Sequences and topology
Genome and proteome informatics
Editorial overview

Peer Bork* and David Eisenberg†

Addresses
*European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69012
Heidelberg, Germany and Max-Delbrück-Center for Molecular
Medicine, Berlin-Buch, Germany; e-mail: bork@embl-heidelberg.de
†UCLA-DOE Laboratory of Structural Biology and Molecular Medicine,
Molecular Biology Institute, University of California Los Angeles,
Box 951570, Los Angeles, CA 90095-1570, USA;
e-mail: david@mbi.ucla.edu

Current Opinion in Structural Biology 2000, **10**:341–342

0959-440X/00/$ – see front matter

The paradigm of research in biological sciences has changed fundamentally in the late 1990s; whereas before the subjects of study at the molecular level were genes and proteins, now they are genomes and proteomes. Whereas biochemists used to focus on functions of individual genes and proteins, they can now study function at a more global level. The change has come from the advent of various high-throughput technologies. The most advanced of these is large-scale sequencing, which allows the determination of the complete genetic information of entire organisms.

The result is the exponential growth in the publication of complete prokaryotic genomes. In addition, deciphering the genetic information of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster* has opened the door for comparative sequence analysis of eukaryotic genomes, which will be crucial for understanding the human genome (reviewed by Fraser and co-workers, pp 343–348). In 2000, we can expect a continued exponential increase in the determination of entire genomes, with probably more than 100 genomes sequenced and probably more than 50 publicly accessible.

All these data beg for methods that are able to catalog and synthesize the information into biological knowledge. After all, this information is vital for bridging the gap between the genotype and the phenotype. However, many practical and theoretical problems must be solved before one can move on to complex topics such as the understanding of cellular networks. For example, annotation remains a serious practical problem, as reviewed by Lewis, Ashburner and Reese (pp 349–354) for the genome of *Drosophila*. The problem of finding the genes in higher eukaryotes is far from being solved. Probably, novel genomes, more than novel algorithms, will increase the accuracy of finding genes, because knowledge of conserved regions will help greatly. Even more challenging is the functional characterization of the predicted gene products. Beyond the problem that the term *function* is only loosely defined (what are we aiming for?) and is applied in very different contexts, common terms for individual functional features need to be defined and applied. Furthermore, for the vast amount of incoming sequence data, function is currently annotated almost entirely via inferences made by homology. The extent to which this is justified is not clear yet.

Despite all the new information from sequenced genomes, the evolution of genomes and proteomes remains puzzling. Doolittle (pp 355–358) summarizes ideas on the subject that arose from genome-wide analyses, but opinions remain diverse. Massive lateral (horizontal) gene transfer might hamper our current analysis and some researchers conclude that there may not even have been a last common ancestor. Doolittle also notes (pp 355–358 and references therein) that, in evolutionary biology, much debate is centered around basic terms that are not well defined, such as 'homology' and 'species'. This resembles the problem with the term 'function', mentioned above. Despite persistent problems, many elegant theories on genome evolution have recently been put forward that would have been impossible without the flood of novel data.

The advent of completely sequenced genomes has also stimulated many new approaches to theoretical and practical problems. For example, soon after the arrival of the first sequenced prokaryotic genomes, it became clear that comparative analysis of these would allow the inference of functional features. The common idea behind these methods is that functional association can be inferred among proteins by exploiting the genomic context of their respective genes. Marcotte (pp 359–365) and Huynen *et al.* (pp 366–370) review different aspects of these novel methods. The addition of further genomes should increase the predictive power of such gene context methods. Because in these new methods, functional information is not inferred by homology, they complement the classical methods and are a first step towards the prediction of higher order functions: entire pathways and complexes can be considered, rather than individual proteins or genes.

Completely sequenced genomes also offer more precise benchmarks of what we know about gene products in terms of their three-dimensional structures. *Mycoplasma* has been widely used as a platform for various fold predictions and we can see a continuous improvement of fold prediction methods. Furthermore, quantitative distribution of fold types has been estimated, with implications for the evolution of the proteome and its functions. Jones (pp 371–379) summarizes various methodological developments and their limitations. Despite the fact that we will soon be able to reliably predict the folds of about half of the proteins in all genomes, this knowledge is often insufficient to zoom into molecular detail. Sometimes, the protein sequence similarity between the new sequence and that of the closest known structure is well below 20% identity, so that the resulting homology model will be very crude.

Nevertheless, fold prediction has been an important step in structural genomics, one of the new types of large-scale efforts aimed at characterizing proteomes. Because some current projects in structural genomics focus on proteins for which the fold is unknown, sequence-based methods are crucial in target prioritization. Kim (pp 380–383) describes several concepts and current views on structural genomics, with the aim of obtaining information on the three-dimensional structure of each protein in an organism. Structural genomics should also enhance our knowledge of the functions of proteins, because many methods have been developed that make use of structural information to derive functional features that are complementary to those obtained by sequence-based methods (reviewed by Moult and Melamud, pp 384–389). Many essential molecular details can be revealed only if structural information is available.

Even if we can learn the identities of all the proteins and their post-translational modifications, and even if we can learn their structures and their individual functional features and interaction partners, there remains a long path to understanding cells as a whole. The start of this path is the computational analysis of data from genomics and proteomics, as introduced in this section. This work will direct the next steps along the path. Full simulations of complex networks will be successful only if we develop a solid understanding of all the components involved and of their interactions.