Chapter VI
Data Integration Through Protein Ontology

Amandeep S. Sidhu
University of Technology Sydney, Australia

Tharam S. Dillon
University of Technology Sydney, Australia

Elizabeth Chang
Curtin University of Technology, Perth

ABSTRACT

Traditional approaches to integrate protein data generally involved keyword searches, which immediately excludes unannotated or poorly annotated data. An alternative protein annotation approach is to rely on sequence identity, structural similarity, or functional identification. Some proteins have a high degree of sequence identity, structural similarity, or similarity in functions that are unique to members of that family alone. Consequently, this approach can not be generalized to integrate the protein data. Clearly, these traditional approaches have limitations in capturing and integrating data for protein annotation. For these reasons, we have adopted an alternative method that does not rely on keywords or similarity metrics, but instead uses ontology. In this chapter we discuss conceptual framework of protein ontology that has a hierarchical classification of concepts represented as classes, from general to specific; a list of attributes related to each concept, for each class; a set of relations between classes to link concepts in ontology in more complicated ways then implied by the hierarchy, to promote reuse of concepts in the ontology; and a set of algebraic operators for querying protein ontology instances.

INTRODUCTION

Since the first efforts of Maxam (Maxam & Gilbert, 1977) and Sanger (Sanger et al., 1977), the DNA sequence databases have been doubling in size every 18 months or so. This trend continues unabated. This forced the development of systems of software and mathematical techniques for managing and searching these collections. In the past decade, there has been an explosion in the amount of DNA sequence data available, due to very rapid progress of genome sequencing projects. There are three principal comprehensive databases of nucleic acid sequences in the world today.

- The European Molecular Biology Laboratory (EMBL) database is maintained at European Bioinformatics Institute in Cambridge, UK (Stoesser et al., 2003).
- GenBank is maintained at the National Center for Biotechnology Information in Maryland (Benson et al., 2000).
- The DNA Databank of Japan (DDBJ) is maintained at National Institute of Genetics in Mishima, Japan (Miyazaki et al., 2003).

These three databases share information and hence contain identical sets of sequences. The objective of these databases is to ensure that DNA sequence information is stored in a way that is publicly, and freely, accessible and that it can be retrieved and used by other researchers in the future.

Clearly, we have reached a point where computers are essential for the storage, retrieval, and analysis of biological sequence data. The sheer volume of data made it hard to find sequences of interest in each release of sequence databases. The data were distributed as collection of flat files, each of which contained some textual information (the annotation), such as organism name and keywords as well as the DNA sequence. The main way of searching for the sequence of interest was to use a string-matching program. This forced the development of relational database management systems in the main database centers but the databases continued to be delivered as flat files. One important system that is still in use, for browsing and searching the databases, was ACNUC (Gouy et al., 1985), from Manolo Gouy and colleagues in Lyon, France. This was developed in the mid-eighties and allowed fully relational searching and browsing of database annotation.

Another important type of biological data that is exponentially increasing is data of protein structures. Protein Data Bank (PDB) (Bernstein et al., 1977, Weissig & Bourne, 2002, Wesbrook et al., 2002) is a database of protein structures obtained from X-ray crystallography and NMR experiments. At the time of writing this chapter PDB contained over 35,000 structures. It consists of collections of fixed format records that describe the atomic coordinates, chemical and biochemical features, experimental details of structure determination, and some structural features, such as hydrogen bonds and secondary structure assignments. In recent years, dictionary based representations emerged to give data a consistent interface, making it easier to parse. A widely used dictionary-format is the macromolecular crystallographic information file (mmCIF).

The problem of managing biological macromolecular sequence data is as old as the data themselves. In 1998, a special issue of *Nucleic Acids Research* lists 64 different databanks covering diverse areas of biological research, and the nucleotide sequence data alone at over 1 billion bases. It is not only the flood and heterogeneity that make the issues of information representation, storage, structure, retrieval, and interpretation critical. There also has been a change in the user community. In the middle 1980s, fetching a biological entry on a mainframe computer was an adventurous step that only a few dared. At the end of the 1990s, thousands of researchers make use of biological databanks on a daily basis to answer queries, for example, to find sequences similar to a newly sequenced gene, or to retrieve