Sequences and topology Genes and structures in context

Editorial overview

Peer Bork and Christine A Orengo

Current Opinion in Structural Biology 2004, **14**:261–263

0959-440X/$ – see front matter
© 2004 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.sbi.2004.05.008

Peer Bork

EMBL, Meyerhofstrasse 1, 69012 Heidelberg,
Germany
e-mail: bork@embl.de

Peer's work centres on function prediction, mostly by comparative analysis. This involves methodologies ranging from genome annotation and protein network analysis to the automatic retrieval of information from literature. Concentrating on sequence analysis in the past, he now uses various other large-scale data sets to analyse molecular and cellular aspects of function and evolution.

Christine A Orengo

Biomolecular Structure and Modelling Unit,
Department of Biochemistry and Molecular
Biology, University College London,
Gower Street, London WC1E 6BT, UK
e-mail: orengo@biochemistry.ucl.ac.uk

Christine works on the analysis and prediction of evolutionary relationships using largely structural and sequence data. To enable these analyses, research in the group has also focused on the development of algorithms for structural classification and the clustering of complete genomes into families. These resources are now being exploited in the mining of functional genomics data.

Recent years have focused attention on many novel aspects of proteins and their structures, some of which are highlighted by the following reviews. They range from molecular aspects, such as alternative splicing (which sometimes affects only a single amino acid), to the cellular layout of macromolecular assemblies. The reviews all reveal the shift in the field to include qualitatively novel data in the analysis and to merge disciplines, be it the introduction of chemical thinking to biological entities or the combination of molecular and cellular techniques to uncover the structural framework of a cell. So, it is novel context that brings new roles for sequences and structures, and increases the impact of the subsequent research.

However, one needs some kind of coordinate system to bridge this increasingly interdisciplinary field and integrate heterogeneous data. Genomes are well suited to this task, but, although we know the sequences of many of them, there are still problems with their proper interpretation. One of the most important and pressing needs is to know the complete and correct set of genes encoded by a genome. After all, genes are the parts lists from which interaction networks and cellular assemblies are built.

Despite progress, particularly in higher eukaryotes, gene identification remains a tough job. Brent and Guigó describe the state of the art in gene prediction and the impact of improvements from comparative genome analysis. Dual-genome predictors currently offer the greatest hope for improvements in the short term and the application of pseudogene detection methods has been important in eliminating many false positives. Multiple-genome efforts will hopefully also yield improvements within a few years, especially those that model variations in evolutionary rates for different regions of the genome.

However, significant challenges remain. One of these challenges is the identification of all the transcripts that a single gene can encode and the context in which particular splice forms are expressed. Brenner and colleagues review recent progress in the detection of alternative splicing and also provide the emerging biological context. Alternative splicing affects at least 50% of human genes. As well as increasing proteome diversity, many new biological roles have been shown or postulated for alternative splicing, for example, in the regulation of gene expression. The authors present some intriguing data illustrating mechanisms whereby alternative splicing regulates gene expression by splicing transcripts into unproductive mRNAs targeted for degradation. Interestingly, this regulation, which is conserved across species and protein families, may also provide a mechanism for regulation of alternative splicing. Thus, minimal input could trigger a cascade of regulated splicing effects, thereby effecting a large change in the proteome.

The context of expression can only be understood if regulatory networks leading to expression behaviour are identified and analysed. Progress in array technology and comparative genomics are only two cornerstones that enable the identification and reconstruction of regulatory gene networks. Teichmann, Gerstein and co-workers give an overview of the structure of current regulatory networks and introduce an evolutionary angle that has been revealed by the integration of various large-scale data available for several species. For example, such analyses reveal that, although transcription factors are drawn from a small set of ancient conserved superfamilies, their relative abundance shows dramatic variation among different phylogenetic groups. Large portions of these networks have clearly arisen through extensive duplication of transcription factors and targets, often with inheritance of regulatory interactions from the ancestral gene.

In close relation and complementary to regulatory gene networks are protein interaction networks — the subject of the review by Bork, Marcotte and co-workers. The considerable attention focused on these networks over the past few years has revealed that they are all small world, scale free and modular. Although the well-studied baker's yeast is still the preferred model organism for an increasing number of protein network studies, the authors show that attention is moving to higher eukaryotes and is about to approach human. In order to be efficient for this organism, reliable knowledge bases have to be assembled to guide predictive efforts and to ensure quality control for both theoretical and experimental efforts. Encouragingly, results suggest that the reconstruction of networks can be improved by combining all the available data, even noisy data, provided that it is integrated appropriately.

Partially overlapping with protein interaction networks are metabolic networks, although in this case a crucial intermediate, the metabolite, plays an essential role. These, mostly chemical, compounds cannot be seen in genomes, just the enzymes that catalyse the respective metabolic reactions. This has led, in the past, to an enzyme- or protein-centric view of pathways or metabolic networks, as enzymes identified in genome projects were seen as good markers for the respective pathways. However, different species can have different chemical solutions for the synthesis and degradation of metabolites, and enzymes can switch substrates. Hatzimanikatis *et al.* discuss current views from a chemical perspective, but also connect enzymes and metabolite structures. Combining these chemical and structural perspectives will enhance our understanding of metabolic networks and could help improve the prediction of enzyme function. It might also guide us in the design of novel pathways for medical and biotechnological purposes.

A structural understanding of the protein universe, in even more global terms, is the goal of many structural proteomics projects. These aim to provide structural representatives and models for all major protein families and families of special biological or medical interest. Crucial in such projects is again the interplay between theoretical understanding and experimental protocols. Thus, Godzik and co-workers concentrate on the role of fold prediction methods to enable better target selection for the large-scale determination of three-dimensional protein structures. Improvements in structural modelling techniques are critical for reducing the number of necessary structural targets. This is an important consideration as, to date, only about 600 structures have been solved by these initiatives. Encouragingly, though, by applying current modelling and prediction techniques to small bacterial genomes, nearly 70% of the sequences can now be associated with a good or at least a 'fuzzy' structural model, which may help assign function.

The more structures that become available, the more it is important to consider their context. Structures often function in complexes and have well-defined surfaces for interacting with other structures within the complex. Structures of large complexes or transient interactions are often difficult to capture and thus modelling of complexes, as well as determining the structural basis of protein interactions in general, is becoming an important task. Russell, Sali and co-workers discuss the interplay between protein interaction networks, as described above, and three-dimensional protein structures. In order to obtain real structure-based networks in which individual interactions can be translated into molecular docking events, iterations of experimental and computational approaches are needed, and compromises between accuracy and coverage have to be made. Some notable successes have been achieved in docking component structures into electron microscopy maps by adapting homology modelling techniques to help improve fit. As the structural proteomics initiatives expand to sample novel protein–protein and domain–domain interactions, we can expect further improvements in these approaches.

To explore an interaction network comprising three-dimensional protein complexes and exploit this as a bridge to the cellular level, it would be best to use the components to build up an entire three-dimensional cell. For this, a novel coordinate system is needed beyond the genome: the structural framework of an entire cell. Frangakis and Förster describe the current status of cryo-electron tomography, which is now approaching this goal by using lower resolution three-dimensional images of entire cells to place protein complexes and large structural assemblies into a cellular framework. Once this is achieved, such a knowledge base would us allow to take the cell and focus, via protein complexes, in on tiny amino acid changes and any functional consequences caused by alternative splicing, which can then be pinpointed to the genome. This is obviously only a dream at present, but we are probably not that far away from seeing the impact of cellular context information on sequences and structures, and vice versa.

Taking all these reviews together, one key message that emerges is the importance of considering and expanding the context of the data that we obtain. As the parts lists of sequences (genes) and three-dimensional structures draw closer to completion, it is crucial to connect them with each other, and also with chemical and cellular knowledge to be able to understand the cell as a system and to navigate through it.