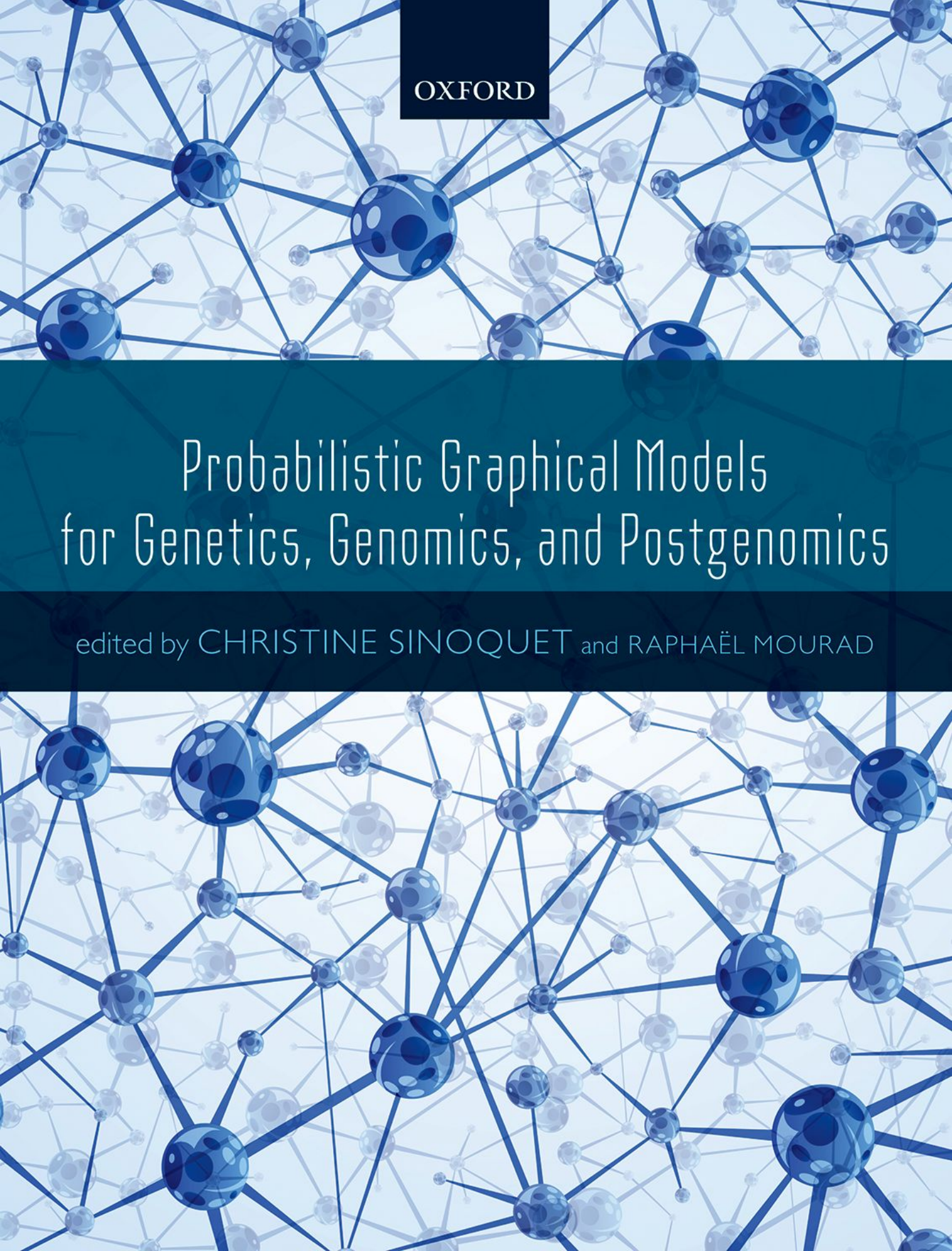
OXFORD
Probabilistic Graphical Models
for Genetics, Genomics, and Postgenomics
edited by CHRISTINE SINOQUET and RAPHAËL MOURAD

Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics

Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics

Edited by

CHRISTINE SINOQUET

Editor in chief

and

RAPHAËL MOURAD

Editor

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2014

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2013953773

ISBN 978–0–19–870902–2

Printed in Great Britain by
Clays Ltd, St Ives plc

A NOTE FROM THE EDITOR

To my loved ones.

The idea of editing a collective book about probabilistic graphical models in genetics arose in the spring of 2011. This project was fortunate to obtain the support of researchers at the forefront of innovation in this domain. From then on, in the back of my mind was always present the concern of honoring the confidence of the invited authors by achieving the project within a decent time frame. May they all be warmly thanked for their trust and their deep investment in this project, as well as for all the intellectually stimulating exchanges we had.

A collective book—not proceedings—is much more than the compendium of the scientific contributions that supports it, however invaluable these contributions are by themselves; and this comes at a cost. The edition and compilation of this book drew on any time reserve that could be ferreted out of a researcher's timetable. Using a metaphor borrowed from carpentry, sanding, smoothing, and polishing again and again the job took quite a while before I was able to apply the undercoat paint layers and the top varnish.

I was therefore converted into a sort of Benedictine monk, of the specific kind that monitors a whole reviewing process, reads two or three successive versions of each chapter, writes a submission package to gain the support of the prestigious publishing group targeted, controls bibliographical references, checks figures, tables, captions, homogenizes the presentation throughout the whole draft, indexes the whole book, and benedictinely runs the LaTeX compiler until it does not scream anymore. As I confess a fierce determination to separate professional and private lives, this book has been elaborated at my office at the university, during innumerable weekends as well as countless late, or even very late, evenings. By the way, this specific time schedule offered me the opportunity to frequently hear the owl living in the little wood in front of the lab, and to catch sight of such furtive animals as badgers and foxes, which one would never think would live in a university campus.

Fortunately, these months of labor have reached their term within the time the tribe I ordinarily belong to was still able to recognize me. May they all be thanked for their patience and their attentive listening and concern about the progress of the project.

I am in special debt to Keith Mansfield from Oxford University Press (OUP), for his support of the project from the very start, and not least for his encouragement and his valuable advice and guidance in the preparation of the proposal dossier for OUP. Complying to high standards is

the lot if one wishes to publish with OUP. Driven by the confidence of the invited authors of the project and of my joint editor, I had therefore an obligation: obtain the sesame to be allowed to press ahead.

I also wish to warmly thank Clare Charles from Oxford University Press for her efficient management and attentive monitoring of the production step.

C.S., June, 2014

PREFACE

At the crossroads between statistics and machine learning, probabilistic graphical models provide a powerful formal framework to model complex data. Examples of probabilistic graphical models are Bayesian networks and Markov random fields, which represent two of the most popular classes of such models. With the rapid advancements of high-throughput technologies and the ever decreasing costs of these next-generation technologies, a fast-growing volume of biological data of various types–the so-called omics–is in need of accurate and efficient modeling methods, prior to further downstream analysis. As probabilistic graphical models are able to deal with high-dimensional data and non-linear dependences, it is foreseeable that such models will have a prominent role to play in advances in genome-wide analyses.

Currently, few people are specialists in the design of cutting-edge methods using probabilistic graphical models for genetics, genomics, and postgenomics. This seriously hinders the diffusion of such methods. The prime aim of this book is therefore to bring the concepts underlying these advanced models within understanding of a broader audience of scientists, engineers, and graduate students.

If they are not specialists of probabilistic graphical models, bioinformaticians, statisticians, biostatisticians, and experts in statistical genetics with an intuition that their solution to a problem should involve such models are compelled to glean incomplete information from publications. We are not even talking of surveys whose consultation will never allow launching out into the design of advanced methods. Some academic courses may well be delivered here and there, that dwell on cutting-edge approaches using probabilistic graphical models for the targeted topics; neither are such courses widely available for the potentially interested audience, nor do they cover a sufficiently illustrative set of models and applications.

The target readers of this book include researchers and engineers as well as graduate students starting a master's or a PhD thesis. Besides, if there is one area where transdisciplinarity is the daily lot, it is the advanced analysis of genome-wide data. Constructive cooperation with a domain specialist requires the ability to hold a productive dialogue, which therefore demands a deep understanding of the models as well as a solid background regarding these models. Often, scientists from different fields such as genetics, statistics, or computer science do not use the same scientific language, and this might lead to confusion and misunderstanding. Bridging the gap between different scientific worlds thus helps scientists to better communicate, and from a

higher perspective, contributes to the emergence of new fields of research. Currently, the only solution for such people to gain a deep understanding is finding spare time to gather information to learn from it. The book intends to spare such readers this task.

Hopefully, this book will be of equal interest, if still not higher, for the graduate students supervised by members of the aforementioned audience. Depending on their academic institution, students taught computational methods for genetics, genomics, or postgenomics rarely have access to a course presenting the advanced use of probabilistic graphical models in such fields. One reason for this lies in the fact that these models and their potentialities have only rather recently created renewed interest in genetics in the broad sense. Another reason might be the lack of experts possessing this two-fold skill in these students' institutions. Besides, a few hours taught on the subject are not sufficient to provide both enough material and hindsight on the topic. This book attempts to fill this gap.

This book is also designed to help experts in machine learning grasp the interest in designing advanced methods based on probabilistic graphical models in transdisciplinary collaborations.

This book arises out of a six-year collaboration between its scientific editors. Our various interests in computer science, machine learning, applied mathematics, Bayesian statistics, applications in genetics, genomics, and postgenomics have found in probabilistic graphical models a breeding ground for both our own investigations and the preparation and direction of this book. Besides, coming from different backgrounds, we found a common ground in demanding the highest self-containedness in the contributions of the invited authors. In addition to the intrinsic richness of these contributions, our guiding thread was then providing added value through accessibility for non-specialists of probabilistic graphical models, with no concession on the informativeness of the book's contents.

We have been fortunate to obtain the widest consent regarding invited authors' participation in our project. We subsequently enjoyed a fruitful period of dense exchanges with these authors, who accepted this extra workload.

The book is divided into a general introduction, a tutorial on probabilistic graphical networks, and six main sections devoted to specific application fields in genetics (in the broad sense). The introductory chapter aims at providing a minimal background for readers that are not familiar with biology or need information about the high-throughput biological data addressed by the models described in the book. Moreover, such terms and expressions as genetics, genomics, postgenomics, systems biology, and integrative biology are clarified. Indeed, a leitmotif of the book is the integration of heterogeneous sources of omics data, to boost downstream biological applications. Finally, this introduction provides the motivation for using probabilistic graphical models to handle high-throughput biological data and provides a brief evocation of the use of probabilistic graphical networks in the six applications highlighted by the book: gene network inference, causality discovery, association genetics, epigenetics, detection of copy number variations, and prediction of outcomes from high-dimensional genomic data.

The essentials for understanding probabilistic graphical models are offered in a tutorial at the beginning of the book. This tutorial was carefully designed to be accessible to the largest audience. Since the concepts and techniques presented in this tutorial may require broader and non-trivial knowledge, accessibility and self-containedness were again the targeted objectives.

Together with a thorough review chapter focusing on selected domains in genetics, fourteen chapters illustrate the design of advanced approaches, for the six abovementioned applications. This book offers a lot of new insights that could only be gleaned from the literature available through excruciating labor. The chapters are self-contained, and they can be read independently of each other.

C. S. and R. M.

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 PACHTER
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: Methyltyping 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

17.5. Evaluation Protocol 438
17.5.1. Data set 438
17.5.2. Protocol 438
17.6. Results 439
17.7. Conclusion 440

Index *447*

ABBREVIATIONS

A	Adenine
aCGH	array comparative genomic hybridization
AIC	Akaike information criterion
AUC	area under the receiver operating characteristic curve
BD	Bayesian Dirichlet
BDe	Bayesian Dirichlet equivalent
BDeu	Bayesian Dirichlet equivalent uniform
BIC	Bayesian information criterion
BN	Bayesian network
BNPP	Bayesian network posterior probability
C	cytosine
cDNA	complementary deoxyribonucleic acid
CGH	comparative genomic hybridization
CNP	copy number polymorphism
CNV	copy number variation
CRF	conditional random field
DAG	directed acyclic graph
DDAG	direct directed acyclic graph
DGM	decomposable graphical model
DNA	deoxyribonucleic acid
D-map	dependence map
EM	expectation-maximization
ER	estrogen receptor
ER^+	estrogen receptor positive
ER^-	estrogen receptor negative
eQTL	expression quantitative trait loci

FDR	false discovery rate
FISH	fluorescence *it situ* hybridization
FLTM	forest of latent tree models
G	Guanine
GGM	Gaussian graphical model
GOGE	genetics of gene expression
GWAS	genome-wide association study
HCGR	homogeneous conditional Gaussian regression
HMM	hidden Markov model
HRMF	Hidden random Markov field
IC	inductive causation
IG	interval graph
I-map	independence map
i.i.d.	identically and independently distributed
KEGG	Kyoto Encyclopedia of Genes and Genomes
LARS	least angle regression
LASSO	least absolute shrinkage and selection operator
LCM	latent class model
LCMS	likelihood-based causality model selection
LD	linkage disequilibrium
LOD	log-odds ratio
LTM	latent tree model
LLS	log-likelihood score
MDL	minimum description length
mRNA	messenger ribonucleic acid
MCMC	Markov chain Monte Carlo
MME	mixed model equations
MRF	Markov random field
PCR	polymerase chain reaction
PDAG	partially directed acyclic graph
PGM	probabilistic graphical model
MAP	maximum *a posteriori*
QTL	quantitative trait loci
RNA	ribonucleic acid
RNA-seq	RNA sequencing
ROC	receiver operating characteristic

SBM	stochastic block model
SCT	stochastic causal tree
SEM	structural equation model
SEM	structural expectation-maximization
SNP	single nucleotide polymorphism
SSR	sum of squares
SSTO	total sum of squares
T	thymine
UG	undirected graph
VLMC	variable length Markov chain

LIST OF CONTRIBUTORS

Abel, Haley J.
Division of Statistical Genomics
Washington University School of Medicine
St. Louis, USA

Antal, Péter
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary

Chaibub Neto, Elias
Department of Computational Biology
Sage Bionetworks
Seattle, USA

Charbonnier, Camille
LaMME (Laboratoire de Mathématique et Modélisation d'Evry)
UMR CNRS 8071, USC INRA
Évry, France

Current address:
CNR-MAJ, Rouen, Lille and Paris Salpetriere
University Hospitals
Rouen, France

Chen, Min
Department of Mathematical Sciences
University of Texas at Dallas
Richardson, USA

Chipman, Kyle
Department of Computer Science & Biomolecular Science and Engineering
University of California
Santa Barbara, USA

Chiquet, Julien
LaMME (Laboratoire de Mathématique et Modélisation d'Evry)
UMR CNRS 8071, USC INRA
Évry, France

Cho, Judy
Icahn School of Medicine at Mount Sinai
New York, USA

Deng, Xinwei
Department of Statistics
Virginia Polytechic Institute and State University
Blacksburg, USA

Falus, András
Department of Genetics, Cell and Immunobiology
Semmelweis University
Budapest, Hungary

Gézsi, András
Department of Genetics, Cell and Immunobiology
Semmelweis University
Budapest, Hungary

Guedj, Mickaël
Department of Bioinformatics and Biostatistics
Pharnext
Issy-les-Moulineaux, France

Hajós, Gergely
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary

Houseman, Andrés E.
College of Public Health and Human Sciences
Oregon State University
Corvallis, USA

Hullám, Gábor
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary

Jeanmougin, Marine
LaMME (Laboratoire de Mathématique et Modélisation d'Evry)
UMR CNRS 8071, USC INRA
Évry, France

Current address:
Department of Immunology, Institut Curie

INSERM U932
Paris, France

Jiang, Xia
Department of Biomedical Informatics
University of Pittsburgh
Pittsburgh, USA

Kiiveri, Harri
CSIRO Computational Informatics
The Leuwin Centre
Floreat, Australia

Li, Jing
Electrical Engineering and Computer Science Department
Case Western Reserve University
Cleveland, USA

Millinghoffer, András
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary

Moon, Jee Young
Department of Statistics
University of Wisconsin, Madison
Madison, USA

Current address:
Department of Genetics and Genomic Sciences
Mount Sinai School of Medicine
New York, USA

Mourad, Raphaël
LINA, UMR CNRS 6241
Computer Science Institute of Nantes-Atlantic
Nantes University/Polytechnic Institute
Nantes, France

Current address:
Computational Biology Institute
Mantpellier, France

Neapolitan, Richard E.
Division of Biomedical Informatics
Department of Preventive Medicine
Northwestern University Feinberg School of Medicine
Chicago, USA

Pachter, Lior
Department of Mathematics
University of California Berkeley
Berkeley, USA

Rodriguez Zas, Sandra L.
Department of Animal Sciences
University of Illinois Urbana-Champaign
Urbana, USA

Rosa, Guilherme J. M.
Department of Animal Sciences,
Department of Biostatistics & Medical Informatics,
University of Wisconsin-Madison
Madison, USA

Sárközy, Péter
Department of Measurement and Information Systems
Budapest University of Technology and Economics
Budapest, Hungary

Singer, Meromit
Computer Science Division
University of California Berkeley
Berkeley, USA

Singh, Ambuj
Department of Computer Science
Department of Biomolecular Science and Engineering
University of California Santa Barbara
Santa Barbara, USA

Sinoquet, Christine
LINA, UMR CNRS 6241
Computer Science Institute of Nantes-Atlantic
University of Nantes
Nantes, France

Southey, Bruce R.
Department of Animal Sciences
University of Illinois Urbana-Champaign
Urbana, USA

Szalai, Csaba
Department of Genetics, Cell and Immunobiology
Semmelweis University
Budapest, Hungary

Thomas, Alun
Division of Genetic Epidemiology
University of Utah
Salt Lake City, USA

Valente, Bruno D.
Department of Animal Sciences
University of Wisconsin-Madison
Madison, USA

Visweswaran, Shyam
Department of Biomedical Informatics
University of Pittsburgh
Pittsburgh, USA

Yandell, Brian S.
Department of Statistics and Horticulture
University of Wisconsin–Madison
Madison, USA

Yin, XiaoLin
Electrical Engineering and Computer Science Department
Case Western Reserve University
Cleveland, USA

Zhao, Hongyu
Department of Biostatistics
Yale School of Public Health
New Haven, USA

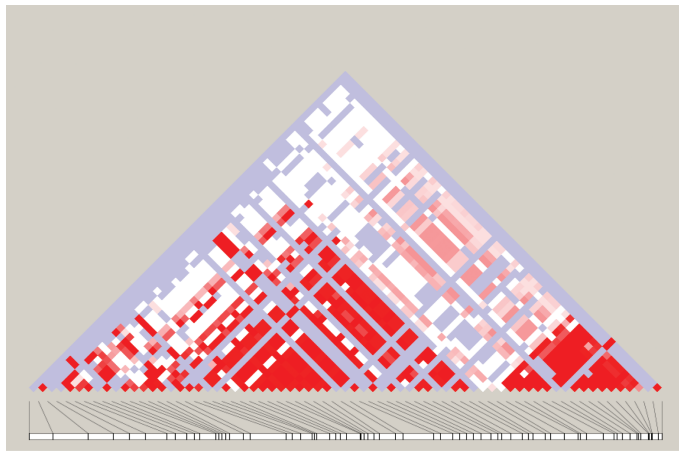

Fig 9.1 Linkage disequilibrium (LD) plot of a 500 kb SNP sequence. Human genome, chromosome 1, region [10 000 kb - 10 500 kb]. LD is revealed through the matrix of pairwise dependences between genetic markers. For a pair of SNPs, the color shade is all the darker as the correlation between the two SNPs is high.

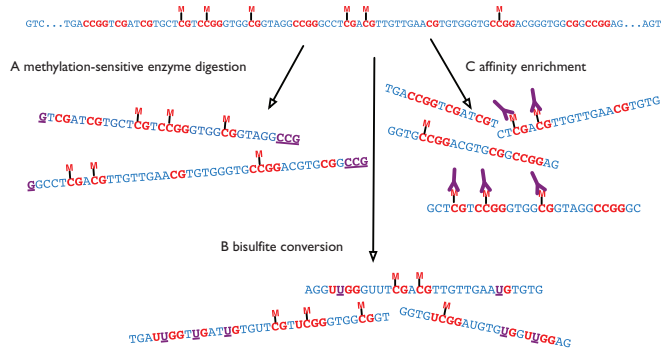

Fig 14.1 Three common techniques for genome-scale annotation of DNA methylation. (A) Enzyme digestion: the genomic DNA is digested with a methylation-sensitive restriction enzyme such as *HpaII*, which digests unmethylated CCGG sites. (B) Bisulfite conversion: converts cytosines that are not methylated to uracil. (C) Affinity enrichment: methylated cytosines in methylated regions are bound by antibodies or methyl-CpG binding proteins. M denotes methylation site.

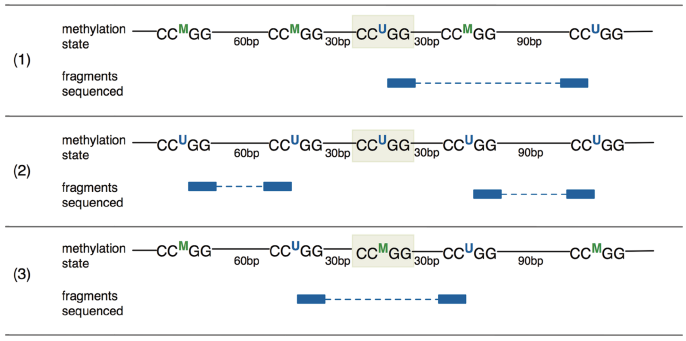

Fig 14.2 The methylation state of a site cannot always be determined from the number of fragments that originated at that site. In many cases, the methylation state of a site cannot be determined from the extent to which it was present at the end of sequenced fragments but can be determined by integrating sequencing data from its neighborhood. bp, base pair; M, methylated; U, unmethylated.

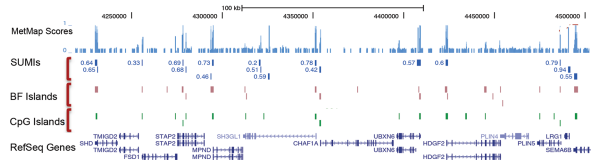

Fig 14.5 A section of the genome showing site-specific methylation scores (top panel) and unmethylated clusters (SUMIs, second panel) as inferred from one of human neutrophil samples. For the site-specific scores, a score of 0 determines a site as fully methylated. The third and fourth panels show BF islands as annotated by [5] and UCSC islands, respectively. While there is substantial overlap between SUMIs and the islands inferred by the sequence-based methods, a few novel SUMIs are seen in this figure, one of them at a transcription start site. RefSeq denotes genes annotated in the National Center for Biotechnology Information reference sequence database.

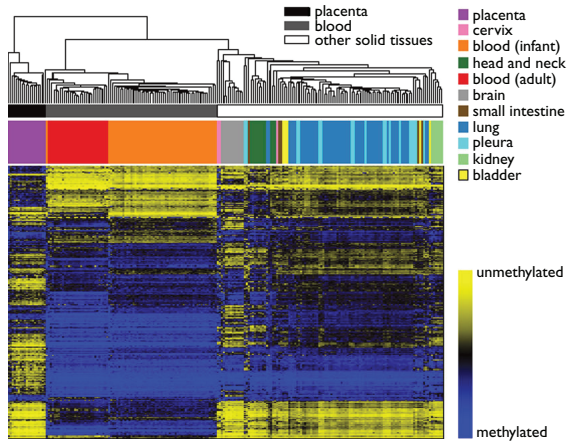

Fig 15.1 Clustering heat map showing DNA methylation patterns for 11 normal tissues [8]. Each cell represents an average beta value from the GoldenGate assay (Illumina). Rows represent one of 500 CpG dinucleotides, columns represent one of 211 individual samples.

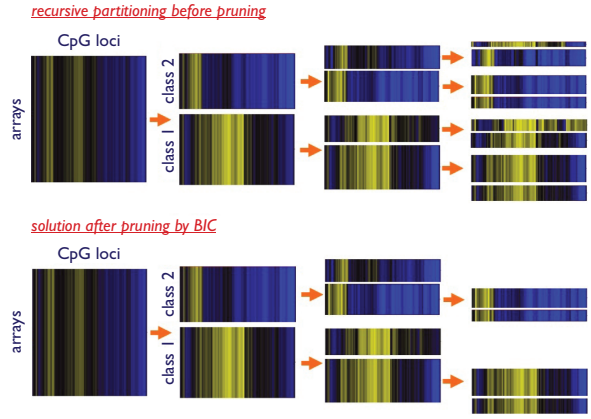

Fig 15.3 Schematic representation of the RPMM. Rows represent individual specimens or arrays, columns represent individual CpG loci. Initially, each array is assumed to be drawn from the same multivariate distribution consisting of a distinct distribution for each CpG (indicated by color). The data set is partitioned recursively into component data subsets using a two-part mixture model. Along the way, BIC is used to prune the tree, so that partitions that are likely to be unstable are never attempted.

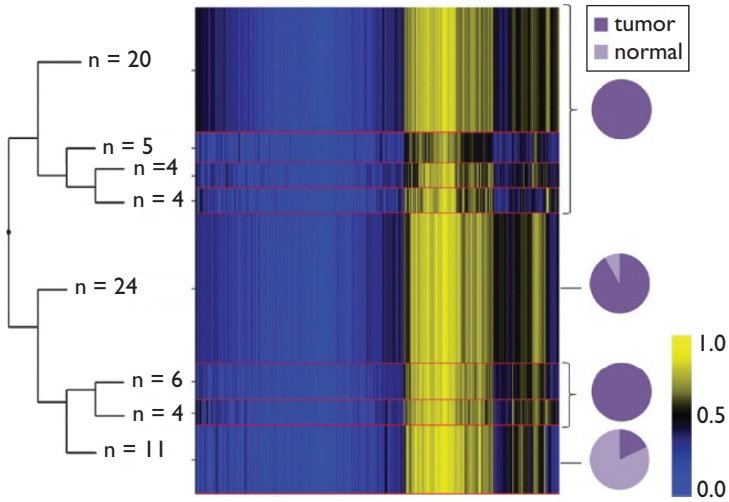

Fig 15.4 Recursive partitioning mixture model classification of normal and tumor head and neck tissues. The model was based on methylation values of 1413 autosomal loci measured using the GoldenGate assay produced by Illumina, and resulted in eight classes whose average methylation values are represented in the heat map. Distribution of normal and tumor samples within each class is depicted in pie charts on the right. Reproduced from [28], Figure 1B.

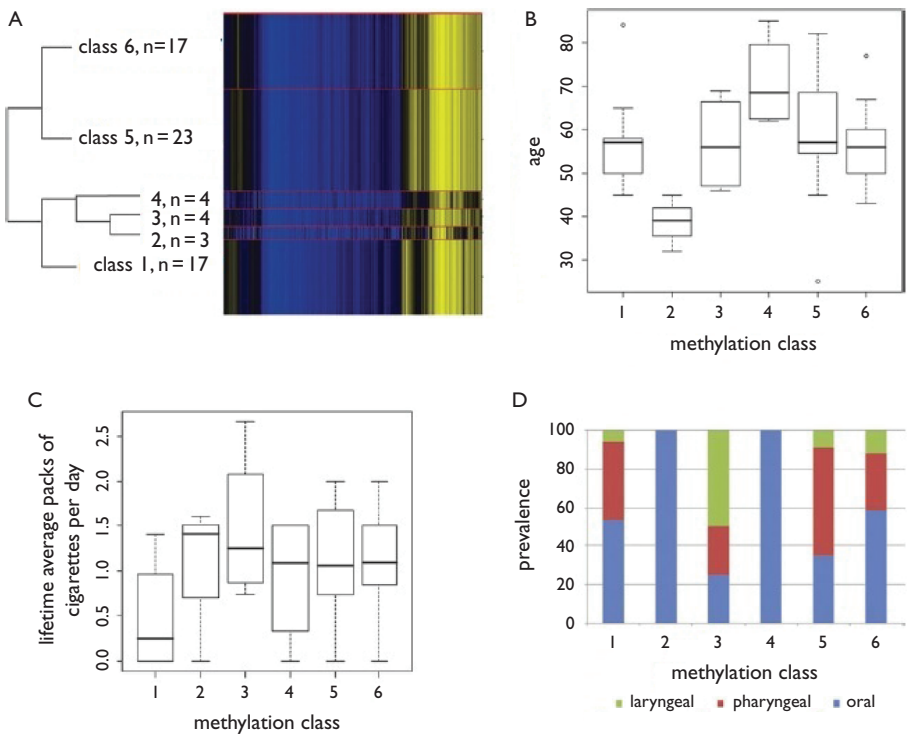

Fig 15.5 Recursive partitioning mixture model classification of HNSCCs. A) Six classes with average methylation values across loci depicted in the heat map. Associations of class membership with age, lifetime average packs of cigarettes smoked per day, and tumor location are shown in B, C, and D, respectively. Reproduced from [28], Figure 2A–D.

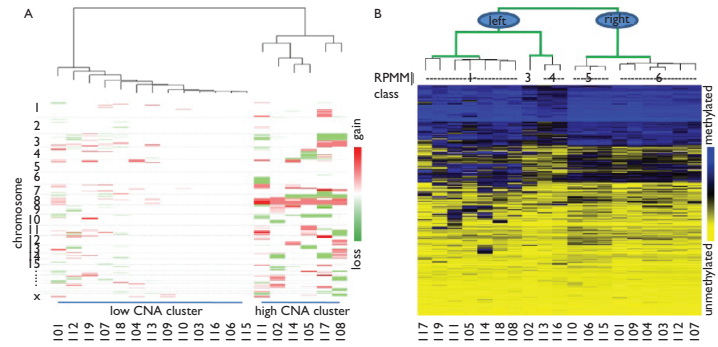

Fig 15.6 A) DNA copy number states are arranged by chromosome for 500 000 SNP loci. Copy number is red for amplified regions with three or more copies, white for two normal copies, and green for allele loss (no copies). Tumors are ordered by unsupervised hierarchical clustering and are dichotomized into low/high clusters of copy number alterations (CNAs). B) Methylation loci (more methylated = blue, less methylated = yellow) are grouped by Euclidean distance, and tumors samples are ordered first by RPMM class structure (green branches) then by simple hierarchical clustering (black branches). Tumor IDs are provided below each plot and "high CNA" samples are colored orange for reference. Reproduced from [33].

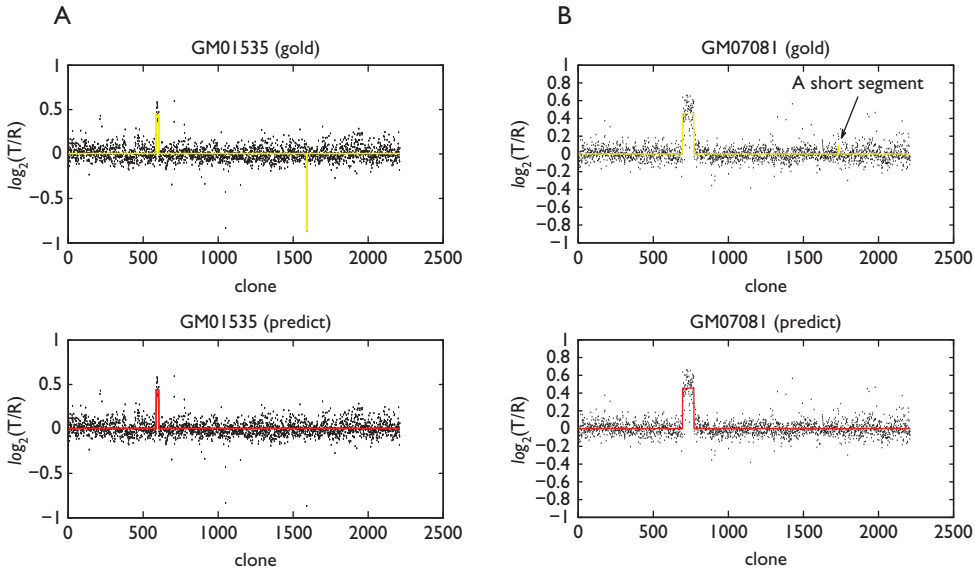

Fig 16.3 Predicted breakpoints by CRF-CNV (bottom) versus true breakpoints (top) on the two cell lines A) GM01535 and B) GM07081. R, gene expression level in the reference sample; T, gene expression level in the testing sample.

PART I

Introduction

CHAPTER 1

Probabilistic Graphical Models for Next-generation Genomics and Genetics

CHRISTINE SINOQUET

The explosion in "omics" and other types of biological data has increased the demand for solid, large-scale statistical methods. These data can be discrete or continuous, dependent or independent, and from many individuals or tissue types. There might be millions of correlated observations from a single individual or observations at different scales and levels, in addition to covariates. The study of living systems encompasses a wide range of concerns, from prospective to predictive and causal questions, reflecting the multiple interests in understanding biological mechanisms, disease etiology, predicting outcomes, and deciphering causal relationships in data. Precisely, probabilistic graphical models provide a flexible statistical framework that is suitable to analyze such data. Notably, graphical models are able to handle dependences within data, which is an almost defining feature of cellular and other biological data.

This introductory chapter aims at providing a minimal background for readers that are not familiar with biology or need information about the high-throughput biological data the models described in the book deal with. The chapter also provides the motivation for using probabilistic graphical models to handle high-throughput biological data. The chapter is organized as follows. Section 1.1 describes the fine-grained components studied by molecular biology and provides the definitions of key terms. The biological information allows to conduct studies in the fields of genetics, genomics, and postgenomics. The respective scopes of these three domains are first defined. In these domains, various types of analyses allow inference of knowledge about one or several levels of description of living systems. Section 1.2 then focuses on the multiple levels of biological organization of living systems to which the chapters of this book are connected. This section takes the opportunity to clarify which definition of the expression "systems biology" the

Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics. First Edition. Christine Sinoquet & Raphaël Mourad (Eds). Published in 2014 by Oxford University Press.

book is interested in. The definition of "integrative biology" is also clarified. In the era of modern genomics, the data are provided by high-throughput technologies; Section 1.3 briefly surveys the types of data covered by the book. Finally, this section emphasizes the complexity of the biological data available nowadays, and stresses various issues encountered when handling such data. This emphasis serves as a transition to Section 1.4, which starts advocating the use of probabilistic graphical models in genetics, genomics, and postgenomics: thus can be evidenced and exploited dependences within various biological components, with the aim of explanation and prediction. The chapter ends with a brief evocation of the use of probabilistic graphical networks in the six applications highlighted by the book: gene network inference, causality discovery, association genetics, epigenetics, detection of copy number variations, and prediction of outcomes from high-dimensional genomic data.

1.1 Fine-grained Description of Living Systems

1.1.1 DNA and the Genome

Except in viruses, the cell is the smallest structural unit of all living organisms that is capable of independent functioning, through metabolic activities. The metabolism encompasses all chemical transformations within the cell. In contrast with the procaryotic cell (e.g., bacteria), the eukaryotic cell is typically described as possessing a nucleus isolated by a membrane from the rest of the cell (cytoplasm); the nucleus contains the majority of the hereditary material, called the genome. In prokaryotes, the hereditary material is not bound within a nucleus. All the applications described in this book address the human genome, which explains our focus on eukaryotic cells. Under the influence of environmental factors, the genome plays an important role in the development of the individual's observable features (also called **phenotypes**). For instance, it is known that genes influence race, hair and eye color, gender, height, and weight.

In each eukaryotic cell of a living organism, the same genetic information is encoded in a biochemical molecule, the DNA. The DNA molecule is double-stranded, and it is twisted into a helix. Each strand consists of a long polymer of nucleotides (or bases). The genome is encoded through an alphabet of four bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The two strands of the DNA molecule are paired, based on hybridization properties: A and T (respectively C and G), on opposite strands, are physically connected together as *complementary* bases. DNA molecules determine the synthesis of proteins, via intermediary messenger ribonucleic acid (mRNA) molecules: mRNA is produced from DNA through the **transcription** step; proteins are produced from mRNA by the **translation** step. The deshybridization property, which locally frees the two DNA strands, is involved in the replication and transcription processes. The replication step produces a DNA copy from a DNA molecule; the transcription step produces a single-strand RNA molecule from a double-strand DNA molecule. One of the most revolutionary levers in science in the twentieth century is the polymerase chain reaction (PCR), which exploits the hybridization and deshybridization properties; PCR thus allows to obtain several million identical copies from a single DNA fragment.

In eukaryotes, the genome is packaged into chromosomes, each consisting of a specific DNA sequence tightly packed into a complex series of coils thanks to proteins (i.e., histones). The human genome contains approximately 3.4 billion base pairs of DNA packaged into 23 chromosomes. Most cells in the body, except female ova and male sperm, are diploid. Diploidy means that such cells possess two sets of homologous chromosomes. Therefore, each cell contains a total

of 6.8 billion base pairs of DNA. If (virtually) laid end to end, the 46 DNA molecules in each human cell would produce a two-meter long sequence.

Between any two humans, genetic variation roughly amounts to 0.1%. Thus, on average, about one base pair out of every 1000 is different between any two individuals. Various types of DNA polymorphism are known: the most common type is the single nucleotide polymorphism (SNP), where genetic variations consist in single base-pair differences; moreover, another characteristic of SNP is that over the four possible nucleotides (A, G, C, and T), only two variants are exhibited over a studied population. Other less frequent types of polymorphisms include insertions, deletions, duplications, and rearrangements of segments of DNA, as well as differences in the numbers of copies of a given segment.

1.1.2 Genes and Proteins

Any DNA region that produces a functional RNA molecule is called a gene. In addition, the most well-known acception of the term "gene" relates to the class of genes that code for proteins. The human genome contains approximately 20 000 such genes. Proteins are large molecules that play most of roles in an organism. In a multicellular organism, proteins are required for the structure, function, and regulation of the organism's tissues and organs. For instance, **enzymes** catalyze an overwhelming part of the thousands of chemical reactions that occur in a cell; such proteins are thus essential to the production of the remaining organic biomolecules necessary for life. For example, the phenylalanine hydroxylase enzyme converts the amino acid phenylalanine into another amino acid, the tyrosine. Another crucial role is that of **transcription factors**. Such proteins bind to specific DNA sequences, alone or in a complex, to promote (activate) or block (repress) the positioning of the **RNA polymerase enzyme** on the DNA molecule. Both previous types of proteins thereby control the flow of genetic information from DNA to mRNA, and thus the formation of new protein molecules. Other proteins form the **structural** components of the cell. On a larger scale, they also allow an organism to move. For instance, actin filaments are structural proteins built up of multiple subunits; they help cells maintain their shape and are also involved in muscle contraction. **Storage** and **transport** are two other crucial functions performed by proteins: the proteins concerned bind to atoms or small molecules; transport throughout an organism is thus made possible. For instance, ferritin, a protein made up of 24 identical subunits, is involved in iron storage. Some proteins are messengers that transmit signals to coordinate biological processes between different cells, tissues, and organs. An example is the growth hormone, which regulates cell growth. We complete the enumeration of the vital functions fulfilled by proteins with the mention of antibodies. An antibody is a protein that binds to a specific foreign particle, such as a virus or a bacterium, to help protect an organism. For instance, immunoglobulin G is a type of antibody present in the blood.

1.1.3 Phenotype and Genotype

The **phenotype** of an organism is defined as the combination of the organism's observable characteristics or traits. In particular, phenotypes can be described at the lowest level of living systems, that is the cellular level: the definition of phenotype can be extended so as to designate characteristics that are only made observable by some technical procedure. Such traits are connected to the various levels of the scale through which a biological system is observed. Higher level traits include biochemical properties, physiological properties, development and morphology, phenology, and behavior. For instance, phenological traits comprise periodic biological phenomena,

such as flowering, breeding, and migration, in relation to climatic and habitat conditions. Even this level is subject to the influence of the genetic information carried by an organism. Phenotypes result from the expression of an organism's genes as well as the influence of environmental factors and the interplay between the two.

The **genotype** of an organism is defined as the set of alternate variations of genes expressed in some specific traits. Such traits are often expressed through the synthesis of proteins. In genetics, gene expression is the most fundamental level at which the genotype gives rise to the phenotype. In another common usage, the **genotype** of an organism is more systematically defined as the description of DNA variations, based on a set of genetic markers. **Genetic markers** are well-characterized loci of the genome, which represent many short windows in which to observe DNA polymorphism between individuals. In particular, the genotype of a diploid organism accounts for the DNA variants—or **alleles**—present at opposite loci, on the two homologous chromosomes of a pair of chromosomes. Comparing genotypes among a set of organisms of the same species is the key to deciphering the differences in phenotypes observed.

1.1.4 Molecular Biology, Genetics, Genomics, and Postgenomics

Molecular biology is the branch of biology that describes the molecular characteristics of the genome as DNA, RNA, and proteins. Various definitions can be provided for the terms genetics and genomics. In the scope of the present book, the word **genetics** designates the discipline that studies variations between the genomes of individuals in some population; this analysis of variations may focus on simple units (genetic markers) or on more complex units (genes). The definition of **genomics** generally encompasses the range of biotechnological and computational analyses related to genome sequencing, gene mapping, and genome annotation; **functional genomics** is the appropriate expression for this book. Functional genomics focuses on transcription, translation, and interactions between proteins. Notably, functional genomics includes the study of the transcriptome, through DNA chips, to describe and quantify gene expression: for example, gene expression correlation potentially indicates that genes belong to the same gene interaction network; the identification of differentially expressed genes, for instance between affected and unaffected individuals, allows to identify putative causes for a studied disease.

Beyond functional genomics, **postgenomics** takes a step further to encompass an increasingly large range of topics. All such topics essentially aim at teasing higher functional biological understanding out of raw data. These data allow different viewpoints on living organisms. Such viewpoints may be transcriptomics (analysis of gene expression level through mRNAs), proteomics (analysis of gene expression as proteins), and metabolomics (characterization of the small molecules that are intermediates and products of metabolism), to name but a few.

In genetics, genomics, and postgenomics, various types of analyses enable the inference of knowledge about one or several levels of description of living systems. The next section describes the levels that are addressed in this book.

1.2 Higher Description Levels of Living Systems

The activity and state of a multicellular organism may be described from different viewpoints: cell, organ or tissue, system (e.g., cardiovascular, nervous), and whole organism. In this book, methods based on probabilistic graphical models are described that allow the inference of knowledge about

various description levels of living systems. The chapters in this book deal with the following description levels:

- genome,
- transcriptome,
- gene interaction networks,
- phenotype.

Depending on the chapter, knowledge inference addresses a single level or deals with several levels. We introduce this section by giving a flavor of the complexity of the processes and the variety of actors involved in the life of a cell. Then, we focus on the multiple levels of biological organization of living systems to which the chapters of this book are connected.

1.2.1 Complexity in Cells

A eukaryotic cell consists of the nucleus and various organelles—the "organs" of the cell, for short—which are immersed in the cytoplasm (see Fig. 1.1). The cell is surrounded by a semi-permeable membrane. Although no exact number can be provided, the number of cells in an adult human body can be approximated as 10^{14}. A unique cell, the fertilized egg, is the origin of all these cells, through cell division. However, though they bear the same genetic information in their nucleus, the cells of a multicellular organism perform different specific tasks, depending on their location in organs or tissues: red blood cells exchange oxygen, muscle cells expand and contract, and cells in the immune system recognize pathogens. Modifications in gene expression play a key role in guiding and maintaining cell differentiation.

Extrinsic and intrinsic factors regulate gene expression in cells. The first category includes small molecules, secreted proteins, temperature, and oxygen. Within the organism, cells communicate with each other by sending and receiving secreted proteins (e.g., growth factors, morphogens,

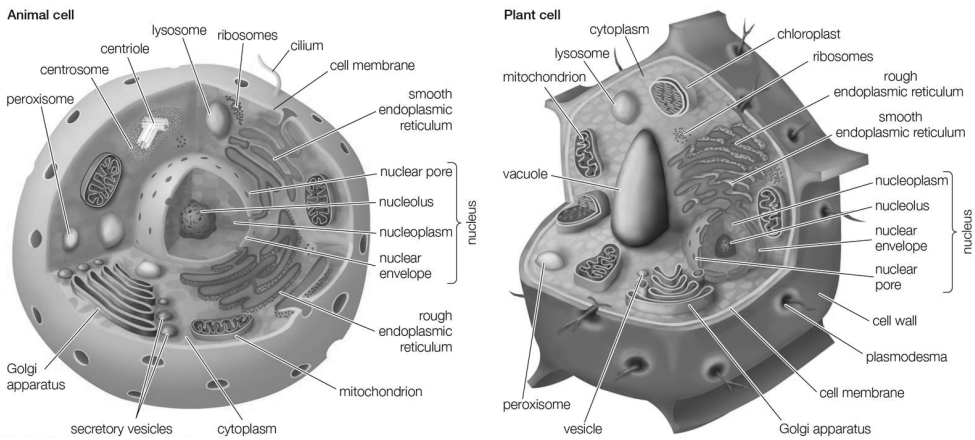

Fig 1.1 Typical animal and plant cells.

2010 Encyclopaedia Britannica, Inc.

cytokines). The receipt of these signaling molecules triggers intercellular signaling cascades that modify the expression of genes. Sequence-specific transcription factors are considered the most important and diverse mechanisms of gene regulation in cells [28]. An example of cell-intrinsic regulation is that of the modification of chromatin (DNA associated with histone proteins) by a cell's own machinery. A possible consequence of chromatin modification is a change in the accessibility of genes to transcription factors; the impact on gene expression may be positive or negative. Two major classes of chromatin modifications include DNA methylation and histone modification.

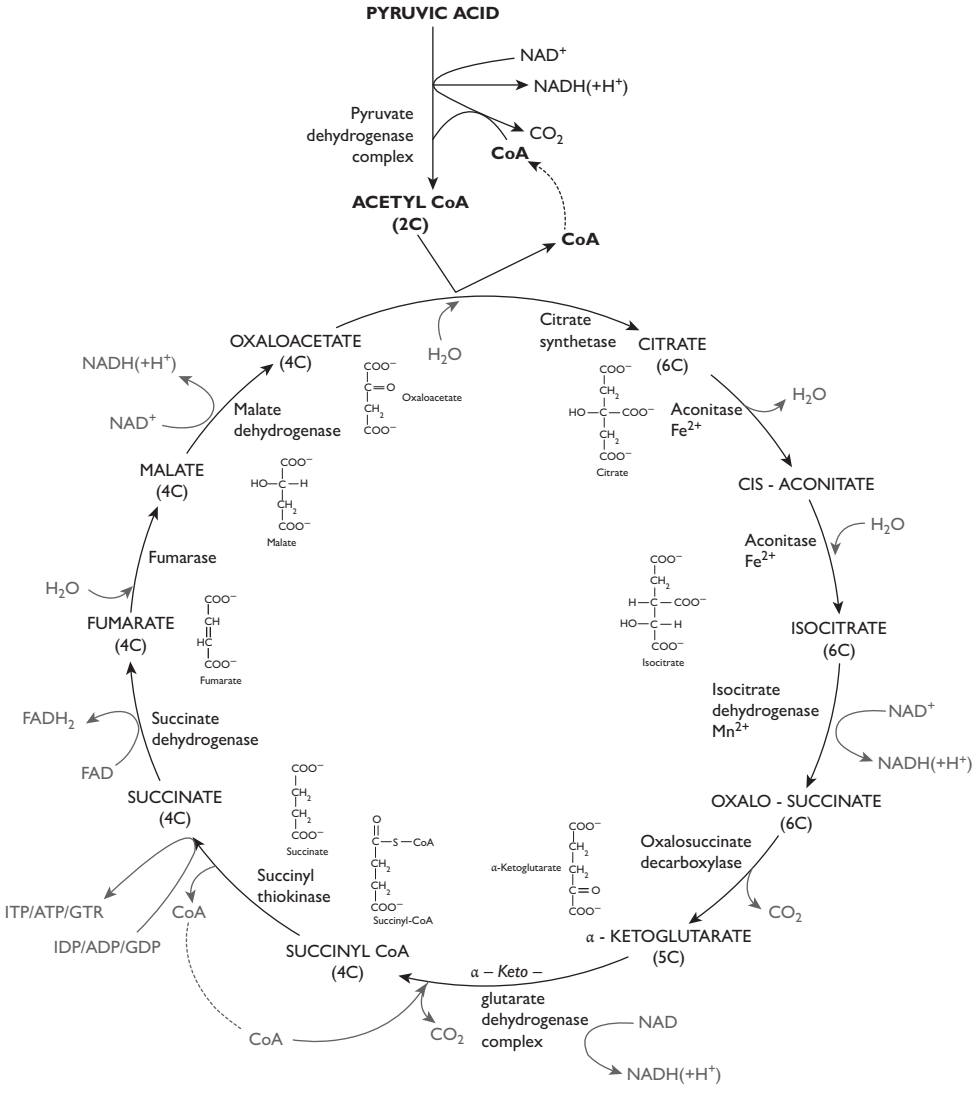

Fig 1.2 Krebs cycle. The first product of Krebs cycle is citric acid (citrate).

Xamplified, Free Online Education Resource

In a cell, the metabolism comprises thousands of complex chemical reactions. These reactions are chained in metabolic pathways. A metabolic pathway consists of a series of biochemical reactions, starting from a substrate S to generate a product P. Each intermediate reaction (that is except the first and the last ones) uses at least one product from another reaction in the pathway as a substrate, and generates the substrate of another reaction in the pathway. For instance, the metabolic pathway most widespread in living systems is glycolysis, which breaks down glucose to produce energy and takes place in the cell cytoplasm. Another example of a metabolic pathway is the Krebs cycle (see Fig. 1.2), whose specificity lies in that one of its basic substrates is also the end product of the pathway. Cells produce and transform the organic molecules that supply both the material and the energy requested for life. Metabolism consists of two opposed processes, catabolism, and anabolism. Catabolism extracts energy from complex molecules (e.g., glucids, lipids) by breaking them into smaller molecules. Anabolism requires energy to synthesize complex molecules from simple molecules. Metabolic pathways are controlled by enzymes. Enzymes are proteins that catalyze chemical reactions; namely they accelerate reactions, even in small amounts, and without participating in these reactions. Since many proteins are enzymes, the control of metabolism and the regulation of gene expression are intimately linked. Sometimes, metabolic control aims at homeostasis, that is, maintaining constant levels of some variables or constant rates of some processes; sometimes adaptation demands change. Gene regulation allows the cell to express protein when needed, thus ensuring the versatility and adaptability of an organism. Regulation of gene expression involves a wide range of mechanisms and actors (proteins, microRNAs, chromatin) and complex dynamics (production, storing, degradation). All steps of gene expression can be modulated, encompassing transcriptional initiation, RNA processing, protein synthesis, and post-translational modification of proteins. Fig. 1.3 (page 10) illustrates the hierarchical organization of living systems.

1.2.2 Genetics, Epigenetics, and Copy Number Polymorphism

Genetics, DNA methylation in epigenetics, and Copy Number Polymorphism all deal with the DNA sequence.

The dependences within genetic data (e.g., SNPs) define the linkage disequilibrium (LD). Faithful models of LD are required for the visualization of LD at various scales—including the genome scale—or to perform downstream analyses such as association studies (see Subsection 1.2.6). LD occurs because DNA variants close on the chromosome are scarcely separated by the shuffling of chromosomes (recombination) that takes place during sex cell formation. Such variants are therefore transmitted together (as a haplotype) from parent to child. Such patterns are at the basis of the so-called haplotype block structure [12]: "blocks", where statistical dependences between loci are high, alternate with shorter regions characterized by low statistical dependences, the recombination hotspots.

Epigenetic features, such as DNA methylation and histone modifications, the two most studied, are known to be heritable across cell divisions. It has been shown that epigenetic mechanisms influence phenotype through the regulation of gene expression. Epigenetic features differ across different tissues and cell types. Most of the vertebrate genome is methylated. Unmethylated sites show a propensity to cluster together along the genome; unmethylated clusters are often present in the regulatory regions of many genes. DNA methylation is an important regulator of gene transcription and is tightly linked with cellular differentiation. Besides, in many diseases, abnormal hypermethylation of these clusters results in transcriptional silencing of the nearby genes. Specifically, associations between altered methylation states and various cancers have been

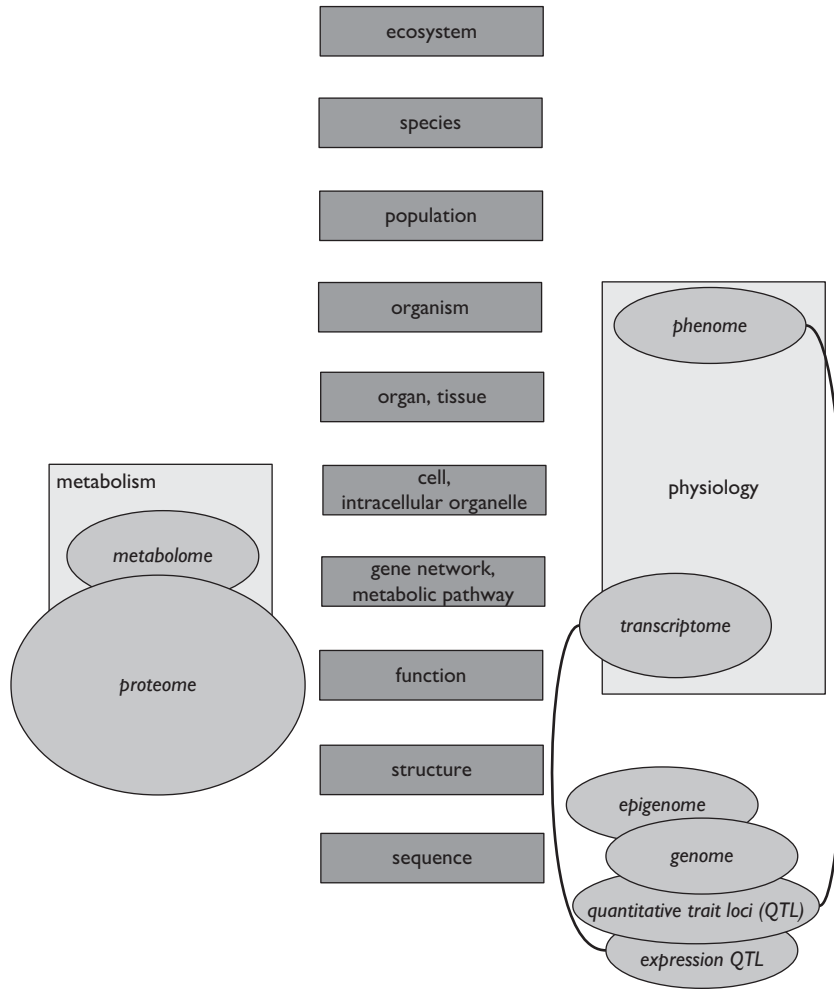

Fig 1.3 Hierarchical organization of living systems.

reported. Moreover, DNA methylation in tumor cells encodes phenotypic information about the tumor or the tumor subtypes. In analogy with the difference between genome sequencing and genotyping, where only a small subset of an individual's nucleotides are assayed, methyltyping suffers from low resolution in comparison with methylome sequencing. Thus, methyltyping poses a challenge in DNA methylation profiling. Modeling DNA methylation to exhibit subtypes in a population is another challenge.

Two chapters are dedicated to the genome-scale modeling of dependences within genetic data:

Chapter 9: Modeling Linkage Disequilibrium and Performing Association Studies through Probabilistic Graphical Models: A Visiting Tour of Recent Advances (C. Sinoquet and R. Mourad), and

Chapter 10: Modeling Linkage Disequilibrium with Decomposable Graphical Models (H. Abel and A. Thomas).

The two chapters above deal with Single Nucleotide Polymorphism. The chapter below addresses another kind of DNA polymorphism, DNA Copy Number Variations:

Chapter 16: Detection of Copy Number Variations from Array Comparative Genomic Hybridization Data using Linear-chain Conditional Random Field Models (X. Yin and J. Li).

In diploid genomes, for each gene, or more generally for each genomic segment, each individual inherits one copy from its father and one copy from its mother. Thus, in principle, the total number of copies is two. However, copy number mutations may occur: the total number of copies may be one (deletion), or three or more (amplifications/insertions).

Copy number alterations have been reported to be associated with numerous diseases. In particular, such chromosal aberrations as amplifications and deletions have led to the discovery of important oncogenes or tumor suppress genes. Array comparative genomic hybridization (aCGH) is a technology that allows the identification of copy number alterations across genomes.

In aCGH, which is an array-based technology, fluorescence is used to measure indirectly the number of copies for each DNA fragment in the array. Analyzing copy number polymorphisms using aCHG data consists of two tasks: detection of the boundaries where the copy number exhibits changes, and inference of the copy number state for any such designated regions. Basic data integration within the genomic level is performed in this case, since it is necessary to align the regions targeted by the array, and thus to refer to the genome sequence.

In another category, but again in the line of methods addressing the lowest level of biological organization—the DNA sequence—two other chapters address DNA methylation profiling. The chapter below analyzes DNA methylation profiles to cluster data:

Chapter 15: Latent Variable Models for Analyzing DNA Methylation (E. Andrés Houseman).

The other chapter will be mentioned in the next section. As it relies on prior genomic knowledge, it is an example of data integration.

1.2.3 Epigenetics with Additional Prior Knowledge on the Genome

In the chapter mentioned below, knowledge integration is performed within the same level of description (DNA):

Chapter 14: Bayesian Networks in the Study of Genome-wide DNA Methylation (M. Singer and L. Pachter).

Therein, more information is incorporated from the genomic data. Basically, the genomic structure is used as a prior on methylation status: in vertebrates, unmethylated sites tend to cluster together. Besides, the so-called CpG sites, which are unmethylated, are more conserved than other sites. Therefore, when experimental annotation of CpG sites is available, the richness of genomic regions in CpG clusters tends to point out unmethylated regions.

1.2.4 Transcriptomics

The sequence of an mRNA mirrors the sequence of the DNA from which it was transcribed. Consequently, by analyzing the entire collection of RNAs (transcripts) in a cell, transcriptomics

can determine which gene is turned on or off in the cells and tissues of an organism. Different cells show different patterns of gene expression. Transcriptomics examines gene expression microarrays, in which individuals are observed for a common set of genes.

One aim of transcriptomics is to determine how gene expression changes under the pressure of various factors such as tissue type, stage of development, drugs, or disease status. Differentially expressed genes are genes whose mean expression over a group of individuals sampled under a given condition (treatment, disease status (affected)) is significantly higher or lower than the mean expression over the control group (e.g. unaffected). For rigorous differential expression assessment, there is much more information in a microarray data set than the usual analysis extracts: the correlation structure between genes should be taken into account. The usual simplifying assumption of no correlation is unreasonable as genes are known to be connected in pathways or networks.

Since proteins can be transcription factors for other genes, genes' interplays may be summarized in gene regulatory networks. Genes targeted (in their regulatory regions) by the same transcription factors tend to show similar expression patterns along time. Thus, genes that are simultaneously co-expressed in some experimental or physiological condition (that is, genes that are highly correlated since they have similar expression profiles) are likely to be co-regulated by the same gene or genes.

However, gene network inference is far more complicated than identifying clusters of co-expressed genes. Genes expressed or inhibited in similar conditions or time points are likely to interact together. Yet, gene network reconstruction requires distinguishing between the correlation of two genes due to direct causal relationships and the correlation that originates from intermediate genes. Therefore, it is necessary to evaluate the correlation between genes conditioning on other genes. Through the exhibition of direct causal relationships, gene network inference highlights potential regulations or chains of regulations. For example, it is crucial to identify hubs, those key genes that regulate many other genes. On the other hand, modules—or communities—of genes are main contributors to the robustness and evolvability of biological networks; a module is defined as a set of interacting genes, whose function is separable from the function of other modules. The role of biologists remains to validate the gene network inferred or to clarify which are the exact paths corresponding to regulatory chains.

Observation across various conditions sheds light on the constants or variations of these dependences, that is, on the flexibility of the gene regulation network. In contrast to gene relationships unique to particular conditions or samples, some interactions may be shared across conditions or samples. Potentially complex distributions of gene expression across a wide range of conditions may be described through mixture models.

In some cases, merging different experimental conditions mainly aims at enlarging the number of observations available to infer a gene network. In this case, heterogeneity among microarray experiments represents an issue to cope with. A remedy is to study multiple networks simultaneously with an incentive to share interplays across conditions.

One chapter focuses on the acknowledgment of gene correlation in the assessment of differential expression:

Chapter 3: Graphical Models and Multivariate Analysis of Microarray Data (H. Kiiveri).

Two other chapters address gene network inference:

Chapter 4: Comparison of Mixture Bayesian and Mixture Regression Approaches to Infer Gene Networks (S.L. Rodriguez-Zas and B.R. Southey), and

Chapter 5: Network Inference in Breast Cancer with Gaussian Graphical Models and Extensions (M. Jeanmougin, C. Charbonnier, M. Guedj and J. Chiquet).

1.2.5 Transcriptomics with Prior Biological Knowledge

Chapter 5 is another example of the integrative approach described by this book. Therein, prior knowledge on the latent gene network structure is used. Many sources can be used as a biological prior on the network structure. For instance, prior knowledge may come from the gene level, as information about metabolic pathways or about which genes code for other genes' transcription factors. Metabolic pathways are available from the KEGG (Kyoto Encyclopedia of Genes and Genomes) [20] or BioCarta databases (http://www.biocarta.com/genes/index.asp); the connection between two genes is promoted or penalized depending on whether the genes belong to the same pathway or not. In addition, binding sites of transcription factors point out which genes are potentially regulated by the transcription factors.

All other chapters in the book infer knowledge through the integration of various sources of data.

1.2.6 Integrating Data from Several Levels

This transition provides the opportunity to define the concepts of **integrative biology** and **systems biology**. According to some scientists, integrative biology denotes multidisciplinary research (cross-disciplinary, transdisciplinary) incorporating chemistry, physics, mathematics, and computer science, as appropriate. At the interfaces, significant issues are discussed among scientists bringing together diverse but specific skills. As each chapter in this book takes a machine learning approach to deal with genetics, genomics, or postgenomics, this first definition of integrative biology holds for the book.

To other researchers, integrative biology means using a panel of various techniques and approaches to fulfill their own research programs. The previous definition includes the hierarchical approaches that deal with integration across levels of biological organization. At the extreme, such integrative frameworks describe life from molecules to the biosphere, with diversity across taxa, encompassing viruses, bacteria, plants, and animals. The availability of omics data (genomics, transcriptomics, proteomics, metabolomics, phenomics . . .) allows the implementation of integrative approaches across as many levels of biological organization [19]. In this book, all chapters not already mentioned in the present section fit this specific latter definition of integrative biology.

To some extent, the above definition meets the concept of **systems biology**. Systems biology is an approach in biology and biomedical research meant to understand living systems as wholes, be they an organism, a tissue, or a cell. In the more traditional so-called reductionist biology, a system's pieces are studied separately. In contrast, the purpose of systems biology is to put a system's pieces together, in a holistic perspective. Through this integration, systems biology aims at discovering the emergent properties of cells, tissues, and organisms functioning as systems. Evidencing such emergent properties is ideally addressed by observing multiple components simultaneously and by rigorously integrating data, based on mathematical models. These emergent properties mainly describe the complex interactions within biological systems, as illustrated by gene regulation networks, causal phenotype networks, and associations between genotype and phenotype. All three previous topics are at the core of fourteen of the chapters in this book. The hierarchies of biological levels that are spanned therein may not appear very deep as they connect the genomic

and gene levels, relying on genetics, transcriptomics, and phenomics. Nonetheless, the integrative approaches depicted require advanced models.

INTEGRATING GENETICS AND PHENOMICS

Four chapters in this book deal with quantitative genetics, which is the understanding of how genotype contributes to phenotype. In the biomedical research domain, an association study aims at identifying a causal relation between some genomic locus or loci and a disease status (affected/unaffected). Genome-wide association studies (GWASs) tackle the issue of unraveling such genotype-phenotype dependences from massive data. Such data usually describe thousands or ten thousands of subjects with a few hundred thousands to one or two millions of SNPs. The two following chapters address GWAS strategies:

Chapter 9: Modeling Linkage Disequilibrium and Performing Association Studies through Probabilistic Graphical Models: A Visiting Tour of Recent Advances (C. Sinoquet and R. Mourad),

Chapter 11: Scoring, Searching, and Evaluating Bayesian Network Models of Gene-phenotype Association (X. Jiang, S. Visweswaran and R.E. Neapolitan).

In addition, one chapter thoroughly reviews various refined concepts of association, whereas another chapter takes the slightly different viewpoint of *predicting* phenotypes from GWAS data:

Chapter 13: Bayesian, Systems-based, Multilevel Analysis of Associations for Complex Phenotypes: From Interpretation to Decision (P. Antal, A. Millinghoffer, G. Hullám, G. Hajós, P. Sárközy, A. Gézsi, C. Szalai and A. Falus), and

Chapter 17: Prediction of Clinical Outcomes from Genome-wide Data (S. Visweswaran).

INTEGRATING GENETICS, PHENOMICS, AND PRIOR KNOWLEDGE ON BIOLOGICAL PATHWAYS

In the quantitative genetics domain, one chapter of the book illustrates integration across three levels of biological organization:

Chapter 12: Graphical Modeling of Biological Pathways in Genome-wide Association Studies (M. Chen, J. Cho and H. Zhao).

In this chapter, a standard GWAS provides a list of genes associated with a studied disease. On the other hand, some other genes, not surveyed by the GWAS, are known to belong to the same biological pathways as the previous genes. The purpose is to estimate the probability that these other genes may be associated with the disease.

INTEGRATING GENETICS AND TRANSCRIPTOMICS

The phenotypes dealt with in the above cited chapters are discrete variables (affected/unaffected status). In Section 1.1, it was recalled that an organism's phenotype consists in the expression of its genotype, under defined environmental conditions. The expression of observable characteristics includes features that are only observable through the aid of technology. Transcriptomics provides gene expression levels for the genes targeted by a microarray. These expression levels represent as many continuous—or quantitative—phenotypes.

A quantitative phenotype (or trait) is defined as any physical, physiological, or biochemical quantitative feature that may be observed for organisms. The purpose of quantitative trait loci (QTL) mapping is to identify the genomic regions, named QTLs, where genotype variation entails phenotype variation. The definition of QTLs is straightforwardly transposed to expression QTLs (eQTLs) for which the continuous phenotype is a gene expression level.

Dissecting the causal relationships among *expression* traits involved in the same biological pathways—and therefore correlated—is a current research topic. Assumptions about the causal structure of observed variables are often represented in a *directed* acyclic graph. In causality inference, the identification of the eQTLs causal to each phenotype is of prime importance. The genetic architecture (GA) of a given phenotype denotes the locations and effects of its (directly) causal QTLs. The inference of a causal phenotype network (CPN) has to benefit from the knowledge about the genetic architecture: adding causal QTL nodes to a phenotype network allows the inference of causal relationships between phenotypes that could not be distinguishable using phenotype data alone. Conversely, GA inference may be refined based on the information borne by the CPN.

Three chapters in the book are dedicated to the inference of causal phenotype networks. Two of them rely on the mere integration of genetics and transcriptomics:

Chapter 6: Utilizing Genotypic Information as a Prior for Learning Gene Networks (K. Chipman and A. Singh), and

Chapter 8: Structural Equation Models for Studying Causal Phenotype Networks in Quantitative Genetics (G.J.M. Rosa and B.D. Valente).

INTEGRATING GENETICS, TRANSCRIPTOMICS, AND PRIOR BIOLOGICAL KNOWLEDGE

To reconstruct causal phenotype networks, the chapter below implements further data integration:

Chapter 7: Bayesian Causal Phenotype Network Incorporating Genetic Variation and Biological Knowledge (J. Young Moon, E. Chaibub Neto, X. Deng and B.S. Yandell).

Prior biological knowledge is incorporated, which may originate from various sources of biological information. One possible source of information is chromatin immunoprecipitation with microarray experiments (ChIP on chip), which is used to investigate the interaction of proteins and DNA *in vivo*. This technology is employed to generate putative lists of target genes for a given transcription factor; it evidences that a given transcription factor binds to some putative target. Regulation inference from knock-out data and protein–protein interaction can also be used as a prior. Knock-out gene technology allows to inactivate specific genes within an organism. Genes are knocked out by modifying the region of the gene that codes for the protein. Thus can be determined the effect of this gene on the functioning of the organism. Pathway information can also guide to refine the causal phenotype network. Finally, information from the Gene Ontology (GO) [6] may contribute to the biological prior. The GO is a specific vocabulary of terms describing the molecular functions, biological processes, and cellular components of a gene. The GO terms annotate a large fraction of genes. A similarity measure between genes may be defined in this GO

framework, that enables the connection of genes in a gene network. This network is subsequently used as a prior for causal phenotype network inference.

Finally, not only does the method described in Chapter 7 perform data integration, it also performs process integration; whereas most approaches conduct GA inference and CNP reconstruction separately, these two processes are intertwined.

1.2.7 Recapitulation

Table 1.1 offers a summarized description of the various data sources involved in the integrative approaches described in this book.

1.3 An Era of High-throughput Genomic Technologies

In the previous section, we emphasized the integrative dimension present in all chapters in this book but one. For readers not familiar with data originating from high-throughput technologies, we now briefly describe the data and the genesis of the data dealt with by the chapters of this book. This section may be skipped by other readers.

Various technologies can be used to generate genome-scale data that provide measurements at various levels of biological organization. These so-called "omics" data offer an unprecedented potential to gain insights on the workings of living systems.

1.3.1 Genotyping

In the broad sense, genotyping is the process of determining the genetic composition of an organism by inspecting its DNA sequence.

Genotyping can be achieved through a variety of methods, depending on the polymorphism of interest (e.g., SNPs, insertions, deletions, duplications, and rearrangements) and the resources available. Copy Number Variations (CNVs), which result from duplications and deletions, will be addressed in Subsection 1.3.2. In the present section, we concentrate on SNPs. SNP-based genotyping focuses on a small subset of nucleotide locations, known to exhibit variety within a population of subjects. The characteristic of SNP lies in that, in each such location, when *referring* to *one* of the two DNA strands, only two variants are observable among the four possible nucleotides. According to the international HapMap project [7], the estimated number of SNPs in the human genome amounts to 10 millions. SNP genotyping is associated with low cost but low resolution techniques. The use of genotyping chips or arrays is an efficient and accurate option for examining many loci simultaneously. In addition, next-generation technologies have reduced the costs of DNA sequencing down to the point that genotyping by sequencing is now feasible.

Subsection 1.3.4 is devoted to the presentation of the principle used in array techniques.

DNA sequencing aims to determine the exact sequence of a given region of DNA. Such regions may cover a short piece, the whole genome, or parts of the genome (e.g., the "exome", which is the 2% or so of the human genome that contains genes). If the targeted DNA stretches encompass SNPs, DNA sequencing may fulfill the purpose of genotyping. The remaining part of this section succinctly explains the technology behind DNA sequencing.

DNA polymerase is the enzyme involved in DNA replication, the biological process that enables the generation of DNA copies from a DNA template molecule. The molecule produced consists