Bioinformatics Data Management and Data Mining

Boris Galitsky
Birbeck College University of London, UK

INTRODUCTION

Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting, and utilizing information from biological sequences and molecules. The focus of bioinformatics is the application of computer technology to the management of biological information. Specifically, it is the science of developing computer databases and algorithms to facilitate and expedite biological research, particularly in genomics. It has been mainly stimulated by advances in DNA sequencing and genome mapping techniques (Adams, Fields & Venter, 1994). Genomics is the discipline that studies genes and their functions, including the functional study of genes, their resulting proteins, and the role played by the proteins in the biochemical processes, as well as the study of human genetics by comparisons with model organisms such as mice, fruit flies, and the bacterium *E. coli*.

The Human Genome Project has resulted in rapidly growing databases of genetic sequences. Genome includes all the genetic material in the chromosomes (loci of genetic information) of a particular organism; its size is generally given as its total number of base pairs. New techniques are needed to analyze, manage, and discover sequence, structure, and functional patterns or models from these large sequence and structural databases. High performance data analysis algorithms are also becoming central to this task (Pevsner, 2000; Wilkins et al., 1997).

Bioinformatics provides opportunities for developing novel data analysis methods. Some of the grand challenges in bioinformatics include protein structure prediction, homology search, multiple alignment, and phylogeny construction (the evolutionary relationships among organisms; the patterns of lineage branching produced by the true evolutionary history of the organisms being considered). Proteins are defined as large molecules composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene that codes for the protein. Proteins are required for the structure, function, and regulation of the body’s cells, tissues, and organs; and each protein has unique functions. Examples are hormones, enzymes, and antibodies. Amino acids are the building blocks of proteins; only about 20 amino acids are used to build the thousands of kinds of proteins needed by living cells. Nucleotides are the building blocks of DNA and RNA, consisting of a nitrogenous base, a five-carbon sugar, and a phosphate group. Together, the nucleotides form codons (groups of three nucleotides), which when strung together, form genes that link to form chromosomes.

The other challenges of bioinformatics include genomic sequence analysis and gene finding, as well as applications in the data analysis of gene expression (transcriptional activity of a gene resulting in one or more RNA products and, usually following translation, one or more protein products), drug discovery in the pharmaceutical industry, and so forth (Waterman, 1995). For example, in protein structure prediction, the problem is in determining the secondary (division of sequence into fragments), tertiary (3D space), and quaternary structure of proteins, given their amino acid sequence (Figure 1) (Protein Structure).

Homology search aims at detecting increasingly distant homologues, that is, proteins related by evolution from a common ancestor. Multiple alignment and phylogenetic tree construction are interrelated problems. Multiple alignment aims at aligning a whole set of sequences to determine which subsequences are conserved. This works best when a phylogenetic tree of related proteins (illustration of how some of these proteins were inherited by others) is available. Finally, gene finding aims at locating the genes in a DNA sequence.

The following database-related problems are arising in bioinformatics:

- Data mining and warehousing as applied to biology;
- Data types and modeling needed for biological analysis;
- Interactive exploration and visualization for biology; and
- New indexing and search structures with applications to biology.
As to innovative approaches to database design, the following biological problems are worth mentioning:

1. Evolution and phylogenetic analysis (Maier et al., 2003). The demands of biodiversity and ecosystem research can advance one’s understanding and use of information technologies;
2. Protein structure prediction;
3. Molecular sequence management and alignment;
4. Recognition of genes and regulatory elements;
5. Interpretation of large-scale gene expression data;
6. Whole genome comparative analysis and synthesis;
7. Modeling of biochemical pathways (complex network of interactions between proteins, Pathogenomics); and

We also outline a number of challenges in representing biological data:

- The inherent complexity of biological data;
- Domain knowledge barrier;
- The evolution of domain knowledge; and
- The lack of expert data modeling skills.

**BACKGROUND**

Data management for molecular and cell biology involves the traditional areas of data generation and acquisition, data modeling, data integration, and data analysis. In industry, the main focus of the past several years has been the development of methods and technologies supporting high-throughput data generation, especially for DNA sequence and gene expression data. New technology platforms for generating biological data present data management challenges arising from the need to: (1) capture, (2) organize, (3) interpret, and (4) archive vast amounts of experimental data. Platforms keep evolving with new versions benefiting from technological improvements, such as higher density arrays and better probe selection for microarrays.

This technology evolution raises the additional problem of collecting potentially incompatible data generated using different versions of the same platform, encountered both when these data need to be integrated and analyzed. Further challenges include qualifying the data generated using inherently imprecise tools and techniques and the high complexity of integrating data residing in diverse and poorly correlated repositories.

The data management challenges mentioned above, as well as other data management challenges, have been