Computational Biology

Volume 18

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

Małgorzata Bogdan
Institute of Mathematics
University of Wrocław
Wrocław
Poland

David Ramsey
Department of Operations Research
Wrocław University of Technology
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

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 — Florian Frommlet
Wrocław — Małgorzata Bogdan
August 2015 — David Ramsey

Contents

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