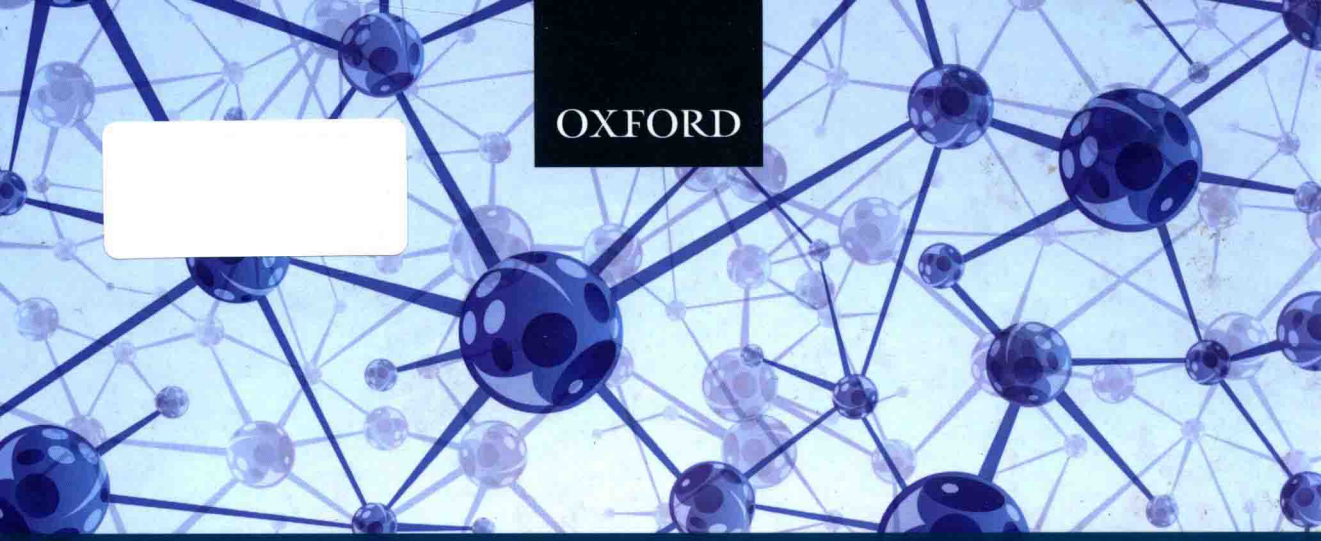
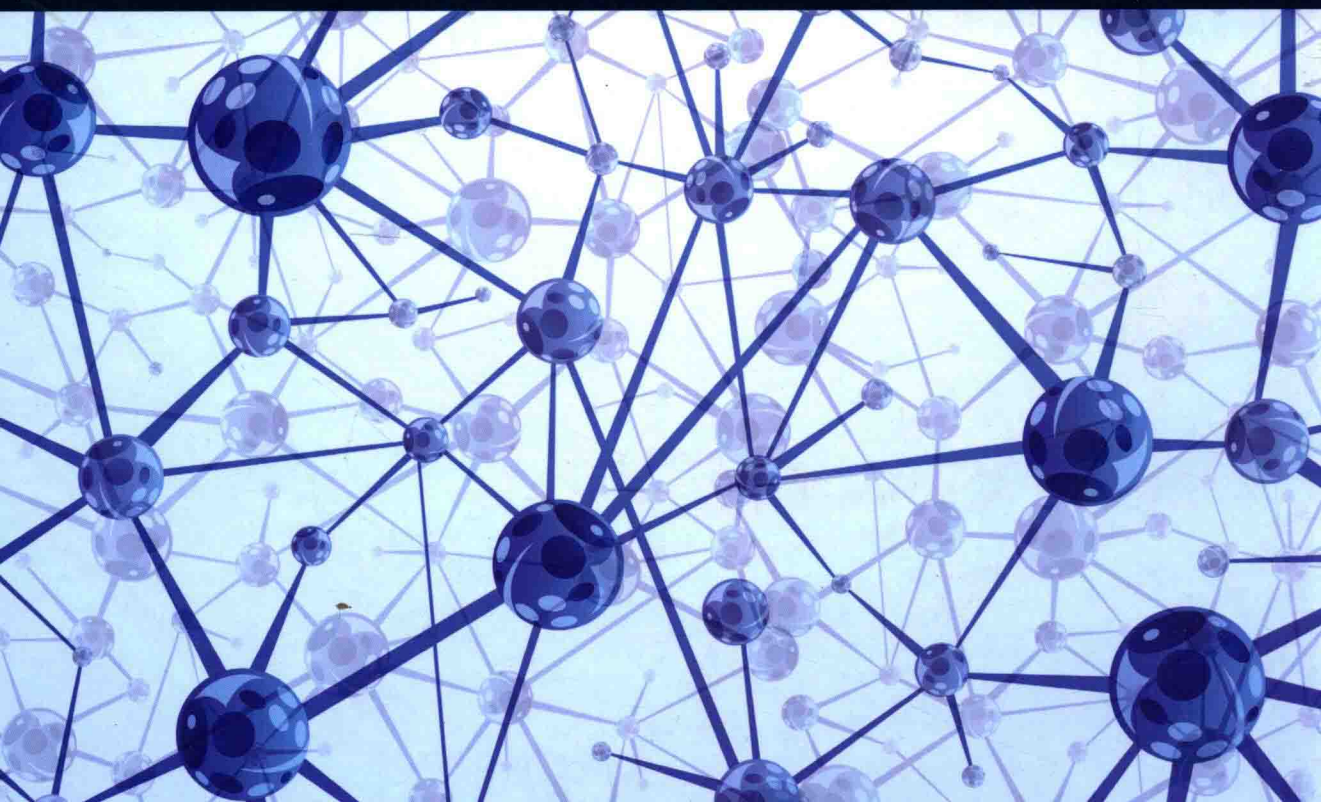


OXFORD

Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics

edited by CHRISTINE SINOQUET and RAPHAËL MOURAD



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

Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics

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.

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