

Contents

Preface XIII

List of Contributors XVII

1	Introduction to DNA Microarrays	1
	<i>Chris Seidel</i>	
1.1	Introduction	1
1.1.1	The Genome is an Information Scaffold	2
1.1.2	Gene Expression is Detected by Hybridization	3
1.1.2.1	Hybridization is Used to Measure Gene Expression	4
1.1.2.2	Microarrays Provide a New Twist to an Old Technique	5
1.2	Types of Arrays	5
1.2.1	Spotted Microarrays	6
1.2.2	Affymetrix GeneChips	6
1.2.2.1	Other <i>In Situ</i> Synthesis Platforms	7
1.2.2.2	Uses of Microarrays	8
1.3	Array Content	11
1.3.1	ESTs Are the First View	11
1.3.1.1	Probe Design	12
1.4	Normalization and Scaling	14
1.4.1	Be Unbiased, Be Complete	18
1.4.2	Sequence Counts	18
	References	19
2	Comparative Analysis of Clustering Methods for Microarray Data	27
	<i>Dongxiao Zhu, Mary-Lee Deque�ant, and Hua Li</i>	
2.1	Introduction	27
2.2	Measuring Distance Between Genes or Clusters	28
2.3	Network Models	34
2.3.1	Boolean Network	34
2.3.2	Coexpression Network	34
2.3.3	Bayesian Network	36

2.3.4	Co-Occurrence Network	37
2.4	Network Constrained Clustering Method	38
2.4.1	Extract the Giant Connected Component	38
2.4.2	Compute “Network Constrained Distance Matrix”	39
2.5	Network Constrained Clustering Results	39
2.5.1	Yeast Galactose Metabolism Pathway	40
2.5.2	Retinal Gene Expression Data	43
2.5.3	Mouse Segmentation Clock Data	46
2.6	Discussion and Conclusion	47
	References	48
3	Finding Verified Edges in Genetic/Gene Networks: Bilayer Verification for Network Recovery in the Presence of Hidden Confounders	51
	<i>Jason E. Aten</i>	
3.1	Introduction: Gene and Genetic Networks	51
3.2	Background and Prior Theory	53
3.2.1	Motivation	53
3.2.2	Bayesian Networks Theory	53
3.2.2.1	d-Separation at Colliders	55
3.2.2.2	Placing Genetic Tests Within the Bayesian Network Framework	56
3.2.3	Learning Network Structure from Observed Conditional Independencies	58
3.2.4	Prior Work: The PC Algorithm	58
3.2.4.1	PC Algorithm	58
3.2.5	Prior Work: The Local Causal Discovery Algorithm	59
3.2.5.1	LCD Algorithm	60
3.3	New Theory	61
3.3.1	Novel Algorithm: The RVL Algorithm for Learning DAGs Efficiently	61
3.3.1.1	Algorithm: Recursive v-Structure Location	61
3.3.2	Novel Theory: Bilayer Verification and the RVV Algorithm for Verifying Graphs in the Presence of Unobserved Confounders	62
3.3.2.1	Algorithm: Recursive v-Structures with Verification	67
3.4	Methods	68
3.4.1	C3H/HeJ $\times$ C57BL/6J Microarray Data	68
3.4.2	Parameters and the v-Structure Test	68
3.4.2.1	Mechanics of the v-Structure Test	69
3.4.3	Special Handling for Genotypes	70
3.5	Results and Further Application	70
3.5.1	Estimating α False-Positive Rates for the v-Structure Test	70
3.5.2	Learning an Aortic Lesion Network	77
3.5.3	Further Utilizing Networks: Assigning Functional Roles to Genes	77
3.5.4	Future Work	79
	References	80

4	Computational Inference of Biological Causal Networks – Analysis of Therapeutic Compound Effects	83
	<i>Jing Yu, Gabriel Helmlinger, Muriel Saulnier, and Anna Georgieva</i>	
4.1	Introduction	83
4.2	Basic Theory of Bayesian Networks	84
4.2.1	Bayesian Scoring Metrics	86
4.2.2	Heuristic Search Methods	87
4.2.3	Inference Score	88
4.3	Methods	88
4.3.1	Experimental Design	88
4.3.2	Tissue Contamination	88
4.3.3	Gene List Prefiltering	89
4.3.4	Outlier Removal	89
4.3.5	Further Screening of the Gene List	90
4.3.6	Data Pair-Up for Using DBN	91
4.3.7	Applying DBN	92
4.4	Results	93
4.4.1	Computational Results	93
4.4.2	Biological Findings	93
4.5	Discussion and Conclusions	96
	References	97
5	Reverse Engineering Gene Regulatory Networks with Various Machine Learning Methods	101
	<i>Marco Grzegorzczuk, Dirk Husmeier, and Adriano V. Werhli</i>	
5.1	Introduction	101
5.2	Methods	102
5.2.1	Relevance Networks	103
5.2.2	Gaussian Graphical Models	104
5.2.3	Bayesian Networks	106
5.2.3.1	Introduction to Bayesian Networks	106
5.2.3.2	Learning Causal Relationships	108
5.2.3.3	Bayesian Network Scoring Metrics	109
5.2.3.4	The Gaussian BGe Scoring Metric	110
5.2.3.5	Structure Learning Via MCMC Simulations	113
5.2.3.6	Learning Bayesian Networks from Interventional Data	118
5.3	The RAF Signalling Pathway	120
5.4	Criteria for Evaluating Learning Performances	122
5.5	Data	125
5.6	Simulations	128
5.7	Results	129
5.8	Discussion	131
5.9	Conclusion	140
	References	140

6	Statistical Methods for Inference of Genetic Networks and Regulatory Modules	143
	<i>Hongzhe Li</i>	
6.1	Introduction	143
6.2	Network Inference Based on Gaussian Graphical Models	145
6.2.1	Gaussian Graphical Models	146
6.2.2	Threshold Gradient Descent Regularization	146
6.2.3	Model Selection by Cross-Validation and Bootstrap	148
6.2.4	Simulation Results and Application to Real Data Set	149
6.3	Methods for Identifying Regulatory Modules	151
6.3.1	The SRMM for Identifying Transcriptional Modules	151
6.3.2	An EM Algorithm Based on Lasso	152
6.3.3	Selection of the Number of Modules K and the Tuning Parameter s	153
6.3.4	Application to Yeast Stress Data Set	154
6.4	Inference of Transcriptional Networks	155
6.4.1	Functional Response Model with Time-Varying Coefficients for MTC Gene Expression Data	156
6.4.2	Estimation Using B-Splines	157
6.4.3	A Group SCAD Penalization Procedure	157
6.4.4	Numerical Algorithm, Properties, and Application	158
6.5	Discussion, Conclusions, and Future Research	160
6.5.1	Incorporating Network Information into Analysis of Microarray Gene Expression Data	160
6.5.2	Development of Statistical and Computational Methods for Integrating Gene Expression Data and Epigenomic Data	163
6.5.3	Final Remarks	163
	References	164
7	A Model of Genetic Networks with Delayed Stochastic Dynamics	169
	<i>Andre S. Ribeiro</i>	
7.1	Introduction	169
7.2	Experimental Observations of Gene Expression	171
7.2.1	The Stochastic Nature of Gene Expression	172
7.2.2	Time Delays in Transcription and Translation	173
7.3	The Delayed Stochastic Simulation Algorithm	176
7.3.1	Stochastic Simulation Algorithm	176
7.3.2	The Delayed Stochastic Simulation Algorithm	178
7.4	Modeling Gene Expression as a Multiple Time-Delayed Stochastic Event	179
7.5	A Gene Regulatory Network Model	180
7.6	Applications	186
7.6.1	Modeling Single Gene Expression	186
7.6.2	Bistability of a Toggle Switch as a Result of Time Delays in Transcription	190

7.7	A Model of the P53–Mdm2 Feedback Loop Network	194
7.8	Summary, Conclusions, and Applications	200
	References	201
8	Probabilistic Boolean Networks as Models for Gene Regulation	205
	<i>Yufei Huang and Edward R. Dougherty</i>	
8.1	Introduction	205
8.2	Modeling Gene Regulation with Probabilistic Boolean Networks	207
8.2.1	Preliminaries	207
8.2.2	Probabilistic Boolean Networks	210
8.2.2.1	Context-Sensitive PBNs and PBNs with Random Perturbation	213
8.3	Reverse Engineering Regulatory Networks with PBN-Based Microarray Expression Data	215
8.3.1	A Disjoint Bayesian Solution of Constructing Probabilistic Boolean Networks	216
8.3.1.1	Experimental Results	218
8.3.2	A Full Bayesian Solution	219
8.3.2.1	Melanoma Application	220
8.4	Optimal Control of Context-Sensitive PBN	221
8.4.1	Introduction to Network Intervention	221
8.4.2	Defining the Transition Probability of a Context-Sensitive PBN	223
8.4.3	External Intervention with Finite-Horizon Control	224
8.4.3.1	Melanoma Application	227
8.4.4	External Intervention with Infinite-Horizon Control	228
8.4.4.1	The Discounted Approach	230
8.4.4.2	The Average-Cost-Per-Stage Approach	233
8.4.5	Melanoma Application	235
	References	240
9	Structural Equation for Identification of Genetic Networks	243
	<i>Momiao Xiong</i>	
9.1	Introduction	243
9.2	Models	245
9.3	Covariance Matrix	249
9.4	Estimation	250
9.4.1	Likelihood Function	250
9.4.2	Maximum Likelihood Estimators	251
9.4.3	Asymptotic Properties of the Maximum Likelihood Estimators and Test Statistics	254
9.4.4	Two-Stage Least Square Method	255
9.4.4.1	Reduce Form	256
9.4.4.2	Two-Stage Least Squares Estimation	256
9.4.4.3	Unweighted Least Squares (ULS)	257
9.4.4.4	Generalized Least Squares (GLS)	257
9.5	Model Selection	258

9.5.1	Model Selection Criterion	258
9.5.2	Genetic Algorithms (GAs)	259
9.5.3	Illustration of Structural Equations for Modeling Genetic Networks	260
9.6	Identification of Differentially Expressed Genetic Networks	267
9.6.1	The Generalized T^2 Statistic for Testing the Differential Expression of Genetic Networks	267
9.6.2	Nonlinear Tests for Identifying Differentially Expressed Genetic Networks	268
9.6.3	Examples	269
9.7	Differentially Regulated Genetic Networks	272
9.7.1	Index for Measuring Difference in Regulation of Genetic Networks	272
9.7.2	Examples	274
9.8	Conclusions	279
	References	280
10	Detecting Pathological Pathways of a Complex Disease by a Comparative Analysis of Networks	285
	<i>Frank Emmert-Streib and Matthias Dehmer</i>	
10.1	Introduction	285
10.2	Outline of Our Method	287
10.3	Detecting Pathological Pathways	288
10.3.1	Pathways as Undirected Dependency Graphs	289
10.4	Measuring the Similarity of Pathways	292
10.4.1	GED for Graphs With Unique Vertex Labels	295
10.4.2	Statistical Significance of the GED	298
10.5	Results for the Chronic Fatigue Syndrome	298
10.5.1	Influence of Measurement Errors	300
10.6	Discussions and Conclusions	302
	References	303
11	Predicting Functional Modules Using Microarray and Protein Interaction Data	307
	<i>Yu Chen and Dong Xu</i>	
11.1	Introduction	307
11.2	Materials and Methods	309
11.2.1	Data sets	309
11.2.2	Protein Function Annotation and GO Index	310
11.2.3	Construction of Probabilistic Functional Network	310
11.2.4	Identification of Functional Modules by Clustering the Network	311
11.2.5	Evaluation of Topological and Functional Properties of Modules	312
11.3	Results	314
11.3.1	Modules Discovered from the Probabilistic Functional Network	314
11.3.2	Evaluation of Modules	316

11.3.3	Module Organization in Yeast Gene Interaction Network	320
11.4	Discussion	324
	References	326
12	Computational Reconstruction of Transcriptional Regulatory Modules of the Yeast Cell Cycle	331
	<i>Wei-Sheng Wu, Wen-Hsiung Li, and Bor-Sen Chen</i>	
12.1	Introduction	331
12.2	Methods	332
12.2.1	Data Sets	332
12.2.2	Temporal Relationship Identification Algorithm	333
12.2.3	The Module Finding Algorithm	334
12.3	Results	337
12.3.1	Validation of the Identified Modules	337
12.3.2	Identification of Important Cell Cycle TFs and Their Combinations	338
12.3.3	The M/G1 Phase	338
12.3.4	The G1 Phase	341
12.3.5	The S Phase	342
12.3.6	The SG2 and G2/M Phases	342
12.4	Discussion	343
12.4.1	Relationships Between Two TFs of a Module	343
12.4.2	Advantages of MOFA	343
12.4.3	Parameter Settings of MOFA	345
12.4.4	Refining Clusters from Spellman <i>et al.</i>	346
12.5	Conclusions	347
	References	350
13	Pathway-Based Methods for Analyzing Microarray Data	355
	<i>Herbert Pang, Inyoung Kim, and Hongyu Zhao</i>	
13.1	Introduction	355
13.2	Methods	356
13.2.1	Random Forests Based Approach	356
13.2.1.1	Random Forests Classification	356
13.2.1.2	Random Forests Regression	358
13.2.2	Regression Model Based Approach	359
13.2.2.1	Bayesian Hierarchical Model	359
13.2.2.2	A Bayesian MCMC Approach	359
13.3	Real Data Analysis	360
13.3.1	Pathways and Gene Sets	361
13.3.2	Data Analysis Using Random Forests	361
13.3.2.1	Canine Data Set	361
13.3.2.2	Breast Cancer Data Set	371
13.3.2.3	Diabetes Data Set	372
13.3.2.4	Comparison with Other Machine Learning Approaches	373

13.3.3	Data Analysis Using Bayesian Approach	374
13.4	Conclusions and Discussion	378
	References	380

14 **The Most Probable Genetic Interaction Networks Inferred from Gene Expression Patterns** 385

Timothy R. Lezon, Jayanth R. Banavar, Marek Cieplak, Nina V. Fedoroff, and Amos Maritan

14.1	Introduction	385
14.2	Entropy Maximization	386
14.3	Recovering the Data	391
14.4	Integrating Over Interactions	393
14.5	Higher Order Interactions	395
14.6	Network Analysis	398
14.6.1	Metabolic Oscillations in Yeast	398
14.6.2	Polishing of the Data and Selection of Subsets of Genes	399
14.6.3	The Nature of the Network	400
14.6.4	The Biological Interpretation of the Network	404
14.6.5	The Larger Subset of Genes	407
14.6.6	Metabolic Oscillations with Longer Periods	407
14.6.7	Three-Gene Interactions	409
14.7	Conclusion	409
	References	410

Index 413