

Computational Biology

Volume 18

Editors-in-Chief

Andreas Dress, CAS-MPG Partner Institute for Computational Biology, Shanghai, China
Michal Linial, Hebrew University of Jerusalem, Jerusalem, Israel
Olga Troyanskaya, Princeton University, Princeton, NJ, USA
Martin Vingron, Max Planck Institute for Molecular Genetics, Berlin, Germany

Editorial Board

Robert Giegerich, University of Bielefeld, Bielefeld, Germany
Janet Kelso, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
Gene Myers, Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany
Pavel A. Pevzner, University of California, San Diego, CA, USA

Advisory Board

Gordon Crippen, University of Michigan, Ann Arbor, MI, USA
Joe Felsenstein, University of Washington, Seattle, WA, USA
Dan Gusfield, University of California, Davis, CA, USA
Sorin Istrail, Brown University, Providence, RI, USA
Thomas Lengauer, Max Planck Institute for Computer Science, Saarbrücken, Germany
Marcella McClure, Montana State University, Bozeman, MO, USA
Martin Nowak, Harvard University, Cambridge, MA, USA
David Sankoff, University of Ottawa, Ottawa, ON, Canada
Ron Shamir, Tel Aviv University, Tel Aviv, Israel
Mike Steel, University of Canterbury, Christchurch, New Zealand
Gary Stormo, Washington University in St. Louis, St. Louis, MO, USA
Simon Tavaré, University of Cambridge, Cambridge, UK
Tandy Warnow, University of Texas, Austin, TX, USA
Lonnie Welch, Ohio University, Athens, OH, USA

The *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

More information about this series at <http://www.springer.com/series/5769>

Florian Frommlet · Małgorzata Bogdan
David Ramsey

Phenotypes and Genotypes

The Search for Influential Genes

 Springer

Florian Frommlet
Center for Medical Statistics, Informatics,
and Intelligent Systems
Section for Medical Statistics
Medical University of Vienna
Vienna
Austria

David Ramsey
Department of Operations Research
Wrocław University of Technology
Wrocław
Poland

Małgorzata Bogdan
Institute of Mathematics
University of Wrocław
Wrocław
Poland

ISSN 1568-2684

Computational Biology

ISBN 978-1-4471-5309-2

ISBN 978-1-4471-5310-8 (eBook)

DOI 10.1007/978-1-4471-5310-8

Library of Congress Control Number: 2015959940

© Springer-Verlag London 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer-Verlag London Ltd.

Preface

In the past 20 years we have witnessed revolutionary technological development in the fields of biology/genetics and computing. This has enabled the success of the Human Genome Project and the sequencing of a huge proportion of the human genome. However, this achievement has not reduced the number of questions related to the influence of genes on a multitude of traits and the general well-being of living organisms, although the availability of new tools has enabled us to identify complicated genetic mechanisms, such as DNA methylation or gene–gene regulation.

The systematic increase in the availability of good quality genetic data has aided efforts toward a more complete description of the genetic background of complex traits, i.e., those that are determined by many genes, often interacting with each other. Research in this area is rapidly expanding, since, apart from extending knowledge in the field of biology, it addresses many socially/economically important problems. Marker (gene) assisted selection is currently widely applied to identify promising individuals for breeding programs among domesticated animals, leading to increased efficiency in production or enhancing the quality of food products such as milk or meat. In the context of human genetics, the identification of influential genes allows us to evaluate an individual’s susceptibility to certain diseases, design tools for early diagnosis, and produce new efficient medicines or personalized therapies.

As a result of this technological breakthrough, bioinformatics has appeared as a new scientific discipline, where the most effective research is performed by collaboration between biologists, computer scientists, and statisticians. While search through large and rapidly expanding genetic databases enables the identification of new genetic effects, it also creates a multitude of computational and statistical problems. Concerning statistical issues, the large dimension of statistical data often results in an erroneous description of reality when oversimplified statistical tools are used for their analysis. A full understanding of the properties of statistical/bioinformatics methods in such a high-dimensional setting is needed to accelerate progress in this field and requires further intensive research.

Understanding the properties of various methods for analyzing high-dimensional data requires advanced mathematical tools, while the development of efficient computational methods requires advanced knowledge in computer science. Therefore, the main intended audience of this book is students/researchers with a background in mathematics or computer science, who would like to learn about problems in the field of statistical genetics and statistical issues related to the analysis of high-dimensional data. Thus, we expect that readers possess some mathematical or computer science skills. On the other hand, the genetic material is explained starting at a basic level. For those who are not totally familiar with the fundamentals of statistics, an extensive statistical appendix is presented for reference.

While bioinformatics and statistical genetics deal with a variety of complex questions in the field of genetics, in this book we concentrate on methods for locating influential genes. Thus, we mainly discuss methods of identifying the associations between the *genotypes* of genetic markers and interesting traits (*phenotypes*). Also, we do not discuss methods based on pedigree analysis or family relationships, often applied in studies on humans or domesticated animals. Instead, we cover in detail methods of gene mapping in experimental crosses, as well as genome wide association studies, which are based on a random sample of individuals from outbred populations (e.g., from general human populations). We summarize classical and modern methods for gene mapping and point toward related statistical and computational challenges. We believe that the knowledge contained in our monograph forms an excellent starting point for becoming involved in the exciting world of this field of research and hope that at least some of our readers decide to take this invitation and participate in the ongoing journey to develop a better understanding of the role of genetics in the biology of living organisms.

Vienna
Wrocław
August 2015

Florian Frommlet
Małgorzata Bogdan
David Ramsey

Contents

1	Introduction	1
	References	8
2	A Primer in Genetics	9
2.1	Basic Biology	9
2.1.1	Phenotypes and Genotypes	9
2.1.2	Meiosis and Crossover	12
2.1.3	Genetic Distance	12
2.1.4	The Haldane Mapping Function	13
2.1.5	Interference and Other Mapping Functions	14
2.1.6	Markers and Genetic Maps	16
2.2	Types of Study	17
2.2.1	Crossing Experiments	17
2.2.2	The Basics of QTL Mapping	21
2.2.3	Association Studies	23
2.2.4	Other Types of Study	27
	References	29
3	Statistical Methods in High Dimensions	31
3.1	Overview	31
3.2	Multiple Testing	32
3.2.1	Classical Procedures Controlling FWER	33
3.2.2	Permutation Tests and Resampling Procedures	37
3.2.3	Controlling the False Discovery Rate	41
3.2.4	Multiple Testing Under Sparsity. Minimizing the Bayesian Risk in Multiple Testing Procedures	42
3.3	Model Selection	51
3.3.1	The Likelihood Function	52
3.3.2	Information Theoretical Approach	55
3.3.3	Bayesian Model Selection and the Bayesian Information Criterion	57

3.3.4	Modifications of BIC for High-Dimensional Data Under Sparsity	59
3.3.5	Further Approaches to Model Selection	61
	References	69
4	Statistical Methods of QTL Mapping for Experimental Populations	73
4.1	Classical Approaches	73
4.1.1	Single Marker Tests	73
4.1.2	Power of a Test Based on a Single Marker as a Function of the Distance Between the Marker and a QTL	74
4.1.3	Genome Wide Search with Tests Based on Single Markers	76
4.2	Interval Mapping	79
4.2.1	Interval Mapping Based on the Mixture Model	79
4.2.2	Regression Interval Mapping	81
4.2.3	Nonparametric Version of Interval Mapping	82
4.2.4	Specific Models	83
4.2.5	Overestimation of Genetic Effects	83
4.3	Model Selection	84
4.3.1	QTL mapping with mBIC	84
4.3.2	Robust Version of mBIC	87
4.3.3	Version of mBIC Based on Rank Regression	89
4.3.4	Extensions to Generalized Linear Models	90
4.3.5	mBIC for Dense Markers and Interval Mapping	92
4.4	Logic Regression	97
4.5	Applying mBIC in a Bayesian Approach	101
4.6	Closing Remarks	101
	References	102
5	Statistical Analysis of GWAS	105
5.1	Overview	105
5.2	Inferring Genotypes	107
5.2.1	Genotype Calling	107
5.2.2	Imputation	110
5.3	Single Marker Tests	111
5.3.1	Case-Control Studies	111
5.3.2	Quantitative Traits	115
5.3.3	Covariates and Population Stratification	118
5.3.4	Multiple Testing Correction	121
5.3.5	Rare SNPs	122
5.4	Model Selection	124
5.4.1	Motivation	124
5.4.2	HYPERLASSO	129

5.4.3	GWASelect.	132
5.4.4	MOSGWA	133
5.4.5	Comparison of Methods	135
5.4.6	Mixed Models.	138
5.5	Admixture Mapping	140
5.6	Gene–Gene Interaction	144
5.6.1	Analyzing Gene–Gene Interaction via ANOVA	144
5.6.2	Multifactor Dimensionality Reduction	147
5.6.3	Logic Regression in GWAS	149
5.7	Other Recent Advances and the Outlook for GWAS.	151
	References	156
6	Appendix A: Basic Statistical Distributions	163
6.1	Normal Distribution	163
6.2	Important Distributions of Sample Statistics	165
6.2.1	Chi-Square Distribution	165
6.2.2	Student’s t-Distribution.	166
6.2.3	F-distribution	167
6.3	Gamma and Beta Distributions	168
6.3.1	Exponential Distribution.	168
6.3.2	Inverse Gamma Distribution	168
6.3.3	Beta Distribution	169
6.4	Double Exponential Distribution and Extensions	169
6.4.1	Asymmetric Double Exponential (ADE) Distribution.	169
6.5	Discrete Distributions	170
6.5.1	Binomial Distribution.	170
6.5.2	Poisson Distribution.	170
6.5.3	Negative Binomial Distribution	171
6.5.4	Generalized Poisson Distribution	171
6.5.5	Zero-Inflated Generalized Poisson Distribution	172
	Reference	172
7	Appendix B: Basic Methods of Estimation.	173
7.1	Basic Properties of Statistical Estimators.	173
7.1.1	Statistical Bias.	173
7.1.2	Mean Square Error	174
7.1.3	Efficiency of Estimators	174
7.1.4	Method of Moments	175
7.1.5	Maximum Likelihood Estimation.	176
7.2	Estimates of Basic Statistical Parameters.	177
7.2.1	Mean and Variance	177
7.2.2	Pearson Correlation Coefficient	177
	Reference	178

8	Appendix C: Principles of Statistical Testing	179
8.1	Basic Ideas of Statistical Testing: The Z-test	179
8.2	The Family of t-tests	182
8.2.1	One Sample t-Test	182
8.2.2	Two Sample t-Test	183
8.2.3	Paired t-Test	184
8.2.4	Robustness of t-Tests	184
8.3	Classical Approach to ANOVA and Regression	185
8.3.1	One-Way Analysis of Variance	185
8.3.2	Two-Way ANOVA. Interactions	186
8.3.3	Two-Way ANOVA with No Interactions	189
8.3.4	Extensions to a Larger Number of Factors	190
8.3.5	Multiple Regression	190
8.3.6	Weighted Least Squares	192
8.4	General Linear Models	192
8.4.1	Cockerham's Model	193
8.4.2	Robustness of General Linear Models	194
8.5	Generalized Linear Models	194
8.5.1	Extensions of Poisson Regression	197
8.6	Linear Mixed Models	197
8.7	Nonparametric Tests	200
8.7.1	Wilcoxon Signed-Rank Test	202
8.7.2	Rank Regression	202
8.8	Tests for Categorical Variables	203
8.8.1	Chi-Square Goodness-of-Fit Test	203
8.8.2	Chi-Square Test of Independence	204
8.8.3	Fisher's Exact Test	205
	References	206
9	Appendix D: Elements of Bayesian Statistics	207
9.1	Bayes Rule	207
9.2	Conjugate Priors	208
9.3	Markov Chain Monte Carlo	208
9.3.1	Gibbs Sampler	209
9.3.2	Metropolis–Hastings Algorithm	210
9.3.3	Hierarchical Models	210
9.3.4	Parametric Empirical Bayes	211
9.4	Bayes Classifier	212
	References	212
10	Appendix E: Other Statistical Methods	215
10.1	Principal Component Analysis	215
10.2	The EM Algorithm	216
	References	217
	Index	219

Acronyms

ABOS	Asymptotic Bayes optimality under sparsity
AIC	Akaike's information criterion
BH	Benjamini Hochberg procedure
BIC	Bayesian information criterion
BLUP	Best linear unbiased predictor
CAT	Cochran Armitage trend test
CDCV	Common disease—common variant
CDRV	Common disease—rare variant
CEU	HapMap Population: Utah, Europe ancestry
CHB	HapMap Population: Han Chinese in Beijing
CNV	Copy number variations
DNA	Deoxyribonucleic acid
EBIC	Extended Bayesian information criterion
FDR	False discovery rate
FWER	Family wise error rate
gFWER	Generalized family wise error rate
GLM	General linear model
gLM	Generalized linear model
GWAS	Genome wide association study
HWE	Hardy-Weinberg equilibrium
IBD	Identical by descent
JPT	HapMap population: Japanese in Tokyo
KL	Kullback-Leibler
LASSO	Least absolute shrinkage and selection operator
LD	Linkage disequilibrium
LMM	Linear mixed model
LRT	Likelihood ratio test
mBIC	A modification of the Bayesian information criterion
ML	Maximum likelihood
MOSGWA	Model selection for genome wide association (software)
mRNA	Messenger RNA

MTP	Multiple testing procedure
NEG	Normal exponential gamma distribution
PCA	Principal component analysis
PCER	Per-comparison error rate
PFER	Per-family error rate
QTL	Quantitative trait locus
REML	Restricted maximum likelihood
RNA	Ribonucleic acid
SD	FDR controlling step-down procedure
SIS	Sure independence screening
SNP	Single nucleotide polymorphisms
YRI	HapMap population: Yoruba in Ibadan