Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Update on the Pfam5000 Strategy for Selection of Structural Genomics Targets

J.-M. Chandonia and S.E. Brenner

Abstract— **Structural Genomics is an international effort to determine the three-dimensional shapes of all important biological macromolecules, with a primary focus on proteins. Target proteins should be selected according to a strategy that is medically and biologically relevant, of good financial value, and tractable. In 2003, we presented the "Pfam5000" strategy, which involves selecting the 5,000 most important families from the Pfam database as sources for targets. In this update, we show that although both the Pfam database and the number of sequenced genomes have increased in size, the expected benefits of the Pfam5000 strategy have not changed substantially. Solving the structures of proteins from the 5,000 largest Pfam families would allow accurate fold assignment for approximately 65% of all prokaryotic proteins (covering 54% of residues) and 63% of eukaryotic proteins (42% of residues). Fewer than 2,300 of the largest families on this list remain to be solved, making the project feasible in the next five years given the expected throughput to be achieved in the production phase of the Protein Structure Initiative.**

I. Introduction

STRUCTURAL genomics aims at the discovery, analysis, and dissemination of three-dimensional structures of protein, RNA, and other biological macromolecules representing the entire range of structural diversity found in nature [1-5]. In the United States, the National Institutes of Health are supporting structural genomics projects at 9 pilot centers through the Protein Structure Initiative (PSI). The primary focus at the PSI centers is on proteins, with targets for experimental structure determination being conducted independently at each center. Beginning later this year, the PSI project moves into a production phase, with the total throughput expected to increase to approximately 600-1,000 protein structures per year. In the production phase, the majority of targets for all centers are expected to be chosen using a more centralized strategy (http://grants2.nih.gov/grants/guide/rfa-files/RFA-GM-05-001.html) [6].

In November 2003, in order to help inform future strategic decisions, we quantified the potential impact that several proposed strategies for target selection will have on our future ability to model the structures of proteins from a wide range of sequenced genomes [7]. We examined genome-centric strategies such solving the structures of all proteins from a single pathogenic genomes [8, 9] or a comparable number of human proteins. We found that these strategies would have little impact on our structural knowledge of other proteomes, since a significant fraction of each proteome is classified in small families, which may have little overlap with other species of interest. As an alternative, we proposed the "Pfam5000" strategy, or choosing targets from a regularly updated index of the 5,000 most important, tractable families in the Pfam database [10] at a given point in time.

In the original report, we projected that successful solution of targets from the largest 5,000 Pfam families would accurate fold assignment for approximately 68% of all prokaryotic proteins (covering 59% of residues) and 61% of eukaryotic proteins (40% of residues). More fine-grained coverage which would allow accurate modeling of these proteins would require an order of magnitude more targets. The Pfam5000 strategy may be modified in several ways, for example to focus on larger families, bacterial sequences, or eukaryotic sequences; as long as secondary consideration is given to large families within Pfam, coverage results vary only slightly.

Since our first analysis, Pfam has grown by over 1,000 families and numerous additional genomes have been sequenced. In this report, we repeat the analysis on the expanded data set in order to examine the stability of our projections. We find that the conclusions in the original study are still valid, despite major changes in the data we analyzed.

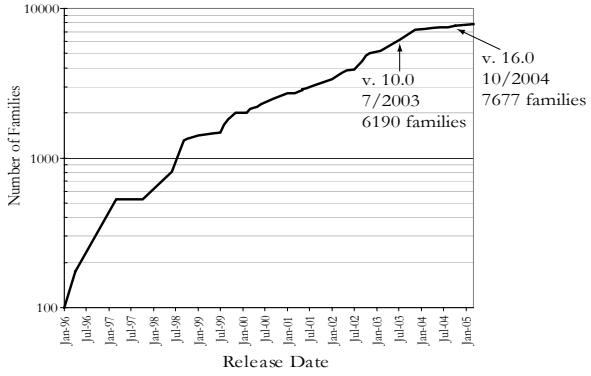

Fig. 1. The Pfam database has been growing exponentially since its inception in 1996. The versions used in the previous and current studies are indicated.

Manuscript received May 9, 2005. This work is supported by grants from the NIH (1-P50-GM62412, 1-K22-HG00056) and the Searle Scholars Program (01-L-116), and by the U.S. Department of Energy under contract DE-AC03-76SF00098.

J.-M. Chandonia is with the Berkeley Structural Genomics Center, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA (e-mail: JMChandonia@lbl.gov).

S. E. Brenner is with the Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA (phone: 510-643-9131; fax: 413-280-7813; e-mail: brenner@compbio.berkeley.edu).

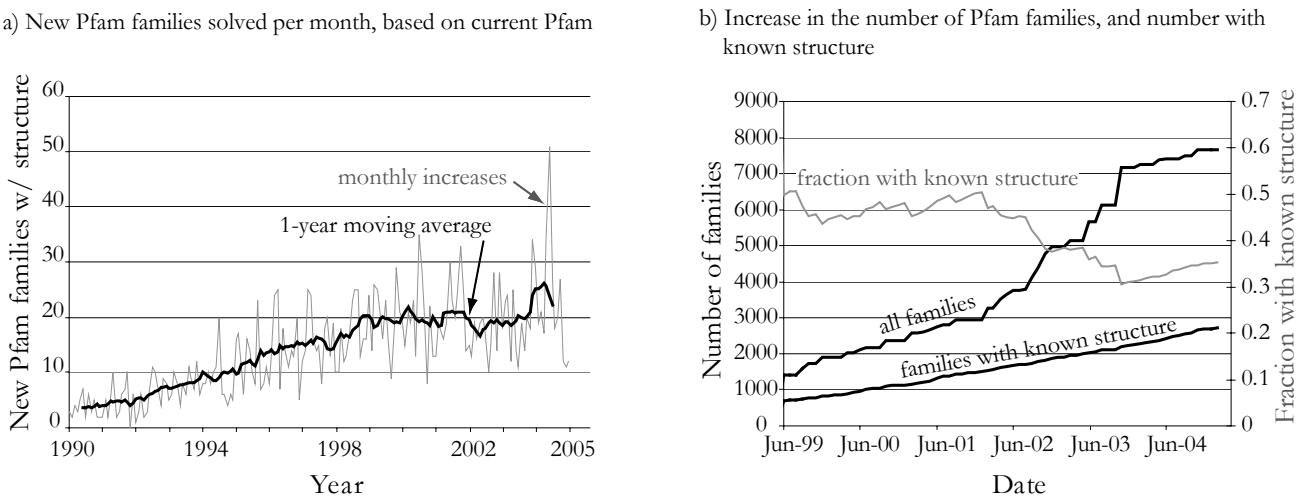

Fig. 2. Increase in structural coverage of Pfam. a) shows how the coverage of the current Pfam-A by known structure has increased over time, based on deposition dates of PDB entries and reported solution dates by structural genomics centers. 141 current Pfam families had known structure prior to 1990. b) shows the total number of annotated Pfam-A families, and the number and fraction of Pfam-A families with a known structure, from release 4.0 in May 1999 until the present.

II. Data and Methods

A. *Pfam*

This study is based on Pfam 16.0, released in October 2004. Our previous study [7] was based on Pfam 10.0, released in July 2003. Fig. 1 shows the growth of Pfam. Although Pfam has been growing exponentially for years, its rate of growth appears to have slowed somewhat in the last several years.

Between versions 10.0 and 16.0 of Pfam, the number of sequences classified has risen from 984,936 to 1,557,840, an increase of 58%. The number of families has risen from 6,190 to 7,677, an increase of 24%. Most (94%) of the new sequences are classified in new families, with the remaining proteins classified as newly sequenced members of previously annotated families. Although the Pfam curators typically add new families in descending order by size, this is not always the case: 349 families were added between Pfam 10.0 and Pfam 16.0 that were larger than the median family size (33 sequences) in Pfam 10.0. A much larger number (1,203) of families at or below the median size were added. A small number of families (65) present in Pfam 10.0 were removed or merged with other families before the release of Pfam 16.0.

B. *The Pfam5000*

The specific families chosen for the Pfam5000 should be based on their importance and tractability. The simplest definition for importance is size: the number of proteins that belong to a family may be taken as a proxy for its significance. As we showed in the original study [7], so long as size is a secondary criterion in the current Pfam database, the selected set of proteins is relatively insensitive to a wide variety of primary criteria, including presence of representatives in particular proteomes of interest. In the current analysis, we present results on a version of the Pfam5000 biased towards currently solved structures. This means that the Pfam5000 includes all families with a structure currently solved. Families without a solved structure are added in decreasing order by size to fill out the remaining families.

C. *Proteome Database*

In the original study, we measured progress on structural completion of proteomes from the Proteome Analysis database at the European Bioinformatics Institute [11]. This database has since been merged into the Integr8 database [12]. Integr8 version 12 was used in this analysis.

The Integr8 database contains annotations for 238 complete proteomes, including 19 eukaryotes. This is almost double the number of proteomes we analyzed in our previous study of 131 species and 10 eukaryotes. Species added include several new types of yeast, the malaria parasite, and partial annotation of the zebrafish. The proteome for each organism is composed of proteins curated from the Swiss-Prot and TrEMBL databases. All proteins were annotated with hidden Markov models [13, 14] from the InterPro [15] database. Since InterPro includes models from Pfam, we used the supplied InterPro annotations to map Pfam domains onto each protein. The version of InterPro used to annotate Integr8 version 12 includes Pfam 16.0

D. *Measures of Progress*

The ultimate goal of structural genomics is to provide structural information for the complete repertoire of biological macromolecules. We measure progress towards that goal as "coverage." Coverage of a proteome is the fraction of its sequences or residues that are covered. Per-sequence coverage is measured as the fraction of sequences that have at least one domain that belongs to a family with a

TABLE I
COVERAGE OF SELECTED PROTEOMES BY CURRENTLY KNOWN STRUCTURES

Organism	Number of Proteins	Current Coverage			Coverage from original study		
		Families	% Proteins	% Residues	Families	% Proteins	% Residues
A. thaliana	26,149	1369	52.4	31.4	1147	47.8	27.5
B. rerio (partial)	9,249	1236	67.3	45.4	*not included in original study*		
C. elegans	21.926	1301	41.2	29.5	1102	36.5	25.0
C. glabrata	5180	1057	51.5	30.2	*not included in original study*		
D. melanogaster	19,474	1335	49.9	30.3	1128	46.1	27.3
E. coli	4,338	1164	58.1	56.0	969	51.0	49.2
H. sapiens	38,492	1562	52.3	34.1	1292	45.4	29.7
K. lactis	5,307	1084	49.1	31.0	*not included in original study*		
M. jannaschii	1,782	588	50.2	45.3	503	42.7	38.6
M. musculus	32,226	1539	56.1	37.4	1288	52.5	35.3
M. tuberculosis	3,947	922	51.5	47.8	804	47.9	43.1
M. genitalium	486	337	65.4	54.1	302	60.9	47.4
M. pneumoniae	687	355	55.2	46.1	319	46.1	38.1
P. falciparum	5,251	782	35.2	15.8	*not included in original study*		
S. cerevisiae	6,218	1080	47.8	30.3	923	43.1	26.6
S. pombe	4,956	1092	55.1	33.7	932	48.8	29.3
Prokaryote Average			51.3	43.8		46.5	38.3
Eukaryote Average			51.7	33.1		47.2	30.4

Coverage of proteomes by structural biology efforts to date, compared to the original Pfam5000 study. Number of Proteins is the number annotated in Integr8. Families are the number of Pfam families in each proteome that also contain a known structure. Averages are weighted by the total size of the proteome, and are based on all proteomes in Integr8.

representative whose structure is to be experimentally characterized; this would allow the relevant domain to have its fold assigned [16-18].

To identify Pfam families with currently known structures, we ran all Pfam-A models against our database of sequences of known structure. This database includes sequences of all proteins currently in the PDB [19], sequences of proteins on hold in the PDB (when made available by the authors), and sequences of proteins reported as solved by structural genomics centers in the TargetDB database [20]. This database was updated on 21 March 2005. The "trusted cutoff" for each Pfam family was used as a cutoff for determining which hits were significant.

Per-residue coverage by Pfam families was calculated using the beginning and end residues annotated in the Proteome Analysis databases. All residues between the endpoints were annotated as part of the matching family, ignoring any potential gaps. In the original report [7], we tested several variant methods of calculating per-residue coverage of proteomes. The simplest method is to divide the total number of residues in regions matched by Pfam by the total length of all proteins in the proteome. However, we found it more informative to measure progress on the portion of the proteome expected to be interesting and tractable to study under high-throughput experimental conditions. In the latter calculation of per-residue coverage, we ignore unmatched regions of the proteome predicted to be in transmembrane regions, low complexity regions, or coiled coil, as well as short regions of fewer than 50 consecutive unmatched residues between transmembrane regions and/or Pfam hits. This variant of per-residue coverage was used in the current study.

The "seg" program [21] (version dated 5/24/2000) was run on all sequences in Integr8 to identify putative low complexity regions. The "ccp" program [22] (version dated 6/14/1998) was used to predict coiled coil regions in all sequences, and TMHMM 2.0a [23] was used to predict the locations of transmembrane helices. Default options were used for all programs.

III. RESULTS

A. Known Structures in Pfam

Currently, 2729 of 7677 (36%) of Pfam families have at least one member of known structure. We identified the earliest structural representative from each family using the deposition dates of the structures. We found that the rate of structural characterization of new families rose steadily throughout the 1990s, but has leveled off at around 20 new families per month since 1999 (Fig. 2a), even as the total number of structures solved continues to increase exponentially. The fraction of Pfam families with known structure (Fig. 2b) has varied over recent years from approximately 50% to as low as 30%, and is currently increasing as new families are being solved faster than the rate at which new families are classified in Pfam.

B. Increase in Proteome Coverage

We calculated the coverage of proteomes in Integr8 by families with currently known structures. Results for selected proteomes are given in Table I, and results for all 238 proteomes are available from the authors upon request. Several results are apparent from the table. First, structural coverage of most prominent organisms has increased by 4 to 7% in the 1.5 years since the original study. Second, the average coverage of prokaryotes has increased by 4.8%, from 46.5% of all proteins to 51.3%., while the percentage of residues covered has increased from 38.3% to 43.8% (a

TABLE II
PROJECTED COVERAGE OF SELECTED PROTEOMES BY PFAM5000

Organism	Number of Proteins	Current Projection			Original Projection		
		Families	% Proteins	% Residues	Families	% Proteins	% Residues
A. thaliana	26,149	2043	69.7	43.8	2008	69.2	42.9
B. rerio (partial)	9,249	1750	78.7	54.2	*not included in original study*		
C. elegans	21.926	1898	56.7	40.1	1844	53.7	37.4
C. glabrata	5180	1463	62.8	37.8	*not included in original study*		
D. melanogaster	19,474	1967	60.9	37.4	1890	59.9	36.0
E. coli	4,338	1685	76.7	69.6	1590	74.2	67.3
H. sapiens	38,492	2288	61.2	41.5	2224	56.7	38.8
K. lactis	5,307	1500	60.1	38.9	*not included in original study*		
M. jannaschii	1,782	766	63.5	57.2	781	64.7	58.3
M. musculus	32,226	2250	65.8	45.2	2225	64.8	45.1
M. tuberculosis	3,947	1219	67.6	60.0	1153	66.3	57.0
M. genitalium	486	384	73.7	60.7	374	74.9	58.8
M. pneumoniae	687	407	61.6	52.2	397	70.0	54.5
P. falciparum	5,251	1048	46.0	21.7	*not included in original study*		
S. cerevisiae	6,218	1501	59.5	38.2	1448	57.7	37.6
S. pombe	4,956	1505	66.8	41.6	1466	64.6	40.4
Prokaryote Average			65.1	54.2		63.9	51.7
Eukaryote Average			63.4	41.5		61.6	41.2

Projected coverage of proteomes by the current Pfam5000 families, compared to the original Pfam5000 study. Number of Proteins is the number annotated in Integr8. Families are the number of Pfam5000 families that contain at least one member in the proteome. Averages are weighted by the total size of the proteome, and are based on all proteomes in Integr8.

5.5% increase). The average increase is lower than for the organisms listed in the table, because the average includes many proteomes sequenced in the past several years. These organisms tend to have been less well studied than the prominent organisms listed in the table (e.g., *E. coli*).

Coverage figures for most prominent eukaryotes listed in the table have increased by 4 to 6% over the past 1.5 years. The average coverage of eukaryotes has increased by 4.5%, from 47.2% of all proteins to 51.7%., while the percentage of residues covered has increased from 30.4% to 33.1% (a 2.7% increase). The proteome of the malaria parasite, *P. falciparum*, stands out as particularly poorly covered by currently known structures. As observed in the original study, per-residue coverage of eukaryotes is lower than for prokaryotes even at comparable per-protein coverage; this difference is presumably due to eukaryotes having more multi-domain proteins.

C. Stability of Coverage Projections

Table II shows projected coverage by the families in the current Pfam5000, compared to the 5,000 families from the original study 1.5 years ago. Both sets were biased towards known structures; thus, the current Pfam5000 is composed of the 2,729 families with currently known structures, along with the 2,271 largest families that do not have a current structural representative. The original set contained 2,108 families of known structure and the remaining 2,892 largest families.

The stability of the projections is quite good: for the prominent organisms listed in the table, projected coverage at the completion of structural characterization of the Pfam5000 has generally increased by 1-2%, with only one organism (*M. jannaschii*) showing a projected decrease. The slight increase in most projections is due to the classification of additional large families in Pfam: these large families replaced smaller families in the Pfam5000, allowing slightly more coverage projected upon completion.

More importantly, the average projection for coverage across a wide range of prokaryotic and eukaryotic organisms has remained stable, despite a near-doubling (131 to 238) in the number of sequenced organisms we analyzed.

In the previous study, we reported unweighted statistics on average projected coverage for organisms. Since that time, a number of very small eukaryotic genomes have been sequenced, including *Cryptococcus neoformans* and *Guillardia theta*. We therefore report statistics weighted by proteome size (the number of proteins for per-protein statistics, and the number of residues for per-residue statistics). We expect completion at least one structure from each of the current Pfam5000 families to allow accurate fold assignment for approximately 65% of all prokaryotic proteins (covering 54% of residues) and 63% of eukaryotic proteins (42% of residues).

IV. DISCUSSION

The families in Pfam5000 represent a tractable yet extremely useful set of targets to study in the production phase of the PSI. If all structures in Pfam5000 were solved, we would have at least partial information on the folds of nearly 2/3 of all then-known proteins. This goal appears to be feasible within the next five years, especially as the number of families of known structure has already increased by over 600 in the 1.5 years since our initial study. Completion of the largest 5,000 families would have a broad impact on our structural understanding of both currently sequenced genomes and on genomes yet to be sequenced. If modeling and threading methods enjoy similar advances in

the next five years, we will be able to produce accurate models for these proteins as well as fold assignments.

Given that the projections appear stable even as the number of sequenced proteomes continues to increase, further prioritization of the families would be useful. We showed in our original study that the projected coverage results are relatively insensitive to specific secondary methods of prioritization, as long as family size remains the main criterion. Further prioritization of the large families most likely to yield significant biomedical breakthroughs would also be of great interest.

Another major problem not yet examined is prioritization by tractability. Most groups assume that low complexity, coiled coil, and transmembrane proteins are less tractable to study by high throughput experimental methods. It would be useful to prioritize the Pfam5000 families according to predicted tractability, in order to get the most results out of our current state-of-the-art technology for high throughput experimentation. Some methods to prioritize individual family members for high-throughput experimentation have already been developed (e.g., prioritizing proteins from thermophiles and halophiles), and further systematic development of these methods will also be essential to achieving further breakthroughs in experimental throughput.

REFERENCES

[1] S. K. Burley and J. B. Bonanno, "Structural genomics," *Methods Biochem Anal*, vol. 44, pp. 591-612, 2003.

[2] T. L. Blundell and K. Mizuguchi, "Structural genomics: an overview," *Prog Biophys Mol Biol*, vol. 73, pp. 289-95, 2000.

[3] S. E. Brenner, "A tour of structural genomics," *Nat Rev Genet*, vol. 2, pp. 801-9, 2001.

[4] G. T. Montelione, "Structural genomics: an approach to the protein folding problem," *Proc Natl Acad Sci U S A*, vol. 98, pp. 13488-9, 2001.

[5] M. R. Chance, A. R. Bresnick, S. K. Burley, J. S. Jiang, C. D. Lima, A. Sali, S. C. Almo, J. B. Bonanno, J. A. Buglino, S. Boulton, H. Chen, N. Eswar, G. He, R. Huang, V. Ilyin, L. McMahan, U. Pieper, S. Ray, M. Vidal, and L. K. Wang, "Structural genomics: a pipeline for providing structures for the biologist," *Protein Sci*, vol. 11, pp. 723-38, 2002.

[6] "PSI-phase 1 and beyond," *Nat Struct Mol Biol*, vol. 11, pp. 201, 2004.

[7] J. M. Chandonia and S. E. Brenner, "Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches," *Proteins*, vol. 58, pp. 166-79, 2005.

[8] A. Matte, J. Sivaraman, I. Ekiel, K. Gehring, Z. Jia, and M. Cygler, "Contribution of structural genomics to understanding the biology of Escherichia coli," *J Bacteriol*, vol. 185, pp. 3994-4002, 2003.

[9] C. W. Goulding, M. Apostol, D. H. Anderson, H. S. Gill, C. V. Smith, M. R. Kuo, J. K. Yang, G. S. Waldo, S. W. Suh, R. Chauhan, A. Kale, N. Bachhawat, S. C. Mande, J. M. Johnston, J. S. Lott, E. N. Baker, V. L. Arcus, D. Leys, K. J. McLean, A. W. Munro, J. Berendzen, V. Sharma, M. S. Park, D. Eisenberg, J. Sacchettini, T. Alber, B. Rupp, W. Jacobs, Jr., and T. C. Terwilliger, "The TB structural genomics consortium: providing a structural foundation for drug discovery," *Curr Drug Targets Infect Disord*, vol. 2, pp. 121-41, 2002.

[10] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy, "The Pfam protein families database," *Nucleic Acids Res*, vol. 32 Database issue, pp. D138-41, 2004.

[11] M. Pruess, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, E. Kriventseva, V. Mittard, N. Mulder, I. Phan, F. Servant, and R. Apweiler, "The Proteome Analysis database: a tool for the in silico analysis of whole proteomes," *Nucleic Acids Res*, vol. 31, pp. 414-7, 2003.

[12] P. J. Kersey, L. Morris, H. Hermjakob, and R. Apweiler, "Integr8: enhanced inter-operability of European molecular biology databases," *Methods Inf Med*, vol. 42, pp. 154-60, 2003.

[13] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler, "Hidden Markov models in computational biology. Applications to protein modeling," *J Mol Biol*, vol. 235, pp. 1501-31, 1994.

[14] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics*, vol. 14, pp. 755-63, 1998.

[15] N. J. Mulder, R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. Sigrist, R. Vaughan, and E. M. Zdobnov, "The InterPro Database, 2003 brings increased coverage and new features," *Nucleic Acids Res*, vol. 31, pp. 315-8, 2003.

[16] S. E. Brenner and A. Berry, "A quantitative methodology for the de novo design of proteins," *Protein Sci*, vol. 3, pp. 1871-82, 1994.

[17] S. E. Brenner, "Target selection for structural genomics," *Nat Struct Biol*, vol. 7 Suppl, pp. 967-9, 2000.

[18] D. Baker and A. Sali, "Protein structure prediction and structural genomics," *Science*, vol. 294, pp. 93-6, 2001.

[19] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," *Nucleic Acids Res*, vol. 28, pp. 235-42, 2000.

[20] L. Chen, R. Oughtred, H. M. Berman, and J. Westbrook, "TargetDB: a target registration database for structural genomics projects," *Bioinformatics*, 2004.

[21] J. C. Wootton, "Non-globular domains in protein sequences: automated segmentation using complexity measures," *Comput Chem*, vol. 18, pp. 269-85, 1994.

[22] A. Lupas, "Prediction and analysis of coiled-coil structures," *Methods Enzymol*, vol. 266, pp. 513-25, 1996.

[23] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *J Mol Biol*, vol. 305, pp. 567-80, 2001.