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

Managing Biological Sequence and Protein Structure Data

ZHIPING WENG

MY LABORATORY is primarily concerned with sequence (DNA, RNA and protein) and protein three-dimentional (3D) structure data. In this whitepaper, I discuss the most frequently encountered tasks performed on these types of data that may need new development in database technology.

DATA STRUCTURE

We need a unified data model for biological sequences and their annotations. So far, every sequence analysis algorithms take its favorite format. Even though FASTA seems to be a widely accepted format, it is very limited when it comes to annotations. XML helps with data structure, but the community needs to agree upon a set of most commonly used tags. Substantial effort is spent on improving annotation, but there is no simple way to compare the output of an analysis algorithm with the curated annotations of the same sequence (such as in GenBank RefSeq). How to represent a multiple sequence alignment is far from trivial. BLAST provides seven options for viewing an alignment, all of which become hard on the eyes when there are many sequences and the alignment is gappy. It is unclear how to best store alignments in a database.

We also need a unified data model for protein 3D structures. Presently, each structure is represented as a downloadable file in PDB. The user can search for structures by overall annotations (e.g., author, resolution, structural determination method), but cannot search by sub-structure features; for example, one cannot obtain all structures containing a sub-structure with the secondary structure ordering of α-β-α-β. Or, if one is interested in obtaining 10 residues before and after a motif, which has the sequence of ATTC and occurs roughly at position 50, there is no way to do so except for writing a script to parse all PDB files. There are many data types that are associated with 3D structures, such as the electrostatic potential of a protein structure and surface patches of binding sites. There is no easy way of representing them in a database.

Uncertain, incomplete and inconsistent data are always a problem with biology. Examples include missing coordinates of a PDB file (because these atoms were disordered or too flexible in the crystal structure or in solution). Dealing with alternative conformations of a side chain in crystal-structures is a tough test for PDB parsers. Annotations are typically of varying degrees of certainty, and some may contradict with one another. How to deal with these problems so that an analysis algorithm can spend minimal amount of effort on parsing deserves attention.

SIMILARITY-BASED QUERIES

Sequence similarity search is routinely performed to identify homology. I would like to argue that it is more fundamental than that. When the structures of two homologous proteins are determined, it is quite likely that corresponding residues are numbered differently in the two structures. Therefore, the only way

Department of Biomedical Engineering, Bioinformatics Program, Boston University, Boston, Massachusetts.

of identifying the correspondence, which is necessary for transferring annotations, is to perform a sequence (or structure) alignment. Querying a database by sequence alignment, should be as basic as Boolean operators, similarly for comparing shapes in a 3D structure database. If there can be built-in functions in a database that allow quick similarity-based queries, it would be very useful. Such functions do not need to be sensitive enough to detect remote homology. They should focus on aligning very similar sequences (or structures) and should be very fast. In this sense, they are extended versions of pattern search.

DATA INTEGRATION

There is substantial redundancy in sequence and structure databases. Also, sequences are related by homology. Each sequence carries annotations. It would be very useful if we could easily map the annotated features on one sequence onto the exact corresponding positions of the other. An example would be that two labs are both studying single nucleotide polymorphisms (SNPs) for the same gene, and individually submit their findings as a sequence record. From the user's point of view, it would be ideal if there is a database function to merge all SNPs from these two sequence records straightforwardly. Another example is to transfer the 3D structure information from one sequence to another, which is essentially the homology modeling problem. Of course, one must be very careful with such transfer, especially when there are repeats or the sequence similarity is low. Also, it is not trivial to distinguish orthologs from paralogs. The user needs to make intelligent decisions, but at least the database should provide necessary functions to facilitate the decision making.

TERMINOLOGY MANAGEMENT

Naming is a big problem in biology (of course this gets into the field of ontology): one gene has multiple names, and different genes can have the same name. Sorting this out obviously is not the focus of this workshop. But provided that is done, it would be very useful if there are database functions that can automatically incorporate the correspondences into the query system. There may be conflicts, and functions need to be developed to resolve them.

Address reprint requests to:
Dr. Zhiping Weng
Department of Biomedical Engineering
Boston University
44 Cummington Street
Boston, MA 02215

E-mail: zhiping@bu.edu