ELSEVIER

Sequences and topology: a decade of genomes

Editorial overview

Steven E Brenner and Anna Tramontano

Current Opinion in Structural Biology 2005, **15**:245–247

Available online 1st June 2005

0959-440X/$ – see front matter
Published by Elsevier Ltd.

DOI 10.1016/j.sbi.2005.05.009

Steven E Brenner

Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720-3102, USA
e-mail: brenner@compbio.berkeley.edu

Anna Tramontano

Pasteur Institute-Cenci Bolognetti Foundation, Department of Biochemical Sciences, University of Rome ''La Sapienza'', P. le Aldo Moro, 5 00185 Rome, Italy
e-mail: anna.tramontano@uniroma1.it

It has been a decade since the first complete genome sequences were revealed. We have taken this anniversary as an opportunity to reflect on what has been accomplished, as well as the most significant challenges for the future.

We are deluged with millions of nucleotide and inferred protein sequences, but where are we in terms of unraveling their biological meaning? How much have we learned from them? How many unexpected problems have we encountered? The reviews in this section try to answer these and other questions brought about by the genomic revolution by outlining the discoveries, breakthroughs and challenges of the past decade and highlighting the challenges ahead.

The section begins with a review by Doolittle, who highlights some of the most dramatic insights that complete genomes have offered. Most obviously, the deluge of sequences has provided a cornucopia of new gene and inferred protein sequences, which have yielded millions of proteins for individual study and, in aggregate, have offered demographics of the protein universe. With the entire repertoire of genes from numerous species, it has been possible to construct superior phylogenetic trees reflecting the histories of extant species. The data have allowed us to better imagine the common ancestor at the root of this tree, and to trace gene loss and gene flow amongst the lineages. Comparative studies, especially of reduced genomes, have provided increasingly refined estimates of what genes would be required to create truly minimal genomes. However, genomes have created new questions and have left others as yet unanswered. We do not yet know the root of the evolutionary tree of all organisms nor can we be certain of many branchings. Similarly, the history and role of introns remain unresolved. Most glaringly, the huge number of ORFans — apparent genes with no identifiable homology to any others — highlights the limits of our knowledge even of individual genes, much less their interactions and regulation. For all their bounty, complete genomes herald more exciting discoveries ahead.

The primary tool for understanding genomes is sequence database searching, as reviewed by Pearson and Sierk. Although the basic idea of sequence comparison dates back 50 years ago [1], genomic sequence analysis required refinements in the reliable statistical scoring of gapped local alignment that have only become available in the past decade. A flurry of more sophisticated methods, including hidden Markov models, have enhanced abilities to detect distant evolutionary relationships. Critical to this enterprise has been the development of effective means to evaluate sequence comparison methods. Although these show the consistently increasing power of sequence analysis, Pearson and Sierk write, sequence comparison still fails

to find many ancient homologs identified by protein structure and even automated structure alignment programs cannot compete with manual studies for distinguishing homologs from analogs. This review thus offers a clarion call for enhancements in homology detection approaches, even as the imperfect current approaches have become ubiquitous and essential for molecular biology.

Once a set of homologous proteins are in hand, multiple alignment typically offers the most effective means of understanding the sequences' significance and history. Bizarrely for such a computationally accessible data set, the process of optimally matching characters has long been beyond the ken of automated programs, which typically provided only a rough alignment that could be readily enhanced by manual inspection and refinement. Wallace, Blackshields and Higgins review recent developments that may finally allow programs to succeed alone. They briefly describe a new generation of programs that are dramatically more accurate and efficient than their predecessors, as well as the new assessment techniques that have provided confidence in the programs' abilities. Hearkening to a future in which most proteins are homologous to a known protein structure, the review also introduces a method for multiple alignment enhanced by structural knowledge.

The review by Valencia contains an insightful discussion about the effect of the data deluge on the reliability of the information contained in biological databases. Functional information on the large majority of proteins is extrapolated from a very limited set of known cases, as we have gathered experimental data only on a very tiny fraction of the protein universe. Most of the proteins in our databases have a functional annotation inherited from the functional annotation of a related protein, which, in turn, has been derived on the basis of an inferred evolutionary relationship with a third protein and so on. This is an error-prone process that becomes more dangerous as the amount of data increases, because it becomes impossible for experts to carefully analyze the results of the computational methods. Moreover, the intrinsic complexity of biology adds further complications. For example, enzyme substrate specificity is very poorly conserved during evolution, even among proteins that share as much as 70% sequence identity. As there is no magic solution to the problem, we have to trade quality for quantity, if we want to investigate complete genomes and biological systems. However, as Valencia points out, we should at least assign a reliability value to functional predictions, on the basis of what we have understood about the limits of evolution-based inference of function. This is an important take-home message for developers and users of biological computational tools.

Once upon a time, the structure determination of a protein represented the final step of its characterization, coming after the protein had been isolated on the basis of its function and extensively characterized: the X-ray structure was used to gain insight into the detailed mechanism of the catalyzed reaction, or the specific recognition pattern between the molecule under study and its cognate molecules. But the '-omic' revolution has also affected this aspect of biochemistry: structural genomics projects, aimed at determining the structure of as many proteins as possible, are flourishing and producing structures of proteins whose function is yet to be discovered. In the fifth review, Watson, Laskowski and Thornton discuss what we can do when we are presented with 'structures without a history'. These cases do represent a sizeable fraction of the protein structure database. They review methods based on fold matching (whereby the structure of the protein is used to attempt the detection of evolutionary relationships impossible to highlight by the comparison of their sequences alone), on the identification of clefts and binding pockets on the surface of the proteins, and on machine learning techniques. The very existence of methods based on different criteria suggests that it might be sensible to combine different methods and derive a 'consensus' prediction. Thornton and co-workers, who have been working in this area, indeed conclude that, following this route, the accuracy of function assignment is higher.

As discussed in the review by Thornton and colleagues, structure-based methods are essential ingredients in the quest for the molecular function of unknown proteins. However, even though the field of structure determination has made impressive progress, the experimental elucidation of the structure of a protein is still a lengthy and resource-intensive process. Therefore, we would like to infer (or predict, as we usually say) the three-dimensional structure of a protein given its amino acid sequence. In 1994, John Moult proposed a worldwide experiment named CASP (Critical Assessment of Protein Structure Prediction), aimed at establishing the current state of the art in protein structure prediction, identifying what progress has been made and highlighting critical future research needs. Every two years, structural biologists who are about to solve a protein structure are asked to make the sequence of the protein available, together with a tentative date for the release of the final coordinates. Predictors produce and deposit models of these proteins before the structures are made available, and a panel of assessors compares the models with the experimentally solved structures. The exercise provides a detailed evaluation of model quality, as well as conclusions about the state of the art of the different methods. The results are discussed in a meeting where assessors and predictors convene, and the conclusions are made available to the whole scientific community. There have been CASP experiments since 1994 and there is no sign of a decrease in interest in the experiment, with more than 200 groups taking part in the last challenge. The lessons

of these ten years are the subject of the review by Moult. Historically, methods for predicting protein structure are distinguished according to the relationship between the target protein(s) and proteins of known structure. Comparative modeling can be applied when a clear evolutionary relationship between the target and a protein of known structure can be easily detected from sequence. We catalogue as 'fold recognition' those methods that can be applied when the structure of the target protein turns out to be related to that of a protein of known structure. Finally, when neither the sequence nor the structure of the target protein are similar to that of a known protein, we classify the methods as techniques for new fold prediction. This review describes the bottlenecks of all of these approaches and, interestingly, concludes that better refinement techniques at the atomic level would provide improvements over the spectrum of available methods. Indeed, although this is not mentioned in the review, some methods adopted the strategy of funneling into each subsequent step both the optimal and suboptimal intermediate results, and evaluating the final models at the atomic level. They seem to produce better results, suggesting that it is easier to evaluate the quality of a full atomic model with respect to the reliability of each of the intermediate steps of the procedure.

The next review, written by Lecomte, Vuletich and Lesk, is a beautiful example of how much can be learned through a low-throughput, careful and manual analysis of protein structures. It reminds us that, even if an expert cannot cope with the impressive amount of available data, he or she can still select a representative example and use it to derive insights into the intriguing and fascinating process of protein structure and evolution. Even though they have been studied for decades, globins keep surprising us, by appearing in different forms and functions, hinting at, and helping to unravel, complex biological mechanisms of general validity. This review demonstrates that ''There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy''.

One of the most surprising discoveries of the past decade has been the importance of RNA molecules, as catalysts, regulators and structural elements. Holbrook shows how our understanding of these fascinating molecules has been enhanced by structural studies. The flurry of new RNA structures continues unabated, but we are now increasingly able to interpret these in the context of recurrent motifs. As Holbrook explains, however, new structures continue to offer surprises, and also have suggested re-evaluation and new insights into structures solved years ago. Reliable secondary structure prediction of large molecules remains out of reach but is being increasingly enhanced by the availability of homologous sequences, whereas fully automated tertiary structure prediction is only now emerging as a possibility. Even as individual RNAs are revealing their secrets and commonalities between distinct structures are being interpreted, Holbrook calls for further computational and experimental studies to bring our understanding of the repertoire, role and mechanism of RNA biology to the genomic scale.

Together, these reviews demonstrate the power of interdisciplinary research, whereby traditionally different fields, such as protein sequence analysis, structure prediction and structure analysis, all come together to try to face the new challenges posed by the genomic era. These ten years have been exciting, challenging and intellectually stimulating. These reviews tell us that the fun is not over yet and that many more fascinating discoveries lie ahead of us.

Reference

1. Zuckerkandl E, Pauling L: **Molecules as documents of evolutionary history**. *J Theor Biol* 1965, **8**:357-366.