The future of bioinformatics

Predicting the future of any scientific discipline is almost impossible, especially for one as young as bioinformatics, but in my view the future of biology will increasingly be shaped by the interplay of bioinformatics and ‘classical’ biology. Currently, the field is undergoing an enormous expansion, as witnessed by the astonishing increase in the number of advertisements for bioinformatics personnel during the past year. To date, most jobs are in the commercial sector, either in large pharmaceutical companies, who can afford to establish internal bioinformatics groups, or in small start-up companies. In academia, most posts are located in large institutes, which provide either sequence data or bioinformatics services. Relatively few positions have been seen in conventional university departments, although recently, at least in the UK, the discipline is becoming increasingly recognized, with several academic lectureship appointments. One current problem is a lack of trained personnel in this interdisciplinary field, which bridges biology, medicine, mathematics, statistics and computer science. As noted by Mark Boguski (pp. 1–3), most practitioners are currently self-taught, having migrated either from wet biology into computing or from the more physical or mathematical sciences into biology, attracted by the explosion in data and the intrinsic value and importance of the data. It is crucial to improve the training provided at all levels, from the casual user to the specialist. For a biologist, bioinformatics expertise is no longer an ‘optional extra’ but a core skill. If biology is to benefit fully from the genome data, establishing adequate training programs in bioinformatics, from undergraduate level upwards, is essential.

However, the more important question is, ‘How will bioinformatics develop and what will it add to classical in vitro/in vivo biology?’ There is no doubt that its immense power and attraction in the future lies in the ability to bring together disparate data from different organisms and different disciplines to unify biology as we know it today. Biologists are continually surprised by the similarities observed between species and across the kingdom of life. In addition, the complexity and sophistication of biological pathways and molecular interactions are astounding. The dissection and interpretation of these data, using tools provided by bioinformatics, will be the crucial ‘clearing house’ from which a modern understanding of the evolution of life will emerge.

Data explosion

In the immediate future, we are faced with a data explosion – not only in primary sequences, which will be the tip of the iceberg, but increasingly with the improvement of proteomics technology, as expression profiles, time correlations, tissue-specific proteins (normal and abnormal), disease-related proteins and ‘personal’ genomes. Such studies will provide challenges in data organization, accessibility and, most importantly, interpretation.

The importance of the Internet for accessibility to data cannot be overstated; its fortuitous parallel development, just as bioinformatics emerged, has radically affected the way data are provided, handled and analyzed. This powerful combination of data and tools, which allow easy access and analysis, has changed and will continue to change our approach to the design and practice of biological research.

A library of protein families

We already have complete genomes for several microorganisms, and the complete human genome should be determined by 2005. In isolation, such an abundance of data is difficult to rationalize. However, one

The consequences of evolution and the physicochemical principles of protein structure on the numbers of genes, sequences and structures

Although we know, given the number of species etc., that there are a large number of different genes (x = 5), the number of domain sequences is much smaller. Furthermore, the number of folds is almost certainly in order of magnitude less, as is the number of different architectures (describing how the sheets and helices pack together in a structure, regardless of sequential coherence). In this extreme, the number of structural supersecondary motifs that constitutes the folds is actually very small. Biological complexity is achieved by using local variation together with a combinatorial approach at all levels, including combining domains to create different proteins and combining proteins to make different complexes. (The numbers represent orders of magnitude only.)

Janet M. Thornton

thornton@biochem.ucl.ac.uk
The biological universe is finite and the hope is that the study of "model organisms" will reveal most if not all protein families. Therefore, following from the genome projects, a major focus for the next few years will be to compile and annotate the dictionary of protein families. This is similar in the biological world to creating a lexicon: a collection of words with their definitions and descriptions. However, in addition to the "dictionary", we will also need a "thesaurus" to highlight relationships between different families.

Work on both these projects has been started in many laboratories. As more families are discovered, these relationships become more complex, with large extended families, and it becomes more difficult to distinguish the boundaries between different groups. The challenge to bioinformatics is to "capture" the biological information for each family and make it readily accessible.

Implications and challenges

The applications and commercial ramifications of bioinformatics are considerable. In the past, computer experts have often been regarded as part of the service environment. In the future, the crucial management decisions on drug discovery programs will be made by individuals who not only understand the biology but can also use the bioinformatics tools and the knowledge they release to develop hypotheses and identify quality targets.

The data explosion modifies the old challenges in computational biology and presents new exciting prospects. As more sequences are determined, the identification of remote homologs will become easier, as intermediate sequences will provide the "missing links". The long-terms goal of predicting structure from sequence ab initio will become more academic, because it will be possible to model most structures from a relative. The emphasis will therefore shift to understanding the principles and control of biological function and the interactions between molecules. Modeling cellular processes, such as signaling and metabolic pathways, will become increasingly important, especially as more proteomic data become available.

Understanding and modeling function is essential to enable the rational design or modification of proteins or ligands for new functions. In my view, this is the greatest challenge for bioinformatics in the next millennium.

References


Fig. 1. The distribution of structures for all domains in the Protein Data Bank (pdb), classified according to class, architecture and fold. The plot attempts to capture our current knowledge of the world of protein structures. It is derived from a classification of structures using sequence and structure comparison methods, and comprises a set of concentric pie charts. The colours define the class: red, mainly as green, mainly b, yellow, mixed a/b and blue, low secondary structure. The inner circle represents the distribution between different architectures, and the outer circle represents different topologies (folds). The size of each segment is proportional to the number of homologous families for that fold/architecture.