Chapter 12
Scalable Index Structures for Biological Data

Ambuj K. Singh

Summary

Bioinformatics holds great promise for the advancement of agriculture, public health, drug design, and the understanding of complex medical and biological systems. For this promise to come to fruition, new query algorithms, data models, and data management techniques need to be developed that can provide access to the varied kinds and large amounts of biological data. This chapter presents scalable index structures for DNA/protein sequences, protein structures, and pathways. After a brief discussion of sequences and structures, the focus shifts to pathways. Modeling of pathways along with their qualitative and quantitative characteristics is considered. Techniques that allow comparison and querying of static and dynamic aspects of pathways are presented.

12.1 Introduction

As a result of the recent spurt in high-throughput techniques, new biological data are being acquired at phenomenal rates. With such a rapid growth, biological datasets (e.g., sequence, structure, expression array, pathway) have become too large to be readily accessible for homology searches, mining, adequate modeling, and integrative understanding. Scalable and integrative tools that access and analyze these valuable data need to be developed.

The growth in genomic information has spurred increased interest in large-scale comparison of genetic sequences. Comparative genomics analyzes and compares the genetic material of different species to identify genes and predict their functions. Genome analysis involves comparison of sequences as large as the whole genome of a species. Phylogenetics and evolutionary studies are other important applications that use complete genetic information

of different species. Shotgun assembly of a genome also requires rapid identification of overlaps across millions of reads. It is obvious that new approaches for large-scale comparison of sequences are needed.

Akin to the growth of sequence databases, protein structure databases, expression array databases, and pathway databases have also been recording significant growth. These databases are intrinsically different from sequence databases. For example, in the case of protein structures, common queries ask for the best alignment (in terms of root mean square distance) of a given query protein to a set of target database proteins. The desired alignment can be either global (i.e., using the entire query, say, for the construction of evolutionary trees or classifications), or local (i.e., using parts of the query to find the active sites). Computing the best alignment is an expensive step if it has to be repeated for all protein structures in PDB [39] or for the larger number of predicted structures [344]. Structure comparison defines the conserved core of a protein family by isolating the common ancestry of proteins. This allows one to go beyond the "twilight zone" where similarities cannot be detected reliably using sequence alone. Predicting the function of proteins *in silico* is of great benefit since it is faster and cheaper than experimentation. Characterization and understanding of protein structures is important for identification of functional motifs and understanding of principles underlying the structure and dynamics of proteins.

Just as sequence and structure databases require the design of new techniques to access, manipulate, and mine datasets, pathway databases require the design of new techniques for accessing, comparing, and manipulating large graphs. There is significant semantics attached to the nodes (substrates, products) and edges (enzymes, reaction control) of such graphs. There is also a need to identify common motifs such as modules in the constructed pathways and to make predictions based on them.

It is evident that the exploding growth in biological data is on a collision course with current database query techniques, presenting new challenges to biological database design. The new generation of databases have to (a) encompass terabytes of data, often local and proprietary, (b) answer queries involving large and complex inputs such as a complete genome, and (c) handle highly complex queries that access more than one dataset (e.g. "find all genes that are structurally similar to a given gene and express similarly over a specific DNA microarray dataset"; "find all proteins that are structurally similar to a given protein, used in a given pathway, and are expressed similarly as another given protein in a given experiment"; "find all protein pairs that are less than 30% similar at a sequence level, share a given active site, and cooccur in some metabolic pathway").

The complexity, heterogeneity, and quantity of biological data also raise difficult issues in the area of data models. Flexible and complex access to biological databases require a model in which information can be stored and queried. There is a need to develop new data models that are sensitive to