

CONTENTS

Abbreviations	xix
List of Contributors	xxiii

Part I. INTRODUCTION

1. Probabilistic Graphical Models for Next-generation Genomics and Genetics	3
CHRISTINE SINOQUET	
1.1. Fine-grained Description of Living Systems	4
1.1.1. DNA and the Genome	4
1.1.2. Genes and Proteins	5
1.1.3. Phenotype and Genotype	5
1.1.4. Molecular Biology, Genetics, Genomics, and Postgenomics	6
1.2. Higher Description Levels of Living Systems	6
1.2.1. Complexity in Cells	7
1.2.2. Genetics, Epigenetics, and Copy Number Polymorphism	9
1.2.3. Epigenetics with Additional Prior Knowledge on the Genome	11
1.2.4. Transcriptomics	11
1.2.5. Transcriptomics with Prior Biological Knowledge	13
1.2.6. Integrating Data from Several Levels	13
1.2.7. Recapitulation	16
1.3. An Era of High-throughput Genomic Technologies	16
1.3.1. Genotyping	16
1.3.2. Copy Number Polymorphism	19
1.3.3. DNA Methylation Measurements	19
1.3.4. Gene Expression Data	20
1.3.5. Quantitative Trait Loci	21
1.3.6. The Challenge of Handling Omics Data	23
1.4. Probabilistic Graphical Models to Infer Novel Knowledge from Omics Data	23
1.4.1. Gene Network Inference	24
1.4.2. Causality Discovery	24
1.4.3. Association Genetics	26

1.4.4.	Epigenetics	26
1.4.5.	Detection of Copy Number Variations	26
1.4.6.	Prediction of Outcomes from High-dimensional Genomic Data	26
2.	Essentials to Understand Probabilistic Graphical Models: A Tutorial about Inference and Learning	30
	CHRISTINE SINOQUET	
2.1.	Introduction	32
2.2.	Reminders	32
2.3.	Various Classes of Probabilistic Graphical Models	38
2.3.1.	Markov Chains and Hidden Markov Models	38
2.3.2.	Markov Random Fields	39
2.3.3.	Variants around the Concept of Markov random field	41
2.3.4.	Bayesian networks	41
2.3.5.	Unifying Model and Model Extension	45
2.4.	Probabilistic Inference	46
2.4.1.	Exact Inference	46
2.4.2.	Approximate Inference	51
2.5.	Learning Bayesian networks	57
2.5.1.	Parameter Learning	58
2.5.2.	Structure Learning	61
2.6.	Learning Markov random fields	69
2.6.1.	Parameter Learning	69
2.6.2.	Structure Learning	72
2.7.	Causal Networks	75
2.8.	List of General Monographs and Focused Chapter Books	77
	Part II. GENE EXPRESSION	
3.	Graphical Models and Multivariate Analysis of Microarray Data	85
	HARRI KIIVERI	
3.1.	Introduction	85
3.2.	The Model	87
3.3.	Model Fitting	88
3.3.1.	Maximum Likelihood Estimation when the Zero Pattern is Known	89
3.3.2.	Determining the Pattern of Zeroes in the Inverse Covariance Matrix	90
3.4.	Hypothesis Testing	92
3.4.1.	Null Distributions by Permutation	92
3.4.2.	A Multivariate Test Statistic	93
3.4.3.	Partitioning of the Test Statistic	94
3.4.4.	Testing Strategies	95
3.5.	Example	96
3.6.	Discussion and Conclusions	99
4.	Comparison of Mixture Bayesian and Mixture Regression Approaches to Infer Gene Networks	105
	SANDRA L. RODRIGUEZ-ZAS AND BRUCE R. SOUTHEY	
4.1.	Introduction	106

4.2. Methods	107
4.2.1. Mixture Bayesian Network	107
4.2.2. Mixture Regression Approach	108
4.2.3. Data	110
4.3. Results	112
4.3.1. Comparison of Mixtures	112
4.3.2. Mixture Modeling of Changes in Gene Relationships	112
4.3.3. Interpretation of Mixtures	114
4.3.4. Inference of Large Networks	116
4.4. Conclusions	116
5. Network Inference in Breast Cancer with Gaussian Graphical Models and Extensions	121
MARINE JEANMOUGIN, CAMILLE CHARBONNIER, MICKAËL GUEDJ, AND JULIEN CHIQUET	
5.1. Introduction	122
5.2. Modeling of Gene Networks by Gaussian Graphical Networks	123
5.2.1. Simple Gaussian graphical network	123
5.2.2. Extensions Motivated by Regulatory Network Modeling	127
5.3. Application to Estrogen Receptor Status in Breast Cancer	134
5.3.1. Context	134
5.3.2. Biological Prior Definition	135
5.3.3. Network Inference from Biological Prior: Application and Interpretation	139
5.4. Conclusions and Discussion	141
 Part III. CAUSALITY DISCOVERY	
6. Utilizing Genotypic Information as a Prior for Learning Gene Networks	149
KYLE CHIPMAN AND AMBUJ SINGH	
6.1. Introduction	149
6.2. Methods	151
6.2.1. eQTL Data sets	151
6.2.2. LCMS Method for Learning a Prior Matrix of Causal Relationships	151
6.2.3. Bayesian Network Structure Learning	154
6.2.4. Integrating the Prior Matrix	155
6.2.5. Stochastic Causal Tree Method	156
6.3. Conclusion	161
7. Bayesian Causal Phenotype Network Incorporating Genetic Variation and Biological Knowledge	165
JEE YOUNG MOON, ELIAS CHAIBUB NETO, XINWEI DENG, AND BRIAN S. YANDELL	
7.1. Introduction	166
7.2. Joint Inference of Causal Phenotype Network and Causal QTLs	167
7.2.1. Standard Bayesian Network Model	168
7.2.2. HCGR Model	169
7.2.3. Systems Genetics and Causal Inference	170

7.2.4.	QTL Mapping Conditional on Phenotype Network Structure	172
7.2.5.	Joint Inference of Phenotype Network and Causal QTLs	173
7.3.	Causal Phenotype Network Incorporating Biological Knowledge	174
7.3.1.	Model	175
7.3.2.	Sketch of MCMC	178
7.3.3.	Summary of Encoding of Biological Knowledge	180
7.4.	Simulations	183
7.5.	Analysis of Yeast Cell-Cycle Genes	185
7.6.	Conclusion	188
8.	Structural Equation Models for Studying Causal Phenotype Networks in Quantitative Genetics	196
	GUILHERME J. M. ROSA AND BRUNO D. VALENTE	
8.1.	Introduction	196
8.2.	Classical Linear Mixed-effects Models in Quantitative Genetics	197
8.3.	Mixed-effects Structural Equation Models	202
8.4.	Data-driven Search for Phenotypic Causal Relationships	204
8.4.1.	General Overview	204
8.4.2.	Search Algorithms	206
8.5.	Inferring Causal Structures in Genetics Applications	207
8.5.1.	Genotypic information as Instrumental Variable	207
8.5.2.	Accounting for Polygenic Confounding Effects	208
8.6.	Concluding Remarks	210
	Part IV. GENETIC ASSOCIATION STUDIES	
9.	Modeling Linkage Disequilibrium and Performing Association Studies through Probabilistic Graphical Models: a Visiting Tour of Recent Advances	217
	CHRISTINE SINOQUET AND RAPHAËL MOURAD	
9.1.	Introduction	218
9.2.	Modeling Linkage Disequilibrium	219
9.2.1.	General Panorama	221
9.2.2.	Decomposable Markov Random Fields	221
9.2.3.	Bayesian Network-based Approaches without Latent Variables	223
9.2.4.	Bayesian Network-based Approaches with Latent Variables	224
9.2.5.	Recapitulation	226
9.3.	Single-SNP Approaches for Genome-wide Association Studies	228
9.3.1.	Integration of Confounding Factors	228
9.3.2.	GWAS Multilocus Approach	230
9.3.3.	Strengths and Limitations	235
9.4.	Identifying Epistasis at the Genome Scale	237
9.4.1.	Bayesian Network-based Approaches	237
9.4.2.	Markov Blanket-based Method	239
9.4.3.	Recapitulation	240
9.5.	Discussion	241
9.6.	Perspectives	242

10. Modeling Linkage Disequilibrium with Decomposable Graphical Models	247
HALEY J. ABEL AND ALUN THOMAS	
10.1. Introduction	248
10.2. Methods	249
10.2.1. Decomposable Graphical Models	249
10.2.2. Estimating Decomposable Graphical Models	251
10.2.3. Application to Diploid Data by Phase Imputation	254
10.2.4. Estimation on the Genome-Wide Scale	256
10.3. Applications	258
10.3.1. Phasing	258
10.3.2. Unconditional Simulation	260
10.3.3. Phenotypes and Covariates	261
10.3.4. Admixture Mapping	263
10.4. Application to Sequence Data	265
11. Scoring, Searching and Evaluating Bayesian Network Models of Gene-phenotype Association	269
XIA JIANG, SHYAM VISWESWARAN, AND RICHARD E. NEAPOLITAN	
11.1. Introduction	270
11.2. Background	270
11.2.1. Epistasis	270
11.2.2. Genome-wide association studies	271
11.3. A Bayesian Network Model	272
11.4. Scoring Candidate Models	273
11.4.1. Bayesian Network Scoring Criteria	273
11.4.2. Experiments	275
11.5. Searching over the Space of Models	278
11.5.1. Experiments	280
11.6. Determining Whether a Model is Sufficiently Noteworthy	280
11.6.1. The Bayesian Network Posterior Probability (BNPP)	282
11.6.2. Prior Probabilities	285
11.6.3. Experiments	287
11.7. Discussion and Further Research	290
12. Graphical Modeling of Biological Pathways in Genome-wide Association Studies	294
MIN CHEN, JUDY CHO, AND HONGYU ZHAO	
12.1. Introduction	295
12.2. MRF Modeling of Gene Pathways	296
12.3. A Bayesian Framework	300
12.3.1. Prior Specification and Likelihood Function	300
12.3.2. Posterior Distribution	302
12.3.3. Making Inference Based on the Posterior Distribution	304
12.3.4. Numerical Studies	305
12.3.5. Real Data Example—Crohn's Disease Data	309
12.4. Discussion	312

13. Bayesian, Systems-based, Multilevel Analysis of Associations for Complex Phenotypes: from Interpretation to Decision	318
PÉTER ANTAL, ANDRÁS MILLINGHOFFER, GÁBOR HULLÁM, GERGELY HAJÓS, PÉTER SÁRKÖZY, ANDRÁS GÉZSI, CSABA SZALAI, AND ANDRÁS FALUS	
13.1. Introduction	319
13.2. Bayesian network-based Concepts of Association and Relevance	320
13.2.1. Association and Strong Relevance	320
13.2.2. Stable Distributions, Markov Blankets and Markov Boundaries	322
13.2.3. Further relevance types	323
13.2.4. Necessary Subsets and Sufficient Supersets in Strong Relevance	326
13.2.5. Relevance for Multiple Targets	327
13.3. A Bayesian View of Relevance for Complex Phenotypes	328
13.3.1. Estimating the Posteriors of Complex Features	330
13.3.2. Sufficiency of the Data for Full Multivariate Analysis	332
13.3.3. Rate of Learning: Effect of Feature and Model Complexity	333
13.3.4. Bayesian network-based Bayesian Multilevel Analysis of Relevance	336
13.3.5. Posteriors for Multiple Target Variables	339
13.3.6. Subtypes of Strong and Weak Relevance	340
13.3.7. Interaction-redundancy Scores Based on Posteriors of Strong Relevance	342
13.4. Bayes Optimal Decisions about Multivariate Relevance	344
13.4.1. Optimal Decision about Univariate Relevance	344
13.4.2. Optimal Bayesian Decision to Control FDR	345
13.4.3. General Bayes Optimal Decision about Multivariate Relevance	348
13.5. Knowledge Fusion: Relevance of Genes and Annotations	350
13.6. Conclusion	352
Part V. EPIGENETICS	
14. Bayesian Networks in the Study of Genome-wide DNA Methylation	363
MEROMIT SINGER AND LIOR PACTHER	
14.1. Introduction to Epigenetics	364
14.2. Next-generation Sequencing and DNA Methylation	365
14.2.1. Assaying Genome-wide DNA Methylation	366
14.2.2. The methyl-Seq Method	368
14.3. A Bayesian network for methyl-Seq Analysis	370
14.3.1. Notation	371
14.3.2. A Generative Model	371
14.3.3. Parameter Learning and Inference of Posterior Probabilities	372
14.4. Genomic Structure as a Prior on Methylation Status	375
14.5. Application: Methylation Typing the Human Neutrophil	379
14.5.1. Unmethylated Clusters	379
14.6. Conclusions	381

15. Latent Variable Models for Analyzing DNA Methylation	387
E. ANDRÉS HOUSEMAN	
15.1. Introduction	388
15.2. Latent Variable Methods for DNA Methylation in Low-dimensional Settings	390
15.2.1. Discrete Latent Variables	391
15.2.2. Continuous Latent Variables	392
15.3. Latent Variable Methods for DNA Methylation in High-dimensional Settings	396
15.3.1. Model-based Clustering: Recursively Partitioned Mixture Models	396
15.3.2. Semi-Supervised Recursively Partitioned Mixture Models	399
15.4. Conclusion	401

Part VI. DETECTION OF COPY NUMBER VARIATIONS

16. Detection of Copy Number Variations from Array Comparative Genomic Hybridization Data Using Linear-chain Conditional Random Field Models	409
XIAOLIN YIN AND JING LI	
16.1. Introduction	410
16.2. aCGH Data and Analysis	411
16.2.1. aCGH Data	411
16.2.2. Existing Algorithms	412
16.3. Linear-chain CRF Model for aCGH Data	413
16.3.1. Feature Functions	415
16.3.2. Parameter Estimation	417
16.3.3. Evaluation Methods	421
16.4. Experimental Results	421
16.4.1. A Real Example	421
16.4.2. Simulated Data	424
16.5. Conclusion	425

Part VII. PREDICTION OF OUTCOMES FROM HIGH-DIMENSIONAL GENOMIC DATA

17. Prediction of Clinical Outcomes from Genome-wide Data	431
SHYAM VISWESWARAN	
17.1. Introduction	431
17.2. Challenges with Genome-wide Data	432
17.3. Background	433
17.3.1. The Naive Bayes Model	433
17.3.2. Bayesian Model Averaging	434
17.3.3. Alzheimer's Disease	434
17.4. The Model-Averaged Naive Bayes (MANB) Algorithm	435
17.4.1. Overview of the MANB Algorithm	435
17.4.2. Details of the MANB Algorithm	436

