Chapter 5
An Overview of Biological Data Mining

Seetharaman Balaji
Manipal University, India

ABSTRACT

The largest digital repository of information, the World Wide Web keeps growing exponentially and calls for data mining services to provide tailored web experiences. This chapter discusses the overview of information retrieval, knowledge discovery and data mining. It reviews the different stages of data mining and introduces the wide spread biological databanks, their explosion, integration, data warehousing, information retrieval, text mining, text repositories for biological research publications, domain specific search engines, web mining, biological networks and visualization, ontology and systems biology. This chapter also illustrates some technical jargon with picture analogy for a novice learner to understand the concepts clearly.

INTRODUCTION

As stated by John Naisbitt, "We are drowning in information but starved for knowledge." There is a sea change in science due to the tsunami of sequences sweeping over databases. The genome sequencing projects are producing the vast amount of data challenging scientists to potentially reveal the structure and functional relationships of genes and proteins. This phenomenal growth of biological data (biodata) has been witnessed owing to genomic and proteomic technologies such as high-throughput sequencing (HTS) and mass spectrometry, genome-wide two-hybrid screening, DNA microarray, etc. The tremendous amount of sequence information accumulating worldwide have been organized in the form of databanks. However, the data present in the databanks is heterogeneous in nature, such as genomic data, DNA, RNA and protein sequences, protein structures, sequence patterns, sequence annotations, metabolomic data, gene expression data, protein-protein interactions, cross-references, etc. Most of these data are available only on the World Wide Web (Etzioni, 1996). The explosion of these data is impossible to print because it is voluminous and moreover the information in it can only be acquired and assimilated with the aid of data mining tools. With the advent of computers with huge storage capacity and enhanced processing speed, it became easier to use and share information in an electronic format.

DOI: 10.4018/978-1-5225-8903-7.ch005

The digital revolution has made digitized information easy to capture, process, store, distribute and transmit (Fayyad & Uthurusamy, 1996; Inmon, 1996). Besides, the internet allows rapid data transfer and sharing across the globe in an inexpensive manner. The emergence of data mining research such as cluster analysis, outlier analysis, pattern analysis, data visualization and analysis tools contributes to the development of more efficient and scalable methods for knowledge discovery in databanks (Wang et al., 2005). Data mining is a natural evolution of information technology along the path of data collection, database creation, database management, and data analysis and interpretation (Han & Kamber, 2001) in a new field of study known as Bioinformatics. Bioinformatics addresses problems related to the storage, retrieval, and analysis of information about the biological sequence, structure, and function (Altman, 1998).The main idea of biodata mining is to discover knowledge out of sequence, structure and functional information from the web.

Since structural biology has an enormous impact on our understanding of biology and medicine, some of the examples used in this chapter are from EMBL-EBI's Macromolecular Structure Database (MSD; www.ebi.ac.uk/msd). Recently, the MSD group has been changed its name to the Protein Databank in Europe (PDBe; http://www.ebi.ac.uk/pdbe/), to reflect its close partnership with the wwPDB project. The services and tools name have been changed, but the EBI maintain all existing URLs for external references to the MSD resource. For the 40thanniversary of PDB (2011), PDBe has turned its attention to a focus on the fundamental problem faced by the structural biology community: "how to make the wealth of structural data available to the larger biomedical community?" (Velankar et al. 2011; Golovin et al. 2004)

BACKGROUND

Information Retrieval

Users approach large information spaces like the Web with different motives, to search for a specific piece of information, or to gain familiarity with some general topic or domain, or to navigate something appealing to them. Usually, the needs and preferences of the users are of varying interest. When they navigate through large web structures, they frequently miss their goal of inquiry. Information retrieval uses the Web (and digital libraries) to access information repositories consisting of mixed media and metadata. Information retrieval based on content implies some amount of summarization or compression (Baeza-Yates & Ribeiro-Neto, 1999). The user can give a query so that the information system retrieve relevant documents related to the given query. Further indexing techniques can be used for effective retrieval. Information retrieval can be made efficient by utilizing data mining tools to infer new biological knowledge from existing data (Figure 1).

There are four main components of information retrieval (Manning, Raghvan & Schutze, 2008). They are as follows:

1. **Indexing:** Generates a representation of the document.
2. **Querying:** User preferences are expressed in natural language with logical operators.
3. **Evaluation:** Match user query and document representation.
4. **User Profile Generation:** Record the user preferences to enhance better user retrieval during future access.