Chapter 1
Introduction to Data Mining in Bioinformatics

**Jason T. L. Wang, Mohammed J. Zaki,
Hannu T. T. Toivonen, and Dennis Shasha**

Summary

The aim of this book is to introduce the reader to some of the best techniques for data mining in bioinformatics in the hope that the reader will build on them to make new discoveries on his or her own. The book contains twelve chapters in four parts, namely, overview, sequence and structure alignment, biological data mining, and biological data management. This chapter provides an introduction to the field and describes how the chapters in the book relate to one another.

1.1 Background

Bioinformatics is the science of managing, mining, integrating, and interpreting information from biological data at the genomic, metabalomic, proteomic, phylogenetic, cellular, or whole organism levels. The need for bioinformatics tools and expertise has increased as genome sequencing projects have resulted in an exponential growth in complete and partial sequence databases. Even more data and complexity will result from the interaction among genes that gives rise to multiprotein functionality. Assembling the tree of life is intended to construct the phylogeny for the 1.7 million known species on earth. These and other projects require the development of new ways to interpret the flood of biological data that exists today and that is anticipated in the future.

Data mining or knowledge discovery from data (KDD), in its most fundamental form, is to extract interesting, nontrivial, implicit, previously unknown and potentially useful information from data [165]. In

bioinformatics, this process could refer to finding motifs in sequences to predict folding patterns, to discover genetic mechanisms underlying a disease, to summarize clustering rules for multiple DNA or protein sequences, and so on. With the substantial growth of biological data, KDD will play a significant role in analyzing the data and in solving emerging problems.

The aim of this book is to introduce the reader to some of the best techniques for data mining in bioinformatics (BIOKDD) in the hope that the reader will build on them to make new discoveries on his or her own. This introductory chapter provides an overview of the work and how the chapters in the book relate to one another. We hope the reader finds the book and the chapters as fascinating to read as we have found them to write and edit.

1.2 Organization of the Book

This book is divided into four parts:

I. Overview
II. Sequence and Structure Alignment
III. Biological Data Mining
IV. Biological Data Management

Part I presents a primer on data mining for bioinformatics. Part II presents algorithms for sequence and structure alignment, which are crucial to effective biological data mining and information retrieval. Part III consists of chapters dedicated to biological data mining with topics ranging from genome modeling and gene mapping to protein and chemical mining. Part IV addresses closely related subjects, focusing on querying and indexing methods for biological data. Efficient indexing techniques can accelerate a mining process, thereby enhancing its overall performance. Table 1.1 summarizes the main theme of each chapter and the category it belongs to.

1.2.1 Part I: Basics

In chapter 2, Peter Bajcsy, Jiawei Han, Lei Liu, and Jiong Yang review data mining methods for biological data analysis. The authors first present methods for data cleaning, data preprocessing, and data integration. Next they show the applicability of data mining tools to the analysis of sequence, genome, structure, pathway, and microarray gene expression data. They then present techniques for the discovery of frequent sequence and structure patterns. The authors also review methods for classification and clustering in the context of microarrays and sequences and present approaches for the computational modeling of biological networks. Finally, they highlight visual data mining methods and conclude with a discussion of new research issues such as text mining and systems biology.