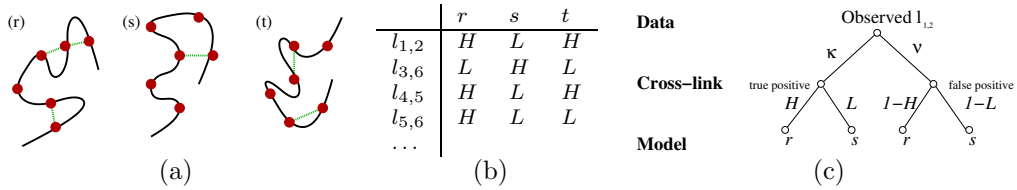


Figure 1: Model discrimination by cross-linking. (a) Different predicted models of a protein have different patterns of feasible cross-links (green dotted lines). (b) Cross-link maps represent feasibilities with conditional relationship for cross-links (rows) given models (columns), here shown as either high (H) or low (L). (c) Experimental identification of a cross-link $l_{1,2}$ provides evidence for and against models r and s, based on consistency with cross-link maps and modulated by the capture and noise rates of the experimental method (here uniformly κ and ν, respectively). Other terms arise from other cross-links, both observed and unobserved.

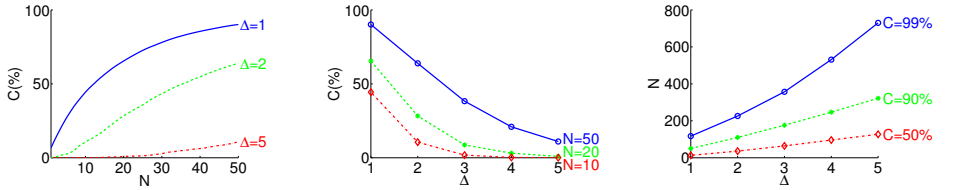


Figure 2: The relationship between coverage percentage C (%), discriminability Δ, and number of experiments N in disulfide experiments planned for 103 Tfa models. Pairs of parameters are varied, while the third is blocked at the indicated values. Our greedy plans are roughly within a factor of two of a simplistic lower bound (data not shown).

References

[1] D.S. Johnson. Approximation algorithms for combinatorial problems. *J Comput System Sci*, 9:256–278, 1974.

[2] M.A. Kurowski and J.M. Bujnicki. Genesilico protein structure prediction meta-server. *Nucleic Acids Res*, 31(13):3305–7, 2003. `http://genesilico.pl/meta`.

[3] S. Potluri, A.A. Khan, A. Kuzminykh, J.M. Bujnicki, A.M. Friedman, and C. Bailey-Kellogg. Geometric analysis of cross-linkability for protein fold discrimination. In *Pac Symp Biocomp*, pages 447–458, January 2004.

[4] A. Scaloni et al. Topology of the calmodulin-melittin complex. *J Mol Biol*, 277:945–958, 1998.

[5] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J Mol Biol*, 268:209–25, 1997.

[6] P.L. Sorgen, Y. Hu, L. Guan, H.R. Kaback, and M.E. Girvin. An approach to membrane protein structure without crystals. *PNAS*, 99(22):14037–14040, 2002.

[7] M.M. Young et al. High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *PNAS*, 97:5802–5806, 2000.

J2. Analyzing Protein Structure Using Almost-Delaunay Tetrahedra

Deepak Bandyopadhyay[1], **Jack Snoeyink** [1] **Alexander Tropsha** [2]

Keywords: almost-Delaunay, neighbors, protein structure, Delaunay probability, SNAPP, four-body statistical potential, motif detection, secondary structure

1 Introduction.

The Delaunay tessellation (DT) of a protein structure [6] collects representative points of four "neighboring" residues into tetrahedra. The DT has many applications in the analysis of protein structure, of which we consider two in detail – scoring folded proteins to distinguish the native state from decoys [2, 4] and detecting motifs of local structure [7].

The Delaunay tessellation is defined using an "empty sphere" criterion (Delaunay, 1934) – the circumspheres of Delaunay tetrahedra contain no other points. However, protein atom coordinates are subject to uncertainties from rounding, measurement, conformational change and motion, and small changes in the coordinates may cause large changes in the DT.

The almost-Delaunay tetrahedra [1] expand the set of Delaunay tetrahedra to account for perturbation or motion of point coordinates. We define a quadruple of points to be in the set of *almost-Delaunay tetrahedra* with parameter ϵ, denoted AD(ϵ), if there is a perturbation of all points by at most ϵ that makes its circumsphere empty. We denote the minimum such perturbation for a quadruple its *AD threshold*. The Delaunay tetrahedra have threshold 0.

2 Experiments and Results

Our implementation in MATLAB and C++ can calculate the AD tetrahedra for typical proteins of 100-1000 residues in a few seconds to a few minutes, for typical values of two parameters: the maximum edge length (*prune*) and maximum perturbation allowed (*cutoff*). By studying a large number of point sets with different structure, we observed that there are fewer AD tetrahedra at low thresholds in proteins than in random point sets; hence the DT is more stable in proteins.

Simplicial Neighbor Analysis of Protein Packing (SNAPP) [2, 4] scores protein structures by summing the frequencies with which the four-tuples of amino acids observed as Delaunay neighbors, occur in native protein structures. We calculated SNAPP scores using AD tetrahedra for 6 proteins and their decoys from the *4state_reduced* [5] set. Overall, the modified scores were as successful as the original at distinguishing proteins from decoys, and made a slightly stronger distinction between proteins and highest-scoring decoys. Thus, decoy discrimination using the DT is robust.

We observed that the histogram distribution of AD tetrahedra vs. threshold for an ideal α-helix has sharp peaks at $\epsilon = 0.3$, 0.7 and 1.2. These values of ϵ correspond to specific patterns in the residue sequence numbers, as shown in Figure 1. Histograms for proteins containing α-helices reveal the same peaks and patterns; we can mark the corresponding tetrahedra as α-helical, and determine the residues in α-helical conformation using a heuristic. We may identify β-sheets and β-turns similarly by decoding patterns present in their AD tetrahedra. For details of the methods, see http://www.cs.unc.edu/~debug/

[1]Department of Computer Science, University of North Carolina at Chapel Hill. E-mail: {debug, snoeyink}@cs.unc.edu

[2]Laboratory of Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill. E-mail: tropsha@email.unc.edu

papers/AlmDel. Secondary structures assigned using the AD method match the widely used DSSP [3] assignments in most cases (a few are shown in Table 1). They are consistent and robust in cases where DSSP is not, and are closer to visual assignments done by a human expert.

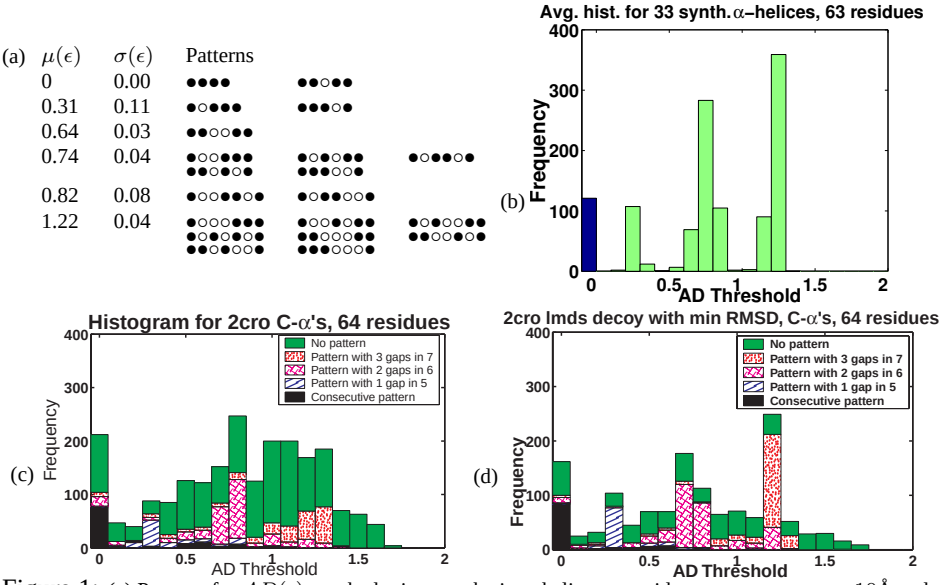


Figure 1: (a) Patterns for $AD(\epsilon)$ tetrahedra in a synthetic α-helix. $\bullet$=residue, $\circ$=gap, prune $= 10$Å and cutoff $\epsilon < 2$Å. (b) Histogram showing α-helical peaks, also seen in (c) 2cro and (d) a decoy with same secondary structure.

PDB ID	**#**	α-helix		β-sheet		β-turn	
/chain	**resid**	**DSSP**	**AD**	**DSSP**	**AD**	**PRO**	**AD**
1brx	209	158	158	10	8	12	10
1lrv	233	90	100	0	0	29	36
1timA	247	106	101	42	51	15	18
1bg5	254	70	102	0	12	68	32
1ejdA	418	128	138	105	134	43	40
1oen	524	133	112	126	138	86	94

Table 1: α-helical, β-sheet and β-turn residues assigned by DSSP [3] and by our AD patterns for 6 protein chains with varying lengths and CATH architectures.

References

[1] D. Bandyopadhyay and J. Snoeyink. Almost-Delaunay simplices : Nearest neighbor relations for imprecise points. In *ACM-SIAM Symposium On Discrete Algorithms*, pages 403–412, 2004.

[2] C. W. Carter, B. C. LeFebvre, S. Cammer, A. Tropsha, and M. H. Edgell. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *Journal of Molecular Biology*, 311(4):625–638, 2001.

[3] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[4] B. Krishnamoorthy and A. Tropsha. Development of a four-body statistical pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, 19(12), 2003.

[5] R. Samudrala and M. Levitt. Decoys 'R' Us: A database of incorrect conformations to improve protein strucure prediction. *Protein Science*, 9:1399–1401, 2000. http://dd.stanford.edu.

[6] R. Singh, A. Tropsha, and I. Vaisman. Delaunay tessellation of proteins. *J. Comput. Biol.*, 3:213–222, 1996.

[7] H. Wako and T. Yamato. Novel method to detect a motif of local structures in different protein conformations. *Protein Engineering*, 11:981–990, 1998.

J3. Avoiding Local Optima in Single Particle Reconstruction

Marshall Bern, [1] **Jindong (JD) Chen,** [2] **H. Chi Wong** [3]

Keywords: Electron microscopy, cryo-EM, maximum likelihood, EM algorithm.

1 Introduction

Cryo-Electron Microscopy (cryo-EM) uses a transmission electron microscope to acquire 2D projections of a specimen preserved in vitreous ice. A 3D electron density map can then be reconstructed from the 2D projections computationally. In "single-particle" cryo-EM, the specimen consists of many ostensibly identical copies of randomly oriented particles, and the reconstruction process must estimate the unknown orientations at the same time that it estimates the 3D structure.

The solution to this chicken-and-egg problem, developed over the last 20 years [1], is a computationally intensive process that starts from an initial guess of the shape and then iteratively aligns images to the current shape and reconstructs a new shape from the aligned images. In the standard approach, taken by all three of the major software packages (IMAGIC, SPIDER, and EMAN [2]), each particle image contributes equally to the reconstruction, with orientation set by the image's best alignment, typically determined by maximum correlation. The standard approach is remarkably successful, giving correct structures from images that to the human eye appear to be almost pure noise. Yet occasionally, depending upon the initial guess, the standard approach gives a completely incorrect structure, close to a fixed point for the iteration but nowhere near the true structure. Figure 1 gives an example.

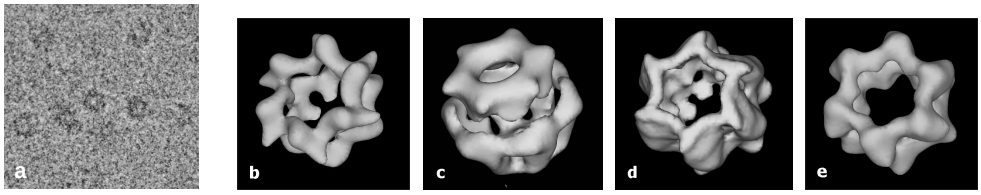
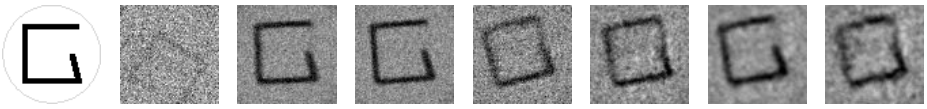


Figure 1: (a) A piece of a micrograph showing images of p97 AAA ATPase [3] particles (the dark blobs). (b) An isosurface of a correct reconstruction of p97 made from about 4000 particle images. (c) An incorrect reconstruction made from about 600 images, using EMAN's default initial guess. What went wrong is that some of the "top" and "bottom" views (hexagonal rings) of the particle were mistaken for "side" views. (d) A better reconstruction using EMAN with the same initial guess, along with multiresolution refinement and "annealing" of image orientations. (e) A "best-possible" reconstruction from the 600 images using the reconstruction from (b) as the initial guess.

An alternative to the standard approach is a more theoretically grounded (and even more computationally intensive) approach, due to Sigworth [4]. Sigworth's approach assumes a simple probabilistic model of cryo-EM imaging (i.i.d. Gaussian pixel noise) and seeks the maximum-likelihood (ML) reconstruction. The well-known Expectation Maximization algorithm (EM applied to EM!) is used to maximize the likelihood; this amounts to adding each image into the reconstruction for each possible orientation (that is, "soft" rather than "hard" orientation assignments), weighted according to the probability of that orientation. Sigworth demonstrated his ML approach on an analogous 2D problem: reconstructing an image from a number of noisy copies, randomly rotated and translated. The ML approach

[1] Palo Alto Research Center, 3333 Coyote Hill Rd., Palo Alto, CA 94304 E-mail: bern@parc.com
Work performed in part while visiting Departamento de Ciência da Computação, UFMG, Belo Horizonte, Brazil.
[2] Palo Alto Research Center E-mail: jchen@parc.com
[3] Palo Alto Research Center and DCC, UFMG, Brazil. E-mail: hcwong@dcc.ufmg.br



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 2: (a) The original particle, designed to have an incorrect local optimum (a square). (b) A data image with 10% signal, meaning that the image is a weighted average of a randomly rotated and translated copy of the original particle and 90% Gaussian noise. Both the standard algorithm (c) and the ML algorithm (d) give correct reconstructions from 100 images with 10% signal. (e) The standard algorithm starts to fail with 6% signal. The ML algorithm (f) starts to fail with 5% signal, but the ML algorithm with multiresolution (g) still gives good results. (h) The ML algorithm with multiresolution starts to fail with 4% signal.

shows reduced sensitivity to the initial guess and can recover structures from images with lower signal-to-noise ratio. In this poster we address the question: can the maximum-likelihood approach also help avoid incorrect local optima?

2 Our Experiments

We implemented the standard approach and a speeded-up version of the ML approach for the 2D problem. The speeded-up version does not use all possible assignments of orientation, but rather only the single best translation for each rotation; it gives results indistinguishable from the full ML approach. Figure 2 shows that this version of ML can indeed avoid an incorrect local optimum at a lower signal-to-noise ratio than the standard approach; this result turns out to be essentially independent of the initial guess. (We believe that because each 2D data image contains a full copy of the original, the 2D and 3D problems differ in their sensitivities to the initial guess.) We made another modification to the ML approach that turned out to be a substantial improvement (Figure 2(g) and (h)): this "multiresolution" algorithm low-pass filters each image before determining orientation probabilities and adding it into the reconstruction, and gradually reduces the filtering in later iterations. The rationale is to smooth out the optimization and remove local optima. (Low-pass filtering is also helpful because particle shape is often stronger relative to noise at low spatial frequencies.)

Will these results carry over to 3D? We modified EMAN in an attempt to correct the problem shown in Figure 1(c). Ludtke (personal communication) had previously tried to implement Sigworth's ML approach in EMAN, but could not find a robust definition of probability for the possible orientations of an image. (The experiments with synthetic data do not suffer from this problem, because the synthetic noise really does conform to the probabilistic model.) We tried a crude hack instead: simply adding a random component to EMAN's matching score for each image-orientation pair. Thus suboptimal asssignments occasionally win by chance. The size of the random component is reduced in later iterations, as in simulated annealing. Our modified version of EMAN also includes the multiresolution idea described above, and both "annealing" and multiresolution seem to be necessary to avoid the local optimum for the p97 data set. Figure 1(d) shows our 3D results.

References

[1] J. Frank. *Three-dimensional electron microscopy of macromolecular assemblies*. Academic Press, 1996.

[2] S.J. Ludtke, P.R. Baldwin, and W. Chiu. EMAN: Semiautomated software for high-resolution single-particle reconstructions. *J. Struct. Biol.* 128 (1999), 82–97.

[3] I. Rouiller, B. DeLaBarre, A.P. May, W.I. Weis, A.T. Brunger, R. Milligan, E.M. Wilson-Kubalek. Conformational changes of the multifunction p97 AAA ATPase during its ATPase cycle. *Nature Struct. Biol.* 9 (2002), 950–957.

[4] F.J. Sigworth. A maximum-likelihood approach to single-particle image refinement. *J. Struct. Biology* 122 (1998), 328–339.

J4. Discrete-event Simulation of Self-assembly Systems

Sue Yi Chew, [1] **Rorianne Rohlfs,** [1] **Russell Schwartz** [2]

Keywords: self-assembly, simulation, discrete-event, queuing

1 Motivation.

Self-assembly is the spontaneous and autonomous organization of components into patterns or structures. Self-assembly processes are common throughout nature and technology [8]. Naturally occurring self-assembly systems include viral protein shells, cellular cytoskeletons, and bacterial flagella and pili. Self-assembly has also emerged as a promising technology for nanometer-scale fabrication. The mathematical principles behind self-assembly have thus attracted considerable recent attention. Yet the kinetics of non-trivial self-assembly systems — the progress of the systems over time — remain extremely difficult to analyze due to the many physical details that are difficult to capture in abstract theoretical models and the propensity of these systems to exhibit "emergent" behaviors that are not easily predictable solely from a knowledge of low-level binding interactions. Without a firmer understanding of these kinetic properties, we cannot hope to design novel systems to have both rapid growth and high fidelity. Nor can we hope to understand biological self-assembly systems, which rarely exist at thermodynamic equilibrium. Simulation methods are therefore likely to be vital in understanding emergent properties of naturally occurring self-assembly systems, performing rapid *in silico* experimentation with possible interventions into those systems, and screening hypothetical novel systems for unforeseen difficulties prior to fabrication.

There have been many past approaches to simulating self-assembly, but there remain significant obstacles to the development of general, quantitative simulation tools. Differential equation models have provided quantitative simulations of large numbers of subunits [9, 4] but require substantial simplifications to be computationally tractable and cannot generally easily be adapted from one system to another. Simplified discrete-event models [1, 2] can provide efficient simulation and easy adaptability to new systems but have generally lacked quantitative accuracy and realistic models of binding kinetics. Highly detailed models that attempt to capture the physical forces underlying self-assembly phenomena can resolve many of these problems [6, 7, 5], but are difficult to parameterize and computationally intractable for all but very small systems. Our goal is to fill in important gaps in the range of systems these various approaches can model through the development of a quantitative simulation tool for generic self-assembly systems that is robust, efficient, and platform-independent.

2 Overview.

We have developed algorithms and data structures for fast, quantitatively accurate self-assembly simulation on biologically-relevant scales and incorporated them into a prototype simulation tool. We adapted the calendar queue [3] data structure to properties of a discrete event model of generic chemical assembly events to allow expected linear-time event processing across a broad range of self-assembly systems. This method further allows us to maintain realistic quantitative reaction rates even at very small system scales.

[1] School of Computer Science, Carnegie Mellon University. 5000 Forbes Avenue, Pittsburgh, PA 15213 USA. E-mail: {syc,rrohlfs}@andrew.cmu.edu

[2] Department of Biological Sciences, Carnegie Mellon University. 4400 5th Avenue, Pittsburgh, PA 15213 USA. E-mail: russells@andrew.cmu.edu

```
<SubunitType>
        <Name>
                subt1
        </Name>
        <Domain>
                position: 0.0,0.0,0.0
                <Conformation>
                        name: subt1only
                        E value: 6.6
                        <BindingSite>
                                <BindingSiteType>
                                        name: in
                                        tolerances: 1.2,0.0,3.5
                                        compat sites: out
                                </BindingSiteType>
                                position: 1.0,0.0,0.0
                                orientation: 0.0,0.0,0.0,0.0
                        </BindingSite>
                </Conformation>
                <TransitionTimes>
                        -
                </TransitionTimes>
        </Domain>
        <SubunitMass>
                3.4
        </SubunitMass>
</SubunitType>
```

Figure 1: Example specification of a subunit type for the simulator

We have combined these techniques with a generic representation of self-assembly subunits in an extensible Java object model. This approach allows us to model a broad range of system types using a single code base. Figure 1 shows how a specific type of assembly subunit can be specified for this model. Furthermore, work is currently underway on a versatile 3-D graphical interface to allow fine control over simulation specification and progress. This prototype simulation tool will provide a first step toward the availability of an efficient, generic, quantitative platform for *in silico* experimentation with self-assembly systems.

This work was supported by the Merck Company Foundation and by NSF award #0320595.

References

[1] Berger, B., Shor, P. W., Tucker-Kellog, L. and King, J. 1994. Local rule-based theory of virus shell assembly. *Proceedings of the National Academy of Sciences USA* 91:7732–7736.

[2] Berger, B., King, J., Schwartz, R., and Shor, P. W. 2000. Local rule mechanism for selecting icosahedral shell geometry. *Discrete Applied Mathematics* 104:97–111.

[3] Brown, R. 1988. Calendar queues: a fast O(n) priority queue implementation for the simulation event set problem. *Communications of the ACM* 31:1220–1227.

[4] Endres, D. and Zlotnick, A. 2002. Model-based analysis of assembly kinetics for virus capsids or other spherical polymers. *Biophysical Journal* 83:1217–1230.

[5] Rapaport, D. C., Johnson, J. E., and Skolnick, J. 1999. Supramolecular self-assembly: molecular dynamics modeling of polyhedral shell formation. *Computer Physics Communications* 121–122:231–235.

[6] Reddy, V. S., Giesing, H. A., Morton, R. T., Kumar, A., Post, C. B., Brooks, C. L. and Johnson, J. E. 1998. Energetics of quasiequivalence: computational analysis of protein-protein interactions in icosahedral viruses. *Biophysical Journal* 74:546–558.

[7] Schwartz, R., Shor, P. W., Prevelige, P. E. and Berger, B. 1998. Local rules simulation of the kinetics of virus capsid self-assembly. *Biophysical Journal* 75:2626–2636.

[8] Whitesides, G. M. and Grzybowski, B. 2002. Self-assembly at all scales. *Science* 295:2418–2221.

[9] Zlotnick, A. 1994. To build a virus capsid: an equilibrium model of the self-assembly of polyhedral protein complexes. *Journal of Molecular Biology* 241:59–67.

J5. A Dynamical Monte Carlo Algorithm to Study Protein Folding Pathways

Andres Colubri[1]

Keywords: Dynamical Monte Carlo, Kinetic Monte Carlo, Protein Folding, Folding Pathways, Kinetic Barriers, Structure Prediction

1 Abstract.

The most straightforward computational method to generate protein folding pathways consists in running Molecular Dynamics (MD) simulations. The time-dependent behavior of the atoms that form the protein and the surrounding solvent is obtained by solving the Newton's equations of motion given all the interaction forces present in the system. The disadvantages of this approach are well known: first, the time step of the simulations must be chosen small enough to accurately represent the fastest modes of atomic motion, hence a huge number of steps are needed in order to reach the timescales of interest, and second, a realistically sized protein contains thousands of atoms, which makes the calculation of the forces an extremely expensive task. Nevertheless, MD simulations can still be used to study problems simpler than finding the pathway that connects the unfolded state with the native structure. For example, determining the structural changes induced by single point mutations in the primary sequence [1], or more traditionally, refining the atomic coordinates obtained from experimental techniques such as X-ray diffraction or NMR.

These limitations observed in MD, caused by the high dimensionality of the system and by the presence of extremely fast motion modes, prompted the development of simplified or coarse-grained models to represent protein structure. The "standard" simplifications consist in treating the solvent implicitly and using the torsional coordinates of the protein backbone as the only degrees of freedom. However, these reductions have a cost, mainly the low resolution of the simulated structures and the need of using additional data-base information in order to generate reasonable conformations. With regards to the search algorithm, all these models usually rely on some variation of Monte Carlo (MC) sampling, frequently combined with a cooling schedule on the temperature to allow wider moves at the beginning of the simulation, but "freezing" when getting closer to a "good" structure. Important progress has been made with these techniques, particularly in the prediction of native folds given the primary sequence [2, 3].

Because MC evolves the system towards the equilibrium conformation, regardless of the real pathway, the simulated transitions cannot be interpreted dynamically and there is no clear correlation between the MC steps and real time. Leaving aside the question of whether or not the native structure of a protein corresponds to the state of global equilibrium, it is evident that a different approach is required to generate folding pathways that can be matched with real folding events. Even though the sole prediction of the native fold from the sequence is generally regarded as the central objective of any protein folding algorithm, the prediction of additional dynamical information has considerable practical relevance, as shown by the important role played by disordered and misfolded proteins in numerous biological processes [4].

Under the light of all these facts, we propose a coarse-grained model to codify the main structural features of a protein, which, as many other previous models, has an implicit treatment of the solvent and considers the backbone torsional angles (φ, ϕ) as the only coordinates to characterize

[1] Searle Chemistry Laboratory, University of Chicago, 5735 S. Ellis Ave. #126, Chicago, IL, 60637. E-mail: acolubri@uchicago.edu

folding. For a detailed discussion of this model, see ref. [4]. But instead of using MC as the sampling scheme, we have introduced a Dynamical Monte Carlo (DMC) algorithm [5, 6] to compute transitions in torsional space. DMC (also known as Kinetic Monte Carlo, KMC) is aimed to describe the time evolution of processes directed mainly by the kinetic barriers between local configurations of the system, which we assume is one of the main features of protein folding.

In order to apply DMC in our model, we have discretized the torsional angles (φ, ϕ). For each amino acid k along the protein chain, the integer variable R_k is defined as the Ramachandran basin that contains the point (φ_k, ϕ_k) (see ref. [4]). The DMC process is defined at the level of these discrete variables by generating transitions between the Ramachandran basins available to each amino acid.

For the DMC algorithm to be complete, the inter-basin transition rates must to be specified in such a way that they satisfy detailed balance. We have defined these transition rates using the thermally-exited process approach, in which the rates are barrier controlled:

$$W(\text{amino acid k goes from basin i to basin j}) = \exp\{-\beta B_k(i, j)\}$$

where $B_k(i, j)$ is the kinetic barrier associated to the basin change. This barrier can be estimated by computing the interactions that are "affected" (either by being formed or dismantled) by a basin change at site k.

Given that the sequence of basin hops for each amino acid k is independent, it results that the time evolution of the variables R_k follows a Poisson process. More importantly, it is possible to find a correspondence between the simulation steps and real time, given by:

$$\tau_i = - \ln U / R$$

where U is an uniform variable in [0, 1] and R is the total rate of the process at simulation step i.

We have implemented a protein simulation framework, called Open Protein Simulator (OPS) in which this DMC algorithm has been coded to test different functional forms for the kinetic barriers $B_k(i, j)$. Results in terms of folding pathways and applicability to predict native structure from the sequence alone are discussed.

2 References

[4] Colubri, A. 2004. Prediction of Protein Structure by Simulating Coarse-grained Folding Pathways, a Preliminary Report. In: *Journal Biomolecular Structure & Dynamics*. Accepted for publication.

[2] Fang, Q. and Shortle D. 2003. Prediction of Protein Structure by Emphasizing Local Side-Chain/Backbone Interactions in Ensembles of Turn Fragments. In: *Proteins: Structure, Function and Genetics* 53: 480-485.

[5] Fichthorn, K. A. and Weinberg, W. H. 1991. Theoretical foundations of dynamical Monte Carlo simulations. In: *Journal of Chemical Physics* 95:1090-1096.

[6] Johnson, D. 2001. Kinetic Monte Carlo: Bare Bones and a Little Flesh. Course material available online at: http://www.mcc.uiuc.edu/SummerSchool01/Duane%20Johnson/johnson.htm

[3] Jones, D. T. and McGuffin, L. J. 2003. Assembling Novel Protein Folds From Super-secondary Structural Fragments. In: *Proteins: Structure, Function and Genetics* 53: 480-485.

[1] Mooney, S. D. and Klein, T. E. 2002. Structural Models of Osteogenesis Imperfecta-associated Variants in the COL1A1 gene. In: *Molecular & Cellular Proteomics* 1:868-875.

3 Online resources

Source code implementing the Dynamical Monte Carlo algorithm and simulations generated with it are available at:
http://sosnick.uchicago.edu/aifoldlab.html

J6. Assessment of Replica Exchange Method for Protein Structure Refinement

G. Dent[1], A. K. Royyuru, P. Athma, and R. Zhou

Keywords: protein structure refinement, replica exchange, molecular dynamics, continuum solvent

1 Introduction.

Protein structure prediction has been of great interest recently, as witnessed by the past five worldwide events in Critical Assessment of Protein Structure Prediction (CASP). Even with enormous efforts from various groups, protein structure prediction remains a challenging and unsolved problem[1]. A derived and arguably equally challenging problem is how to refine the model structures from homology modeling or fold recognition methods. Even for homology models, the typical RMSDs from the native structures are about 3Å (even larger for typical fold recognition targets). However, structures with higher resolution (1.5-2.0Å RMSD) are often needed for many important studies, such as ligand-protein binding affinity prediction in drug design. In this study, we will use an extensive sampling technique, the replica exchange method, to systematically examine the structure refinement protocol with a wide range of model protein structures.

2 Methods.

The replica exchange method (REM)[2] is a powerful tool for efficient sampling of conformational space. Normal molecular dynamics (MD) or Monte Carlo (MC) methods are usually not very efficient for sampling because protein systems are often trapped in many local minima at room temperature. With REM, however, the high temperature replicas can traverse energy barriers more effectively, thus providing a mechanism for low temperature replicas to overcome the energy barriers they would otherwise encounter. The method can be briefly described as a two-step algorithm: (i) Each replica, i = 1, 2…M at fixed temperature T_m (m=1, 2, ..., *M*), is simulated *simultaneously* and *independently* for a certain number of MC or MD steps. (ii) Pick a pair of replicas, and exchange them with the acceptance probability:

$$T(x \to x') = \min\{1,\ \exp(-\Delta)\}, \tag{1}$$

$$\Delta = (\beta - \beta')[U(x') - U(x)], \tag{2}$$

β and β' are the two reciprocal temperatures; and U (x) and U (x') are potential energies at the configurations {x} and {x'}. In this study, a total of 18 replicas are used for each protein system with a temperature range from 300K to 502K. Each replica is run with MD for 2ns and conformations are saved every 1ps (it takes on average ~2 months CPU on IBM-Power3-375MHz SP nodes for each replica). Only neighboring replicas are attempted for swap with an acceptance ratio of about 20-40%. The total energy of protein conformations are calculated with the widely used OPLS-AA force field and a continuum solvent model, the Surface Generalized Born (SGB)[3] model:

$$U = U_{vac} + U_{SGB} + U_{cav} \tag{3}$$

where U_{vac} is the vacuum potential energy from the OPLS-AA force field, and U_{SGB} is the electrostatic contribution to the solvation energy,[3] and U_{cav} is the nonpolar contribution to the solvation energy.[3] The structure with the lowest total energy will be picked as the best one.

[1] IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, E-mail: gdent@us.ibm.com

3 Results and Discussion.

Protein sequence targets were selected for structure prediction and refinement from a non-redundant subset of PDB structures with a pairwise sequence identity below 25%. A predicted initial model was created for each target using homology modeling.[4] Three different classes of model structures were selected for this systematic study of refinement based on the backbone RMSD from the native structures: (1) RMSD < 3Å, (2) 3Å < RMSD < 8Å, and (3) RMSD >8Å, to represent a wide range of qualities in the initial models.

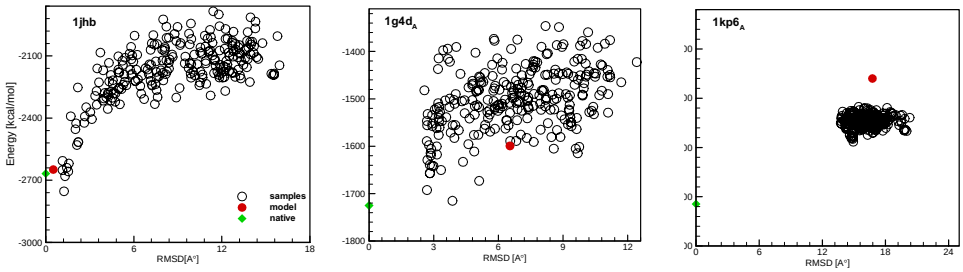


Figure 1: Total energy vs. RMSD for all conformations in each category: (a) initial RMSD < 3Å; (b) 3Å< RMSD < 8Å; and (c) RMSD >8Å.

The conformations sampled from REM is then clustered and minimized. The final results from three representative structures, one from each category, are shown in Figure 1. The first model, 1jhb (106 residues) has a starting RMSD of 0.5Å. The REM sampling shows only two structures having a lower energy than the model, but none of them has a lower RMSD. Other models in this category show similar limited improvements. The second model, 1g4d_A (69 residues) with a starting RMSD of 6.5Å, is found to have many structures with a lower energy and a smaller RMSD. This is a clear instance where refinement has improved the structure significantly, from 6.5Å to ~3Å. The last model, 1kp6_A (79 residues) with an initial RMSD of 16.8Å, shows some improvement as well, about 2-3Å. Overall, we do observe successes in refining some protein models, particularly those with larger initial RMSDs. Also, we found that with only a few exceptions the native structures are indeed the lowest total energy structures. However, the improvement is limited for the highly homologous models with very small RMSDs. This is probably the most useful category of structures with relevance to drug discovery applications, thus, further work is needed in this direction. Also, interestingly, the majority of the conformations sampled are further away from the native structure, indicating that the total sampling space is enormous and the total simulation time for effective refinement can be very long, perhaps comparable to the folding time. Further improvement and efficiency is likely through biased sampling to constrain the sampling around "native-like fragments" based on a suitable confidence measure.

References

[1] Moult J, Fidelis K, Zelma A, and Hubbard T. 2003. Critical Assessment of methods of Protein Structure Prediction (CASP) – Round V, Proteins, 53: 334-339 and references therein.
[2] Hukushima K and Nemoto K. 1996. J. Phys. Soc. Jpn. 65:1604; Zhou R, Berne BJ, and German R. 2001. The free energy landscape for b-hairpin folding in explicit water. Proc. Nat. Acad. Sci. 98: 14931-14936
[3] Ghosh.A, Rapp CS, Friesner RA. 1998. Generalized Born model based on a surface integral formulation. J. Phys. Chem. 102: 10983
[4] Zhou R, Silverman BD, Royyuru AK, and Prasanna A. 2003. Spatial profiling of protein hydrophobicity: Native vs decoy strucutres. 52: 561-572

J7. Predicting disulfide bond partners

F. Ferrè[1] and P. Clote[2]

Keywords: disulfide bond, neural network, machine learning.

1 Introduction

Disulfide bonds (covalently bonded sulfur atoms from nonadjacent cysteine residues) play a critical role for protein functionality and in stabilizing the protein structure. A number of relatively good algorithms have been developed to determine whether a cysteine is *reduced* (sulfur occurring in reactive sulfhydryl group SH) or *oxidized* (sulfur covalently bonded)[3], reaching 88% accuracy [5]. Despite this success there has been little progress in the problem of determining whether two half-cystines form a disulfide bond with each other – the *disulfide bond partner prediction* problem. In [1] a neural network is used to predict the probability of a disulfide bond between two half-cystines, using flanking sequence information, and subsequently, maximum weight matching is applied to pair those most likely partners.

Starting from the observation that there is a bias in the secondary structure preferences of free cysteines and half-cystines, we develop a neural network to learn disulfide bond preferences of both amino acid residues and secondary structure assignment of the symmetric flanking regions centered at partner half-cystines. Considering the secondary structure of pairs of half-cystines known to form a disulfide bond, some combinations are preferred, presumably indicating a sort of structural complementarity. This novel approach, as calibrated using receiver operating characteristic (ROC) curves [2], shows a marked improvement over previous work of Fariselli and Casadio [1]. Our final stand-alone program uses a neural network on the symmetric flanking residues about both cysteines of a potential disulfide bond, along with the PsiPred-determined secondary structure of the residues and PsiBlast-determined evolutionary information.

2 Methods and Results

We built a database by extracting flanking residues from the symmetric window of size w centered at each half-cystine in each mono-chain peptide domain from the nonredundant collection PDBselect25 [3], using DSSP to determine the cysteine oxidation state. Given two size w windows centered at an N- resp. C-terminus half-cystines, we then extracted DSSP [4] secondary structure annotations for each of the $2w$ residues; subsequently we ran PsiBlast to produce a profile, consisting of frequencies $f(i,a)$, for each of the 20 amino acids a and each position $1 \leq i \leq 2w$, obtained from the multiple sequence alignment of homologous proteins. The resulting input to our neural network consisted of $2w \cdot 20$ frequencies, along with $2w \cdot 3$ additional binary inputs, which latter encode in unary the secondary structure (H, C, E) of each of the $2w$ residues. When training the neural network, we used output value of 1 for an input corresponding to a valid disulfide bond, as determined by DSSP, and 0 for a pair of half-cystine flanking regions for incorrectly paired half-cystines. Altogether, there were $O(N)$ many positive [resp. $O(N^2)$ many negative] training examples.

[1]Department of Biology, Boston College, Chestnut Hill, MA 02467, ferref@bc.edu

[2]Departments of Biology and Computer Science (courtesy appointment), Boston College, Chestnut Hill, MA 02467, clote@bc.edu

[3]Disulfide-bonded cysteines are known as *half-cystines*, while reduced cysteines are also called *free* cysteines.

For the resulting neural network, trained with evolutionary information and secondary structure preferences, we tested a variety of possible network architectures. Of those tested, two architectures showed the best results in 20-fold cross-validation experiments with our database (see discussion above) using a window size of $w = 11$ residues. The first architecture had one hidden layer with two units, while the second had two hidden layers with 5 and 2 units, respectively. See Figure 1 for a summary of the statistics of our neural network, as well as a ROC curve comparison of our method with that of Fariselli and Casadio [1]. For the latter, we parsed the Fariselli-Casadio CONPRED neural network scores for likelihood of disulfide bond formation, without using their additional application of maximum weight matching.

Accuracy	76.58%
True positive rate%	81.05%
Falso positive rate %	26.57%
Correlation coefficient	53.66%
Sensitivity	81.05%
Specificity	73.43%

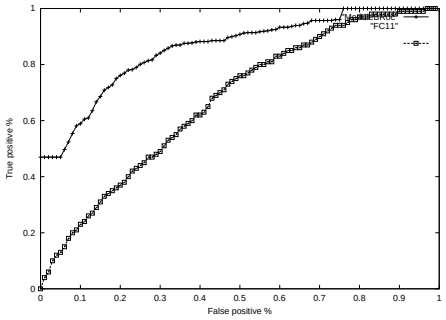


Figure 1: *(i)* Performance of the neural network disulfide connectivity prediction using secondary structure and evolutionary information. *(ii)* ROC curve for our method, described in this paper, compared with that of CONPRED – our method is the upper curve. Window size for both algorithms is $w = 11$.

After training, our prediction software works as follows. Given an input peptide along with user-designated half-cystine positions, our program uses PSIBLAST to obtain a profile and PSIPRED to predict secondary structure for the flanking residues. Our software then calls the neural network described in this paper.

References

[1] P. Fariselli and R. Casadio. Prediction of disulfide connectivity in proteins. *Bioinformatics*, 17(10):957–964, 2001.

[2] M. Gribskov and N.L. Robinson. The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers Chem.*, 20:25–34, 1996.

[3] U. Hobohm and C.Sander. Enlarged representative set of protein structures. *Proteins Science*, 3:522, 1994.

[4] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637, 1983.

[5] P.L. Martelli, P. Fariselli, L. Malaguti, and R. Casadio. Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. *Protein Eng.*, 15(12):951–953, 2002.

J8. Visualization and Analysis of Eukaryotic Gene Signals on Protein Structures

Vivek Gopalan[1], Shang Liang[2], Tin Wee Tan[3], Shoba Ranganathan[4]

Keywords: evolution, protein, gene, structure

1 Introduction.

Any relation between eukaryotic gene signals on protein structures is essential for analyzing the origin and evolution of genes. Here we present an approach for analyzing the correspondence between boundaries of the protein structural elements such as go-modules, domains and secondary structures and the intron positions in homologous eukaryotic genes. As a first step, we have developed a program that visualizes the various structural elements and provides detailed statistics about the significance of the intron positions in a single or set of protein structures.
The results from the analysis can be used for testing the Exon Theory of Genes (ETG) or inron early theory, which advocates that genes in primordial cells contain introns and proposes exon shuffling as the main process for protein diversity.

2 Methods

The intron positions and phases of eukaryotic genes are extracted from Xpro [1], which is eukaryotic protein encoding database based on GenBank (version 137)[2]. The algorithm for prediction of go module boundaries is based on de Souza et al [3]. The intron positions and phases are mapped on the PDB protein chain sequences based on conserved region in the BLAST alignment against the subset of protein sequences that represent intron containing genes in Xpro database. The E-value cutoff for the BLAST search was kept as 10^{-4} [4]. The program is written in Java 2 and Java 3D API is required for visualization of protein structures. The cartoon model of the protein structure is obtained based on the MOLSCRIPT program [5]. The secondary structures of the protein chains are obtained from the definitions in the PDB file.

3 Results

Figure 1 shows the Go plot for the chain A of the PDB code 1TIM. The module regions and their boundaries are shown as rectangular blocks in the hypotenuse of the Go plot. Figure 2 shows the screen shot of the of the protein structure for the PDB code 1TIM with Go-module mapped on it.

[1] Department of Biochemistry, National University of Singapore, Singapore 119260. E-mail: vivek@bic.nus.edu.sg
[2] Department of Biochemistry, National University of Singapore, Singapore 119260. E-mail: sci11162@nus.edu.sg
[3] Department of Biochemistry, National University of Singapore, Singapore 119260. E-mail: tinwee@bic.nus.edu.sg
[4] Department of Biochemistry, National University of Singapore, Singapore 119260. E-mail: shoba@bic.nus.edu.sg

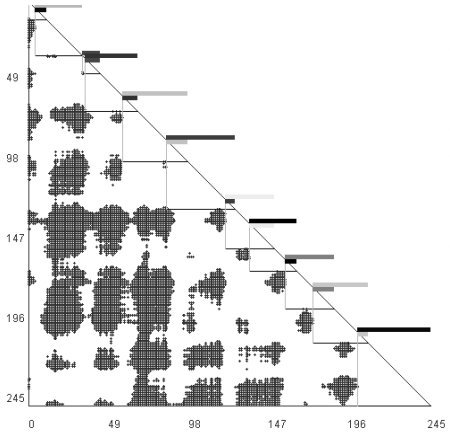


Figure 1: Screen shot showing Go plot for chain A for PDB code 1TIM at 28 Å diameter.

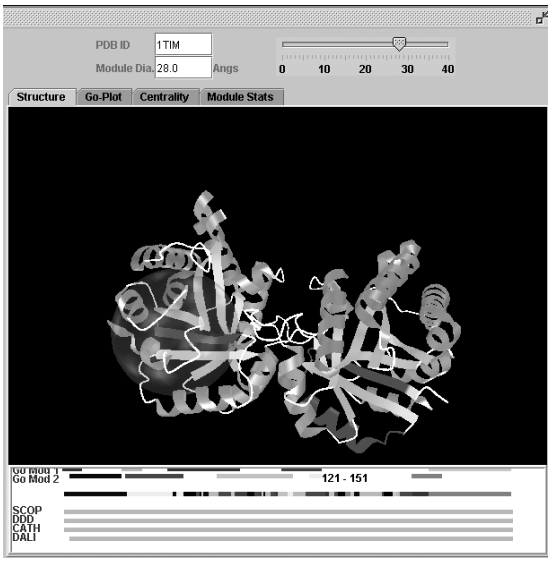


Figure 2: Cartoon model of a PDB structures with Go-modules mapped on protein 3D structure. The spherical region in the left chain indicates a Go-module (PDB code – 1TIM, chains – A,B).

The lower region of figure 2 shows the Go-modules mapped a PDB chain in addition to the mapped intron positions and domain definitions from various domain definition databases such as CATH, SCOP, DALI, and 3Dee. Provisions for changing the Go-module diameter and visualizing each of the domain regions, exon encoded regions or module boundaries are built-in within the program. In conclusion, we have developed an integrated platform for analyzing and visualizing features related to protein structures and eukaryotic gene signals.

4. References

[1] Gopalan, V., Tan, T.W, Lee, B.T. and Ranganathan, S. 2004. Xpro: database of eukaryotic protein-encoding genes. *Nucleic Acids Research* 32:D59-63.

[2] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. 2003. GenBank. Nucleic Acids Research. 31:23–27.

[3] de Souza, S.Long, M., Schoenbach, L. and Gilbert, W. 1996. Intron positions correlate with module boundaries in ancient proteins. *Proceedings of the National Academy of Sciences USA* 93: 14632-14636.

[4] Fedorov, A., Cao, X., Saxonov, S., de Souza, S.J., Roy, S.W., Gilbert, W. 2001. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. *Proceedings of the National Academy of Sciences USA* 93: 14632-14636 98:13177-82.

[5] Kraulis, P.J. 1991. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. Journal of Applied Crystallography 24: 946-950.

J10. Modelling and Simulation Studies of the Intracellular Domains of the Inwardly Rectifying K^+ Channels

Shozeb Haider[1], Frances Ashcroft[2]. Mark S P Sansom[1]

Keywords: mgirk, Kir-ATP, K^+ channels, molecular dynamics

1 Introduction.

Cells communicate with their extracellular environment through a diverse range of proteins that are present in the membrane. Membrane proteins account for the ~25% of the genes in most genomes. Ion channels are membrane proteins that have a ubiquitous presence in cells and reflecting their roles in cellular physiology. Potassium (K^+) channels form a large and diverse family responsible for a range of functions in various cell types and tissues, including control of cell electrical excitability in the nervous and cardiovascular systems. K^+ channels that are responsible to maintain and stabilize the resting membrane potential are the inwardly rectifying potassium channels (Kir). ATP-sensitive K^+ channels (K_{ATP}) belong to this group and are responsible in coupling the membrane electrical activity to energy metabolism. In the beta cells of the pancreas, they couple the rate of insulin release to the blood glucose levels and it has been well established that type II diabetes results from defective metabolic regulation of K_{ATP} channels. Despite the recent advances in the understanding of the K^+ channel structure and function, the molecular mechanisms underlying the ATP-dependent inhibition of Kir channels still remain elusive.

2 Results.

The crystal structure of the C-terminal intracellular domain of the inwardly rectifying K^+ channel Kir3.1 has recently been published at 1.8Å [1]. In the present study, molecular dynamics simulations were performed to investigate the conformational dynamics of both monomeric and homotetrameric forms of the protein. The structure of the Kir3.1 domain also provides a template for homology modeling of the equivalent domain of other Kir channels, and thus offers the prospect understanding their mechanisms of inhibition and gating. Models of the monomeric and tetrameric forms of the intracellular domains of the mammalian Kir6.2 channels were built based on the x-ray structure of the C-terminal domain of the Kir3.1 channel. Automated docking was employed to identify the residues and the binding site involved in the inhibition of Kir6.2 channels by ATP [2]. Extensive molecular dynamics simulations were performed to investigate the stability of these structures, with and without bound ATP in both the monomeric and tetrameric forms. Simulations were carried out for a total of 10 ns using GROMACS. The overall RMSD (root mean square deviation) of the Cα atoms for the simulations are described in Table 1. The loop regions form the most flexible part of the monomeric structure. They are stabilised in the tetramer by interactions with the adjacent subunits. The RMSD for the core region (i.e. excluding the loops) is low, suggesting the stability of the core structure. It has been experimentally observed that the mutation of Lys185 to Asp/Glu reduces ATP sensitivity considerably [2]. The distance between Lys185 and

[1] Department of Biochemistry, South Parks Road, The University of Oxford, Oxford, OX1 3QU. E-mail: shozeb@biop.ox.ac.uk

[2] Department of Physiology, South Parks Road, The University of Oxford, Oxford OX1 3QU. E-mail: frances.ashcroft@physiology.ox.ac.uk

the ß-phosphate from ATP averaged over the entire trajectory for the structures was between 3.1-3.5A suggesting that Lys185 interacts with ß-phosphate of ATP. These simulations help in studying the nature of residues that contribute to the binding of ATP to the channel and thus could be involved in the structural and functional properties. Furthermore, such computational studies aid in understanding and interpreting the mutation data and relate the structure to the physiological function at an atomic level.

3 Figures and tables.

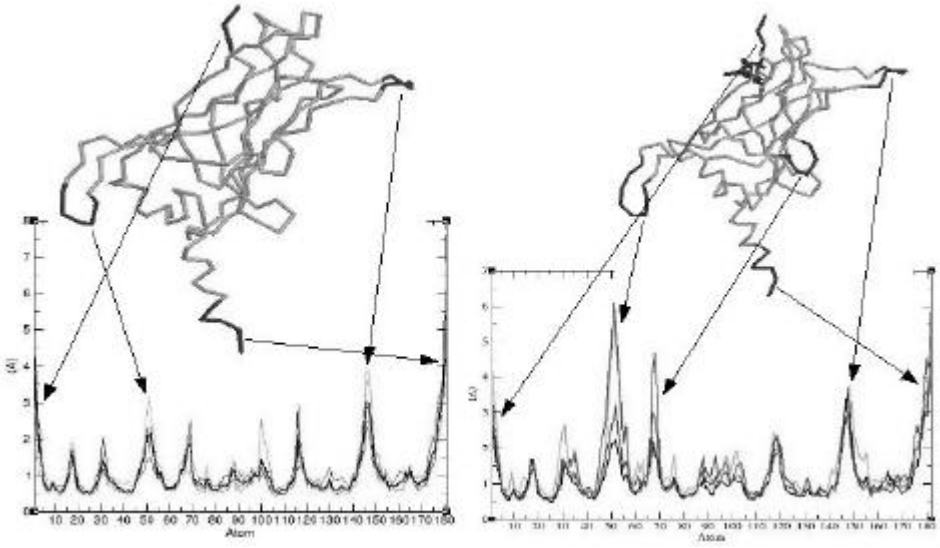


Figure 1: RMSF (root mean square fluctuation) of residues in each subunit of the intracelleular domain from (a) Kir3.1 and (b) Kir6.2. The RMSF peaks correspond to loops in the structure and the corresponding regions are indicated using arrows.

Protein	Ligand	Time (ns)	RMSD (Overall, Å)	RMSD (cutoff at 2.5 Å)
Kir3.1 monomer	-	5	2.8	2.2
Kir3.1 tetramer	-	5	2.6	1.9
Kir6.2 monomer	-	5	3.8	2.7
Kir6.2 monomer	ATP	5	3.2	2.0
Kir6.2 tetramer	-	5	3.2	2.9
Kir6.2 tetramer	ATP	5	3.7	2.4

Table 1: Summary of simulations.

4 References and bibliography.

[1] Nishida, M and Mackinnon, R. 2003. Structural basis of inward rectification: Cytoplasmic pore of the G protein-gated inward rectifier GIRK1 at 1.8Å resolution. *Cell.* Vol 111. No. 7 pp. 957-965.

[2] Trapp, S., Haider, S., Jones, P., Sansom, M.S.P. and Ashcroft, F. 2003. Identification of residues contributing to the ATP binding site of Kir 6.2. *EMBO J.* Vol 22 No.12 pp. 2903-2912.

J11. Hybrid Probabilistic Roadmap and Monte Carlo Methods for Biomolecule Conformational Changes

Li Han [1]

Keywords: Conformation space, conformational changes, Monte Carlo, probabilistic roadmaps.

1 Introduction

Biomolecule conformational changes play important roles in biological functions and are an integral part of various important structural problems addressed in computational biology such as ligand binding and (mis)folding of proteins and nucleic acids. The current structural biology experimental techniques are not yet as powerful for providing information on dynamic molecular motion as on static molecular structures; but there has been a large body of work using computational approaches to study conformational changes. Even with the growth of computer power and the impressive results generated from computational study, computer modeling and simulation of biomolecules remain challenging in general, since molecular systems follow complex physical laws and involve high dimensional conformation spaces. In this work, we have developed a hybrid Probabilistic Roadmap and Monte Carlo planner for biomolecule conformation study. The planner has been designed to integrate the strengths of both probabilistic roadmap and Monte Carlo simulation methods, and its effectiveness has been demonstrated in our promising preliminary results.

2 Prior Work

Monte Carlo (MC) methods have been successfully applied to statistical physics and chemistry where the motion of complex systems of thousands of atoms is studied. Given an initial conformation of a molecular system, a new conformation is generated through a perturbation of the initial conformation. If the energy of the new conformation is lower than that of the initial one, the new conformation is accepted. Otherwise, the new conformation is accepted with some probability: the higher the new energy, the lower the acceptance rate. Such a method can be used to generate a trajectory of molecular conformations obeying the Boltzmann distribution. While the algorithmic simplicity and theoretical grounding have made MC appealing to a wide variety of problems, MC methods are quite computationally expensive. For large molecules, random Monte Carlo moves generally result in high rejection rates and long running time. Furthermore, each simulation run is generally treated independently from other runs, without using any knowledge generated from prior runs. Regarding the high rejection rate problem, it has been identified that perturbing one or few atoms at each step generally leads to a low percentage of rejections. This type of moves also facilitates efficient computation of molecular structures and energy[4] since a small number of perturbed atoms will partition a molecule chain into a small number of subchains, where the perturbation only causes non-rigid relative motion between the subchains and only the energy terms involving atoms of different subchains need to be updated. As for reusing prior computation results to improve the simulation efficiency, we believe that probabilistic roadmap (PRM) methods provide a natural framework.

[1]Department of Mathematics and Computer Science, Clark University, Worcester, MA 01610, USA. E-mail: lhan@clarku.edu

PRM [2] was originally developed for robot motion planning problems and has been successfully adapted to studying computational biology problems[1, 3]. Given a molecular system, the general methodology of PRM planners is to construct a graph(roadmap) to capture the properties of the conformation space. Roadmap vertices are randomly sampled conformations with their inclusion in the roadmap determined by some acceptance criteria, and roadmap edges correspond to transitions between 'nearby' vertices found with simple local planning methods such as the linear interpolation, again with inclusion of edges subject to some acceptance conditions such as the Metropolis criteria. After generating a roadmap, numerous trajectories between pairs of conformations can be extracted from the graph. This is one distinct feature of PRMs as compared to Monte Carlo simulation, where each successful simulation run generates one trajectory and many simulation runs simply fail. The success of PRM can be partly attributed to its efficient reuse of nodes and edges in the roadmap.

3 Our Methods

PRM planners generally need to sample and connect a large number of conformations in order to build a roadmap that reflects the properties of the conformation space. Conventionally, roadmap nodes are generated independently from each other. Such an approach need to compute the energy of each conformation from scratch and move most, if not all, atoms to connect each pair of conformations, which contribute to long computation times and low acceptance rates.

We have recently developed a hybrid Probabilistic Roadmap and Monte Carlo (PRM-MC) planner to combine the favorable properties of PRM – capturing the connectivity of conformations and facilitating the reuse of already identified conformations and conformational changes – with those of MC – efficiently updating molecular conformations and energy after perturbations of a small number of atoms at each step. First the planner generates a small number of seed conformations. These seed conformations can be experimentally determined structures stored in PDB or random conformations generated from some conformation space sampling schemes. Then the planner uses Monte Carlo simulation to generate and connect conformations originated from the perturbation of the seed conformations. Finally the planner uses PRM type connection strategies to try to find more energetically feasible conformational changes that have not been identified in the Monte Carlo step. Our preliminary simulation results have indicated that the PRM-MC planner can efficiently generate roadmaps useful for the study of molecule conformation space and conformational changes.

References

[1] N. M. Amato, Ken A. Dill, and G. Song. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. In *Proc. Int. Conf. Comput. Molecular Biology (RECOMB)*, pages 2–11, 2002.

[2] L. Kavraki, P. Svestka, J. C. Latombe, and M. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Automat.*, 12(4):566–580, August 1996.

[3] S.M. LaValle, P.W. Finn, L.E. Kavraki, , and J.C. Latombe. A randomized kinematics-based approach to pharmacophore-constrained conformational search and database screening. *Journal of Computational Chemistry*, 21(9):731–747, 2000.

[4] I. Lotan, F. Schwarzer, D. Halperin, and J.C. Latombe. Algorithm and data structure for efficient energy maintenance during monte carlo simulation of proteins. *Journal of Computational Biology*, 2003.

J12. A Physical Scoring Function Based on the AMBER Force Field and the Poisson-Boltzmann Implicit Solvent for Protein Structure Prediction

Mengjuei Hsieh [1] **and Ray Luo** [2]

Keywords: protein folding, ab initio, molecular mechanics, finite difference, surface area, decoys.

1 Introduction.

Protein structure prediction at atomic detail, an important aspect of the protein folding problem, remains one of the fundamental unsolved problems in the field of computational molecular biology. The final stage of protein structure prediction usually involves ranking or evaluating a protein model with a scoring function, an algorithm that gives a score for an input structure to its fitness, that is used to judge the models likelihood of being the native structure, or at least of being close to the native. There are two classes of scoring functions: knowledge-based and physics-based approaches[1, 2]. The two scoring functions are constructed from very different starting points. Knowledge-based approaches are derived from distributions of experiment structural data[3]. Physics-based approaches assume that the protein potential energy function can be broken down into terms of bond stretching, angle bending, torsional and nonbonded interactions.

2 Methods and results.

A well-behaved physics-based all-atom scoring function for protein structure prediction is analyzed with several widely used all-atom decoy sets. This scoring function, termed as AMBER/PB, is based on a refined AMBER force field[4] for intramolecular interactions and an efficient Poisson-Boltzmann model for solvation interactions[5, 6]. Testing on the chosen decoy sets shows that the scoring function, designed to consider detailed chemical environments, is able to consistently discriminate all 62 native crystal structures after considering the heteroatom groups, disulfide bonds, and crystal packing effects that are not included in the decoy structures. When NMR structures are considered in the testing, the scoring function is able to discriminate 8 out of 10 targets. In the more challenging test of selecting near-native structures, the scoring function also performs very well: for the majority of the targets studied, the scoring function is able to select decoys that are close to the corresponding native structures as evaluated by ranking numbers and backbone Cα RMSD. Overall summary is shown in Table 1. Various important components of the scoring function were also studied to understand their discriminative contributions towards the rankings of native and near-native structures. It was found that neither the non-polar solvation energy as modelled by the SA model nor a higher protein dielectric constant improve its discriminative power. The terms remained to be improved are related to 1-4 interactions. We found that the most troublesome term is the large and highly fluctuating 1-4 electrostatics term, but

[1]Department of Molecular Biology and Biochemistry, University of California Irvine, Irvine CA 92697-3900. E-mail: mengjueh@uci.edu

[2]Department of Molecular Biology and Biochemistry, University of California Irvine, Irvine CA 92697-3900. E-mail: rluo@uci.edu

not the torsion-angle term. These data support ongoing efforts in the community to develop protein structure prediction methods with physics-based potentials that are competitive with knowledge-based approaches.

3 Tables.

Decoys Sets	Discrimination	RMSD Rank	Best RMSD Deviation
EMBL misfolded	18 / 19	N/A	N/A
CASP1 homology	5 / 5	1∼2 / 10	0.00∼0.89
Globin homology	30 / 30	1∼6 / 29	0.20∼2.98
Park & Levitt ab initio	7 / 7	2∼28 / 650	0.00∼0.68
MMPBSA ab initio	11 / 12	6∼20 / 30	0.58∼0.84

Table 1: Performance of the AMBER/PB scoring function on decoy sets. Discrimination: number of native structures discriminated out of the total decoy number. RMSD rank: ranking position of the lowest Cα RMSD by the scoring function out of total number of decoys. Best RMSD: refers to the lowest Cα RMSD in the top 5 ranked decoys. Best RMSD Deviation: The deviation from the lowest RMSD in the decoy set to the Best RMSD (Å). n/a: not applicable due to insufficient data.

References

[1] Moult, J. 1997. Database Potentials and Molecular Mechanics Force Fields. In: *Current Opinion in Structural Biology,* 7(2):194199.

[2] Lazaridis, T., Karplus, M. 2000. Effective Energy Functions for Protein Structure Prediction. *Current Opinion in Structural Biology,* 10(2):139145.

[3] Sippl, M. J. 1995. Knowledge-Based Potentials for Proteins. In: *Current Opinion in Structural Biology,* 5(2):229235.

[4] Lu, J. Q., Luo, R. 2003 Optimization of the Main Chain Torsional Term for Protein Simulations. *Journal of Physical Chemistry,* submitted.

[5] Luo, R., David, L., Gilson, M. K. 2002. Accelerated Poisson-Boltzmann Calculations for Static and Dynamic Systems. *Journal of Computational Chemistry.* 23:12441253.

[6] Lu, J. Q., Luo, R. 2003. A Poisson-Boltzmann Dynamics Method with the Non-periodic Boundary Condition. *Journal of Chemical Physics.* 119(21):000000.

J13. Mining Spatial Motifs from Protein Graph Databases

Jun L. Huan[1], **Wei Wang**[1], **Deepak Bandyopadhyay**[1], **Jack Snoeyink**[1], **Jan Prins,** [1]
Alexander Tropsha [2]

Keywords: subgraph mining, almost-Delaunay, protein classification

1 Introduction.

Finding recurring structural features among protein three-dimensional (3D) structures is an important problem in bioinformatics. We apply a novel subgraph mining algorithm to three related graph representations of the sequence and proximity characteristics of a protein's amino acid residues. The subgraph mining algorithm is used to discover spatial motifs that can be used to discriminate among proteins in different families found in the SCOP database.

Protein structure may be modeled using a variety of graph representations [6]. Our approach uses a labeled graph in which the nodes represent the amino-acid residues comprising the protein, and the edges represent proximity relations among the residues. Two types of edges are identified: a bond edge connects two residues that are contiguous in the primary sequence, and a proximity edge connects two (non-bonded) residues within a given distance δ of each other. Spatial motifs appear as recurring subgraphs among a set of proteins represented in this fashion.

A fundamental challenge for applying frequent subgraph mining to find patterns from a group of proteins is the huge list of frequent subgraphs. As the underlying operation of subgraph isomorphism testing is NP-complete, it is critical to minimize the number of frequent subgraphs that need to be considered. We investigate techniques to reduce the number of edges in a contact graph (CG). In particular, we choose edges from the Delaunay tessellation and its recently developed extension to almost-Delaunay [1]. The Delaunay tessellation graph (DG), based on Delaunay tessellation of the protein backbone residues, captures neighbor relations between points representing residues or atoms. It has been used to analyze packing [3, 5] and structure [2, 4] in proteins. The almost-Delaunay edges (AD) expand the set of Delaunay edges to account for perturbation or motion of point coordinates, controlled by a parameter ϵ. A property of these representations is that $DG \subseteq AD(\epsilon) \subseteq CG$ for all $\epsilon \geq 0$.

The results indicate that the Delaunay and almost-Delaunay subsets of the contact graph are robust, sparse, and adequate to produce simplified graphs for mining spatial motifs, yielding motifs with discrimination qualities similar to, or better than, those obtained from the full contact graph.

2 Experimental Results

We applied the frequent subgraph mining (support is set to 90%) on the Delaunay graph constructed from the SCOP serine protease family using 10 Å for distance prune. The mined subgraphs are then used to query the PDB-select 60 dataset, which contains 4672 diverse protein structures with pair-wise similarity between any two sequences is no more than 60%. Those subgraphs having low background frequency (0.6%) are selected and reported. In Figure 1, we present the largest motif which is specific to the serine protease family according to our setup.

[1]Department of Computer Science, University of North Carolina at Chapel Hill, E-mail: {huan, weiwang, debug, Snoeyink, prins}@cs.unc.edu
[2]Laboratory of Molecular Modeling, School of Phamacy, University of North Carolina at Chapel Hill, E-mail: tropsha@email.unc.edu

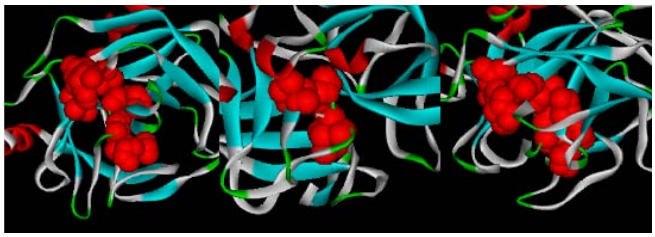


Figure 1: the largest spatial motif which are specific to the serine protease (SP) family. We show the motif's occurrence in three members of the SP family: Human Kallikrein 6 (Hk6) (1L06, left), Porcine Pancreatic Elastase Acyl Enzyme (1GVK, middle) and Human Coagulation Factor Ixa (1RFN, right).

We also applied the coherent subgraph mining algorithm, a postprocessing technique to extract features fast from a graph database, to multiple SCOP families. The extracted features are then used for binary classification. The first dataset (C_1) included two protein families from different SCOP superfamilies: the nuclear receptor ligand-binding domain proteins and the prokaryotic serine protease family. The second dataset (C_2) included the families of eukaryotic serine proteases and the prokaryotic serine proteases, which belong to the same SCOP superfamily.

Table 1 summarizes the total number of features we identified from different graph representations and the classification accuracy. From the table, we can see that the Delaunay graph, which mines significantly small number of features, still preserves the high classification accuracy.

Data SetC_1	Features	Dist. F.	Accu%
DG	20,646	934	100
AD(0.1)	23,130	1093	100
AD(0.25)	26,943	1234	96
AD(0.5)	32,463	1582	100
AD(0.75)	37,394	1674	96
CG	40,274	1859	95

Data SetC_2	Features	Dist. F.	Accu%
DG	15,895	20	95
AD(0.1)	18,491	29	95
AD(0.25)	23,288	35	93
AD(0.5)	29,083	35	95
AD(0.75)	32,569	36	95
CG	34,697	20	98

Table 1: Binary Classification Results - number of distinguishing features and classification accuracy with different graph representations: DG (Delaunay) and AD (almost-Delaunay, with perturbation in the parenthesis), CG (contact graph).

References

[1] D. Bandyopadhyay and J. Snoeyink. Almost-Delaunay Simplices : Nearest Neighbor Relations for Imprecise Points. Symposium On Distributed Algorithms. 403-412. 2004

[2] J. Liang, H. Edelsbrunner, P. Fu, P.V. Sudhakar and S. Subramaniam. Analytical shape computing of macromolecules I: molecular area and volume through alpha shape. Proteins 33 1-17. 1998.

[3] F. M. Richards. The interpretation of protein structures: total volume, group volume distributions, and packing density. J. Molecular Biology 82. 114, 1974.

[4] R.K. Singh, A. Tropsha and I.I. Vaisman. Delaunay tessellation of proteins. J. Comput. Biol. 3. 212-22. 1996.

[5] J. Tsai, R. Taylor, C. Chothia and M. Gerstein. The Packing Density in Proteins: Standard Radii and Volumes. J. Molecular Biology 290. 253-66. 1999.

[6] S. Vishveshwara, K. V. Brinda and N. Kannan. Protein Structure: Insights from graph theory. J. of Theo. and Comp. Chem. 2002.

J14. Molecular dynamics simulation of branch migration in RuvA tetramer - Holliday junction DNA complex

Hisashi Ishida[1], and Nobuhiro Go[1,2]

Keywords: branch migration, RuvA tetramer, Holliday junction DNA, molecular dynamics

1 Introduction.

Homologous recombination is important in the generation of genetic diversity and in DNA repair in all organisms. One of the crucial steps in recombination is branch migration of a Holliday junction. The binding of a tetrameric RuvA to the Holliday junction DNA is followed by the loading of RuvB which in turn drives the branch migration. However, the structure of Holliday junction DNA complexed with RuvB, has not yet been solved by X-ray crystallography. Therefore, in order to perform the molecular dynamics (MD) simulation of branch migration, it is necessary to substitute the force generated by RuvB as a result of ATP hydrolysis with an artificial force.

In order to understand in the molecular level how RuvA recognizes the Holliday junction and how branch migration occurs, MD simulations of RuvA – Holliday junction DNA complex were performed with and without external forces on the Holliday junction DNA.

2 Methods

First, MD simulations of (1) an uncomplexed Holliday junction DNA (25 nucleotides × 4 strands), and (2) RuvA tetramer – Holliday junction DNA complex (PDB code: 1C7Y [1]) were performed for 1nano second at a constant pressure, one bar, and temperature, 300K.

Second, MD simulations of modeled RuvA tetramer – Holliday junction DNA Complex were performed by substituting the DNA sequence of 1C7Y with the sequences of poly(dA)-poly(dT), poly(dT)-poly(dA), poly(dC)-poly(dG), poly(dG)-poly(dC) for 1nano second at the same pressure and temperature. Then the external screw forces of 5pN in the direction of DNA groove were applied on 24 phosphor atoms of 15A-20A, 6B-11B, 15C-20C and 6D-11D (see Fig.1) along the translocation. The MD simulations of branch migration were performed for 500ps at a constant pressure, one bar, and temperature, 300K. These simulations were carried out using the ABMER7 program.

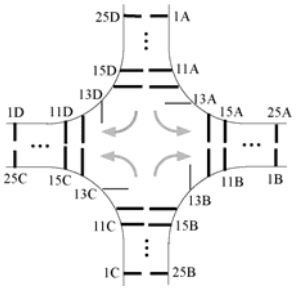


Figure 1: Schematic picture of Holliday junction DNA with numbering from A-strand to D-strand. 13th bases are unpaired. Arrows at the center indicate the direction of branch migration.

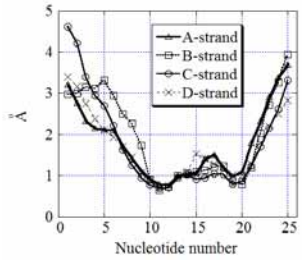
3 Results

1. Atomic fluctuations in thermal equilibrium

[1] Center for Promotion of computational Science and Engineering, Japan Atomic Energy Institute, Japan. E-mail: ishida@apr.jaeri.go.jp

[2] Graduate School of Information Science, Nara Institute of Science and Technology, Japan. E-mail: go@apr.jaeri.go.jp, ngo@is.aist-nara.jp

It was found that the hole at the center of the uncomplexed Holliday junction became smaller while the hole at the center of RuvA tetramer – Holliday junction DNA complex remained wide. Moreover, it was found that the atomic fluctuations of the 11th and 19th nucleotides interacting with HhH motifs of RuvA were small and the atomic fluctuations of nucleotides between the 12th and 18th nucleotides were rather large (Fig. 2). The same results were obtained for the modeled RuvA tetramer – Holliday junction DNA Complex with the DNA sequences of poly(dA)-poly(dT), poly(dT)-poly(dA), poly(dC)-poly(dG), poly(dG)-poly(dC).



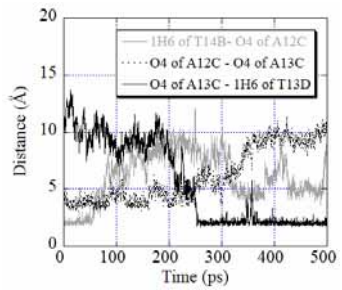
It is considered that the acidic pins (the carboxyl side chains of Glu55 and Asp56) located in the center of the RuvA tetramer kept the DNA junction center wide and mobile to facilitate the branch migration. However, the spontaneous branch migration is unlikely to be observed in the scale of a 1nano scale simulation. Therefore, in order to observe the branch migration by molecular dynamics, steered MD simulations were carried out by applying artificial external forces which facilitate the branch migration.

Figure 2: Average RMS fluctuations for heavy atoms of the bases in Holliday junction DNA against the nucleotide number

2. Initial process of branch migration

In the steered MD simulation, the initial process of disconnecting and reforming of hydrogen bonds at Holliday junction center was seen in the system of poly(dA)-poly(dT), poly(dT)-poly(dA) although it was not observed in the system of poly(dC)-poly(dG), poly(dG)-poly(dC). This may imply that the three hydrogen bonds between cytosine and guanine are strong enough to tolerate the external force for 500ps.

In the system of poly(dA)-poly(dT), for example, Fig. 3 shows that the disconnection and reformation of hydrogen bonds between T14B and A12C (14^{th} Thymine of B-strand and 12^{th} Adenine of C-strand), A13C and T13D. A pair of T14B and A12C was broken quickly at 50ps although A12C-A13C stack remained until the reformation of hydrogen bonds between A13C and T13D at 250ps. It is considered that the disconnection of the hydrogen bonds preceding the unpaird 13^{th} bases triggered the separation of the two strands, and could facilitate the complete disruption of base pairing and stacking that accompanies strand exchange at the cross-over point.



It is suggested that the acidic pins modulate the process of the branch migration [2]. During the simulation of migration, the change of interaction between the unpaired nucleotides and the acidic pins through water-mediated bonding was seen when A12C-A13C stack started to break. It is considered that the water between the unpaired nucleotides and the acidic pins may be crucial for the process of the branch migration.

Figure 3: Time evolution of the distance between 1H6 of T14B and O4 of A12C, O4 of A12C and O4 of A13C, O4 of A13C and 1H6 of T13D

References

[1] Ariyoshi, M., Nishino, T., Iwasaki, H., Shinagawa, H. and Morikawa, K. 2000. Crystal structure of the Holliday junction DNA in complex with a single RuvA tetramer. *Proceedings of the National Academy of Sciences USA* 97:8257-8262.

[2] Ingleston, S.M., Sharples, G.J. and Lloyd, R.G. 2000. The acidic pin of RuvA modulates Holliday junction binding and processing by the RuvABC resolvasome. *The EMBO Journal* 19:6266-6274.

J15. Flexible Docking of Peptides to MHC

Joo Chuan Tong[1], Lesheng Kong[2], Tin Wee Tan[3], Shoba Ranganathan[4]

Keywords: Major Histocompatibility Complex, Epitope prediction, Flexible Docking

1 Introduction.

Major histocompatibility complex (MHC) molecules are highly polymorphic cell surface molecules that present antigenic peptides to cells of the T-cell compartment of the immune system. In the design of molecular vaccines for the treatment of diseases, identification of T-cell epitopes from immunologically relevant antigens is an important prerequisite. However, the experimental identification of T cell epitopes is a time consuming and expensive process due to the large number and diverse nature of MHC alleles and candidate peptides. This study presents a new protocol for rapid and precise docking of peptides to MHC class I and class II receptors.

2 Methods

Our docking procedure consists of three steps: (i) peptide residues near the ends of the binding groove are docked by using an efficient pseudo-Brownian rigid body docking procedure followed by (ii) loop enclosure of the remaining backbone structure by satisfaction of spatial constraints and subsequently (iii) refinement of the entire backbone and ligand interacting side-chains and receptor side-chains around atomic clash regions between MHC receptor and peptide.

3 Results

Evaluation of our modeling procedure is performed systematically in the following two steps: (1) Rebuilding of 40 test case complexes; and (2) docking of 15 solved peptides into templates of appropriate alleles. In the first test, 33 out of 40 MHC class I and class II peptides can be modeled with Cα RMSD < 1.00 Å. In the second test, 11 out of 15 MHC-peptide complexes were modeled with Cα RMSD < 1.00 Å. To the best of our knowledge, these results represent up to five-fold increase in accuracy compared to available techniques in the remodeling of MHC-peptides and up to three times increase in speed.

[1] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260.
Email: victor@bic.nus.edu.sg
[2] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260.
Email: lesheng@bic.nus.edu.sg
[3] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260.
Email: tinwee@bic.nus.edu.sg
[4] Research Institute for Biotechnology, Macquarie University, NSW 2109, Australia.
Email: shoba@bic.nus.edu.sg

4 Figures and tables.

Peptide	Technique	Author	RMSD[a]	RMSD[b]
TLTSCNTSV	Simulated Annealing	Rognan et al [5]	1.04	0.46, 0.58
FLPSDFFPSV	Simulated Annealing	Rognan et al [5]	1.59	1.10, 1.48
GILGFVFTL	Simulated Annealing	Rognan et al [5]	0.46	0.32, 0.38
ILKEPVHGV	Simulated Annealing	Rognan et al [5]	0.87	0.87, 0.73
LLFGYPVYV	Simulated Annealing	Rognan et al [5]	0.78	0.33, 0.69
RGYVYQGL	Combinatorial Buildup Algorithm	Desmet et al [4]	0.56	0.32, 0.66
FAPGNYPAL	Multiple copy Algorithm	Rosenfeld et al [6,7]	2.70	0.40, 0.90
GILGFVFTL	Multiple copy Algorithm	Rosenfeld et al [6,7]	1.40	0.32, 0.38
LLFGYPVYV	Combinatorial Buildup Algorithm	Sezerman et al [8]	1.40	0.33, 0.69
ILKGPVHGV	Combinatorial Buildup Algorithm	Sezerman et al [8]	1.30	0.87, 0.73
GILGFVFTL	Combinatorial Buildup Algorithm	Sezerman et al [8]	1.60	0.32, 0.38
TLTSCNTSV	Combinatorial Buildup Algorithm	Sezerman et al [8]	2.20	0.46, 0.58

Table 1: Benchmarking of our MHC-peptide procedure with previously published works in MHC class I peptide modeling. [a]RMSD of peptide backbone atoms obtained from the works of respective authors. [b]RMSD of peptide backbone atoms obtained in our work from redocking bound complexes and docking into single template respectively.

5 References

[1] Abagyan R, Maxim Totrov. 1999. Ab Initio Folding of Peptides by the Optimal-Bias Monte Carlo Minimization Procedure. J. Comput. Phys 151: 402-421.

[2] Fernández-Recio J, Totrov M, Abagyan R. 2002. Soft protein-protein docking in internal coordinates. Protein Sci. 11, 280-291.

[3] Sali A, Blundell TL. 1993. Comparative Protein Modelling by Satisfaction of Spatial Restraints. J.Mol.Biol. 234, 779-815.

[4] Desmet J, Maeyer MD, Spriet J, Lasters I. 2000. Flexible Docking of Peptide Ligands to Proteins. Methods Mol Biol 143, 359-376.

[5] Rognan D, Laumoeller SL, Holm A, Buus S, Tschinke V. 1999. Predicting Binding Affinities of Protein Ligands from Three-Dimensional Models: Application to Peptide Binding to Class I Major Histocompatibility Proteins. J. Med. Chem. 42, 4650-4658.

[6] Rosenfeld R, Zheng Q, Vajda S, DeLisi C. 1995. Flexible docking of peptides to class I major-histocompatibility-complex receptors. Genet Anal. 12, 1-21.

[7] Rosenfeld R, Zheng Q, Vajda S, DeLisi C. 1993. Computing the Structure of Bound Peptides: Application to Antigen Recognition by Class I Major Histocompatibility Complex Receptors. J.Mol.Biol. 234, 515-521.

[8] Sezerman U, Vajda S, DeLisi C. 1996. Free energy mapping of class I MHC molecules and structural determination of bound peptides. Protein Sci. 5, 1272-1281.

J16. Protein Families Classification using Support Vector Machine

Joo Chuan Tong[1], Khar Heng Choo[2], Teck Kwong Lee[3], Lesheng Kong[4], Soon Heng Tan[5], Tin Wee Tan[6], Shoba Ranganathan[7]

Keywords: Support Vector Machines, Transporter Protein Families, Protein Structure

1 Introduction.

Proteins play an important role in biological processes and detailed knowledge about protein functions is fundamental to understand the complex biological pathways that occur in living organisms. A large number of sequence information has been experimentally determined and deposited in numerous databases. However, only a small fraction of protein sequences have been experimentally characterized. In this context, theoretical prediction of protein functions is becoming critically important in furthering our understanding of biological processes. In recent years, several groups have adopted the use of SVM as a prediction tool for protein functional family classification. This study investigates the 11 amino acid attributes (surface tension, hydrophobicity, normalized Van der Waals volume, relative mutability, polarity, polarizability, charge, bulkiness, solvent accessibility, predicted secondary structure and predicted trans-membrane region) to the accuracy of transporter protein family classification using the leave-one-property-out procedure.

2 Methods

Our dataset currently covers 81 protein functional families from the class 2A super family of TCDB [6]. These include 774 Swiss-Prot protein records from TCDB, 1065 Swiss-Prot records collected from Pfam, and 9764 NCBI's non-redundant protein database records. The program SVM^{light} [3] was adopted in this study for SVM calculations. Secondary structure assignment was performed using PSIPRED [4] and trans-membrane assignment was achieved using MEMSAT [5].

3 Results

[1] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Email: victor@bic.nus.edu.sg
[2] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Email: justin@bic.nus.edu.sg
[3] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Email: bernett@bic.nus.edu.sg
[4] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Email: lesheng@bic.nus.edu.sg
[5] Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613. Email: soonheng@i2r.a-star.edu.sg
[6] Department of Biochemistry, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260. Email: tinwee@bic.nus.edu.sg
[7] Research Institute for Biotechnology, Macquarie University, NSW 2109, Australia. Email: shoba@bic.nus.edu.sg

A three-fold cross validation was adopted to investigate the contribution of each amino acid property on the accuracy of the classifier. From the amino acid properties used in our current study, 11 iterations were performed using the leave-one-property-out validation with the exclusion of a specific property during each train-test phase in order to determine the significance of each property on the accuracy of prediction. Our study reveals that radial basis function presents the best performing kernel function and the exclusion of polarity (#5) and polarizability (#6) from the protein-chain descriptors resulted in the optimal prediction accuracy of 84.00% to 99.95% for 81 functional families from the 2A family of transporter in Transporter Commission Database (TC-DB). Refined versions of our SVM can serve as a useful tool for transporter protein classification of the entire TC-DB. In this way, the effort to create a universal classification system for all currently identified and yet-to-be recognized transport proteins could be greatly facilitated.

4 Figures and tables.

Property	All	Exclude Property										
		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
Ave. Acc. (%)	99.31	99.10	99.14	99.29	99.28	99.43	99.40	99.27	99.29	99.28	99.21	99.26
Ave. Acc. of top 5 biggest subclasses (%)	97.04	95.76	96.15	96.95	96.93	96.34	96.21	96.94	96.99	96.93	96.41	96.94
Prec/Recall Score Freq	0.543	0.193	0.218	0.523	0.519	0.860	0.852	0.482	0.535	0.519	0.420	0.457

Table 1: SVM prediction accuracy using the leave-one-property-out validation

5 References

[1] Cai CZ, Han LY, Ji ZL, Chen X, and Chen YZ. 2003. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res*. 31, 3692-3697.

[2] Dubchak I, Muchnik I, Holbrook SR, and Kim SH. 1995. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U. S. A. 92*, 8700-8704.

[3] Joachims T. 2001. Learning To Classify Text Using Support Vector Machines - Methods, Theory, *Algorithms*. Kluwer Academic Publishers

[4] Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*. 292, 195-202.

[5] Jones DT, Taylor WR, and Thornton JM. 1994. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry (Mosc)*. 33, 3038-3049.

[6] Saier MH Jr. 2000. A functional-phylogenetic classification system for transmembrane solute transporters. *Microbiol. Mol. Biol. Rev*. 64, 354-411.

J17. Combining structure and function information in a local alignment search tool for sequence-sequence comparison.

Maricel Kann[1], Paul Thiessen[1], Anna Panchenko[1], Alejandro Schaffer[1], Stephen F. Altschul[1] and Stephen H. Bryant[1].

Keywords: protein sequence alignment, alignment algorithms, statistical significance, similarity search.

1 Introduction.

With thousands of recently sequenced proteins, sequence-sequence comparison methods have become the most widely used tool in bioinformatics and related fields. Several popular methods (i.e., PSI-BLAST and IMPALA) [1, 2] implement the search using a position-specific scoring matrix (PSSM) as a scoring scheme, which captures important information about the protein sequence, and achieves a fast and effective search for sequence homologies. However, explicit information about the structure and function of proteins in the databases during the search is difficult to include. One could incorporate such information by modeling the indels, represented by gaps in the alignment, and/or by correctly aligning all biologically relevant residues. In most of the algorithms for sequence comparison, however, the choice of the gap penalties is determined ad-hoc, gaps occur anywhere, and key residues are aligned only if this increases the total score for the alignment. In this paper, we introduce an algorithm, Structure-based Local Algorithm Method or SLAM; that uses a new approach for the placements and penalization of gaps with a novel approach for the definition of aligned regions. This algorithm produces a local alignment between a query protein sequence and a database of PSSMs, and uses a scoring scheme based on the differences in conservation along the protein sequences. Highly-conserved regions (or blocks) will be aligned from start to end without any gaps and without penalties for the insertion of gaps (up to a maximum length) between those blocks. For the classification of the families, definition of the blocks or conserved domains (CDs) and PSSMs, we use an approach developed by our group in which multiple sequence alignment of related sequences have been manually curated to represent the function of that family. The scores obtained using SLAM follow an extreme value distribution which allows the correct estimation of their statistical significance. The fact that CD database have been manually curated with key residues necessary for the specific function of each of the families and inter-block regions carefully selected, suggests that SLAM's alignments are highly biologically significant. SLAM's performance in the search for biological relationships, alignment accuracy and speed is very similar to IMPALA's. Based on these results, SLAM has been successfully included into Cn3D (tool for visualization of protein structures) as an alternative choice for the alignment of new sequences.

References

[1] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 17: pp. 3389-3402.

[1] Computational Biology Branch, National Center for Biotechnology Information, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA.
E-mail: kann@mail.nih.gov

[2] A. A. Schaffer, Y. I. Wolf, C. P. Ponting, E. V. Koonin, L. Aravind and S. F. Altschul. 1999. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics. 12: pp. 1000-1011.

J18. Native and non-native oligopeptide fragments biased to alpha-helical formation

Gelena Kilosanidze[1], Alexey Kutsenko[2]

Keywords: protein secondary structure, structure prediction, molecular mechanics, energy contributions

1 Introduction.

Secondary structure prediction from protein sequences is the most studied problem in computational biology. Current matter of thinking implies that a complete solution to all aspects of the protein structure problem require that the information available from the analysis of large structural and sequence databases be combined with the physical-chemical principles of structure formation.
The approach of secondary structure prediction proposed by us [2,3] is *ab initio* method and uses a molecular mechanics for the analysis of standard conformations in proteins. A model for prediction of α-helical regions in amino acid sequences has been tested on the mainly-α protein structure class. The modeling represents the construction of a continuous hypothetical α-helical conformation for the whole protein chain, and was performed using molecular mechanics program ICM [1] that represents macromolecules in internal coordinate system. The results of the modeling clearly demonstrated that the profile of energy along the model α-helical protein reveals minima corresponding to real α-helical segments in the native protein.

2 Method and results.

In spite of good performance for α-helices the method is rather computer time consuming (about 6h for protein of 100 residues on a regular PC). To move to the fast statistical prediction based on molecular modelling we have undertaken an analysis of an energy distribution of all possible short oligopeptides, natural and non-natural, involved in α-helical pattern. For oligopeptide in length of 4 we modeled 20^4 = 160 000 structures. In order to reveal a contribution of every tetrapeptide to forming α-helix we constructed and optimized polyalanine α-helices with analyzed tetrapeptide built into the central area of such model – $Ala_{10}\mathbf{X_4}Ala_{10}$. The energy profiles of all possible combinations of tetrapeptides were calculated (Fig. 1). The central, 11th point of the energy profiles corresponds to value of energy of a tested combination of four amino acids. Analysis of profiles results in some conclusions:

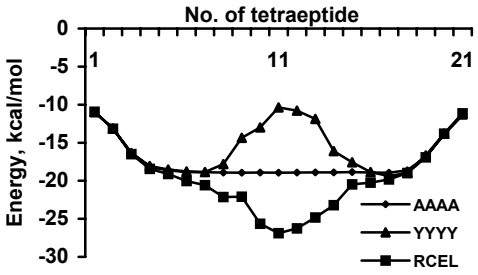


Figure 1: Energy profiles of optimized α-helical models of polyalanine chains with insertions of AlaAlaAlaAla, TyrTyrTyrTyr or ArgCysGluLeu at the 11th position.

[1] Microbiology and Tumor Biology Center, Karolinska Institute, Stockholm, Sweden. E-mail: Gelena.Kilosanidze@mtc.ki.se
[2] Microbiology and Tumor Biology Center, Karolinska Institute, Stockholm, Sweden. E-mail: Alexey.Kutsenko@mtc.ki.se

(1) The fact that end-regions of profiles coincide at different models after optimization of their structure, suggests that the central sites of these profiles can be compared with each other.
(2) Energy of polyalanine tetrapeptide fragments on the middle part of a chain, not subjected to edge effects (tetrapeptides 5th to 17th), is -18,9 kcal/mol and does not change.
(3) Energy of the other tetrapeptide fragments in the middle part of a chain can accept values above or below the polyalanine level and unambiguously characterize given tetrapeptide.

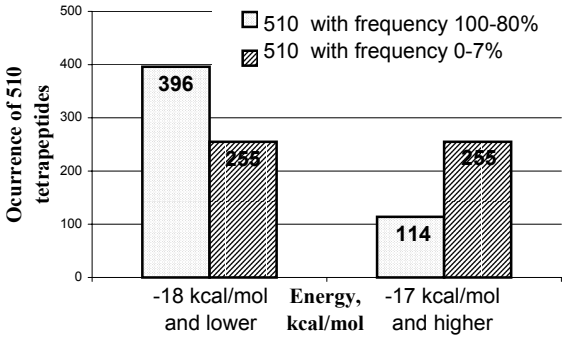


Figure 2: Energy distribution of modeled in α-helix tetrapeptides from two groups (of 510 combinations in each) that differ by occurrence in the nature.

This approach promises great benefit – exploring non-natural peptides. The estimated number of tetrapeptides that occur in native proteins is about 9000 [4], which is significantly less than the all-possible combinations of oligopeptide in length of four. Some of the native tetrapeptide fragments always prefer to be in α-helical conformation, others never adopt that type of secondary structure. Energies of model tetrapeptides will result in bias of given oligopeptides to the specific type of secondary structure. After the energy analysis of all oligopeptides one could suggest such amino acid combination that have not yet been found in nature but can adopt some kind of secondary structure with high efficiency, which will be of interest for design of *de novo* proteins.
We performed the exhaustive and statistical analyses of oligopeptide occurrence of in different conformations in PDB database [4]. That distribution was compared with calculated energy distribution.
The analysis shows that tetrapeptide fragments occurring with probability of 80 to 100% in the nature in α-helices preferably have energy values of –18 kcal/mol and lower in the model (Fig.2). Thus, this energy level serves as a cutoff benchmark to divide tetrapeptide fragments into favorable and unfavorable for α-helices formation.

References

[1] Abagyan, R.A. and Totrov, M. 1999. *Ab Initio* Folding of Peptides by the Optimal-Bias Monte Carlo Minimization Procedure. *J. Comp.Phys.*, 151:402-421.

[2] Kilosanidze, G.T., Kutsenko, A.S., Esipova, N.G. and Tumanyan, V.G. 2002. Use of molecular mechanics for secondary structure prediction. It is possible to reveal α-helix? *FEBS Lett.* 510:13-16.

[3] Kilosanidze, G.T., Kutsenko, A.S., Esipova, N.G. and Tumanyan, V.G. 2003. Analysis of forces leading the helix forming in α-proteins. *Protein Science* (in press).

[4] Vlasov, P.K., Kilosanidze, G.T., Ukrainskii, D.L., Kuz'min, A.V., Tumanian, V.G. and Esipova, N.G. 2001. Left-handed conformation of poly-L-proline type II in globular proteins. *Biofizika*. 46:573-576.

J19. Web-based Prediction of Membrane Spanning β-strands in Outer Membrane Proteins

M. Michael Gromiha[1], Shandar Ahmad[2], Makiko Suwa[1]

Keywords: transmembrane strand, neural network, cation-π interaction, hydrophobicity

1 Introduction

Outer membrane proteins perform a variety of functions, such as mediating non-specific, passive transport of ions and small molecules, selectively passing the molecules like maltose and sucrose and are involved in voltage dependent anion channels. These proteins contain β-strands as their membrane spanning segments (known as transmembrane strand proteins, TMS). The amino acid sequence in the membrane part of TMS proteins contains several polar and charged residues compared with the stretch of hydrophobic residues in transmembrane helical proteins. Hence, most predictive schemes, which are successful in predicting transmembrane helical segments, fail to predict the transmembrane strand segments. On the other hand, the amino acid sequence of TMS proteins is somewhat similar to all-β globular (BG) proteins and the discrimination of all-β globular and TMS proteins is another task. In this work, we have systematically analyzed the characteristic features of amino acid residues in TMS and BG proteins and revealed the similarities and differences between them. Further, a neural network based algorithm has been developed for predicting the membrane spanning β-strands in outer membrane proteins, which shows an excellent agreement with experimental observations. A web interface has been set up for online prediction and it is available at *http://psfs.cbrc.jp/tmbeta-net/*.

2 Database

A database of TMS proteins was derived from the information about their three-dimensional structures available in literature (1). The residues in membrane spanning β-strands of TMS proteins have been assigned from their three-dimensional structures. The PDB codes of the TM proteins used in the present study are, 1A0SP, 1BXWA, 1BY5A, 1E54A, 1EK9A, 1FEPA, 1OPF, 1OSMA, 1PHO, 1PRN, 1QD6C, 1QJ9A, 2MPRA, 2POR and 7AHLA. These proteins contain 5487 residues and 204 membrane-spanning β-strands. Further, we have set up a database of 138 β-domains for representing BG proteins, containing 23081 residues and 1500 β-strands. These proteins are selected in such a way that they are non-redundant (sequence similarity < 30%) and are solved at high resolution (< 2.5 Å).

3 Structural analysis of outer membrane proteins

A conformational parameter set for the 20 amino acid residues in BG and TMS proteins has been developed from the ratio between the frequency of occurrence (%) of amino acid residues in the β-strand part of globular/membrane proteins and that of the whole protein. We found that all aromatic residues have the preference to be in β-strands of BG and TMS proteins. Pro has the lowest β-value

[1] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST)., Aomi Frontier Building 17F, 2-43 Aomi, Koto-ku, Tokyo 135-0064, Japan
[2] Department of Biochemical Engineering and Science, Kyushu Institute of Technology, Iizuka, Japan

in both BG and TMS proteins. Ile has the highest preference to be in the β-strand of BG proteins and an opposite trend is observed in TMS proteins (1).

The surrounding hydrophobicity of a residue in a protein is computed as the sum of the experimental hydrophobicity of the residues that are occurring within the radius of 8Å from the central residue. We found that the aromatic and hydrophobic residues in BG proteins have higher surrounding hydrophobicity than that in TMS proteins.

We have delineated the inter-residue contacts in BG and TMS proteins based on the sequence and structural information. Residues that are close to each other in space and are distant in sequence are termed as long-range contacts (2). We found that both the positively charged residues are making significant contacts with other residues in the membrane. We have further evaluated the cation-π interactions (CPI) in TMS proteins. Each TMS protein contains an average of 5 CPIs and a good correlation is observed between number of amino acid residues and number of CPIs. Further, most of the CPIs are influenced by long-range contacts (3,4).

4 Prediction of membrane spanning β-strand segments

We have developed an algorithm based on neural networks for predicting the transmembrane β-strands in TMS proteins. We introduced the concept of "residue probability" for assigning residues in transmembrane β-strand segments. The performance of our method is evaluated with single residue accuracy, correlation, specificity and sensitivity.
We observed a good agreement between predicted β-strand segments and experimental observations (5). We have developed a web interface in which users can input the amino acid sequence and obtain the probable membrane spanning segments along with the probability of each residue to be in transmembrane segment (Fig 1). It is available at *http://psfs.cbrc.jp/tmbeta-net/*.

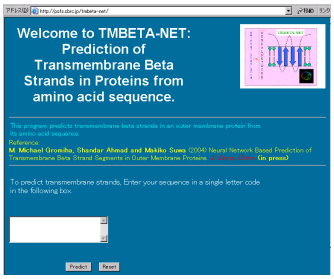


Figure 1: Online prediction of membrane spanning β-strand segments in outer membrane proteins.

5 References

[1] Gromiha, M.M. and Suwa, M. 2003. Variation of amino acid properties in all-β globular and outer membrane Protein Structures. *Int. J. Biol. Macromol.* 32: 93-98.

[2] Gromiha, M.M. and Selvaraj, S. 2004. Inter-residue Interactions in Protein Folding and Stability. *Prog. Biophys. Mol. Biol.* (in press).

[3] Gromiha, M.M. 2003. Influence of cation-π interactions in different folding types of membrane proteins. *Biophys. Chem*. 103, 251-258.

[4] Gromiha, M.M. and Suwa, M. 2004. Structural analysis of residues involving cation-π interactions in different folding types of membrane proteins. *Biopolymers* (submitted).

[5] Gromiha, M.M., Ahmad, S. and Suwa, M. 2004. Neural network based prediction of membrane spanning β-strands in outer membrane proteins. *J. Comp. Chem*. (in press).

J20. Computational Studies of Thioredoxin Superfamily

Efrosini Moutevelis[1]. Jim Warwicker[1]

Keywords: electrostatics calculations, CXXC, rotamers, pH, redox potential

Electrostatic interactions play important roles in diverse biological phenomena, controlling the function of many proteins. Their alterations are reflected in pKa values of ionisable, functional groups and in redox potential. Polar molecules can be studied with the FDPB method solving the Poisson-Boltzmann equation on a Finite Difference grid [1]. We present a method for the prediction of pK_a and redox potentials in the thioredoxin superfamily. The results are compared with experimental pK_a data where available and predictions are made for members lacking experimental pK_a data. The CXXC motif of this superfamily is essential for the catalysis of redox reactions and exhibits extensive variation of redox equilibria. A noteworthy example is the difference between *E. coli* thioredoxin E_0'=-270mV and *E. coli* DsbA E_0'=-122mV [2]. We show how our model, which includes sidechain rotamer variation for the CXXC motif, can be an effective predictive tool for pK_as and redox potential in the superfamily. A qualitative, rather than quantitative, correlation between cysteine pK_a and redox potential indicates that a pH-independent factor also plays a role in determining redox potentials across the superfamily [3]. A possible molecular basis for this feature is proposed. A clustering method that uses matrices of distances improves the speed of our calculations without loosing any accuracy. A pK_a spectrum across members of the superfamily is presented.

[2] Åslund, F. and Beckwith J. 1999. The Thioredoxin Superfamily: Redundancy, Specificity, and Gray-Area Genomics. *JOURNAL OF BACTERIOLOGY* 181:1375-1379

[3] Chivers, P.T., Prehoda, K.E. and Raines, R.T. 1997. The CXXC Motif: A Rheostat in the Active Site. *Biochemistry* 36:4061-4066

[1] Honig B. and Nicholls, A. 1995. Classical Electrostatics in Biology and Chemistry. *SCIENCE* 268:1144-1149

[1] Department of Biomolecular Sciences, UMIST, Manchester, M60 1QD, UK.
E-mail: E.Moutevelis@postgrad.umist.ac.uk

J21. Bounding A Protein's Free Energy In Lattice Models Via Linear Programming

Robert Carr[1], **William E. Hart**[1], **Alantha Newman**[2]

Keywords: Protein structure prediction, linear programming, lattice models, HP model

1 Introduction

The established HP lattice 2D and 3D models have been useful abstractions in understanding protein structure prediction. In these models, a protein folds to maximize H-H contacts (minimize free energy). We analyze and compare integer programming models for the 2D lattice, whose linear relaxations provide non-trivial upper bounds on the maximum number of contacts. These bounds can be used in a branch-and-bound approach to solve the problem optimally and could potentially be used to obtain improved approximation algorithms. In particular, we seek to beat the simple combinatorial bound that arises from the lattice being bipartite.

2 Problem formulation

The Hydrophilic-Hydrophobic (HP) model, introduced by Dill [4], abstracts the dominant force in protein folding: the hydrophobic interaction. The hydrophobicity of an amino acid measures its affinity for water, and the hydrophobic amino acid residues of a protein form a tightly clustered core. In the HP model, each amino acid is classified as an H (hydrophobic) or a P (hydrophilic). The model further simplifies the problem by restricting the feasible foldings to the 2D or 3D square lattice. An optimal conformation for a string of amino acid residues in the HP model is the one that maximizes the number of H-H contacts, which are formed by pairs of H's that occupy adjacent lattice points but are not adjacent on the string.

3 Our approach

We discuss discrete optimization approaches to the problem of protein folding in the Hydrophobic-Hydrophilic (HP) model. We formulate several different integer programs for the problem of protein folding in the 2D HP model and compare the relative strengths of their respective linear programming relaxations. One way to measure the quality of an integer program for a maximization problem is to determine the upper bound guaranteed by its linear relaxation. A linear programming relaxation provides an upper bound on a maximum integral solution and can be solved much more efficiently than an integer program. In general, the tighter (better) the bound provided by the linear relaxation, the higher the quality of the integer programming formulation.

Such methods have been posed previously as a potential approach to protein folding in lattice models [3, 5]. However, the strengths of the proposed LP relaxations were not addressed. For example, we prove that the linear programming relaxation for a natural integer program (described in [3]) provides a solution with value at least twice as much as

[1]Discrete Algorithms and Math Department, Sandia National Laboratories, Albuquerque, NM. E-mail: rdcarr, wehart@cs.sandia.gov

[2]CSAIL, MIT, Cambridge, MA. E-mail: alantha@theory.lcs.mit.edu

the simple combinatorial upper bound for *every* string. However, a strengthened version of this linear program with *backbone constraints* provides a bound that is provably no worse than the simple combinatorial upper bound. We propose additional constraints that may further strengthen these linear programs.

4 Experimental results

For our experiments, we used benchmarks for the problem in the 2D HP model that were taken from: www.cs.sandia.gov/ tech_reports/compbio/tortilla-hp-benchmarks.html. We ran one of our linear programs (LP_3 in [2]) on the following strings:

1. `hphpphhphpphphhpphph`
2. `hhpphpphpphpphpphpphpphh`
3. `pphpphhpppphhpppphhpppphh`
4. `ppphhpphhpppppphhhhhhpphhpppphhpphpp`
5. `pphpphhpphhpppppphhhhhhhhhhpppppphhpphhpphpphhhhh`
6. `hhhpphphphpphphphpph`

String	length	upper bound	LP_3	Opt
1	20	11	10.67529996	9
2	24	11	11	9
3	25	8	8	8
4	36	16	14.89908257	14
5	48	25	24.88770748	22
6	20	11	10.76264643	10

5 Discussion

The challenge that we introduce here is to compute better upper bounds for the 2D folding problem using linear programming or otherwise. Our integer and linear programming models provide a promising direction for solving the 2D folding problem to optimality using branch-and-bound. However, because of the large size of the linear program (i.e. number of variables), we likely need tighter linear programming bounds to make these techniques practical.

Another possible application of our integer and linear programming formulations is to find actual foldings that are better than those obtained in approximation algorithms but perhaps not provably optimal. Backofen has used exact methods from constraint logic programming to obtain compact conformations, i.e. solutions, for these folding problems [1]. If we can further constrain our integer programs to the solution space of compact foldings, then we may be able to reduce the time needed to find a solution.

References

[1] Rolf Backofen, "Optimization Techniques for the Protein Structure Prediction Problem", *Ph.D. Thesis, Ludwig-Maximilians-Universität München* (2000).

[2] Robert Carr, William E. Hart, and Alantha Newman, "Discrete Optimization Models for Protein Folding", Technical Report, Sandia National Laboratories, 2003.

[3] V. Chandru, A. DattaSharma, and V. S. A. Kumar, "The algorithmics of folding proteins on lattices", *Discrete Applied Mathematics* (2003) Vol. 127(1):145-161.

[4] K. A. Dill, "Dominant Forces in Protein Folding, *Biochemistry* (1990) Vol. 29:7133-7155.

[5] H. J. Greenberg, W. E. Hart, and G. Lancia, "Opportunities for Combinatorial Optimization in Computational Biology", *INFORMS Journal of Computing, To appear.*

J22. Protein Structure Alignment by Principle Component Analysis

Sung-Hee Park, Soo-Jun Park, Seon-Hee Park[1]

Keywords: alignment, protein structure, principle component analysis

1 Introduction.

One of the most difficult problems in structural bioinformatics is aligning two or more protein structure geometrically. Our structure alignment method involves the principle component analysis (PCA). PCA is well-known as one of the statistical methods for multivariate analysis. In general, PCA is used to reduce dimensionality of multi dimensional data. However, in this research work, PCA is used to align the two protein structure.

Many protein structure alignment methods have been proposed to compare with protein structures. Among them, one is a method using distance matrix for *alpha carbons of proteins structure*[1] while another is a method using distance matrix for *the mass center of the alpha carbons consisting of a segment* which consist of several residue[2]. That is, the former was improved into the latter.
And the other research approach aligns protein structure by vectors of secondary structures.[3] Then it measures a similarity with the vector representation. A recently proposed precise method aligns by incremental combinatorial extension(CE) of the optimal path[4].

Methods mentioned above maximize similarity function to optimize the alignment. They may perform optimization process repeatedly. This is a weak point of above methods.

But, it is the advantage of proposed method that it does not need to optimize similarity function to align structure as other methods.

2 Method.

PCA is a popular statistical method for dimensionality reduction. However, we approach the point of geometrical view of PCA. Geometrically, PCA transforms all atom coordinates on the original coordinate space to those on a new one where the variant with largest variation of covariance matrix is first principle component. Two protein structures aligned by PCA is used to measure similarity.

3 Application: nearest neighbor search.

Here, we used bond line distribution as similarity measure. We named bond line distribution the 3 dimensional edge histogram. 3D edge histogram is a distribution of 10 edge patterns that stands for atom bond.

The flowchart of protein structure comparison by PCA alignment is shown in Fig.1

[1] Bioinformatics Team, Electronic and Telecommunication Research Institute, Daejeon, South Korea. E-mail: {sunghee, psj, shp}@etri.re.kr

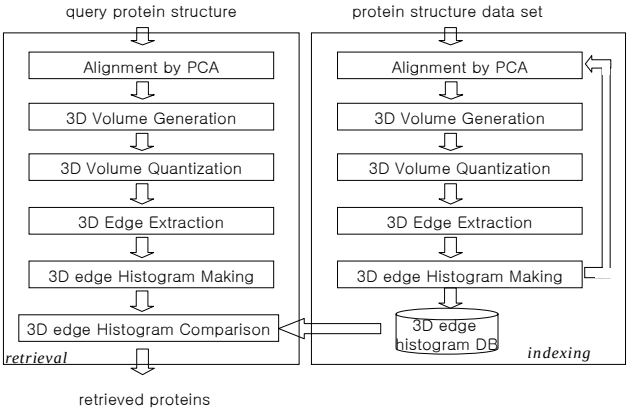


Figure 1: Flowchart of protein structure comparison by PCA alignment.

4 Discussion.

We verified the effectiveness of alignment by PCA through application to nearest neighbor search. We propose a protein structure alignment method of new concept by PCA. The proposed method does not need to optimize similarity function to align structure, while other existing methods may perform optimization process repeatedly. This reduces the time cost for alignment and makes the nearest neighbor search from huge database rapid.

References

[3]Amit P. Singh and Douglas L. Brutlag, "Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representation", *Proc. Intelligent Systems for Molecular Biology*, 1993

[1]Lholm and C.Sander, "Protein Structure Comparison by alignment of distance matrices", *Journal of Molecular Biology*, Vol. 233, pp. 123-138, 1993

[2]Rabian Schwarzer and Itay Lotan, "Approximation of Protein Structure for Fast Similarity Measures", *Proc. 7th Annual International Conference on Research in Computational Molecular Biology(RECOMB2003)*, pp. 267-276, 2003

[4]Ilya N.Shindyalov and Philip E.Bourne, "Protein structure alignment by incremental combinatorial extension(CE) of the optimal path", *Protein Engineering* vol.11 no.9 pp.739-747, 1998

J23. Protein Fold Recognition Using an Optimal Structure-Discriminative Amino Acid Index

J. Ben Rosen[1]. Robert H. Leary[2]. Per Jambeck[3]. Connie X. Wu[4].

Keywords: protein folding, fold recognition, combinatorial extension, supervised learning, amino acid index

1 Introduction.

Identifying the fold class of a protein sequence of unknown structure is a fundamental problem in modern biology. We have applied a supervised learning algorithm FoldID [1] to the classification of protein sequences of length 64 to 300 with low sequence identity from a library of 174 structural classes created by the Combinatorial Extension (CE) structural alignment methodology [2,3]. A class of rules is considered that assigns test sequences to structural classes based on the closest match of an amino acid index profile of the test sequence, which is a numerical vector of dimension N for sequences of length N, to a numerical profile centroid vector for each class. A mathematical optimization procedure is applied to determine an amino acid index of maximal structural discriminatory power by maximizing the ratio of between-class to within-class profile variation. The optimal index is computed as the solution to a generalized eigenvalue problem, and its performance for fold classification of aligned sequences in the training library is compared to that of other published indices using cross-validation techniques.

The algorithm is also tested on raw, unaligned sequences using a combination of local and global alignment techniques to align the numerical profile vectors corresponding to raw sequences with the various class centroid vectors.

2 Computational Results.

For aligned sequences in the CE training library, the optimal index has significantly more structural discriminatory power than all currently known indices in the AAindex database [4], including average surrounding hydrophobicity, which it most closely resembles (Figure 1). It demonstrates more than 70% cross validation classification accuracy on the structurally aligned sequences in the CE fold library over all folds, and nearly 100% accuracy on many individual folds with distinctive conserved structural features such as the EF hand fold family of calcium binding proteins.

For unaligned raw sequences, the corresponding numerical profile vector is first aligned with each class centroid using both local and global dynamic programming techniques. Sufficiently high scoring subsequences that match particular fold class centroid profiles are tentatively assigned to that structural class. The method has proved to be successful in many cases for properly classifying

[1] Dept. of Computer Science and Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0114, USA. E-mail: jbrosen@cs.ucsd.edu

[2] San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0505, USA. E-mail: leary@sdsc.edu

[3] Dept. of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093, USA. E-mail: jambeck@bioeng.ucsd.edu

[4] San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA, 92093-0505, USA. E-mail: cxwu@sdsc.edu

subsequences of the raw sequences on which the CE library is based and extracting the CE structural alignments based only on the sequence, particularly where the number of gaps and/or deletions is relatively small. Tests on raw sequences outside the CE library are encouraging, with successful structural identifications (as determined by subsequent CE analysis) having been made for structural classes with highly conserved features. For example, all seven test instances of EF hand structures not in the original CE library were correctly identified and aligned using FoldID.

3 Figure.

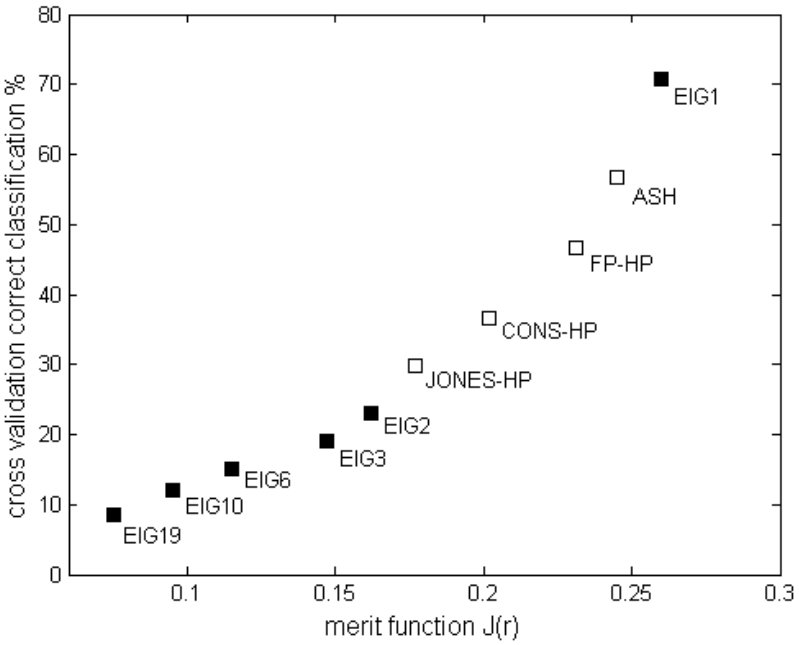


Figure 1. Relative performance of various amino acid indices for CE fold classification. Each EIGi index is computed by FoldID as the i-th eigenvector of the generalized eigenproblem. The remaining indices are various members of the hydrophobicity family.

4 References and bibliography.

[4] Kawashima, S. and Kanehisa, M. 2000. AAindex: amino acid index database. *Nucleic Acids Research* 28:334.

[1] Leary, R.H., Rosen, J.B. and Jambeck, P. 2004. An optimal structure-discriminative amino acid index for protein fold recognition. *Biophysical Journal* 86:411-419.

[2] Shindyalov, I.N. and Bourne, P. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering* 11:739-747.

[3] Shindyalov, I.N. and Bourne, P. 2000. An alternative view of protein fold space. *Proteins* 38:247-260.

J24. Catalytic and Structural Properties of Carp D-Amino Acid Oxidase

Md. Golam Sarower[1], Shigeru Okada[1]. Hiroki Abe[1]

Keywords: D-amino acid oxidase, kinetic parameters, inhibition, three-dimensional structure

1 Introduction.

D-Amino acid oxidase (DAO, EC 1.4.3.3) catalyzes the enantioselective oxidation of a variety of D-amino acids to the corresponding α-keto acids. DAO has been known to distribute widely in nature from microbes to mammals. The physiological functions of DAO, however, remain unclear because its substrates, D-amino acids, are not abundant in vertebrates. Possible functions proposed so far include the catabolic means of bacterial cell wall components, regulator of D-serine, an agonist of *N*-methyl-D-aspartate receptor in mammalian brain, and metabolizing agents of exogenous and/or endogenous free D-amino acids. Recently, we cloned a gene encoding DAO from carp for the first time in animals other than mammals [1]. In view of the enigmatic nature and the industrial potential of DAO, more enzymes from different sources are required to pave the way for further research on biological functions and applications of DAO. Thus, we purified and characterized recombinant carp hepatopancreas DAO (chDAO).

2 Materials and Methods.

The expression vector containing carp DAO cDNA was constructed with pET 11c vector and transformed into a host *E. coli* strain AD494(DE3)pLysS. DAO expressed in *E. coli* was purified with DEAE-Toyopearl, Phenyl-Toyopearl, and HiPrep^{TM} 16/60 Sephacryl^{TM} High Resolution gel filtration columns. Three-dimensional model was constructed by using primary sequence of chDAO with ProModII supplied by Swiss-Model [2].

3 Result and discussion.

Recombinant chDAO was purified to 5.6-fold with a yield of 50%. chDAO had a specific activity of 293 unit/mg protein. It showed high activity against D-alanine with a *K*m of 0.227 (mM) and *K*cat of 190 (S^{-1}) (Table 1). Whereas, *K*cat values for pork kidney (pkDAO) and *Rhodotorula gracilis* (RgDAO) are about 10 (S^{-1}) and 300 (S^{-1}), respectively. The optimum temperature and pH were 35°C and 8.5, respectively. This enzyme exhibited a good thermal and pH stability. It was completely inhibited by Ag^{+} and Hg^{+}. The inhibition by creatinine, *p*-chloromercuribenzoate, and benzoate was competitive with a *K*i of 5.1, 6.0, and 12.5 mM, respectively. Three-dimensional (3D) model of chDAO was analogous to that of pkDAO [3] and RgDAO [4] (Figure 1). Two main topological differences were observed from pkDAO as well as RgDAO. One is the presence of shorter active site loop (9 residues in chDAO *versus* 13 in pkDAO). The active site loop contains an important residue Tyr224 probably involved in a broad range of substrates/products fixation and interaction with substrate α-amino group. In yeast this active site loop is not present, however, Tyr238 (corresponding to Tyr224) found at a similar position plays the same role. Another is the absence of a long C-terminal loop found in RgDAO (6 residues in chDAO and 4 in pkDAO *verses* 21 in RgDAO). This long C-

[1] Department of Aquatic Bioscience, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Bunkyo, Tokyo 113-8657, Japan. E-mail: sarower@gawab.com

terminal loop allows “head to tail” dimer mode that is more stable than “head to head” one found in mammals and carp. The conformational change in large size active site loop controls the overall rate of turnover of the mammalian enzyme, where product release is rate-limiting. Comparison of the 3D structures of chDAO with pkDAO and RgDAO suggests that evolutive pressure has led to the conformational change from microbes to mammals DAOs that share the same chemical process, but use different kinetic efficiency for catalysis.

Table 1: Kinetic parameters of chDAO for some representative D-amino acid oxidase

Substrate	Relative activity (%)	*V* max (U/mg)	*K* m (mM)	*K* cat (s^{-1})	*K* cat/*K* m (s^{-1}/mM)
D-Alanine	100	292	0.227	189.8	836.1
D-Valine	89.9	262	0.263	170.3	647.5
D-Proline	75.1	219	1.13	142.35	126.0
D-Phenylalanine	63.0	183	0.357	118.95	333.2

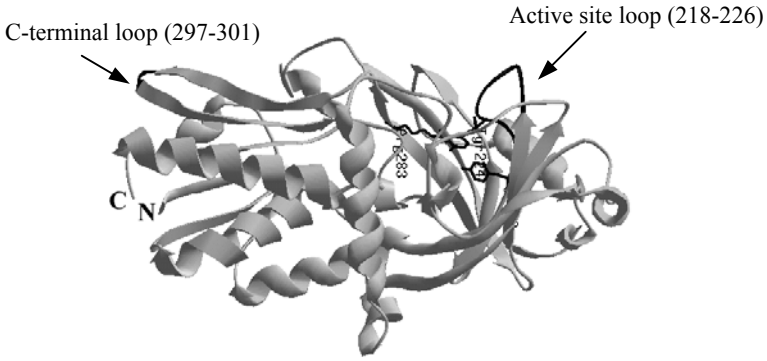


Figure 1. Ribbon representation of chDAO structure. Arg283, Tyr224, and Tyr228 are thought to be key catalytic residues in pkDAO [3]. Active site loop and C-terminal loop are shown in black.

References

[1] Sarower, M.G., Okada, S. and Abe, H. 2003. Molecular characterization of D-amino acid oxidase from common carp *Cyprinus carpio* and its induction with exogenous free D-alanine. *Archives of Biochemistry and Biophysics* 420:121-129.

[2] Peitsch, M.C. 1996. ProMod and Swiss-model: Internet-based tools for automated comparative protein modeling. *Biochemistry Society Transaction* 24:274-279.

[3] Mattevi, A., Vanoni, M.A., Todone, F., Rizzi, M., Teplyakov, A., Coda, A., Bolognesi, M. and Curti, B. 1996. Crystal structure of D-amino acid oxidase: a case of active site mirror-image convergent evolution with flavocytochrome b2. *Proceedings of the National Academy of Sciences USA* 93:7496–7501.

[4] Pollegioni, L, Diederichs, K., Molla1, G., Umhau, S., Welte, W., Ghisla, S. and Pilone, M.S. 2002. Yeast D-amino acid oxidase: structural basis of its catalytic properties *Journal of Molecular Biology* 324:535–546.

J26. Introducing a new protein structure comparison website that reports alternative alignments including structure permutations

Edward S.C. Shih [1], Richie Gan [2]. Ming-Jing Hwang [3]

Keywords: structure comparison, alternative alignment, structure permutation, permuted index, structural bioinformatics, OPAAS

1 Introduction.

The torrential protein 3D structural data are a rich resource for understanding protein evolution and function. To make maximal use of this resource, there is a pressing need for a protein structure comparison (PSC) method that is not only very fast but also versatile enough to detect and characterize permuted and alternative alignments. Here, we introduced a new PSC tool, called OPAAS for Optimal, Permuted, and Alternative Alignment of protein Structure, and constructed a website that is clear, comprehensible and informative for displaying the results of the algorithm. Two main services, database search of one structure and comparison of two structures, are provided on the internet. Our website offers several advantages over other PSC methods. Firstly, a precomputed result of database comparison is ready for users to retrieve and interrogate. Secondly, a hierarchical representation including a list of alternative alignments, diagramed permuted alignments, real-time 3D protein structure superimpositions and detailed structure-based sequence alignments, allows users to acquire most relevant information and knowledge. Thirdly, a permuted index is devised for showing the complexity of an alignment, and with it user can infer the non-topological relationship of two structures. We believe this website will give structural/molecular biologists the most useful and comprehensive information from structure comparison of proteins (URL: http://gln.ibms.sinica.edu.tw/software.php).

2 Methods.

OPAAS,[1] as its predecessor FLASH,[2] is a protein structure comparison method by probability-based matching of secondary structure elements. Besides being very fast, FLASH can also report multiple alignments of two protein structures, if any, and distinguish them as the optimal or alternative alignment according to their statistical similarity scores. OPAAS, the new version, inherits all the merits from FLASH, and in addition introduces non-topological mechanisms to identify permuted alignments, which are quite common but usually neglected in structure comparison studies.

[1] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. E-mail: shihds@gate.sinica.edu.tw
[2] Institute of bioinformatics, National Yang-Ming University, Taipei, Taiwan. E-mail: g39203001@ym.edu.tw
[3] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. E-mail: mjhwang@ibms.sinica.edu.tw

References

[1] Shih, S.C. E. & Hwang,M-J. Alternative alignments from comparison of protein structures. (Submitted).

[2] Shih, S.C. E. and Hwang,M-J. 2003. Protein structure comparison by probability-based matching of secondary structure elements. *Bioinformatics*. **19**, 735-741.

J27. Comprehensive Protein Database Representation

Amandeep S. Sidhu[1], Tharam S. Dillon[1], Baldev S. Sidhu[2] Henry Setiawan[1]

Keywords: BIODB, BIOMAP, Bioinformatics, Protein, Sequence, Structure, Databases.

1 Introduction

Protein Structure data is stored in geographically dispersed databases. Most of these databases have flat files as exchange format and each of them have a different representation. We propose a database representation BIOlogical DataBase (BIODB) which provides a unified structured representation for protein structure records covering all the aspects of protein structure description. It contains information about residue sequences, atomic structure, bonding and folding for every protein structure, collected from existing protein databases – PDB [1], SWISS-PROT [2], PROSITE [3] and PIR [4]. It is not dependent on any particular data format.

We created a representation of BIODB that will be used to construct an efficient similarity detection method for protein structures, in our project BIOlogical MAPping (BIOMAP). Each of the documents generated from BIODB database record corresponds to a protein structure description. These documents possess tree like structure. Associations can be built up among trees rather than single atomic values. The tree like structure of document is very helpful in comparing the Protein Structures among themselves and with other biological structures from the type of information that is being searched. It allows definition of the sub-trees which contain that information. It provides a faster way of searching the candidates and also organizes the vast amount of data available about the Proteins. Using the containment relationship of the elements in the documents generated we also found large similar patterns with high levels of confidence. This representation also provides a way of creating a map between the BIODB database and the data structures of existing Protein Databases. This greatly helps in creating interoperability and sharing information from different databases. Thus the information can be represented in existing Protein Database formats if needed.

2 The BIODB Database

The BIODB contains information about Protein Structures in a way that represents the relationships between various Protein Structure elements. It takes into greater consideration formation of the ultimate protein conformation and the relationships that exist in the data, rather then just storing the data.

Table	Description
DatabaseEntry	Protein Database Entry Record Details from the Protein Database from where entry is fetched.
Entry	Protein Entry Description Details.
Compounds	Details of Molecular Compounds present in Protein.
Source	Organism Source from where Protein is taken and its description.
Revisions	Revisions done to Protein.

[1] Faculty of IT, University of Technology Sydney, Australia
E-Mail : {asidhu, tharam, henryws}@it.uts.edu.au
[2] Member, Board of Studies for Biology, Punjab State Education Department, India
E-Mail : bssidhu@biomap.org

Citations	Citation of Protein in Publications.
Conflicts	Conflicts between Atom records and Database Entry of Protein.
DBRef	Databases that parts of Protein Sequence refer to.
ATOMSequence	Details of each Atom Structure in a Chain of Residue Sequences of Protein.
ModifiedResidues	Residues that have been modified during Protein Evolution.
Helices	Helices present in the Protein Structure.
Turns	Turns present in the Protein Structure.
Sheets	Sheets present in the Protein Structure.
NonStandardResidues	Description of Non Standard Residues in Protein Structure.
NonStandardAtomGroups	Details about Atom Structure within a Chain of Non Standard Residue Groups in Protein Sequence.
AtomListEnd	Identification of end of Atom list for all Standard Residues in Protein.
DisulphideBonds	Disulphide Bonds in Protein Structure.
HydrogenBonds	Hydrogen Bonds in Protein Structure.
ResidueLinks	Residue Links in Protein Structure.
SaltBridges	Salt Bridges in Protein Structure.
CISPeptides	Peptides found in CIS Conformation of Protein.
Sites	Important Binding Sites and their description in Protein Structure.
UnitCell	Crystallographic information Parameters of Protein Structure.

Table 1: Tables of BIODB Database

We intend to construct a Complete Protein Data Representation with interfaces, available on the Internet. This representation will be extended to other types of biological data (RNA, DNA, etc.) and a complete map to understand the relationship between these biological elements will be created to understand the cell physiology.

3 Bibliography

1. M.Berman, H., et al., The Protein Data Bank. Nucleic Acids Research, 2000. 28(1): p. 235-242.
2. Bairoch, A. and R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Research, 1997. 25(1): p. 31-36.
3. Bairoch, A., P. Bucher, and K. Hofmann, The PROSITE database, its status in 1997. Nucleic Acids Research, 1997. 25(1): p. 217-221.
4. George, D.G., et al., The Protein Information Resource (PIR) and the PIR-International Protein Sequence Database. Nucleic Acids Research, 1997. 25(1): p. 24-27.

J28. Protein-protein recognition: relationship between domain and interface cores in immunoglobulins

Vladimir Potapov[1], Vladimir Sobolev[2], Marvin Edelman[3], Alexander Kister[4], Israel Gelfand[5]

Keywords: molecular recognition, protein flexibility, protein binding site

1 Introduction.

Protein-protein recognition plays a major role in cell function. A question of key importance is how the two proteins interact (i.e., which residues form the interface and what interactions primarily stabilize the complex). There are several theoretical approaches to predict if a residue (or a set of residues) is included in a protein-protein interface. These approaches are founded on empirical rules derived from an analysis of protein-protein, or domain-domain, interfaces. Such predictions still are restricted in accuracy. As a result, there are currently intensive efforts to extract additional levels of information relating to interacting surfaces from the structural database. A general aim in these studies is to clearly recognize the site of interaction. However, in most cases, predicting a unique interface region has remained elusive, probably because the interface as a whole has approximately the same character as the surface as a whole. Thus, a different approach may be needed to distinguish the interface residues of a protein dimer. In our research, we are developing a prediction approach that is based on analysis of the relationship between protein sequence/structure and interface properties. In the present study we use a statistical approach to derive the protein-protein interface core and further analyze its structural conservativity. A set of 47 immunoglobulin structures was evaluated to determine the residue positions that play a primary role in protein - protein dimer association of the heavy and light chains (VL-VH and CL-CH1 associations).

2 Results and Discussion.

Recently, we discovered in beta sandwich-like proteins an invariant supersecondary substructure called 'interlock' [1]. Two interlocked pairs of beta strands located on separate sheets (B, E on one and C, F on the other) are situated at the center of each domain. Eight conserved, hydrophobic, residue positions (B6, B8, C4, C6, E8, E10, F6 and F8) within the interlocked strand pairs form the common geometric core of all beta sandwich-like proteins, including immunoglobulins. The C-alpha atoms of residues at these eight conserved positions were used as reference points for superimposition of structures.

Database. Two hundred and eighty one structures of Fab fragments from the PDB were extracted using the SCOP database. Structures of resolution 2.3 Å or better were retained and Fab fragments

[1] Plant Sci. Dept., Weizmann Inst. of Sci., Israel. E-mail: vladimir.potapov@weizmann.ac.il
[2] Plant Sci. Dept., Weizmann Inst. of Sci., Israel. E-mail: vladimir.sobolev@weizmann.ac.il
[3] Plant Sci. Dept., Weizmann Inst. of Sci., Israel. E-mail: marvin.edelman@weizmann.ac.il
[4] Dept. of Health Informatics, Univ. of Medicine and Dentistry, New Jersey, USA and Dept. of Math., Rutgers University, USA. E-mail: kisterae@umdnj.edu
[5] Dept. of Math., Rutgers University, USA. E-mail: igelfand@math.rutgers.edu

selected such that none had more than 80% sequence identity with any other. The final database consists of 47 structures.

Interface core definition. The core is defined in four steps: 1. All inter-domain contacts are determined for each structure using CSU software [2]; 2. The average contribution to the summated (virtual) interface surface area of all structures was calculated for each interface position; 3. The minimal set of contacts that form 80% of the average contact surface area was selected; 4. The positions constituting the interface core are then chosen based on: (i) frequency at the interface; (ii) conservation of residue character (hydrophobic, hydrophilic or neutral) for both residues of a contact pair; (iii) physical-chemical compatibility of a contact at the atomic level; (iv) spatial conservation (RMSD of C-alpha atoms ≤2.0 after superimposition of all structures). The interface between beta sheets within a domain (VL, VH, CL, CH1), and domain core, were derived in the same manner as above.

VL-VH interface core. Analysis of our database showed that, in summation, residues at 51 VL and 46 VH positions are involved in the inter-domain virtual interface. For any given Ig structure, approximately 24 residues in each domain make up the interface, yielding an average contact surface of 1074 Å2, while 11 VL and 7 VH positions constitute the interface core. The dominant contacts in the core are hydrophobic interactions (formed by hydrophobic residues or by aliphatic part of hydrophilic ones). Similar sort of data were obtained for the CL-CH1 interface and each of the four domain cores.

Relations between interface and domain cores. We found that: (i) almost all residues of the interface core occupy positions on beta strands rather than loops; (ii) seven of eight interlock positions are bordered by interface core residues in VL-VH and CL-CH1; (iii) 66% of interface core residues are immediately adjacent to residues forming the domain core. The probability for such grouping ranges from ~10^{-3} to ~10^{-4}.

3 Conclusions.

We demonstrated the existence of a highly conserved interface core between domains for Ig-like proteins that is considerably smaller than the interface itself. We discovered that the interface surface is geometrically wedded to the domain core, suggesting that the latter controls interface rigidity in immunoglobulin (and probably other) domains.

References

[1] Kister A.E., Finkelstein A.V., Gelfand I.M. 2002. Common features in structures and sequences of sandwich-like proteins. *Proc. Natl. Acad. Sci. USA* 99:14137-14141.
[2] Sobolev V., Sorokine A., Prilusky J., Abola E.E., Edelman M. 1999. Automated analysis of interatomic contacts in proteins. *Bioinformatics* 15:327-332.
[3] Sobolev V., Wade R.C., Vriend G., Edelman M. 1996. Molecular docking using surface complementarity. *Proteins* 25:120-129.
[4] Chothia C., Gelfand I., Kister A. 1998. Structural determinants in the sequences of immunoglobulin variable domain. *J. Mol. Biol.* 278:457-479.

J29. Protein Structural Repeats Revealed in Alternative Alignments of Self Structure Comparison

Ching-Shu Suen, Edward S.C. Shih and Ming-Jing Hwang[1]

Keywords: protein repeats, structural comparison, alternative alignment, solenoid

1 Introduction.

It has been reported, based on sequence analysis, that about 14% of all proteins contain repeating amino acid segments, which occur more in eukaryotic proteins than in prokaryotic proteins owing to unique functions of eukaryotes [1]. However, a method for comprehensively detecting internal repeats in protein 3D structures is lacking until recently. One investigation used wavelet transformations to identify motif repeats in both protein sequences and structures and found many well-known repeated proteins [2]. Another study analyzed a smoothed score matrix in structure alignment and applied Fourier transformations to detect protein repeats [3].

Here, we applied a new protein structure comparison (PSC) method, called OPAAS for Optimal, Permuted, Alternative Alignments of protein Structures (Shih & Hwang, submitted), to detect internal repeats in protein structures. OPAAS differs from most of existing PSC methods by the ability to identify distinctive alternative alignments in a very efficient way. Using alternative alignments, protein structures with internal repeats can be readily identified by structure comparison to itself.

2 Methods.

The SCOP database release 1.55 was analyzed in this study. A total of 4940 domains (class a-f) sharing less than 90% sequence identity and with ≥ 3 SSEs (Secondary Structure Element) were selected and structure-compared to itself.

After generating alternative alignments in each pair of self-aligned protein domains, as overlap ratio of the number of aligned residues shared by two distinct alignments of the same protein structure was calculated. When the overlap ratio is greater than or equal to 0.5, the alternative alignment is indicative of a structure containing a sub-structural internal repeat. The alignment was then manually inspected, the internal repeat extracted and characterized.

3 Results.

Applying to SCOP, we identified 401 repeat-containing domains in 96 SCOP folds. We catalogued these repeats into four types, spiral (solenoid), circular, irregular, and duplicate, based on the ways how the repeat units are arranged spatially. We identified all the prolific repeat proteins that have been observed previously; in addition, we discovered many protein repeats not reported in the literature.

4 References.

[1] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. E-mail: mjhwang@ibms.sinica.edu.tw

[1] Marcotte, E. M., Pellegrini, M., Yeates, T. O. & Eisenberg, D. 1998. A census of protein repeats. *J. Mol. Biol.* 293:151-160.

[2] Murray, K. B., Gorse, D. & Thornton J. M. 2002. Wavelet transforms for the characterization and detection of repeating motifs. *J. Mol. Biol.* 316:341-363.

[3] Taylor, W. R., Heringa, J., Baud, F. & Flores, T. P. 2002. A fourier analysis of symmetry in protein structure. *Protein Engineer.* 15:79-89.

J30. Assignment of structural domains in proteins: why is it so difficult?

Stella Veretnik[1], Ilya N. Shindyalov[1]. Phillip E. Bourne, [1,2]

Keywords: 3D protein structure, detection of 3D protein domains, automatic domain assignment methods, consensus approach, curated domain resources.

1 Introduction.

Structural domains are often considered to be basic units of protein structure. Assignment of structural domains from atomic coordinates is crucial for understanding protein evolution and function. Currently there is no good agreement among different assignment methods for what constitute the basic structural unit, underscoring the complexity of structural domain assignment. This work discusses tendencies of individual methods and highlights the problematic areas in assignment of structural domains by experts as well as by fully automated methods.

2 Methods.

Domain assignments were analyzed for three automatic methods (DALI[1], DomainParser[2], PDP[3]) and three expert methods (AUTHORS[4], CATH[5], SCOP[6]), using a 467-chains dataset assigned by all 6 methods. The following features were investigated: agreement on the number of assigned domains, agreement on domain boundaries, distribution of domain sizes and tendency toward assignment of discontinuous domains. Consensuses among automatic, expert and all methods were defined and used during comparison to tease out the behaviors specific to individual assignment methods or groups of methods.

3 Results and Discussion.

We observe that unambiguous domain assignments (when all methods agree on domain assignment) are confined predominantly to one-domain chains. Agreements among all methods in multi-domain chains are infrequent; in all cases the domains are compact and clearly spatially separated. For the majority of multi-domain proteins, there is no agreement on domain assignment among all methods. From the consensus analysis we observe that the majority of the difficulties of fully automated methods stem from overwhelming reliance on the structural cues (compactness/contact density) during domain assignments and the lack of functional/evolutionary information. Thus the cases in which domains are positioned close together are difficult or impossible for automatic methods to resolve. On the other hand, the differences in expert methods arise from different philosophical approaches underlying the specific methods. Authors of the structures (AUTHORS method) tend to define domains based on functionality, which may produce small and structurally not clearly defined domains. The creators of SCOP, on the other hand, often look for the largest common structure (fold) as a domain, which often consists of several distinctive structural units. The CATH method appears to strike a balance between sometimes contradictory structural, functional and evolutionary information. The inconsistencies in expert assignments are well reflected in the propensities of different fully automated methods, as those are trained and validated using a specific expert method, thus reflecting its philosophical biases. Detailed analysis

[1] San Diego Supercomputer Center, University of California, San Diego, 9500 Gilman Dr. La Jolla 92093-0537

[2] Department of Pharmacology, University of California, San Diego, 9500 Gilman Dr. La Jolla 92093

of structures which do not have consensus between the assignment methods regarding the number of assigned domains indicates the following problematic areas: (1) assignment of small domains, (2) discontinuous domains and unassigned regions in the structure, (3) splitting of the secondary structure elements between domains (if required), (4) convoluted domain interfaces and complicated architectures. Comprehensive domain re-definition, which takes into account the above issues is overdue and will be a great step toward improvement of domain definitions in multi-domain proteins, which represent (by an estimation [7]) 66-75% of the sequence database. Also, the intensive growth of 3D protein data demands fully automated approaches to be used to maintain currency and uniformity of domain information relative to the PDB.

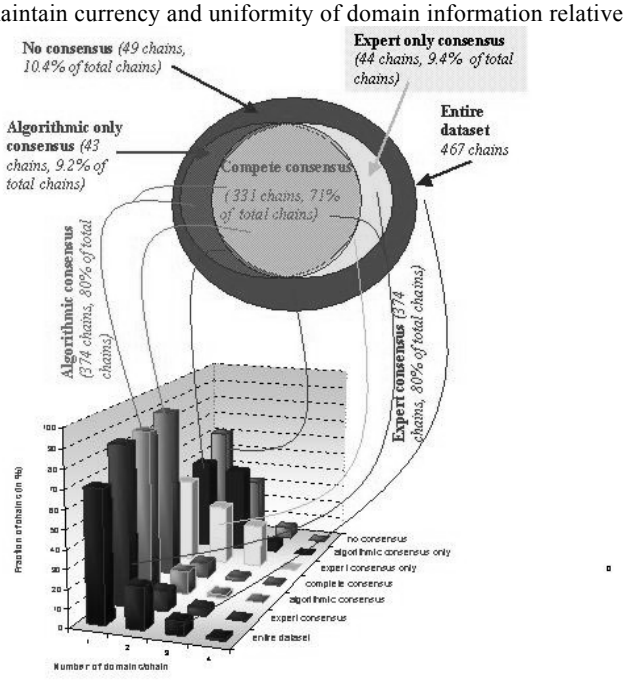


Figure1. Distribution of single- and multi-domain structures within different consensuses of domain assignment methods.

References

[1] Holm L., S. C. 1996 Mapping the protein universe. *Science* 273, 595-602.

[2] Guo, J-T. Xu, D. Kim, D. Xu, Y. 2003 Improving the Performance of DomainParser for Structural Domain Partition Using Neural Network, *Nucleic Acids Res*.31(3), 944-952.

[3]. Alexandrov, N. & Shindyalov, I. 2003 .PDP: protein domain parser. *Bioinformatics* 19, 429-430.

[4] Islam, S. A., Luo, J. & Sternberg, M. J. 1995 Identification and analysis of domains in proteins. *Protein Eng* 8, 513-25.

[5] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. 1997 CATH—a hierarchic classification of protein domain structures. *Structure* 5, 1093-108.

[6] Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247, 536-40.

[7] Chothia C, Gough J, Vogel C, Teichmann SA. 2003 Evolution of the protein reprertoire. Science 300, 1701-03.

J31. Training Hidden-Markov Models on Sequences of Local Structural Alphabets for Protein Fold Assignment

Shiou-Ling Wang[1,2], Chung-Ming Chen[1] and Ming-Jing Hwang[2]

Keywords: structural alphabet, protein fold assignment, HMM, structural alignment, fold signature

Introduction

Recurring local structures of proteins, which may be represented by a set of structural alphabets or libraries of structural motifs, are increasingly used to study the relationship between sequence and structure and to predict protein three-dimensional (3D) structures. We have recently derived a set of protein local structural alphabets (LSA) from clustering > 130,000 fragments, each of five residues in size, excised from ~1,000 non-redundant and diverse known protein structures [1]. In the present study, we employed the derived LSA for fold assignment, i.e. assigning the SCOP fold for a given protein structure, and evaluated the size of LSA required for optimal performance of the assignment.

Methods

With LSA, we can approximate a protein 3D structure and converted it into a 1D character string, or sequence, of LSA. To evaluate to what extent the LSA sequence representation can capture the essence of a protein 3D fold; we tested the fold assignment performance by training Hidden-Markov models (HMM) on 43 populated SCOP fold families, each having at least 20 member structures. For each fold family selected, we identified a reference structure, and aligned all the other member structures onto it using a fast structure comparison algorithm FLASH [2]. The HMM was trained on this multiple structural alignment, which was represented in the form of a multiple LSA sequence alignment. A protein structure can then be assigned to one of the 43 SCOP folds, i.e. the HMM having the maximal probability score. For evaluation of the fold assignment performance, we conducted a 5-fold cross-validation on a dataset with less than 40% pair wise sequence identity chosen according to the ASTRAL Compendium database.

Results

The HMM was run on different sets of LSA, in size of 5, 10, 15, 20, 25, 33 and 40 alphabets, respectively. The 5-fold cross-validation results showed that a performance plateau was reached at 20 alphabets, beyond which improvement was negligible. Furthermore, the use of a substitute matrix giving different substitution scores for different alphabets elevated the assignment accuracy by ~7% for all the different alphabet sets, and yielded an accuracy of 82% for the set of 20 alphabets. A comparison with the results of Cootes et al. [3], which used a very different approach to capture fold signatures, showed that our method performed better in three of the four major protein classes. The less-optimal results for $\alpha + \beta$ structures can be attributed in large part to a gross misalignment of long helices in the form of 1D LSA sequence for, particularly, the

[1] Institute of Biomedical Engineering, Taiwan University, Taipei, Taiwan.

[2] Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. E-mail:mjhwang@ibms.sinica.edu.tw

Zincin-like fold. Our results suggested that protein fold signatures can be largely captured by local structures even if they are represented in the form of 1D alphabet sequences.

References

[1] Soong, T.T. 2002. Clustering and characterizing local protein structures by an Expectation-Maximization (EM)-assisted approach. *Master Thesis*, National Taiwan University.

[2] Shih, E. S. and M. J. Hwang. 2003. Protein structure comparison by probability-based matching of secondary structure elements. *Bioinformatics.* 19:735-741.

[3] Cootes, A. P., S. H. Muggleton, and M. J. Sternberg. 2003. The automatic discovery of structural principles describing protein fold space. *J.Mol.Biol.* 330:839-850.

J32. A Probability-Based Similarity Measure for Saupe Alignment Tensors with Applications to Residual Dipolar Couplings in NMR Structural Biology

Anthony K. Yan [1]
Christopher J. Langmead [2]
Bruce Randall Donald [3 4 5 6]

Keywords: SO(3), rotations, subgroup method, orthogonal image, alignment tensor, residual dipolar couplings, RDC, Saupe matrix, NMR structural biology, resonance assignment

1 Introduction

High-throughput NMR structural biology and NMR structural genomics pose a fascinating set of geometric challenges. A key bottleneck in NMR structural biology is the resonance assignment problem. We seek to accelerate protein NMR resonance assignment and structure determination by exploiting *a priori* structural information. In particular, a method known as Nuclear Vector Replacement (NVR) has been proposed as a method for solving the assignment problem given a priori structural information [4, 5]. Among several different kinds of input data, NVR uses a particular type of NMR data known as *residual dipolar couplings (RDCs)*. The basic physics of residual dipolar couplings tells us that the data should be explainable by a structural model and set of parameters contained within the *Saupe alignment tensor*.

In the NVR algorithm, one estimates the Saupe alignment tensors and then proceeds to refine those estimates. We would like to quantify the accuracy of such estimates, where we compare the estimated Saupe matrix to the correct Saupe matrix. In this work, we propose a way to quantify this comparison. Given a correct Saupe matrix and an estimated Saupe matrix, we compute an upper bound on the probability that a randomly rotated Saupe tensor would have an error smaller than the estimated Saupe matrix. This has the advantage of being a quantified upper bound which also has a clear interpretation in terms of geometry and probability.

While the specific application of our rotation probability results is given to NVR, our novel methods can be used for any RDC-based algorithm to bound the accuracy of the estimated alignment tensors. For example, there is research on using RDCs in resonance assignment, as well as structure determination and refinement [1, 2, 3, 4, 5, 6, 7, 8].

Furthermore, our method is a general framework for characterizing the accuracy of rotations. As a result, it may be applicable to many of areas, including X-ray crystallography or molecular docking to quantitate the accuracy of calculated rotations of proteins, protein domains, nucleic acids, or small molecules.

[1]Dartmouth Computer Science Department, Hanover, NH 03755, USA. E-mail: yan@cs.dartmouth.edu

[2]Department of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh PA 15213, USA. E-mail: cjl@cs.cmu.edu

[3]Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. Phone: 603-646-3173. Fax: 603-646-1672. Email: brd@cs.dartmouth.edu

[4]Dartmouth Chemistry Department, Hanover, NH 03755, USA.

[5]Dartmouth Department of Biological Sciences, Hanover, NH 03755, USA.

[6]This work is supported by grants to B.R.D. from the National Institutes of Health (R01 GM-65982), and the National Science Foundation (IIS-9906790, EIA-0102710, EIA-0102712, EIA-9818299, and EIA-9802068, EIA-0305444).

References

[1] Al-Hashimi, H.M. and Gorin, A. and Majumdar, A. and Gosser, Y. and Patel, D.J. Towards Structural Genomics of RNA: Rapid NMR Resonance Assignment and Simultaneous RNA Tertiary Structure Determination Using Residual Dipolar Couplings. *J. Mol. Biol. 318* (2002), 637–649.

[2] Al-Hashimi, H.M. and Patel, D.J. Residual dipolar couplings: Synergy between NMR and structural genomics. *J. Biomol. NMR 22*, 1 (2002), 18.

[3] Hus, J.C. and Propmers, J. and Brüschweiler, R. Assignment strategy for proteins of known structure. *J. Mag. Res 157* (2002), 119–125.

[4] C. J. Langmead, A.K. Yan, L. Wang, R. Lilien and B. R. Donald. A Polynomial Time Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments. *The Seventh Annual International Conference on Computational Molecular Biology (RECOMB)* (2003), Berlin, Germany April 10-13. p. 176–187

[5] C. J. Langmead, A.K. Yan, L. Wang, R. Lilien and B. R. Donald. A Polynomial Time Nuclear Vector Replacement Algorithm for Automated NMR Resonance Assignments. *Journal of Computational Biology* (2003).

[6] Losonczi, J.A. and Andrec, M. and Fischer, W.F. and Prestegard J.H. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson 138*, 2 (1999), 334–42.

[7] Tjandra, N. and Bax, A. Direct Measurement of Distances and Angles in Biomolecules by NMR in a Dilute Liquid Crystalline Medium. *Science 278* (1997), 1111–1114.

[8] Wedemeyer, W. J. and Rohl, C. A. and Scheraga, H. A. Exact solutions for chemical bond orientations from residual dipolar couplings. *J. Biom. NMR 22* (2002), 137–151.

J33. When and where do protein folds come from? an evolutionary view

Song Yang[1], Phil Bourne[2]

Keywords: SCOP, fold evolution, disulfide bond, phylogenetic tree

Compared with the fast-growing numbers of protein sequences, the number of possible protein structures is quite limited. It has been proposed that there may be only 400-4000 protein folds existing in all organisms. Thus structure represents a more conserved and alternative measure by which evolution can be studied. Thus as more and more complete genomes of organisms in every kingdom are sequenced, it is possible to compare protein folds across various genomes in the tree of life to gain an evolutionary view.

We have used existing databases (SUPERFAMILY[1] and PEDANT[2]), which contain fold assignments for complete genomes (based on the protein fold classification defined by SCOP[3] and using the homology search methods PSI-BLAST or HMMs), to obtain protein fold counts in 17 Archaea, 123 Bacteria and 16 Eukaryota[4]. The accompanying graph shows a Venn diagram of fold distribution (defined as the second level in SCOP) in the three kingdoms. Among the total 753 folds, only one fold, d.199, is unique to Archaea. Since the only representative of d.199 in PDB is a transcription factor from bacteriophage T4, it is likely that Archaea obtained this fold from a virus. Therefore, Archaea have hardly any unique folds, if any. In contrast, Bacteria and Eukaryota appear to have invented a substantial number of new folds since the divergence of the three kingdoms. About 40% of the folds (24 of 61) unique to Eukaryota contain disulfide bonds, whereas the percentage of folds containing disulfide bonds found only in Bacteria is 31% (5 of 16), those folds found in both Eukaryota and Bacteria is 22% (36 of 165), and those common to all kingdoms is only 6% (30 of 491). This suggests that many protein domains stabilized by disulfide bonds emerged after the atmosphere became more oxygen-rich, under which conditions Eukaryota generated more disulfide bond-containing folds than the other groups.

On another front, we were able to build a phylogenetic tree based on the existence and abundance of folds in each genome. The tree is similar and comparable to phylogenetic trees built with gene sequences or other features. Detailed analyses of certain folds were also performed.

1 Department of chemistry and biochemistry, University of California, San Deigo
2 San Diego Supercomputer Center, Department of Pharmacology, UCSD, Burnham Institute

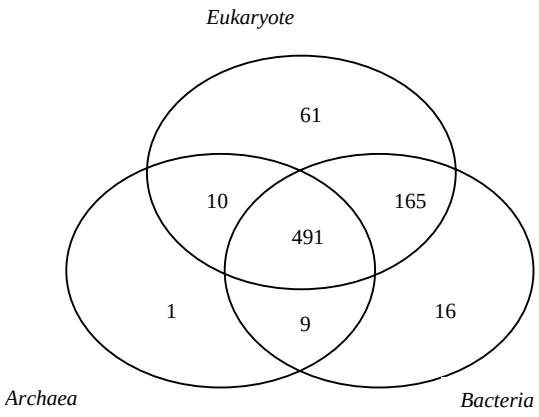


Figure 1: Protein fold distribution among three kingdoms

References

[1] Gough, J., Karplus, K., Hughey, R. and Chothia, C. 2003. Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure. *J. Mol. Biol*., 313: 903-919

[2] Frishman, D etc. 2003. The PEDANT genome database. *Nucleic Acids Research* 31: 207-211.

[3] Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. 2002. SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30: 264-267.

[4] Gustavo Caetano-Anollés and Derek Caetano-Anollés. 2003. An Evolutionarily Structured Universe of Protein Architecture. *Genome Res.* 13: 1563-1571.

J35. Hydrophobic Moment of Multi-Domain Proteins: Magnitude and Spatial Orientational Bias

R. Zhou[1], A. Royyuru[1], P. Athma[1] and B. D. Silverman[1]

Keywords: hydrophobic moment, spatial orientation, hydrophobic ratio, multi-domain

1 Introduction.

The understanding of multi-domain protein folding and structures is of great interest recently. Protein domains are generally hypothesized to demarcate compact units that fold independently. This suggests that the distribution of residue hydrophobicity for individual domains as well as aggregate sets of domains should exhibit special properties. If domains do fold independently, one expects individual domains to exhibit a core composed predominantly of hydrophobic residues with an exterior composed predominantly of hydrophilic residues. Furthermore, as a consequence of domain coalescence within the aqueous environment one expects the amphiphilicity of spatially adjacent domains to reveal a bias of hydrophobic residues in the region of domain contact. In this study, we examine the hydrophobic spatial profiling[2-3] of all single-chain multi-domain proteins in PDB to see if such expected bias is reflected in the hydrophobic moment of the individual domains.

2 Hydrophobic Moments.

Hydrophobicity is widely used to examine the protein structures. Each amino acid exhibits a different degree of hydrophobicity h_i based upon its solubility in water[4]. As a good approximation, the protein shapes can be represented by an ellipsoidal,[2-3]

$$x^2 + g_2' y^2 + g_3' z^2 = d^2$$

where g_2', g_3' are normalized moments of geometry, and d is a generalized ellipsoidal radius. Then one can define the 0th-, 1st- and 2nd-order hydrophobic moments:

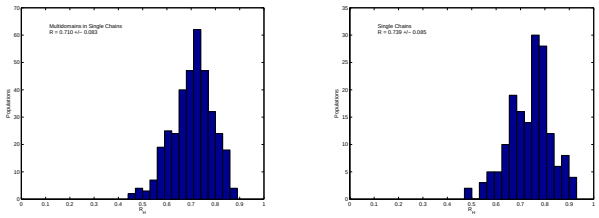
$$H_0(d) = \sum_{r<d} h_i' \qquad \vec{H}_1 = \frac{1}{n}\sum_i h_i(\vec{r}_i - \vec{r}_c) \qquad H_2(d) = \sum_{r<d} h_i'(x_i^2 + g_2' y_i^2 + g_3' z_i^2)$$

The prime designates the value of hydrophobicity of each residue after normalization (mean zero with standard deviation of 1) to enable comparison between different proteins[2-3]. When the value of d is just sufficiently large enough to collect all of the residues, the net hydrophobicity of the protein vanishes, i.e., H_0(d) vanishes at this maximum value of d_0. And the location at which the 2nd-order moment vanishes is defined as d_2. The hydrophobic-ratio is then defined as, $R_H = d_2 / d_0$ (consult previous papers[2-3] for more details). Surprisingly, R_H is found to be relatively constant, and a value of 0.71 ± 0.08 was found for all native globular soluble protein domains in PDB[2-3].

3 Results and Discussion.

A total of 162 non-redundant multi-domain proteins (single chains) and 358 corresponding domains have been selected from PDB following a similar procedure as described previously[3]. Surprisingly, these multi-domain complexes profile like the individual domains with a well-defined hydrophobic ratio R_H. Fig. 1 shows the distribution of R_H for both the individual domains and the multi-domain complexes, with R_H= 0.71 ± 0.08 for individual domains, and 0.74 ± 0.08 for complexes.

[1] Computational Biology Center, BM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, E-mail: ruhongz@us.ibm.com

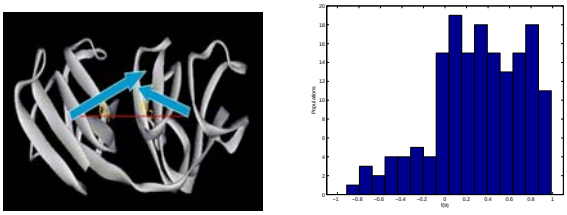


(a) (b)

Figure 1: Hydrophobic ratio distribution: (a) individual domains; (b) the multi-domain complexes.

These otherwise surprising results could occur only if during folding and the exclusion of water from the region between domains, a predominantly hydrophobic interface was formed. This, in turn, would appear as an imbalance in the distribution of residue hydrophobicity across the individual domains which could be characterized by the magnitude and direction of the 1st-order hydrophobic moment. Fig. 2a shows one example of the 1st-order moments for protein 1dhy. The vector orientations show that the hydrophobic moments point towards each other, reflecting the prevalence of hydrophobic residues near the domain interface. Thus, an orientation score f(θ) is defined to quantify such orientation bias. For a two-domain protein, it is simply,

$$\cos(\theta_1) = (\vec{H}_1 \cdot \vec{r}_{12})/\left|\vec{H}_1\right|\left|\vec{r}_{12}\right|; \quad \cos(\theta_2) = (\vec{H}_2 \cdot \vec{r}_{21})/\left|\vec{H}_2\right|\left|\vec{r}_{21}\right|; \quad f(\theta) = \frac{1}{2}(\cos(\theta_1) + \cos(\theta_2))$$



(a) (b)

Figure 2: (a) The 1st-order hydrophobic moments of protein 1dhy; (b) The distribution of the function f(θ).

For more than two-domains, it is just the average over all pairs of domains. Fig. 2b shows the distribution of the scoring function f(θ). If no preference in the direction of the 1st-order moments, a random distribution of f(θ) would occur with a mean value of ~0. The results, however, show that 79.4% of complexes yield a value that is greater than zero (with a mean of 0.32). Thus, individual domains distribute a greater number of hydrophobic residues in the vicinity of the domain interface, suggesting that during the multi-domain protein folding each individual domain develops its own hydrophobic core while generating the hydrophobic imbalance at the domain interface, required for the final assembled complex. There is a strong tendency for individual domains to re-orientate themselves in order to bury more hydrophobic residues during the last stage of multi-domain protein folding. This would explain the origin of the well-defined hydrophobic profiles and hydrophobic ratios obtained for the multi-domain single-chain complexes.

References

[1] Jaenicke R. 1998. Stability and folding of domain proteins, Prog. Biophys. & Mol. Biol. 1999, 71: 155-241
[2] Silverman BD. 2001. Hydrophobic moments of protein structures: Spatially profiling the distribution. Proc. Natl. Acad. Sci. USA 98: 4996-5001
[3] Zhou R, Silverman BD, Royyuru A, Athma P. 2003. Spatial profiling of protein hydrophobicity: Native vs. decoy structures. Proteins: Struc. Funct. & Genetics 52: 561-572
[4] Eisenberg D, Weiss RM, Terwilliger TC. 1982. The helical hydrophoic moment: a measure of the amphiphilicity of a helix. Nature 299: 371-374

J36. Analytical Model for the Prediction of NMR Methyl-side Chain Order Parameters in Proteins

Dengming Ming[1,2] and Rafael Brüschweiler[2]

Keywords: Protein backbone dynamics, NMR relaxation, S^2 order parameters, methyl group dynamics, side-chain dynamics, prediction of protein mobility

1 Introduction.

An analytical model is presented for the prediction of NMR order parameters of methyl groups in proteins. The model, which is an extension of the local contact model for backbone order parameter prediction, uses a static 3D protein structure as input. It expresses the methyl-group S2 order parameters as a function of local contacts of the methyl carbon with respect to the neighboring atoms in combination with a term that takes into account the number of consecutive mobile dihedral angles between the methyl group and the protein backbone. For six out of seven proteins the prediction results are good when compared with experimentally determined methyl-group S2 values with an average correlation coefficient r=0.65±0.14. For cytochrome c2, which is an unusually rigid protein, no correlation between prediction and experiment is found. Despite its simplicity, it represents a first comprehensive relationship between protein NMR side-chain dynamics and protein structure.

2 Figures and tables.

The analytical model reads: $S_i^2 = \tanh\left(a \cdot C_i / n^b\right) - c$, where a, b, c are empirical parameters, C_i is the local contact experienced by methyl carbon i, and n_i is the number of mobile covalent bonds between the methyl carbon i and the backbone.

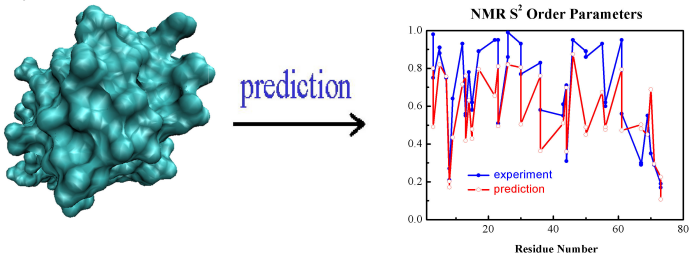


Figure 1: Prediction of ubiquitin methyl group order parameters using analytical model

References

[1] Zhang, F. and Brüschweiler, R. 2002. *J. Am. Chem. Soc.*, **124**:12654-12655.

[1] Computer and Computational Sciences Division, Mail Stop B256, Los Alamos National Lab, Los Alamos, New Mexico, 87545 USA. E-mail: dming@lanl.gov

[2] Carlson School of Chemistry and Biochemistry, Clark University, Worcester, MA 01610 USA E-mail: bruschweiler@nmr.clarku.edu

J37. Significance of conformational biases in Monte Carlo simulations of protein folding

Teresa Przytycka[1]

Keywords: protein folding, Metropolis-Hasting algorithm, detailed balance principle, biased sampling, hierarchical folding, minimalist models

Despite significant effort, the problem of predicting a protein's three-dimensional fold from its amino-acid sequence remains unsolved. An important line of attack is to treat folding as a statistical process, using the Markov chain formalism, implemented as a Metropolis Monte Carlo algorithm. A formal prerequisite of this approach is the *condition of detailed balance*, the plausible requirement that at equilibrium, the transition from state *i* to state *j* is traversed with the same probability as the reverse transition from state *j* to state *i*. Surprisingly, some relatively successful methods that use biased sampling fail to satisfy this requirement [1,2,4,5]. Is this compromise merely a convenient heuristic that results in faster convergence? Or, is it instead a cryptic energy term that compensates for an incomplete potential function? I explore this question using Metropolis-Hasting Monte Carlo simulations. Results from these simulations suggest the latter answer is more likely. The simulations are carried using a new full atom minimalist model. We argue that this model is more natural than the Go model [3]

References

[1] Bonneau R, Tsai J, Ruczinski I, Chivian D, Rohl C, Strauss CEM, Baker D. Rosetta in casp4: progress in ab initio protein structure prediction. *Proteins* 2001;Suppl 5:119 –126.

[2] Bradlay P, Chivian D, Meiler J, Misura KMS, Rohl CA, Shief WR, Wederneher WJ, Schueler-Furman O, Murphy P, Schnburn J, Strauss EM, Baker D. Rosetta Predictions in CASP 5: Successes, Failures and Prospects for Complete Automation. *Proteins: Structure, Functions and Genetics* 53:457-468.

[3] Go N. Protein folding as a stochstic process. J. Stat Physics 1983: 413-4

[4] Srinivasan R., Rose GD. LINUS - A Hierarchical Procedure to Predict the Fold of a Protein. *Proteins* 1995, 22(2): 81-99.

[5] Srinivasan R, Rose GD. A physical basis for protein secondary structure. *P. NATL. ACAD. SCI.* USA 1999, 96(25): 14258-14263

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health 8600 Rockville Pike, Bethesda, MD 20894;
E-mail: przytyck@ncbi.nlm.nih.gov

J38. Structure-based assessment of missense mutations in the HMGB domain of SRY identified in 46,XY females with sex reversal

Sharmila Banerjee-Basu[1] and Andreas D. Baxevanis[1]

Keywords: SRY, HMGB domain. mutation, molecular modeling

1 Summary

The SRY (sex-determining region of the Y chromosome) plays a key role in mammalian sex determination, as expression of the SRY gene initiates the process of testicular differentiation [1]. Mutations in the SRY gene are responsible for ~15% of 46,XY male-to-female sex reversal in humans. A total of 28 mutant proteins harboring sex-reversal missense mutations located in the conserved high-mobility group box (HMGB) domain of SRY were examined here. Comparative model building techniques were used to generate atomic structures of mutant proteins based on the NMR solution structure of HMGB domain of human SRY-DNA complex [2]. The impact of the missense mutations on the three-dimensional structure, stability, and surface electrostatic charge distribution of the HMGB domain of SRY are examined here. Seventeen missense mutations are located on the inner concave face of the HMGB domain; this region is involved in making contacts with the DNA recognition site as well as in nuclear localization. Nine specific missense mutations interfere with the pairwise interactions needed to stabilize the hydrophobic core of the HMGB domain. The mutant models have been compared to the wild-type protein in order to better understand the structural factors underlying these sex-reversal mutations.

References

[1] Berta P., Hawkins J., Sinclair A., Taylor A., Griffiths B., Goodfellow P., and Fellous M. 1990. Genetic evidence equating SRY and the testis-determining factor. *Nature* 348: 448-50.

[2] Murphy E., Zhurkin V., Louis J., Cornilescu G., and Clore GM. Structural basis for SRY-dependent 46-X,Y sex reversal: modulation of DNA bending by a naturally occurring point mutation. 2001 *Journal of Molecular Biology* 312: 481-99.

[1] Computational Genomics Program, Genome Technology Branch, National Human Genome Research Institute, NIH, Bethesda, MD 20892-8002, USA. E-mail: sharmib@nhgri.nih.gov; andy@nhgri.nih.gov

J39. Sequence and Structural Templates for Protein Protein Recognition Motifs

Owen V. Lancaster[1], Jo Avis[2] & Simon J. Hubbard[3]

Current methodologies for recognizing protein sequence patterns are predominantly based upon homology relationships between protein and nucleotide sequences. These methods are widely exploited to annotate new genomes and assign putative functions to new genes. However they are usually based on sequence data alone. More recent approaches have made use of available structural data and incorporated this information into methods to improve upon the predictions compared to just sequence based methods alone. So far these approaches have not been widely exploited in bioinformatics.

We have examined atest system containing degenerate but short, repeating motif, the tetratricopeptide repeat (TPR). Sequence analysis was done to assess the effectiveness of common search tools for finding TPR motifs. These methods included Blast, PSI-Blast and Hidden Markov Models . A full structural analysis was also performed. The simple repeat nature of the TPR motif allowed structural information to be obtained, and structurally conserved features in TPRs comprising conserved interacting residue pair positions were revealed.

Current work has involved building and evaluating models of all TPR sequences with unknown structures (over 7000) to check if they fit the TPR motif structure. From these and other models the interaction energy of structurally adjacent residue pairs has been calculated f residues. These models were generated by mutating residues in key conserved positions to all possible amino acid combinations. The energy is then evaluated for all these pair combinations - 20x20. This energy will then be integrated into sequence based methods such as Hidden Markov Models or other profiles with the aim of improving TPR predicting. Future work will be to go on to apply these methods to other simple repeating motifs.

[1] University of Manchester Institute of Science and Technology, Department of Biomolecular Sciences , Manchester, England
E-mail: O.Lancaster@postgrad.umist.ac.uk
[2] University of Manchester Institute of Science and Technology, Department of Biomolecular Sciences, Manchester, England
[3] University of Manchester Institute of Science and Technology, Department of Biomolecular Sciences , Manchester, England
E-mail: Simon.Hubbard@umist.ac.uk

J40. Partition Function and Base-Pairing Probability Algorithms for Nucleic Acid Secondary Structure including Pseudoknots

Robert M. Dirks[1] **and Niles A. Pierce**[2]

Keywords: DNA, RNA, secondary structure, pseudoknots, partition function, base-pairing probabilities, recursion probabilities

Abstract

Nucleic acid secondary structure models usually exclude pseudoknots due to the difficulty of treating these non-nested structures efficiently in structure prediction and partition function algorithms. Here, the standard secondary structure energy model is extended to include the most physically relevant pseudoknots. We describe an $O(N^5)$ dynamic programming algorithm, where N is the length of the strand, for computing the partition function and minimum energy structure over this class of secondary structures [3]. Furthermore, we describe a general method for transforming the partition function algorithm to compute a series of quantities termed recursion probabilities [4]. These, in turn, can be used to calculate base-pairing probabilities with or without pseudoknots. The partition function and base-pairing probabilities are useful for analyzing or designing the ensemble properties of RNA and DNA molecules, as illustrated for a human telomerase RNA implicated in dyskeratosis congenita and for a synthetic DNA nanostructure.

Introduction

Over the last two decades, polynomial-time dynamic programming algorithms have been developed to predict the minimum energy secondary structure [14, 9, 17, 5] of an RNA or single-stranded DNA molecule based on a nearest-neighbor empirical potential function [11, 7]. By eliminating redundancy in the recursive process, it is also possible to efficiently compute the partition function and base-pairing probabilities over secondary structure space [8]. In their original forms, these algorithms exclude the possiblity of pseudoknots, a biologically relevant class of secondary structures [13] that also arises in DNA nanotechnology applications [15, 16]. Pseudoknots present a major obstacle to dynamic programming methods because they destroy the locality implied by the nesting of loops in a secondary structure (see Figure 1). In fact, the structure prediction problem becomes *NP*-hard if all pseudoknots are included in the ensemble of secondary structures [6]. Recent extensions to the structure prediction [10, 1, 3] and partition function [3] algorithms allow the inclusion of certain physically-relevant pseudoknots while maintaining polynomial-time complexity.

The ensemble equilibrium can also be characterized by the matrix of base-pairing probabilities with entries $p_{i,j}$ corresponding to the probability that base i is paired with base j. We describe a general method for mechanically transforming the new pseudoknot partition function algorithm [3] to compute recursion probabilities, which can be used in turn to compute base-pairing probabilities [4]. The transformation approach is generalizable to any future partition function extensions that follow the same dynamic programming paradigm.

[1]Chemistry, Caltech. E-mail: dirks@caltech.edu

[2]Applied and Computational Mathematics, Bioengineering, Caltech. E-mail: niles@caltech.edu

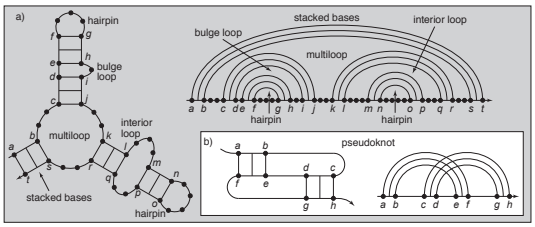


Figure 1: a) Canonical loops of nucleic acid secondary structure. These loops structures are all nested so there are no crossing arcs in the polymer graph with the polymer backbone drawn as a straight line. b) A sample pseudoknot with base pairs $a{\cdot}f$ and $c{\cdot}h$ (with $a < c$) that fail to satisfy the nesting property $a < c < h < f$. This leads to crossing arcs in the polymer graph.

Base pairing probabilities assist in the analysis of biologically relevant pseudoknots. Here, we examine human telomerase RNA, which exists at equilibrium in both hairpin and pseudoknotted forms [2]. A two-point mutation, implicated in the disease dyskeratosis congenita, alters the thermodynamic balance between these competing structures [12]. This shift in equilibrium is clearly identifiable when the base-pairing probabilities for the two sequences are compared. Base-pairing probabilities that permit pseudoknots are also useful in analyzing sequences designed for DNA nanotechnology applications [15, 16].

References

[1] T. Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Appl. Math.*, 104:45–62, 2000.

[2] L.R. Comolli, I. Smirnov, L. Xu, E.H. Blackburn, and T.L. James. A molecular switch underlies a human telomerase disease. *Proc. Natl. Acad. Sci. USA*, 99(26):16998–17003, 2002.

[3] R.M. Dirks and N. A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, 24:1664–1677, 2003.

[4] R.M. Dirks and N. A. Pierce. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. submitted. 2004.

[5] I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Chemical Monthly*, 125:167–188, 1994.

[6] R.B. Lyngso and C.N.S. Pedersen. RNA pseudoknot prediction in energy-based models. *J. of Comput. Biol.*, 7(3/4):409–427, 2000.

[7] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[8] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[9] R. Nussinov, J.R. Pieczenik, J.R. Griggs, and D.J. Kleitman. Algorithms for loop matchings. *SIAM J. Appl. Math.*, 35:68–82, 1978.

[10] E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[11] J. SantaLucia, Jr. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry*, 35:3555–3562, 1996.

[12] C.A. Theimer, L.D. Finger, L. Trantirek, and J. Feigon. Mutations linked to dyskeratosis congenita cause changes in the structural equilibrium in telomerase RNA. *Proc. Natl. Acad. Sci. USA*, 100(2):449–454, 2003.

[13] F.H.D. van Batenburg, A.P. Gultyaev, C.W.A. Pleij, and J. Ng. Pseudobase: a database with RNA pseudoknots. *Nucleic Acids Res.*, 28:201–204, 2000.

[14] M.S. Waterman. Secondary structure of single-stranded nucleic acids. In *Studies in foundations and combinatorics: Advan. in Math. Suppl. Studies*, volume 1, pages 167–212. Academic Press, New York, 1978.

[15] E. Winfree, F. Liu, L.A. Wenzler, and N. C. Seeman. Design and self-assembly of two-dimensional DNA crystals. *Nature*, 394:539–544, 1998.

[16] H. Yan, T.H. LaBean, L. Feng, and J.H. Reif. Directed nucleation assembly of DNA tile complexes for barcode-patterned lattices. *Proc. Natl. Acad. Sci. USA*, 100(14):8103–8108, 2003.

[17] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9(1):133–147, 1981.

J41. Profile–profile methods provide improved fold–recognition. A study of different profile-profile alignment methods.

Arne Elofsson [1] **Tomas Ohlson** [2] **Björn Wallner** [3]

Keywords: fold recognition, profile–profile alignment, PSI-BLAST, homology detection, sequence alignments

1 Introduction.

To improve the detection of related proteins it is often useful to include evolutionary information for both the query and target proteins. One method of including more evolutionary information is the use of profiles in profile–profile alignments. Profile–profile alignments can be implemented in several fundamentally different ways. In this study we present a large scale comparison of different profile-profile alignment methods. We show that the profile–profile methods perform 30 to 50% better than standard sequence-profile methods both in their ability to recognize superfamily related proteins and in the quality of the obtained alignments. Among the fold recognition methods the Log_Aver method developed by von Öhsen and Zimmer and a probaility based scoring method similar to PICASSO by Heger and Holm seem to be preferred as they can create good alignments and show a good fold recongition capacity using one set of gap–penalties gap–penalties, while the other methods need to use different parameters to obtain similar performances.

2 Conclusions

In this study we have shown that several different profile–profile methods perform significantly better than standard sequence–profile methods. The profile–profile methods show a greater ability to identify structurally related residues, provide better recognition and better alignments than standards sequence–profile methods. The different profile–profile methods perform quite similar if the gap–penalties are optimized individually for alignment and fold recognition abilities. However, it is possible that Log_Aver and prob_score has a slight advantage as these are the only method that show good performance in both fold recognition and alignment quality using identical parameters.

What are the reasons to the good performance of Log_Aver and prob_score ? Our assumption was that a good ability to identify related residues would be a requirement for good performance. However, this is obviously not the case, as prob_score show among the lowest ability to identify the related residues, while Log_Aver is the best method at identifying related residues. The distribution of the scores in DP appears not to be ideal as it is necessary

[1]Stockholm Bioinformatics Center, Stockholm University, 106 91 Stockholm, Sweden. E-mail: arne@sbc.su.se

[2]Stockholm Bioinformatics Center, Stockholm University, 106 91 Stockholm, Sweden. E-mail: tomas.ohlson@sbc.su.se

[3]Stockholm Bioinformatics Center, Stockholm University, 106 91 Stockholm, Sweden. E-mail: bjorn@sbc.su.se

to use quite different gap–penalties to obtain high quality alignments in comparison with the optimal penalties for fold recognition and vice–versa.

This study highlights that a better understanding of how different profile–profile methods perform should be useful for improving these methods. It is not obvious how the best distribution separating related and unrelated residues should look like, even if our intuition would tell us that the Gaussian distributions are good. The alignment quality performance seems to show some correlation with the ability to detect distantly related residues, while fold recognition ability seems not to be. It is promising that the quite different profile–profile methods perform quite well compared with PSI-BLAST. However, it is clear that a better understanding in what factors that affect the performance of fold recognition and alignment qualities has to be obtained if better methods are going to be developed.

References

[1] von Öhsen, N and Zimmer, R 2001 Improving profile-profile alignments via log average scoring, WABI, 11-26

[2] Heger, A. and Holm, L. 2003Exhaustive enumeration of protein domain families , J. Mol. Biol 328: 749-67

K1. The Probability of Occurrence in a Single Sequence

Ezekiel F. Adebiyi [1]

Keywords: motifs, gene finding, regular expression, deterministic finite automaton(DFA), statistical significance.

1 Introduction

The problem considered here is to determine p_s, the probability that a single random sequence X (whose characters are drawn from the set of characters in Σ), of length L contains at least one occurrence of s that is Edit distance at most D from s. The computation of p_s helps to determine statistical significance in a variety of pattern searches such as motif searches and gene finding[3, 6].

Tompa[6] developed the first direct and efficient algorithm to compute p_s, but for the case of at most one substitution and Markov chains of order 1, with no insertions or deletions. The algorithm takes $O(|X| \cdot |s|^2)$ time. Atteson[3] presented an algorithm for the exact computation of the probability that a random string of a certain length matches a given regular expression, allowing insertions and deletions. The problem considered in [3] is in the class of NP-hard problems, but it is fixed-parameter tractable. Thus his algorithm that runs in time $O(|\Sigma||X|e^{|s|})$, is of linear order in the length of the sequence for small patterns. Note that, the running time is often much smaller, infact polynomial for some restricted forms of regular expressions, than the worst case for most expression which occur in practice. Briefly, Atteson method calculates the probability that a random string of length L, matches a given regular expression by computing the DFA of the regular expression and multiplying the resulting sparse Markov chain transition matrix by the initial vector $|X|$ times.

A possible combination and extension of Atteson[3] and Tompa[6] ideas requires that, we solve a basic problem: the development of regular expressions (henceforth REs) that are able to represent the D-neighborhood of s. For our analysis below, we adopted the D-neighborhood of s according to Edit distance as defined in [5]. Let $\delta(V, W)$ be the Edit distance between V and W. The *D-neighborhood of a string W* is the set of all strings with Edit distance at most D from W, *i.e.*, $N_D(W) = \{V : \delta(V, W) \leq D\}$ and the *condensed D-neighborhood of W* is the set of all strings in the D-neighborhood of W that do not have a prefix in the neighborhood, *i.e.*, $\overline{N_D}(W) = \{V : V \in N_D(W)$ and no prefix of V is in $N_D(W)\}$. The application of the above concept with the combination of [3] and [6] ideas lead to the algorithm outline in the following sections. The remainder of this paper concentrates on the use of Edit distance, but note that the resulting approach is also applicable to Hamming distance. What changes in the whole system is the generation of D-neighborhood of s according to Hamming distance instead of Edit distance.

2 Formulating Regular Expressions and Computation of their single DFA

Given pattern s, the condensed D-neighborhood of s according to the definition above is generated and the words are concatenated but separated by special symbols for recognition, to form a string. We then build a suffix tree for this string. Finally, we use the algorithm of Gusfield[4] for finding left diverse nodes, to group the concatenated words into regular

[1]Department of Computer and Information Technology, College of Science and Technology, Covenant University, P.M.B 1023, Ota, Nigeria. E-mail: adebiyi@informatik.uni-tuebingen.de

expression classes, based on some maximal common substring that they shared. No two groups are allow to share the same word, that is, each word belong to only one RE class. Let t be the number of words in each RE class. To avoid over-representation of the D-neighborhood words by their resulting RE classes, let $t \leq |\Sigma|$. Using the regular expression definition of [7](page 238), the resulting REs that represent each class can either be simple or complex. For example, $(A|G|T|C)ATAATA$, $(\epsilon|C|G)ATAATA$ are complex REs, while $CTATAGT$ is a simple RE.

Wilhelm and Maurer[7] in section 7.4.2 presented an algorithm that compute a single DFA for a sequence of regular definitions. To the sequence of regular expression classes derived for a given pattern s, we apply their algorithm to compute the required single DFA.

3 Computation of the p_s and Discussion

Atteson[3] showed how the resulting single DFA above can be use to calculate p_s when the subject sequence is an $i.i.d$ random sequence or a rth-order Markov chain. We adopt his method, as it is over the more general set of all DFA's.

The full application of our method to finding the statistical significance of a pattern, for example, motifs will be done in [1]. We show in tables 1 and 2, how many simple and complex REs can be derived from a given pattern s. This tables also discuss the ratio of the size of condensed neighborhood of s, $\overline{N_D}(s)$, to the number of derivable regular expression classes. In a further task, we dramatically reduced the number of REs derivable from a longer pattern, by considering not all condensed D- neighborhood words of s but those that occur in the input sequences. This has been mentioned earlier on by Tompa[6]. This result is shown in table 2. The patterns used in the following tables are motifs found in at least half of the sequences of B. Subtilis sequences used in [2]. The entries of column 2 of each table indicate the number of complex and simple REs respectively.

Table 1:

s for $D=1$	# REs	$\lvert\overline{N_D}(s)\rvert$
TATAGT	14/1	34
GATATA	12/1	31
GTGACA	15/0	33
GTTGAG	13/1	32
TATAAT	14/0	33

Table 2:

s for $D=2$	# REs	$r.\ \lvert\overline{N_D}(s)\rvert$
TAAGAAAAA	49/0	124
TATTTAGAA	51/1	124
ATAAATGAA	65/2	158
GTGTTAAAA	48/1	122
CAAATATAA	59/0	132

References

[1] Adebiyi, Ezekiel F., and Kaufmann, M. *Extracting common motifs using consensus and weighted matrix models under the edit distance: Theory and Experimentation.* In preparation, 2004.

[2] Adebiyi Ezekiel F. and Kaufmann, M. *Extracting common motifs under the levenshtein measure: Theory and Experimentation.* 2nd Intl. Workshop, WABI, 140-156, 2002.

[3] Atteson, K. *Calculating the exact probability of language-like patterns in biomolecular sequences.* 6th Intl. Conf. Intelligent Systems for Molecular Biology, 17-24, 1998.

[4] Gusfield D. *Algorithms on strings, trees and sequences.* Cambridge University Press, New York, 1997.

[5] Myers E. *A sub-linear algorithm for approximate keyword matching.* Algorithmica 12, 4-5, 345-374, 1994.

[6] Tompa, M. *An exact method for finding short motifs in sequences, with application to the ribosome binding site problem.* 7th Intl. Conf. Intelligent Systems for Molecular Biology, 262-271, 1999.

[7] Wilhelm, R. and Maurer, D. *Compiler Design.* Addison-Wesley Publishing Company, 1996.

K2. A Minimization Entropy-Based Bipartite Algorithm with Application to PXR/RXRα Binding Sites

Chengpeng Bi[1]. Carrie A. Vyhlidal[2]. J. Steve Leeder[3]. Peter K. Rogan[4]

Keywords: motif discovery, information theory, entropy, bipartite module, multiple alignment

1 Introduction.

We developed a new method for the bipartite *cis*-regulatory module based on Shannon's entropy [1] minimization principle and applied to a set of known PXR/RXRα binding sites [2]. This work is an extension to [3]. A bipartite module is an independent functional unit on the upstream of a regulated gene and recognized by a protein binding complex such as PXR/RXRα heterodimer. We assume that two proteins (PXR and RXRα) cooperatively bind to the module with constrained spacers. The heterodimer binding controls the expression of co-regulated genes such as *CYP3A4*, which is involved in detoxification of drugs and xenobiotics [2]. We built the models for different motif widths and validated them based on the relative binding strength of the testing sequences.

2 Bipartite Motif Model.

A bipartite module has two components, left and right motifs, and the associated gap function, $g(d)$ defined as $-log(n(d)/n)$ and $n(d)$ is the number of sites with d. Shannon's entropy or uncertainty [1] was used to define the objective function (total information content, *IC*) which is given as,

$$IC = IC(left \mid d) + IC(right \mid d) - g(d) \tag{1}$$

$$IC(m \mid d) = \sum_{l=1}^{J_m} \left(E(H_{nb}) - H_m(l)\right), \qquad m \in \{left, right\} \tag{2}$$

$$E(H_{nb}) = \log_2 |D| - e(n), \quad D = \{A, C, G, T\} \tag{3}$$

Here J_m is the width of motif m and $e(n)$ is a sample correction [3]. Obviously the left and right motif sub-models are subject to a gap constraint. These two motifs are not allowed to be overlapping and the gap size (d) is set to a limited range [d_{min}, d_{max}] based on biological observation. The entropy $H_m(l)$ for motif m at position l is expressed as,

$$H_m(l) = -\sum_{b \in D} f^{(m|d)}(b,l) \log_2 (f^{(m|d)}(b,l)), \qquad D = \{A, C, G, T\} \tag{4}$$

$$f^{(m|d)}(b,l) = (c_{l,b} + \beta_b) \Big/ (n + \sum\nolimits_{b \in D} \beta_b) \tag{5}$$

$c_{l,b}$ is the count of symbol b at position l, β_b is the pseudo-count [2] of b and n is the number of the DNA training sequences.

[1] Laboratory of Human Molecular Genetics, Children's Mercy Hospital, Gillham Road, Missouri. E-mail: cbi@cmh.edu

[2] Division of Clinical Pharmacology & Therapeutics, Children's Mercy Hospital, Gillham Road, Missouri. E-mail: cvyhlidal@cmh.edu

[3] Division of Clinical Pharmacology & Therapeutics, Children's Mercy Hospital, Gillham Road, Missouri. E-mail: sleeder@cmh.edu

[4] Laboratory of Human Molecular Genetics, Children's Mercy Hospital, Gillham Road, Missouri. E-mail: progan@cmh.edu

3 Entropy Minimization in Bipartite Model.

The goal is to estimate the model parameters $f = (f^{(left|d)}, f^{(right|d)})$ that maximize *IC*. This problem can be reduced to minimize the total Shannon's entropy or uncertainty for each model,

$$f^{(m|d)*} = \arg\min_{(b,l)\in\Theta} \left\{ -\sum_{l=1}^{J_m} \sum_{b\in D} f^{(m|d)}(b,l) \log_2 (f^{(m|d)}(b,l)) \right\} \quad (6)$$

where $\Theta = \{a_k^{(m)} \in I\}, I = \{1, \cdots, L_k - J_m + 1\}, k = 1, \cdots, n$, $a_k^{(m)}$ is a set of start positions for motif *m*, and L_k is the length of sequence *k*. We applied a greedy algorithm to search the multiple local alignment space (Θ). The same idea can be applied to homogeneous model without gap.

4 Model Validation and Selection

The individual *IC* values (*Ri*) based on bipartite and homogeneous models were computed. The linear regression models were built to fit the *IC* change fold (X_i) to the binding strength (Y_i),

$$Y_i = K(t_i)/K(ref), \qquad \forall i \in \{1,2,\ldots,n\} \quad (7)$$

$$X_i = 2^{(Ri(t_i) - Ri(ref))}, \qquad \forall i \in \{1,2,\ldots,n\} \quad (8)$$

K is a competition constant at the equilibrium, t_i and *ref* is the tested and reference sequences respectively. The Akaike's information criterion, $AIC = n\log(SSE/n) + 2k$, was used to validate and compare the models, here *k* is number of parameters and *SSE* is sum of squared error.

5 Discussion.

The algorithms were implemented in C++/Perl and successfully applied to finding homogeneous and bipartite motifs using a set of PXR/RXRα protein binding sites. To validate the models, we scanned a set of testing sequences (binding strength data to be presented) using homogeneous and bipartite models and calculated their *Ri* values. Based on the regression models, we computed the correlation coefficients (r) and *AICs* for each model. The best model is the one with the smallest *AIC* values. There are twelve homogeneous model candidates with width ranging from 12 to 23 bps. Given the biologically defined gap range [0, 6], we set the half-site width ranging from 6 to 9 bps. Each bipartite model is one combination of half-sites separated by gap. Among the best bipartite models that fit the experimental data are 7<3>7 (7 bps on left and right half-sites with dominant spacer of 3 bps), 8<2>7 and 9<2>7 (r > 0.83). Therefore the width varies from 16 to 18 bps.

The results may support our hypothesis that PXR and RXRα transcription factors cooperatively bind to two adjacent motifs with variable spacing. Laboratory validation studies show that binding by the 16 bps motif reasonably approximates the prediction of the bipartite model.

References

[1] Shannon, C.E. 1948. A mathematical theory of communication. *Bell Systems Tech. J.* 27:379-623.

[2] Kliewer, S.A., Goodwin, B. and Willson, T.M. 2002. The nuclear pregnane X receptor: a key regulator of xenobiotic metabolism. *Endocrine Reviews*. 23:687-702.

[3] Shultzaberger, R.K., Bucheimer, R.E., Rudd, K.E. and Schneider, T.D. 2001. Anatomy of *Escherichia coli* ribosome binding sites. *J. Mol. Biol.* 313: 215-228.

K3. Identification of Regulatory Elements in Archaea using Self-Organizing Maps

Alan P. Boyle[1], John A. Boyle[2], Susan M. Bridges[3]

Keywords: self-organizing maps, regulatory elements, archaea, operon, *Sulfolobus solfataricus*

1 Introduction.

An explosion in the amount of available genomic data has changed the ways in which these data are analyzed. It has become possible to locate transcription and translation regulatory regions based on large scale comparisons of regions upstream of Open Reading Frames (ORFs). The use of self-organizing maps (SOMs) can aid in this search for cis regulatory elements in an organism by segregating similar patterns in different parts of the map.

A SOM is a basic neural network algorithm based on unsupervised learning. It implements reduced dimensionality mapping of the training set to produce a map that follows the probability density function of the data.[1] This unsupervised training system provides a relatively fast clustering that is, in many ways, better than traditional clustering models. The process of clustering upstream regions attempts to divide sequences of DNA into different groups based on feature vector values derived from a a positional weight matrix.[2] The use of this approach in the study of 5' flanking regions of *Sulfolobus solfataricus* ORFs has produced specific clustering that reveals different regulatory features associated with sets of ORFs.

It has been previously found that archaea use both eukaryotic and eubacterial means of transcription and translation.[3,4]. The use of SOM clustering has enabled us to identify a Shine-Dalgarno (S-D) region associated with some ORFs is complimentary to the 3' end of 16S ribosomal RNA in *Sulfolobus solfataricus*. We have observe an A box and a B box in a relatively fixed position upstream of the start of translation. By analysis of known data sets, we show that ORFs that are internal members of operons cluster together and generally lack the transcriptional feature we have identified. Conversely, first ORFs in operons cluster and have the A and B boxes.

2 Approach.

We used an approach that supports dynamic exploration of regulatory patterns in clusters of ORFs in *Sulfolobus solfataricus* by use of positional weight matrices and Kohonen's self-organizing map (SOM) algorithm. We have explored the data using different dimensions in the SOM lattice. We have also varied the extent of the windows to be used in scrutinizing the 5' flanks. ORFs are seen to cluster into groups with and without the regular TATA feature (A and B boxes) and with and without the Shine-Dalgarno sequence.

[1] Department of Computer Science and Engineering and Department of Biochemistry and Molecular Biology, Mississippi State University, Box 9637, Mississippi State, MS 39759, E-mail: apb22@cse.msstate.edu

[2] Department of Biochemistry and Molecular Biology, Mississippi State University, Box 9650, Mississippi State, MS 39759, E-mail: jab@ra.msstate.edu

[3] Department of Computer Science and Engineering, Mississippi State University, Box 9637, Mississippi State, MS 39759,. E-mail: bridges@cse.msstate.edu

The implications of these regulatory features with respect to the location of the ORFs in operons are considered.

3 Results.

ORFs were classified as either Distant or Nearby depending on the location of the their translation start sites relative to the stop codon of the nearest ORF. ORFs with starts located within ±25 nucleotides of the nearest stop codon were classed as Nearby. First ORFs in operons and Internal ORFs represent sets of about 100 ORFs each identified by inspection of the genome.

	First ORFs in Operons			Internal ORFs in Operons		
Features	S-D	TATA	Mixed	S-D	TATA	Mixed
	21%	60%	19%	62%	21%	17%
	Distant ORFs			Nearby ORFs		
	30%	63%	7%	57%	33%	10%

Table 1: Percentage of ORFs in clusters with identifiable cis elements. Window was -5 to -40 from translation start codon. Dimensions of SOM were 2 X 5.

Changes in window size and location had little impact on the clustering. Use of higher dimensions in the SOM allowed finer discrimination of cluster features but past some number of dimensions, the weight matrices become too noisy to analyze.

4 Conclusions

SOM used in conjunction with positional weight matrices allows for visualization of patterns of cis regulatory elements in genomes. Here we use *Sulfolobus solfataricus* as an example and show that the preponderance of ORFs internal to operons have only an S-D element as a recognizable feature. The preponderance of first genes in operons lack an S-D sequence but have a TATA box in a relatively fixed location. Other ORFs may lack this element or, more likely, have it in another location relative to the start codon. They may also have significant deviation from the recognizable consensus sequence. Division of ORFs into Nearby and Distant gives a similar clustering of cis elements as seen in the operon data. It would be expected that Nearby ORFs are much more likely to be members of an operon as compared to Distant ORFs.[5]

References

[1] Vesanto, J. (1999) SOM-based data visualization methods. Intelligent-Data-Analysis, 3: 111–126.

[2] Staden, R. (1984) Measurements of the effects that coding for a protein has on a DNA sequences and their use for finding genes. Nucleic Acids Res 12: 551-567.

[3]Kyrpides NC and Ouzounis CA (1999) Transcription in archaea. Proc Natl Acad Sci USA 96: 8545-8550

[4]Kozak M (1999) Initiation of translation in prokaryotes and eukaryotes. Gene 234: 187-208

[5] Wan, Xiufeng, Susan M. Bridges, and John A. Boyle. (2004) Revealing gene transcription and translation initiation patterns in arcaea using an interactive clustering model, Under review.

K4. Gene finding in the presence of RNA editing

Ralf Bundschuh, [1] **Jonatha Gott** [2]

Keywords: gene finding, RNA editing, Hidden Markov Model, sequence alignment

1 Introduction.

The central dogma of molecular biology states that the genetic information on the DNA is transcribed into messenger RNA which in turn is translated into proteins. For most organisms the messenger RNA has the same sequence as the DNA — possibly after eliminating introns — and the knowledge of the genetic code allows a one-to-one assignment of genomic sequences on the DNA and the protein sequences they correspond to. However, there are organisms that edit their messenger RNA sequences, i.e., a dedicated biological machinery in these organisms, substitutes, inserts, or deletes single or dinucleotides of the messenger RNA [4]. While the editing machinery for some of these organisms is known, it remains a complete mystery in other organisms. One of the organisms in which hardly anything is known about the editing machinery is the mitochondrion of the slime mold *Physarum polycephalum*. Its most frequent editing event is the insertion of single cytidines [3] and its mitochondrial genome has been sequenced [5].

In order to understand the RNA editing machinery in such organisms it is desirable to identify as many examples of sites at which the messenger RNA is edited as possible. However, the very existence of RNA editing complicates the application of traditional techniques to such organisms: In order to amplify and sequence a given messenger RNA exactly complementary primers are required. Due to the RNA editing, even the knowledge of the whole genomic sequence of an organism does not allow the construction of such primers because the edited sites are not contained in the genomic sequence of a gene. Also, the computational identification of genes within the genome is hindered by RNA editing since the unedited sequence in the genome lacks the features computational gene finders look for. Here, we present an approach that (i) is able to predict the location of genes on a genome of an organism with RNA editing and that (ii) predicts the position of editing sites in such genes thereby enabling primer selection

2 Approach.

Since the mitochondrial genes of a large number of organisms are known we choose a comparative approach to gene finding. We start with the protein sequence of a mitochondrial gene from a different organism and build a profile of the amino acid sequences of this gene using PSI-BLAST [1]. This protein sequence profile implies a hidden Markov model for the protein family along the lines of the models in the PFAM database [2].

We first turn this hidden Markov model which describes amino acid sequences into a hidden Markov model that describes underlying nucleotide sequences. Then, we modify this hidden Markov model in such a way that it takes into account the possibility of the insertion of individual cytidines. Scoring a piece of the genomic sequence of the mitochondrion

[1]Department of Physics, The Ohio State University, 174 West 18th Avenue, Columbus, Ohio 43210-1106, USA. E-mail: bundschuh@mps.ohio-state.edu

[2]Center for RNA Molecular Biology, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA. E-mail: jmg13@po.cwru.edu

of *Physarum polycephalum* with this hidden Markov model corresponds to calculating the probabilities for the translations of each possible way to insert cytidines into the genomic sequence to be members of the protein family in question and choosing the most probable among these ways of inserting cytidines. The dynamic programming algorithm of the hidden Markov model finishes this in principle exponential task with a time complexity that is quadratic in the sequence length.

The location of the gene on the genome can be determined by finding the region of the genome that scores the highest under local alignments against the hidden Markov model. In addition, tracing back through the model produces the most likely position of the inserted cytidines. The predictions are improved by incorporating biological information like the codon frequencies and typical sequence patterns in the vicinity of editing sites.

3 Results.

First, we apply our algorithm to the six mitochondrial genes of *Physarum polycephalum* for which mRNA sequences including the editing sites are available in GeneBank. We use the nad7 gene to optimize the two free parameters of the model and apply the algorithm to the other five genes, namely cox1, cox3, cytb, atp, and pL. We find that in total over 90% of the amino acids and over 70% of the actual editing sites are correctly predicted by the algorithm.

Second, we turn to the four genes nad2, atp8, nad4L, and nad6 that had been reported to be completely missing from the mitochondrial genome in Takano's analysis [5]. Our algorithm is able to find the location of each of these genes. We use the predictions for the editing sites to design primers and experimentally verify that all four genes were indeed present where predicted and had simply escaped gene prediction programs that do not take RNA editing into account in the past.

References

[1] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402.

[2] Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L.L. 2000. The Pfam contribution to the annual NAR database issue. *Nucleic Acids Research* 28:263-266.

[3] Mahendran, R., Spottswood, M.R. and Miller, D.L. 1991. RNA editing by cytidine insertion in mitochondria of *Physarum polycephalum*. *Nature* 349:434-438.

[4] Smith, H.C., Gott, J.M. and Hanson, M.R. 1997. A guide to RNA editing. *RNA* 3:1105-1123.

[5] Takano, H., Abe, T., Sakurai, R., Moriyama, Y., Miyazawa, Y., Nozaki, H., Kawano, S., Sasaki, N. and Kuroiwa, T. 2001. The complete DNA sequence of the mitochondrial genome of *Physarum polycephalum*. *Mol. Gen. Genet.* 264:539–545.

K5. Computational Identification of Noncoding RNA Genes through Phylogenetic Shadowing

Kushal Chakrabarti[1] **and Daniel L. Ong**[2]

Keywords: noncoding RNAs, gene prediction, comparative genomics

1 Introduction

Although fairly accurate databases exist for protein-coding genes, little is known about another important class of genes known as *noncoding RNA genes*. These genes, which have been implicated in a wide variety of critical biochemical pathways including brain development [7] and viral defense [5], are not translated into polypeptides. Instead, their transcribed RNAs fold into stable, base-paired secondary and tertiary structures that confer catalytic ability. For the purposes of this paper, it is especially important to note that these secondary structures cause noncoding RNA genes to contain pseudo-palindromic sequences.

Unfortunately, these pseudo-palindromic and other signals are not statistically sufficient for the computational identification of such genes [9]. Because they are difficult to detect even through biological techniques, it is important that accurate computational approaches be developed [10]. Although many heuristic and specialized methods have been suggested, comparative genomics approaches [8] have shown particular promise. However, even these approaches are primitive. For instance, current comparative genomics approaches are limited to two sequences, despite recent work showing the importance of using several related species [2]. Other problems include various heuristic approximations and poor scaling.

Here, we briefly present a machine learning approach to genome-wide noncoding RNA gene prediction with multiple sequence alignments. The algorithm we describe scales linearly with respect to the length and number of genomes. More importantly, the approach is statistically sound, and allows the direct computation of probabilities through modular protein-coding, noncoding RNA, and intergenic sequence models.

2 Theory

We have developed a graphical model that integrates *probabilistic context-free grammars (PCFGs)* and phylogenetic trees within a *generalized hidden Markov model (GHMM)* framework. The latter of these models, the GHMM, is a straightforward generalization of the HMM in which a state can explicitly choose the length of its emission (according to a specified length prior). Our GHMM can be conceptually considered as three different states, each of which emits sequence alignments based on protein-coding, noncoding RNA, or intergenic sequence models.[3] The protein-coding and intergenic models emit multiple alignment columns according to standard GHMMs [1] and phylogenetic trees.

On the other hand, the noncoding RNA model emits columns according to PCFGs and phylogenetic trees that define a joint probability distribution over pairs of columns. This simultaneous emission of multiple, distant columns allows the PCFG to model the evolutionary dependencies between base-paired positions within noncoding RNA genes [9]. In addition, the inclusion of phylogenetic trees permits us to model the expectation that base-paired

[1]Dept. of Computer Science, Univ. of Calif., Berkeley. E-mail: kushalc@uclink.berkeley.edu
[2]Dept. of Computer Science, Univ. of Calif., Berkeley. E-mail: dlong@ocf.berkeley.edu
[3]In practice, this is not strictly true because of noncoding RNAs encoded within introns.

positions within noncoding RNA genes will undergo *compensatory mutation*. Perhaps more importantly, alignments of several, closely related genomes can be used (*phylogenetic shadowing*) [2]; because noncoding RNA genes tend to mutate faster than protein-coding genes, such sequence sets prevent pathological alignments and increase the accuracy of comparative genomics approaches. Despite these apparent benefits, phylogenetic trees that define joint distributions over pairs of columns have only been recently developed [6].

Unfortunately, because PCFGs are too inefficient for long sequences, they must be approximated. Although Rivas and Eddy [9] use a windowing approach, it is both inefficient and probabilistically unsound. We instead enforce a standard constraint on the GHMM length priors, ie. that it be zero for all lengths beyond some constant C. This constraint allows us to straightforwardly and efficiently compute the necessary PCFG probabilities in time $O(C^3 + LC^2)$, where L is the length of the multiple alignment. For instance, the inside probability of the first C bases can be computed in the standard fashion in time $O(C^3)$, while the necessary probabilities for each remaining base can be incrementally computed in time $O(C^2)$. We briefly note that this approximation is "perfect," ie. the probability computed in this fashion is exactly equivalent to the probability computed naively.

3 Experimental Results

Seven yeast species were downloaded and analyzed. Genome-wide homology maps were first generated for these species [4], which were then used to align the genomes [3]. Aligned genomic DNA for annotated ORFs and noncoding RNAs was then extracted, and the remainder was classified as null sequence. The model was then used to reclassify these sequences; these model reclassifications were then compared against the known (true) classification. Although initial results are very promising, we intend to systematically study the performance of our algorithm against other noncoding RNA genefinders in the future.

References

[1] Alexandersson, M., Pachter, L., and Cawley, S. 2003. SLAM: Cross-species gene finding and alignment with a generalized pair hidden markov model. *Genome Research*, 13:496-502.

[2] Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., and Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299:1391-1394.

[3] Bray, N. and Pachter, L. MAVID: Constrained ancestral alignment of multiple sequences. Genome Research, in press.

[4] Dewey, C. Personal communication.

[5] Hamilton, A., Voinnet, O., Chappell, L., and Baulcombe, D. 2002. Two classes of short interfering RNA in RNA silencing. *The EMBO Journal*, 21(17):46714679.

[6] Jow, H., Hudelot, C., Rattray M., and Higgs, P.G. 2002. Bayesian Phylogenetics Using an RNA Substitution Model Applied to Early Mammalian Evolution. *Molecular Biology and Evolution*, 19(9):1591–1601.

[7] Krichevsky, A.M., King, K.S., Donahue, C.P., Khrapko, K., and Kosik, K.S. 2003. A microRNA array reveals extensive regulation of microRNAs during brain development. *RNA*, 9(10):12741281.s

[8] McCutcheon, J.P. and Eddy, S.R. 2003. Computational Identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics. *Nucleic Acids Research*, 31(14):4119-4128.

[9] Rivas, E. and Eddy, S.R. 2000. Secondary structure alone is generally not statistically sufficient for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583-605.

[10] Storz, G. 2002. An expanding universe of noncoding RNAs. *Science*, 296(5571):1260-1263.

K6. Discovering Transcription Factor Binding Sites in the Yeast *Saccharomyces Cerevisiae*

Xue-wen Chen[1], Jianwen Fang[2*], Xinkun Wang[3*]

Keywords: transcription factor binding site, genome sequence, microarray, fuzzy clustering

1 Introduction.

With the availability of whole genome sequence information and the large amount of gene expression profiling data from high throughput functional genomics approaches, it becomes clear that genome wide computational tools are needed for the analysis of transcript factors (TF) and for the identification of their binding sites. Even without knowing either a set of binding sites for a particular TF or a set of co-regulated sequences, computational algorithms are capable of predicting binding sites and finding their locations in sequences. Typical binding site identification algorithms find co-regulated genes using clustering algorithms. As microarray data are typically noisy, expression profile based clustering is fuzzy by nature. Thus, it is desirable to allow each gene to associate to all clusters with some degrees of certainty. In this paper, we introduce a fuzzy clustering based method for discovering binding signals. An initialized fuzzy C-means algorithm (initFCM) [1] is used for clustering genes in terms of their expression profiles; a web-based interface is developed to search for gene upstreams; furthermore, a standard motif discovery tool MEME [2] is used to find consensus patterns. The statistical significance of consensus patterns is measured in terms of both p-value and E-value, as this helps find distinctive patterns with little chance to be the background.

2 Methods.

The proposed system consists of the following components: fuzzy clustering, upstream search, and multiple sequence analysis. Gene expression profiles are first analyzed by the initFCM algorithms to find co-regulated genes. This algorithm is an iterative partitioning method and gives well-separated initial centers while avoiding the choice of outliers. Instead of assigning each gene to one and only one cluster, we allow each gene to have a membership value associated to each cluster. This cluster membership measures the degree of believe that a gene belongs to that particular cluster. Only genes with a certain degree of believe are considered to be coregulated. A web-based interface connected to a local yeast genome database is developed to retrieve the upstreams of co-regulated genes. It takes a list of gene ID, the length of an upstream sequence, and the position of the upstream starting site as inputs. The output is upstream sequences in FASTA format, which are fed to a multiple local alignment program, MEME, to identify the motifs and thus putative binding sites. The statistical significance of these motifs are evaluated in terms of their E-values (expectation values) and *p*-values. The E-value is an estimate of the number of motifs that would have equal or higher log likelihood ratio if the training set sequences have been generated randomly, while the *p*-value is the probability of a random sequence having the same match score or higher. The consensus sequences identified are considered to be potential regulatory signals.

[1] Corresponding author. Information and Telecommunication Technology Center, Electrical Engineering and Computer Science Department, The University of Kansas, Lawrence, KS 66045. E-mail: xwchen@ku.edu.

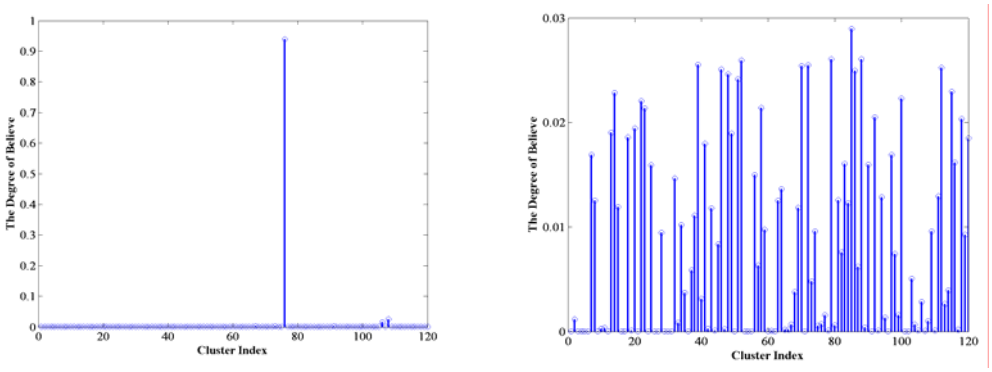
[2]Molecular Biosciences Department, The University of Kansas, Lawrence, KS 66045. E-mail: jwfang@ku.edu.

[3] Higuchi Biosciences Center, The University of Kansas, Lawrence, KS 66045. E-mail: xwang@ku.edu.

[*] Both authors contributed equally.

3 Results.

The data set analyzed here is the cDNA microarray data of *Saccharomyces Cerevisiae* in cell structures which were collected for 6221 genes under 80 different experimental conditions (e.g., cell cycle, sporulation, and diauxic shift) [3]. The initFCM algorithm is first applied for clustering yeast genes. The number of clusters is chosen to be 120. After clustering, some genes can be easily assigned to a cluster, while some can not. Figure 1 (a) and (b) show the degrees of believe versus the cluster index for two ORFs, YAL023C and YAL008W, respectively. For each gene *j*, we calculate the ratio of the maximum value of u_{kj} (k = 1, … , 120) and the second largest value of u_{kj}. If this ratio is no less than two, we assign this gene j to cluster $k = \arg\max_{k} u_{kj}$. Otherwise, we conclude that the cluster membership of gene *j* is fuzzy, and the corresponding gene will not be considered further. Clusters of sizes of 20 or more genes will remain for further analysis. For each gene in a cluster, the 600 bp upstreams are extracted using our web based interface; these upstreams are then fed to the MEME program to identify regulatory signals. Further analysis shows that most of the binding sites identified by our system are either verified by biological experiments or found in TRANSFAC database. For example, one cluster with 48 co-regulated ORFs, identified by our approach, has a consensus upstream sequence GAGGAAATTGAA (E-value = 8.1×10^{-17}), which is a substring of the experimentally verified sequence GAAGAGGAAATTGAA in SCPD database [4]. This sequence is related to the transcript factor GAL4 MCM1 UAS2CHA UASH.



(a) (b)

Figure 1: The degree of believes in clusters for (a) YAL023C and (b) YAL008W.

References

[1] Chen, X. 2002. Clustering Gene Expressing Data with Min-max-median Initialized Fuzzy C-Means Algorithms. *Workshop on Genomics Signal Processing and Statistics* (*GENSIPS 2002), NC: IEEE.*

[2] Bailey, T. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California.

[3] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. 1998. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* **9**, 3273-3297.

[4] Zhu, J. and Zhang, M. 1999. SCPD: a promoter database of the yeast Saccharomyces Cerevisiae. *Bioinformatics* 15 (7/8), pp. 607-611.

K7. Mammalian Promoter Database: Information resource of mammalian gene promoters

Hao Sun[1], Saranyan K. Palaniswamy[2], Twyla T. Pohar[3], Ramana V. Davuluri[4]

Keywords: promoters, transcription factors, *cis*-regulatory elements, comparative genomics

1 Introduction.

The control of gene expression, i.e. the transfer of biological information from DNA to RNA (transcription) and RNA to protein (translation), is central to all mammalian cellular processes. The majority of gene regulation occurs at the level of transcription initiation. This information is hardwired in the gene regulatory regions formed by *cis*-regulatory elements that bind specific TFs. Examples of the process of gene expression have been described for several organisms [1]. The basal transcriptional machinery of RNA polymerase II assembles at the core promoter, which is a minimum stretch of DNA sequence (from –35 to +35 nt of the transcription start site (TSS)) that is sufficient to direct the transcription initiation [2]. The proximal promoter region (from –500 to +250 nt of TSS) contains the *cis*-regulatory elements of most of the TFs. Enhancers and silencers are located several kbp upstream of the TSS. Extensive molecular research has provided a wealth of such information about experimentally characterized proximal promoter sequences, TFs and their binding sites. This information is dispersed throughout various databases, such as GenBank, PubMed, TRANSFAC and DBTSS. The integration of such essential information with the human and rodent genome sequences is one of the major challenges of the post-genome era. The database may be searched for promoter sequences, TFs, and their direct target genes through a user-friendly web interface at http://bioinformatics.med.ohio-state.edu/MPromDb. A facility for batch download of a set of promoter sequences is also available at this website.

2 MPromDb.

We developed a novel database MPromDb, which consists of mammalian promoters with annotation of first exon (transcription start site to first donor site) and experimentally supported *cis*-regulatory elements. Promoter sequences of orthologous genes from different species are linked with each other and displayed in same annotation image. The current release of MPromDb (release 1.0) contains experimentally supported (Table 1) promoters and first exons, 6,088 TF binding sites (3,319 of human & 2,769 mouse) and 1,503 transcription factors with links to corresponding PubMed and GenBank references. Currently, MPromDb contains 9,907 pairs of human-mouse orthologous genes. The corresponding record displays both of the promoters of the orthologous genes, allowing a visual platform for comparison (**Figure 1**). We have implemented in house developed JAVA™ application framework called Genome Data Visualization Tool Kit (GDVTK) [3] for MPromDb database management and information presentation in the form of an image map of gene regulatory regions with interactive contextual menus for easy navigation. A Web interface to the MPromDb has been developed using the J2EE technology (JSP and Servlet). Users can search the database and retrieve the

[1] Div. of Human Cancer Genetics, Dept. of Mol.. Virology, Immunol and Med. Genetics, 420 West 12th Ave. Room 570A, Columbus, Ohio, USA. E-mail: sun.143@.osu.edu

[2] Div. of Human Cancer Genetics, Dept. of Mol.. Virology, Immunol and Med. Genetics, 420 West 12th Ave. Room 570A, Columbus, Ohio, USA. E-mail: palaniswamy-1@medctr.osu.edu

[3] Div. of Human Cancer Genetics, Dept. of Mol.. Virology, Immunol and Med. Genetics, 420 West 12th Ave. Room 570A, Columbus, Ohio, USA. E-mail: pohar-2@medctr.osu.edu

[4] Div. of Human Cancer Genetics, Dept. of Mol.. Virology, Immunol and Med. Genetics, 420 West 12th Ave. Room 524, Columbus, Ohio, USA. E-mail: davuluri-1@medctr.osu.edu

promoter sequence and associated annotation information of a specified gene in several ways. For example, a user may obtain the promoter of a gene by searching with Gene Name or Symbol, Locus Link, UniGene or GenBank Accession IDs. Alternatively, a user may obtain TF information, including binding site position, binding sequence and promoter annotation of target gene, by simply searching with its corresponding name.

3 Figures and tables.

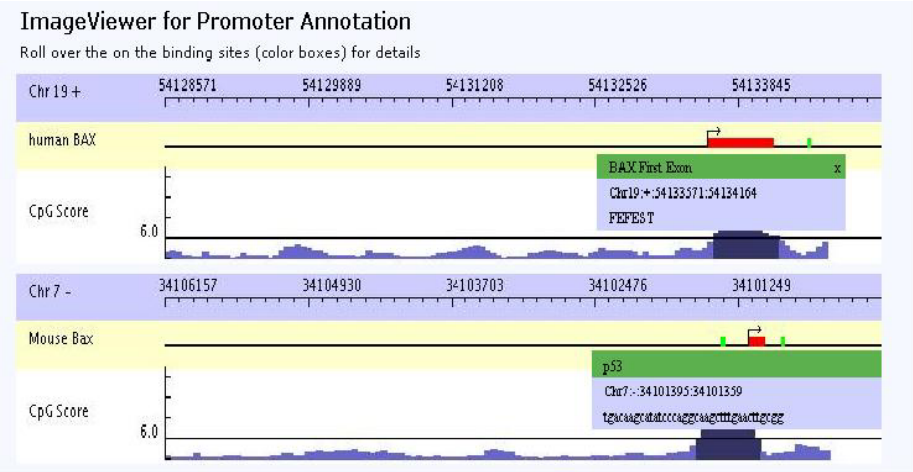


Figure 1: Screen shot of the genome view associated with the BAX promoter annotation

Organism	Promoter/first-exon mapping supported by					Total number of records
	full-length 5'UTR/ mRNA* e.g.	exon (numuber=1) e.g.	prime-transcript e.g.:A01198	DBTSS	Bidirectional promoter	
Human	5,213	1,109	174	4,010	3,274	13,780
Mouse	3,637	886	215	479	3,806	9,023
Rat	3,494	925	230	472	30	5,151
Total number of records	12344	2920	619	4961	7110	27954

Table 1: Promoters/first-exons in MPromDb that are supported by different types of experimental data

4 References.

[1] Orphanides G, Reinberg D. 2002. A unified theory of gene expression. *Cell.* 108: 439-451.

[2] Butler, J.E. and Kadonaga, J.T. 2002.The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev.* 16: 2583-2592.

[3] Sun, H, Davuluri, RV. 2003. Java-Based Application Framework for Gene Regulatory Region Annotations. *Bioinformatics.* In Pres.

K8. From Motif-Finding to Promoter Structure

Chun (Jimmie) Ye [1] **and Eleazar Eskin** [2]

Keywords: motif-finding, transcription factor binding sites, promoter regions

1 Motif-Finding

The discovery of transcription factor binding sites (TFBSs) by the analysis of promoter sequences of motif-finding is one of the most well studied problems in computational biology. Motif-finding algorithms discover statistically significant signals in the promoter regions. These signals are typically short patterns[3] or profiles[1] representing a motif or candidate TFBS. These signals are often statistically significant based on several different criteria. The signals may be over represented based on some background model. The signals may have stronger conservation in aligned genomes then what is expected or may be enriched in a set of genes with respect to a functional group or gene expression data. Each of these criteria, implicitly assume a different underlying model for the motif-finding problem. Each of these models finds different types of signals and their underlying statistical tests of often incompatible.

Once a motif-finder discovers a set of over represented signals, these are used as candidate TFBSs. However, while the algorithms for discovering over represented signals are often very sophisticated, the generation of a set of candidate TFBSs from a set of over represented signals is often done in a very ad-hoc way. The over represented signals are clustered together and heuristics are used to determine the TFBSs length.

In this project, we propose a set of post-processing techniques to make predictions for a set of TFBSs from a set of over represented signals. These techniques include statistical significance tests which incorporates many different models for motif significance. These statistical tests assign p-values for the motif under different models using different types of information such as a background set of sequences, gene expression data, positional tendencies of the signals, spacial relations of the signal with other known or predicted motifs and finally over representation in aligned genomes. This combination of statistical tests gives a complete picture of the evidence for whether a motif is an actual biological signal or just an artifact of the motif-finding algorithm.

We apply a variant of the MITRA[2] algorithm to discovering transcription factor binding sites to efficiently evaluate every possible pattern using each of the statistical tests.

In addition, in many cases, especially with signals that consist of patterns, often there are many closely related over represented signals. These signals may be either shifted variants of each other, signals of different length, or slightly different signals. We propose techniques for deciding how to merge multiple signals into a single prediction for a TFBS.

2 Promoter Modeling

Often when a researcher is interested in discovering TFBSs in a set of co-expressed genes, the researcher is interested in both known and unknown TFBSs. Typically, the a motif-finder is used to discover these TFBSs which look for any motifs.

[1] University of California, San Diego. E-mail: yimmieg@ucsd.edu
[2] University of California, San Diego. E-mail: eeskin@cs.ucsd.edu

The statistical models for known and unknown motifs are fundamentally different. This is because the number of known motifs is much smaller than the number of possible motifs. The statistical models for motif-finding algorithms make the assumption that the the motifs that they are looking for are unknown. Intuitively, if we are looking for a known transcription factor binding site, a weaker signal may still be strong evidence for the presence of the motif.

In this project, we introduce a similar set of statistical tests for evaluating the significance of observing a known TFBS in a set of sequences using various types of additional information.

By combining the algorithm and statistics for finding novel motifs with the algorithm and algorithm for finding the over represented known motifs we obtain a much richer picture of the promoter region and can quantify with p-values every part of our prediction.

3 Software Availability

The algorithm and statistical tests are available via webserver at `http://www.calit2.net/compbio/mitra`.

References

[1] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21:51, 1995.

[2] E. Eskin and P. A. Pevzner. Finding composite regulatory patterns in dna sequences. In *Special Issue Proceedings of the Tenth International Conference on Intelligent Systems for Molecular Biology (ISMB-2002) Bioinformatics.*, pages 1:S354–63, 2002.

[3] M. Sagot. Spelling approximate or repeated motifs using a suffix tree. *Lecture Notes in Computer Science*, 1380:111–127, 1998.

K9. A Whole-genome Analysis of Transcription Factor Binding Sites for Human and Mouse Orthologs

Caroline S Finnerty[1], Dr. James O McInerney[1]

Keywords: transcription factor binding sites, gene regulation, human and mouse orthologs.

1 Introduction.

We aim to test the hypothesis that dissimilarities in the control regions of genes are responsible for species-specific variation. We will do this by comparing human and mouse orthologs and analyse their upstream regions for presence or absence of various transcription factor binding sites (TFBS). We hope to see species-specific variation in TFBS that may be responsible at some level for species-specific differences. The biological significance of this work will be to determine if two genes which share a particular subset of transcription factors will have a similar function and if they will be expressed to the same level and conversely, if two genes have very divergent upstream sequences will their functions and expression levels also be dissimilar? The ultimate goal is to infer gene function from regulatory sequence.

2 Methods.

Our approach is to analyse, on a genome-wide scale the upstream regions of human genes with an emphasis on transcription factor binding sites. In order to take a conservative approach we have only used human genes, which have a corresponding mouse ortholog. This is commonly known as phylogenetic footprinting. Using existing databases such as TRANSFAC [2] to retrieve transcription factor binding site data we have recoded each transcription factor binding site with a different number so that each upstream region is identified by a different string of numbers. Using these newly recoded vectors, pairwise alignments using SWNumString.java (In-house software) have been carried out.

3 Discussion and Future Work.

These approaches should enable one to identify homologous upstream regions as well as those that are divergent which we can then analyse further.
We also hope to examine interspecies variation by performing multivariate analysis [3] on this dataset

[1]National University of Ireland, Maynooth, Co. Kildare, Ireland E-mail: caroline.s.finnerty@may.ie

4 References and bibliography.

[1] Dermitzakis, E.T. and Clarke, A.G. 2002. Evolution of transcription factor binding sites in mammalian gene regulatory regions: Conservation and turnover. *Molecular Biology and Evolution* 19(7):1114-1121

[2] Thioulouse, J., Chessel, D., Doledec, S., Olivier, J. M. 1997 ADE-4: A multivariate analysis and graphical display software *Statistics and Computing* 7(1)75-83

[3] Wingender et. al., 2000 TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Research* 28(1):316-319

K10. Identification of Regulatory Controls for Sets of Co-expressed Genes

Shannan J. Ho Sui[1], James Mortimer[2], Brian P. Kennedy[2], Chris J. Walsh[1], Wyeth W. Wasserman[1]

Keywords: promoter, transcription factor binding sites, comparative sequence analysis, gene expression, microarray, regulatory network

1 Introduction.

The use of large-scale gene expression profiling experiments to decipher underlying transcriptional networks is an intriguing and challenging area of research in bioinformatics. Creative computational algorithms are required to elucidate the transcription factors (TFs) that give rise to observed co-expression patterns by searching for shared *cis*-regulatory motifs in the regulatory regions of co-expressed genes. We have developed an integrated approach that combines cross-species comparisons with promoter motif identification tools for the automated detection of significantly over-represented transcription factor binding sites (TFBS) in sets of coordinately expressed genes.

The use of position-specific scoring matrices (PSSMs) to detect known TFBS is well-established (reviewed in [1]). However, these methods typically yield a large number of false positive predictions due to the short, variable nature of TFBS. Dramatic improvements in the specificity of TFBS prediction are attained by limiting the search space to regions of conserved, non-coding DNA using a comparative genomics approach known as phylogenetic footprinting [2].

2 Methods.

The promoter regions of human genes (defined as 5kb upstream and 1kb downstream of the annotated transcription start sites), were aligned to the corresponding promoter regions of their mouse orthologs (as defined by EnsEMBL). Regions of the alignments with greater than 75% sequence conservation were searched for matches to 75 vertebrate-specific TF binding profiles present in the JASPAR database [3]. Computational methods utilized the TFBS suite of regulatory analysis Perl modules [4].

Two statistical measures were calculated to determine which, if any, TFBS were over-represented in the set of promoters for co-expressed genes. These represent two distinct models for counting the occurrences of binding sites.

The z-score uses a simple binomial distribution model to compare the *frequency of occurrence of a TFBS* in the set of co-expressed genes to the expected frequency estimated from a background set containing all genes on the microarray chip. For a given TFBS, let the random variable X denote

[1] Centre for Molecular Medicine and Therapeutics, Vancouver, BC, Canada. E-mail: (shosui, cjwalsh, wyeth)@cmmt.ubc.ca

[2] Department of Biochemistry and Molecular Biology, Merck Frosst Centre for Therapeutic Research, Point-Claire, Dorval, Quebec H9R 4P, Canada. E-mail: (james_mortimer, brian_kennedy)@merck.com

the number of predicted binding site nucleotides in the conserved non-coding regions of the co-expressed genes. Let p be the rate of occurrence of predicted binding site nucleotides in the background sequences. Using a binomial model with n events, where n is the total number of nucleotides examined from the co-expressed genes, and p is the probability of success, the expected value of X is $\mu = np$, with standard deviation $\sigma = \sqrt{np(1-p)}$. Let x be the observed number of binding site nucleotides in the conserved non-coding regions of the co-expressed genes. By applying the Central Limit Theorem and using the normal approximation to the binomial distribution with a continuity correction, the z-score is calculated as $z = \frac{x - \mu - 0.5}{\sigma}$. Then, the probability of observing x or more binding site nucleotides in the conserved non-coding regions of the co-expressed genes is given by $\Pr(X \geq x) \cong \Pr(Z \geq z)$.

In contrast, the one-tailed Fisher exact probability compares the *proportion of co-expressed genes* containing a particular TFBS to the proportion of the background set that contains the site to determine the probability of a non-random association between the co-expressed gene set and the TFBS of interest. It is calculated using the hypergeometric probability distribution that describes sampling without replacement from a finite population consisting of two types of elements [5]. Therefore, the number of times a TFBS occurs in the promoter of an individual gene is disregarded, and instead, the TFBS is considered as either present or absent.

3 Results.

The method was validated on a number of reference sets, and then applied to genes significantly down-regulated in cells treated with a compound known to inhibit the NF-κB signaling pathway. TFBS that were significantly over-represented in the down-regulated set relative to the background set are shown in Table 1.

	TFBS	TF Class	z-score p-value	Fisher p-value		TFBS	TF Class	z-score p-value	Fisher p-value
1	NF-κB	Rel/ NF-κB	0.0e+00	3.2e-09	7	SPI-B	ETS	1.7e-17	2.2e-03
2	p65	Rel/ NF-κB	0.0e+00	4.0e-08	8	HFH-2	Forkhead	8.5e-17	2.0e-03
3	c-Rel	Rel/ NF-κB	0.0e+00	1.5e-04	9	FREAC-4	Forkhead	3.8e-16	5.9e-04
4	p50	Rel/ NF-κB	0.0e+00	5.5e-04	10	Max	bHLH-ZIP	2.4e-07	8.6e-03
5	Pbx	Homeo	3.3e-32	2.1e-03	11	SRY	HMG	1.8e-04	8.8e-03
6	Sox-5	HMG	1.5e-18	4.1e-03					

Table 1: Significant TFBS detected in genes down-regulated by treatment with the inhibitor (p-values less than 0.01 for both the z-score and Fisher measures).

4 References and bibliography.

[1] Wasserman, W.W. and Krivan, W. 2003. In silico identification of metazoan transcriptional regulatory regions. *Naturwissenschaften* 90:156-66.

[2] Lenhard, B., Sandelin, A., Mendoza, L., Engstrom, P., Jareborg, N. and Wasserman, W.W. 2003. Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* 2:13.

[3] Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. and Lenhard, B. 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32:D91-4.

[4] Lenhard, B. and Wasserman, W.W. 2002. TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics* 18:1135-6.

[5] Fleiss, J.L. 1981. *Statistical methods for Rates and Proportions*. New York: John Wiley.

K11. Splign: a Hybrid Approach to Spliced Sequence Alignments

Yuri Kapustin[1]; Alexander Souvorov, Tatiana Tatusova[2]

Keywords: gene prediction, exon assembly, spliced alignment, genome annotation, dynamic programming

1 Introduction.

As the number of available mRNA and EST sequences continues to grow, ability to accurately determine gene structure based on comparison with spliced sequences becomes a corner stone in eukaryotic genome annotation. Alignments with spliced sequences provide the most objective criteria for building gene models.

Although several programs have been developed in recent years aiming at the same goal, there is still a demand for a tool that would be computationally effective, comprehensive and accurate. To address this challenge, we developed a program called Splign utilizing a hybrid approach, where specialized global alignment procedures are applied locally to parts of sequences selected with Blast.

2 Methods.

At the core of the method is a modification of the Needleman-Wunsch algorithm that specifically accounts for conventional splice signals and imposes limits on lengths of introns. To facilitate proper handling of cases when sequencing errors affect splice signals, dynamic programming recurrences maintain scores associated with different levels of damage to every splice signal.

Before applying accurate, but costly dynamic programming procedure, the spliced sequence is aligned with Blast against its genomic counterparts to localize candidate locations on the genomic sequence. At this step, a procedure is used which based on the analysis of mRNA coverage by hits coming from different locations. After the candidate locations have been established, they are further refined with the dynamic programming algorithms limited to those regions that were missed or ambiguously aligned by Blast.

3 Comparisons.

In a series of comparisons performed over a set of mRNA sequences from human chromosomes 7 and 22, Splign has been compared to Sim4, Est_Genome and Spidey. Initial set of mRNA sequences was selected so that all the methods under comparison generated exactly the same models on it. Then random mutations were introduced to either mRNA, Genomic or to both sequences. For every mutation level, Splign has demonstrated the best accuracy and error tolerance running within the same time bounds as the other programs.

[1] MSD Inc., Vienna, Virginia, USA. E-mail: kapustin@ncbi.nlm.nih.gov

[2] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA.

K12. In silico studies of the transcriptional regulation of the genes coding for the novel IL28A, IL28B, and IL29 protein family: A computational screening approach applicable on a genomic scale.

William Krivan,[1] Brian Fox,[2] Emily Cooper,[2] Teresa Gilbert,[2] Frank Grant,[2] Betty Haldeman,[2] Katherine Henderson,[2] Wayne Kindsvogel,[2] Kevin Klucher,[2] Gary McKnight,[2] Patrick O'Hara,[2] Scott Presnell,[2] Monica Tackett,[2] David Taft,[2] Paul Sheppard[2]

Keywords: interleukins, interferons, transcriptional regulation, phylogenetic footprinting

The novel IL28A, IL28B, and IL29 protein family consists of three non-allelic human proteins, and homologous mouse proteins, which are distantly related to interferons and IL-10 [1]. We use this protein family to illustrate an approach to the computational identification and characterization of putative transcriptional regulatory regions that consists of a combination of available and novel techniques (see [2] for a review) that can be applied on a genomic scale.

Insights into the regulatory mechanisms of the novel IL28A,B and IL29 protein family may be gained from comparisons of their potential regulatory regions with the regulatory regions of characterized cytokines such as IFN-alpha, beta, and gamma. In metazoans, however, it is in general not feasible to study co-regulation of paralogous genes by simply performing alignments of the upstream genomic sequences, an approach that has been successfully pursued for yeast and bacteria. Comparisons of potential regulatory regions must reveal subtle similarities such as individual transcription factor binding sites. However, the low binding specificity of transcription factors results in a high rate of false predictions in the computational analysis of genes from metazoan species. The number of predicted sites can be reduced by about one order of magnitude to a set more likely to have sequence-specific functions by means of phylogenetic footprinting, a conservation-based filter based on the biological observation that regulatory regions are often more highly conserved between species than other non-coding regions [3]. Another technique that can be used for the selection of presumably functional motifs is motivated by the observation that groups of transcription factors rather than single factors are required for biologically functional regulatory regions and is based on the hypothesis that statistical significance of clusters of sites is correlated with biological function [4, 5].

We illustrate the combined application of these techniques for the characterization of putative regulatory regions of IL28A,B and IL29. We also present results from genomic screens using a liver-specific model [6] and a model for immune-related gene function [7].

References

[1] Sheppard, P. et al. 2003. IL-28, IL-29 and their class II cytokine receptor IL-28R. *Nat. Immunol.* 4:63–68.

[2] Wasserman, W.W. and Krivan, W. 2003. In silico identification of metazoan transcriptional regulatory regions. *Naturwissenschaften* 90:156-166.

[1]ZymoGenetics, Inc. 1201 Eastlake Ave. East, Seattle, WA 98102, USA. E-mail: krivan@zgi.com
[2]ZymoGenetics, Inc. 1201 Eastlake Ave. East, Seattle, WA 98102, USA.

[3] Wasserman, W.W. et al. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet.* 26:225-228.

[4] Wagner, A. 1999. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* 15:776–784.

[5] Frith, M.C., Li, M.C. and Weng, Z. 2003. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.* 31:3666-3668.

[6] Krivan, W. and Wasserman, W.W. 2001. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* 11:1559-1566.

[7] Liu, R., McEachin, R.C. and States, D.J. 2003. Computationally identifying novel NF-κB-regulated immune genes in the human genome. *Genome Res.* 13:654-661.

K13. Gene modeling using cDNA or amino acid to genomic sequence alignments

Roland Luethy[1]

Keywords: gene model, splice sites, exon detection

1 Introduction

The availability of large sequence databases and completed genome sequences have made the mapping of cDNA or protein sequences onto genomic sequences an important genomics analysis technique [4, 7, 2, 1]. Knowledge of the exact location of the exons is necessary to design probes to be used for the detection of polymorphism within a specific exon. At the same time the locations of noncoding regions is important to know for studying regulatory processes. New genes and their exon structure can be predicted by matching the sequences of homologous proteins from other species onto a new genome [2, 1]. Another aspect of gene-modeling is the search for alternative splice forms of a gene. Alternative splicing is thought to be a way to expand the gene repertoire and the alternative splice forms of a gene might have different or attenuated functions. Here we describe a software tool to find the exons and introns of a gene and to construct a gene model with correct splice sites. This is achieved by aligning a cDNA or amino acid sequence to a genomic DNA sequence using a double affine Smith-Waterman algorithm with the second gap extension penalty set to 0. This allows long gaps to span over the introns. These intron-spanning gaps are expected to start and end close to, but not exactly at the appropriate splice donor and acceptor sites. Therefore the alignments are modified such that intron spanning gaps start and end at canonical splice sequence motifs.

2 Methods and Results

A software tool called GeneDetective was developed that uses the following four steps to build correctly spliced gene models:

1. a database search to find the best matching sequences using Tera-BLAST [6].
2. select a region of the genomic sequence around the initial matches.
3. re-align the genomic sequence region from step 2 with the cDNA or amino acid sequences using a variant of the Smith-Waterman algorithm [5], where the gap extension penalty drops to zero after 20 extension steps. This allows for long gaps that span over the intronic regions. In the case of protein to genomic sequence alignments the occurrence of frameshifts is also allowed.
4. the alignments from the previous step are inspected for occurrences of matching GT..AG or CT..AC splice donor-acceptor sites in the neighborhoods of starts and ends of long gaps. If such sites are found, the corresponding gaps are adjusted to start and end at the splice motifs.

A test dataset was derived from a set of 178 genomic sequences and annotated exons from [3]. For testing using cDNA sequences only the 138 sequences with more than one exon were used. Exons were considered correct when they started and ended at the correct splice sites. The first exon was considered correct when the end was correct and vise versa the last exon was counted as correct

[1] Timelogic Corp., 1914 Palomar Oaks Way, Carlsbad, CA 92008, E-mail: rolandl@timelogic.com

when it began at the correct site. With these criteria all 832 exons were correctly recovered using the method described above.
For testing with protein sequences only 127 sequences with the coding regions located on more than one exon were used. Again to be considered correct an exon had to start and end at the correct splice sites. The first exon was considered correct when it started at the translation initiation site and the last exon was correct when it ended at the stop codon. Using GeneDetective it was possible to get the completely correct gene model for 117 genes and 783 out of 803 total exons were found. The following table compares the results obtained by GeneDetective with GeneWise [1]:

	Correct gene models	Correct exons	Predicted exons	Specificity	Sensitivity
GeneDetective	117	783	810	0.97	0.98
GeneWise	103	752	795	0.97	0.94
Annotated	127	803			

Table 1: Comparison of GeneDetective and GeneWise. Correct gene models: the number of gene models where all exons were correctly found; correct exons: exons that had correct start and ends as annotated; Specificity = correct exons/predictged exons; Sensitivity = correct exons/expected exons

The missing exons were all of the kind where three or less amino acids were encoded on the first or last exons and the alignment was therefore missing the first or last exon.

In order to find alternative splice forms a variant of GeneDetective uses a genomic sequence to search an EST or protein sequence database. The matching sequences are pairwise aligned to the genomic sequence as described in methods. The resulting pairwise alignments are then all consolidated into a multiple sequence alignment where the genomic sequence serves as the anchoring sequence. The multiple sequence alignment can then be inspected for alternative exons.

3 Conclusions

GeneDetective provides a solution to find exons in genomic sequences by aligning cDNA or protein sequences of the genomic sequence. It can be used to derive a gene model or to find alternative splice forms of a gene.

4 References

[1] E. Birney, http://www.ebi.ac.uk/Wise2/doc_wise2.html

[2] M. S. Gelfand, A. A. Mironov and P. A. Pevzner, *Gene recognition via spliced sequence alignment*, Proc Natl Acad Sci U S A, 93 (1996), pp. 9061-6.

[3] R. Guigo, P. Agarwal, J. F. Abril, M. Burset and J. W. Fickett, *An assessment of gene prediction accuracy in large DNA sequences*, Genome Res, 10 (2000), pp. 1631-42.

[4] R. Mott, *EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA*, Comput Appl Biosci, 13 (1997), pp. 477-8.

[5] T. F. Smith and M. S. Waterman, *Identification of common molecular subsequences*, J Mol Biol, 147 (1981), pp. 195-7.

[6] TimeLogic, http://www.timelogic.com

[7] S. J. Wheelan, D. M. Church and J. M. Ostell, *Spidey: a tool for mRNA-to-genomic alignments*, Genome Res, 11 (2001), pp. 1952-7.

K14. Detecting Functional Modules of Transcription Factor Binding Sites in the Human Genome

Thomas Manke, Christoph Dieterich and Martin Vingron [1]

Keywords: gene regulation, regulatory modules, transcription factor network

1 Introduction.

Understanding the flexible and robust response of living cells to diverse environmental condition has become a major challenge in functional genomics. The identification of transcription factor binding sites (TFBS) and regulatory DNA elements presents only a first step into this direction. It is often thought that complex regulatory control is achieved due to synergistic action of transcription factor modules, which regulate gene modules with specific function.

Here we report on an *in-silico* approach to find putative modules of human transcription factors and their regulated genes. Our work is based on the identification of known binding motifs for about 400 human transcription factors (TF) in $\approx$ 12.000 conserved upstream regions Below we describe a comparison of our binding data with available experimental data, the extraction of significant and functional TF associations, and a first analysis of the human TF network.

2 Validating *in-silico* Binding Data

The identification of evolutionary conserved binding sites for the human genome has been reported previously in ([Dieterich *et al.*, 2002]). We took a list of TF motifs from the TRANSFAC database [Matys *et al.*, 2003] and exhaustively searched an annotated collection of conserved sequence elements between human and mouse. In order to compare our *in-silico* predictions with biological data we chose as reference the experiment by [Ren *et al.*, 2002], where the authors studied the binding of E2F to the selected promoter regions. Using only exact matches to known binding patterns, our search identifies 240 E2F binding sites, where Ren *et al.* find only 38. This indicates a large false discovery rate and motivates the search for biological more meaningful TF modules. However, already at this point, we are encouraged by a highly significant overlap of 26 binding sites ($p = 6.3 \times 10^{-8}$). We also observe a functional enrichment of the putatively regulated genes in GO-categories for DNA replication and developmental processes.

3 Transcription Factor Associations

We observed functional enrichment also for several other TFs, but our focus is on higher TF-modules. The simplest approach to extracting higher TF modules is to count the pair frequency and score it with respect to an expected random distribution. The resulting rank list of TF-pairs is represented as a threshold graph, in which synergistic TF-pairs often appear to form higher clusters.

To extract biclusters of TFs and genes more directly from the data, we model the binding information as a bipartite graph and apply a greedy biclustering algorithm as originally

[1]Max Planck Institute for Molecular Genetics, Ihnestr. 63-73, D-14195 Berlin, Germany.
E-mail: manke@molgen.mpg.de

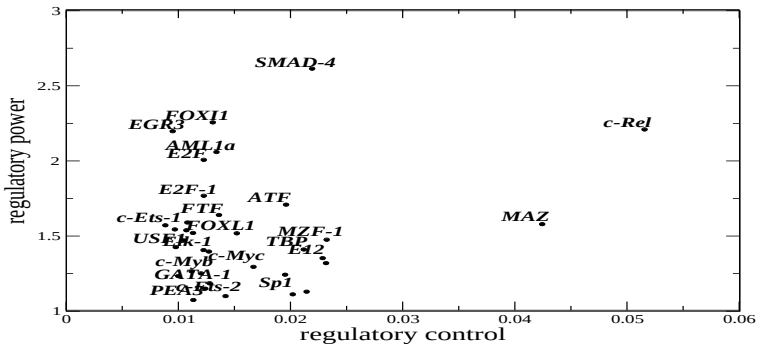


Figure 1: This plot highlights those factors of the TF-network with increased regulatory power (e.g. *SMAD-4*) and those whose genes are themselves under strict control (e.g *c-Rel*). The axis correspond to suitably normalized in- and out-degrees.

suggested for microarray data by Tanay *et al* ([Tanay *et al.*, 2002]). While the bicluster weights can be used to rank them according to their *topological relevance*, we processed the high-ranking clusters further, and screened their gene sets for *biological relevance* against different GO-categories. Since the assignment of functional categories is often somewhat arbitrary we decided to focus on categories with 50-100 member genes - children of a given category are automatically assigned the parental category. We refer to this binned sets as balanced categories. Our analysis yields a number of meaningful modules (sharing up to 10 TF), elements of which have previously been implicated in common biological processes. Other modules must remain speculative, and should be validated by further analysis.

4 Transcription Factor Network

Here we study direct genetic interactions of transcription factors as they bind to promoter regions of other TFs. In reality such cascades may be mediated by other proteins, but in the absence of large-scale interaction data we focus only on this simplest element of a regulatory circuit. Based on our binding data we identified a number of transcription factors with high *regulatory power* (regulating many other TFs) and those which are themselves under strict *regulatory control* (regulated by many). These results can be summarized conveniently in a power-control plot as shown in Figure 1.

References

[Dieterich *et al.*, 2002] Dieterich, C., Cusack, B., Wang, H., Rateitschak, K., Krause, A., & Vingron, M. (2002). Annotating regulatory DNA based on man-mouse genomic comparison. *Bioinformatics,* **Suppl 2**, 84–90.

[Matys *et al.*, 2003] Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A., & Kel-Margoulis, O. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research,* **31**, 374–378.

[Ren *et al.*, 2002] Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R. A., & Dynlacht, B. D. (2002). E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev,* **16** (2), 245–56.

[Tanay *et al.*, 2002] Tanay, A., Sharan, R., & Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics,* **18 Suppl 1**, S136–44.

K15. SHADOWER: A generalized hidden Markov phylogeny for multiple-sequence functional annotation

Jon D. McAuliffe, [1] **Lior Pachter,** [2] **Michael I. Jordan**[1, 3]

Keywords: Functional annotation, gene finding, phylogenetic shadowing, generalized HMM.

The prediction of functional regions in genomic sequences has traditionally been based on the identification of features associated with genes or regulatory regions. Comparison of homologous genomic sequences, e.g. from a pair of species, facilitates such identification [1, 5]. This is because functional regions tend to be conserved in sequences which have evolved from a common ancestor, whereas non-functional regions are more likely to mutate.

One drawback of pairwise comparative approaches to gene prediction is that non-functional regions are required to have diverged to a degree that enables statistical procedures to distinguish them from biologically active regions. These methods are therefore not applicable to discovering features present only at close evolutionary proximity, such as primate-specific genes. The *phylogenetic shadowing* principle of [2] circumvents this problem by seeking to identify conserved regions among multiple closely-related organisms. This has numerous advantages: sequence alignment is straightforward, the relevant phylogenetic tree is easy to infer, and identification of conserved regions is possible using standard evolutionary models.

To provide a systematic computational methodology for annotating genomic sequences based on the principle of phylogenetic shadowing, we have developed the *generalized hidden Markov phylogeny* (GHMP). The GHMP is a probabilistic graphical model [4] that combines conservation-based constraints deriving from multiple genomic sequences with algorithmic ideas that have proven useful in single-organism gene annotation systems. Our approach synthesizes generalized hidden Markov model gene finders, evolutionary models of nucleotide substitution, and phylogenetic trees. Similar ideas have been presented by [6] and [7]. Our extensions include generalized hidden Markov dynamics; a frame- and phase-consistent dual-strand hidden state space, supporting single-exon, multi-exon, and incomplete gene prediction; GC isochore-specific parameters; deterministic constraints on repeats, gaps, and in-frame stop codons; more complete splice site modeling; and an automated iterative procedure for alignment and tree building. The annotation is obtained as the most *a posteriori* probable trajectory through a hidden space of functional states; this trajectory is computed efficiently using algorithms for graphical model inference. Figure 1 shows a subcomponent of the GHMP graphical model corresponding to an aligned forward-strand internal exon.

To limit the number of sequenced organisms required for functional annotation, we have also developed a methodology for species subset selection. The method chooses subsets according to a maximin criterion on the weight of subtrees within the overall phylogenetic tree relating the species. Theory and efficient algorithms for this *maximal Steiner subtree* approach will be described at the conference.

We have implemented SHADOWER, a gene prediction system based on the GHMP. Table 1 shows that, by exploiting the additional constraints from multiple-species conservation, SHADOWER outperforms existing *ab initio* methods on a small dataset of single exons from five separate gene regions, across 13 primates. The data were originally reported by [2]. In addition, an analysis using species subsets of various sizes, each chosen by the maximal Steiner subtree criterion, revealed that SHADOWER needs only five of the available 13 primates to attain the performance reported in Table 1.

[1]Department of Statistics, University of California, 367 Evans Hall, Berkeley, CA 94720.
E-mail: {jon,jordan}@stat.berkeley.edu
[2]Department of Mathematics, University of California, 970 Evans Hall, Berkeley, CA 94720.
E-mail: lpachter@math.berkeley.edu
[3]Division of Computer Science, University of California, 387 Soda Hall, Berkeley, CA 94720.

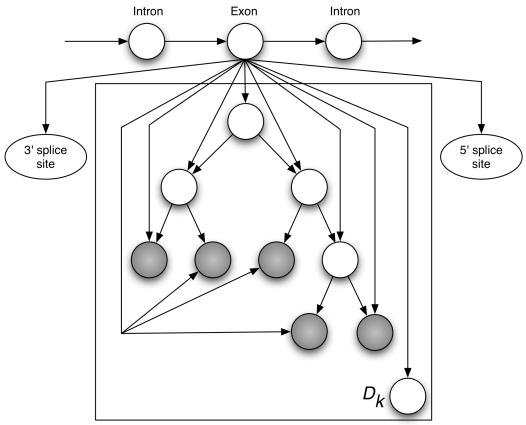


Figure 1: An excerpt of the GHMP graphical model corresponding to an aligned internal exon on the forward strand. The hidden chain of functional states runs along the top. Depicted underneath is a phylogenetic tree of nucleotides, both observed (shaded) and unobserved (unshaded). The bounding box (*plate*) around the phylogenetic tree denotes duplication, D_k times. Each copy of the tree corresponds to an alignment column, which populates the tree's leaves. D_k too is random, allowing the length of aligned exons to follow an arbitrary distribution (thus *generalized* hidden Markov phylogeny). The ovals labeled as splice sites are not part of the language of graphical models; they appear here to reduce visual clutter.

	Nucl.(%)		Exon Partial		Exon Exact	
	Sn	Sp	Sn	Sp	Sn	Sp
GENSCAN	44.7	34.0	2/5	2/3	1/5	1/3
MZEF	37.4	63.2	3/5	3/4	1/5	1/4
SHADOWER	**100.0**	**89.6**	**5/5**	**5/6**	**4/5**	**4/6**
SHADOWER^b	42.7	42.2	2/5	2/5	1/5	1/5
SLAM	80.2	100.0	3/5	3/3	3/5	3/3

Table 1: Sensitivity and specificity of various gene finders on the primate exon datasets. Results are shown at the nucleotide, partial exon (i.e. inexact boundaries), and exact exon level. GENSCAN [3] predicts complete or incomplete genes, using only the human sequence data. MZEF [8] predicts individual internal exons (without frame or phase consistency), using only the human sequence data. SHADOWER employs the GHMP to analyze multiple orthologous sequences. SHADOWER^b excludes exon boundary models, to exemplify a more limited approach based on multiple-species conservation. SLAM [1] uses human-mouse homology in a generalized pair HMM.

References

[1] Alexandersson, M., Cawley, S. and Pachter, L. 2003. SLAM—cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Research* 13:496–502.

[2] Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L. and Rubin, E. M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299:1391–1394.

[3] Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* 268:78–94.

[4] Jordan, M. I., ed. 1999. *Learning in Graphical Models*. Cambridge, MA: MIT Press.

[5] Korf, I., Flicek, P., Duan, D. and Brent, M. R. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* 17:S140–S148.

[6] Pedersen, J. S. and Hein, J. 2003. Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics* 19:219–227.

[7] Siepel, A. and Haussler, D. 2003. Combining phylogenetic and hidden Markov models in biosequence analysis. In: *Proceedings of the Seventh Annual International Conference on Computational Biology (RECOMB 03)*, New York: ACM. pp. 277–286.

[8] Zhang, M. Q. 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proceedings of the National Academy of Sciences USA* 94:565–568.

K16. Longer sequence surrounding motif distinguishes regulatory elements from false positives

Emily Rocke, [1] **James Thomas,** [2]

Keywords: motif, regulatory element, chemosensory, srh gene family, elegans

1 Overview.

The DNA sequence CASSTG[3] is overrepresented upstream of srh-family chemosensory genes in the nematode *C. elegans*. It very likely represents a regulatory binding site of interest to nematode chemosensation, although its regulatory effects and ligands are unknown.[3]

In this abstract, we make the observation that the CASSTG motif is embedded in a much longer (upwards of 40 nucleotides, although we look at only 26), stochastically conserved motif which is not palindromic. Computationally, the conservation of this longer motif allows instances of CASSTG as a regulatory motif to be distinguished from the many expected random instances of CASSTG in the genome. Because the longer motif is loosely conserved, the misidentification rate is high, but the approach is useful in identifying a number of very likely motif candidates. In this abstract we discuss the effectiveness of even a very simple algorithm which uses this additional information.

2 Introduction.

DNA regulatory elements are short DNA sequences (usually 6-20 nucleotides), typically appearing near or in the genes they regulate, that have evolved to be favorable binding sites for specific proteins or RNA molecules. The same type of protein or RNA molecule may bind to similar sites near several related genes; such patterns of similar sites, known as motifs, can often be detected computationally.

Any particular short DNA sequence is expected to occur many times in a genome by chance alone. For example, a given 6-nucleotide sequence is expected to occur about every 4,000 nucleotides, and so many thousands or even millions of times in a typical genome. These chance occurrences of a short pattern may overwhelm the number of legitimate binding sites, causing a difficult identification problem. One solution is to look at clusters of two or more nearby motif instances[1, 2], but this only works if such clusters exist in the data.

3 Method.

The nematode *C. elegans* has a large family of about a thousand genes that encode related 7-transmembrane proteins, presumed to act as chemoreceptors[5]. The large srh gene family[4] has approximately 200 genes and pseudogenes, of which 185 apparently active genes were used as a data set. The 1Kb region preceding each gene was searched for exact instances of CASSTG; 139 of the 185 regions had at least one occurrence of the motif, for 240 total motif occurrences, of which the 115 that fell closest to the genes were used to construct a weight matrix of 10 nucleotides on each side of the small motif.

[1] Genome Sciences Dept., University of Washington, Seattle. E-mail: ecrocke@gs.washington.edu
[2] Genome Sciences Dept., University of Washington, Seattle. E-mail: jht@u.washington.edu
[3] Here S, or strong, means that either nucleotide C or G may appear in this position

C. elegans chromosome V, about 21.7 million bases, was scanned for matches to either CACCTG or CAGGTG. 13,727 instances were found and sorted by the log likelihood score of the 20 surrounding nucleotides belonging to the motif weight matrix versus the background model. The 24 top-scoring instances were selected using a predetermined cutoff score, and each checked for plausibility using the Wormbase database (`http://www.wormbase.org/`). 9 of these top-scoring instances are recovered motifs from the original set, which are ignored since they were used in constructing the motif weight matrix.

4 Results.

Of the remaining 15 highest-scoring motifs, four immediately showed strong indications of being regulatory motifs. The other eleven can not yet be classified.

One interesting motif instance was juxtaposed with a gene that has been assigned to the srz family of chemosensory genes, a cousin of the srh family with about 40 member genes. This finding, while not conclusive, leads to a strong suspicion that the CASSTG motif regulates the srz family as well. We are now collecting data on the upstream regions of the srz family to test this hypothesis.

Two more high-scoring motif instances occur in front of srh-family *pseudogenes*, or inactive gene copies. The strong preservation of the regulatory motif suggests that these are very recent duplications or inactivations of active srh genes, which may be informative on the evolution of this gene family.

The fourth interesting motif instance occurs in the genome directly after the srh gene srh-120, and immediately before a nearly-perfect copy of srh-120. This copy does not appear to be annotated as a gene or pseudogene in Wormbase or in other sources. It is difficult to tell through computational means alone whether this gene is an inactive, very recent copy of srh-120, or whether it is a new active srh gene discovered through this method.

5 Conclusions.

A simple algorithm for scoring the moderately-conserved surroundings of a small, well-conserved motif allowed several new biologically interesting motif instances to be selected out of an overwhelmingly large set of false positives. This approach may help understand the elusive srh gene family, but the implications go beyond *C. elegans* biology. There is an intriguing possibility that other apparently small regulatory motifs are embedded in large, loosely conserved motifs. If this is often true, then computational methods of distinguishing "real" motif instances from false positives using surrounding sequence will become an important toolset in the arsenal of computational techniques.

References

[1] Frith, M. C., Hansen, U. and Weng, Z. 2001. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics*, 17(10):878-889

[2] GuhaThakurta, D. and Stormo, G. D. 2001. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17:608-621.

[3] McCarroll, S. and Bargmann, C. 2002. Personal communication.

[4] Robertson, H. M. 2000. The large srh family of chemoreceptor genes in Caenorhabditis nematodes reveals processes of genome evolution. *Genome Res.*, 10:192-203.

[5] Troemel, E. R., Chou, J. H., Dwyer, N. D., Colbert, H. A. and Bargmann, C. I. 1995. *Cell*, 83(2):207-218.

K17. Gnomon – a multi-step combined gene prediction program

Alexandre Souvorov[1], Tatiana Tatusova[1], David Lipman[1]

Keywords: HMM, gene, prediction, ab-initio, alignments

Gnomon uses a set of heuristics to find the maximal self-consistent set of corresponding transcript and protein alignment data to set the constraints for an HMM-based gene prediction. The goal is to ensure that if a biological expert is presented with the same data, they could not produce an obviously improved gene model. Using this set of heuristics Gnomon predicts the gene structure in genomic DNA sequences in a multi-step fashion.

The program evaluates the coding propensity of the available transcript alignments and determines their most probable coding regions. A single set of non overlapping transcript alignments with better coding propensity is chosen. Then the best matching proteins for these transcript alignments are aligned back on the genomic DNA sequence.

Gnomon makes the first pass of the prediction using the above transcript and protein alignments as the constraints. For the transcript alignments, the program makes sure that the chosen coding region is a part of a putative mRNA than can be extended on both sides of the predicted coding region. For the protein alignments, Gnomon checks that the predicted gene has every exon in the right frame as suggested by the protein alignment. Although in this case, the program is free to choose the splice sites and to introduce another exons between parts of the protein alignment.

The genes that were built using the alignments from the above step are included in the final output. For the rest of the gene models, the best matching proteins are found and then aligned back on the genomic DNA sequence. These protein alignments are used in the second pass of the prediction for refining the models.

While doing the alignment of the best matching proteins, Gnomon finds all cases where two exons of the protein alignment are within 50 bp and have different frames. Since the probability of such a short intron is extremely low, in all these cases the program introduces a frame shift in the genomic sequence allowing for combining the exons into a single one. In some cases protein alignments include a stop codon in the middle of the alignment. These stop codons are disregarded during the prediction and appear as premature stops in the model. Both the models with frame shifts and the models with premature stops are annotated as possible pseudo-genes in the Gnomon output.

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA.

K18. Experimental tools to determine DNA binding sites of KRAB zinc finger proteins in their candidate target genes – a challenge in computational biology of transcriptional regulatory networks

Peter Lorenz[1], Sabine Dietmann[1], Christian Sina[1], Dirk Koczan[1], Steffen Möller[1] and Hans-Juergen Thiesen[1, 2]

Keywords: target detection assay, RNA profiling, Kox1/ZNF10, antisense

1 Introduction.

The ordered expression of genes in transcriptional networks is an important means of living organisms to control biological processes. Transcription factors play pivotal roles in such networks. C2H2 zinc finger proteins form one of the largest protein superfamilies in the human genome with more than 1300 members according to INTERPRO (domain IPR007087**)**. Typically, these proteins contain an array of multiple C2H2 domains that is thought to specify nucleic acid binding and also contribute to protein-protein interactions. One particularly interesting subclass of the C2H2 zinc finger proteins contains the Krueppel-associated box (KRAB; INTERPRO domain IPR001909; more than 400 human members listed) domain that confers potent transcriptional repression activity upon targetted promoters (Thiesen, 1990; Margolin et al., 1995). We have started to combine experimental strategies for the search of target genes with bioinformatic approaches to define target gene signatures and DNA binding motifs of KRAB zinc finger proteins. The challenges to define such signatures and motifs reside in the degenerate binding code of individual zinc fingers, the potential different binding site specificities in a multi-zinc finger array in different parts of the protein and the fact that KRAB-mediated repression can also be elicited from remote positions with respect to the regulated gene.

2 Materials and Methods.

The target detection assay. (TDA; Thiesen and Bach, 1990) employed random 15mer double-stranded oligodeoxynucleotides and recombinant KRAB zinc finger proteins Kox1/ZNF10 (X52332) to enrich for DNA sequences with high binding affinity and to define respective binding matrices. A variation of this approach employed genomic fragments from a PAC clone containing C2H2 zinc finger genes instead of oligonucleotides. The impact of ectopic overexpression of KRAB zinc finger proteins on the global RNA expression profile of cultured human HeLa cells was recorded. Genes influenced in their transcriptional activity by this overexpression should contain direct targets of the KRAB zinc finger proteins as well as genes that are affected as secondary reactions, e.g. in gene networks. Antisense oligonucleotides were employed to downregulate intracellular expression of Kox1 in HeLa cells. Then concomitant changes in RNA expression profiles were monitored. Potential target genes should be to a certain extent relieved from Kox1-mediated repression and thus increase in their expression. Bioinformatics tools have been developed to evaluate and to combine the results of the experimental procedures.

[1] Institute of Immunology, Proteome Center Rostock; www.pzr.uni-rostock.de;
University of Rostock, Schillingallee 70, D-18055 Rostock, Germany
[2] E-mail: hans-juergen.thiesen@med.uni-rostock.de

3 Results.

Oligonucleotide sequences derived from the TDA selection were initially compared with each other to define DNA binding preferences of KRAB zinc finger proteins. In a second TDA approach, randomised oligonucleotides were replaced by a PAC clone encoding zinc finger gene sequences on human chromosome 10. Furthermore, zinc finger specific target genes were identified to be induced or repressed in HeLa cells with ectopically expressed zinc finger proteins. Finally, target genes were determined by KRAB zinc finger genes that had been inactivated by antisense-oligonucleotides. Sequence information of all four distinct approaches were taken to determine putative binding sites by making use of novel software tools. Affinity selection of recombinant Kox1 protein by the TDA resulted in 31 15mer oligonucleotide sequences that did not lead to a homogeneous consensus binding matrix. Since Kox1 contains nine functional C2H2 zinc fingers and each finger potentially contacts 3 nucleotide residues, a panel of 15mers is not sufficient in length to cover all putative binding sites offered by Kox1. Thus, binding activities within the same protein are most likely competing for the oligonucleotides being selected leading to the inhomogeneity observed. Indeed, taking into account the binding frequencies to 3-6mer sequence patterns argued for specific sequence preferences that could be found as well in the genomic DNA sequences selected by the TDA. Following the downregulation of Kox1 gene expression by specific antisense oligonucleotides 81 out of 44928 probed gene sets on Affymetrix microarrays were increased in their RNA levels. These genes constitute a Kox1 candidate target gene list in HeLa cells. Overexpression of Kox1 in HeLa cells did result in the increase of 32 and the decrease of 30 gene transcripts in their abundance. However, none of the genes were also found on the candidate target gene list after downregulation of Kox1 expression by the antisense approach. By applying sophisticated software tools the distribution of putative Kox1 DNA binding sequences were evaluated within the genomic sequences of Kox1 candidate target genes. Seed combinations of the double-stranded oligonucleotides selected by KOX1 proteins led to the identification of sequence motives within the original PAC DNA. Finally, these motives could be determined in target genes detected by Affymetrix microarray analysis as well.

4 Discussion.

The usefulness of our strategy is based on the combination of in-vitro selected DNA binding sequences with target gene signatures in vivo that were the result of perturbed KRAB zinc finger gene expression employing bioinformatic tools. However, the quality and robustness of bioinformatic information has still to be validated and confirmed in experimental settings – a challenge for developing more sophisticated algorithms mimicking DNA-protein interactions of KRAB zinc finger genes.

6 References and bibliography.

References

[1] Margolin, J.F., Friedman, J.R., Meyer, W.K., Vissing, H., Thiesen, H.J., and Rauscher, F.J., 3rd 1994. Krueppel-associated boxes are potent transcriptional repression domains. *Proc Natl Acad Sci U S A* 91:4509-13.

[2] Thiesen, H.J. 1990. Multiple genes encoding zinc finger domains are expressed in human T cells. *New Biol* 2:363-74.

[3] Thiesen, H.J., and Bach, C. 1990. Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucleic Acids Res* 18:3203-9.

K19. A Systematic Analysis of Stress Induced DNA Duplex Destabilization (SIDD) Sites in the *E. coli* Genome: Implications of SIDD Analysis for Promoter and Operon Prediction in Prokaryotes

Huiquan Wang[1] and Craig J. Benham[2]

Keywords: SIDD, promoter, operon prediction, transcription regulation, DNA supercoiling

Introduction

DNA structure and topologically driven structural transitions have been suggested to play important roles in regulating gene expression (1). Stress induced DNA duplex destabilization (SIDD) analysis exploits the known structural and energetic properties of DNA to predict the sites which are susceptible to become separated under superhelical stress (2, 3). Experimental results show that this analysis is quantitatively accurate in predicting transcriptional regulatory regions, matrix/scaffold attachment sites and replication origins (4, 5). Here we report a systematic analysis of the SIDD profile of the *E. coli* genome using a new algorithm specific for long genomic DNA sequences (6).

Results

1. Less than 7% of the *E. coli* genome has the propensity to be destabilized at the physiological superhelical densities (Figure 1).
2. Sites with high destabilization potential are statistically significantly associated with divergent and tandem intergenic regions, but not with convergent intergenic regions, and they strongly avoid coding regions (Figure 2).
3. More than 80% of the intergenic regions containing experimentally characterized promoters are found to overlap these SIDD sites (Figure 3).
4. A large majority of SIDD sites overlap long tandem intergenic regions, suggesting a potential role of SIDD sites in defining operon boundaries (Figure 4).
5. Strong SIDD sites are also found in the 5' upstream regions of genes regulating stress responses in *E. coli*, suggesting a possible link between their locations, the degrees of their destabilization, and the functioning of these genes (Table 1).

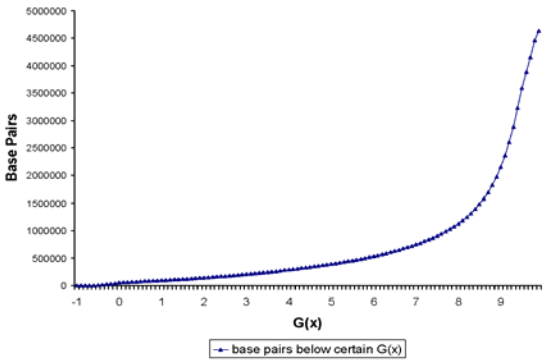


Figure 1. The cumulated G(x) distribution, the number of base pairs destabilized below the specified value.

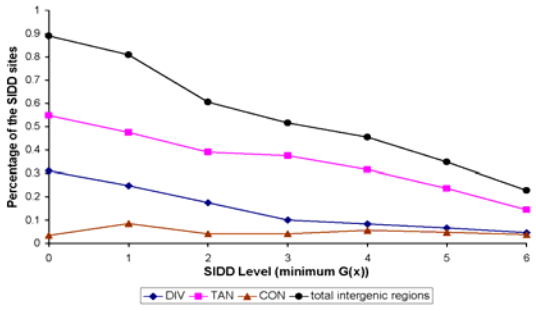


Figure 2. Percentage of SIDD sites at each level overlapping intergenic regions in E. coli genome

---

[1,2] UC Davis Genome Center, University of California, One Shields Avenue, Davis, CA 95616. Email: hqwang@ucdavis.edu; cjbenham@ucdavis.edu

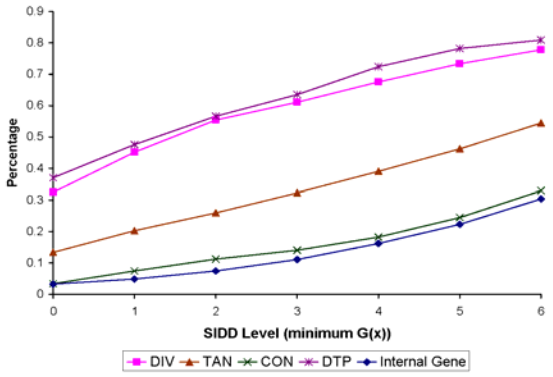


Figure 3: Percentage of DIV, TAN, CON, DTP and internal gene regions overlapping SIDD sites at each level in E. coli genome

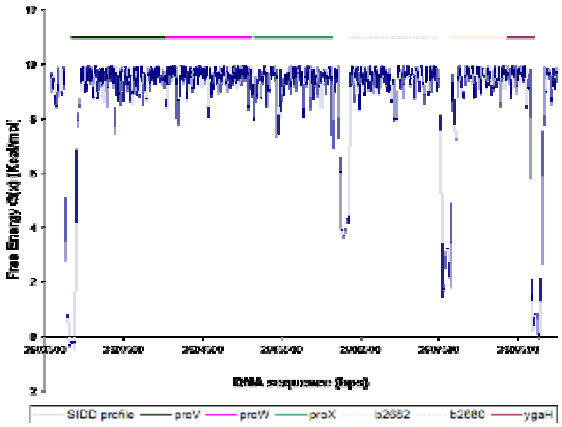


Figure 4. Strong SIDD sites are located at the boundaries of the proU operon. The genes are given above the graph and transcribe directly.

Table I. Global transcriptional regulators for stress responses with their 5' upstream regions overlapping the highly destabilized SIDD sites

Stress	SIDD	Gene	Function
Osmotic, nutrition starvation, cold	0	rpoS	Sigma S (sigma 38) factor of RNA polymerase, major sigma factor during stationary phase
Same as above	2	gyrA	DNA gyrase, subunit A, typeII topoisomerase
Same as above	0 0	hupA hupB	DNA-binding protein HU-alpha (HU-2), plays a role in DNA replication and in rpo translation
Same as above	1	H-NS	Transcriptional regulator, DNA-binding protein HLP-II, increases DNA thermal stability
Same as above?	1	crp	Transcriptional regulator, cyclic AMP receptor protein (cAMP-binding family), interacts with RNAP
Aerobic/anaerobic	0	fnr	Transcriptional regulator of aerobic, anaerobic respiration, osmotic balance (cAMP-binding family)
Aerobic/anaerobic	0	narX	Sensory histidine kinase in two-component regulatory system with NarL, regulation of anaerobic respiration and fermentation, senses nitrate/nitrite
Osmotic shock	0	ompR	response regulator in two-component regulatory system with EnvZ, regulates ompF and ompC expression (OmpR family)

References

1. Hatfield GW and Benham CJ, 2002. DNA Topology-Mediated Control of Global Gene Expression in Escherichia coli, *Annu. Rev. Genet.* **36,** 175-203.

2. Benham CJ, 1992. The energetics of the strand separation transition in superhelical DNA. *J. Mol. Biol.* **225**, 835-847.

3. Benham CJ, 1996. Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions. *J. Mol. Biol.* **255**, 425-34.

4. Sheridan SD, Benham CJ and Hatfield GW, 1998. Activation of gene expression by a novel DNA structural transmission mechanism that requires supercoiling-induced DNA duplex destabilization in an upstream activating sequence. *J. Biol. Chem.* **273**, 21298-21308.

5. Leblanc BP, Benham CJ and Clark DJ, 2000. An initiation element in the yeast CUP1 promoter is recognized by RNA polymerase II in the absence of TATA box-binding protein if the DNA is negatively supercoiled. *Proc. Natl. Acad. Sci. USA* **97**, 10745-50.

6. Benham CJ and Bi C-P, 2004. The Analysis of Stress-Induced Duplex Destabilization in Long Genomic DNA Sequences *J. Comput. Biol.* to appear

K20. Blind Operon Finding in Genomes with Insufficient Training Data

Ben Westover[1], Jeremy Buhler[1], Jeff Gordon[2], and Justin Sonnenburg[2]

Keywords: operon finding, *Bacteroides thetaiotaomicron*, transcriptional regulation

1 Introduction.

Several methods have been presented for detecting operons in bacteria, but the majority of these methods require using a training set of known operons to learn a model of what operons look like. Examples include [5] and [1], which describe operon finders trained for *E. coli*, and [4], which describes an operon finder trained for *B. subtilis*. We have developed a system for detecting operons that does not rely on previous knowledge of operon organization within a genome but instead bases its predictions on *analytical* criteria derived from simple *a priori* assumptions about the properties of operons. We chose this approach to enable reliable operon predictions in bacteria containing few well-characterized operons, in particular *Bacteroides thetaiotaomicron* (*B. theta*) [7], a prominent, yet relatively uncharacterized member of the human gut microbiota.

2 Methods.

For each pair of adjacent, same-stranded genes in the input genome, our operon finder computes the probability that the pair is co-transcribed. Our predictions make use of four specific types of information: distance between adjacent ORFs, functional relatedness of gene pairs, regulatory sites predicted in the intergenic space between pairs, and homologous gene clusters occurring in related species. We represent each source of information as a random variable X that can take on a number of discrete observable values $\{x_1, x_2, \ldots, x_n\}$. For each of the four sources, we calculate $\Pr(O|X_i = x_j)$, which represents the probability that a pair of genes belong to the same operon given that information source i has the value x_j. We merge all information sources to make a prediction using a naive Bayesian classifier.

A key feature of our software is the way *a priori* assumptions and statistical techniques are used to generate operon predictions in the absence of a training set. We use the assumptions and a variation of statistical methods introduced in work by Ermolaeva et al [3].

Another important feature of our approach is our use of homologous clusters. We have developed a novel technique for detecting and scoring clusters of genes occurring close to one another in many genomes. In a fashion similar to [2] we use the hypergeometric distribution to derive p-values for the clusters we detect.

3 Results.

In order to assess the performance of our blind operon finder on a dataset of known operons, we performed operon prediction for *E. coli* (Figure 1A) and compared our results against

[1]Dept. of Computer Science and Engineering, Washington University, One Brookings Drive, St. Louis, Missouri 63130, USA. E-mail: {ben,jbuhler}@cse.wustl.edu

[2]Dept. of Molecular Biology and Pharmacology, Washington University School of Medicine, 660 South Euclid Avenue St. Louis, Missouri 63110, USA. Email: {jgordon,jsonnenb}@molecool.wustl.edu

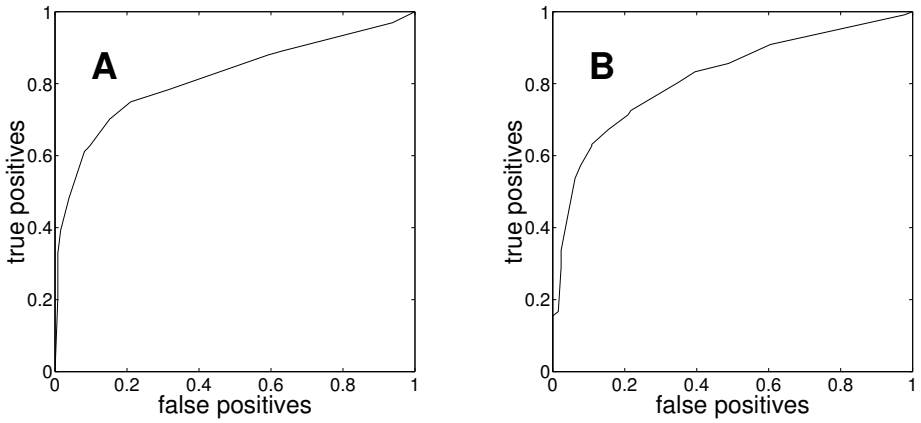


Figure 1: (A) ROC curve for operon prediction in *E. coli*, using RegulonDB as the reference dataset. (B) ROC curve for operon prediction in *B. theta* using gene expression data to generate the reference dataset.

the known operons contained in RegulonDB [6]. We used 914 gene pairs known to belong to the same operon and 257 same-stranded gene pairs known not to belong to the same operon.

To evaluate our software's performance in *B. theta* (Figure 1B), we used a set of time-course gene expression experiments to generate a set of putative predictions. Pairs of adjacent genes with strongly correlated expression were labeled as likely members of the same operon. We used 911 putative pairs of the same operon and 129 same-stranded putative pairs not of the same operon.

Our software correctly classifies around 75% of adjacent gene pairs at a false positive rate of 20%. We are performing experiments to validate some of our predictions in *B. theta*.

References

[1] Bockhorst J., Craven M., Page D., Shavlik J., and Glasner J. 2003. A Bayesian network approach to operon prediction. In: *Bioinformatics.* Jul 1;19(10):1227-35.

[2] Durand, D. and Sankoff, D. 2003. Tests for gene clustering. In: *J Comput Biol.* 10(3-4):453-82.

[3] Ermolaeva, M.D., White, O., and Salzberg, S.L. 2001. Prediction of operons in microbial genomes. In: *Nucleic Acids Research* vol. 29 pp. 1216-1221.

[4] De Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N., and Miyano, S. 2004. Predicting the operon structure of Bacillus subtilis using operon length, intergenic distance, and gene expression information. In : *Proceedings of the Pacific Symposium on Biocomputing (PSB 2004)* , in press.

[5] Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., and Collado-Vides, J. 2000. Operons in Escherichia coli: genomic analyses and predictions. In: *Proc.Natl. Acad. Sci.* vol. 97 pp. 6652-6657

[6] Salgado H., Gama-Castro S., Martinez-Antonio A., Diaz-Peredo E., Sanchez-Solano F., Peralta-Gil M., Garcia-Alonso D., Jimenez-Jacinto V., Santos-Zavaleta A., Bonavides-Martinez C., and Collado-Vides J. 2004. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. In: *Nucleic Acids Res.* Jan 1;32(1):D303-6.

[7] Xu J., Bjursell M.K., Himrod J., Deng S., Carmichael L.K., Chiang H.C., Hooper L.V., and Gordon J.I. 2003. A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. In: *Science* Mar 28;299(5615):2074-6.

K21. Bayesian Variable Selection to Identify Quantitative Trait Loci

Dabao Zhang[*], Min Zhang[†], Kristi L. Montooth[‡], Martin T. Wells[§], Carlos Bustamante[¶], Andrew G. Clark[‖]

January 1, 2004

Mutiple quantitative trait loci (QTL) regulating a specific phenotype can be approximately identified by using available genotypic markers. Classical approaches are usually developed using multivariate regression models with trait values regressed on marker genotypes. However, missing values of either trait values or marker genotypes and large number of marker genotypes relative to sample size challenge the identification of QTL using classical approaches based on multivariate regression models. We will take advantage of Bayesian inference on small datasets to develop a Bayesian variable selection approach to identify QTL, which is built up on a more natural Bayesian framework than the Bayesian methods in literatures.

[*]Department of Biostatistics and Computational Biology, University of Rochester Medical Center, 601 Elmwood Avenue, Box 630, Rochester, NY 14642

[†]Department of Statistical Science, Cornell University, Ithaca, NY 14853

[‡]Department of Molecular Biology and Genetics, Cornell Univeristy, Ithaca, NY 14853

[§]Departments of Statistical Science and Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853

[¶]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853

[‖]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853

K22. The Estimations of Motif Effects with Longitudinal Mixed Model in Temporal Gene Expression Analysis

Song Jiuzhou, Jaime Bjarnason and Michael G. Surette

Keywords: Regulatory analysis, Motif discovery, Longitudinal mixed model, Gene expression analysis

1 Introduction

The identification and testing of relevant transcription factor (TF) binding sites is one of the most important and greatest challenges in a functional genomics era. Traditionally, TF binding sites have been characterized by different experimental methods, a very slow and inefficient process. Although the similarity comparisons and multiple alignments of upstream sequences can find many significant repeats or conserved sequences upstream of the coding region, the statistically significant meaning of the putative motifs is based only on the frequencies of the nucleotides or patterns against the genome species[1-5]. It doesn't indicate the probability that the putative motifs are TF binding sites or that they have biological relevance for gene expression. Real TF binding sites must be confirmed by wet-bench genetic analysis. How to screen the motif candidates is becoming a critical issue. The motif effect indicates the regulatory extent of the motif for a given gene expression, so the estimation of the individual and combinational motif effects on gene expression will provide alternative support for the motif candidates, and improve the quality and efficiency of the screening process. We propose a longitudinal mixed model to estimate motif effects in temporal gene expression analysis,

2 Material and Methods

In a temporal gene expression experiment of iron responsive genes in *S. typhimurium*,[6] we clustered genes on the basis of their expression profile across four conditions and time points via cluster analysis. We adopted the Mismatch Tree Algorithm (MITRA) approach to obtain composite regulatory patterns [7]. We then assume that Y_{ij} satisfies $Y_{ij} = b_{1j} + b_{2i}t_{ij} + b_{3i}t_{ij}^2 + e_{ij}$, where n_i is the number of longitudinal measurements available for the ith gene, and where all error components e_{ij} are assumed to be independently normally distribution with mean zero and variance s^2, Y_i equals $(Y_{i1}, Y_{i2}, ..., Y_{in_i})'$, e_i equals $(e_{i1}, e_{i2}, ..., e_{in_i})'$, b_i equals to $(b_{1i}, b_{2i}, ..., b_{3i})'$, and Z_i is the $(n_i \times 3)$ matrix, the columns of which contain only ones, all time points t_{ij} and all squared time points t_{ij}^2. Ones have believed that $\hat{b}_{i,OLS}$ are good approximations to the real subject-specific regression parameters b_i, that is $b_i = B_i + b_i$. Then a linear mixed-effects model can be defined as any model which satisfies mixed model equations $Y_i = X_i b + Z_i b_i + e_i$.

3 Result and Analysis

We found three motif candidates in the analysis, A1: GATAATAATTATT, A2: TAATGATAATCATT and A3: ATAATTATTATCA, B: GCGT_ACGC and motif C: GCCGGA. In the mixed model analysis, the Table 1 shows the significant tests of the fixed effects, the motif candidates and the interactions among motifs, and time and quadratic time are very significant. Then the maximum likelihood (ML) and restricted maximum likelihood (REML) are used to estimating for all parameters in the longitudinal mixed model as shown in Table 2. The motif candidates A1, A2 and A3 are similar to the Fur binding site, GATAATGATAATCATTATC [8]. The

Department of Biochemistry and Molecular Biology, Health Sciences Centre. University of Calgary, Calgary, AB, Canada, T2N 4N1
Email: songj@ucalgary.ca , bjarnaso@ucalgary.ca, surette@ucalgary.ca,

estimates for the parameters shows that significant effects seem to be present among the motif candidates A and B, although they have opposite effects. The motif candidate C has the weakest effects (0.0438). There are significant positive interactions between motif candidate B and time effects, and weaker interactions between motifs A and C and time effects. The table also indicates that the interaction of all motif candidates and quadratic time effects are negative and weak. Those results suggest that motif effects are strongly influenced by gene expression level over time. The results indicate the evaluation of motif candidates is possible via longitudinal analysis.

Table 1 Type 3 tests of fixed effects

Effects	DF	Den DF	F value	Pr>F
Motif	3	262	9.99	<0.0001
Time*Motif	3	262	23.67	<0.0001
$Time^2$ *Motif	3	262	18.00	<0.0001

Table 2. The estimations of main effects and interaction of the motif candidates

Effects	ML(s.e.)	REML(s.e.)
Motif A	0.3278(0.0854)	0.3278(0.0869)
Motif B	-0.3569(0.0887)	-0.3569(0.0901)
Motif C	0.0438(0.1482)	0.0438(0.1507)
Time * Motif A	0.0636(0.0436)	0.0636(0.0443)
Time * Motif B	0.2965(0.0452)	0.2965(0.0460)
Time * Motif C	0.6006(0.1129)	0.6006(0.1148)
$Time^2$ * Motif A	-0.0036(0.0047)	-0.0036(0.0048)
$Time^2$ * Motif B	-0.0166(0.0049)	-0.0166(0.0050)
$Time^2$ * Motif C	-0.1222(0.0185)	-0.1222(0.0188)

4 Reference

[1] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat Genet*, vol. 22, pp. 281-5, 1999.

[2] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church, "Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae," *J Mol Biol*, vol. 296, pp. 1205-14, 2000.

[3] J. D. H. Frederick P. Roth, Preston W. Estep, and George M. Church, "Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation," *Nature Biotechnology*, vol. 16, pp. 939-945, 1998.

[4] V. R. Hao Li, Carol Gross, and Eric D. Siggia, "Identification of the binding sites of regulatory proteins in bacterial genomes," *PNAS*, vol. vol.99, pp. 11772-11777, 2002.

[5] G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics*, vol. 15, pp. 563-77, 1999.

[6] J. Bjarnason, C. M. Southward, and M. G. Surette, "Genomic profiling of iron-responsive genes in Salmonella enterica serovar typhimurium by high-throughput screening of a random promoter library," *J Bacteriol*, vol. 185, pp. 4973-82, 2003.

[7] E. Eskin and P. A. Pevzner, "Finding composite regulatory patterns in DNA sequences," *Bioinformatics*, vol. 18 Suppl 1, pp. S354-63, 2002.

[8] C. F.Earhart, "Uptake and Metabolism of Iron and Molybdenum," *Frederick C.Neidhardt, Editor in Chief. Escherichia coli and Salmonella Cellular and Molecular Biology*, vol. 1, pp. 1075-1090, 1996.

K23. Excess Information at T7-like Promoters and Classification of T7-like Phages

Zehua Chen, [1] **Thomas D. Schneider,** [2]

Keywords: T7-like promoters, T7 group, excess information, individual information.

1 Introduction.

Transcription plays a key role in the expression of genetic information. Most double-stranded DNA bacteriophages utilize the transcription system of their hosts, while the T7 group of phages mainly utilize their own transcription system, which consists of a single subunit RNA polymerase (ssRNAP) and a set of promoters located on the phage genome. Molecular information theory can be used to precisely characterize the sequence conservation at DNA binding sites, and has been widely applied to many genetic systems. As early as 1984, molecular information theory was applied to analyze the T7 promoters. It was noted that a roughly two fold excess information exists at T7 promoters [1]. In this work, we extended analysis to the other T7-like phage promoters and built promoter models for seven phages. The results show that excess information exists for all seven models.

Besides the ICTV taxonomic system and a genome-based classification which was recently proposed [2], some unique features could also be used for classification for certain group of phages. We propose that the T7-like promoters and the phage specific RNAP are two key features which can be used to classify a phage as a member of the T7 group or not.

2 Software and Methods.

Most programs used in this work are available at `http://www.lecb.ncifcrf.gov/toms/`. Promoter models were built with the programs **delila**, **alist**, **encode**, **rseq**, **dalvec** and **makelogo**. The programs **scan** and **lister** were used for genome scanning. **Genhis** and **genpic** were used to plot individual information distribution. Phylogenetic analyses were conducted with the programs **diffribl**, **neighbor** and **drawtree**.

3 Results and Discussion.

A total of seven promoter models were created for phages T7, ϕA1122, T3, YeO3-12, gh-1, K11 and SP6, and different combined models were also built. The ϕA1122 model is almost the same as the T7 model and the YeO3-12 model is almost the same as the T3 model. The other three models are diversified except for the DNA region from -7 to -4. Information analysis shows that excess information exists for all seven promoters, with excess ratios ranging from 1.52 to 1.85 (Table 1).

The seven models and the combined 93-sites model were used for genome scanning. When a model was used to scan its own genome or a closely related genome, a significant gap of individual information distribution was observed between real promoters and background. When the combined model was used to scan the seven phage genomes and other possible T7-like phages, ϕKMV, P60, VpV262, SIO1 and PaP3, 92 of the 94 promoters (including

[1]LECB, National Cancer Institute at Frederick, Frederick, MD. E-mail: `chenze@ncifcrf.gov`
[2]LECB, National Cancer Institute at Frederick, Frederick, MD. E-mail: `toms@ncifcrf.gov`

one T7 promoter in the T3 genome) above 12 bits can be picked up, while no sites above 12 bits can be picked up for the other phages, indicating that they may not belong to the T7 group. Also, a total of 17 host genomes and closely related genomes were scanned with each of the models, and 15 T7-like promoters were identified in 6 pathogens.

Phylogenetic analysis with the T7-like models show that the eight T7-like phages can be classified into five subgroups. Combining with some data from literature, we propose that T7-like phages can be classified as shown in Figure 1.

Phage	Number of sites	Rs (bits)	Range	Host	Genome size(Mb)	Rf (bits)	Rs-Rf (bits)	Rs/Rf
T7	17	34.9	-20,+5	*E. coli*	4.64	19.1	15.8	1.83
ϕA1122	17	35.2	-20,+5	*Y. pestis*	4.60	19.1	16.1	1.85
T3	14	33.6	-18,+5	*E. coli*	4.64	19.3	14.3	1.74
YeO3-12	15	34.2	-18,+5	*Y. enterocolitica*	4.62	19.2	15.0	1.78
gh-1	10	33.5	-18,+5	*P. putida*	6.18	20.2	13.3	1.65
SP6	11	32.6	-18,+5	*S. typhimurium*	4.86	19.8	12.8	1.65
K11	9	31.0	-17,+4	*Klebsiella*	6.0	20.4	10.6	1.52

Table 1: Excess information at T7-like promoters.

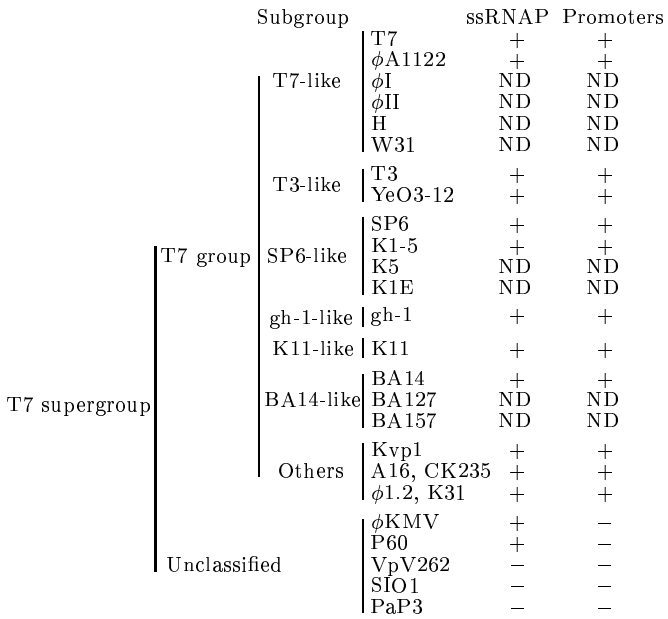


Figure 1: Classification of T7-like phages

References

[1] Schneider, T. D., Stormo, G. D., Gold, L. and Ehrenfeucht, A. 1986. Information content of binding sites on nucleotide sequences. *J Mol Biol*, 188: 415-431.

[2] Rohwer, F. and Edwards, R. 2002. The phage proteomic tree: a genome-based taxonomy for phage. *J Bacteriol*, 184:4529-4535.

L1. Analysis of Ataxin-2 and other Lsm domain proteins

Mario Albrecht[1], Markus Ralser[2], Hans Lehrach[2], Sylvia Krobitsch[2], Thomas Lengauer[1]

Keywords: neurodegenerative disorder, structural model, function prediction, RNA metabolism

Abstract

Sm and Sm-like proteins of the RNA-binding Lsm (Like Sm) domain family are generally involved in important processes of RNA metabolism including RNA modification, splicing, and degradation [1]. Lsm proteins can occur in large complexes and form cyclic hetero- or homo-oligomers [2]. While recent research has focused on the function and structure of small Sm and Sm-like protein (not more than about 150 residues), little is known about Lsm domain proteins with additional domains and sequence motifs.

Two very long Lsm domain proteins with both N- and C-terminal sequence extensions are the human ataxin-2 (1312 residues) of unknown cellular function and its yeast homolog PBP1 (PAB1-binding protein 1) [3,4]. A polyglutamine expansion in the N-terminal region of ataxin-2 is causative of the inherited neurodegenerative disorder spinocerebellar ataxia type 2 (SCA2) [5]. This disorder belongs to a heterogeneous group of trinucleotide repeat disorders including Huntington's disease and several other spinocerebellar ataxia types such as SCA1, SCA3, and SCA7 [6].

The C-terminal regions of the homologs ataxin-2 and PBP1 contain binding sites for A2BP (ataxin-2 binding protein) of unknown function or PABP (poly(A)-binding protein) [7,8], respectively. Both A2BP and PABP possess N-terminal RNA recognition motifs (RRMs) and may be evolutionarily and functionally related. The C-terminal tail of PBP1 has been observed in experiment to bind the C-terminal PABC domain of PABP and to regulate polyadenylation in mRNA splicing [4]. The absence of PBP1 leads to incomplete poly(A) tails, although the 3'-end of pre-mRNA is properly cleaved.

In order to obtain functional hypotheses on ataxin-2, we performed a comprehensive bioinformatics analysis on the structure and function of ataxin-2 and on the yeast interaction network of PBP1. Further experiments including a yeast-2-hybrid screen with ataxin-2 support our predictions of an important role of ataxin-2 in RNA metabolism. We also found at least five novel and evolutionarily conserved Lsm domain proteins with as yet uncharacterized C-terminal tails. Based on yeast interaction data, we could assign putative functions such as RNA methylation and mRNA degradation to some of the new Lsm domain proteins.

Our project web site provides further information: http://www.mpi-sb.mpg.de/~mario/medbioinf/

[1] Max-Planck-Institute for Informatics, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany. E-mail: {mario.albrecht, lengauer}@mpi-sb.mpg.de

[2] Max-Planck-Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany. E-mail: {ralser, lehrach, krobitsc}@molgen.mpg.de

Acknowledgements

Part of the research has been funded by the German Research Foundation (DFG) under contract number LE 491/14-1.

References

[1] He, W. and Parker, R. (2000) Functions of Lsm proteins in mRNA degradation and splicing. *Curr Opin Cell Biol*, 12:346-350.

[2] Mura, C., Phillips, M., Kozhukhovsky, A. and Eisenberg, D. (2003) Structure and assembly of an augmented Sm-like archaeal protein 14-mer. *Proc Natl Acad Sci USA*, 100:4539-4544.

[3] Neuwald, A.F. and Koonin, E.V. (1998) Ataxin-2, global regulators of bacterial gene expression, and spliceosomal snRNP proteins share a conserved domain. *J Mol Med*, 76:3-5.

[4] Mangus, D.A., Amrani, N. and Jacobson, A. (1998) Pbp1p, a factor interacting with Saccharomyces cerevisiae poly(A)-binding protein, regulates polyadenylation. *Mol Cell Biol*, 18:7383-7396.

[5] Pulst, S.M., Nechiporuk, A., Nechiporuk, T. *et al.* (1996) Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nat Genet*, 14:269-276.

[6] Cummings, C.J. and Zoghbi, H.Y. (2000) Trinucleotide repeats: mechanisms and pathophysiology. *Annu Rev Genomics Hum Genet*, 1:281-328.

[7] Shibata, H., Huynh, D.P. and Pulst, S.M. (2000) A novel protein with RNA-binding motifs interacts with ataxin-2. *Hum Mol Genet*, 9:1303-1313.

[8] Kozlov, G., Trempe, J.F., Khaleghpour, K. *et al.* (2001) Structure and function of the C-terminal PABC domain of human poly(A)-binding protein. *Proc Natl Acad Sci USA*, 98:4409-4413.

L2. Secondary structure prediction of RNA pairs

Mirela Andronescu and Anne Condon [1]

Keywords: RNA secondary structure determination, pairs of RNA, ribozymes

1 Introduction

In this work we present *PairFold*, an algorithm for *secondary structure prediction of pairs of RNA molecules*. Given two RNA sequences S_1 and S_2, *PairFold* finds the pseudoknot-free secondary structure R into which S_1 and S_2 can fold, such that R has the smallest free energy change under a standard thermodynamic model. The solution R may include base pairing between the two molecules, as well as base pairing within each molecule. *PairFold* is available online at `http://www.RNAsoft.ca` [1]. A simple extension of Zuker and Stiegler's algorithm [8], *PairFold* is currently being used by several research groups for pre-mRNA splicing, DNA word design and thermal statistics of RNA oligomers. Other motivations include predicting interactions between (1) a ribozyme and an RNA target [2, 4, 5, 6, 7]; (2) a probe or primer and a target RNA molecule [7]; (3) pairs of strands in biomolecular nanostructures; and (4) molecular tags in a polymer library. When testing our algorithm on pseudoknot-free RNA duplexes shorter than 120 nucleotides, we correctly predict approximately 90% of the base pairs on average. Our predictions are limited to pseudoknot-free structures, and do not take into consideration non-canonical base pairs or other tertiary interactions, all these being open questions for future work.

2 Algorithm

PairFold algorithm is a simple extension of Zuker and Steigler's algorithm [8]. Consider that we concatenate the two input sequences, S_1 and S_2, yielding S_1S_2, and that we denote the linkage location with b (b equals the last position in S_1). With some modifications, we can apply Zuker and Stiegler's algorithm [8] on S_1S_2. The possible structures are the same as for a single molecule (stacked pairs, hairpin loops, internal loops, multi-loops and external loops [3]), with the exception that they can be "broken" by the molecular link b between them. In these situations, the free energy is calculated in a way similar to the calculation of external loops, and an *intermolecular initiation penalty* is added. The recurrences underlying the algorithm are more complex than those of Zuker and Steigler [8], especially for the cases of "broken" multi-loops. However, the time and space complexities remain $O(n^3)$ and $O(n^2)$, respectively, where n is the sum of the input sequences lengths.

3 Results

Here we report on the accuracy of our method, when tested on several experimentally determined structures of RNA duplexes drawn from the literature. First, we evaluated *PairFold* on a combinatorial library of ribozymes, whose cleavage ability has been experimentally tested on a highly structured viral mRNA [7]. In this experiment [7], the library of ribozymes has been targeted to oligos, as well as to sequences of length 1.1 kb, and 15 of the targets have been cleaved. *PairFold* correctly predicted on average 97.8% of the base pairs

[1]Department of Computer Science, University of British Columbia, Vancouver, B.C., V6T 1Z4, Canada. E-mail: {andrones,condon}@cs.ubc.ca

when the oligos are targeted, while only four out of the 15 long targets have been predicted to fold into the typical structure. These results show that *PairFold* performs well on short sequences, but as the sequence length increases, the accuracy goes down.

Second, we report prediction accuracy results on a set of eight ribozyme-target duplexes drawn from the literature, where no structure has pseudokots, but some have non-canonical base pairs. Table 1 gives the references, sequence lengths, fraction of non-canonical base pairs, the number of correctly predicted base pairs out of the total number of base pairs, an accuracy measurement showing the fraction of base pairs that are correctly predicted, and the predicted free energy. *PairFold* gives 90% prediction accuracy on average. Note that only 72% of the base pairs of the third duplex are predicted correctly, and it is evident this is directly correlated to the 22% of non-canonical base pairs contained in this structure.

Reference	Len(S_1)	Len(S_2)	Fr(NC)	#bp/tbp	Accuracy	$\Delta G(kcal/mol)$
[5] Fig. 1d	14	55	0.00	15/18	0.83	-16.30
[5] Fig. 1c	14	65	0.00	19/19	1.00	-21.10
[6] Fig. 2	16	36	0.00	18/19	0.95	-22.60
[6] Fig. 3	17	38	0.06	17/18	0.94	-20.00
[4] Fig. 1b	21	92	0.22	33/46	0.72	-58.70
[2] Fig. 1 left	25	46	0.06	28/31	0.90	-45.80
[5] Fig. 1e	34	120	0.00	40/43	0.93	-58.20
[2] Fig. 1 right	70	100	0.03	63/69	0.91	-142.10

Table 1: Measurement of PairFold accuracy on a set of mRNA target - ribozyme pairs. About 90% of the base pairs are predicted correctly for these structures.

In conclusion, *PairFold* is a useful tool that can be used to predict secondary structures of RNA duplexes. It gives good accuracy for short pseudoknot-free strands. Future work includes improving our method to better predict more complex RNA structures.

References

[1] Andronescu, M., Aguirre-Hernandez, R., Condon, A. and Hoos, H. 2003. RNAsoft: a suite of RNA secondary structure prediction and design software tools. *Nucl. Acids. Res.* 31: 3416-3422.

[2] Hormes, R., Homann, M., Oelze, I., Marschall, P., Tabler, M., Eckstein, F. and Sczakiel, G. 1997. The subcellular localization and length of hammerhead ribozymes determine efficacy in human cells. *Nucl. Acids. Res.* 25: 769-775.

[3] Mathews, D. H., Sabina, J., Zuker, M. and Turner, D. H. 1999. Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure. *J. Mol. Biol.* 288: 911-940.

[4] Rupert, P. B. and Ferre-d'Amare, A. R. 2001. Crystal structure of a hairpin ribozyme - inhibitor complex with implications for catalysis. *Nature.* 410: 780-786.

[5] Schmidt, C., Welz, R. and Müller, S. 2000. RNA double cleavage by a hairpin-derived twin ribozyme. *Nucl. Acids. Res.* 28: 886-894.

[6] Vaish, N. K., Kore, A. R. and Eckstein, F. 1998. Recent developments in the hammerhead ribozyme field. *Nucl. Acids. Res.* 26: 5237-5242.

[7] Yu, Q., Pecchia, D. B., Kingsley, S. L., Heckman, J. E. and Burke, H. M. 1998. Cleavage of highly structured viral RNA molecules by combinatorial libraries of hairpin ribozymes. *J. Biol. Chem.* 273 (36): 23524-23533.

[8] Zuker, M. and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids. Res.* 9: 133-148.

L3. A comparison of transmembrane topologies greatly improves the comprehensive functional classification and identification of prokaryotic transmembrane proteins

Masafumi Arai[1,2,*], Kosuke Okumura[1], Masanobu Satake[2,3], Toshio Shimizu[1]

Keywords: transmembrane protein, transmembrane topology, functional classification and identification, prokaryotic species, genome-wide analysis

1 Introduction.

Many of proteins have not yet been annotated, with about one half of all proteome sequences being classified as functionally "putative" or "unknown" at best [1]. Such is the case, in particular, for transmembrane (TM) proteins, which account for as much as 20-30% of proteomes in individual species [2]. This is partly because TM protein sequences of known function are much less compared with soluble proteins. Recent studies, however, revealed that TM protein functions are closely correlated to their TM topologies, i.e., the number of TM segments (TMSs), positions of TMSs and N-tail location [3]. In this study, we propose a new method for the comprehensive classification and identification of TM protein functions by a clustering approach based on TM topology similarity. Prior to performing the clustering, we first investigate the current status of the functional identification of TM proteins based on sequence similarity.

2 Materials and Methods.

Out of 239,359 protein sequences of 87 sequenced prokaryotic (72 bacterial and 15 archaean) species in the GenBank database, 51,044 sequences were extracted as TM protein and their TM topologies (1-12 TMSs) (~21%) were predicted, by using SOSUI [4] (TM protein sequence prediction, ≥98% accuracy), DetecSig (signal peptide prediction and removal, 88% accuracy) [5] and ConPred (TM topology prediction, 69.6% and 83.3% accuracies for the number of TMSs & TMS positions and N-tail location, respectively) [6]. The procedures and the genome-wide analysis of TM topologies are described in detail in our previous paper [2].

The obtained TM protein sequences were classified into three categories, i.e., "known", "putative" and "unknown", according to the level of functional annotations in the SWISS-PROT database by homology search and sequence similarity comparison (details not shown here). Then, these annotated sequences were clustered by the single-linkage method based on TM topology similarity between sequences with the same number of TMSs. The TM topology similarity between sequences 1 and 2, $S_{1,2}$ is calculated as:

$$S_{1,2}\,(\%) = 100 \times \sum_{i=1}^{n+1} \min(l_{1,i},\, l_{2,i}) / \sum_{i=1}^{n+1} \max(l_{1,i},\, l_{2,i}),$$

[1] Department of Electronic and Information System Engineering, Faculty of Science and Technology, Hirosaki University, Hirosaki 036-8561, Japan

[2] Department of Developmental Biology and Neuroscience, Graduate School of Life Sciences, Tohoku University, Sendai 980-8577, Japan

[3] Department of Molecular Immunology, Institute of Development, Aging and Cancer, Tohoku University, Sendai 980-8575, Japan

[*] E-mail address: d01603@si.hirosaki-u.ac.jp

where, n, $l_{1,i}$ and $l_{2,i}$ are the number of TMSs, the length of the i-th loop in sequences 1 and 2, respectively, and $\min(l_{1,i}, l_{2,i})$ and $\max(l_{1,i}, l_{2,i})$ are the lengths of the shorter and longer loops in $l_{1,i}$ and $l_{2,i}$, respectively. The thresholds of TM topology similarity were determined so that the sequences included in the representative clusters (with ≥10 sequences) would occupy over 50% out of all the sequences.

3 Results and Discussion.

Using our clustering approach, the functionally classified and identified TM proteome sequences was increased from 24.3% to 60.9%. Almost half of them used to be "unknown" sequences before applying the clustering method. Additional analysis of the TM topologies in the clusters provided important information regarding TM protein functions that cannot be ascertained from sequence similarity.

Table 1: The results of the functional classification and identification of TM proteins (1-12 TMSs) from the 87 prokaryotic species based on sequence similarity and TM topology similarity.

TMSs	Total sequences	Based on sequence similarity				Based on TM topology similarity						
		Functionally annotated sequences			Identified[1]	Threshold TM topology similarity	In the representative clusters (with ≥10 sequences)					Classified and identified[2]
		"Known"	"Putative"	"Unknown"			Clusters	Sequences	Functionally annotated sequences			
									"Known"	"Putative"	"Unknown"	
1	14,590	584	2,191	11,815	19.0%	98%	74	7,337	332	1,295	5,710	58.2%
2	6,928	229	785	5,914	14.6%	92%	46	3,660	157	534	2,969	57.5%
3	4,059	105	602	3,352	17.4%	85%	32	2,281	75	426	1,780	61.3%
4	4,493	130	813	3,550	21.0%	84%	41	2,515	97	561	1,857	62.3%
5	3,643	131	923	2,589	28.9%	81%	33	1,923	76	625	1,222	62.5%
6	4,628	180	1,411	3,037	34.4%	85%	27	2,464	108	1,024	1,332	63.2%
7	2,076	82	515	1,479	28.8%	75%	25	1,075	44	330	701	62.5%
8	1,965	82	572	1,311	33.3%	73%	26	1,037	52	398	587	63.2%
9	2,015	100	704	1,211	39.9%	74%	30	1,033	67	501	465	63.0%
10	2,061	89	525	1,447	29.8%	74%	31	1,090	42	293	755	66.4%
11	2,045	94	625	1,326	35.2%	75%	23	1,087	62	400	625	65.7%
12	2,541	132	794	1,615	36.4%	82%	22	1,286	80	499	707	64.3%
Total	51,044	1,938	10,460	38,646	24.3%	-	410	26,788	1,192	6,886	18,710	60.9%

[1] "Known" and "putative" sequences are counted.
[2] Sequences in the representative clusters and "known" and "putative" sequences in other clusters (with 1-9 sequences) are included.

References

[1] Serres, M.H., Gopal, S., Nahum, L.A., Liang, P., Gaasterland, T. and Riley, M. 2001. A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.* 2:RESEARCH0035.

[2] Arai, M., Ikeda, M. and Shimizu, T. 2003. Comprehensive analysis of transmembrane topologies in prokaryotic genomes. *Gene* 304:77-86.

[3] Sugiyama, Y., Polulyakh, N. and Shimizu, T. 2003. Identification of transmembrane protein functions by binary topology patterns. *Protein Eng.* 16:479-488.

[4] Hirokawa, T., Boon-Chieng, S. and Mitaku, S. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14:378-379.

[5] Ikeda, M., Arai, M., Lao, D.M. and Shimizu, T. 2002. Transmembrane topology prediction methods: a re-assessment and improvement by a consensus method using a dataset of experimentally-characterized transmembrane topologies. *In Silico Biol.* 2:19-33.

[6] Lao, D.M. and Shimizu, T. 2001. Methods for detecting the signal peptide in transmembrane and globular proteins. In: Matsuda, H., Miyano, S., Takagi, T. and Wong, L., editors, *Proceedings of the Twelfth International Conference on Genome Informatics (GIW 2001)*, Universal Academy Press, Tokyo. pp. 340-342.

L4. Insertions and deletions in protein alignment

Charlotte Deane[1], Jacob Pedersen[2]. Gerton Lunter[1]

Keywords: Protein evolution, indels, gap penalties, protein structure prediction, Ooi number

1 Introduction.

For the purpose of aligning protein sequences, the amino acid substitution process has been extensively studied, but the arguably more important process of insertions and deletions (indels) has received less attention. Here we investigate the relationship between indel propensity and protein structural features. We find a particularly strong correlation with the Ooi coordination number, and a twice-reduced propensity for indels to occur in beta strands, as compared to alpha helices.

Amino acid sequence alignments are widely used for the analysis of protein structure, gene function prediction, inference of phylogeny and other aspects of bioinformatics. An alignment of homologous protein sequences gives a concise presentation of which residues share ancestry, and is used to infer structural or functional characteristics from known proteins. Using the observation that homologous sequences are similar, the formal problem, for two given sequences, is to insert gaps into the sequences in order to arrive at the ``best" alignment. The quantity measuring this goodness-of-fit, termed the score function, depends on the similarity of the aligned residues, and much effort has been expended on finding good so-called scoring matrices. Obviously, the goodness-of-fit must also include the number, position and length of the gaps themselves. However, this part of the score function has received far less attention.

We have undertaken a systematic study of what features in a protein sequence affect the propensity of amino acids to be deleted (or inserted). The results confirm a correlation with gap propensity for the conventional features used in current alignment algorithms, but reveal new and previously unused features exhibiting strong correlations, which could help improve structural alignment algorithms.

2 Structural properties and indel propensities.

We undertook a survey of all structural features collected in the HOMSTRAD database [2] to ascertain which have the clearest relationship to indel propensities. The features used were those calculated by Joy [1].

In current fold recognition and structure-sequence alignment software, two levels of gap penalty are usually employed, with the higher gap penalty imposed for the insertion of gaps into secondary structure elements. Our results illustrate that, as expected, indels are most abundant in coil regions, and are about twice as likely to occur there compared to alpha helices. However, the propensity for indels to occur in alpha helices is again more than twice as high as in beta sheets, suggesting that

[1] Bioinformatics group, Dept. of Statistics, Univ. of Oxford, 1 South Parks Road, Oxford OX1 3TG, United Kingdom. E-mail: {lunter,deane}@stats.ox.ac.uk

[2] Bioinformatics Research Center, Dept. of Genetics and Ecology, The Institute of Biological Sciences, Univ. of Aarhus, Building 550, Ny Munkegade, 8000 Aarhus C, Denmark. E-mail: jsp@daimi.au.dk

separation of the secondary structure class into alpha helix and beta sheet could improve alignments.

It is possible that the local nature of the hydrogen bonding in helices leads to their easier insertion/deletion. If a beta strand residue is deleted, a corresponding residue, perhaps distant in sequence, must be lost from its pairing strand in order not to disrupt the remaining residues of the strand, or alternative hydrogen bonding arrangements must be made. Thus, the non-local nature of beta strand bonding makes the loss of its component residues less likely.

The relationship between indel propensity and the Ooi number is the most pronounced of all structural features considered. The Ooi number [3] is a simple approximation to solvent accessibility or coordination number, and counts the number of other C-alpha atoms. Within a sphere of certain radius, in this case 14Å, of the given residue's own C-alpha. The correlation of the Ooi number to indel propensity is stronger than to more involved solvent accessibility measures; in fact, the logarithm of the indel propensity is very nearly linearly related to the Ooi number.

This linear relationship over a long range is surprising. A naive model to explain this, is to suppose that each C-alpha atom within the sphere of influence of the central atom has a fixed probability to have an important structural relationship to it, which would be disrupted when this residue was deleted, or displaced by an insertion. If one supposes these probabilities to be independent, the observed linear relationship follows.

3 Conclusion.

We investigated the relationship of indel propensities to various structural features, and identified the most informative of such features to use for alignment. Although structure dependent gap penalties are used in many alignment algorithms, this is the first systematic study of the data to inform the choice of structural parameters.

While we confirmed the utility of hydrophobicity in predicting gap propensity, other structural variables, such as the Ooi coordination number, revealed an even stronger correlation with indel propensities.

We found that beta strands are far less likely to contain a gap than alpha helices, a fact which has not been exploited in structural alignment algorithms, and which, we believe, could significantly improve such algorithms, with potentially important implications for structure and function prediction.

4 References and bibliography.

1. Mizuguchi, K., et al., *JOY: protein sequence-structure representation and analysis.* Bioinformatics, 1998. **14**(7): p. 617-23.
2. Mizuguchi, K., et al., *HOMSTRAD: a database of protein structure alignments for homologous families.* Protein Sci, 1998. **7**(11): p. 2469-71.
3. Nishikawa, K. and T. Ooi, *Radial locations of amino acid residues in a globular protein: correlation with the sequence.* J Biochem (Tokyo), 1986. **100**(4): p. 1043-7.

L6. Predicted Secondary Structure Slightly Enhances Ortholog Detection [1]

Ying Lin, [2] **John Case,** [3] **Hsing-Kuo Kenneth Pao,** [4] **Joan Burnside** [5]

Keywords: ortholog, secondary structure, machine learning.

1 Introduction.

Biological primary (nucleic acid or amino acid) sequences from different species are called *orthologs* just in case they evolved from a sequence of a common ancestor species and they have the same biological function. *Secondary structure* (SS) consists of a sequence of local protein foldings traditionally of three types: α-helix, β-sheet, and coil. *Homologs* are defined similarly to orthologs except they may or may not share the same biological function. In [1] it was found that the use of *experimentally determined* (secondary and) tertiary structure to increase primary alignment accuracy does did not aid homolog detection with profile HMMs. The present study complements [1]. We use *predicted* SS based machine learning attributes; employ decision tree induction with boosting, as implemented in C5.0 (http://www.rulequest.com); and are concerned with *ortholog* detection. We discuss below our results employing *PSIPRED*, arguably the best protein SS predictor (http://predictioncenter.llnl.gov/casp3/Casp3.html). *In our setting*, predicted secondary structure *is* slightly helpful for ortholog detection. The *relative* helpfulness of our set of SS based attributes is assessed compared with other, more standard primary sequence alignment based attribute sets. This comparative evaluation is done by: 1. cross-validation with C5.0 and 2. Area Under the Curve (AUC) from Receiver Operating Characteristic (ROC) analysis.

2 Background.

We employ *for classification* a positive training data set of 570 ortholog triples (from human, mouse, and chicken), curated as in [3], and available at http://polaris.cis.udel.edu:3000/ortho/ortholog_list/. For negative data, the selection criterion is that of [2] but with scoring scheme from [3] and restricted to cases where mouse and human are orthologous but chicken is not. This produced a total of 2214 triples. The classification task is to decide of a triple whether the chicken is orthologous to the two orthologous mammals. For each triple, 21 attribute values are computed re identity percentages based on alignments of primary sequences or predicted SS sequences. For primary sequence alignment, we employ Gotoh's algorithm, Smith-Waterman's (SW) algorithm, and Clustal W. Dynamic programming (DP) is employed for SS alignment

[1] This work is supported by USDA IFAFS Grant #01-04145.

[2] Department of Computer and Informatics Sciences, University of Delaware, Newark, DE 19716, U.S.A. E-mail: ylin@cis.udel.edu

[3] Department of Computer and Informatics Sciences, University of Delaware, Newark, DE 19716, U.S.A. E-mail: case@cis.udel.edu

[4] Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, 106, Taiwan. E-mail: pao@mail.ntust.edu.tw

[5] Department of Animal and Food Sciences, University of Delaware Newark, DE 19717, U.S.A. E-mail: joan@udel.edu

using a scoring scheme which empirically produces the best separation between positive and negative training data. Also, we employ one more attribute describing the biological function *class*, e.g., `defense` is its value for immune system proteins.

3 Results.

Cross-validation with C5.0.

We employ C5.0 with 10-tree boosting, 10-fold cross-validation, and 10 repeats. The experiments are conducted in two ways: only-one-in and at-most-one-out. [6] The first way is to compare each attribute set with one another by employing each such attribute set one at a time. The second evaluates the effect of leaving up to one attribute set out. The results are shown in Table 1. N.B. The Standard Error in the percent errors is $< 0.05\%$.

Attribute Set	**Missed in Testing: +/-(%)**
1A: only one in (& *class*)	
w. *CLUSTAL W*	56.2 / 37.0 (4.21%)
w. SM alignment	61.9 / 42.3 (4.71%)
w. Gotoh's alignment	68.3 / 39.3 (4.86%)
w. *SS DP*	**75.3 / 42.3 (5.31%)**
1B: at most one out	
w/o *CLUSTAL W*	46.6 / 12.7 (2.68%)
w/o *SM alignment*	44.2 / 12.1 (2.54%)
w/o Gotoh's alignment	45.3 / 10.8 (2.53%)
w/o *SS DP*	**43.9 / 10.6 (2.46%)**
w. all attribute sets	**43.7 / 8.9 (2.38%)**

Table 1: Our SS attribute set is the least useful of the attribute sets, *but* it is *better to use it with the others* (2.38% errors) than leave it out (2.46% errors)!

ROC analysis

We construct ROC curves for *single* individual attributes and compute *and compare* their AUC values. The best predicted SS based attribute has AUC = 0.901 which is less than that of about 50% of the non-SS based single attributes. However, when we combine two different single attributes a_i, a_j by the formula $b = a_i \cos\theta + a_j \sin\theta$, with θ chosen to maximize the AUC value of b, the best combination of attributes which includes at least one predicted SS based attribute has AUC = 0.968. This beats the overall best single attribute which has AUC = 0.952. *Again* we see that while our SS attribute set is not by itself spectacular, it *is* helpful when combined with other attribute sets.

References

[1] Sam Griffiths-Jones and Alex Bateman. 2002. The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. *Bioinformatics*. 18(9): 1243-1249.

[2] M. Ouyang and J. Case and J. Burnside. 2001. Divide and conquer machine learning for a genomics analogy problem (progress report). *Discovery Science : 4th International Conference*. pages: 290-303.

[3] M. Ouyang, J. Case, V. Tirunagaru and J. Burnside. 2003. 565 Triples of Chicken, Human, and Mouse Candidate Orthologs. *Journal of Molecular Evolution*. 57: 271-281.

[6] Exception: the *class* attribute is always in.

L7. Application of Variable Order Markov Models to Identifying CpG Islands

Zhenqiu Liu [1] **Dechang Chen** [2] **Jaques Reifman** [3]

Keywords: DNA sequence, CpG island, Markov chain, PST, identification, classification

1 Introduction

Identifying the location and function of human genes in a long sequence of genome is difficult due to lack of sufficient information about genes. Much research has been conducted in identifying CpG islands in DNA sequences using different models. First order Markov model and hidden Markov model (HMM) are among the most popular tools currently used [5]. Because of complexity of the real life sequence, the short memory assumption of the first order Markov chain usually is not satisfied. The HMM model, on the other hand, is more complex and can be slow for complex problems. It has been proved that HMMs can not be trained in polynomial time in the alphabet size. In addition, the algorithm of HMM can only be guaranteed to converge to a local minimum. Here we introduce one alternative model called variable order Markov chain. In the variable order Markov chain, the order of the Markov chain is not fixed [1], [4]. Variable order Markov models can be explained by probability suffix automata [2], [3], [6]. They are more succinct than higher order Markov chains and can be used to overcome the drawback that the size of Markov chain grows exponentially with its order. In addition, variable order Markov models are easy to compute and usually have high identification accuracies.

2 Discrimination with Markov models

A Markov model is fully defined by its states and state transition matrix. A variable order Markov chain is derived from the probability suffix tree. In a probability suffix tree, nodes are defined by the occurrence of symbols in DNA sequence. The nodal relationship is based on the parent being a suffix of its children. Each node also contains the probability of symbols that follow the given symbol in the sequence. An advantage of suffix tree model is that it is parsimonious. The tree order is kept to a minimum through excluding nodes that do not provide more stochastic information. The variable order Markov model is created entirely from the information provided in the probability suffix tree. The first step is to add all of the leaf nodes in the suffix tree to the Markov chain as states. The second step is to create a state that corresponds to the root of the suffix tree, and add any intermediate nodes from the root to leaf as additional new states. The third step is to find the transition probability for the given Markov states.

Markov models can be used to identify the CpG islands for DNA sequences. In order to do so, we need to train two Markov models separately: one for the CpG island, the other for the non-CpG island. For simplicity, denote CpG and non-CpG regions by '+' and '-',

[1]Bioinformatics Cell, TATRC, 110 North Market Street, Frederick, MD 21703, USA. E-mail: liu@bioanalysis.org

[2]Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, 4301 Jones Bridge Road, Bethesda, MD 20814, USA. E-mail: dchen@usuhs.mil

[3]Bioinformatics Cell, TATRC, U.S. Army Medical Research and Materiel Command, Ft. Detrick, MD, USA. E-mail: reifman@tatrc.org

respectively. Let a_{ij}^{+} and a_{ij}^{-} represent the transitional probability for the CpG island non-CpG island, respectively. Given a test DNA sequence x of length n, we can discriminate the CpG island from non-CpG island by using the following log-likelihood ratio:

$$R(x) = \log \frac{P(x|model+)}{P(x|model-)} = \sum_{i=1}^{n} \log \frac{a_{ij}^{+}}{a_{ij}^{-}} = \sum_{i=1}^{n} \log a_{ij}^{+} - \sum_{i=1}^{n} \log a_{ij}^{-}.$$

If $R(x) > C$, where C is a predetermined positive constant, the sequence is the CpG island.

3 Computational results

We have identified the CpG islands from hundreds of DNA sequences and found that simple models can usually lead to high prediction accuracies. Here we present an example. One test sequence is a human collagen alpha-1-IV and alpha-2-IV genes, exons 1-3 (HSCOLAA). This sequence has 2184 symbols. We split it into subsequences with 100 symbols. The step for the window to move forward is 1. HMM, the first order Markov chain (MC1), the third order Markov model (MC3), and variable order Markov model (VMC) are all applied to the same sequence. The testing results for the HSCOLAA sequence are given in Table 1, which indicates that the simple variable Markov model with 58 states has the best performance.

Table 1: **Identification with Different Order Markov Chains**

True Islands	MC1	HMM	MC3	VMC
49-877	15-858	1-862	7-833	46-887
953-1538	908-1460	908-1445	916-1444	919-1489
1765-2100	1910-2015	1859-2011	1858-2007	1769-2007

References

[1] Apostolico, A. and Bejerano, G. 2000. Optimal amnesic probabilistic automata or how to learn and classify proteins in linear time and space. *RECOMB 2000.*

[2] Bejerano, G. and Yona, G. 1999. Modeling protein families using probabilistic suffix trees. *RECOMB 99* 15-24.

[3] Kermorvant, C. and Dupont, P. 2002. Improved smoothing for probabilistic suffix trees seen as variable order Markov chains. *ECML'02*, pp. 1-27.

[4] Laird, P. and Saul, R. 1994. Discrete sequence prediction and its applications. *Machine Learning,* 15:43-68.

[5] Lio, P. and Vannucci, M. 2000. Finding Pathogenecity islands and gene transfer events in genome data. *Bioinformatics*, vol. 16, no. 10-2000, pp. 932-940.

[6] Ron, D., Singer, Y. and Tishby, N. 1996. The power of amnesia: learning probabilistic automata with variable memory length. *Machine Learning*, 25, 117-142.

L8. Using an RNA Secondary Structure Partition Function to Determine Confidence in Base Pairs Predicted by Free Energy Minimization

David H. Mathews[1]

Keywords: RNA Partition Function, RNA Secondary Structure, Statistical Mechanics

1 Introduction.

RNA plays many diverse roles in Biology, including catalyzing peptide bond formation [4, 10], catalyzing RNA splicing [2], localizing protein [11], and flagging development [5, 6]. New roles are being found for RNA, and the completion of whole genome projects provides the opportunity to find many new functional non-coding RNA sequences [3].

To understand the detailed mechanism of action of an RNA sequence, the structure of that RNA must be determined. Secondary structure, the sum of canonical base pairs, is usually determined by the comparative analysis of homologous sequences. In the absence of homologous sequences, free energy minimization by dynamic programming can be used to predict the structure of a single sequence with an average of 73% sensitivity for known pairs [7, 8]. This accuracy is sufficient to serve as a starting point for building an alignment for comparative sequence analysis or as an aid for designing RNA sequences, but improvements in the accuracy of base pair predictions would clearly be useful.

The predicted minimum free energy (MFE) structure provides a single best guess for the secondary structure, but it assumes that the secondary structure is at equilibrium, that there is a single conformation for the RNA, and that the thermodynamic parameters for evaluating conformation free energies are without error. One method to represent other possible or competing structures is to sample suboptimal secondary structures with free energies similar to the lowest free energy structure [12]. Another method to demonstrate a diversity of structures, pioneered by McCaskill, is to determine the pairing probabilities of all possible base pairs using a partition function calculated with dynamic programming [9].

2 Results.

A partition function calculation for RNA secondary structure is presented that uses a current set of nearest neighbor parameters for conformational free energy at 37 ºC [7, 8]. The calculation includes free energy increments for the coaxial stacking of helices, but remains $O(N^3)$ in time, where N is the number of nucleotides. The calculation is rapid, e.g. the base pairing probabilities for a 433 nucleotide *Tetrahymena* group I intron can be calculated in 19 seconds with a Pentium 4, 3.06 GHz processor.

For a diverse database of RNA sequences with known secondary structure [8], base pairs in the predicted minimum free energy structure that are predicted by the partition function to have high base pairing probability have a significantly higher positive predictive value for known base pairs.

[1] Center for Human Genetics and Molecular Pediatric Disease, Aab Institute of Biomedical Sciences, University of Rochester Medical Center, 601 Elmwood Avenue, Box 703, Rochester, NY 14642

For example, the average positive predictive value, 65.8% is increased to 90.7% when only base pairs with 99% or above probability are considered.

The recursions were written to allow constraints on base pairing determined by experiments, such as enzymatic cleavage, flavin mononucleotide cleavage, or chemical modification. The quality of base pair predictions are increased by the addition of experimentally determined constraints. For example, the percentage of highly probable pairs (greater than or equal to 95%) for the Dog SRP RNA increases from only 9.9% to 57.0% by including experimentally determined constraints in the calculation [1].

3 Summary.

The partition function calculation presented here does not replace the method of RNA secondary structure prediction by free energy minimization, but provides adjunct information that can be used to infer confidence in predicted base pairs. These data can be superimposed on predicted secondary structures using color annotation to quickly demonstrate high probability base pairs.

References

[1] Andreazzoli, M. and Gerbi, S.A. 1991. Changes in 7SL RNA conformation during the signal recognition particle cycle. *EMBO J*, 10: 767-777.

[2] Doudna, J. and Cech, T. 2002. The chemical repertoire of natural ribozymes. *Nature*, 418: 222-228.

[3] Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nature Reviews*, 2: 919-929.

[4] Hansen, J.L., Schmeing, T.M., Moore, P.B., and Steitz, T.A. 2002. Structural insights into peptide bond formation. *Proc. Natl. Acad. Sci. USA*, 99: 11670-11675.

[5] Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science*, 294: 853-858.

[6] Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans. Science*, 294: 858-862.

[7] Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M., and Turner, D.H. In preparation. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.

[8] Mathews, D.H., Sabina, J., Zuker, M., and Turner, D.H. 1999. Expanded sequence dependence of thermodynamic parameters provides improved prediction of RNA Secondary Structure. *J. Mol. Biol.*, 288: 911-940.

[9] McCaskill, J.S. 1990. The equilibrium partition function and base pair probabilities for RNA secondary structure. *Biopolymers*, 29: 1105-1119.

[10] Nissen, P., Hansen, J., Ban, N., Moore, P.B., and Steitz, T.A. 2000. The structural basis of ribosomal activity in peptide bond synthesis. *Science*, 289: 920-930.

[11] Walter, P. and Blobel, G. 1982. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature*, 299: 691-698.

[12] Zuker, M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science*, 244: 48-52.

L9. Annotation of 3D Protein Chains in PDB with GO terms via Structural Homology

Julia V. Ponomarenko[1], Philip E. Bourne[2], Ilya N. Shindyalov[3]

Keywords: GO (Gene Ontology) annotation, 3D protein structure, structural homology

1 Introduction

Annotation of biological molecules is one of the demanding tasks of a modern bioinformatics. Particularly, annotation through well-defined ontologies allowing data management and exchange between people and computers with minimal loss of information value. Gene Ontology™ (GO) [1] is one of widely used bio-ontologies to describe molecular function, biological process and cellular component of biological entities. GOA (GO Annotation) resources (*http://www.geneontology.org/*) provide protein annotation with GO terms performed by biological experts. Since this routine is expensive and time-consuming there is a need in automated assignment of GO term to proteins through extension of manual annotation. Currently, the extension of protein annotation is centered only on sequence homology with GO-annotated proteins [2]. However, many proteins with low sequence similarity possess a similar fold and related function [3]. In this work, we address the problem of extension of protein annotation with GO terms using homology in 3D structures.

2 Materials and Methods

Two major sources of data used during this study were as follows: (i) 3D protein structures from PDB [4]; (ii) GOA resource provided by the EBI [5] assigning GO terms to the protein chains from PDB. Structural alignments of protein chains all against all were calculated by CE algorithm [6]. To compare two chains four similarity parameters provided by the CE algorithm were used:

1. *Rmsd,* root mean square deviation between two structurally aligned chains.
2. *Z-score,* statistically founded score, it characterizes the significance of the alignment.
3. *Rnar,* ratio of the number of aligned residues to the length of the shortest chain.
4. *Rseq,* sequence identity calculated for the structurally aligned residues.

For two protein chains *A* and *B* with all calculated values (*Rmsd, Z-score, Rnar, Rseq*) and given thresholds we define *sequence-structural similarity criterion* (SSC):

$$SSC_{AB}=(Rmsd<Rmsd_{threshold}) \wedge (Z\text{-}score>Z\text{-}score_{threshold}) \wedge (Rnar>Rnar_{threshold}) \wedge (Rseq>Rseq_{threshold}) \quad (1)$$

$\wedge$ - denotes logical AND. Chains were clustered using *SSC* when for every two chains *i* and *j* in the cluster SSC_{ij} holds true. GO terms for the chains in a cluster with at least one chain with GO term(s) can be assigned based on their co-appearance in the same clusters (Fig. 1). The *specificity* of such assignment is defined as the ratio of the number of TP (true positive) chains i.e. chains with GO terms in so-called "positive clusters" (with non-contradictory annotations) to all chains with GO terms in all clusters. To take into account incompleteness of the initial GO annotation, we introduce three different ways to calculate specificity by providing different definitions of a "positive cluster". The performance of extended annotation is also measured by the *coverage* which is defined as the ratio of newly annotated chains to all chains with no annotation, i.e. with no GO terms assignments in the EBI annotation.

3. Results and Discussion

34,698 protein chains (excluding theoretical models and short chains) were taken from the PDB. 29,734 of them had GO annotations if from the EBI.

[1]San Diego Supercomputer Center, University of California, San Diego, U.S.A. & Institute of Cytology and Genetics, Novosibirsk, RUSSIA. E-mail: jpon@sdsc.edu
[2]San Diego Supercomputer Center, University of California, San Diego, U.S.A. E-mail: bourne@sdsc.edu
[3]San Diego Supercomputer Center, University of California, San Diego, U.S.A. E-mail: shindyal@sdsc.edu

PDB ID (chain)	Protein name (as in PDB)	Specie	GO term	GO term definition
1hjb (C,F)	Runt-related transcription factor 1; residues 60-182.	*Homo sapiens*	**3677**	**(F) DNA binding**
1io4 (C)	Runt-related transcription factor 1; runt domain.		**5524**	**(F) ATP binding**
1hjc (A,D)	Runt-related transcription factor 1; residues 60-182.	*Mus musculus*	**5634**	**(C) nucleus**
1ean (A)	Runt-related transcription factor 1; runt domain residues 46-185		**6355**	**(P) regulation of transcription, DNA-dependent**
1eao (A,B) 1eaq (A,B)	Runt-related transcription factor 1; runt domain residues 36-185			
1e50 (A,C, E,G,Q,R)	Core-binding factor alpha subunit; runt domain residues 50-183	*Homo sapiens*	3700	(F) transcription factor activity
1cmo (A)	Polyomavirus enhancer binding protein 2; runt domain.		7275	(P) development
1co1 (A)	Core binding factor alpha; runt domain.		8151	(P) cell growth and/or maintenance
1ljm (A,B)	Runx1 transcription factor; runt domain.		**3677**	**(F) DNA binding**
			5524	**(F) ATP binding**
			5634	**(C) nucleus**
			6355	**(P) regulation of transcription, DNA-dependent**
1h9d (A,C)	*Core-binding factor alpha subunit1; runt domain*	*Homo sapiens*	*no GO terms*	

Fig. 1. The example of a cluster. GO annotation are provided according to the EBI assignment. (P), biological process, (F), molecular function, (C), cellular component. GO terms common for all chains which have GO terms are shown in bold. Other three GO terms could be assigned as new added GO terms to chains in the upper gray box. Chains 1h9dA and 1h9dC could be newly annotated by if assigning them seven GO terms (white box).

The performance of the approach extending GO annotation was studied with respect of different choices for the threshold values (in Eq. 1). Fig 2 demonstrates the performance for threshold values when the highest specificity-3 value (99.9%) could be achieved (coverage of 47.9%, Rseq ≥ 90%). This allows to extend annotation for 2,371 protein chains from the PDB [4] (amongst 4,964 chains with no annotation) by assigning 13,519 new "GO term – protein chain" associations. Also, 3,962 new "GO term – chain" associations were added to the existing annotation of 1,449 chains previously annotated by EBI [5]. See Fig.1 for the example.

The effect of similarity definition and impact of "superfolds" (folds with multiple functions) are considered as well as comparison with SCOP for definition levels in protein hierarchy involved in our annotation process have been done.

A number of chains with contradictory annotations were revealed. Some of them can be explained by incompleteness of GO annotation, some may be the case a miss-annotation and require further attention.

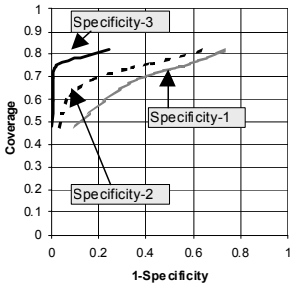


Fig. 2. Performance of the approach extending EBI GO annotation using pair-wise structural similarity for following threshold values of structural similarity parameters, Rmsd ≤ 5.0Å, Z-score ≥ 3.8, Rnar ≥ 70%.

The results of GO annotation of PDB protein chains are available at http://spdc.sdsc.edu/.

4 References

[1] Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.,* **11(8)**: 1425-1433.

[2] Zehetner G. (2003) OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Res.,* **31(13)**: 3799-3803.

[3] Shindyalov I.N., Bourne P.E. (2000) An alternative view of protein fold space. *Proteins,* **38(3)**: 247-260.

[4] Berman H.M., et al. (2000) The Protein Data Bank. *Nucleic Acid Res.,* **28:** 235-242.

[5] Camon E., Magrane M., Barrell D., et al. (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.,* **13(4)**: 662-672.

[6] Shindyalov I.N., Bourne P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) optimal path. *Protein Eng.,* **11**: 739-747.

L10. A Target Selection Informatics Resource for Structural Genomics

Ana Rodrigues[1], Guy G. Dodson[12], Roderick E. Hubbard[13]

Keywords: structural genomics, target selection, informatics resource

A large number of structural genomics programs have been established worldwide with the common aim of large-scale, high-throughput protein structure determination. Due to the considerable challenges posed by the experimental methods of structural determination (primarily X-ray crystallography and nuclear magnetic resonance spectroscopy), it is important to select and prioritize candidate molecules that will maximize the information gained from each new structure.

We are currently developing an informatics resource capable of performing target selection through the implementation of a number of sequence analysis protocols. This framework aims to facilitate the selection and prioritization of candidate proteins for structural determination, by enabling structural biologists to select targets from their genomic sequences of interest, and according to their own research needs.

The work is being developed on three distinct levels. The first focuses on the identification of the kinds of annotations that can be devised for a nucleotide or protein sequence and the assessment of the usefulness of such information for structural genomics. The second involves the establishment of an informatics framework to support these calculations and data through a relational database. The third concerns the refinement of the system and its application to real structural genomics programs.

Preliminary results, driven by the Structural Biology Laboratory's participation in structural and functional genomics projects (such as the EU funded Structural Proteomics IN Europe project, and the Wellcome Trust funded MalariaVac consortium) indicate that the resource will prove useful in performing target selection on a genomics scale [1, 2].

References

[1] Rodrigues, A. and Hubbard, R.E. 2003. Making decisions for structural genomics. *Briefings in Bioinformatics* 4(2):150-167.

[2] URL: http://www.ysbl.york.ac.uk/~rodrigues/targets.html

[1] Structural Biology Laboratory, University of York, York YO10 5YW, United Kingdom. E-mail: rodrigues@ysbl.york.ac.uk

[2] National Institute for Medical Research, Mill Hill, London NW7 1AA, United Kingdom. E-mail: ggd@ysbl.york.ac.uk

[3] Vernalis Ltd, Granta Park, Abington, Cambridge CB1 6GB, United Kingdom. E-mail: rod@ysbl.york.ac.uk

L11. Structural Kinomics – Structural Genomics of the Human Kinome

Kenneth D Schwinn*, Christopher R Hansen*, Ian M Miller, Shane Atwell, Sean G Buchanan, and J Michael Sauder

SGX (Structural GenomiX, Inc.)
10505 Roselle St, San Diego, CA 92121

Keywords: kinase, sequence analysis, structure alignment, Kinator, RoKI

1 Structural Kinomics

We describe bioinformatics tools and databases that facilitate global analysis of public and proprietary sequence/structure data on all human kinases (the "kinome" [1]). Kinator™ is a program that rapidly analyzes kinase sequences, while our database, RoKI™, stores a wealth of kinase-specific information for easy retrieval and mining. Alignator™ performs structure alignments of public and proprietary structures, highlights functionally important residues, and identifies residues that interact with natural ligands or inhibitors. These tools simplify selectivity analysis for kinome-based drug discovery.

2 RoKI™

The Repository of Kinase Information (RoKI™) is an intuitive web-based navigation system for rapid searching and browsing of information related to the sequence and structure of all known human kinases; it also serves as an interface to a comprehensive LIMS (Laboratory Information Management System), tracking experimental methods and results regarding cloning, protein purification, mass spectrometric analysis, biochemical assays, (co-)crystallization, and structure determination.

RoKI can be accessed via keyword search (using gene names/descriptions, HUGO aliases, LocusLink IDs, PDB IDs, or LIMS identifiers) or browsing (both textual and graphical). A detailed analysis page is available for each kinase, including literature references, RefSeq data, Pfam domain identification, related 3D protein structures, and annotation of key residues, secondary structure elements, and subdomains. These annotations are mapped onto the 3D structure in a web-based molecular graphics viewer, automatically identifying the activation loop, catalytic residues, hinge region, bound ligands, etc.

Experimental data (using user-selected criteria such as cloning success, purification yields, crystal quality, number of structures, biochemical assay data) can be mapped onto the kinome tree, providing an instant snapshot of the latest results for every kinase. Kinase similarity can be compared graphically by choosing a sequence identity threshold (e.g., 90%, 70%, 50%); all kinases that share at least that level of similarity are linked together on the kinome tree.

3 Kinator™

Our program, Kinator™, identifies important structural and functional residues within kinase sequence(s). Kinator uses a Hidden Markov Model (HMM) to identify the catalytic domain and to

pinpoint critical residues. The program automatically distinguishes serine/threonine and tyrosine kinases and highlights the P-loop, ATP-binding site, catalytic loop, activation loop, hinge region, and gatekeeper residue, as well as the 12 subdomains defined by Hanks & Hunter [2]. It also labels all secondary structure elements according to their canonical nomenclature (e.g., "helix C"). It is simple to determine whether a kinase has a large insertion, unusual termini, or has variant residues that are likely to affect the function.

Kinator can be used to perform comparisons of certain residues or regions within all kinases. For example, using these tools it is easy to tabulate the length of all activation loops, determine the identity of all gatekeeper residues, or calculate the sequence identity between a target and the ATP-binding site residues of all other kinases. Furthermore, our interactive web-based structure alignment program, Alignator™, is able to superimpose multiple structures and highlight residues of functional interest.

Figure 1 shows some of the views available from RoKI, Kinator, and Alignator.

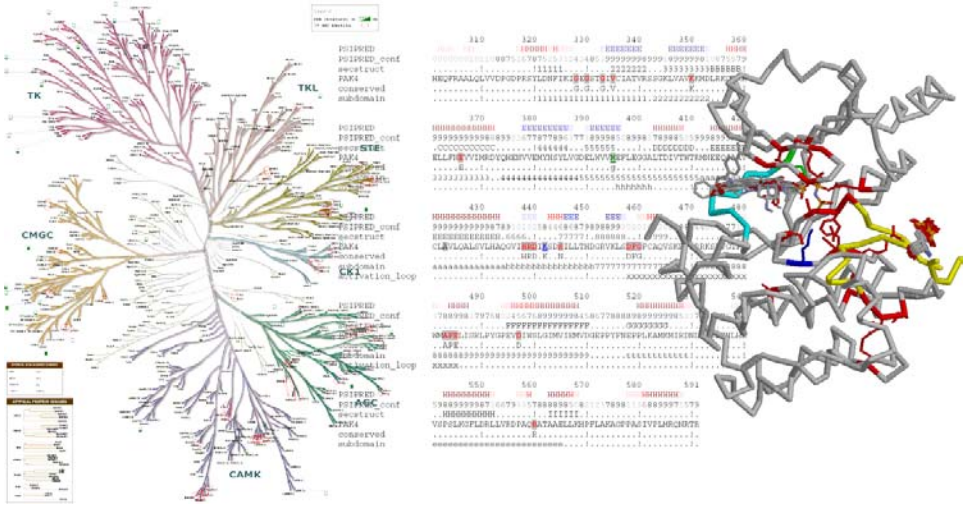


Figure 1: PDB structures (green squares) mapped onto the kinome tree [1] with 90% identity halos (red ovals) using RoKI™ (left), kinase sequence annotation using Kinator™ (middle), and Alignator output of ~30 co-complex structures with Kinator-based color-coding (right).

References

[1] Manning, G., et al 2002. The protein kinase complement of the human genome. *Science* 298:1912-1934.

[2] Hanks, S.K. and Hunter, T. 1995. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *Faseb J* 9:576-596.

*These authors contributed equally to this work.

L12. The SWISS-MODEL Repository of annotated 3-dimensional protein structure homology models

Jürgen Kopp[1], Torsten Schwede[1]

Keywords: large-scale homology modelling, protein model database, annotated protein models.

1 Introduction.

Three-dimensional protein structures are key to a detailed understanding of the molecular basis of protein function. Combining sequence information with 3D structure gives invaluable insights for the development of effective rational strategies for experiments such as site directed mutagenesis, studies of disease related mutations, or the structure based design of specific inhibitors. Techniques for experimental structure solution have made great progress in recent years. However, experimental structure determination is still a time-consuming process without guaranteed success. This is reflected by the fact that the number of structurally characterized proteins is about two orders of magnitude smaller than the number of known protein sequences in the UniProt database [1], which holds more than one million entries. Thus, no experimental structural information is available for the vast majority of protein sequences. Therefore, theoretical methods for protein structure prediction aiming to bridge this structure knowledge gap have gained much interest in recent years. As shown during the biannual CASP experiments, homology modeling is the only computational approach, that can generate reliable three-dimensional models for a protein.

2 Homology Modeling and SWISS-MODEL.

If a target protein shares significant amino acid sequence similarity to at least one experimentally solved three-dimensional structure (template), homology or comparative modeling can be applied to construct a three-dimensional model for the new protein. Homology modeling of protein structures consists of four steps: template selection, target-template alignment, model building, and model evaluation. The huge and constantly growing number of structurally uncharacterized protein sequences together with the increasing number of available template structures motivated the development of automated, stable and reliable modeling methods. Storing and organizing results of large-scale automated modeling in a database gives instant and queryable access to pre-computed and annotated comparative models through a model repository. This also helps to enrich other database projects with structural information, e.g. sequence knowledge bases like UniProt, or databases dedicated to specific organisms, protein families, or cellular functions. Here we describe the SWISS-MODEL Repository [2,3], a database of annotated three-dimensional protein models created by the SWISS-MODEL server pipeline [4,5].

3 Accessing the Repository.

The aim of the SWISS-MODEL Repository is to provide access to an up-to-date collection of annotated models generated by automated homology modelling, bridging the gap between sequence and structure databases. All models in the repository are publicly accessible via our interactive

[1] Biozentrum & Swiss Institute of Bioinformatics, University Basel, CH 4056 Basel, Switzerland, Juergen.Kopp@unibas.ch, Torsten.Schwede@unibas.ch

website at **http://swissmodel.expasy.org/repository.** A graphical "model navigator" provides an overview of the models that have been generated for a selected sequence, allowing fast and easy navigation for the different regions in the protein, for which three-dimensional models are available. The "model info" section contains information about the template structure and target-template sequence alignment on which the modelling has been based. The interactive display allows expanding detailed views of the target-template sequence alignment, the force field based assessment of the model, and the modelling log files. Model coordinates can be downloaded or displayed directly from within the web browser. Moreover, complete DeepView (Swiss-PdbViewer) [4] modeling projects can be exported. These project files contain the final model superposed to the template structures, which allows to manually adjusting the alignment and re-submitting to the SWISS-MODEL server for further model building. The repository can be queried for protein or gene name, Swiss-Prot accession codes, protein description key words, E.C. numbers, and organism names. The search interface allows combining all these different descriptors to complex queries, e.g. searching the repository for all models of a certain enzyme in several organisms.

4 Repository content and update.

The SWISS-MODEL Repository has been implemented using relational database technology. During the modeling process it communicates with the SWISS-MODEL server pipeline and keeps track of the workflow for individual target sequences. The models in the SWISS-MODEL Repository are computed by a modified version of the SWISS-MODEL server pipeline. The quality of each model is assessed using a partial Gromos96 force field implementation, and the empirical Anolea mean force potential in order to select reliable models to be entered into the database. Functional annotation is done by mapping InterPro [6] descriptors to individual models.

As of January 2004, the Swiss-Model repository contained 362,904 models for 324,571 different UniProt sequence entries. The length of the models varies from 45 to 1,524 residues with an average model size of 205. The Repository is updated regularly to take into account new sequences, modifications of existing sequence entries, and new template structures released by the PDB [7] that might allow the construction of models for previously un-modeled proteins, or might provide a better template for already existing model entries. Also, fundamental changes and improvements of the modeling pipeline initiate a new update cycle.

References and Links

[1] Apweiler, R., Bairoch, A., *et al.* 2004. UniProt: the Universal Protein Knowledgebase *Nucleic Acids Res.* 32: D115-D119.

[7] Berman, H.M., Westbrook, J., *et al.* 2000. The Protein Data Bank. *Nucleic Acids Research*, 28:235-242.

[4] Guex, N. and Peitsch, M.C. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714-2723.

[2] Kopp, J. and Schwede, T. 2004. The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Research* 32:D230-D234.

[6] Mulder, N.J., Apweiler, R., *et al.* 2003. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research* 31:315-318.

[5] Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. 2003. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Research* 31:3381-3385.

[3] SWISS-MODEL Repository **http://swissmodel.expasy.org/repository/**

L13. Quantifying Structure-Function Uncertainty: A Graph Theoretical Exploration Into the Origins and Limitations of Protein Annotation

Boris E Shakhnovich[1] J. Max Harvey.
[1]Bioinformatics Program, Boston University, Boston MA, 02215

Keywords: Protein Domain, Annotation, Graph Theory, Database, Structure-Function, Evolution.

Introduction.

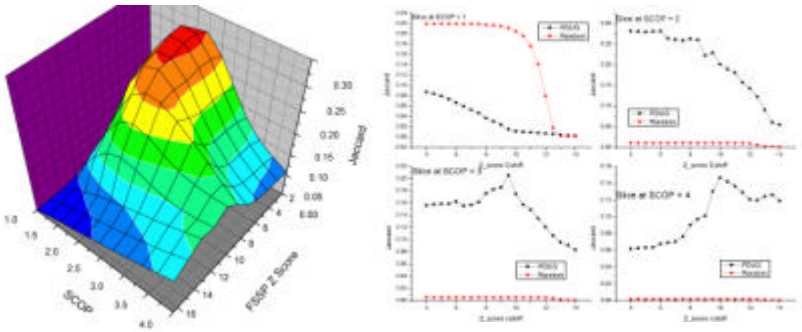
Since the advent of investigations into structural genomics, research has focused on correctly identifying domain boundaries, as well as domain similarities and differences in the context of their evolutionary relationships. As the science of structural genomics ramps up adding more and more information into the databanks, questions about the accuracy and completeness of our classification and annotation systems appear on the forefront of this research. A central question of paramount importance is how structural similarity relates to functional similarity. In this paper we begin to rigorously and quantitatively answer these questions by first exploring the consensus between the most common protein domain structure annotation databases CATH, SCOP and FSSP. Each of these databases explores the evolutionary relationships between protein domains using a combination of automatic and manual, structural and functional, continuous and discrete similarity measures. In order to thoroughly examine the issue of consensus, we build a generalized graph out of each of these databases and hierarchically cluster these graphs at interval thresholds. We then employ a distance measure to find regions of greatest overlap. Using this procedure we were able not only to enumerate the level of consensus between the different annotation systems, but also to define the graph-theoretical origins behind the annotation schema of Class, Family and Superfamily by observing that the same thresholds that define the best consensus regions between FSSP, SCOP and CATH correspond to distinct, non-random phase-transitions in the structure comparison graph itself. To investigate the correspondence in divergence between structure and function further, we introduce a measure of functional entropy that calculates divergence in function space. First, we use this measure to calculate the general correlation between structural homology and functional proximity. We extend this analysis further by quantitatively calculating the average amount of functional information gained from our understanding of structural distance and the corollary inherent uncertainty that represents the theoretical limit of our ability to infer function from structural similarity. Finally we show how our measure of functional "entropy" translates into a more intuitive concept of functional annotation into similarity EC classes.

Databases as graphs

Through graph-morphing procedures for SCOP[1], CATH[2] and FSSP[3] we end up with three *weighted* graphs, one for each database. The nodes in each graph are the protein domains and the edges are the relationships defined by distances or proximity from each database. We proceed to cluster these graphs at regular interval cutoffs. For example, for FSSP we build a graph at each threshold from **Z**=2 to 16 with step .5. In order to do this, we pick a cutoff and keep all edges that are larger than this cutoff[4].

Compare graphs.

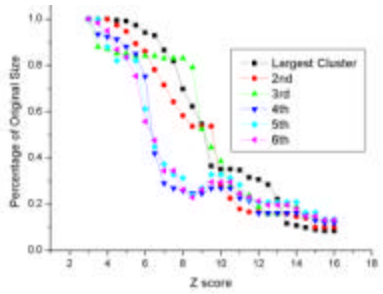
After TP, FP, TN and FN quantities have been defined, the distance measure between two graphs is merely a calculation of how many true positives the two graphs share with respect to false negatives and false positives. This measure is meant to calculate the level of agreement between the two graphs with respect to how many domain pairs they classify in the same cluster. Four Slices of the 3-D graph depicted in The cusps and maxima are easily discernable from these slices. At SCOP cutoff 1 Jaccard is actually smaller than the random control indicating that this level of annotation is probably not indicative of real evolutionary homology and may not indicate meaningful annotation. At SCOP Cutoff 2,3,4 the Jaccard distance between FSSP and SCOP is many thousands



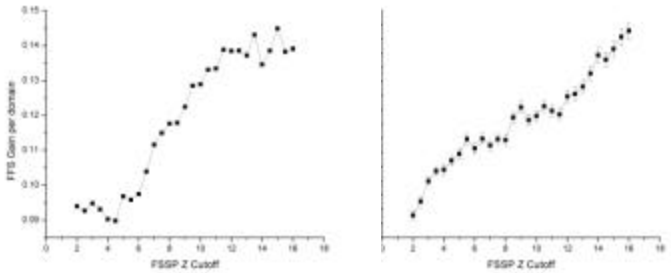
standard deviations away from random. At SCOP cutoff 2 (Fold level) the cusp occurs at Z=6, at SCOP cutoff 3 (Superfamily level) the maximum occurs at Z = 9 and SCOP cutoff 4 (Family level) the maximum occurs at Z = 10-11.

Phase transitions

The size of the largest clusters in FSSP graph plotted against the similarity cutoff threshold at which the graph is clustered. It is worth noting that the size of the largest cluster in the random graph is larger than the largest cluster in FSSP until the end of the phase transition at Z = 12. This is due to the power-law nature of the FSSP graph[4]. The size of the first six largest clusters plotted together as percentage of their original size. The computation was done by ordering the sizes of the clusters at each cutoff and plotting the largest six. The largest six clusters account for vast majority of the domains that are not orphans (singletons). It is worth observing that all the phase transitions occur between Z=6 and Z=9. The behavior of the size of the largest cluster (Fig. 6) and its difference with random bears a striking resemblance to the maxima we just observed on the distance landscapes between the three databases (Figs. 3,4). We can see that there are two very pronounced phase transitions in the size of the largest cluster. The first is from FSSP Z=6 to Z= 9 and the second is from Z=10 to Z=14. These represent the starting and ending points where the largest cluster "suddenly" breaks up into much smaller clusters the largest of which is almost fifty percent of the "parent". The size of the largest cluster in the random graph is always much larger than the size of the largest cluster in the real graph up until Z > 12. Because of this we will argue that the third and final non-random transition occurs at around Z = 11. The behavior of the other clusters closely mirrors that of the largest cluster thus showing that the phase transition is not just the function of the major superfolds but of the majority of the PDUG graph. It is interesting that the first three largest clusters transition at around Z=9 while the smaller three transition closer to Z = 6.



Function Uncertainty



a. b.

a,b. The FFS[5] gain per domain with respect to structural similarity threshold. FFS of each cluster is compared to that expected by random for a cluster that size and added to the gain at that threshold (Eqs 6, 7). The final FFS gain is normalized by the number of domains annotated in the graph. The majority of the functional information is gained from Z = 6 to Z = 11, before and after those thresholds the information content obtained from structural comparison plateaus. Thus we can quantify the amount of function information gained by correctly annotating a domain to its Fold as .095 bit per domain while correctly identifying the Superfamily yields around .15 bits per domain of functional information. The intrinsic uncertainty with which we can expect annotation of function at a given structural similarity. For example, at Z = 6 (Fold level) on average the domain function cannot be annotated to be more precise than 1.6 bits per level on the GO tree. Note that there are two plateaus where the FFS does not significantly change with respect to Z score: the first starting from Z = 5 to Z = 8 and the other starting from Z =9 all the way to Z = 11 showing an intrinsic correlation between structure and function at the Fold and Superfamily Level of annotation. This once again confirms the theoretical origins of this annotation by showing the conservation of function at those levels of structural comparison.

References:

1. Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res* 30, 264-7.
2. Orengo, C. A., Pearl, F. M. & Thornton, J. M. (2003). The CATH domain structure database. *Methods Biochem Anal* 44, 249-71.
3. Dietmann, S., Park, J., Notredame, C., Heger, A., Lappe, M. & Holm, L. (2001). A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res* 29, 55-7.
4. Dokholyan, N. V., Shakhnovich, B. & Shakhnovich, E. I. (2002). Expanding protein universe and its origin from the biological Big Bang. *Proc Natl Acad Sci U S A* 99, 14132-6.
5. Shakhnovich, B. E., Dokholyan, N. V., DeLisi, C. & Shakhnovich, E. I. (2003). Functional fingerprints of folds: evidence for correlated structure-function evolution. *J Mol Biol* 326, 1-9.

L14. Use of Limited Suboptimal Alignment in Homology Modeling

Christopher L. Tang[1], Donald S. Petrey[1], Marc Fasnacht[1], Mickey Kosloff[1], Emil Alexov[1,2], Barry Honig[1,2]

Keywords: suboptimal alignment, homology modeling, protein energetics

1 Introduction.

Improving the accuracy of alignment between a query sequence and a template protein of known structure remains a significant challenge in homology modeling [1]. Although maximal alignments perform sufficiently well to detect similarity relationships between proteins in database search applications, often the correct alignment from a structural standpoint deviates from the maximal alignment. A possible reason for this is that linear alignment schemes may not contain sufficient information to properly align a query sequence onto a template – tertiary information must be evaluated, for instance, by using physical energy terms to discriminate between native and non-native alignments [2]. Such a strategy requires a method for sampling and generating suboptimal alignments – alignments which are not maximal. In recent years, "consensus methods", which look at sequence alignments from programs written by many different laboratories, have been proposed as one solution to sampling alignments (e.g. [3]). However, such methods may not sample sufficiently and may miss the correct alignment if the alignment is "counterintuitive" to the alignment programs [4]. An alternative is near optimal alignment, which may be used to generate alignments whose scores are within a preset distance of the maximal [5]. However, the space of near optimal alignments is very large and often inconsequential differences in alignment will be reported [1]. Other variations of alignment sampling methods include iterative masking [6], [7] and parametric sampling [8]. However, these do not necessarily guarantee *k*-best solutions for any given set of sequences, and there is no unique set of parameters.

True *k*-best algorithms have historically been avoided for their computational expense [4] but in the advent of increasing computer power it is possible to revisit these. In addition, it is possible to supply detailed limits over what regions of a protein should be sampled thoroughly by suboptimal alignment; for instance, it may be desired to concentrate on how secondary structure elements should be aligned between query and template. Here we present a method for generating limited *k*-best suboptimal alignments, which allows us to sample suboptimal alignments deeply in regions of interest and hence more meaningfully for homology modeling. We build models for these alignments and seek to determine whether we can find closer-to-native alignments in our pool of suboptimal alignments.

References

[1] Department of Biochemistry and Molecular Biophysics, Columbia University, New York, New York, 10032. E-mail: clt47@columbia.edu

[2] Howard Hughes Medical Institute, Columbia University, New York, New York, 10032. E-mail: bh6@columbia.edu

[3] Ginalski, K., Rychlewski, L. 2003. Protein structure prediction of CASP5 comparative modeling and fold recognition targets using consensus alignment approach and 3D assessment. *Protein Structure, Function &Genetics*. 53:410-417.

[1] Jaroszewski, L., Li, W. and Godzik, A. 2002. In search for more accurate alignments in the twilight zone. *Protein Science* 11:1702-1713.

[2] Petrey, D.S. and Honig, B. 2000. Free energy determinants of tertiary structure and the evaluation of protein models. *Protein Science*. 9:2181-2191.

[4] Petrey D.S., Xiang, Z., Tang, C.L., Xie, L., Gimpelev, M., Mitros, T., Soto, C.S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I., Alexov, A. and Honig, B. 2003. Using multiple structure alignments, fast homology building, and energetic analysis in fold recognition and homology modeling. *Protein Structure, Function &Genetics*. 53:430-435.

[6] Saqi, M.A. and Sternberg, M.J. 1991. A simple method to generate non-trivial alternate alignments of protein sequences. *Journal of Molecular Biology* 219:727-732.

[5] Waterman, M.S. 1983. Sequence alignments in the neighborhood of the optimum with general application to dynamic programming. *Proceedings of the National Academy of Sciences USA*. 80:3123-3124.

[7] Waterman, M.S. and Eggert, M. 1987. A new algorithm for best subsequence alignments with application to tRNA-rRNA comparisons. *Journal of Molecular Biology*. 197:723-728.

[8] Waterman, M.S., Eggert, M. and Lander, E. 1992. Parametric sequence comparisons. *Proceedings of the National Academy of Sciences USA*. 89: 6090-6093.

L15. Contact Map Prediction via Maximum Entropy

Degui Zhi[1], **Charles Elkan**[2]

Keywords: machine learning, protein structure prediction

1 Introduction

Contact map prediction is an important intermediate goal for protein structure prediction. Previous research has applied neural networks [1, 3] and hidden Markov models [3, 4], but the complexity of those models is hard to justify. Our work applies the maximum entropy method, a simple statistical model with elementary features, namely primary sequence and (predicted) secondary structure, to predict contacts between off-diagonal residues in a protein. When trained on experimentally solved structures in PDB, the model shows a convincing ability to separate contacting from non-contacting residue pairs. When used to predict contact maps for novel proteins of length greater than 170, the performance of the model is comparable to the best result reported previously.

2 Methods and Results

Maximum entropy models use a set of predefined features. A *feature* f_j is a binary function that describes a characteristic of a data point. A maximum entropy model has the *log-linear* form $P(y|x) = \frac{1}{Z(x)} \exp(\sum_j \lambda_j f_j(x, y))$, where λ_j is a weight for the feature f_j, and $Z(x)$ is a normalizing factor that ensures a proper probability. The weights λ_j can be estimated using iterative scaling algorithms [2].

Given a pair of residues at position i and j of a protein sequence, our maximum entropy predictor uses three set of features for contact map prediction:

1. atomic features: amino acids a_i and a_j, secondary structures s_i and a_j, and the sequence separation $d = |i - j|$;
2. pairs (i.e. cross products) of atomic features: $s_i \times s_i$ and $a_i \times a_j$;
3. d paired with other features: $d \times s_i$ and $d \times a_i$, $d \times s_i \times s_j$, and $d \times a_i \times a_j$.

The contacts between residues in a protein are not independent. For example, between two beta strands, if $(i-1, j-1)$ and $(i+1, j+1)$ are in contact, then (i, j) is very likely to be in contact also. We have designed a simple smoothing filter that capture dependencies between neighboring contacts.

Increased modeling power is seen as more features are used (Figure 1). Predictions are dramatically improved after filtering, as filtered predictions capture clumps in the map corresponding to contacts between secondary structures.

Table 1 reports the performance of the maximum entropy method using all three feature sets. The model is trained on all the positive examples (contacts) from the 92 PDB chains and an equal number of randomly chosen negative examples. Our result is comparable to those of the best methods[1, 3, 4] in the literature. Most notable in Table 1 is the non-deteriorating performance for the long chains. In [1], an off-diagonal accuracy of 21% for

[1]Bioinformatics Program, University of California, San Diego, La Jolla, CA 92093-0412, USA, dzhi@ucsd.edu.

[2]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093-0114, USA, elkan@cs.ucsd.edu.

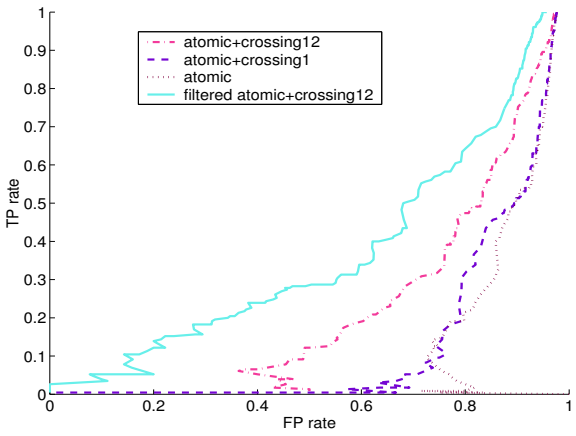


Figure 1: ROC curves for protein $1a1x$ using the model trained on this protein. Feature sets used are basic, basic + first crossing set, basic + both crossing sets. The top left (best) curve is obtained after filtering of predictions based on basic + both crossing sets.

Independent test set	all ($100 < L < 300$)	short ($L < 170$)	long ($L \geq 170$)
Number of chains	59	25	34
MECOP	21%	22%	20%
Smoothed-MECOP	24%	26%	23%

Table 1: Accuracy as defined by Fariselli *et al.* of our method (MECOP). The model is trained on a balanced dataset from 92 PDB chains for 3 iterations. The test set consists of 59 different PDB chains of length 100 to 170, and 171 to 300.

chains of length 100 to 170 is reported, compared to only 15% for chains longer than 170. Our method performs almost equally well for long chains. We hypothesize that contacts between distant residues in long chains are modeled relatively well by using features that combine separation d and other basic features. Also, smoothing improves the prediction accuracy a further 3% to 4%.

References

[1] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein Engineering*, 14:835–43, 2001.

[2] S. D. Pietra, V. J. D. Pietra, and J. D. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

[3] G. Pollastri and P. Baldi. Prediction of contact maps by giohmms and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, S1(1):1–9, 2002.

[4] M. J. Zaki, S. Jin, and C. Bystroff. Mining residue contacts in proteins. In *IEEE International Conference on Bioinformatics and Biomedical Egineering*, pages 168–175, 2000.

L16. Application of a Two-stage Method to Identify Protein-Protein Interface Residues

Changhui Yan[1,2,5*], Vasant Honavar[1,2,4,5] and Drena Dobbs[1,3,4,5]

Keywords: protein-protein interaction, SVM, Bayesian method, interaction sites

Virtually all cellular processes depend on precisely orchestrated interactions between proteins. Genome-wide proteomics studies are providing large sets of potentially interacting proteins. However, experimental elucidation of the molecular details of these interactions by examining protein complexes using X-ray and NMR methods lags far behind. Therefore, methods for computational prediction of protein interaction sites are becoming increasingly important. Here we present a sequence-based method for predicting which surface residues of a protein participate in protein-protein interactions. In the first stage, a support vector machine (SVM) classifier is trained to predict whether or not a surface residue is an interface residue using as input the identities of 9 amino acids (the target amino acid plus 4 amino acids on each side). In the second stage, a Bayesian classifier is trained to predict whether or not a surface residue is an interface residue based on the class labels (interface or non-interface residue) of its 8 neighbor residues (4 on each side). In the testing phase, the outputs of SVM classifier are used as inputs to Bayesian classifier to make new predictions. Our results show that the two-stage method presented here outperforms previously published sequence-based methods. We have applied the two-stage method to predict interface residues on CAPRI (Critical Assessment of PRedicted Interactions) targets. The success of the predictions is validated by examining the predictions in the context of the 3-dimensional structures of protein complexes.

[1]Artificial Intelligence Research Laboratory, Iowa State University, Ames, Iowa 50011
[2]Department of Computer Science, Iowa State University, Ames, Iowa 50011
[3]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa 50011
[4]Laurence H Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa 50011
[5]Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, Iowa 50011
[*]Address of the corresponding author: Atanasoff Hall 226, Iowa State University, Ames, IA 50011-1040, USA.
Email: chhyan@iastate.edu; Phone: +1-515-294-7331; Fax: +1-515-294-0258

L17. High-Throughput 3D Homology Detection via NMR Resonance Assignment

Christopher James Langmead[*] **Bruce Randall Donald**[†‡§¶‖]

January 15, 2004

Abstract

One goal of the structural genomics initiative is the identification of new protein folds. Sequence-based structural homology prediction methods are an important means for prioritizing unknown proteins for structure determination. However, an important challenge remains: two highly dissimilar sequences can have similar folds — how can we detect this rapidly, in the context of structural genomics? High-throughput NMR experiments, coupled with novel algorithms for data analysis, can address this challenge. We report an automated procedure, called HD, for detecting 3D structural homologies from sparse, *unassigned* protein NMR data. Our method identifies 3D models in a protein structural database whose geometries best fit the unassigned experimental NMR data. HD does not use, and is thus not limited by sequence homology. The method can also be used to confirm or refute structural predictions made by other techniques such as protein threading or homology modelling. The algorithm runs in $O(pn^{5/2} \log(cn) + p \log p)$ time, where p is the number of proteins in the database, n is the number of residues in the target protein and c is the maximum edge weight in an integer-weighted bipartite graph. Our experiments on real NMR data from 3 different proteins against a database of 4,500 representative folds demonstrate that the method identifies closely related protein folds, including sub-domains of larger proteins, with as little as 10-30% sequence homology between the target protein (or sub-domain) and the computed model. In particular, we report no false-negatives or false-positives despite significant percentages of missing experimental data.

[*]Carnegie Mellon Department of Computer Science
[†]Dartmouth Computer Science Department, Hanover, NH 03755, USA.
[‡]Dartmouth Chemistry Department, Hanover, NH 03755, USA.
[§]Dartmouth Biological Sciences Department, Hanover, NH 03755, USA.
[¶]Dartmouth Center for Structural Biology and Computational Chemistry, Hanover, NH 03755, USA.
[‖]Corresponding author: 6211 Sudikoff Laboratory, Dartmouth Computer Science Department, Hanover, NH 03755, USA. Phone: 603-646-3173. Fax: 603-646-1672. Email: brd@cs.dartmouth.edu

L19. Efficiently computing the landscape of locally optimal RNA secondary structures

P. Clote[1]

Keywords: RNA, secondary structure, energy landscape, biopolymer folding.

1 Introduction.

We make a novel contribution to the theory of biopolymer folding, by developing an efficient algorithm to compute, for each k, the number of k-locally optimal secondary structures of an RNA molecule, with respect to the Nussinov-Jacobson energy model. Additionally, we apply our algorithm to analyze the folding landscape of selenocysteine insertion sequence (SECIS) elements, hammerhead ribozymes from Rfam, and tRNAs from Sprinzl's database.

2 Locally optimal secondary structures.

A *secondary structure* for an RNA sequence $a = a_1 \cdots a_n \in \{A, C, G, U\}^n$ is an expression $s = s_1 \cdots s_n$ involving dot, left and right parenthesis, which is well-balanced, such that nucleotides corresponding to matching parentheses are either Watson-Crick complements or GU wobble pairs. We say that a secondary structure has threshold θ, if hairpin loops have at least θ unpaired bases. The Nussinov-Jacobson energy model [5] stipulates that each base pair of a secondary structure contributes (negative) stabilizing energy of -1, and that the *optimal* secondary structure has the maximum possible number of base pairs. For instance, with threshold $\theta = 3$, the RNA sequence `ACGUACGUACGU` of length 12 has predicted minimum free energy (mfe) secondary structure `((((....))))` with mfe of -4; here, the parenthesis notation means that the base pairs are (A_1, U_{12}), (C_2, G_{11}), (G_3, C_{10}), (U_4, A_9). In [5] Nussinov and Jacobson introduced a dynamic programming algorithm to compute the optimal secondary structure of an input RNA sequence of length n in time $O(n^3)$. This algorithm is the basis for the more realistic Zuker algorithm [6] using the Turner [4] energy model as implemented in either `mfold` or in Vienna RNA package `RNAfold` [3]. See [1] for more details. Using `mfold` or `RNAfold` the previous sequence `ACGUACGUACGU` has predicted minimum free energy (mfe) secondary structure `((((....))))` with mfe of -1.20 kcal/mol.

A *k-locally optimal* secondary structure has a *maximal* number of base pairs, in that no additional base pairs can be added to the structure without violating the definition of secondary structure (i.e. introducing a pseudoknot), yet the structure has k fewer base pairs than the maximum number possible. For example, for $\theta = 3$ and RNA sequence `AAAAUUUUU`, there are three 2-locally optimal structures: `(...)....` and `(....)...` and `...(....)`. Our notion of locally optimal is related to Zuker's notion of *saturated* secondary structure, in which "stacking regions extend maximally in both directions" and there are no isolated base pairs (see [2] as well). Note that these notions are distinct; for instance, there exist saturated secondary structures whose only locally optimal extension includes an isolated base pair.

In this paper, we announce an $O(n^4)$ time $O(n^3)$ space algorithm which computes, for each k, the number of k-locally optimal secondary structures for a given RNA nucleotide sequence, with respect to the Nussinov-Jacobson energy model [5].

[1]Departments of Biology and Computer Science (courtesy appt.), Boston College, Higgins 355, Chestnut Hill, MA 02467, `clote@bc.edu`.

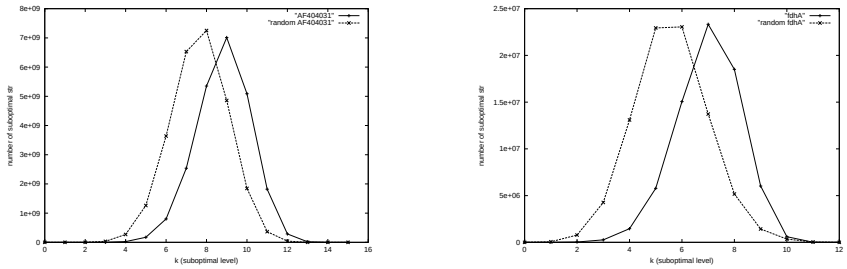


Figure 1: *(i)* Number of k-locally optimal secondary structures of 115 nt. hammerhead ribozyme AF404031 from Rfam, versus average number of k-locally optimal secondary structures of 100 115 nt. random RNAs of same dinucleotide frequency. (Left curve for random RNA.) *(ii)* Average number of k-locally optimal secondary structures of 49 nt. SECIS element `fdhA`, from A. Böck (Ludwig-Maximilians-Universität München, personal communication), versus average number of k-locally optimal secondary structures of 100 49 nt. random RNAs of same dinucleotide frequency. (Left curve for random RNA.)

As is clear from Figure 1, real RNA has a markedly distinct profile for k-locally optimal substructures than does random RNAs of the same dinucleotide frequency. Specifically, random RNA has a greater number of locally optimal structures than does real RNA, but more importantly, for small values of k (corresponding to very low energy *kinetic traps*), random RNA has far more k-locally optimal structures than does real RNA. This profile, has been observed by the author for all examined classes of structurally important RNA.

Our algorithm is not a simple modification of any of the usual secondary structure prediction algorithms [5, 6] but instead requires the introduction of additional parameters, $VisNucl$ and $VisPos$, corresponding respectively to the *visible set* of nucleotides and *visible* positions $VisPos$ from the $3'$ terminal nucleotide, in order to inductively count structures. The idea is then to use dynamic programming to compute the number k-locally optimal, s, b-visible secondary structures, where $VisNucl = s \subseteq \{A, C, G, U\}$ and $b \leq \theta + 1$. Further details and proof of correctness are lengthy and complicated, so cannot be presented here.

References

[1] P. Clote and R. Backofen. *Computational Molecular Biology: An Introduction.* John Wiley & Sons, 2000. 279 pages.

[2] D.J. Evers and R. Giegerich. Reducing the conformation space in RNA structure prediction. In *German Conference on Bioinformatics (GCB'01)*, 2001.

[3] I.L. Hofacker, W. Fontana, P.F. Stadler, L.S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsch. Chem.*, 125:167–188, 1994.

[4] D.H. Matthews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[5] R. Nussinov and A.B. Jacobson. Fast algorithm for predicting the secondary structure of single stranded RNA. *Proceedings of the National Academy of Sciences, USA*, 77(11):6309–6313, 1980.

[6] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9:133–148, 1981.

L20. Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation

Alissa Resch[1], Yi Xing[1], Alexander Alekseyenko[1], Barmak Modrek[1], Christopher Lee[1]

Keywords: comparative genomics, evolution, alternative splicing, exon, bioinformatics

Recently there has been much interest in assessing the role of alternative splicing in evolution. We have sought to measure functional selection pressure on alternatively spliced single exon skips, by calculating the fraction that are an exact multiple of three nucleotides in length and therefore preserve protein reading-frame in both the exon-inclusion and exon-skip splice forms. The frame-preservation ratio (defined as the number of exons that are an exact multiple of three in length, divided by the number of exons that are not) was slightly above random for both constitutive exons and alternatively spliced exons as a whole in human and mouse. However, orthologous exons that were observed to be alternatively spliced in the EST data from two or more organisms showed a substantially increased bias to be frame-preserving. This effect held true only for exons within the protein coding region, and not the untranslated region. In five animal genomes (human, mouse, rat, zebrafish, *Drosophila*), we observed an association between these conserved alternative splicing events and increased selection pressure for frame-preservation. Surprisingly, this effect became stronger as a function of decreasing exon inclusion level: for alternatively spliced exons that were included in a majority of the gene's transcripts, the frame-preservation bias was no higher than that of constitutive exons, whereas for alternatively spliced exons that were included in only a minority of the gene's transcripts, the frame-preservation bias increased nearly twenty-fold. These data indicate that a subpopulation of modern alternative splicing events was present in the common ancestors of these genomes, and was under functional selection pressure to preserve protein reading frame.

[1]UCLA-DOE Center for Genomics and Proteomics. Boyer Hall Rm 609, 611 Charles E. Young Drive East, Los Angeles, CA 90095-1570. Email:aresch@mbi.ucla.edu

L22. Hardness of RNA Secondary Structure Design

Rosalía Aguirre-Hernández[1], Holger H Hoos[1], Anne Condon[1]

Keywords: ribonucleic acids, secondary structure, RNA Designer

1 Introduction.

Ribonucleic acids (RNA) are macromolecules that play fundamental roles in many biological processes and their structure is essential for their biological function. This work is focused on the design of RNA strands that is predicted to fold to a given secondary structure, according to a standard thermodynamic model such as that of Mathews et al. [5]. This problem is relevant because it will facilitate the characterization of biological RNAs by their function and the design of new ribozymes that can be used as therapeutic agents [2]. There are also applications in nanobiotechnology in the context of building self-assembling structures from small RNA molecules [4].

One solution to the RNA secondary structure design problem is provided by Hofacker et al. [3], the implementation of which is included in the Vienna RNA Secondary Structure Package. A more recent stochastic local search algorithm, the RNA Designer of Andronescu et al. [1] shows a better performance.

The purpose of this work is to understand better the factors that make RNA structures hard to design. Such understanding provides the basis for improving the performance of RNA Designer and for characterising its limitations. We will describe a modification of the RNA Designer that improves the performance of the algorithm. Furthermore, it is not known whether there is a polynomial time algorithm for RNA secondary structure design. Therefore, to gain insights into the practical complexity of the problem, we present a scaling analysis to investigate the hardness of the problem on random RNA structures using the improved RNA Designer.

2 Algorithm.

The RNA Designer is a stochastic local search (SLS) procedure that uses a hierarchical decomposition of the given structure [1]. A structure is split into two substructures that do not contain multiloops. Notice that it is necessary to connect the two free ends created by the split such that both resulting substructures have exactly two free ends. To create structural boundary conditions at the split points that are similar to those of the original structure, this connection is achieved by merging the free ends with those of a static cap structure, which is a small hairpin loop of size four (five paired, four unpaired and five paired bases); furthermore, two unpaired bases are added to the two remaining free ends of each substructure if it contains a bulge directly after the first base pair.

We found that there are some structures difficult to design by using this approach. This is the case, for example, for structures in which two loops are separated by a very short stem. But it is possible to improve the performance of the algorithm by introducing a dynamic cap structure and dynamic dangling ends in order to create structural boundary conditions at the split points that are exactly the same to those of the original structure. The number of paired bases in the hairpin loop (cap structure), added to one of the substructures will

[1]Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4,Canada. E-mail: {rosalia,hoos,condon}@cs.ubc.ca

depend on the number of paired bases at the beginning of the other substructure. On the other hand, we will add one dangling end to the $5'$ ($3'$) end of a substructure if, and only if, its adjacent base in the original structure is a free base. Figure 1 (a) shows the performance correlation between the two versions of the algorithm for 60 structures of length 75. Notice that there are two outliers, which correspond to two structures for which the dynamic cap structure and dynamic dangling ends are crucial.

3 Scaling analysis.

In order to investigate the empirical complexity of solving RNA secondary structure design problems with the improved version of RNA Designer and with the Vienna algorithm, we performed a scaling analysis with random structures of length 50, 75, 100, 150 and 200. As can be seen from Figure 1 (b), the median expected run-time of both, RNA Designer and the Vienna algorithm scales polynomially with the size of the random structures. In future work, we will extend our analysis to biologically more realistic structures.

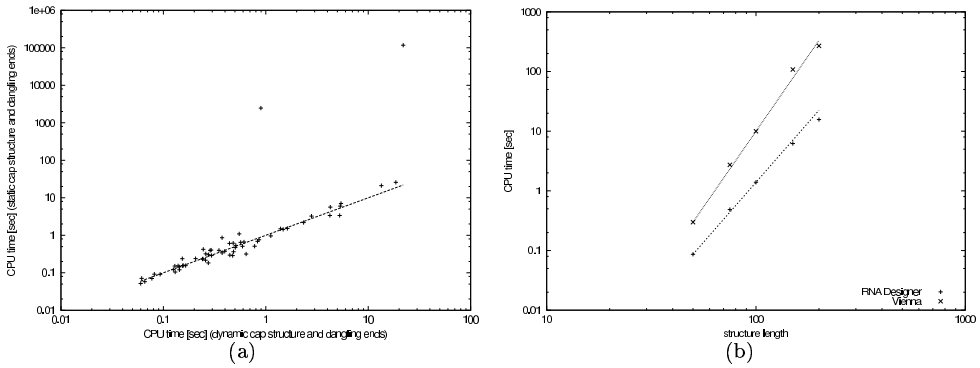


Figure 1: (a) Correlation between improved and original RNA Designer. (b) Median expected run-time over test-set of 1000 instances for length 50-75, and 100 instances for size 100-150.

References

[1] Andronescu, M., Fejes, A. P., Hutter, F., Condon, A. and Hoos, H. H. A New Algorithm for RNA Secondary Structure Design. *Journal of Molecular Biology.* To appear.

[2] Breaker, R. R. 1996. Are engineered proteins getting competition from RNA? *Curr. Opin. Biotech.* 7:442–448.

[3] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M. and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Chemical Monthly.* 125:167–188.

[4] Jaeger, L., Westhof, E. and Leontis, N. B. 2001. TectoRNA: modular assembly units for the construction of RNA nano-objects *Nucleic Acid Research.* Vol. 29, No. 2, 455–463.

[5] Mathews, D. H., Sabina, J., Zuker, M. and Turner, D. H. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology.* 288:911–940.

L23. Data Mining Atomic Motions from Computer Simulations of Nucleic Acids: A Wavelet Study of the Differential Bending of $d[GA_4T_4C]_n$ and $d[GT_4A_4C]_n$

Elijah Gregory[1], Thomas E. Cheatham, III[2] and Julio C. Facelli[3]

Keywords: Wavelet Transform, DNA, Molecular Dynamics, A-tract bending

1 Introduction.

The precise structural effect and role of A-tracts in DNA bending at the molecular level is not fully understood [2] and [3]. Anomalous runtimes in gel experiments suggest that DNA is bent by ~17-21º per A-tract in the direction of the minor groove at the center of the A-tract stretches, but do not provide molecular details. Various experimental techniques, ranging from crystallography to NMR and Raman spectroscopy have probed the nature of A-tract bending; however significant controversy regarding the molecular picture is still evident. Our goal has been to apply advanced atomistic simulation methods and novel analysis methods to provide better microscopic insight. One of these methods is wavelet analysis. Wavelet analysis is a well proven tool to extract regular patterns in time series [1] that may provide the desired information on the microscopic behavior of DNA without the biases associated with the manual analysis of extensive molecular dynamic simulations. In this poster we present our first attempt to use wavelet transforms to extract motional patterns in DNA simulations of phased A-tract DNA sequences. Contrary to the interpretation of the mobility experiments which suggest that A_4T_4 repeats move faster than T_4A_4 repeats due to enhanced bending in the former, our results suggest that the T_4A_4 repeats move slower due to enhanced interaction with the gel matrix.

2 Simulations.

A series of five sets of ~40 ns-lengths molecular dynamics (MD) simulations were performed on two models of phased A-tract DNA sequences in various salt environments using the AMBER suite of simulation programs. The sequences studied were $d[CGA_4T_4CGA_4T_4CG]_2$ (A_4T_4) and $d[CGT_4A_4CGT_4A_4CG]_2$ (T_4A_4) in explicit TIP3P solvent with varying salt concentrations (~200 mM) and identities (NaCl vs. KCl with Mg^{2+}) using the Cornell *et. al* force field, more recent variants, and varied ion parameters. All simulations utilized the particle mesh Ewald method in truncated octahedral unit cells for proper treatment of the long ranged electrostatic interactions and conservative model building (based on fiber B-DNA models of Arnott) and equilibration procedures. Preliminary results on one of these sets of simulations is discussed in our previous work [2]. In each of these simulations, contrary to what has been seen in earlier simulation work, we do not observe stronger bending in the A_4T_4 repeat compared to the T_4A_4 repeat. The wavelet analysis was done using the wavelet toolbox of MATLAB version 6. In the analysis we used Morlet wavelets with scales ranging from 1 to 511, calculating every other scale. A pass band center frequency of 0.8519 Hz was used in the calculation of the frequencies [1].

[1] Center for High Performance Computing, University of Utah, Salt Lake City, Utah 84112. E-mail: dwee@chpc.utah.edu

[2] Departments of Medicinal Chemistry and of Pharmaceutics and Pharmaceutical Chemistry, University of Utah, Salt Lake City, Utah 84112. E-mail: tec3@utah.edu

[3] Center for High Performance Computing and Department of Medical Informatics, University of Utah, Salt Lake City, Utah 84112. E-mail: Julio.Facelli@utah.edu

3 Results and Discussion.

Figure 1 shows the spectral power of the wavelet analysis of the simulations of the T_4A_4 and A_4T_4 sequences. From the figure it is apparent that T_4A_4 is significantly more dynamic than A_4T_4 on the nanosecond time scale. In particular, the strong resonances at ~2.7 10^{+11} and 2.2 10^{+9} Hz suggest that T_4A_4 has significant mobility at these frequencies. This suggests an alternate interpretation of the anomalous mobilities due to enhanced mobility of T_4A_4 rather than greater bending of A_4T_4. Our results indicate that T_4A_4, on average, is less bent than A_4T_4, but T_4A_4 undergoes faster bending oscillations. These fast oscillations effectively increase the area that the DNA fragment presents to the gel and can explain why the T_4A_4 presents anomalous larger retention times in gels.

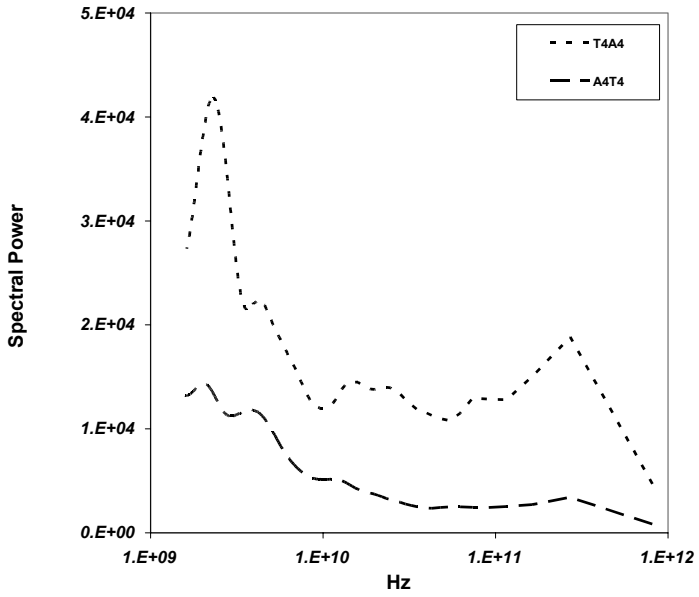


Figure 1: Comparison of the spectral power as function of wavelet pseudo frequency for the T4A4 and A4T4 simulations.

References

[1] Addison, P.S. 2002. The Illustrated Wavelet Transform Handbook. Bristol and Philadelphia: Institute of Physics.

[2] Cheatham III, T.E., and Young M.A. 2001. Molecular Dynamics Simulation of Nucleic Acids: Successes, Limitations, and Promise. *Biopolymers (Nucleic Acid Sciences) 56*:232-256.

[3] Sprous D., Young M.A., and Beveridge D.L. 1999. Molecular Dynamics Studies of Axis Bending in d(G5-(GA4T4C)2-C5) and d(G5-(GT4A4C)2-C5): Effects of Sequence Polarity on DNA Curvature. *J Mol Biol 285*:1623-1632.

L24. The Alternative Splicing Gallery (ASG) – Visualizing Gene Structure and Alternative Splicing

Jeremy Leipzig[1], Steffen Heber[1]

Keywords: splicing graph, alternative splicing, EST assembly

1 Introduction.

Alternative splicing is a major link between the estimated 30,000 genes and the myriad of proteins found in humans [1]. Existing *ab initio* gene prediction programs only infer information about one or a small number of most likely transcripts. The majority of evidence for alternative splicing and gene structure originates from the analysis of large collections of full-length mRNAs and expressed sequences tags (ESTs).

Conventionally, this transcript data is clustered into gene-specific sets and stored in gene indices. Due to the fragmentary nature and inconsistent quality of ESTs, biologists often assemble them into consensus sequences before using them for further analyses. In the presence of alternative splicing, this practice might yield a large number of consensus sequences and analyzing them in a case-by-case fashion is often very inefficient and error-prone.

We use a different approach and integrate all transcripts derived from a gene into a single splicing graph. Each transcript corresponds to a path in the graph, and alternative splicing is displayed by bifurcations. Loosely speaking, splicing graphs are built by 'projecting' transcribed sequences onto their genomic templates and 'overlaying' these projections [2]. Our approach integrates the information of all (even divergent) transcripts of a gene into a single, uniquely defined data structure, rather than handling them separately. This representation preserves the relationships between different splicing variants and allows us to systematically investigate all possible putative transcripts.

2 Material and Methods.

We built splicing graphs for all known human ENSEMBL genes based on transcript information derived from ENSEMBL, RefSeq, UniGene, and TIGR human EST databases. We analyzed these graphs for alternative splicing events (Table 1) and developed the Alternative Splicing Gallery (ASG), a web-based visualization tool (http://statgen.ncsu.edu/asg, Figure 1, [3]). ASG allows users to display splicing graphs, to interactively assemble transcripts, and to access their sequences. We constructed for each gene (except for 89 genes having more than 5000 assemblies each) an exhaustive precomputed catalog of putative transcripts – in total more than 1.2 million sequences. We found that about 65% of the investigated genes show evidence for alternative splicing, and in 5% of the cases, a single gene might produce over 100 transcripts.

[1] Department of Computer Science, College of Engineering, North Carolina State University, Raleigh NC 27695-7566, USA. E-mail: jnleipzi@ncsu.edu

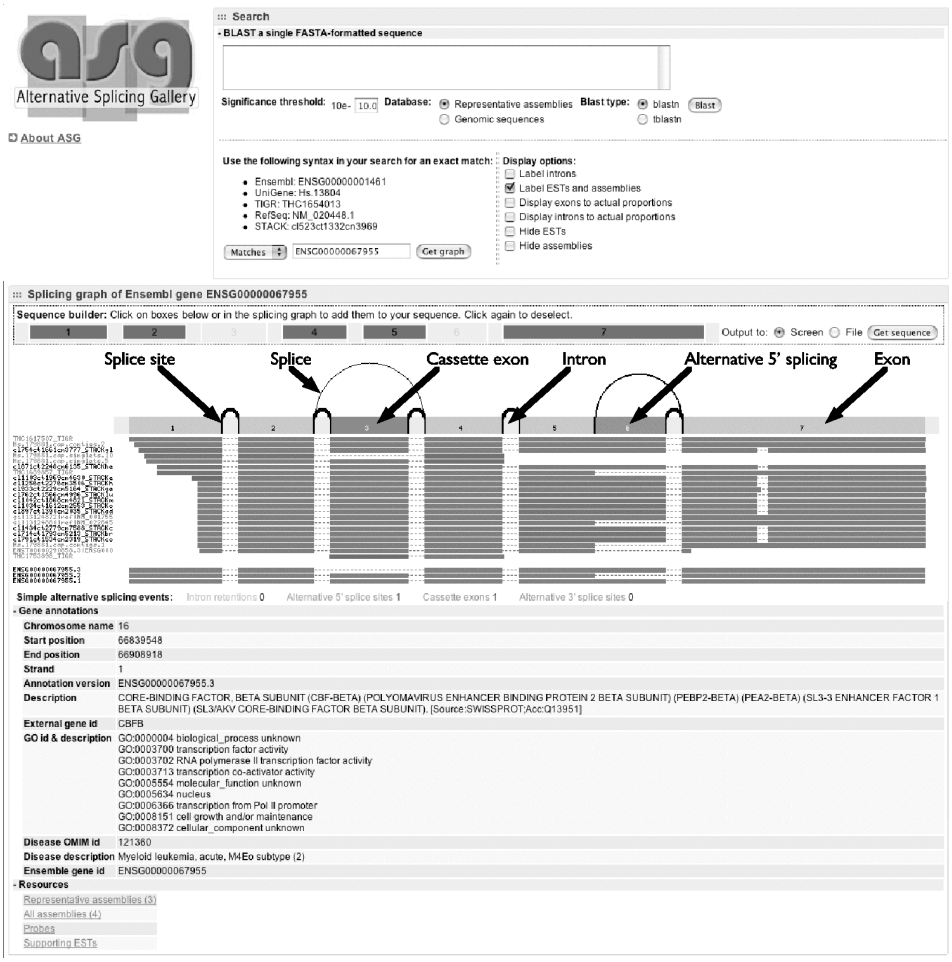


Figure 1: Visualization of the splicing graph of the human *CBFB* gene with Ensembl gene identifier ENSG00000067955 together with aligned input transcripts, representative transcript reconstructions, and gene annotation. Some of the transcripts show a small deletion in exon 6 (vertex 7) which is not bordered by splice sites and therefore not included in the splicing graph.

	total number	percent of genes	number of genes
cassette exons	13466	34.3	7590
alternative 5' splice sites	4169	16.2	3583
alternative 3' splice sites	4319	16.0	3533
retained introns	12777	31.0	6856

Table 1: Tabulation of simple alternative splicing events and number of genes where they are found.

References

[2] Heber, S., Alekseyev, M., Sze, S.H., Tang, H. and Pevzner, P.A. 2002. Splicing graphs and est assembly problem. *Bioinformatics*, 18 Suppl. 1, 181-188.

[3] Leipzig, J., Pevzner, P.A., and Heber, S. 2004. The Alternative Splicing Gallery (ASG): Bridging the gap between genome and transcriptome. (*Submitted*).

[1] Modrek, B. and Lee, C. 2001. A genomic view of alternative splicing. *Nature Genetics* 30: 13-19.

L25. Computer Modeling of DNA Unknotting by TypeII Topoisomerases

Barath Raghavan[1], Diana Nguyen[2], Javier Arsuaga[3], Mariel Vazquez[4]

Keywords: Topoisomerase II, DNA knots, polymer models, BFACF algorithm

Type II topoisomerases (Topo II) are essential enzymes common to all organisms. Their cellular functions include maintaining the levels of chromosome supercoiling and ensuring proper segregation at cell division. Topo II performs strand passage on its substrate DNA. This action has been well characterized at the molecular level [2]. Topo II binds a dsDNA segment called the G-segment, it introduces a double strand break on the G-segment allowing the passage of another DNA fragment called the T-segment. The break is resealed, and both T and G segments are released. When acting on knotted DNA molecules TopoII is known to unknot DNA below thermodynamic equilibrium [3]. That is, the steady-state fraction of knotted molecules produced by topoII is much lower that that obtained by random cyclization of linear DNA (which depends only on the conformations adopted by DNA in solution)[3]. Different biophysical models have been proposed to explain this phenomenon [4][5][6].

Here we address the question of whether the crossings acted on by topoII are selected at random or not (illustrated in Figure 1). Our study is based on Mote-Carlo computer simulations of DNA unknotting.

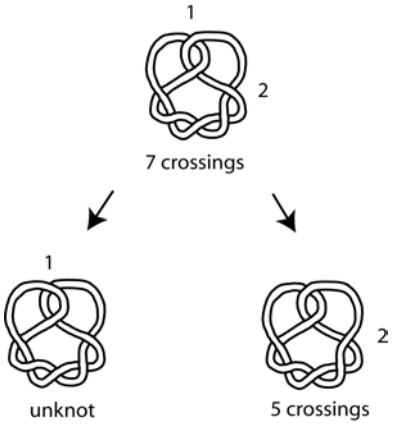


Figure 1: The figure shows two possible strand-passage events on a 7-crossing twist knot K. The pathway on the left indicates one strand-passage at the location "1" taking the 7-crossing K to the unknot in one step. The pathway on the right converts K into a 5-crossing knot by performing strand-passage at the location "2", thus three of these events are required to reach the unknot.

[1] Department of Computer Science and Engineering, UC San Diego, 9500 Gilman Dr., San Diego, CA. E-mail: barath@ucsd.edu

[2] Bioengineering Department, UC Berkeley, Berkeley, CA. E-mail: dianan@uclink.berkeley.edu

[3] Cancer Research Institute UCSF, 2340 Sutter, San Francisco, CA. E-mail: jarsuaga@cc.ucsf.edu

[4] Mathematics Department, UC Berkeley, 970 Evans Hall, Berkeley, CA. E-mail: mariel@math.berkeley.edu

We model the action of topo II as a finite state Markov chain in which each state is a knot type (with crossing number less than 8) and whose transition probabilities are estimated by Monte-Carlo computer simulations of random strand-passage on knotted molecules of fixed length. DNA molecules are modeled as a polygonal chain in the simple cubic lattice and their state space is sampled using the BFACF algorithm [1]. Strand passage simulations are performed symbolically at the Dowker-code level (an integer-entry vector whereby each entry is assigned to a crossing on a fixed knot projection). To each knot K corresponds an infinite family of Dowker-codes D_k with n,n+1,n+2,.. entries where n is the crossing number for K. However to each embedding of the knot K, with fixed length L, corresponds a finite subfamily $D_{K,L}$ and a probability distribution $P_{K,L}$ that assigns a probability to each Dowker code. We use BFACF to generate the pair ($D_{K,L}$, $P_{K,L}$) as a function of K and L. Given ($D_{K,L}$, $P_{K,L}$), we simulate random strand passage on K. We compute the transition probabilities of the Markov chain by repeating the strand-passage simulation until convergence is achieved.

References

[1] Madras, N. and Slade, G. 1993. *The self-avoiding walk.* Boston Birkhauser

[2] Roca, J., Berger, J.M., Harrison, S.C., Wang, J.C. 1996. DNA transport by a type II topoisomerase: direct evidence for a two-gate mechanism. *Proceedings of the National Academy of Sciences USA*. 93(9): 4057-62

[3] Rybenkov, V.V., Ullsperger, C., Vologodskii, A.V., Cozzarelli, N.R. 1997. Simplification of DNA topology below equilibrium values by type II topoisomerases. *Science.* 277(5326):690-693.

[4] Trigueros, S., Salceda, J., Bermúdez, I., Fernández, X and Roca, J., 2004 Asymmetric removal of supercoils suggests how topoisomerase II simplifies DNA topology. *Journal of Molecular Biology* 335:723-731.

[5] Vologodskii, A.V., Zhang, W., Rybenkov, V.V., Podtelezhnikov, A.A., Subramanian, D., Griffith, J.D., Cozzarelli, N.R. 2001. Mechanism of topology simplification by typeII DNA topoisomerases. *Proceedings of the National Academy of Sciences USA* 98(6)3045-49:5849-5856.

[6] Yan, J., Magnasco, M.O., Marko, J.F. 1999. A kinetic proofreading mechanism for disentanglement of DNA by topoisomerases. Nature 401(6756):932-935

L26. Analysis of Shotgun Sequence Data from Microbial Ecosystems

Peter Salamon[1], Mya Breitbart[2], James Nulton[3], Joe Mahaffy[4], Ben Felts[5], Beltran Rodriguez Brito[6], David Bangor[7], Forest Rohwer[8]

Keywords: ecological genomics, shotgun sequencing, contigs, Lander-Waterman

Until recently, we had no idea whether the ocean contained 10 or 10^{11} species of phage. In an attempt to rectify this situation, two of us (MB and FR) shotgun sequenced a sample of DNA extracted from only the viruses present in a 200 liter sample of seawater. Since we could only sequence a small fraction of what was there (about 1000 fragments of length approximately 663 bp long), we did not expect any contigs, i.e. any significant part of our sampled sequences to occur more than once. Much to our surprise, we found several overlaps. Thus began the collaboration which led to the present work [1].

Our tolerance for deciding that two fragments are portions of the same genome was to have 98% identity over at least 20 bps. With this criterion, the contig spectrum of the sample was [1021,17, 2], i.e. there were 1021 fragments that did not overlap with any others, 17 pairs of fragments that overlapped and two groups of three fragments that overlapped to make three contiguous pieces. Our modeling of the population structure began with the numbers of contigs expected from sampling one genome n times.

For one genome, predicting the contig spectrum is just the classic Lander-Waterman problem [2] which gives the following probability that a randomly selected fragment participates in a q contig

$$w_q = qp^{q-1}(1-p)^2 \tag{1}$$

where p is the probability of an overlapwhere

$$p = 1 - e^{-nx/L} = 1 - e^{-0.01286n} . \tag{2}$$

Here $x = 663 - 20 = 643$ is the distance between starting points of sequenced fragments required for no observed overlap and $L = 50000$ is the length of the genome. The formula follows by noting that randomly sampled points on a line are exponentially separated [3] and a q contig requires $q - 1$ overlapping fragments surrounded by 2 not overlapping fragments. The resulting expected number of samples participating in q contigs is nw_q.

[1] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: salamon@saturn.sdsu.edu
[2] Department of Biology, San Diego State University, San Diego, California, 92182-4614.E-mail: mya@sunstroke.sdsu.edu
[3] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: jnulton@mail.sdsu.edu
[4] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: mahaffy@math.sdsu.edu
[5] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: bfelts@myth.sdsu.edu
[6] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: brodrigu@rohan.sdsu.edu
[7] Department of Mathematical Sciences, San Diego State University, San Diego, California, 92182-7720.E-mail: heimdalle@yahoo.com
[8] Department of Biology, San Diego State University, San Diego, California, 92182-4614.E-mail: forest@sunstroke.sdsu.edu

Our next step was to combine this with simple parametrized models of the rank-abundance relationship for the population. Assuming that our population consists of n_i copies of genome i, we can infer that the expected number of q contigs for the population is

$$c_q = \sum_{i=1}^{M} n_i w_{qi} \tag{3}$$

where M is the number of distinct genotypes in the population. Comparing these predictions with the observed contig spectrum C_q enabled us to fit parametrized models of the rank-abundance relationship. For example, for a power law model

$$n_i = a/b^i, 1 \leq i \leq M, \tag{4}$$

we were able to find maximum (quasi) likelihood values of a, b, and M. For the fit, the contig spectrum was taken to be $C_q = [1021, 34, 6, 0, 0, 0]$. Padding with additional zeros for the higher contigs seemed to have negligible effect.

The quasi likelihood function $\mathcal{L}$ was taken to be

$$\log(\mathcal{L}) = \sum_{q=1}^{6} \frac{(c_q - C_q)^2}{2\sigma_q^2} \tag{5}$$

where the variances σ_q^2 were estimated using binomial variances

$$\sigma_q^2 = \sum_{i=1}^{M} n_i w_{qi} (1 - w_{qi}). \tag{6}$$

Note that this amounts to minimizing a variance weighted sum squared error, and corresponds to a quasi-likelihood that is the product of normal distributions. Using the crude model above, we were able to say with reasonable confidence that the number of distinct phage genotypes was on the order of only a few thousand. In follow up modeling efforts we were able calculate the mean and variance of the distribution of q contig values exactly, but the much more complicated calculations [4] yield almost identical results on a variety of samples [5, 6].

References

[1] Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segal, A.M. Mead, D., Azam, F. and Rohwer, F 002. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences USA* 99:14250–14255.

[2] Lander E.S.and Waterman M.S. 1988. Genomic mapping by fingerprinting random clones. *Genomics* 2:231–239.

[3] Feller, W.1971. *An Introduction to Probability Theory and Its Applications* New York: John Wiley & Sons, Inc.

[4] Nulton, J.D., Salamon, P., Breitbart, M., Mahaffy, J.M., Felts, B., Rodrigues, B., Bangor, D and Rohwer, F. 2004. A New Tool for Enumerative Combinatorics? In: *Proceedings of the Eighth Annual International Conference on Computational Molecular Biology (RECOMB 04).*

[5] Breitbart, M., Felts, B., Kelley, S., Mahaffy, J.M., Nulton, J.D., Salamon, P. and Rohwer, F. 2004. Diversity and population structure of a nearshore marine sediment viral community. Proceedings of the Royal Society B, in press

[6] Bangor, D., Rodrigues, B., Salamon, P., Nulton, J.D., Felts, B., Mahaffy, J.M., Breitbart, M. and Rohwer, F. 2004. Modeling Phage Species Abundance In: *Proceedings of the Eighth Annual International Conference on Computational Molecular Biology* (RECOMB 04).

M1. A Full-length HIV-1 Integrase: Molecular Modeling and Molecular Dynamics Simulations

Atchara Wijitkosoom[1] . Somsak Tonmunphean[1] . Vudhichai Parasuk[1] . Supot Hannongbua[1] . Thanh N. Truong[2]

Keywords: HIV-1 integrase, full-length, molecular dynamics simulation

1 Introduction.

HIV-1 integrase (IN), a 288 amino acid residues protein, is an enzyme catalyzing an integration of viral DNA into host chromosome. The integration process consists of three steps namely 3'-processing, strand transfer and integration. [1] IN can also catalyze a disintegration reaction, which is a reversal of the integration, *in vitro*. Proteolysis studies show that the IN composes of three domains, central region (core domain) and two terminal ends (N-terminal and C-terminal domains). [2] The core region only is sufficient for the disintegration but all three regions are required for integration process. [3] Consequently, the overall structure composing these three domains provides a potentially powerful target for rational drug design inhibiting at this process. However, such the structure has not been experimentally solved yet. The structures of each individual domain were experimentally elucidated by X-ray crystallography or NMR techniques. [4] There are two crystal structures of two-domain fragment. The first structure is N-terminal domain connected to core domain. [5] The second one is a structure of C-terminal domain connected to core domain. [6] In this study, we present a model of full-length IN. The model structure was built base on two crystal structures of two-domain fragment. A molecular dynamics (MD) simulation was applied to the model structure of full-length IN with the aims to obtain the reasonable full-length structure of the enzyme and explore its dynamical behavior.

2 Computational details.

A. Molecular Modeling

The two crystal structures of two-domain fragment were taken from the Protein Data Bank, 1K6Y and 1EX4. The missing residues (sequence-only) region, which are 47-55 and 140-148 for 1K6Y and 142-145 for 1EX4, were modeled using the Insight II. The complete two-domain fragment structures were then superimposed onto each other using the core part as a reference point by the SPDBV program. The peptide linkage between residues was created and refined using Insight II. An ionization state of the Asp, Glu, Arg, Lys and His residues were taken into consideration, as well as the positive and negative charges at N-terminal residue, Phe1, and C-terminal residue, Asp270.

B. MD simulation

All modeled parts were relaxed in order to eliminate bad atomic contacts. The structure was then solvated in a rectangular box consisting of 12435 explicit water molecules with a dimension of 103.07

[1] Department of Chemistry, Faculty of Science, Chulalongkorn University, Phyathai Road, Prathumwan, Bangkok 10330 Thailand. E-mail: i_am_atchara@hotmail.com

[2] Henry Eyring Center for Theoretical Chemistry, Department of Chemistry, University of Utah, 315 South 1400 East, Salt Lake City, Utah 84112. E-mail: truong@chemistry.chem.utah.edu

x 64.74 x 77.68 ล ^{3}. The whole system was energy-minimized using 100 steps of steepest descent and then follows by conjugate gradient for 9900 steps. The system was thermalized for 100 ps and was equilibrated for 1 ns. The long-range interaction was treated by a cut off value of 10 ล . The MD simulation was carried out at 300 K. The external coupling bath with coupling constant of 0.2 ps was employed. A 2-fs time step was applied with SHAKE. Data containing a set of coordinates of enzyme structure and water molecules was collected every 100 ps. The energy calculations, as well as MD simulation and analysis of MD trajectory were explored using the Amber 7.

3 Results.

The stability of the system as well as our model structure was examined by monitoring the thermodynamics properties; energies and temperature. The mean temperature for the system was 299.85 ฑ 1.28 K. The main chain fluctuation with respect to the equilibrium structure was shown in Fig. 1. The root-mean-squared deviation (RMSD) for backbone atoms over the MD trajectory was in the range of 1-4 ล . Among the three domains, the C-terminal domain has highest flexibility thus providing the main contribution to the RMSD of all domains.

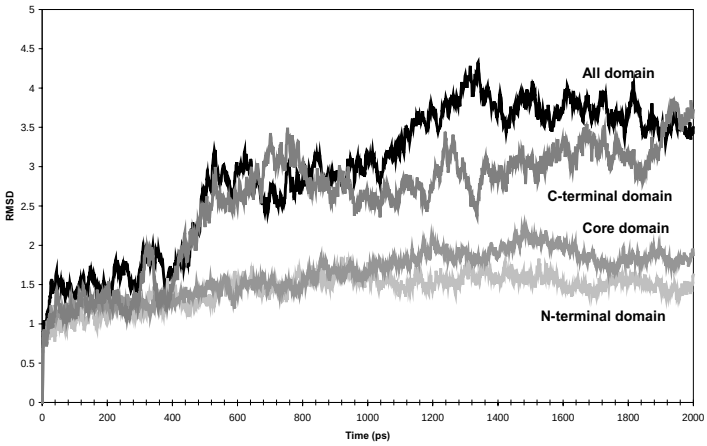


Figure 1: RMSD over the 2-ns MD trajectory.

4 References.

[1] Vink, C., Oude Groeneger, A. A. M., and Plasterk, R. H. A. 1993 *Nucleic Acids Research* 21:1419-1425.

[2] Chow, S. A., Vincent, K. A., Ellison, V., and Brown, P. O. 1992 *Science (Washington, DC, United States)* 255:723-726.

[3] Drelich, M., Wilhelm, R., and Mous, J. 1992 *Virology* 188:459-468.

[4] Goldgur, Y., Dyda, F., Hickman, A. B., Jenkins, T. M., Craigie, R., and Davies, D. R. 1998 *Proceedings of the National Academy of Sciences USA* 95:9150-9154.

[5] Wang, J.-Y., Ling, H., Yang, W., and Craigie, R. 2001 *EMBO Journal* 20:7333-7343.

[6] Chen, J. C., Krucinski, J., Miercke, L. J., Finer-Moore, J. S., Tang, A. H., Leavitt, A. D., and Stroud, R. M. 2000 *Proceedings of the National Academy of Sciences USA* 97:8233-8238.

M2. Large-Scale Biopathway Modeling and Simulation

Masao Nagasaki[1,4], Atsushi Doi[2], Kazuko Ueno[4]
Eri Torikai[4], Hiroshi Matsuno[3], Satoru Miyano[4]

Keywords: biopathways, simulation, database, Petri net, pathway modeling

1 Introduction.

Biopathway databases have been developed, such as KEGG [1] and EcoCyc [2], that compile interaction structures of biopathways together with biological annotations. However, these biopathways are not directly editable and simulatable on a personalized environment. Thus, we are developing an application called BioPathway Executer (BPE) [3] that reconstructs these two major biopathway databases to XML formats of modeling and simulation platforms. BPE is developed with JAVA and has a database of executable biopathways that integrates some parts of biopathway information, KEGG and BioCyc, and other databases, e.g. MIPS and BRENDA. Currently, BPE employs the XML format (GONML) of a Hybrid Functional Petri net (HFPN) for the output. The features of HFPN are: (i) biopathways that contain discrete and continuous processes can be modeled, (ii) all biopathways that are modeled with ordinary differential equations (ODEs) can be remodeled, (iii) biopathways can be modeled while keeping human readability. Other XML formats of biopathways, SBML [4] and CellML [5] can be described as subsets of GONML. Thus, BPE can bridge major biopathway databases and major modeling and simulating softwares.

2 Results and Discussion.

To demonstrate the effectiveness/usability of BPE, two examples are created and simulated on Genomic Object Net [6] which is based on the HFPN architecture [7,8,9]. Fig. 1(a) is a snapshot of the executable large-scale metabolic pathway with 2D plotting graphs and animations by BPE. The map compiles thirty maps that are categorized into carbohydrate metabolism in KEGG. The executable map contains more than 10000 HFPN components. Many substrates and products exist on the map but not displayed, because they are also removed in original KEGG map for human readability. Fig. 1(b) is an executable metabolic pathway with gene regulatory networks. The biopathway consists of right and left boxed parts. The right part is a metabolic pathway created by BPE. The pathway consists of two KEGG maps: glycolysis/gluconeogenesis and galactose metabolism. The left part is the gene regulatory network of lac operon that is modeled by a user.

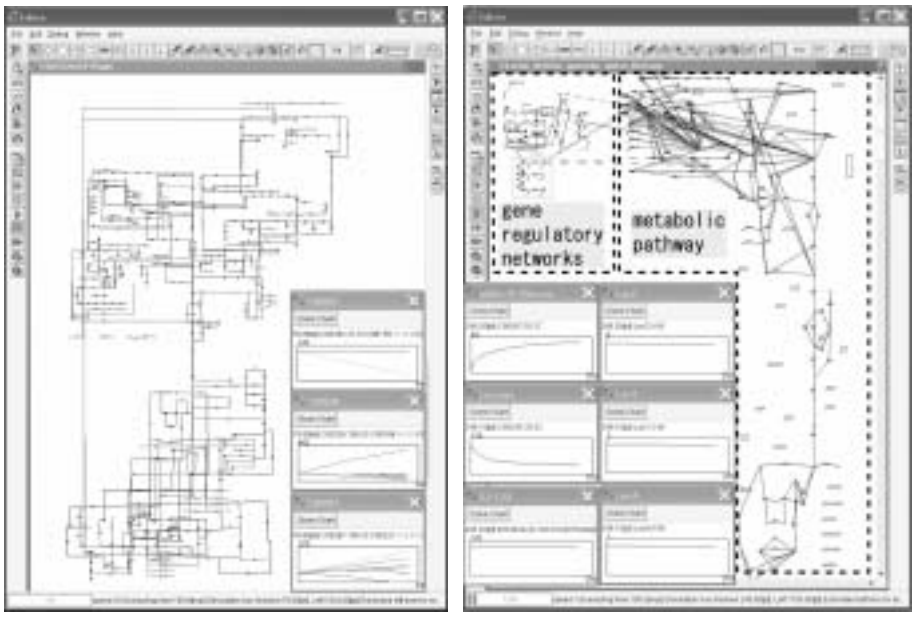
[1] Graduate School of Information Science, University of Tokyo, 7-3-1, Hongo, Bunkyoku, Tokyo, 113-0033, Japan. E-mail: masao@ims.u-tokyo.ac.jp

[2] Graduate School of Science and Engineering, Yamaguchi University, 1677-1 Yoshida, Yamaguchi, Japan. E-mail: atsushi@ib.sci.yamaguchi-u.ac.jp

[3] Faculty of Science, Yamaguchi University, 1677-1 Yoshida, Yamaguchi, Japan. E-mail: matsuno@sci.yamaguchi-u.ac.jp

[4] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo, 108-8639, Japan. E-mail:{ueno, eritori, miyano}@ims.u-tokyo.ac.jp

These examples show that BPE is a useful tool for integrating biopathway databases for large-scale modeling and simulation.



(a) (b)

Figure 1: (a) KEGG 2nd metabolic pathway map recreated with the BPE. (b) A BPE generated metabolic pathway with gene regulatory networks.

References

[9] Doi, A., Nagasaki, M., Matsuno, H., Miyano, S. 2003. Genomic Object Net: II. Modelling biopathways by hybrid functional Petri net with extension. *Applied Bioinformatics*, 2(3):185-188.

[5] http://www.cellml.org/

[2] http://www.ecocyc.org/

[3] http://www.genomicobject.net/BPE/

[6] http://www.genomicobject.net/

[1] http://www.kegg.org/

[4] http://www.sbml.org/

[7] Matsuno, H., Doi, A., Nagasaki, M., Miyano, S. 2000. Hybrid Petri net representation of gene regulatory network. *Genome Informatics*, 10:341-352.

[8] Nagasaki, M., Doi, A., Matsuno, H., Miyano, S. 2003. Genomic Object Net: I. a platform for modeling and simulating biopathways. *Applied Bioinformatics*, 2(3):181-184.

M5. Monte-Carlo Simulation of Metabolic Fluxes: Implications for Making Informative Experimental Measurements and Evaluating Systemic Impact of Enzymopathies

Jan Schellenberger[1], Nathan D. Price[2], Bernhard O. Palsson[3]

Keywords: metabolic network, constraint-based modeling, erythrocyte

Abstract

Genome-scale models provide a concise representation of available data about a biological process and can provide predictions of cellular behaviors which are difficult to observe. Constraint-based modeling is an approach that constrains cellular behavior through the imposition of physico-chemical laws, resulting in a solution space in which a cell's behavior must lie. Uniform random sampling of this constrained solution space allows for the unbiased appraisal of the implications of the imposed physico-chemical constraints upon the reconstructed metabolic network. The in silico sampling procedure was applied to the steady state flux space of the human red blood cell metabolic network under simulated physiologic conditions yielded the following key results: 1) probability distributions for all metabolic fluxes were computed for all fluxes and showed a wide variety of shapes that could not have been inferred without computation; 2) correlation coefficients were calculated between all fluxes, determining the level of independence between any two fluxes, and identifying highly correlated reaction sets; and 3) the system-wide effects of the change in one (or a few) variables (i.e. a simulated enzymopathy or setting a flux range based on measurements of physiological considerations) were computed, showing that not only do the ranges allowed to various fluxes change, but also their probability distributions and the correlations between metabolic fluxes. Taken together, this in silico sampling procedure provides a maturing of the constraint-based approach to modeling by allowing for the unbiased and detailed assessment of the impact of the applied constraints on the reconstructed network.

[1] University of California, San Diego, Bioengineering Department, Mathematics Department E-mail: jschelle@ucsd.edu
[2] University of California, San Diego, Bioengineering Department. E-mail: nprice@ucsd.edu
[3] University of California, San Diego, Bioengineering Department, E-mail: bpalsson@ucsd.edu

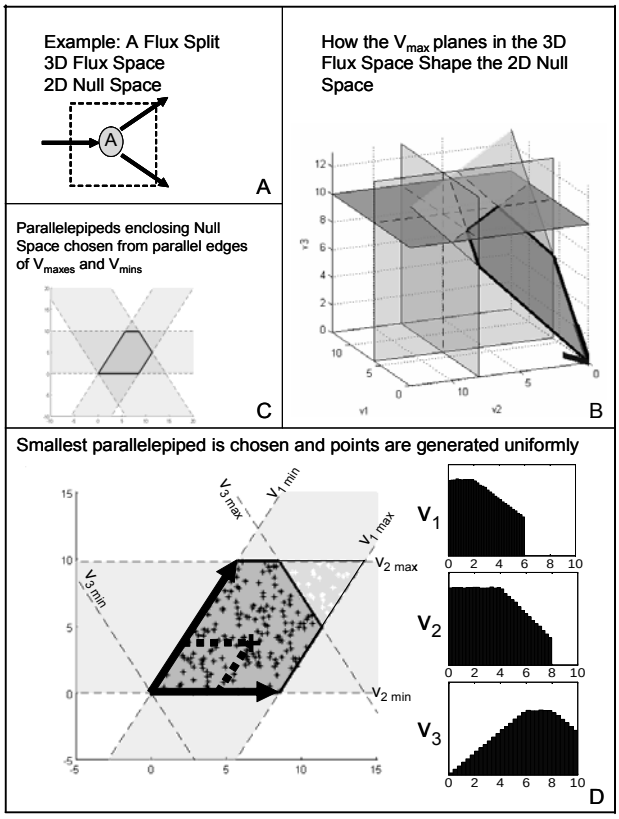


Figure 1. Algorithm for boxing in solution space with parallelepiped and generating uniform random samples. A simple flux split was used as an example to demonstrate how the *in silico* sampling procure works (A). The two dimensional null space is constrained by the Vmax planes corresponding to the three reactions in the network (B). Once the null space is capped off by the reaction Vmax values, combinations choosing two of the three sets of parallel constraints leads to forming three potential parallelepipeds (C). The smallest of these parallelepipeds is chosen and uniform random points within the parallelepiped are generated (D) based on uniform weightings on the basis vectors defining the parallelepiped (shown as black arrows). Points within the solution space are kept and those that fall out of the solution space are discarded. The fraction of the points generated inside the parallelepiped that fall within the solution space is called the "hit fraction." The hit fraction multiplied by the volume of the parallelepiped yields the volume of the solution space.

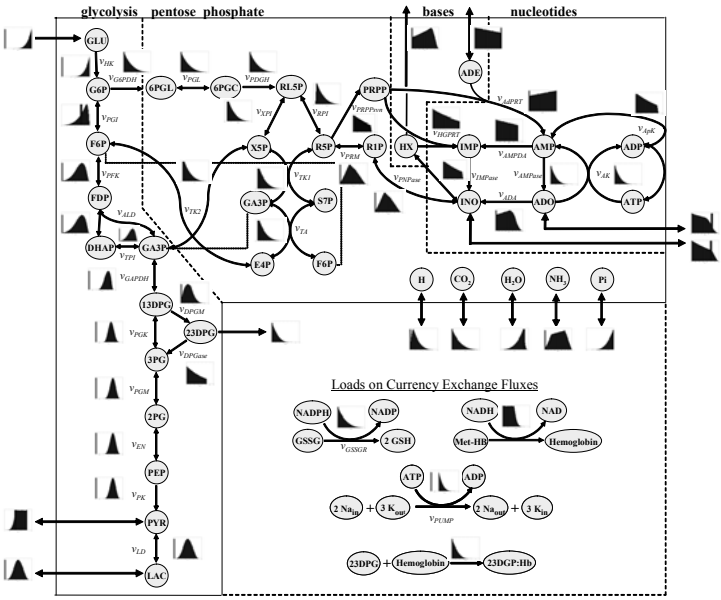


Figure 2: Probability flux distributions for human red blood cell.

The red blood cell model with maximum flux constraints was sampled using the *in silico* algorithm. The histograms next to each reaction represents the number of solutions satisfying the constraint conditions at each flux value. This gives information about the solution space's sensitivity to each constraint.

M6. Dynamic Pathway Modeling of Sphingolipid Metabolism

Peter A. Henning[1], Geoffrey Wang[2], Alfred H. Merrill, Jr.[2]*, May D. Wang[1]*

Keywords: Computational Metaboliomics, Sphingolipids, Dynamic Pathway Map, Mass Spectrometry, Systems Biology, and Simulation.

1 Background

Tremendous advances in molecular biology, both in understanding and developing of high throughput data acquisition techniques such as mass spectrometry, provide ample support for studying complex biological control networks [2, 3]. Even with recent substantial advances in data management, genomics, and robotics, the discovery of new pharmaceutical agents has not accelerated over the last few years. Although the drug industry may approach a saturation point for single targeted drugs, simple drug treatments that effectively target a control network still hold a great deal of promise[4]. Mathematical simulations of these complex networks provide researchers a means of linking the biochemistry of a reaction pathway to not only the resulting healthy phenotype but also to the dysfunctional disease state[5-7]. While still in its early years, the principles of Systems Biology could have profound effects leading to more hypothesis-driven research in drug discovery [8] and other medical treatments.

Sphingolipids perform a wide variety of biological functions: formation of specialized structures, participation in cell-cell and cell-substratum interactions, modulation of the behavior of cellular proteins and receptors, and signal transduction both as extracellular agonists and intracellular mediators. They are synthesized *de novo* via a common sphingoid base backbone (sphinganine) that is modified to produce ceramides and more complex phospho- and glycosphingolipids, some of which are covalently attached to proteins[1]. The many known functions with complex undetermined consequences make sphingolipid pathways an excellent choice to study with new modeling techniques such as system approaches.

2 System Development

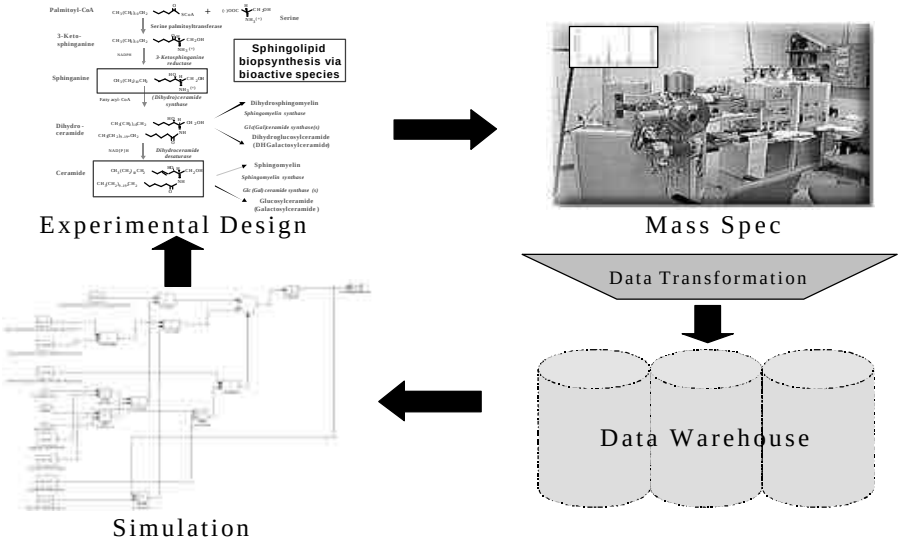
The goal of this research is to develop system approach for biology discovery. Specifically, we want to quantitatively characterize the sphingolipid metabolism pathway in order to understand their roles in normal physiological processes and investigate possible treatment options for a variety of diseases including cancer. As shown in Figure 1, first, the latest proposed sphingolipid metabolism pathways are carefully examined and evaluated to insure adequate experimental design. The focus then shifts to developing optimized mass spectrometry protocols to obtain accurate measurements of the sphingolipid composition of cultured cells at distinct time points. After the data is collected, the experimental results are stored in a database and are managed by DBMS. The governing system equations are formed, and the parameters are identified through a series of regression sequences. Then a time-based dynamical simulation is created. The simulation offers powerful insight into the

[1] Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, 313 Ferst Dr., Atlanta, Georgia, USA. Email: maywang@bme.gatech.edu

[2] School of Biology, Georgia Institute of Technology, 315 Ferst Dr., Atlanta, Georgia, USA. Email: al.merrill@biology.gatech.edu

*Authors to which correspondence should be addressed.

mysterious nature of the biological pathway and initiates an iterative process, in which new hypotheses can be quickly formulated and then validated at the bench.



Because of the size and complexity of this pathway, 3-D visualization tools are designed and implemented to enhance understanding of the biological process and to reformulate hypotheses.

3 Future Work

This system was initially built with a series of nonlinear forward flux equations as the mathematical model governing each reaction. Then the biological relevance and accuracy of the predictive equations were improved by including well-documented factors such as enzyme dependence and the reversible nature of reactions. Ultimately, the goal is to extend the integrated experimental and modeling methodologies illustrated in sphingolipid metabolism study to other complex biological process studies such as signal transduction or gene regulation. Another feature of our research is that the 3-D information representation enables the users to orchestrate the simulated pathway in real time.

References

1. Merrill, A.H., *De novo sphingolipid biosynthesis: A necessary, but dangerous, pathway*. Journal of Biological Chemistry, 2002. **277**(29): p. 25843-25846.
2. Gibbs, J.B., *Mechanism-based target identification and drug discovery in cancer research*. Science, 2000. **287**(5460): p. 1969-1973.
3. Kitano, H., *Systems biology: A brief overview.* Science, 2002. **295**(5560): p. 1662-1664.
4. Bailey, J.E., *Lessons from metabolic engineering for functional genomics and drug discovery*. Nature Biotechnology, 1999. **17**(7): p. 616-618.
5. Cascante, M., et al., *Metabolic control analysis in drug discovery and disease*. Nature Biotechnology, 2002. **20**(3): p. 243-249.
6. Sander, C., *Genomic medicine and the future of health care*. Science, 2000. **287**(5460): p. 1977-1978.
7. Noble, D., *Modeling the heart - from genes to cells to the whole organ*. Science, 2002. **295**(5560): p. 1678-1682.
8. Kitano, H., *Computational systems biology*. Nature, 2002. **420**(6912): p. 206-210.

N1. Evaluation of a New Algorithm for Keyword-Based Functional Clustering of Genes

Ying Liu[1], Brian J. Ciliax[2], Alex Pivoshenko[1], Jorge Civera[1], Venu Dasigi[3], Ashwin Ram[1], Ray Dingledine[4], and Shamkant B. Navathe[1]

Keywords: Bond energy algorithm, microarray, MEDLINE, text analysis, cluster analysis

1 Introduction.

DNA microarrays, among the most rapidly growing tools for genome analysis, are introducing a paradigmatic change in biology by shifting experimental approaches from single gene studies to genome-level analyses. Increasingly accessible microarray platforms allow the rapid generation of large expression datasets. One of the key challenges of microarray studies is to derive biological insights from the unprecedented quantities of data on gene-expression patterns. Partitioning genes into closely related groups has become an element of practically all analyses of microarray data [1]. A number of computer algorithms have been developed for this task. Although these algorithms have demonstrated their usefulness for gene clustering, some basic problems remain. We modified the Bond Energy Algorithm (BEA), which is widely accepted in psychology and database design but is virtually unknown in bioinformatics, to cluster genes by functional keyword associations. The results showed that BEA outperformed *k*-means clustering by correctly assigning 25 of 26 genes in a test set of four known gene groups. To evaluate the effectiveness of BEA for clustering genes identified by microarray profiles, 44 yeast genes that are differentially expressed during the cell cycle and have been widely studied in the literature were used as a second test set. Again, BEA performed better than *k*-means. BEA is simple to implement and provides a powerful approach to clustering genes or to any clustering problem where starting matrices are available from experimental observations.

2 Methods.

We used statistical methods to extract keywords from MEDLINE citations, based on the work of Andrade and Valencia [2]. This method estimates the significance of words by comparing the frequency of words in a given gene-related set (Query Set) of abstracts with their frequency in a background set of abstracts. We modified the original method by using a different background set, a different stemming algorithm (Porter's stemmer), and a customized stop list. The output of the keyword selection for all genes in each Test Set is represented as a sparse keyword (rows) x gene (columns) matrix with cells containing z-scores; the z-score value was set to zero if the value was less than a specified threshold. The sparse matrix was then converted to a gene x gene matrix with the cells containing the sum of products of z-scores for shared keywords. Larger values reflect stronger and more extensive keyword associations between gene-gene pairs. The Bond Energy Algorithm (BEA) takes a symmetric matrix as input, permutes its rows and columns, and generates a sorted matrix, which is then partitioned to form a clustered matrix with well-defined boundaries between

[1] College of Computing, Georgia Institute of Technology. E-mail: yingliu@cc.gatech.edu
[2] Dept. of Neurology, Emory University Medical School.
[3] School of computing and Software Engineering, Southern Polytechnic and State University
[4] Dept. of Pharmacology, Emory University Medical School.

clusters[3].

We compared the quality of clusters produced by BEA and *k*-means clustering algorithms by measuring the quality of the resulting clusters using established metrics (purity, entropy, and mutual information).

3 Results.

To determine whether keyword associations could be used to group genes appropriately, we first clustered 26 genes with both BEA and *k*-means. The 26 genes belong to four well-defined functional groups consisting of ten glutamate receptor subunits, seven enzymes in catecholamine metabolism, five cytoskeletal proteins and four enzymes in tyrosine and phenylalanine synthesis. The gene names are listed in Table 1. The BEA clustering algorithm, with z-score threshold = 10, correctly assigned 25 of 26 genes to the appropriate cluster based on the strength of keyword associations. Tyrosine transaminase was the only outlier. While BEA produced clusters very similar to the original functional classes, those produced by *k*-means did not reproduce the functional groups in Table 1.

Glutamate receptor channels		**Cytoskeletal proteins**
		1. *Actin*
1. *GluR1*	6. *KA1*	2. *Alpha-tubulin*
2. *GluR2*	7. *KA2*	3. *Beta-tubulin*
3. *GluR3*	8. *NMDA-R1*	4. *Alpha-spectrin*
4. *GluR4*	9. *NMDA-R2A*	5. *Dynein*
5. *GluR6*	***10.*** *NMDA-R2B*	
Catecholamine synthetic enzymes		**Enzymes in tyrosine and phenylalanine synthesis**
1. *Tyrosine hydroxylase*		1. *Chorismate mutase*
2. *DOPA decarboxylase*		2. *Prephenate dehydratase*
3. *Dopamine beta-hydroxylase*		3. *Prephenate dehydrogenase*
4. *Phenethanolamine N-methyltransferase*		4. *Tyrosine transaminase*
5. *Monoamine oxidase A*		
6. *Monoamine oxidase B*		
7. *Catechol-O-methyltransferase*		

Table 1: Gene sets manually clustered based on functional similarity.

Second, to determine whether our test mining/gene clustering approach could be used to group genes identified in microarray experiments, we clustered 44 yeast genes taken from [4] via [1], using BEA and *k*-means. Again, BEA outperformed *k*-means.

4 References and bibliography.

[2] Andrade, M. and Valencia, A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families, Bioinformatics, 14:600-607.

[1] Cherepinsky, V., Feng, J., Rejali, M., and Mishra, B. (2003) Shrinkage-based similarity metric for cluster analysis of microarray data, Proc. Natl. Acad. Sci. USA, 100:9668-9673.

[4] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acsd. Sci USA, 95: 14863-14868.

[3] Navathe, S., Ceri, S., Wiederhold, G., and Dou, J. (1984) Vertical partitioning algorithms for database design, ACM Trans. On Database Syst ems, 9: 680-710.

N2. A web-based approach to bio-informatics tool integration using MVC design pattern[1]

Sean Huang[2], Lang-Yang Ch'ang[3], Wen-Chang Lin[4], Chung-Shyan Liu[5]

Keywords: genome browser, web-based, tool integration, MVC design pattern

1 Introduction.

With the advances of human genome research and various sequencing projects, large volume of sequence data and their annotations have been generated. Recent progress of proteomics will generate even larger amount of data. To better extract biological signals from these data will require a new generation of software tools that can (1) integrate different tools from different data sources, and (2) allow users to visualize the final results. One such tool is genome browser, which allows users to view genomic data at a level higher than pure nucleic acids. There are already some good genome browsers available, such as UCSC genome browser [1] and Ensemble [2], both have rich datasets and viewing tools, and can graphically display various annotations. However, it is also difficult for users to integrate new tools. Also, most of the information displayed are mostly pre-computed and static, even though they are updated very quickly, some users may need real-time data.

In this paper, we will present the design and implementation of a web-based integration tool set, GeneBench, which uses model-view-controller (MVC) design pattern and can embed laboratory flow concepts implicitly. The tools in GeneBench are loosely coupled and thus it will be easier to add new tools.

2 Results and Discussion.

At present, the central piece of GeneBench is a simple genome browser. In Figure 1, a snapshot is shown. Most of the selections for datasets or processing are displayed, in tree like, at the left hand side. For example, a user can either choose a human chromosome or open up the chromosome selection display of mouse or rat to select the desired chromosome. At the center, the number of genes in each region is displayed. The size of the regions may be adjusted by zooming in or out. One can click on the number to view the identity and location of each gene, as shown in the bottom right. If a gene is selected, its detailed information will be shown. The lower part shows the introns and exons location of a selected gene. The user can then select any combination of introns and exons, each can be in either direction, to compose a transcript. The synthetic protein can then be submitted to selected protein databases for domain predictions or other processing. At the upper part, the results of domain predictions from different database are shown.

A user can also trigger BLAST or other programs to search selected databases. For example, a user may invoke BLAST to search the whole genome of human or other organisms for genes in the same family. The results may be displayed graphically [3]. Although now only BLAST is supported, other programs may be added in the future.

3 Implementation.

[1] Supported in part by NSC-Taiwan under contract number NSC 91-2213-E033-016 and NSC-92-2213-E033-040

[2] Dept. of Information and Computer Engineering, Chung-Yuan C. University, Chung-Li, 320, Taiwan.
[3] Institute of Biomedical Science, Academia Sinica, Taipei, Taiwan, E-mail: lychang@ibms.sinica.edu.tw
[4] Institute of Biomedical Science, Academia Sinica, Taipei, Taiwan, E-mail: wenlin@ibms.sinica.edu.tw
[5] Corresponding author, Dept. of Information and Computer Engineering, Chung-Yuan C. University, Chung-Li, 320, Taiwan, E-mail: liucs@ice.cycu.edu.tw

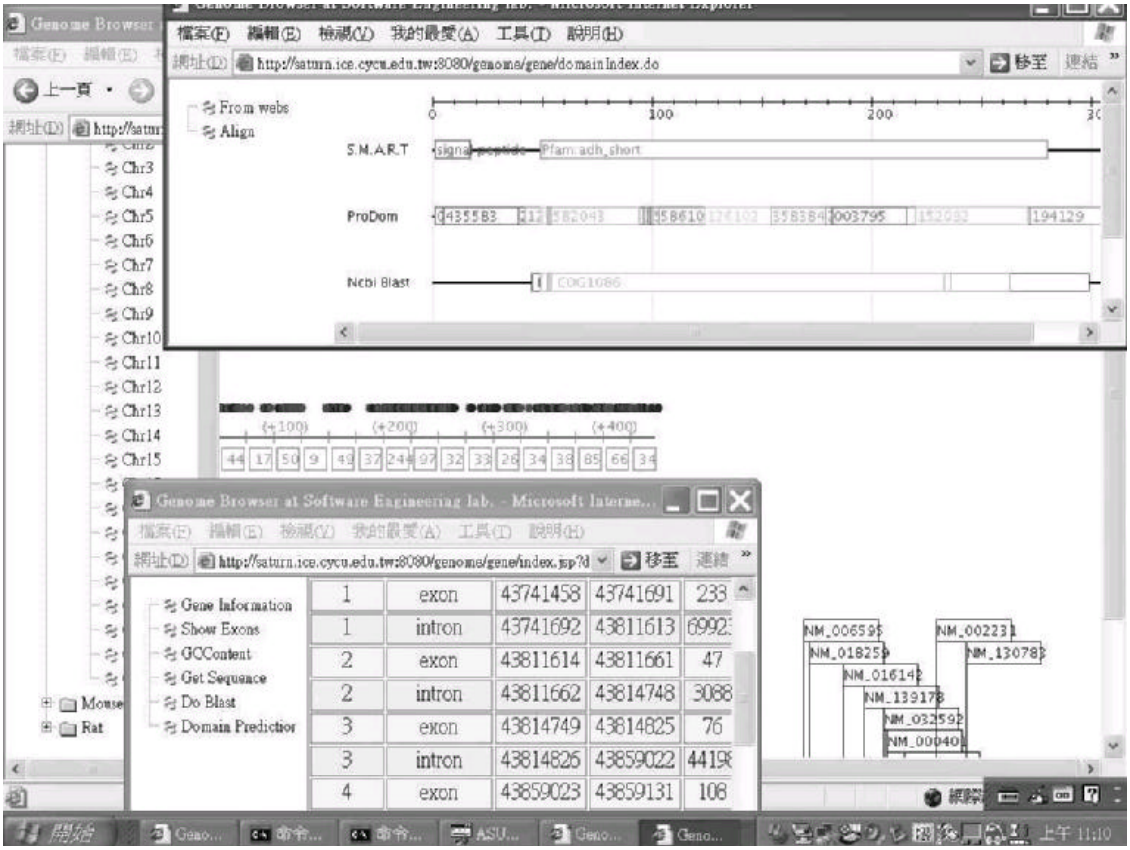


Figure 1: A display of the genome browser.

The system is now running under Tomcat 4.1 and is implemented using the Struts framework [3], which supports MVC design pattern for web applications. All programs are written in JSP and java. A small portion of programs is adopted from BioJava, e.g., computation of GCContent of the gene. The genomic datasets and annotations are from UCSC Genome Bioinformatics. The genome sequences are Human Apr2003, mouse Oct2003, and rat Jun2003. Since UCSC has a quite rich datasets and annotations, we hope to incorporate more of them in the future. The system can be accessed via *http://saturn.ice.cycu.edu.tw:8080/genome/index.jsp*.

References

[1] W. Kent, C. Sugnet, T. Furey, K. Roskin, T. Pringle, A. Zahler, and D. Haussler, 2002, The Human Genome Browser at UCSC, *Genome Research*, 12:996-1006.

[2] T. Hubbard et al, 2002, The Ensemble Genome Database Project, *Nucleic Acids Research*, 30:38-41.

[3] C. Y Yang, L.Y. Ch'ang, W.C. Lin, and C.S. Liu, On the Design of a Simple Genome Browser, Currents in Computational Molecular Biology, RECOMB-03, 31-32.

[4] http://jakarta.apache.org/struts/index.html.

[5] http://www.biojava.org.

N3. Computational analysis of homologous chromosome pairing in fission yeast

Mineo Morohashi, [1,2] **Da-Qiao Ding,** [3] **Ayumu Yamamoto,** [3]
Yasushi Hiraoka, [3] **Shuichi Onami,** [2,4,5] **Hiroaki Kitano** [1,2,5,6]

Keywords: computer simulations, homologous chromosome pairing, fission yeast, meiosis

1 Introduction

Homologous chromomes pairing is an important event during meiosis, which is followed by recombination and proper segregation. How the search for the partner chromosomes is attained has long been a crucial question in cell biology. During meiotic prophase, telomeres form a cluster beneath nuclear envelope with a polarized chromosome arrangement. The arrangement is highly conserved among eukaryotes, and often referred to as "bouquet" [1]. In addition, in fission yeast *Schizosaccharomyces pombe*, a telomere-led dynamic nuclear oscillation is observed [2]. Various mutant analyses demonstrated that inhibition of the bouquet or the nuclear movement shows marked reduction of recombination frequency. The results indicate that both the bouquet and the nuclear movement facilitate pairing, yet their direct contributions is not resolved. In this study, therefore, we examined the direct contributions and mechanisms of the bouquet and nuclear oscillation by making series of computer simulations.

2 Results

A chromosomes is modeled as a set of beads connected by springs (Fig. 1A). The model consists of a set of Langevin equations, each of which represents the dynamics of a bead on the chromosome. Each bead on the string corresponds to a pairing site on a chromosome in our model.

Based on the model, parameter values were initially estimated. The values related to the springs were estimated using image data in which chromosomes are visualized by histone-GFP. The values related to random motion were estimated using image data of thiabendazole-treated chromosomes in which *ade3/lys1* loci were stained. Other values were estimated from references.

In order to test the validity of the model, we simulated the pairing process. We measured the time until when all pairing sites on chromosomes undergo pairing, which we call "pairing time." As the results, the mean pairing time was 83 min, which was within the observed time of nuclear movement to continue *in vivo* (146 min).

To determine the roles of the bouquet and nuclear oscillation against pairing, we simulated the pairing process under three conditions (Fig. 2A–C): (i) with neither bouquet nor nuclear oscillation; (ii) with bouquet but no nuclear oscillation; and (iii) with both bouquet

[1]ERATO-SORST Kitano Symbiotic Systems Project, JST, M-31 6A 6-31-15 Jingumae Shibuya-ku, Tokyo 150-0001, Japan. E-mail: moro@symbio.jst.go.jp
[2]Graduate School of Science and Technology, Keio University, Japan.
[3]CREST Research Project, Kansai Advanced Research Center, Communications Research Laboratory, 588-2 Iwaoka Nishi-ku, Kobe 651-2492, Japan.
[4]Institute for Bioinformatics Research and Development, JST, 3-14-1 Hiyoshi Kohoku-ku, Yokohama 223-8522, Japan.
[5]The Systems Biology Institute, Japan.
[6]Sony Computer Science Laboratories, Inc. Takanawa Muse Bldg. 3-14-13, Higashigotanda Shinagawa-ku, Tokyo 141-0022, Japan.

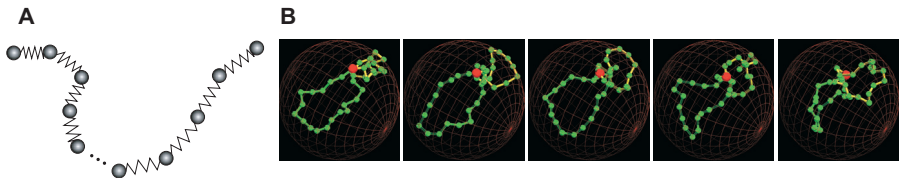


Figure 1: Chromosome model and simulations of pairing process. **A**. Schematic view of the chromosome model. **B**. Snapshots of simulations under condition of telomere clustering.

and nuclear oscillation. The time needed for all the pairing sites to attain pairing is defined as "the pairing time", and is used as a criterion to measure pairing performance. As expected, we found that the pairing time with condition (ii) was 77% shorter than that with condition (i), and the pairing time with (iii) was 94% shorter than that with condition (i) (Fig. 2D). Furthermore, the variation of the pairing time was reduced in both cases. These results indicate that the bouquet directly shortens the pairing time and also reduces its scatter, which is further augmented by nuclear oscillation.

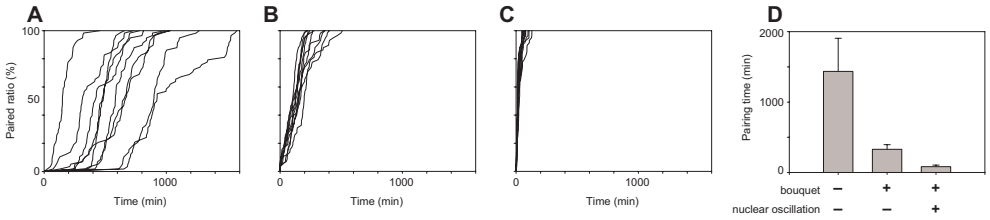


Figure 2: Simulated pairing time under various conditions. **A-C**. Ratio of paired sites as a function of time. The superimposed 10 tracks are illustrated under 3 conditions. **D**. Mean pairing time and its variance. The error bars represent standard deviations.

Further analyses suggest that the bouquet formation contributes to pairing by (a) spatially constraining the chromosomes in the proximity, and (b) disposing the chromosomes in a way that all pairing loci lie within closed-end regions. The nuclear oscillation contributes to pairing probably by aiding the spatial constraint of the chromosomes.

3 Concluding Remarks

In this study, we have examined direct roles and mechanisms of the bouquet and nuclear oscillation upon pairing by computer simulations. In conjunction with cytological and molecular biology approach, our approach will make a significant contribution to our understanding of chromosome dynamics.

References

[1] Scherthan, H. 2001. A bouquet makes ends meet. *Nature Reviews in Molecular Cell Biology* 2:621–627.

[2] Chikashige, Y., Ding, D.-Q., Funabiki, H., Haraguchi, T., Mashiko, S., Yanagida, M. and Hiraoka, Y. 1994. Telomere-led premeiotic chromosome movement in fission yeast Schizosaccharomyces pombe. *Science* 264:270–273.

N4. Primary Human Hepatocytes – A Suitable Tool in Systems Biology

Dieter Runge[1], Dirk Koczan[1], Detlef Haase[1], Hilmar Christoph[1], Peter Lorenz[1], Peter Kohlschein[2], Peter Schuff-Werner[2], Michael O. Glocker[1], and Hans-Jürgen Thiesen[1, 3]

Keywords: EGF, HGF, RNA profiling

1 Introduction.

The human genome encodes 30.000 to 40.000 genes whose expression in space and time reflects the ontogenesis of man. To describe the complexity of life by systems biology approaches, novel tools have to be developed to visualize and to simulate biological processes within a cell, a tissue or a complete organism. To feed virtual models with real data high throughput systems are required which allow data acquisition and data integration at the levels of RNA (transcriptomics), proteins (proteomics), and metabolites (metabolomics). As part of our initiative si-RNA approaches are currently performed to monitor perturbations in time and space.

2 Materials and Methods.

We have established and used a standardized, reproducible culture system for primary hepatocytes that may serve as a suitable human model system in systems biology. A broad variety of parameters was determined to analyze functional maintenance of human hepatocytes and cellular viability. Hepatocellular function was assessed by evaluation of glucose balance, ammonia uptake, and the synthesis and secretion of urea and albumin. Lactate dehydrogenase, glutamate pyruvate transaminase and glutamate oxalacetate transaminase activities in the culture medium were determined to define cellular integrity. For control purposes we followed up on 1.) glycolysis, as a well documented "housekeeping" pathway, and analyzed the RNA expression patterns of glycolytic enzymes, 2.) a variety of genes known to be expressed specifically in hepatocytes that are indicators for hepatocellular function and 3.) a set of genes reflecting proteins involved in signal transduction and liver regeneration. Hepatocyte growth factor (HGF) and epidermal growth factor (EGF) were then used to modulate cellular gene expression patterns since both growth factors have pleiotropic effects and are involved in regulating a variety of cellular pathways. Changes in RNA and protein expression patterns were determined by Affymetrix microarrays and two-dimensional gel electrophoresis. Data obtained by this means are stored and analysed in our ProteoBase data bank system.

3 Results.

Hepatocyte cultures preserved cellular integrity and maintained stabile hepatocellular functions from day 4 to day 10, thus giving us a broad time window to study the effects of exogenously applied modulators as well as siRNA. In general, all hepatocyte cultures examined so far were characterized by well preserved RNA expression patterns in the pathways chosen as internal controls (glycolysis,

[1] Institute of Immunology, Proteome Center Rostock; www.pzr.uni-rostock.de
[2] Institute of Clinical Chemistry and Pathobiochemistry
University of Rostock, Schillingallee 70, D-18055 Rostock, Germany
[3] E-mail: hans-juergen.thiesen@med.uni-rostock.de

signal transduction, membrane receptors, secretory proteins). In most instances, these expression patterns remained unchanged in the presence of growth factors. However, HGF and/or EGF modulated the expression of a variety of proteins involved in growth, differentiation and motility of hepatocytes, thereby confirming the suitability of this culture system as valuable tool in systems biology. In addition, even though only a limited number of different donors was available in our study, we noticed strong donor-dependent differences in the expression of a number of genes, some of them being considered as important players in maintenance of a hepatocytic phenotype.

4 Discussion.

Our initial analyses were performed to confirm the usefulness of primary human hepatocytes in studying the physiology of liver functions in vitro. Studying the RNA expression patterns of iso-enzymes we had to realize that enzymatic activities present in human cells most likely have to be assigned to different RNA species that are shown by Affymetrix microarray analysis to be expressed concomitantly. This observation demonstrates the complexity of regulatory pathways that has to be taken into account once metabolic pathways are going to be modelled and simulated – a prerequisite in computational biology.

5 Summary.

The data obtained in this study clearly demonstrate the feasibility of the established human hepatocyte culture system as a suitable tool in systems biology. The study also underlines the importance and value of high-through-put platform technologies and a laboratory data management system that allow data acquisition and data integration at transcriptome, proteome and metabolome level. This holds especially in the context of a human systems where inter-individual differences itself may modulate the outcome of any given experiment.

6 References and bibliography.

References

[1] Runge D., Michalopoulos G.K., Strom S.C. and Runge D.M. 2000. Recent advances in human hepatocyte culture systems. *Biochem Biophys Res Commun*. 274: 1-3.

[2] Runge D., Runge D.M., Jager D., Lubecki K.A., Beer-Stolz D., Karathanasis S., Kietzmann T., Strom S.C., Jungermann K., Fleig W.E., Michalopoulos G.K. 2000. Serum-free, long-term cultures of human hepatocytes: maintenance of cell morphology, transcription factors, and liver-specific functions. *Biochem Biophys Res Commun*. 269: 46-53.

N5. GO trees: Predicting GO associations from protein domain composition using decision trees

Boris Hayete[1] and Jadwiga R. Bienkowska[2]

Keywords: gene ontology, classification, decision trees, functional domains

1 Introduction.

The Gene Ontology (GO) [1] offers a comprehensive and standardized way to describe a protein's biological role. Proteins are annotated with GO terms based on direct or indirect experimental evidence. Term assignments are also inferred from homology and literature mining. Regardless of the type of evidence used, GO assignments are manually curated or electronic. Unfortunately, manual curation cannot keep pace with the data, available from publications and various large experimental datasets. Automated literature-based annotation methods have been developed in order to speed up the annotation. However, they only apply to proteins that have been experimentally investigated or have close homologs with sufficient and consistent annotation. One of the homology-based electronic methods for GO annotation is provided by the InterPro database. The InterPro2GO/PFAM2GO [2, 3] associates individual protein domains with GO terms and thus can be used to annotate less studied proteins. However, protein classification via a single functional domain demands stringency to avoid large number of false positives. This work broadens the basic approach.

2 Materials and Methods.

We model proteins via their entire functional domain content (PFAM domains [4]) and train individual decision tree classifiers [5] for each GO term using protein assignments. In a conservative manner, we use lack of assignments of proteins to a term, in combination with other evidence, to generate different kinds of negative examples. We train individual OC1 [5] decision tree classifiers for each GO term using this training set both in oblique and in axis-parallel mode.

As a benchmark we compare our results to those generated using a reference list generated by the InterPro database. InterPro2GO associates the protein domains with GO terms [2, 6]. InterPro2GO is the only available approach that uses domain information to predict GO terms. InterPro2GO uses a simple association rule: a protein domain is associated with a GO term if all proteins associated with the term have that domain. We have used the InterPro2GO list to assess the performance of our method by annotating proteins in our testset with GO terms via InterPro2GO's mapping of domains to terms. Only the InterPro mappings which had corresponding PFAM annotations were applied, to properly compare this approach to our technique, which utilized PFAMs.

3 Results.

We demonstrate that our approach is sensitive, specific and precise, as well as fairly robust to sparse data. We have found that our method is more sensitive when compared to the InterPro2GO performance and suffers only some precision decrease. In comparison to the InterPro2GO we have improved the sensitivity by 22%, 27% and 50% for Molecular Function, Biological Process and Cellular GO terms respectively.

Given the relatively small training set, some GO terms will have quite a small number of positive examples. The average number of positives is much smaller for the leaf terms (= 4) than the parent terms (= 67), where leaves are the nodes in the GO graph that do not have any children. To check how the small number of positives affects the performance we have calculated the averaged over the GO terms in the two categories the Precision, Sensitivity and Specificity. The average performance for leaves was: Precision = 93.8 ± 21.0, Sensitivity=88.6±25.3, Specificity=99.99 ± 0.07. The average performance for parents is: Precision = 88.0 ± 24.0, Sensitivity = 81.7 ± 25.3, Specificity = 99.6 ± 4.1. These results show that the performance of the classifier is not affected by a small number of positive examples. In fact the leaf terms have on average better performance as we can expect from the more detailed level of description given by the leaf terms. Results of our initial investigation show that the Decision Tree training approach is a valid and effective method for assigning GO ontology terms to proteins based on the domain composition.

[1] Bioinformatics Program, Boston University, 44 Cummington St, Boston, MA 02215. Email: theboris@bu.edu
[2] Serono Reproductive Biology Institute, One Technology Pl., Rockland MA 02370. Email: jadwiga.bienkowska@serono.com

	InterPro2GO	Decision Tree
Biological Process	prec = 91.5 ± 1.8 sens = 42.6 ± 0.7 spec = 99.9 ± 0.2	prec = 82.9 ± 1.3 sens = 69.1 ± 1.1 spec = 99.9 ± 0.1
Cellular Component	prec = 99.8 ± 3.4 sens = 34.8 ± 1.0 spec = 99.9 ± 0.4	prec = 85.0 ± 1.8 sens = 84.9 ± 1.8 spec = 99.8 ± 0.2
Molecular Function	prec = 98.9 ± 2.0 sens = 58.4 ± 1.0 sepc = 99.9 ± 0.2	prec = 82.6 ± 1.3 sens = 81.0 ± 1.3 spec = 99.9 ± 0.1

Table I. Comparison of the performance of the Interpro2Go and Decision Tree approach. All SwissProt proteins from human, mouse and yeast annotated in GO are our training set. All SwissProt proteins from *D. Melanogaster* and *C.Elegans* constitute the test set.

1. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
2. Mulder, N.J., et al., *The InterPro Database, 2003 brings increased coverage and new features.* Nucleic Acids Res, 2003. **31**(1): p. 315-8.
3. Mulder, N.J., et al., *InterPro: an integrated documentation resource for protein families, domains and functional sites.* Brief Bioinform, 2002. **3**(3): p. 225-35.
4. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2002. **30**(1): p. 276-80.
5. Murthy, S.K., S. Kasif, and S. Salzberg, *A System for Induction of Oblique Decision Trees.* Journal of Artificial Intelligence Research, 1994. **2**: p. 1-32.
6. Camon, E., et al., *The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro.* Genome Res, 2003. **13**(4): p. 662-72.

N6. Development of an Integrated LIMS for Microarray Facility Center

Jianchang Ning[1]

Keywords: microarray, LIMS, bioinformatics

Abstract

Microarray has become a vital tool in functional genomics, providing a high-throughput means of genome-wide analyses. Capturing both the data and metadata that is required for the interpretation of such analyses is a crucial requirement for today's biologists. Adding in some additional requirements makes writing such a tool an interesting challenge. A LIMS is being developing for the Microarray Core Facility Center of Delaware Biotechnology Institute. The LIMS was designed to automate the management of the Center, easy the use of the facility, store the information of the samples, experiments, processes and accounting, and archive the raw image data and the result text data. The LIMS uses Oracle 9 DBMS and also has the capability to work with MySQL and Postgres DBMS. Object-oriented programming approach with Java is used to build the applications and Java Servlet/JSP technology is employed to develop the interface. Thus, it achieves high performance, high security and easy deployability. Although it was originally designed for the DBI Microarray Center, it is easy to be adapted for any other similar microarray facility.

Introduction

Microarray technology is having a significant impact on functional genomics research by allowing scientists to measure the expression level of thousands of genes simultaneously. Due to the complexity of gene expression data, much detailed information besides results needs be recorded during the entire laboratory experiments so that the results can be appropriately interpreted and compared [1]. Because of the high-throughput nature, a central microarray facility needs a Laboratory Information Management System (LIMS) for tracking the operation of its instruments, the processes and the experiments.

LIMS is a system that tracks, manages and stores information associated with a laboratory, such as customers, samples, parameters, results, operators, passwords, etc. Over the past two decades, LIMS has become the workhorse of the laboratory, encompassing laboratory work-flow combined with user input, data collection, instrument integration, data analysis, user notification, and delivery of information and reporting. The broadly accepted microarray data standard called MIAME requiring much detailed information of the study plus security-sensitive accounting information post a challenge for development of such laboratory information management system.

---

[1] Delaware Biotechnology Institute, University of Delaware, 15 Innovation Way, Newark, DE 19711. E-mail: ning@dbi.udel.edu

In this project, the author used the Microarray Core Facility Center at Delaware Biotechnology Institute (DBI) as an example and has been developing an integrated LIMS for the Center. This project includes three phases: 1. track and record the work-flow, facilitate the management and easy the access of the voluminous data; 2. integrate the management of the inventory; 3. incorporate with analysis software. This reports the result of the first phase.

Architecture

The LIMS was designed to facilitate the organization, storage, archive and retrieval of microarray experiment-associated information and data within a secure environment and for ease-of-use and flexibility. The system provides a clear management based on accession rights and user roles. All the access to the system is solely through web browsers. The laboratory manager (operator) creates a client account and assigns a password to the client upon his/her request. Users will receive defined privileges for the access of the system. The system will inform the client of the account information by sending his/her an email and asks the client to initialize personal and billing information in the system. The client can update the information anytime later once the account is created by the operator. The client can also check the sample status. The operator will process the samples and input experimental conditions. The system will assign the chip ID and the coordinates. The operator can update all the experiment-related information and also generate reports of clients, samples, experiments, usage and so on anytime. The system will automatically record the billing information including the balance according to the protocols (Affy GeneChip and customized array), the experimental procedure and the operator's inputs. Once the experiment is done, the raw image files and text files will be copied to the data server and the system will inform the cashier of that by sending a billing statement. Once the balance is received, the cashier can unclock the client account so that the client now can download the results. By default, client accounts are locked if there is any balance remained so that the clients will not be able to view and download the results. But the cashier has the privilege to intervene that. The system administrator has all the privilege to do anything with the system.

Implementation

The LIMS is based on a relational database and has the capability to work with Oracle, MySQL and Postgres. The database of this example is Oracle 9i in the SUN Solaris 9 OS environment. The data processing applications and all the interface are implemented in Java according to Java Servlet/JSP technology.

References

[1] Brazma A, Hingamp P, Quackenbush J, et al. 2001. Minimum information about a microarray experiment (MIAME) – toward standard for microarray data. Nat. Genet. 29(4): 365-71.

N7. Cooperative Biomedical Knowledge Inference

Chun-Hsi Huang, Sanguthevar Rajasekaran, Longde Yin [1]

Keywords: molecular biology, knowledge inference, semantic network

1 Introduction.

The Human Genome Project (HGP) is extracting information from the DNA strands that constitutes human genetic inheritance. The acquisition of a comprehensive human genome sequence imposes unprecedented impact on basic biology, biomedical research, biotechnology, and medicine. It is crucial the massive genomic data produced are well represented so that useful biological information may be efficiently extracted/inferred. The Unified Medical Language System (UMLS) project was initiated in 1986 by the U.S. National Library of Medicine (NLM). The goal of the UMLS is to facilitate associative retrieval and integration of biological and biomedical information so researchers and health professionals can use such information from different (readable) sources [1].

The UMLS project consists of three core components: (1) the **Metathesaurus**, providing a common structure for more than 95 source biomedical vocabularies. It is organized by concept, which is a cluster of terms, *e.g.*, synonyms, lexical variants, and translations, with the same meaning. (2) the **Semantic Network**, categorizing these concepts by semantic types and relationships, and (3) the **SPECIALIST lexicon** and associated lexical tools, containing over 30,000 English words, including various biomedical terminologies. Information for each entry, including base form, spelling variants, syntactic category, inflectional variation of nouns and conjugation of verbs, is used by the lexical tools [2]. The 2002 version of the Metathesaurus contains 871,584 concepts named by 2.1 million terms. It also includes inter-concept relationships across multiple vocabularies, concept categorization, and information on concept co-occurrence in MEDLINE.

In this research work, we focus on designing a distributed UMLS semantic network to cooperatively infer biological and medical information efficiently from distributed information sources.

2 Software and files.

The system construction bases on a task-based and message-driven model to exploit both task and data parallelism while processing queries. Queries are decomposed into tasks and distributed among processors for execution. Other system support activities are also decomposed into system tasks and distributed as well. When a task is completed, a message is generated to either spawn new tasks or trigger further processing, depending on the property and current status of the task. This process is carried out by two collaborating components: the *host system* and the *slave system*. The host system interacts with the user and processes the information for the slave system, while the slave system performs task execution.

The host system is composed of the following major components. The *language front-end* interacts with the user and decomposes the commands into either knowledge or tasks. All the preprocessing and distributing are carried out in the *command processing module*. The *object-oriented packing module* is the communication channel between processors. When the

[1]Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269, USA. E-mail: {huang,rajasek,longde}@cse.uconn.edu

slave module finishes a query, the answer messages are then sent back to the *host answer processing module* of the host system to be merged into a final inference conclusion. Some knowledge is kept in the *host knowledge base* for simple queries. Similarly, the slave system has the following components: the *shared knowledge management module*, the *task execution module*, the *kernel message module*, the *task execution engine*, the *load balancing module*, the *duplicate checking module*, the *slave scheduler* and the *object-oriented packing system*.

3 Figures and tables.

Fig. 1 illustrates the software architecture of the host system. Tests of individual components and the overall performace are being conducted on a local Grid, consisting of three heterogeneous systems: a SUN Cluster, an SGI Origin 3800, and a Dell Pentium Cluster. Preliminary experiments demonstrate promising results.

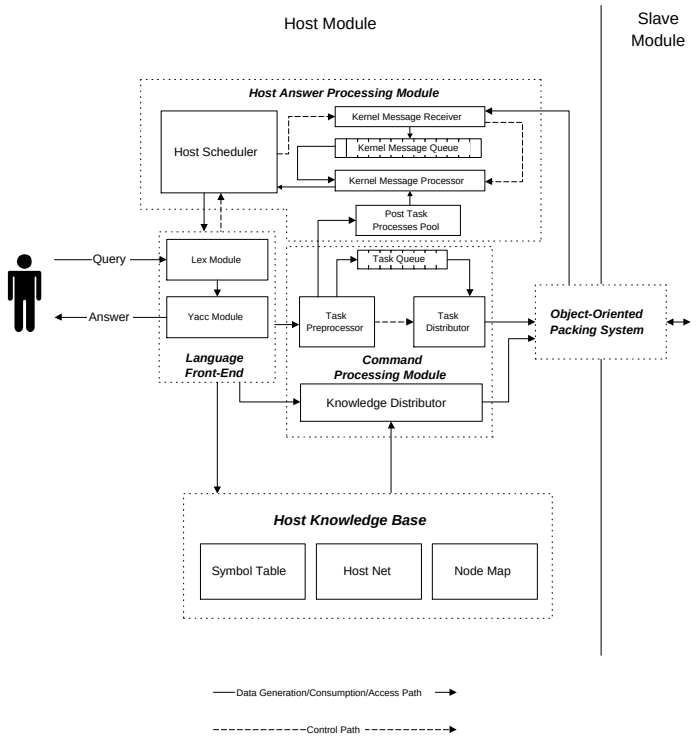


Figure 1: Host System Software Architecture

References

[1] Lindberg, D., Humphreys, B., and McCray, A. The Unified Medical Language System. *Methods Inf. Med. 32*, 4 (1993), 281–291.

[2] McCray, A., Srinivasan, S., and Browne, A. Lexical methods for managing variation in biomedical terminologies. In *Proc. Annual Symposium Compu. Appl. Med. Care* (1994), pp. 235–239.

N8. A property-based model for lung cancer diagnosis

Alma Barranco-Mendoza[1], Deryck R. Persaud[2], Verónica Dahl[3]

Keywords: cancer diagnosis, biomarkers, molecular targets, logic programming, constraint handling rules, concept formation.

1 Introduction.

To this day, lung cancer remains the leading cause of cancer death for both sexes: almost one-third of cancer deaths among men and almost one-quarter among women. [1] Survival rates for lung cancer are low. It is suggested that only about 15% of patients are diagnosed at the early stages. The average 5-year survival rate for patients that are diagnosed early is 48% compared to 15% for those who were diagnosed at the later stages. [2] It is well known that the survival rates can be improved by the early detection of pre-invasive lesions, which are believed to be the possible precursors of malignant tumours. Although new technology is allowing numerous early lesions to be detected, it is becoming clear that only a small percentage of these will actually progress to cancer. Currently a lot of work is being done on the image analysis field, however, at the early development stages of the lesion, the information that can be obtained from lung imaging analysis (X-ray, CAT scans, MRI, PET scans, etc.) is quite limited. To completely understand the evolution of normal epithelium into invasive neoplasia would require the understanding of the genetic relationship of the cells in a pre-invasive neoplastic lesion during the development into invasive cancer. In recent years, biological research has been done in the area of cancer genetics that has shown that cancer results from an accumulation of key mutations in expanding clones originating from tissue-specific stem cells. [3] The recent availability of the human genome sequence, and the development of high throughput genomic technologies and methods for isolating selected cell populations have started to give us the opportunities for understanding how human cancers develop. This information will drastically improve cancer diagnosis and treatment through the discovery of disease-specific molecular targets. As well, recent research has also been focusing on the detection of biomarkers obtained from serum and sputum proteomic analysis [4]. Unfortunately, there is not much work done in terms of computational tools to assist in the analysis of all the genetic and molecular information in addition to the radiological, serum and sputum data that could determine with better accuracy whether an early lesion would progress into cancer or not. As well, for a system to be more valuable it should be able to provide some kind of diagnosis even if given incomplete patient information, as not all tests can or will be done on said patient. Our research intends to address this with the development of a multidisciplinary property-based model for early lung cancer diagnosis.

2 Concept Formation Rules

Concept Formation Rules (CFR) [5] is a directly executable new cognitive model of knowledge construction inspired in constructivist theory as well as in recent natural language processing

[1]School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada. E-mail: abm@cs.sfu.ca

[2] Infogenetica Bioinformatics, 3197 Tahsis Avenue, Coquitlam, BC, V3B 6E2, Canada E-mail: deryck@infogenetica.com

[3] School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada. E-mail: veronica@cs.sfu.ca

methodologies. This system accommodates user definition of properties between concepts as well as user commands to relax their enforcement; accepts concept formation from concepts that violate principles that have been declared as relaxable, and produces a list of satisfied properties and a list of violated properties as a side effect of the normal operation of the rules.
We have used CFR to implement our model for lung cancer diagnosis. As part of the input concepts it accepts the patient's age, smoking history, malignancy history, radiological, serum and sputum data. The knowledge store includes the properties that should be evaluated for each input data element as well as the relations amongst them. The diagnosis is given as a probability of cancer that is calculated as a function of the concepts used in the analysis. As well, the diagnosis will list those diagnostic properties that were satisfied and those that were not. For example:

const(Prob),age(x,**A**),history(x,smoker,**T**),serum_data(x,marker_type,in_range)<=>
marker(x,marker_type,in_range,**P**,B),acceptable(marker(x,marker_type,in_range, **P**),B),
probability(**P**,Prob,x, B) acceptable(probability(**P**,Prob,x),B)| possible_lung_cancer(yes,Prob,x).

relax(marker(x,marker_type,in_range,**P**,B)).

This rule evaluates for a patient x if a specific biomarker, *marker_type*, found in serum data is within a certain value range for a patient with an age of **A** who is a type **T** smoker (**T** depends on the number of cigarettes or cigars smoked daily). If true, then the diagnosis of possible lung cancer is going to be true with a probability increase of **P** (where **P** is a function of the patient's age, health history, and this particular biomarker presence). But if we relax the requirement of the presence of the biomarker, then the system can evaluate patient records that do not have this particular information and report in the diagnosis listing that this information was not included in the record, which could be valuable information as recommended follow-up tests for that particular patient.

3 References

[1] Canadian Cancer Society. *Lung cancer stats*, August 2003
http://www.bc.cancer.ca/ccs/internet/standard/0,3182,3278_14459_371459_langId-en,00.html

[5] Dahl, V. 2004. Concept Formation Rules: an executable cognitive model of knowledge construction. In *Proc. First International Workshop on Natural Language Understanding and Cognitive Science*, Porto, Portugal, April 2004.

[4] Gealy, R., Zhang, L., Siegfried, J.M., Lutetich, J.D., and Keohavong, P. 1999. Comparison of mutations in the p53 and K-ras genes in lung carcinomas from smoking and nonsmoking women. *Cancer Epidemiology, Biomarkers, & Prevention* 8:297-302.

[3] Marx, J. 2003. Mutant Stem Cells May Seed Cancer. News Focus. In: *Science* 301:1308-1310.

[2] Varner, L., Norris, C. 2002. Lung Cancer: Battling the Number One Cancer Killer. Medicine on the Horizon. In: *ExploreHealth with Sentara* http://12.42.224.152/healthnews/MedicineontheHorizon/moth042002.htm

N10. Motif Preservation in Biochemical Pathways

Zachary M. Saul * [1], Vladimir Filkov [1]

Keywords: network motifs, biochemical pathways, KEGG

1 Introduction.

Large biochemical networks are usually derived and analyzed by focusing on the properties of their sub-networks, which are often functionally important for the system. In doing so, the implicit assumption is that these smaller networks can then be put together in a modular fashion, treating each module as a black box, and that the properties are scale-independent and still hold for the whole system.

It has been recently proposed [3] that subgraphs in biochemical networks which occur more often than expected could be elementary functional building blocks. Such over-represented subgraphs, or *network motifs*, represent systematic properties of networks. Here we ask the above question restricted to network motifs: we sought to compare the incidence of over-represented subgraphs in individual pathways to that of an organism's whole biochemical network.

2 Data and Methods.

We retrieved all forty-eight available biochemical pathways for *E. coli* from the KEGG database [2]. These pathways are models for common biochemical functions and are studied as separate modular entities. They range in size from 6 to 38 nodes. Connected together they comprise a network of 503 nodes. We looked for 3 and 4 node motifs in both the individual pathways and their merge, using the software used by Milo et al. [3].

3 Results and Discussion.

Figure 1 shows for the five most over-represented 3-motifs (in the individual pathways). The figure also shows the number of pathways (out of 48) in which each subgraph was a statistically significant motif. The motifs that scored highest in the merged network are marked with a star, and their corresponding z-scores (in the composite network) are shown below them.

The strongest motifs do in fact correspond to the motifs found in the composite pathway. The results above make the case that separating the *E. coli* network into pathways does not change the network motifs detected, a result that is both intuitive and desirable.

In the future, we intend to test a randomly modularized set of pathways in order to asses if the systemic properties are preserved better within functional units (pathways) or are independent of both function and scale. Additionally, we are studying different organisms and different types of networks looking to establish scale invariance of functional sub-networks as an evolutionary imperative across biological systems.

*Corresponding author. E-mail: saul@cs.ucdavis.edu

[1]Department of Computer Science, 2063 Kemper Hall, University of California, Davis, 1 Shields Ave., Davis, CA, 95616.

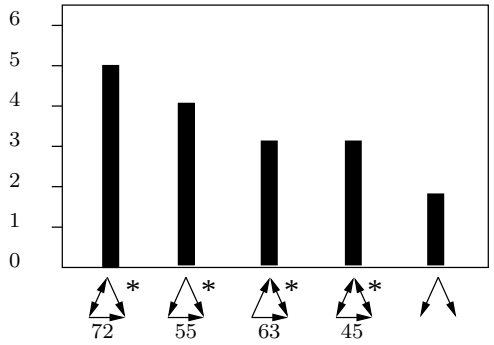


Figure 1: The number of trials out of 48 that each motif scored in the top 20% of subgraphs.

References

[1] Réka Albert and Albert-László Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97.

[2] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya. 2002. The kegg databases at genomenet. *Nucleic Acids Res*, 30:42–46.

[3] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. Network motifs: Simple building blocks of complex networks. *Science*, 298:824–827.

[4] M.E. Newman, S.H. Strogatz, and D.J. Watts. 2001. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64(026118):1–17.