

# CONTENTS

|                                                                                     |             |
|-------------------------------------------------------------------------------------|-------------|
| <i>PREFACE</i>                                                                      | <i>xi</i>   |
| <i>CONTRIBUTORS</i>                                                                 | <i>xiii</i> |
| <b>1 ROAD TO STATISTICAL BIOINFORMATICS</b>                                         | <b>1</b>    |
| Challenge 1: Multiple-Comparisons Issue                                             | 1           |
| Challenge 2: High-Dimensional Biological Data                                       | 2           |
| Challenge 3: Small- <i>n</i> and Large- <i>p</i> Problem                            | 3           |
| Challenge 4: Noisy High-Throughput Biological Data                                  | 3           |
| Challenge 5: Integration of Multiple, Heterogeneous Biological Data Information     | 3           |
| References                                                                          | 5           |
| <b>2 PROBABILITY CONCEPTS AND DISTRIBUTIONS FOR ANALYZING LARGE BIOLOGICAL DATA</b> | <b>7</b>    |
| 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          |
| <b>3 QUALITY CONTROL OF HIGH-THROUGHPUT BIOLOGICAL DATA</b>                         | <b>57</b>   |
| 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          |
| <b>4 STATISTICAL TESTING AND SIGNIFICANCE FOR LARGE BIOLOGICAL DATA ANALYSIS</b>    | <b>71</b>   |
| 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         |
| <b>5</b> | <b><i>CLUSTERING: UNSUPERVISED LEARNING IN LARGE BIOLOGICAL DATA</i></b>                       | <b>89</b>  |
| 5.1      | Measures of Similarity                                                                         | 90         |
| 5.2      | Clustering                                                                                     | 99         |
| 5.3      | Assessment of Cluster Quality                                                                  | 115        |
| 5.4      | Conclusion                                                                                     | 123        |
|          | References                                                                                     | 123        |
| <b>6</b> | <b><i>CLASSIFICATION: SUPERVISED LEARNING WITH HIGH-DIMENSIONAL BIOLOGICAL DATA</i></b>        | <b>129</b> |
| 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        |
| <b>7</b> | <b><i>MULTIDIMENSIONAL ANALYSIS AND VISUALIZATION ON LARGE BIOMEDICAL DATA</i></b>             | <b>157</b> |
| 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        |
| <b>8</b> | <b><i>STATISTICAL MODELS, INFERENCE, AND ALGORITHMS FOR LARGE BIOLOGICAL DATA ANALYSIS</i></b> | <b>185</b> |
| 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        |

|           |                                                                                    |            |
|-----------|------------------------------------------------------------------------------------|------------|
| <b>9</b>  | <b><i>EXPERIMENTAL DESIGNS ON HIGH-THROUGHPUT BIOLOGICAL EXPERIMENTS</i></b>       | <b>201</b> |
| 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        |
| <b>10</b> | <b><i>STATISTICAL RESAMPLING TECHNIQUES FOR LARGE BIOLOGICAL DATA ANALYSIS</i></b> | <b>219</b> |
| 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        |
| <b>11</b> | <b><i>STATISTICAL NETWORK ANALYSIS FOR BIOLOGICAL SYSTEMS AND PATHWAYS</i></b>     | <b>249</b> |
| 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        |
| <b>12</b> | <b><i>TRENDS AND STATISTICAL CHALLENGES IN GENOMEWIDE ASSOCIATION STUDIES</i></b>  | <b>283</b> |
| 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 $\times$ Gene and Gene $\times$ 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        |
| <b>13 R AND BIOCONDUCTOR PACKAGES IN BIOINFORMATICS: TOWARDS SYSTEMS BIOLOGY</b> | <b>309</b> |
| 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        |
| <b>INDEX</b>                                                                     | <b>339</b> |