4 Data Mining for Bioinformatics

A.W.-C. Liew¹, Hong Yan¹,², and Mengsu Yang³

¹ Department of Computer Engineering and Information Technology
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
² School of Electrical and Information Engineering
University of Sydney, NSW 2006, Australia
³ Department of Biology and Chemistry
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

4.1 Introduction

Biological data is exhaustively vast. Extracting any information from it is not an easy task. It is like mining a tiny bit of gold from the voluminous portion of ore taken from a gold mine. The word “mining” is taken from the mining industry, extracts precious metals from which ores. Data mining in bioinformatics implies extracting valuable information from a large amount of incomprehensible, biological data. In other words, it is a process that leads to knowledge discovery.

An intense effort worldwide by a large segment of the scientific community has focused an mining precious information from genetic data. Data mining in bioinformatics deals with different techniques and algorithms to gain knowledge from data of biological sequences, structures and microarrays.

In this chapter, we will discuss the analysis of biomedical, DNA, and protein data. A detailed discussion is made on the major databases, such as the nucleotide sequence database, the protein sequence database, and the gene expression database. In order to make use of the data from these databases, efficient software tools are needed to retrieve data, compare
biological sequences, discover patterns, and visualize the discovered knowledge. The most widely used tools are also covered.

DNA is a very large molecule that contains amazingly complex information in the form of sequences and structures of amino acids. In order to extract such information, one has to understand the DNA sequences. After a brief discussion on the DNA sequences, the popular techniques and tools used to analyze the DNA sequence data are explained.

Finally, we will have to encounter the product that is based on the information stored in DNA, known as protein. The protein is a relatively large molecule that contains a combination of around 20 amino acids. An understanding of the amino acid sequence will help us probe into problems or events for which the protein is responsible. The basics of the protein sequences are briefly discussed. The techniques, algorithms and the associated tools to analyze the protein data are explained in detail in this chapter in the light of the latest developments.

Even though microarray forms an integral part of data-mining, this relatively mature area needs a separate chapter. We will discuss analysis of microarray data in Chapter 12.

4.2 Biomedical Data Analysis

Recent advances in molecular biology and genomic research, such as high throughput sequencing methods and cDNA microarray technology, have generated an unprecedented amount of data. Efficient analysis of this data by computational methods is becoming a major challenge. Many algorithms have been developed for classifying sequences, detecting weak similarities, separating protein coding regions from non-coding regions in DNA sequences, predicting protein structure and function, and reconstructing the underlying evolutionary history.

Biomedical data can exist in many different forms. We are mainly concerned with biomolecular data based on DNA sequence, protein sequence, and gene expression. A DNA sequence consists of four components, namely, adenine (A), cytosine (C), guanine (G) and thymine (T), specifying the genetic code of the organism. A protein sequence consists of 20 amino acids, coded from the coding region of a DNA sequence. Gene expression data measures the expression of a particular gene, whether up-regulated, down-regulated, or non-expressing, under specific conditions in a cell. A major challenge in bioinformatics research is predicting the structure and function of biosequences by analysis of different biomolecular