OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003
© Mary Ann Liebert, Inc.

Querying and Mining Biological Databases

AMBUJ K. SINGH

THANKS TO INTERNATIONAL SEQUENCING EFFORTS, genome datasets have been growing exponentially in the past few years. The GenBank database, for example, has doubled every 15 months. With such a rapid growth, biological datasets (sequence, structure, expression array, pathway) have become too large to be readily accessible for homology searches, mining, adequate modeling, and understanding. Much of the important information in this enormous and rapidly growing gold mine will be wasted if we do not develop proper tools to access and analyze it efficiently.

The growth in genomic information has spurred increased interest in large-scale comparison of genetic strings. Comparative genomics analyzes and compares the genetic material of different species to identify genes and predict their functions. Such genome analysis involves comparison of strings 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 the genome requires rapid identification of overlaps across millions of reads. The assembly of a large genome can take months on a single machine, the most expensive part being the identification of overlaps. It is obvious that, at this scale, we need new approaches for rapid sequence alignment.

Akin to the growth of sequence databases, the protein structure databases, the expression array databases, and the 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 whole query, say for the construction of evolutionary trees or classification) or local (i.e., using parts of the query, say in order to find the active sites). Computing the best alignment is an expensive step if it has to be repeated for all the approximately 20,000 structures in PDB, or for the larger set of predicted structures. 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 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 the sequence and structure databases require the design of new techniques to access, manipulate, and mine datasets, the 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. The growing amount of biological data and metadata along with novel distance metrics require the design of new index structures and search tools. There is also a need to identify common motifs such as modules in the constructed pathways, and to make predictions based on them. Models, languages, and logics for understanding, simulating, and analyzing pathways are also required to keep pace with the amount of growing data.

It is evident that the exploding growth in size and complexity of biological data is on a collision course with current database query techniques, and presents new challenges to biological database design. The new generation of databases have to (a) encompass terabytes of data, often local and proprietary; (b) answer

Department of Computer Science, University of California, Santa Barbara, California.

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 string level, share a given active site, and co-occur in some metabolic pathway").

The complexity, heterogeneity, and the size 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, so there is a need to develop new data models that are sensitive to the novel characteristics of biological data and queries. Current databases use ad-hoc models that can answer a predefined set of queries and meet the requirements of only parts of the community. The static information about the modeled entities (e.g., genes, protein structures, enhancers) needs to be coupled with dynamic information such as metabolic pathways, regulatory networks, feedback mechanisms, and protein trafficking. There is an increasing demand for the integration of diverse information sources in order to answer complex queries. For example, in order to understand the differences in protein localization between normal and diseased cells, one will need to understand and query the entire dynamic process including abnormalities at the DNA level, and events during transcription, translation, and signaling. When the amount of data and the number of experiments become large, it is no longer feasible for a single scientist to track everything. Unified models will facilitate the integration of currently disparate data.

Following is a partial list of research problems focusing on data management for molecular and cell biology that we need to work on:

- Develop index structures with a small memory footprint that aid in local and global alignment of genomes. Develop new I/O scheduling techniques for accessing disk-resident biological data. Use the developed tools for genome comparisons (e.g., human vs. mouse) and shotgun assembly. Extend the above pair-wise comparison techniques to multiple alignment of genomes.
- Develop new database techniques for finding globally and locally similar protein structures. Develop new memory and CPU- efficient algorithms for discovering new motifs, active sites, and profiles in protein databases. Support similarity searches based on a mix of sequence and structure attributes.
- Develop new algorithms for comparison of pathways. Develop techniques for combining information about pathways with sequences and structures.
- Support interactive queries and efficient access to distributed datasets through the use of prediction and statistics to identify meaningful data.
- Build new data models for providing seamless access to heterogeneous and complex data sources.

Address reprint requests to:
Dr. Ambuj K. Singh
Department of Computer Science
University of California
Santa Barbara, CA 93106

E-mail: ambuj@cs.ucsb.edu