PREFACE

Data mining is the process of using computational algorithms and tools to automatically discover useful information in large data archives. Data mining techniques are deployed to score large databases in order to find novel and useful patterns that might otherwise remain unknown. They also can be used to predict the outcome of a future observation or to assess the potential risk in a disease situation. Recent advances in data generation devices, data acquisition, and storage technology in the life sciences have enabled biomedical research and healthcare organizations to accumulate vast amounts of heterogeneous data that is key to important new discoveries or therapeutic interventions. Extracting useful information has proven extremely challenging however. Traditional data analysis and mining tools and techniques often cannot be used because of the massive size of a data set and the non-traditional nature of the biomedical data, compared to those encountered in financial and commercial sectors. In many situations, the questions that need to be answered cannot be addressed using existing data analysis and mining techniques, and thus, new algorithms and methods need to be developed.

Life science is an important application domain that requires new techniques of data analysis and mining. This is one of the first technical books focusing on the data analysis and mining techniques in life science applications. In this introductory chapter, we present the key topics to be covered in this book. In Chapter 1 "Taxonomy of early detection for environmental and public health applications," Chung-Sheng Li of IBM Research provides a survey of early warning systems and detection approaches in terms of problem domains and data sources. The chapter introduces current syndromic surveillance prototypes or deployments and defines the problem domain for three classes: individual and public health level, cellular level, and molecular level. For data sources, they were also categorized into three parts including clinically related data, non-traditional data, and auxiliary data. Furthermore, data sources can be characterized by three dimensions (structured, semi-structured, and non-structured).