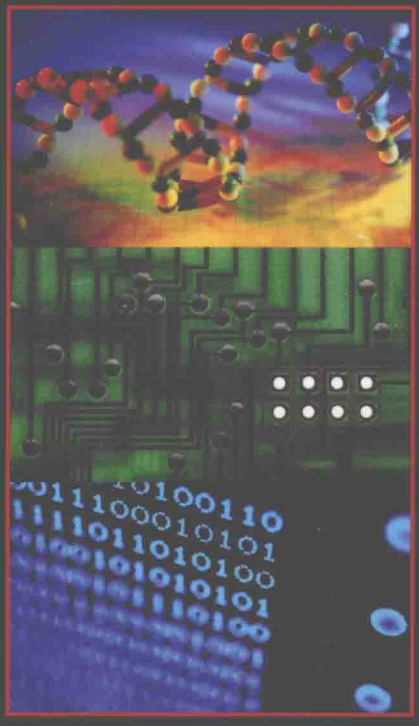
STATISTICAL BIOINFORMATICS

For Biomedical and Life Science Researchers



Edited by JAE K. LEE

WILEY-BLACKWELL

STATISTICAL BIOINFORMATICS

A Guide for Life and Biomedical Science Researchers

Edited by

JAE K. LEE

WILEY-BLACKWELL
A JOHN WILEY & SONS, INC., PUBLICATION

The cover figure signifies the statistical analysis and biological interpretation of high-throughput molecular data. It is the logo of Dr. John N. Weinstein's former research group at the US National Cancer Institute (in which Dr. Jae Lee worked) and his current Genomics & Bioinformatics Group at the M. D. Anderson Cancer Center.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Statistical bioinformatics: a guide for life and biomedical science researchers / edited by Jae K. Lee.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-471-69272-0 (cloth)
1. Bioinformatics—Statistical methods. I. Lee, Jae K.
QH324.2.S725 2010
570.285—dc22

2009024890

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

STATISTICAL BIOINFORMATICS

For Sue, Irene, and Kevin

PREFACE

This book has been constructed primarily as a textbook for a one- or two-semester course in statistical bioinformatics. We hope that it will first serve as a comprehensive introduction to a broad range of topics in this area for life science students and researchers who are motivated to learn statistical analysis concepts and techniques in bioinformatics. Statistical and quantitative science audiences who have not yet dealt with challenging statistical issues in recent information-rich biological and biomedical data analysis may also benefit from this book by efficiently reviewing the different statistical concepts and techniques in such analyses. In particular, the four separate blocks in this book—statistical foundation, high-dimensional analysis, advanced topics, and multigene systems analysis—can be somewhat independently studied and taught in a course based on relevant needs and time restrictions for an effective learning purpose.

A similar outline as organized in this book has been used for several years in a one-semester graduate course which is open to students with diverse backgrounds at the University of Virginia. By publishing this book, we felt that these contents could be significantly enhanced by direct contributions of more specialized experts in the broad field of bioinformatics. In this multiauthor book, we have yet tried to maintain the need of a high-level mathematical and statistical understanding at a minimum. Readers of this book are thus assumed to have not much more than a basic college calculus level of mathematical understanding. A knowledge of matrix algebra would also be useful but is not a prerequisite.

This book is certainly a tribute to the contributions and support of many people. Acknowledgments are first due to the succeeding life science editors of John Wiley & Sons: Luna Han, Thomas Moore, and Karen Chambers. Without their continuous support and encouragement, the publication of this book would not have been possible. Also, all the efforts made by the expert authors who were willing to participate in this book should be highly acknowledged for its successful completion. Finally, the students who have taken this initial course and provided valuable feedback on various topics in this area over the last several years are also significant contributors to the current form of the book. The preparation of this book was supported in part by the National Institutes of Health Research Grant R01 HL081690.

J. K. LEE

Division of Biostatistics and Epidemiology
Charlottesville, Virginia

CONTRIBUTORS

Hongshik Ahn, Department of Applied Mathematics and Statistics, SUNY at Stony Brook, Stony Brook, NY, USA

Merav Bar, Fred Hutchinson Cancer Research Center, Program in Computational Biology, Division of Public Health Sciences, Seattle, WA, USA

Nabil Belacel, National Research Council Canada, Institute for Information Technology, Moncton, NB, Canada

Sooyoung Cheon, KU Industry-Academy Cooperation Group Team of Economics and Statistics, Korea University, Jochiwon 339-700, Korea

Hyung Jun Cho, Department of Statistics, Korea University, Seoul, Korea

Xiangqin Cui, Department of Biostatistics, University of Alabama, Birmingham, AL, USA

Miroslava Cupelovic-Culf, National Research Council Canada, Institute for Information Technology, Moncton, NB, Canada

Ginger Davis, Department of Systems and Information Engineering, University of Virginia, Charlottesville, VA, USA

Robert Gentleman, Fred Hutchinson Cancer Research Center, Program in Computational Biology, Division of Public Health Sciences, Seattle, WA, USA

Debashis Ghosh, Departments of Statistics and Public Health Sciences, Penn State University, University Park, PA, USA

Haseong Kim, Intelligent Systems and Networks Group, Department of Electrical and Electronic Engineering, Imperial College, London, UK

Youngchul Kim, Division of Biostatistics and Epidemiology, University of Virginia, Charlottesville, VA, USA

Michael Lawrence, Fred Hutchinson Cancer Research Center, Program in Computational Biology, Division of Public Health Sciences, Seattle, WA, USA

Jae K. Lee, Division of Biostatistics and Epidemiology, University of Virginia, Charlottesville, VA, USA

Seungyeoun Lee, Department of Applied Statistics, Sejong University, Seoul, Korea

Nolweim LeMeur, Fred Hutchinson Cancer Research Center, Program in Computational Biology, Division of Public Health Sciences, Seattle, WA, USA

Karen Lostritto, Yale School of Public Health, Yale University, New Haven, CT, USA

Annette M. Molinaro, Yale School of Public Health, Yale University, New Haven, CT, USA

Hojin Moon, Department of Mathematics and Statistics, California State University, Long Beach, CA, USA

Annamalai Muthiah, Department of Systems and Information Science, University of Virginia, Charlottesville, VA, USA

Taesung Park, Department of Statistics, Seoul National University, Seoul, Korea

Jinwook Seo, Visual Processing Laboratory, School of Computer Science and Engineering, Seoul National University, Gwanak-gu, Seoul, Korea

Wonseok Seo, Department of Statistics, Korea University, Seoul, Korea

Ben Sheiderman, Human-Computer Interaction Lab & Department of Computer Science, University of Maryland, College Park, MD, USA

Ning Sun, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT, USA

Muneesh Tewari, Fred Hutchinson Cancer Research Center, Program in Computational Biology, Division of Public Health Sciences, Seattle, WA, USA

Christa Wang, National Research Council Canada, Institute for Information Technology, Moncton, NB, Canada

Paul D. Williams, Department of Public Health Science, University of Virginia, Charlottesville, VA, USA

Hongyu Zhao, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT, USA

CONTENTS

PREFACE *xi*

CONTRIBUTORS *xiii*

1 *ROAD TO STATISTICAL BIOINFORMATICS* *1*

Challenge 1: Multiple-Comparisons Issue 1
Challenge 2: High-Dimensional Biological Data 2
Challenge 3: Small-n and Large-p Problem 3
Challenge 4: Noisy High-Throughput Biological Data 3
Challenge 5: Integration of Multiple, Heterogeneous Biological Data Information 3
References 5

2 *PROBABILITY CONCEPTS AND DISTRIBUTIONS FOR ANALYZING LARGE BIOLOGICAL DATA* *7*

2.1 Introduction 7
2.2 Basic Concepts 8
2.3 Conditional Probability and Independence 10
2.4 Random Variables 13
2.5 Expected Value and Variance 15
2.6 Distributions of Random Variables 19
2.7 Joint and Marginal Distribution 39
2.8 Multivariate Distribution 42
2.9 Sampling Distribution 46
2.10 Summary 54

3 *QUALITY CONTROL OF HIGH-THROUGHPUT BIOLOGICAL DATA* *57*

3.1 Sources of Error in High-Throughput Biological Experiments 57
3.2 Statistical Techniques for Quality Control 59
3.3 Issues Specific to Microarray Gene Expression Experiments 66
3.4 Conclusion 69
References 69

4 *STATISTICAL TESTING AND SIGNIFICANCE FOR LARGE BIOLOGICAL DATA ANALYSIS* *71*

4.1 Introduction 71
4.2 Statistical Testing 72
4.3 Error Controlling 78

4.4 Real Data Analysis 81
4.5 Concluding Remarks 87
Acknowledgments 87
References 88

5 *CLUSTERING: UNSUPERVISED LEARNING IN LARGE BIOLOGICAL DATA* *89*

5.1 Measures of Similarity 90
5.2 Clustering 99
5.3 Assessment of Cluster Quality 115
5.4 Conclusion 123
References 123

6 *CLASSIFICATION: SUPERVISED LEARNING WITH HIGH-DIMENSIONAL BIOLOGICAL DATA* *129*

6.1 Introduction 129
6.2 Classification and Prediction Methods 132
6.3 Feature Selection and Ranking 140
6.4 Cross-Validation 144
6.5 Enhancement of Class Prediction by Ensemble Voting Methods 145
6.6 Comparison of Classification Methods Using High-Dimensional Data 147
6.7 Software Examples for Classification Methods 150
References 154

7 *MULTIDIMENSIONAL ANALYSIS AND VISUALIZATION ON LARGE BIOMEDICAL DATA* *157*

7.1 Introduction 157
7.2 Classical Multidimensional Visualization Techniques 158
7.3 Two-Dimensional Projections 161
7.4 Issues and Challenges 165
7.5 Systematic Exploration of Low-Dimensional Projections 166
7.6 One-Dimensional Histogram Ordering 170
7.7 Two-Dimensional Scatterplot Ordering 174
7.8 Conclusion 181
References 182

8 *STATISTICAL MODELS, INFERENCE, AND ALGORITHMS FOR LARGE BIOLOGICAL DATA ANALYSIS* *185*

8.1 Introduction 185
8.2 Statistical/Probabilistic Models 187
8.3 Estimation Methods 189
8.4 Numerical Algorithms 191
8.5 Examples 192
8.6 Conclusion 198
References 199

9 *EXPERIMENTAL DESIGNS ON HIGH-THROUGHPUT BIOLOGICAL EXPERIMENTS* *201*

9.1 Randomization 201
9.2 Replication 202
9.3 Pooling 209
9.4 Blocking 210
9.5 Design for Classifications 214
9.6 Design for Time Course Experiments 215
9.7 Design for eQTL Studies 215
References 216

10 *STATISTICAL RESAMPLING TECHNIQUES FOR LARGE BIOLOGICAL DATA ANALYSIS* *219*

10.1 Introduction 219
10.2 Resampling Methods for Prediction Error Assessment and Model Selection 221
10.3 Feature Selection 225
10.4 Resampling-Based Classification Algorithms 226
10.5 Practical Example: Lymphoma 226
10.6 Resampling Methods 227
10.7 Bootstrap Methods 232
10.8 Sample Size Issues 233
10.9 Loss Functions 235
10.10 Bootstrap Resampling for Quantifying Uncertainty 236
10.11 Markov Chain Monte Carlo Methods 238
10.12 Conclusions 240
References 247

11 *STATISTICAL NETWORK ANALYSIS FOR BIOLOGICAL SYSTEMS AND PATHWAYS* *249*

11.1 Introduction 249
11.2 Boolean Network Modeling 250
11.3 Bayesian Belief Network 259
11.4 Modeling of Metabolic Networks 273
References 279

12 *TRENDS AND STATISTICAL CHALLENGES IN GENOMEWIDE ASSOCIATION STUDIES* *283*

12.1 Introduction 283
12.2 Alleles, Linkage Disequilibrium, and Haplotype 283
12.3 International HapMap Project 285
12.4 Genotyping Platforms 286
12.5 Overview of Current GWAS Results 287
12.6 Statistical Issues in GWAS 290
12.7 Haplotype Analysis 296
12.8 Homozygosity and Admixture Mapping 298
12.9 Gene × Gene and Gene × Environment Interactions 298
12.10 Gene and Pathway-Based Analysis 299

12.11 Disease Risk Estimates 301
12.12 Meta-Analysis 301
12.13 Rare Variants and Sequence-Based Analysis 302
12.14 Conclusions 302
Acknowledgments 303
References 303

13 *R AND BIOCONDUCTOR PACKAGES IN BIOINFORMATICS: TOWARDS SYSTEMS BIOLOGY* *309*

13.1 Introduction 309
13.2 Brief overview of the Bioconductor Project 310
13.3 Experimental Data 311
13.4 Annotation 318
13.5 Models of Biological Systems 328
13.6 Conclusion 335
13.7 Acknowledgments 336
References 336

INDEX *339*

CHAPTER 1

ROAD TO STATISTICAL BIOINFORMATICS

Jae K. Lee

Department of Public Health Science, University of Virginia, Charlottesville, Virginia, USA

There has been a great explosion of biological data and information in recent years, largely due to the advances of various high-throughput biotechnologies such as mass spectrometry, high throughput sequencing, and many genome-wide SNP profiling, RNA gene expression microarray, protein mass spectrometry, and many other recent high-throughput biotechniques (Weinstein et al., 2002). Furthermore, powerful computing systems and fast Internet connections to large worldwide biological databases enable individual laboratory researchers to easily access an unprecedentedly huge amount of biological data. Such enormous data are often too overwhelming to understand and extract the most relevant information to each researcher's investigation goals. In fact, these large biological data are information rich and often contain much more information than the researchers who have generated such data may have anticipated. This is why many major biomedical research institutes have made significant efforts to freely share such data with general public researchers. Bioinformatics is the emerging science field concerned with the development of various analysis methods and tools for investigating such large biological data efficiently and rigorously. This kind of development requires many different components: powerful computer systems to archive and process such data, effective database designs to extract and integrate information from various heterogeneous biological databases, and efficient analysis techniques to investigate and analyze these large databases. In particular, analysis of these massive biological data is extremely challenging for the following reasons.

CHALLENGE 1: MULTIPLE-COMPARISONS ISSUE

Analysis techniques on high-throughput biological data are required to carefully handle and investigate an astronomical number of candidate targets and possible mechanisms, most of which are false positives, from such massive data (Tusher

Statistical Bioinformatics: A Guide for Life and Biomedical Science Researchers. Edited by Jae K. Lee

et al., 2001). For example, a traditional statistical testing criterion which allows 5% false-positive error (or significance level) would identify ~500 false positives from 10K microarray data between two biological conditions of interest even though no real biologically differentially regulated genes exist between the two. If a small number of, for example, 100, genes that are actually differentially regulated exist, such real differential expression patterns will be mixed with the above 500 false positives without any a priori information to discriminate the true positives from the false positives. Then, confidence on the 600 targets that were identified by such a statistical testing may not be high. Simply tightening such a statistical criterion will result in a high false-negative error rate, without being able to identify many important real biological targets. This kind of pitfall, the so-called *multiple-comparisons issue*, becomes even more serious when biological mechanisms such as certain signal transduction and regulation pathways that involve multiple targets are searched from such biological data; the number of candidate pathway mechanisms to be searched grows exponentially, for example, 10! for 10-gene sequential pathway mechanisms. Thus, no matter how powerful a computer system can handle a given computational task, it is prohibitive to tackle such problems by exhaustive computational search and comparison for these kinds of problems. Many current biological problems have been theoretically proven to be NP (nonpolynomial) hard in computer science, implying that no finite (polynomial) computational algorithm can search all possible solutions as the number of biological targets involved in such a solution becomes too large. More importantly, this kind of exhaustive search is simply prone to the risk of discovering numerous false positives. In fact, this is one of the most difficult challenges in investigating current large biological databases and is why only heuristic algorithms that tightly control such a high false positive error rate and investigate a very small portion of all possible solutions are often sought for many biological problems. Thus, the success of many bioinformatics studies critically depends on the construction and use of effective and efficient heuristic algorithms, most of which are based on probabilistic modeling and statistical inference techniques that can maximize the statistical power of identifying true positives while rigorously controlling their false positive error rates.

CHALLENGE 2: HIGH-DIMENSIONAL BIOLOGICAL DATA

The second challenge is the high-dimensional nature of biological data in many bioinformatics studies. When biological data are simultaneously generated with many gene targets, their data points become dramatically sparse in the corresponding high-dimensional data space. It is well known that mathematical and computational approaches often fail to capture such high-dimensional phenomena accurately (Tamayo et al., 1999). For example, many statistical algorithms cannot easily move between local maxima in a high-dimensional space. Also, inference by combining several disjoint lower dimensional phenomena may not provide the correct understanding on the real phenomena in their joint, high-dimensional space. It is therefore important to understand statistical dimension reduction techniques that

can reduce high-dimensional data problems into lower dimensional ones while the important variation of interest in biological data is preserved.

CHALLENGE 3: SMALL-*n* AND LARGE-*p* PROBLEM

The third challenge is the so-called "small-n and large-p" problem. Desired performance of conventional statistical methods is achieved when the sample size, namely n, of the data, the number of independent observations of event, is much larger than the number of parameters, say p, which need to be inferred by statistical inference (Jain et al., 2003). In many bioinformatics problems, this situation is often completely reversed. For example, in a microarray study, tens of thousands of gene transcripts' expression patterns may become candidate prediction factors for a biological phenomenon of interest (e.g., tumor sensitivity vs. resistance to a chemotherapeutic compound) but the number of independent observations (e.g., different patient biopsy samples) is often at most a few tens or smaller. Due to the experimental costs and limited biological materials, the number of independent replicated samples can be sometimes extremely small, for example, two or three, or unavailable. In these cases, most traditional statistical approaches often perform very poorly. Thus, it is also important to select statistical analysis tools that can provide both high specificity and high sensitivity under these circumstances.

CHALLENGE 4: NOISY HIGH-THROUGHPUT BIOLOGICAL DATA

The fourth challenge is due to the fact that high-throughput biotechnical data and large biological databases are inevitably noisy because biological information and signals of interest are often observed with many other random or biased factors that may obscure main signals and information of interest (Cho and Lee, 2004). Therefore, investigations on large biological data cannot be successfully performed unless rigorous statistical algorithms are developed and effectively utilized to reduce and decompose various sources of error. Also, careful assessment and quality control of initial data sets is critical for all subsequent bioinformatics analyses.

CHALLENGE 5: INTEGRATION OF MULTIPLE, HETEROGENEOUS BIOLOGICAL DATA INFORMATION

The last challenge is the integration of information often from multiple heterogeneous biological and clinical data sets, such as large gene functional and annotation databases, biological subjects' phenotypes, and patient clinical information. One of the main goals in performing high-throughput biological experiments is to identify the most important critical biological targets and mechanisms highly associated with biological subjects' phenotypes, such as patients' prognosis and therapeutic response

(Pittman et al., 2004). In these cases, multiple large heterogeneous datasets need to be combined in order to discover the most relevant molecular targets. This requires combining multiple datasets with very different data characteristics and formats, some of which cannot easily be integrated by standard statistical inference techniques, for example, the information from genomic and proteomic expression data and reported pathway mechanisms in the literature. It will be extremely important to develop and use efficient yet rigorous analysis tools for integrative inference on such complex biological data information beyond the individual researcher's manual and subjective integration.

In this book, we introduce the statistical concepts and techniques that can overcome these challenges in studying various large biological datasets. Researchers with biological or biomedical backgrounds may not be able, or may not need, to learn advanced mathematical and statistical techniques beyond the intuitive understanding of such topics for their practical applications. Thus, we have organized this book for life science researchers to efficiently learn the most relevant statistical concepts and techniques for their specific biological problems. We believe that this composition of the book will help nonstatistical researchers to minimize unnecessary efforts in learning statistical topics that are less relevant to their specific biological questions, yet help them learn and utilize rigorous statistical methods directly relevant to those problems. Thus, while this book can serve as a general reference for various concepts and methods in statistical bioinformatics, it is also designed to be effectively used as a textbook for a semester or shorter length course as below. In particular, the chapters are divided into four blocks of different statistical issues in analyzing large biological datasets (Fig. 1.1):

I. *Statistical Foundation* Probability theories (Chapter 2), statistical quality control (Chapter 3), statistical tests (Chapter 4)

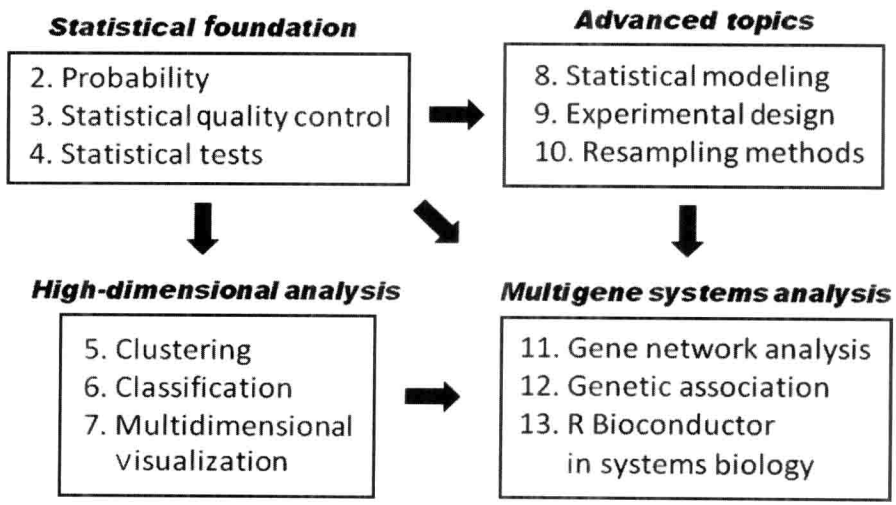


Figure 1.1 Possible course structure.

II. *High-Dimensional Analysis* Clustering analysis (Chapter 5), classification analysis (Chapter 6), multidimensional visualization (Chapter 7)

III. *Advanced Analysis Topics* Statistical modeling (Chapter 8), experimental design (Chapter 9), statistical resampling methods (Chapter 10)

IV. *Multigene Analysis in Systems Biology* Genetic network analysis (Chapter 11), genetic association analysis (Chapter 12), R Bioconductor tools in systems biology (Chapter 13)

The first block of chapters will be important, especially for students who do not have a strong statistical background. These chapters will provide general backgrounds and terminologies to initiate rigorous statistical analysis on large biological datasets and to understand more advanced analysis topics later. Students with a good statistical understanding may also quickly review these chapters since there are certain key concepts and techniques (especially in Chapters 3 and 4) that are relatively new and specialized for analyzing large biological datasets.

The second block consists of analysis topics frequently used in investigating high-dimensional biological data. In particular, clustering and classification techniques, by far, are most commonly used in many practical applications of high-throughput data analysis. Various multidimensional visualization tools discussed in Chapter 7 will also be quite handy in such investigations.

The third block deals with more advanced topics in large biological data analysis, including advanced statistical modeling for complex biological problems, statistical resampling techniques that can be conveniently used with the combination of classification (Chapter 6) and statistical modeling (Chapter 8) techniques, and experimental design issues in high-throughput microarray studies.

The final block contains concise description of the analysis topics in several active research areas of multigene network and genetic association analysis as well as the R Bioconductor software in systems biology analysis. These will be quite useful for performing challenging gene network and multigene investigations in the fast-growing systems biology field.

These four blocks of chapters can be followed with the current order for a full semester-length course. However, except for the first block, the following three blocks are relatively independent of each other and can be covered (or skipped for specific needs and foci under a time constraint) in any order, as depicted in Figure 1.1. We hope that life science researchers who need to deal with challenging analysis issues in overwhelming large biological data in their specific investigations can effectively meet their learning goals in this way.

REFERENCES

Cho, H., and Lee, J. K. (2004). Bayesian hierarchical error model for analysis of gene expression data. *Bioinformatics*, **20**(13): 2016–2025.

Jain, N., et al. (2003). Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics*, **19**(15): 1945–1951.