

Preface

My interest in studying gene expression microarray data originated in 1998, when a faculty member from the School of Pharmacy at the State University of New York at Buffalo approached me for assistance with his microarray data. I was immediately attracted to the interesting problems presented by these data sets. Since that time, I have been closely following this area of research and have been fascinated by the enormous quantity of data being generated and the need for effective approaches to its analysis.

The total volume of microarray data has increased rapidly over the past several years, a trend which is likely to continue. In addition to traditional time-series and gene-sample data sets, microarray data have also appeared in new and challenging formats. For example, some recent microarray data operate in three dimensions, simultaneously addressing the time, gene, and sample components. Many existing analytical approaches focus on various aspects of data preprocessing, including the processing of scanned images, missing-data estimation, normalization, and summarization. Some apply conventional statistical and machine-learning techniques to pattern analysis and class prediction, but these methods often do not transfer well to use with microarray data. More recently, advances in data-mining techniques have been fruitfully applied to the analysis of complex patterns in microarray data.

This book is intended to provide an understanding of the most current methods available for the analysis of gene expression microarray data, with a particular focus on data-mining techniques. Data mining is well-established area of research which develops scientific approaches to the extraction of knowledge from large data sets. These approaches are more conventionally applied in industrial settings, especially in retail, financial, and telecommunications contexts, but have also recently gained acceptance

for biomedical applications such as the analysis of genomic data. The development of techniques for the effective analysis of genomic datasets is a crucial step in the medical application of bioinformatics. This unique merging of computer-science and biomedical expertise is expected to provide the synergy needed to advance biomedical research to the next level. This book is intended to provide a useful in-depth survey for bioinformatics researchers which I hope will guide and stimulate further investigation.

The book assumes some knowledge on the part of the reader of the fields of molecular biology, data mining, and statistics, as well as a basic understanding of microarray technology. To bridge the gap between these disparate fields, brief overviews are provided of each. Chapter 2 borrows from several standard texts on molecular biology to set forth the fundamental concepts of that field. Chapter 3 summarizes the nature of microarray experiments and data-gathering, and Chapter 4 is a brief tutorial on the techniques of statistical analysis typically applied to microarray data.

Some of the materials in this book have been amassed over several years of teaching a bioinformatics course at an advanced graduate level at the State University of New York at Buffalo. Many of the approaches and research papers referenced in this book can be found at <http://www.cse.buffalo.edu/DBGROUP/bioinformatics/research.html>.

Acknowledgments. I would like to express my deepest thanks to Dr. Daxin Jiang, whose dedicated and indefatigable efforts were essential to the preparation of this book. I would also like to thank my other former doctoral students, Dr. Chun Tang, Xian Xu, and Dr. Li Zhang, for their excellent technical contributions. I am also highly appreciative of the editorial work of Rachel Ramadhyani in recasting my text into proper and idiomatic English.

The inspiration for this book was an invitation from Dr. Jason T. L. Wang to prepare a contribution to his Science, Engineering, and Biology Informatics (SEBI) series. I would like to express my special thanks to Dr. Wang. I also would like to thank the commissioning editor, Ms. Yubing Zhai, for guiding the development of this book and to Mr. Ian Selstrup of World Scientific Publishing Co., Inc. for his overall editorial supervision.

Aidong Zhang
Buffalo, New York