METHODS IN MOLECULAR BIOLOGY™

Series Editor
John M. Walker
School of Life Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Evolutionary Genomics

Statistical and Computational Methods, Volume 1

Edited by

Maria Anisimova

Department of Computer Science, Swiss Federal Institute of Technology (ETHZ), Zürich, Switzerland
Swiss Institute of Bioinformatics, Lausanne, Switzerland

Humana Press

Editor
Maria Anisimova, Ph.D.
Department of Computer Science
Swiss Federal Institute of Technology (ETHZ)
Zürich, Switzerland

Swiss Institute of Bioinformatics
Lausanne, Switzerland

The photo used for book cover is made by one of the authors of the book, Wojciech Makalowski.

ISSN 1064-3745 e-ISSN 1940-6029
ISBN 978-1-61779-581-7 e-ISBN 978-1-61779-582-4
DOI 10.1007/978-1-61779-582-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2012931926

Printed on acid-free paper

Humana Press is part of Springer Science+Business Media (www.springer.com)

Preface

Discovery of genetic material propelled the power of classical evolutionary studies across the diversity of living organisms. Together with early theoretical work in population genetics, the debate on sources of genetic makeup initiated by proponents of the neutral theory made a solid contribution to the spectacular growth in statistical methodologies for molecular evolution. The methodology developed focused primarily on inferences from single genes or noncoding DNA segments: mainly on reconstructing the evolutionary relationships between lineages and estimating evolutionary and selective forces. Books offering a comprehensive coverage of such methodologies have already appeared, with Joe Felsenstein's "Inferring Phylogenies" and Ziheng Yang's "Computational Molecular Evolution" among the favorites.

This volume is intended to review more recent developments in the statistical methodology and the challenges that followed as a result of rapidly improving sequencing technologies. While the first sequenced genome (RNA virus Bacteriophage MS2 in 1976) was not even 4,000 nucleotides long, the sequencing progress culminated with the completion of the human genome of about 3.3×10^9 base pairs and advanced to sequence many other species genomes, heading ambitiously towards population sequencing projects such as 1,000 genome projects for humans and *Drosophila melanogaster*. Next-generation sequencing (NGS) technologies sparked the "genomics revolution," which triggered a renewed effort towards the development of statistical and computational methods capable of coping with the flood of genomic data and its inherent complexity.

The challenge of analyzing and understanding the dynamics of large-system data can be met only through an integration of organismal, molecular, and mathematical disciplines. This requires commitment to an interdisciplinary approach to science, where both experimental and theoretical scientists from a variety of fields understand each other's needs and join forces. Evidently, there remains a gap to be breached. This book presents works by top scientists from a variety of disciplines, each of whom embodies the interdisciplinary spirit of evolutionary genomics. The collection includes a wide spectrum of articles—encompassing theoretical works and hands-on tutorials, as well as many reviews with much biological insight.

The evolutionary approach is clearly gaining ground in genomic studies, for it enables inferences about patterns and mechanisms of genetic change. Thus, the theme of *evolution* streams through each chapter of the book, providing statistical models with basic assumptions and illustrated with appealing biological examples. This book is intended for a wide scientific audience interested in a compressed overview of the cutting-edge statistical methodology in evolutionary genomics. Equally, this book may serve as a comprehensive guide for graduate or advanced undergraduate students specializing in the fields of genomics or bioinformatics. The presentation of the material in this volume is aimed to equally suit both a novice in biology with strong statistics and computational skills and a molecular biologist with a good grasp of standard mathematical concepts. To cater for differences in reader backgrounds, *Part I* of *Volume 1* is composed of educational primers to help with fundamental concepts in genome biology (Chapters 1 and 2), probability and statistics (Chapter 3), and molecular evolution (Chapter 4). As these concepts reappear repeatedly throughout the books, the first four chapters will help the neophyte to stay "afloat."

The exercises and questions offered at the end of each chapter serve to deepen the understanding of the material. Additional materials and some solutions to exercises can be found online: http://www.evolutionarygenomics.net.

Part II of this volume reviews state-of-the-art techniques for genome assembly (Chapter 5), gene finding (Chapter 6), sequence alignment (Chapters 7 and 8), and inference of orthology, paralogy (Chapter 9), and laterally transferred genes (Chapter 10). *Part III* opens with a comparative review of genome evolution in different breeding systems (Chapter 11) and then discusses genome evolution in model organisms based on the studies of transposable elements (Chapters 12 and 13), gene families, synteny (Chapter 14), and gene order (Chapters 15 and 16).

Part I of *Volume 2* is the evidence that, since embracing Darwin's tree-like representation of evolution and pondering over the universal Tree of Life, the field has moved on. Nowadays, the evolutionary biologists are well aware of numerous evolutionary processes that distort the tree, complicating the statistical description of models and increasing computational complexity, often to prohibitive levels. Each taking a different angle, the chapters of *Part I*, *Volume 2* discuss how to overcome problems with phylogenetic discordance, as the Tree of Life turns out to be more like a "forest" (Chapter 3). The multispecies coalescent model offers one solution to reconciling phylogenetic discord between gene and species trees (Chapter 1); others pursue probabilistic reconciliation for gene families based on a birth–death model along a species phylogeny (Chapter 2). By some perspectives, constraining the understanding of evolution solely with tree-like structures omits many important biological processes that are not tree-like (Chapter 4).

Most fundamental questions in genome biology strive to disentangle the evolutionary forces shaping species genomes, inferring evolutionary history, and understanding how molecular changes affect genomic and phenotypic characteristics. To this goal, *Part II* of the *Volume 2* introduces methods for detecting and reconciling selection (Chapters 5 and 6) and recombination (Chapters 9 and 10), and discusses the mechanisms for the origins of new genes (Chapter 7) and the evolution of protein domain architectures (Chapter 8). The role of *natural selection* in shaping genomes is a pinnacle of the classical neutralist–selectionist debate and sets an important theme of the book; the "neo-selectionist" model of genome evolution is tested on many counts. This theme is also apparent in *Part III* dedicated to *population genomics*, which starts by discussing models for genetic architectures of complex disease and the power of genome-wide association studies (GWAS) for finding susceptibility variants (Chapter 11). With the availability of multiple genomes from closely related species, gleaning the ancestral population history also became possible, as is illustrated in the following chapter (Chapter 12). Most population genetics problems rely on ancestral recombination graphs (ARG), and reducing the redundancy of the ARG structure helps to reduce the computational complexity (Chapter 13).

Entering the era of postgenomics biology, recent years have seen rapid growth of complementary genomic data, such as data on expression and regulation, chemical and metabolic pathways, gene interactions and networks, disease associations, and more. Considering the genome as a uniform collection of coding and noncoding molecular sequences is no longer an option. To address this, great efforts are currently dedicated to embrace the complexity of biological systems through the emerging "-omics" disciplines—the focus of *Part IV* of this volume. Chapter 14 discusses ways to study the evolution of gene expression and regulation based on data from "old-fashioned" microarrays as well as transcriptomics data obtained with NGS such as RNAseq and ChIPseq. Interactomics is the focus of the next chapter. Indeed, better understanding of genes, their diversity and

regulation comes from studies of interaction between their protein products and networks of interacting elements (Chapter 15). Further topics include metabolomics (Chapter 16), metagenomics (Chapter 17), epigenomics (Chapter 18), and the newly reinvented discipline with a mysterious name—genetical genetics (Chapter 19). Despite the effort, complex dependencies and causative effects are difficult to infer. A way forward must be in the integration of complimentary "-omics" information with genomic sequence data to understand the fundamentals of systems biology in living organisms. This cannot be achieved without studying how such information changes over time and across various conditions. Vast amount of multifaceted data promise a big future for machine learning, pattern recognition and discovery, and efficient data mining techniques, as can be seen from many chapters of this book.

Finally, *Part V* of the second volume focuses on challenges and approaches for large and complex data representation and storage (Chapter 20). The rapid pace of computational genomics, as well as research transparency and efficiency, exacerbates the need for sharing of data and programming resources. Fortunately, some solutions already exist (Chapter 21). Handling ever increasing amounts of computation requires efficient computing strategies, which are discussed in the closing chapter of the book (Chapter 22).

For a novice in the field, this book is certainly a treasure chest of state-of-the-art methods to study genomic and omics data. I hope that this collection will motivate both young and experienced readers to join the interdisciplinary field of evolutionary genomics. But even the experienced bioinformatician reader is certain to find a few surprises. On behalf of all authors, I hope that this book will become a source of inspiration and new ideas for our readers. Wishing you a pleasant reading!

Zürich, Switzerland ***Maria Anisimova, Ph.D.***

Acknowledgments

The foremost gratitude goes to the authors of this book who came together to make this resource possible and who were enthusiastic and encouraging about the whole project. Over 100 reviewers have contributed to improving the quality and the clarity of the presentation with their constructive and detailed comments. Some reviewers have accepted to be acknowledged by their name. With great pleasure, I list them here:

Tyler Alioto, Peter Andolfatto, Miguel Andrade, Irena Artamonova, Richard M. Badge, David Balding, Mark Beaumont, Chris Beecher, Robert Beiko, Adam Boyko, Katarzyna Bryc, Kevin Bullaughey, Margarida Cardoso-Moreira, Julian Catchen, Annie Chateau, Karen Cranston, Karen Crow, Tal Dagan, Dirk-Jan de Koning, Christophe Dessimoz, Mario dos Reis, Katherine Dunn, Julien Y. Dutheil, Toni Gabaldon, Nicolas Galtier, Mikhail Gelfand, Josefa Gonzalez, Maja Greminger, Stephane Guindon, Michael Hackenberg, Carolin Kosiol, Mary Kuhner, Anne Kupczok, Nicolas Lartillot, Adam Leache, Gerton Lunter, Thomas Mailund, William H. Majoros, James McInerney, Gabriel Musso, Pjotr Prins, David A. Ray, Igor Rogozin, Mikkel H. Schierup, Adrian Schneider, Daniel Schoen, Cathal Seoighe, Erik Sonnhammer, Andrea Splendiani, Tanja Stadler, Manuel Stark, Krister Swenson, Adam M. Szalkowski, Gergely J. Szöllősi, Jijun Tang, Todd Treangen, Oswaldo R. Trelles Salazar, Albert Vilella, Rutger Vos, Tom Williams, Carsten Wiuf, Yuri Wolf, Xuhua Xia, S. Stanley Young, Olga Zhaxybayeva, and Stefan Zoller.

My colleagues from the Computational Biochemistry Research Group at ETH Zurich deserve much credit for being a constant source of inspiration and for providing such an enjoyable working environment. Finally, but no less importantly, I would like to thank my family for their love and for tolerating the overtime that this project required.

Contents

Contributors

Tyler Alioto • *Centro Nacional de Análisis Genómico, Barcelona, Spain*

Adrian M. Altenhoff • *Department of Computer Science, ETH Zurich, Zurich, Switzerland; Swiss Institute of Bioinformatics, Switzerland*

Stéphane Aris-Brosou • *Departments of Biology and Mathematics & Statistics and Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, ON, Canada*

Rajeev K. Azad • *Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA, USA; Departments of Biological Sciences and Mathematics, University of North Texas, Denton, TX, USA*

Niko Beerenwinkel • *Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland*

Aidan Budd • *European Molecular Biology Laboratory (EMBL), Heidelberg, Germany*

Christophe Dessimoz • *Department of Computer Science, ETH Zurich, Zurich, Switzerland; Swiss Institute of Bioinformatics, Switzerland*

Colin N. Dewey • *Biostatistics and Medical Informatics and Computer Sciences, Genome Center of Wisconsin, University of Wisconsin-Madison, Madison, WI, USA*

Nadia El-Mabrouk • *Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Montréal, QC, Canada*

Nicolas Galtier • *Institut des Sciences de l'Evolution, UMR5554, Université Montpellier II, Montpellier, France*

Sylvain Glémin • *Institut des Sciences de l'Evolution, UMR5554, Université Montpellier II, Montpellier, France*

Josefa González • *Department of Biology, Stanford University, Stanford, CA, USA; Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain*

Valer Gotea • *National Human Genome Research Institute, National Institutes of Health, Rockville, MD, USA*

Niina Haiminen • *IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA*

Shigehiro Kuraku • *Genome Resource and Analysis Unit, RIKEN Center for Developmental Biology, Chuo-ku, Kobe, Japan*

Jeffrey G. Lawrence • *Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA, USA*

Heewook Lee • *School of Informatics and Computing, Indiana University, Bloomington, IN, USA*

Ari Löytynoja • *European Bioinformatics Institute (EMBL), Hinxton, UK; Institute of Biotechnology, University of Helsinki, Helsinki, Finland*

Izabela Makałowska • *Laboratory of Bioinformatics, Adam Mickiewicz University, Poznań, Poland*

Wojciech Makałowski • *Institute of Bioinformatics, University of Muenster, Muenster, Germany*
Axel Meyer • *Department of Biology, University of Konstanz, Constance, Germany*
Amit Pande • *Institute of Bioinformatics, University of Muenster, Muenster, Germany*
Laxmi Parida • *IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA*
Dmitri A. Petrov • *Department of Biology, Stanford University, Stanford, CA, USA*
Nicolas Rodrigue • *Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, Ottawa, ON, Canada; Quebec Center for Biodiversity Science, McGill University, Montreal, QC, Canada; Agriculture and Agri-Food Canada, Eastern Cereal and Oilseeds Research Center, Central Experimental Farm, Ottawa, ON, Canada*
David Sankoff • *Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada*
Juliane Siebourg • *Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland*
Haixu Tang • *School of Informatics and Computing, Indiana University, Bloomington, IN, USA*