Computing for Comparative Microbial Genomics

Computational Biology

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 discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, to professional text/reference works.

Author guidelines: springer.com > Authors > Author Guidelines

For other titles published in this series, go to http://www.springer.com/series/5769

David W. Ussery • Trudy M. Wassenaar •
Stefano Borini

Computing for Comparative Microbial Genomics

Bioinformatics for Microbiologists

Springer

David W. Ussery, PhD
Department of Systems Biology
The Technical University of Denmark
Lyngby, Denmark
dave@cbs.dtu.dk

Trudy M. Wassenaar, PhD
Molecular Microbiology
and Genomics Consultants
Zotzenheim, Germany
wassenaar_t@yahoo.co.uk

Stefano Borini, PhD
Laboratory of Physical Chemistry
Swiss Federal Institute of
Technology (ETH)
Zurich, Switzerland
stefano.borini@igc.phys.chem.ethz.ch

Computational Biology Series ISSN 1568–2684
ISBN 978-1-84800-254-8 e-ISBN 978-1-84800-255-5
DOI 10.1007/978-1-84800-255-5

Library of Congress Control Number: 2008940847

Printed on acid-free paper

Springer Science+Business Media
springer.com

Preface

Overview and Goals

This book describes how to visualize and compare bacterial genomes. Sequencing technologies are becoming so inexpensive that soon going for a cup of coffee will be more expensive than sequencing a bacterial genome. Thus, there is a very real and pressing need for high-throughput computational methods to compare hundreds and thousands of bacterial genomes.

It is a long road from molecular biology to systems biology, and in a sense this text can be thought of as a path bridging these fields. The goal of this book is to provide a coherent set of tools and a methodological framework for starting with raw DNA sequences and producing fully annotated genome sequences, and then using these to build up and test models about groups of interacting organisms within an environment or ecological niche.

Organization and Features

The text is divided into four main parts: Introduction, Comparative Genomics, Transcriptomics and Proteomics, and finally Microbial Communities. The first five chapters are introductions of various sorts. Each of these chapters represents an introduction to a specific scientific field, to bring all readers up to the same basic level before proceeding on to the methods of comparing genomes. First, a brief overview of molecular biology and of the concept of sequences as biological information are given. The equivalent in the post-genomics era of the 'Central Dogma' of molecular biology (DNA makes RNA makes protein) is that the genome makes the transcriptome, which makes the proteome. Before going on to the details of this, a historical background is provided that pictures the scene of the origins of molecular biology and biological sequences. After this introduction, Chapter 2 describes sequence alignment, the most common procedure used to compare biological sequences. Instead of going into technical details of how exactly these alignments are calculated, the text focuses on their practical use. Chapter 3 introduces bacterial genomes and Chapter 4 deals with the most important databases, whilst Chapter 5 is

an introduction to the computational background of the tools necessary to analyze all of this information.

The second part, on Comparative Genomics (Chapters 6–8), describes some basic methods of comparing genomes. This section introduces various atlases building up to the 'Genome Atlas,' which is our standard visualization tool for representing the DNA sequence of a chromosome in a single figure, mapping the most relevant DNA properties along the chromosome. We have found such atlases very useful for mapping newly sequenced genomes and quickly visualizing regions of potential interest. The value of atlas projections is illustrated by the examples provided.

Part three (Chapters 9–11) takes the reader from genome sequences to RNA sequences (transcriptomics) to proteins (proteomics) and regulation of gene expression. An important overview of experimental results can be obtained by mapping back and visualizing the transcriptomic and proteomic data onto physical chromosomal maps. Examples illustrate how important chromosome location is, and which features can be predicted by careful analysis of genes and their surrounding sequences.

The final part (Chapters 12–14) deals with microbial communities. In a sense this can be thought of as 'population genomics' (as opposed to the more traditional 'population biology' which often focuses on only one or a few genes). First the concept of 'pan-genome' and 'core genome' is introduced (Chapter 12), followed by metagenomics (Chapter 13), and then evolution of microbial communities (Chapter 14). From a larger perspective, population genomics can provide a framework for modeling ecosystems in terms of interacting biological systems.

Target Audiences and Required Background Knowledge

The reader should have basic knowledge about computers and be able to use web interfaces. For programmers, some general knowledge of microbiology is assumed, but it is our hope that both programmers and more 'biology-oriented' readers will find this book helpful. Details on programming were deliberately left out; instead, the text concentrates on the use and interpretation of publicly available web tools. This book has grown out of lectures for the course in Comparative Microbial Genomics,[1] which DWU has taught since 2001 as a full semester length course at the Technical University of Denmark, and as one-week workshops given in Bangkok, Thailand; in Petropolis, Brazil; and in Oslo, Norway.

This book is in a sense merging different scientific languages. The three authors have different scientific and national backgrounds. DWU is from the U.S., studied biochemistry, worked in molecular biology, and for the last 10 years has led a group

[1] http://www.cbs.dtu.dk/dtucourse/programme27444.php

in bioinformatics and genomics. SB is from Italy, studied quantum chemistry with focus on scientific programming, data standardization, and software integration; whereas TMW studied biochemistry and worked in molecular biology and later as a consultant in microbiology. These different backgrounds actually helped to develop a common language in science. The subject area of this textbook is extremely interdisciplinary, covering (bio)chemistry, physics, biology, microbiology, mathematics, and computational science, and by the introduction of concepts (and some jargon) from these various disciplines, the different languages used by specialists are bridged.

This book is meant mainly for people studying bacterial genomes, although of course nearly all of the methods described in the text would work for viral, Archaeal, or Eukaryotic genomes as well. There are two main target audiences. The first is the microbiologist who wants to get the most out of a bacterial genome sequence. This could be a university student, or an experienced laboratory microbiologist who enters the field of genomics. This book enables one to get a handle on how to use high-throughput computational methods to compare only a few, or hundreds of sequenced genomes. The second audience comprises the computer programmers who assist these microbiologists in actually carrying out the analyses. From experience we know there can be communication problems between the experimental bacteriologist who is more laboratory-oriented, and the computer scientist who wants to do everything on computers. Both disciplines are essential in present-day research. This book aims to explain to the computational scientist why and how we want to study bacterial genomes, and what questions we hope to answer. At the same time, it explains to the biologist some of the basics behind the bioinformatic tools that are necessary for research in the field. Bringing these two worlds, scientific interests, and languages together is our ultimate goal.

Notes to the Instructor

There are no exercises or questions at the end of the chapters, although at the end of most chapters textboxes present descriptions of essential methods used. From experience we can say that giving small groups of students a project in which they can choose a recently sequenced bacterial genome and compare it to other similar genomes can produce surprisingly successful results. It is very motivating to work with recently published data (new genome sequence papers are being published on an almost daily basis now), and sometimes the students produce important observations that the authors of the scientific papers had missed! In some occasions, such activities have resulted in a real scientific publication by the students, illustrating how ‘easy’ it is to do these kinds of analyses, as long as one asks relevant questions.

Supplemental Resources

A number of web links are mentioned in the book, and since web addresses are not always stable, a dedicated web page is put up on which all web pages presented in the book are summarized, and as necessary, updated. This can be found at http://comparativemicrobial.com.

Lyngby, Denmark — David Ussery
Zurich, Switzerland — Stefano Borini
Zotzenheim, Germany — Trudy Wassenaar

Acknowledgements

This book is based on input from many people, including our research team and external collaborators. We are grateful for all the advice, assistance, and help we received throughout this project. We thank all current and former members of the Comparative Microbial Genomics group at CBS: in particular, Peter F. Hallin for his excellent programming skills and help with development of many of the programs mentioned in this book; Flemming Hansen for his work on bacterial replication and his vast knowledge of *E. coli*; Henrik J. Nielsen for his help with *E. coli* genomics; Kristoffer Kiil for his help with phylogeny and work on protein function; and Carsten Friis for his assistance with various analyses and for keeping the group running whilst we were writing.

We thank former group members whose work also contributed to this book, including Tim T. Binnewies for his work with *Vibrio* genomes and secretion systems, and Hanni Willenbrock for her work with developing pan-genome microarrays.

We are grateful to external collaborators, notably Thomas Quinn from Denver University for his work on phylogenetic trees whilst on sabbatical in our group; Karin Lagesen from CMBN, Institute of Medical Microbiology, Rikshospitalet University Hospital in Oslo, who has helped with the rRNA and tRNA searches; and Jon Bohlin from the Norwegian School of Veterinary Science, who has helped with analysis of oligonucleotide usage patterns in bacterial genomes.

We would also like to acknowledge help from the many people at CBS, which is currently one of the largest bioinformatics groups in Europe. In particular, we thank Hans Henrik Stærfeldt from the CBS systems administration group, who wrote the original code for the GeneWiz program that is used to construct the atlas plots, and for his help and support over the past 10 years in updating and maintaining GeneWiz. Jannick D. Bendtsen helped us on the secretome, and Thomas Blicher kindly provided wonderful pictures of protein structures. Finally, Søren Brunak, center director for CBS has established a wonderful place to work (including an excellent coffee machine!) and has been supportive of our group since it was formed in 1998.

David would like to thank his students over the years in his Comparative Microbial Genomics course for their many helpful suggestions and comments. He would also like to thank his wife for helpful editorial comments and for her support during the writing of this book.

Stefano would like to thank his parents, Paola Marani and Walter Padovani, for their constant support and trust, and his dear friends Paolo Soriani and Ruggero Paratelli for their life-long support and understanding.

Trudy would like to thank her son Martijn for inventing the analogy of a road to explain DNA strand direction, both of her sons for their understanding and patience, and her husband for his constant support.

Much of the work described in this textbook has been funded by grants over the past decade from the Danish National Research Foundation (Danmarks Grundforskningsfond), Danish Research Councils, and the EU. Many of the calculations presented in this book have been made on our large computer system at CBS, funded in part by money from the Danish Center for Scientific Computing.

Contents

Part III Transcriptomics and Proteomics

PART IV MICROBIAL COMMUNITIES