Chapter 2
Survey of Biodata Analysis from a Data Mining Perspective

Peter Bajcsy, Jiawei Han, Lei Liu, and Jiong Yang

Summary

Recent progress in biology, medical science, bioinformatics, and biotechnology has led to the accumulation of tremendous amounts of biodata that demands in-depth analysis. On the other hand, recent progress in data mining research has led to the development of numerous efficient and scalable methods for mining interesting patterns in large databases. The question becomes how to bridge the two fields, *data mining* and *bioinformatics*, for successful mining of biological data. In this chapter, we present an overview of the data mining methods that help biodata analysis. Moreover, we outline some research problems that may motivate the further development of data mining tools for the analysis of various kinds of biological data.

2.1 Introduction

In the past two decades we have witnessed revolutionary changes in biomedical research and biotechnology and an explosive growth of biomedical data, ranging from those collected in pharmaceutical studies and cancer therapy investigations to those identified in genomics and proteomics research by discovering sequential patterns, gene functions, and protein-protein interactions. The rapid progress of biotechnology and biodata analysis methods has led to the emergence and fast growth of a promising new field: *bioinformatics*. On the other hand, recent progress in data mining research has led to the development of numerous efficient and scalable methods for mining interesting patterns and knowledge in large databases, ranging from efficient classification methods to clustering, outlier analysis, frequent, sequential, and structured pattern analysis methods, and visualization and spatial/temporal data analysis tools.

The question becomes how to bridge the two fields, *data mining* and *bioinformatics*, for successful data mining of biological data. In this chapter, we present a general overview of data mining methods that have been successfully applied to biodata analysis. Moreover, we analyze how data mining has helped efficient and effective biomedical data analysis and outline some research problems that may motivate the further development of powerful data mining tools in this field. Our overview is focused on three major themes: (1) data cleaning, data preprocessing, and semantic integration of heterogeneous, distributed biomedical databases, (2) exploration of existing data mining tools for biodata analysis, and (3) development of advanced, effective, and scalable data mining methods in biodata analysis.

- **Data cleaning, data preprocessing, and semantic integration of heterogeneous, distributed biomedical databases**

  Due to the highly distributed, uncontrolled generation and use of a wide variety of biomedical data, data cleaning, data preprocessing, and the semantic integration of heterogeneous and widely distributed biomedical databases, such as genome databases and proteome databases, have become important tasks for systematic and coordinated analysis of biomedical databases. This highly distributed, uncontrolled generation of data has promoted the research and development of integrated data warehouses and distributed federated databases to store and manage different forms of biomedical and genetic data. Data cleaning and data integration methods developed in data mining, such as those suggested in [92, 327], will help the integration of biomedical data and the construction of data warehouses for biomedical data analysis.

- **Exploration of existing data mining tools for biodata analysis**

  With years of research and development, there have been many data mining, machine learning, and statistics analysis systems and tools available for general data analysis. They can be used in biodata exploration and analysis. Comprehensive surveys and introduction of data mining methods have been compiled into many textbooks, such as [165, 171, 431]. Analysis principles are also introduced in many textbooks on bioinformatics, such as [28, 34, 110, 116, 248]. General data mining and data analysis systems that can be used for biodata analysis include SAS Enterprise Miner, SPSS, SPlus, IBM Intelligent Miner, Microsoft SQLServer 2000, SGI MineSet, and Inxight VizServer. There are also many biospecific data analysis software systems, such as GeneSpring, Spot Fire, and VectorNTI. These tools are rapidly evolving as well. A lot of routine data analysis work can be done using such tools. For biodata analysis, it is important to train researchers to master and explore the power of these well-tested and popular data mining tools and packages.