Sequences and topology
Deriving biological knowledge from genomic sequences
Editorial overview

Peer Bork* and David Eisenberg†

Addresses
*European Molecular Biology Laboratory, Meyerhofstrasse 1, 102209,
D-69012 Heidelberg, 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 1998, **8**:331–332

http://biomednet.com/elecref/0959440X00800331

The nine reviews in this section chart new methods for understanding the biological messages of genome sequences. The accelerating rate at which these sequences are being determined has created a demand for informative analytical methods. The accumulation of new data does not in itself lead to increased knowledge. Rather, it challenges us to improve methods for the filtering and processing of sequences to identify the subtle signals therein. This need is heightened by the advent of sequences of entire genomes; these allow qualitatively new features to be detected and open new views on the evolution of genetic material. The initial progress of this emerging science of functional genomics is impressive and is documented in this set of reviews.

Fortunately, one of the first observations to emerge from comparative genome analysis is the robustness of genetic material that has undergone rearrangement. It may be shuffled, horizontally transferred and disrupted, but nevertheless it often maintains its functionality in different organisms. One of the biological themes seems to be 'modularity', which shows up in noncoding DNA, as well as within the genes, and is also manifest in the three-dimensional structures of their products.

Modularity in DNA is created by duplication events followed by modifications, leading to repetitive segments of DNA. Jerzy Jurka (pp 333–337) reviews the evolution of the repetitive transposable elements that comprise a considerable fraction of the total DNA in eukaryotic genomes. Classification and improved detection is essential for genome annotation and also for cleaning expressed sequence tag databases. Jurka emphasizes that the previous view, that these repeats are merely selfish elements, needs to be expanded. Also, whereas most of the current applications treat repeats only as 'waste' for the reduction of search space, the repeats seem to have diverse roles in the genome that can be exploited in a wide range of applications, ranging from population studies to mapping and genomic engineering.

Detection and analysis of repeats is also a challenge at the protein level. Jaap Heringa (pp 338–345) reviews the shift in focus during the past year from repeats at the protein domain level to much shorter fragments that are associated with protein malfunction and genetic diseases. At both the domain level and the subdomain level, the relationship between sequence repeats and three-dimensional structure remains a puzzle.

After the detection of repeats, it is crucial to identify the genes in the genomes. Christopher Burge and Samuel Karlin (pp 346–354) review the recent progress in method development, and also point out future directions. The problem of finding genes (particularly in eukaryotes) is far from solved. No wonder, because various weak translational, transcriptional and splicing signals in the DNA have to be identified and combined with experimental information, such as from expressed sequence tags and trapped exons.

Identification of genes is essential, but their full value comes only with their functional and structural annotation. Using the first complete prokaryotic genomes, Eugene Koonin and colleagues (pp 355–363) discuss important aspects of this annotation process, such as the identification of orthologs and the assignment of folds and catalytic activities. The power of comparative sequence analysis, well known at the level of individual proteins, is now also found at the genome level.

There is still much, however, that is not evident from sequence. Genetic mechanisms can cause modifications of sequence (such as circular permutations, domain insertions and secondary-structure rearrangements) that are beyond the limits of detection of current sequence analysis methods. Robert Russell and Christ Ponting (pp 364–371) summarize cases that can be deciphered only by the analysis of protein topology. Their review emphasizes a general point; in many cases, only structural information can illuminate some of the phenomena that hamper sequence analysis.

Structural knowledge can increase the sensitivity of sequence searches. Liisa Holm (pp 372–379) shows how one can exploit superposition of three-dimensional structures for the unification of protein sequence families and the detection of remote homologues. Yet structural similarity does not lead to iron-clad functional predictions because the same fold can support numerous functions. This is illustrated by the examples that Alexey Murzin (pp 380–387) presents. These examples also show how a wealth of structural data can be correlated in the light of protein evolution.

The complexity of the course of evolution adds complications to genomic analysis. Structural similarity does not necessarily mean a common evolutionary origin and homologous sequences may evolve into different folds (according to current classification schemes). A single function can be found on similar structural scaffolds, so there are numerous examples of parallel evolution towards a similar functionality, even based on extremely different folds. This adds complexity to sequence annotation, as most of the current knowledge on sequenced genomes (particularly beyond the well characterized yeast and *Escherichia coli* genomes) comes from functional inference via homology searches. Thus we can never be sure that a detected homologue has exactly the same function in different genomes. On the other hand, when we hunt for a particular function in a genome, it is always possible that an unrelated protein has acquired this particular function.

A first step towards clarifying such problems will be reliable functional annotation that discriminates between *in vivo*, *in vitro* and (homology) derived data. Clarification also requires, where possible, a structure-based annotation of functional features. At the start, we need to ask what kind of features can and should be derived and described for each sequence. Functional classifications are essential if we want to describe metabolism and, ultimately, phenotypes. Monica Riley (pp 388–392) summarizes many of the problems in function classification, including semantics, hierarchies and inconsistencies. It is important to reach a consistent annotation level, but will we ever achieve annotation that is both reasonably complete and computer-readable? Function always depends on the context and yet only molecular features can be deduced directly from sequence. Some information comes from the availability of entire genomes; for example, the absence of genes and/or functions can be included in predictions.

Today, what we predict from sequences is at best fragmentary and qualitative, for example, the presence or absence of a certain gene or structure or function or pathway. This is not enough to describe cellular processes. Fortunately, there are experimental tools of growing power for the support and extension of genome predictions, such as direct measures of gene expression and protein interaction. One of the leading techniques is mass spectrometry. Bernhard Küster and Matthias Mann (pp 393–400) describe how mass spectrometry can be used to sequence and identify proteins that have post-translational modifications, even though some cannot yet be predicted from sequence.

Although sequence and structure space is not infinite, we will probably never be able to explore them completely (consider, for example, the extinction of species with their genetic material and the rapid modification of virus sequences). With model genomes from evolutionarily distant species becoming available, however, we can make a start at this exploration for humans and other living organisms. In this endeavor, the methods for analysis and annotation that are being developed today will be of the utmost importance in future attempts to bridge the genotype and phenotype of organisms.