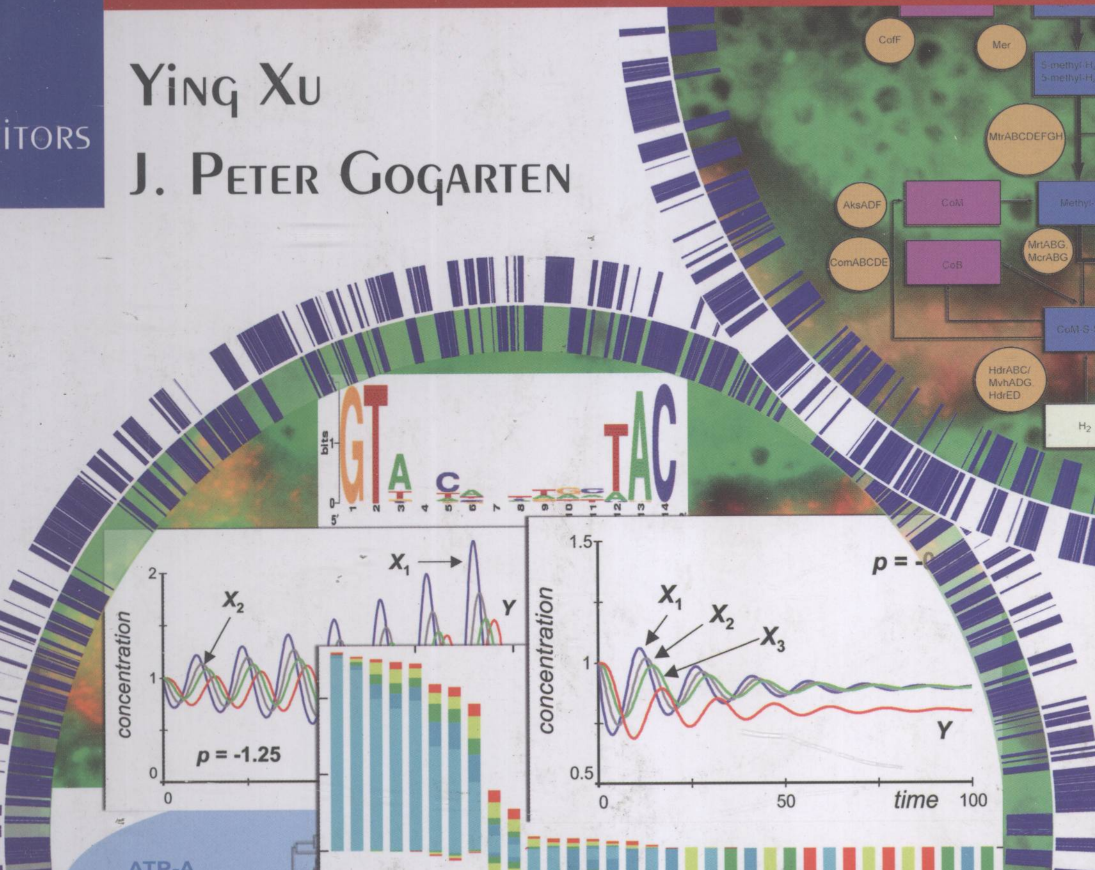
Series on Advances in Bioinformatics and Computational Biology – Volume 7

COMPUTATIONAL METHODS FOR UNDERSTANDING BACTERIAL AND ARCHAEAL GENOMES

EDITORS

Ying Xu

J. Peter Gogarten



Imperial College Press

Series on Advances in Bioinformatics and Computational Biology – Volume 7

COMPUTATIONAL METHODS FOR UNDERSTANDING BACTERIAL AND ARCHAEAL GENOMES

editors

Ying Xu
University of Georgia, USA

J. Peter Gogarten
University of Connecticut, USA

ICP

Imperial College Press

Published by

Imperial College Press
57 Shelton Street
Covent Garden
London WC2H 9HE

Distributed by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

COMPUTATIONAL METHODS FOR UNDERSTANDING BACTERIAL AND ARCHAEAL GENOMES
Series on Advances in Bioinformatics and Computational Biology — Vol. 7

ISBN-13 978-1-86094-982-1
ISBN-10 1-86094-982-7

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed by FuIsland Offset Printing (S) Pte Ltd, Singapore

COMPUTATIONAL METHODS FOR UNDERSTANDING BACTERIAL AND ARCHAEAL GENOMES

SERIES ON ADVANCES IN BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

ISSN: 1751-6404

Vol. 1: Proceedings of the 3rd Asia-Pacific Bioinformatics Conference
Eds: Yi-Ping Phoebe Chen and Limsoon Wong

Vol. 2: Information Processing and Living Systems
Eds: Vladimir B. Bajic and Tan Tin Wee

Vol. 3: Proceedings of the 4th Asia-Pacific Bioinformatics Conference
Eds: Tao Jiang, Ueng-Cheng Yang, Yi-Ping Phoebe Chen and Limsoon Wong

Vol. 4: Computational Systems Bioinformatics 2006
Eds: Peter Markstein and Ying Xu
ISSN: 1762-7791

Vol. 5: Proceedings of the 5th Asia-Pacific Bioinformatics Conference
Eds: David Sankoff, Lusheng Wang and Francis Chin

Vol. 6: Proceedings of the 6th Asia-Pacific Bioinformatics Conference
Eds: Alvis Brazma, Satoru Miyano and Tatsuya Akutsu

Vol. 7: Computational Methods for Understanding Bacterial and Archaeal Genomes
Eds: Ying Xu and J. Peter Gogarten

PREFACE

Sequencing technology has advanced to such a level that large sequencing centers such as the Joint Genome Institute (JGI) of the US Department of Energy can sequence a prokaryotic genome within a day. As of this writing, ~800 prokaryotic genomes have been sequenced and at least 1,000 are in the pipeline being sequenced. Knowing a few large sequencing efforts currently under planning, we could possibly see over 10,000 complete prokaryotic genomes within the next few years. In addition, the sequencing efforts of metagenomes have produced over one billion base-pairs of sequence fragments since 2000, and the efforts are expected to scale up rapidly, soon to produce substantially more genomic sequence data than what we have seen in the past twenty years. While these genomic sequence data have provided unprecedented opportunities for biologists to study and to understand these organisms, it has also raised some highly challenging problems regarding how to "mine" the genomes, extract the information encoded in the genomes, in a much more effective manner than what the existing tools can offer, simply to keep up with the pace of world-wide genome sequencing efforts.

Compared to eukaryotic genomes which are more complex in general, prokaryotic genomes pose a set of unique challenging problems. First, prokaryotic genomes are much more dynamic in terms of their gene compositions than the typical eukaryotic genomes since horizontal gene transfers take place substantially more often in prokaryotes. Second, prokaryotes evolve at much faster rates than eukaryotes in general, hence making their genomes diverge faster. Third, prokaryotes have been found in broader environments than eukaryotes, suggesting their more flexible adaptability to the environments. Fourth, prokaryotes often co-exist with other prokaryotes as a mutually dependent community, parts of whose metabolisms are inter-species, making studies of their biochemistry and their genomes rather unique.

Comparative genomics has been the most effective approach to mining the genomes and deciphering the information encoded in genomes. Numerous computational techniques have been developed based on comparative strategies, to predict genes and the functions of their protein products, to elucidate operonic structures, to detect previously unknown structures such as uber-operons, and to predict regulatory elements and interactions of biochemical pathways. Fundamental to these computational techniques is our understanding about the evolution of genes, molecular interactions, biological processes and genomes. It is the theoretical framework for studying these evolutionary processes that have guided our computational studies of the prokaryotic genomes and their structures. For this reason, we have designed this book in such a way that it tightly integrates computational studies with theoretical discussions of prokaryotic genomes. We believe that it is essential to have a good understanding about both the evolutionary theories for prokaryotes and the possibilities and limitations of computational genome analysis techniques, in order to carry out in-depth computational studies that may generate new insights about these genomes and the information they encode.

In this book, we included a collection of cohesively written chapters on prokaryotic genomes, their organization and evolution, the information they encode, computational approaches to deriving such information, and to understanding their organization and evolution. When appropriate, we attempt to provide a comparative view of the bacterial and archaeal genomes and of how information is differently encoded in these two domains of life. This book is intended to be used as an introductory text book for a graduate-level microbial genomics and bioinformatics course as well as a reference book for researchers working in the area of prokaryotic genome studies

While the chapters are organized in a logical order, each chapter in the book is a self-contained review of a specific subject. Hence a reader does not necessarily have to read through the chapters in their sequential order. Since this is a rapidly evolving field that encompasses an exceptionally wide range of research topics, it is difficult for any individual to write a comprehensive textbook on the entire field. Most of the chapters are written by members of our two labs at the University of Georgia and the University of Connecticut, respectively. The remaining chapters, which, we feel, will help to fill gaps in covering the field, are written by experts who are actively doing research at the forefront of the selected topical areas.

Chapter 1 (*General characteristics of prokaryotic genomes*) explains the notion of a genome as a collection of replicons in a cell, and provides an overview of different types of replicons and their distinguishing characteristics. This is followed by a review of the diversity of prokaryotic chromosomes with respect to their size, gene content, G+C content, codon usage, oligonucleotide composition, amino acid usage, repeat content, and intragenomic compositional heterogeneity emphasizing contrasts between eukaryotes and prokaryotes.

Chapter 2 (*Genes in prokaryotic genomes and their computational prediction*) begins with a historical account of the development of gene prediction methods. These methods exploit either intrinsic information, that is, nucleotide ordering patterns intrinsic to a DNA sequence, or extrinsic information gained from sequence conservation in evolution. The basic idea and models underlying the prediction programs utilizing either intrinsic information or extrinsic information or both are described, followed by a critical comparative assessment of their performance on experimentally validated datasets. The strengths and weaknesses of the current programs and the future challenges are discussed, with suggestions for addressing the remaining core issues in this field.

Chapter 3 (*Evolution of the genetic code: computational methods and inferences*) focuses on the evolution of genetic codes. Ever since its discovery, the genetic code has been one of the most difficult puzzles for evolutionary biology to unravel. As the evolution of early life on Earth is inextricably linked to the evolution of the genetic code, understanding its history has been the focus of numerous investigations over the last several years. We present several of the computational methods that have been applied to this problem, including a brief discussion of their results and significance.

Chapter 4 (*Dynamics of prokaryotic genome evolution*) addresses the dynamics of genome evolution. The ever-increasing amount of genome sequences available in public databases allows a better comprehension of how evolutionary forces act on prokaryotic species. Several methods exist to categorize genes based on their frequency of occurrence among genomes, and are used to predict the genes' roles and modes of evolution. This chapter describes practical approaches used in the comparison of multiple genomes, and discusses the current status of the field of prokaryotic genome evolution.

In Chapter 5 (*Mobile genetic elements and their prediction*), we review mobile genetic elements, with a focus on those that can be computationally predicted within genome sequences. We discuss the features of these elements ranging from the small neutral Insertion Sequence (IS) elements, to the large genomic islands and prophage regions that can encode antibiotic resistance and virulence-related functions. The chapter focuses on the *in silico* prediction of these elements, including the discussion of several existing tools and databases. In addition, we point out the strengths and weaknesses of the existing methods and suggest future avenues of research in this area.

Chapter 6 (*Horizontal gene transfer: its detection and role in microbial evolution*) provides a brief history of the discovery of gene transfer and how it impacted attempts to develop a natural taxonomy for prokaryotes. We review species concepts as applied to bacteria and archaea, look at examples of multilevel selection acting on genes, organisms and communities, and examine biological processes and artifacts that can create conflicts between gene and genome phylogenies. We describe phylogenetic and surrogate approaches that aim to identify transferred genes and discuss the advantages and problems associated with different methods, and provide an outlook on future developments that will allow to trace the history of genes and pathways through the network of organismal history.

In Chapter 7 (*Genome reduction during prokaryotic evolution*), we review the phenomenon of genome reduction, which has taken place independently many times in several prokaryotic lineages. This chapter describes methods to reconstruct ancestral genome contents and to infer genome expansion and contraction. The mutational and selective hypotheses to explain these changes will be discuss. Finally, we describe the gene content of the smallest genomes and the fuzzy boundary between cells and organelles.

Chapter 8 (*Comparative mechanisms on transcription and regulatory signals in archaea and bacteria*) describes our current knowledge about regulation of transcription in archaea and bacteria. This overview emphasizes the main *cis*-regulatory signals, *trans*-regulatory factors (TFs) and alternative RNA polymerase types constituting the machinery needed for proper regulation of genes in response to changing environmental conditions at the level of transcription initiation. In archaea the basal machinery for transcription is related to RNAP II from eukaryotes while the use of TFs for modulating gene transcription is similar to those used by bacteria. We describe the promoter at different levels, including the activity of the basal machinery for transcription and the use of specific transcription factors sensing and responding to particular effectors signals. We also summarize regulation at other levels beyond transcription initiation. There is an inherent limitation in this comparative approach given the limited amount of knowledge of the regulation mechanisms in archaea. Furthermore, these comparisons also bring eukaryotic transcription to the scene when searching for an evolutionary understanding of the diverse puzzle of similarities and differences in gene organization, conservation of proteins and their mechanisms involved in bacteria, archaea and eukaryotes.

Chapter 9 (*Computational techniques for orthologous gene prediction in prokaryotes*) discusses the definitions of homologs, orthologs, paralogs and their subtypes, and reviews the existing gene family databases and the approaches to homology assignment as well as family-superfamily classification systems. The chapter focuses on the automated methods of orthologous gene prediction. It describes the reciprocal best blast hit method, reciprocal smallest distance method, tree reconciliation algorithm, and the phylogenetic clustering algorithm BranchClust.

In Chapter 10 (*Computational elucidation of operons and uber-operons*), we describe the basics about operons as the basic units of transcription regulation as well as their ties to biological pathways and networks. In addition, we discuss a relatively new and less well-studied layer of genomic structure, called uber-operons. The chapter presents a number of basic ideas as well as computational methods for operon and uber-operon prediction, plus relevant prediction servers publicly available on the Internet. We also showcase one study on the evolution of operons, suggesting possible rules for operon evolution.

Chapter 11 (*Prediction of regulons through comparative genome analyses*) introduces first the typical structure of a regulon in prokaryotes, and the special regulon prediction problem as well as the genome-wide *de novo* regulon prediction problem. It then links these problems to the prediction problem of *cis*-regulatory binding sites. The chapter details a widely used phylogenetic foot-printing motif finding algorithm as well as a genome-wide scanning procedure for solving the special regulon prediction problem. For the more challenging genome-wide *de novo* prediction of regulons, the chapter introduces a recently developed phylogenetic foot-printing based algorithm. Examples of practical applications are also given for each described computational procedure.

Chapter 12 (*Prediction of biological pathways through data mining and information fusion*) presents a few approaches to biological pathway construction using multiple sources of information, including transcriptomic data, genomic data, proteomic data as well as protein-protein and protein-DNA interaction data. Several computational algorithms for these predictions are reviewed in this chapter. A comprehensive description on how to estimate parameters in metabolic pathways is also included in this chapter.

Chapter 13 (*Microbial pathway models*) emphasizes the importance of mathematical modeling approaches to understanding microbial systems. It reviews traditional approaches of enzyme kinetics and leads the reader to the state of the art in metabolic modeling. Particular focus is placed on Biochemical Systems Theory, which in several hundred publications has proven an invaluable tool for designing, analyzing, manipulating, and optimizing biological systems, even in situations where information on the target network is uncertain or partially missing.

Chapter 14 (*Metagenomics*) reviews a new scientific endeavor, metagenomics, which is emerging with the convergence of sequencing and computational technologies focused on communities of microbes. The techniques and strategy of metagenomics enable the exploration of the genomes and the natural world of microbes without their prior isolation and cultivation. While building on pioneering research over roughly the past two decades and their enhancement by the convergent technologies, this exceptionally interdisciplinary endeavor, which also brings in biogeochemistry and ecological, evolutionary and environmental biology, as well as the wide range of research in microbiology and genomics, is just being defined. To invite the reader to participate in this endeavor, the chapter provides a brief overview of the current state of knowledge and the opportunities that create the excitement and bring together scientists from such diverse disciplines. Research observations in metagenomics, for example, have already demonstrated the vast diversity of microbial proteins and suggest extensive implications for applied life sciences (from environmental science to medicine), as well as for our fundamental understanding of biology.

Ying Xu and J. Peter Gogarten

LIST OF CONTRIBUTORS

Agustino Martínez-Antonio
Departamento de Ingeniería Genética
Instituto Politécnico Nacional
CINVESTAV, IPN. Irapuato, Gto. México
amartinez@ira.cinvestav.mx

Kayo Arima
University of California San Diego
La Jolla, CA 92093-0043, USA
karima@ucsd.edu

Rajeev Azad
Department of Biological Sciences
University of Pittsburgh
Pittsburgh, PA 15260, USA
rka5+@pitt.edu

Fiona S.L. Brinkman
Department of Molecular Biology and Biochemistry
Simon Fraser University
Burnaby, BC, Canada
brinkman@sfu.ca

Dongsheng Che
Department of Biochemistry & Molecular Biology
University of Georgia
Athens, GA 30602-7229, USA
dsche@uga.edu

I-Chun Chou
Department of Biomedical Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
bigjump@gatech.edu

Julio Collado-Vides
Computational Genomics Program
Universidad Nacional Autónoma de México
Av. Universidad s/n, Col. Chamilpa, C.P. 62210
Cuernavaca, Morelos, México
collado@ccg.unam.mx

Phuongan Dam
Department of Biochemistry & Molecular Biology
University of Georgia
Athens, GA 30602-7229, USA
phd@csbl.bmb.uga.edu

Amber Fedynak
Department of Molecular Biology and Biochemistry
Simon Fraser University
Burnaby, BC, Canada
afedynak@sfu.ca

Greg Fournier
Department of Molecular and Cell Biology
University of Connecticut
Storrs, CT 06269-3125, USA
g4nier@gmail.com

J. Peter Gogarten
Department of Molecular and Cell Biology
University of Connecticut
Storrs, CT 06269-3125, USA
j.p.gogarten@uconn.edu

William W.L. Hsiao
Department of Molecular Biology and Biochemistry
Simon Fraser University
Burnaby, BC, Canada
whsiao@som.umaryland.edu

Morgan G.I. Langille
Department of Molecular Biology and Biochemistry
Simon Fraser University
Burnaby, BC, Canada
mlangill@sfu.ca

Pascal Lapierre
Bioinformatics Facility, Bioservices Center
University of Connecticut
Storrs, CT 06269-3149, USA
pascal.lapierre@uconn.edu

Amparo Latorre
Institut Cavanilles de Biodiversitat i Biologia Evolutiva
and Departament de Genètica
Universitat de València, 46071 Valencia, Spain
amparo.latorre@uv.es

Guojun Li
Department of Biochemistry & Molecular Biology
University of Georgia
Athens, GA 30602-7229, USA
guojun@csbl.bmb.uga.edu

Fenglou Mao
Department of Biochemistry & Molecular Biology
University of Georgia
Athens, GA 30602-7229, USA
fenglou@csbl.bmb.uga.edu

Jan Mrázek
Department of Microbiology and Institute of Bioinformatics
University of Georgia
Athens, GA 30602, USA
Mrazek@uga.edu

Maria Poptsova
Department of Molecular and Cell Biology
University of Connecticut
Storrs, CT 06269-3125, USA
maria.poptsova@gmail.com

Francisco J. Silva
Institut Cavanilles de Biodiversitat i Biologia Evolutiva
and Departament de Genètica
Universitat de València, 46071 Valencia, Spain
francisco.silva@uv.es

Zhengchang Su
Department of Computer Science
University of North Carolina
Charlotte, NC 28233, USA
zcsu@email.uncc.edu

Anne Summers
Department of Microbiology
University of Georgia
Athens, GA 30602, USA
summers@uga.edu

Thao Tran
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
tran@ece.gatech.edu

Siren R. Veflingstad
Department of Biomedical Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
and
Max Planck Institute for Biochemistry
Martinsried, Germany
siren.veflingstad@gmail.com

Eberhard O. Voit
Department of Biomedical Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
Eberhard.voit@bme.gatech.edu

Ping Wan
College of Life Science
Capital Normal University
Beijing, P. R. China
wanp@csbl.bmb.uga.edu

John Wooley
University of California San Diego
La Jolla, CA 92093-0043, USA
jwooley@ucsd.edu

Hongwei Wu
School of Electrical and Computer Engineering
Georgia Institute of Technology
Savannah, GA 31407, USA
hongwei.wu@gatech.edu

Ying Xu
Department of Biochemical and Molecular Biology
and Institute of Bioinformatics
University of Georgia
Athens, GA 30602, USA
xyn@bmb.uga.edu

Olga Zhaxybayeva
Department of Biochemistry and Molecular Biology
Dalhousie University
Halifax, Nova Scotia B3H 1X5, Canada
olgazh@dal.ca

Fengfeng Zhou
Department of Biochemical and Molecular Biology
and Institute of Bioinformatics
University of Georgia
Athens, GA 30602, USA
ffzhou@csbl.bmb.uga.edu

ACKNOWLEDGMENTS

During the editing of this book, we, the editors, have received tremendous help from many friends, colleagues and families, to whom we would like to take this opportunity to express our deep gratitude and appreciation. First we would like to thank Dr. Limsoon Wong of the National University of Singapore, who encouraged us to start this book project on behalf of the World Scientific Publishing Company. We are very grateful to the following colleagues who have critically reviewed the drafts of the chapters of the book at various stages: Siv G.E. Andersson, S.J.W. Busby, Luis Delay, Dan Drell, Jonathan Filee, Greg Fournier, Derrick Fouts, Maria Pilar Francino, Robert Friedman, Joel Graber, Roy Gross, Loren Hauser, Jinling Huang, Anders Krogh, Pascal Lapierre, John Leigh, Shoudan Liang, Bill Majoros, Trina Montgomery, Andres Moya, Jan Mrazek, Thane R. Papke, Haluk Resat, Heidi Sofia, Victor Solovyev, Kristen Swithers, Ariane Toussaint, Kalyan Vinnakota, Xiufeng Wan, Yuri Wolf, Alex Worden, Dong Xu, and Olga Zhaxybayeva. Their invaluable input on the scientific content, on the pedagogical style, and on the writing style helped to improve these book chapters significantly. We also want to thank Ms Kristen Swithers of the University of Connecticut for providing artwork for the book cover; and Ms Joan Yantko of the University of Georgia for her tireless help in numerous fronts in this book project, including taking care of a large number of email communications between the editors and the authors, chasing busy authors to get their revisions and other material, helping putting together the index list, and integrating the reference list. Last but not the least, we want to thank our families for their constant support and encouragement during the process of us working on this book project.

CONTENTS