Contents

Preface XIII
List of Contributors XVII

Part One General Overview 1

1 Control of Type I Error Rates for Oncology Biomarker Discovery with High-Throughput Platforms 3

Jeffrey Miecznikowski, Dan Wang, and Song Liu

1.1 Brief Summary 3
1.2 Introduction 3
1.3 High-Throughput Platforms 4
1.3.1 Gene Expression Arrays 5
1.3.2 RNA-Seq 5
1.3.3 DNA Methylation Arrays 6
1.3.4 Mass Spectrometry Platforms 6
1.3.5 aCGH Arrays 7
1.3.6 Preprocessing HT Platforms 7
1.4 Analysis of Experiments 8
1.4.1 Linear Regression 8
1.4.1.1 Simple Linear Regression 9
1.4.1.2 Multiple Regression 11
1.4.2 Logistic Regression (Y Discrete) 11
1.4.2.1 Multiple Logistic Regression 13
1.4.3 Survival Modeling 13
1.4.3.1 Kaplan–Meier Analysis 13
1.5 Multiple Testing Type I Errors 15
1.5.1 FWER, k-FWER Methods 17
1.5.1.1 Adjusted Bonferroni Method 17
1.5.1.2 Holm Procedure 17
1.5.1.3 Generalized Hochberg Procedure 18
1.5.1.4 Generalized Šidák Procedure 18
1.5.1.5 minP and maxT procedures 19
2 Overview of Public Cancer Databases, Resources, and Visualization Tools
Frank Emmert-Streib, Ricardo de Matos Simoes, Shailesh Tripathi, and Matthias Dehmer
2.1 Brief Overview 27
2.2 Introduction 27
2.3 Different Cancer Types are Genetically Related 28
2.4 Incidence and Mortality Rates of Cancer 29
2.5 Cancer and Disorder Databases 30
2.6 Visualization and Network-Based Analysis Tools 34
2.6.1 Web-Based Software 34
2.6.2 R-Based Packages 34
2.7 Conclusions 35
2.8 Perspective 37
References 37

Part Two Bayesian Methods 41

3 Discovery of Expression Signatures in Chronic Myeloid Leukemia by Bayesian Model Averaging
Ka Yee Yeung
3.1 Brief Introduction 43
3.2 Chronic Myeloid Leukemia (CML) 44
3.3 Variable Selection on Gene Expression Data 44
3.4 Bayesian Model Averaging (BMA) 46
3.4.1 The Iterative BMA Algorithm (iBMA) 47
3.4.2 Computational Assessment 48
3.5 Case Study: CML Progression Data 49
3.6 The Power of iBMA 50
3.7 Laboratory Validation 51
3.8 Conclusions 52
3.9 Perspective 53
3.10 Publicly Available Resources 54
References 54

4 Bayesian Ranking and Selection Methods in Microarray Studies
Hisashi Noma and Shigeyuki Matsui
4.1 Brief Summary 57
4.2 Introduction 57
4.3 Hierarchical Mixture Modeling and Empirical Bayes Estimation 59
4.4 Ranking and Selection Methods 60
4.4.1 Ranking Based on Effect Sizes 60
4.4.1.1 Posterior Mean (PM) 61
4.4.1.2 Rank Posterior Mean (RPM) 61
4.4.1.3 Tail-Area Posterior Probability (TPP) 62
4.4.2 Ranking Based on Selection Accuracy of Differential Genes 63
4.4.2.1 Posterior Probability of Differentially Expressed (PPDE) 63
4.4.2.2 Evaluating Selection Accuracy 64
4.5 Simulations 65
4.6 Application 67
4.7 Concluding Remarks 71
4.8 Perspective 72
4.9 Appendix: The EM Algorithm 72
References 73

5 Multiclass Classification via Bayesian Variable Selection
with Gene Expression Data 75
Yang Aijun, Song Xinyuan, and Li Yunxian
5.1 Brief Summary 75
5.2 Introduction 75
5.3 Matrix Variate Distribution 77
5.4 Method 77
5.4.1 Model 77
5.4.2 Prior Specification 79
5.4.3 Computation 80
5.4.4 Classification 82
5.5 Real Data Analysis 83
5.5.1 Leukemia Data 83
5.5.2 Lymphoma Data 87
5.5.3 Computational Time 89
5.6 Discussion 89
5.7 Perspective 89
References 90

6 Semisupervised Methods for Analyzing High-dimensional
Genomic Data 93
Devin C. Koestler
6.1 Brief Summary 93
6.2 Motivation 93
6.3 Existing Approaches 95
6.3.1 Fully Unsupervised Procedures 96
6.3.2 Fully Supervised Procedures 96
6.3.3 Semisupervised Procedures 97
6.3.3.1 Semisupervised Clustering 99
6.3.3.2 Semisupervised RPMM 100
References 90
8.5.2 Disease-Module-Based Methods 146
8.5.3 Diffusion-Based Methods 147
8.6 Conclusion 147
8.7 Perspectives 148
References 148

9 Inference of Gene Regulatory Networks in Breast and Ovarian Cancer by Integrating Different Genomic Data 153
Binhua Tang, Fei Gu, and Victor X. Jin
9.1 Brief Summary 153
9.2 Introduction 153
9.3 Theory and Contents of Gene Regulatory Network 154
9.3.1 Basic Theory of Gene Regulatory Network 154
9.3.2 Content of Gene Regulatory Network 155
9.3.2.1 Identify and Infer the Structure Properties and Regulatory Relationships of Gene Networks 155
9.3.2.2 Understand the Basic Rules of Gene Expression and Function 155
9.3.2.3 Discover the Transfer Rules of Genetic Information During Gene Expression 155
9.3.2.4 Study on the Gene Function in a Systematic Framework 156
9.4 Inference of Gene Regulatory Networks in Human Cancer 156
9.4.1 The In Silico Analytical Approach 156
9.4.1.1 Study Case 1: Inference of Static Gene Regulatory Network of Estrogen-Dependent Breast Cancer Cell Line 158
9.4.1.2 Study Case 2: Gene Regulatory Network of Genome-Wide Mapping of TGFβ/SMAD4 Targets in Ovarian Cancer Patients 160
9.4.2 A Bayesian Inference Approach for Genetic Regulatory Analysis 164
9.4.2.1 Study Case: ERα Transcriptional Regulatory Dynamics in Breast Cancer Cell 165
9.5 Conclusions 167
9.6 Perspective 168
References 169

10 Network-Module-Based Approaches in Cancer Data Analysis 173
Guanming Wu and Lincoln Stein
10.1 Brief Summary 173
10.2 Introduction 173
10.3 Notation and Terminology 174
10.4 Network Modules Containing Functionally Similar Genes or Proteins 174
10.5 Network Module Searching Methods 175
10.5.1 Greedy Network Module Search Algorithms 175
10.5.2 Objective Function Guided Search 176
10.5.3 Network Clustering Algorithms 176
10.5.4 Community Search Algorithms 177
10.5.5 Mutual Exclusivity-Based Search Algorithms 178
10.5.6 Weighted Gene Expression Network 178
10.6 Applications of Network-Module-Based Approaches in Cancer Studies 179
10.6.1 Network Modules and Cancer Prognostic Signatures 179
10.6.2 Cancer Driver Gene Search Based on Network Modules 179
10.6.3 Using Network Patterns to Identify Cancer Mechanisms 180
10.7 The Reactome FI Cytoscape Plug-in 180
10.7.1 Construction of a Functional Interaction Network 181
10.7.2 Network Clustering Algorithm 181
10.7.3 Cancer Gene Index Data Set 181
10.7.4 Analyzing the TCGA OV Mutation Data Set 182
10.7.4.1 Loading the Mutation File into Cytoscape and Constructing a FI Subnetwork 182
10.7.4.2 Network Clustering and Network Module Functional Analysis 184
10.7.4.3 Module-Based Survival Analysis 186
10.7.4.4 Cancer Gene Index Data Overlay Analysis 187
10.8 Conclusions 189
10.9 Perspective 189

References 191

11 Discriminant and Network Analysis to Study Origin of Cancer 193

Li Chen, Ye Tian, Guoqiang Yu, David J. Miller, Ie-Ming Shih, and Yue Wang

11.1 Brief Summary 193
11.2 Introduction 193
11.3 Overview of Relevant Machine Learning Techniques 194
11.3.1 Fisher’s Discriminant Analysis and ANOVA 194
11.3.2 Hierarchical Clustering 195
11.3.3 One-Versus-All Support Vector Machine and Nearest-Mean Classifier 196
11.3.4 Differential Dependency Network 197
11.4 Methods 198
11.4.1 CNA Data Analysis for Testing Existence of Monoclonality 198
11.4.1.1 Preprocessing 200
11.4.1.2 Assessing Statistical Significance of Monoclonality 200
11.4.1.3 Visualization of Monoclonality 201
11.4.2 A Two-Stage Analytical Method for Testing the Origin of Cancer 201
11.4.2.1 Basic Assumptions 202
11.4.2.2 Tissue Heterogeneity Correction 203
11.4.2.3 Stage 1: Feature Selection and Classification 203
11.4.2.4 Stage 2: Transcriptional Network Comparison 204
11.5 Experiments and Results 204
11.5.1 Monoclonality 204
11.5.1.1 Testing Existence of Monoclonality 204
11.5.1.2 The Significance of Monoclonality 206
11.5.2 Testing the Origin of Ovarian Cancer 207
11.5.2.1 Stage 1 Results 207
11.5.2.2 Stage 2 Results 208
11.6 Conclusion 211
11.7 Perspective 212
References 212

12 Intervention and Control of Gene Regulatory Networks: Theoretical Framework and Application to Human Melanoma Gene Regulation 215
Nidhal Bouaynaya, Roman Shterenberg, Dan Schonfeld, and Hassan M. Fathallah-Shaykh
12.1 Brief Summary 215
12.2 Gene Regulatory Network Models 216
12.3 Intervention in Gene Regulatory Networks 218
12.3.1 Optimal Stochastic Control 219
12.3.2 Heuristic Control Strategies 221
12.3.3 Structural Intervention Strategies 222
12.4 Optimal Perturbation Control of Gene Regulatory Networks 223
12.4.1 Feasibility Problem 226
12.4.2 Optimal Perturbation Control 226
12.4.2.1 Minimal-Energy Perturbation Control 226
12.4.2.2 Fastest-Convergence Rate Perturbation Control 228
12.4.3 Trade-offs Between Minimal-Energy and Fastest Convergence Rate Perturbation Control 228
12.4.4 Robustness of Optimal Perturbation Control 231
12.5 Human Melanoma Gene Regulatory Network 231
12.6 Perspective 235
References 236

Part Four Phenotype Influence of DNA Copy Number Aberrations 239
13 Identification of Recurrent DNA Copy Number Aberrations in Tumors 241
Vonn Walter, Andrew B. Nobel, D. Neil Hayes, and Fred A. Wright
13.1 Introduction 241
13.2 Genetic Background 242
13.2.1 Definitions 242
13.2.2 Mechanisms of DNA Copy Number Change: An Overview 243
13.2.3 CNAs and Cancer 244
13.2.4 Sporadic and Recurrent CNAs 245
13.2.5 Measuring DNA Copy Number 245
13.2.6 Other Issues to Consider When Assessing DNA Copy Number 246
13.3 Analyzing DNA Copy Number: Single Sample Methods 246
13.3.1 Notation 247
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.3.2</td>
<td>Quality Control and Preprocessing</td>
<td>247</td>
</tr>
<tr>
<td>13.3.3</td>
<td>Thresholding</td>
<td>247</td>
</tr>
<tr>
<td>13.3.4</td>
<td>Segmentation Algorithms</td>
<td>248</td>
</tr>
<tr>
<td>13.3.5</td>
<td>Methods Based on Hidden Markov Models</td>
<td>248</td>
</tr>
<tr>
<td>13.4</td>
<td>Analyzing DNA Copy Number Data: Multiple Sample Methods to Detect Recurrent CNAs</td>
<td>249</td>
</tr>
<tr>
<td>13.4.1</td>
<td>Additional Preprocessing and Summary Statistics</td>
<td>249</td>
</tr>
<tr>
<td>13.4.2</td>
<td>Multiple Testing</td>
<td>250</td>
</tr>
<tr>
<td>13.4.3</td>
<td>Assessing Statistical Significance: An Overview</td>
<td>250</td>
</tr>
<tr>
<td>13.5</td>
<td>Analyzing DNA Copy Number Data with DiNAMIC</td>
<td>251</td>
</tr>
<tr>
<td>13.5.1</td>
<td>Cyclic Shifts</td>
<td>251</td>
</tr>
<tr>
<td>13.5.2</td>
<td>Assessing Statistical Significance with DiNAMIC</td>
<td>252</td>
</tr>
<tr>
<td>13.5.3</td>
<td>Peeling</td>
<td>253</td>
</tr>
<tr>
<td>13.5.4</td>
<td>Confidence Intervals for Recurrent CNAs</td>
<td>256</td>
</tr>
<tr>
<td>13.5.5</td>
<td>Bootstrap Test-Based Confidence Intervals in Real Datasets</td>
<td>257</td>
</tr>
<tr>
<td>13.6</td>
<td>Open Questions</td>
<td>258</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>259</td>
</tr>
</tbody>
</table>

14 The Cancer Cell, Its Entropy, and High-Dimensional Molecular Data 261

Wessel N. van Wieringen and Aad W. van der Vaart

14.1 Brief Summary 261
14.2 Introduction 261
14.3 Background 262
14.3.1 Molecular Biology 262
14.3.2 Cancer 263
14.3.3 Measurement Devices 263
14.4 Entropy Increase 264
14.5 Statistical Arguments 266
14.6 Statistical Methodology 268
14.6.1 Experiments 269
14.6.2 Entropy 269
14.6.3 Mutual Information 272
14.7 Simulation 275
14.8 Application to Cancer Data 275
14.8.1 Analyses of Type II Experiments 276
14.8.2 Analyses of Type I Experiments 279
14.8.3 Potential 280
14.8.4 Discussion 282
14.9 Conclusion 283
14.10 Perspective 283
14.11 Software 284
References 284

Index 287