Genomics and Bioinformatics

With the arrival of genomics and genome sequencing projects, biology has been transformed into an incredibly data-rich science. The vast amount of information generated has made computational analysis critical and has increased demand for skilled bioinformaticians.

Designed for biologists without previous programming experience, this textbook provides a hands-on introduction to Unix, Perl and other tools used in sequence bioinformatics. Relevant biological topics are used throughout the book and are combined with practical bioinformatics examples, leading students through the process from biological problem to computational solution. All of the Perl scripts, sequence and database files used in the book are available for download at the accompanying website, allowing the reader to easily follow each example themselves. Programming examples are kept at an introductory level, avoiding complex mathematics that students often find daunting. The book demonstrates that even simple programs can provide powerful solutions to many complex bioinformatics problems.

Tore Samuelsson is a Professor in Biochemistry and Bioinformatics at the Institute of Biomedicine, University of Gothenburg, Sweden. He has been active in bioinformatics research for more than 15 years and has over 10 years' experience of teaching in the field, including the development of web resources for molecular biology and bioinformatics education.

Genomics and

Bioinformatics

An introduction to programming tools for life scientists

Tore Samuelsson

University of Gothenburg, Sweden

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9781107008564

First published 2012

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Samuelsson, Tore, 1951–
Genomics and bioinformatics : an introduction to programming tools for life scientists / Tore Samuelsson.
pages cm
Includes bibliographical references and index.
ISBN 978-1-107-00856-4
1. Genomics – Data processing. 2. Bioinformatics. I. Title.
QH447.S26 2012
572.8′6 – dc23 2012006477

ISBN 978-1-107-00856-4 Hardback
ISBN 978-1-107-40124-2 Paperback

Additional resources for this publication at www.cambridge.org/9781107008564

CONTENTS

PREFACE

We currently see a vast amount of information being generated as a result of experimental work in biomedicine. Particularly impressive is the development in DNA sequencing. As a result, we are now facing a new era of genomics where a lot of different species, as well as many different human individuals, are being analysed. There are many important biological questions being addressed in such genome-sequencing projects, including questions of medical relevance. A critical technical part of all these projects is computational analysis. With the large amount of sequence information generated, computational analysis is often a bottleneck in the pipeline of a genomics project. Therefore, there is great demand for individuals with the appropriate computational competence. Ideally, such individuals should not only be proficient in the relevant mathematical and computer scientific tools, but should also be able to fully understand the different biological problems that are posed. This book was partly motivated by the urgent need for bioinformatics competence due to recent developments in genomics.

A student or scientist may enter into bioinformatics from different disciplines. This book is written mainly for the biologist that wants to be introduced to computational and programming tools. There are certainly books out there already for that type of audience. However, I was attracted by the idea of assembling a book that would cover a large number of relevant biological topics and, at the same time, illustrate how these topics may be studied using relatively simple programming tools. Therefore, an important principle of the book is that it will attempt to convince the reader that relatively simple programming is sufficient for many bioinformatics tasks and that you need not be a programming expert to be effective. Another important principle of the book is that I wanted the bioinformatics examples to be very practical and explicit. Thus, the reader should be able to follow all the details in a procedure all the way from a biological problem to the results obtained through a technical approach. As one demonstration of this principle, all files and scripts mentioned in this book are available for download at www.cambridge.org/samuelsson. This means the reader is able to try it all out on his/her own computer. I also wanted this book to illustrate the interdisciplinary nature of bioinformatics. Therefore, I have chosen to include a substantial amount of biological motivation as well as programming technology. As a result, the book has a number of rather sudden transitions from descriptions of biological topics to very technical computing matters.

PREFACE

This book is intended as a guide to Perl and Unix-based computing tools for students with a background in molecular biology, biochemistry, cell biology or genomics who have no previous background in this type of computing. In addition, PhD students and scientists at all levels in these fields who want to be introduced to such programming tools should hopefully benefit from the book. The computational parts of the text should be easy to understand for a student lacking a background in computer science; the programming examples presented are at a fairly basic level. The book is complemented by exercises for further study, with mixed levels of difficulty. For the benefit of the student without a mathematical background, the book is by and large non-mathematical, avoiding topics such as probability theory. In summary, the reader of this book should be any student or scientist with some insight into biology, but who also wants to learn about bioinformatics at a more technical level. I also think of the book as being of potential interest to the student or scientist with a background in computer science or programming, but who seeks biological motivation and wants to know more about biological problems that are typically addressed in genomics and bioinformatics.

I present a number of biological and medical topics related to DNA and protein sequences and show how they may be exploited using bioinformatics tools. The book inspires from a biological point of view by selecting relevant and interesting examples; some of the examples will be understood also by non-biologists. Many of the biology topics presented in the book are related to human genomics or human disease, emphasizing the importance of bioinformatics in human medicine. The examples chosen are mainly in the field of sequence bioinformatics. This is a classic area of bioinformatics that has been described previously in many textbooks, but is enjoying renewed interest following current developments in genomics. 'Personal genomics', as touched upon in the final chapters, will be an important area in biomedicine and clinical medicine.

The material in this book is divided into a number of major biological or bioinformatical topics; gene technology, human disease, evolution, gene function, information resources, gene identification methods and personal genomes. Within each of these topics there are different examples of problems that require bioinformatics tools.

The material is organized from the perspective of sequence bioinformatics. First, simple sequence operations such as translation and pattern matching are presented in Chapters 1–5. Chapter 6 deals with RNA secondary structure, and there is a discussion of pairwise alignments and sequence similarity searches in Chapters 7 and 11. Multiple alignments and molecular phylogeny are covered in Chapters 8–10. Different methods of functional assignment are discussed in Chapters 11–13, while molecular sequence databases are discussed in Chapter 14. Finally, gene prediction methods are covered in Chapters 15–17.

From a computational point of view the book focuses on the Unix operating system and the Perl programming language as these are the predominant bases of computational tools in the area of bioinformatics. The Perl content is also organized in a specific fashion, with new concepts introduced in each chapter. For this reason, it is a good idea to read the chapters of the book in the order they are presented. Should a reader contemplate studying the chapters in a different order, Appendix III, providing a short reference guide to Perl, might be helpful. The Perl examples are at a fairly simple level throughout the book, although the Perl code tends to get somewhat more complex towards the end. As mentioned above, a major principle in the design of the book is to convince the reader that relatively simple programming is sufficient to handle many common biological problems. It should also be pointed out that this book is not a complete Unix or Perl reference. In addition, there are more advanced areas of Perl programming that are not covered, such as references and object-oriented programming. A reader seeking information on such topics should consult additional books, such as those listed in Appendix III.

ACKNOWLEDGEMENTS

Nick Lane says in his book *Power, Sex, Suicide: Mitochondria and the Meaning of Life* that 'writing a book sometimes feels like a lonely journey into the infinite, but that is not for lack of support . . . '. In the same vein there are a number of people that I am indebted to in the context of my own journey into the infinite. They are listed below in a (partly) random order.

A number of people provided help on specific chapters. For the sections on NCBI Entrez and BLAST, I received information and comments from Peter Cooper, Dennis Benson and Eric Sayers, all at the NCBI. Marie–Claude Blatter of the Swiss–Prot communication team provided information about Swiss–Prot. Sean Eddy, at HHMI Janelia Farm Research Campus, provided helpful information regarding HMMER and Infernal. Gunnar Hansson, University of Gothenburg, with whom I collaborated on mucin bioinformatics, had helpful comments on the chapter dealing with these proteins. I'm grateful to Magnus Alm Rosenblad of the Department of Cell and Molecular Biology, University of Gothenburg, for discussions about chloroplast RNAs and many other topics that unfortunately would not fit into this book. Stefan Washietl allowed me to use his code generating double-stranded DNA shown in Appendix III. I'm grateful to Russell Doolittle for feedback on the chapter about blood clotting and to Joe Felsenstein for information about Dnapars. For the story on the HIV criminal case, my sources of information included an article by Pam Lambert in *People* (http://www.people.com) and one article by Stephen G. Michaud at truTV.com (http://trutv.com). I'm also indebted to a large number of anonymous Wikipedia authors.

For the chapter on thylacine, I had much help from Robert Paddle of the Australian Catholic University. In addition, his book, *The Last Tasmanian Tiger*, was a great source of information. Caroline Freeman of the University of Tasmania provided comments on the thylacine chapter, and also supplied a copy of the Burrell photograph. Thanks also to Ellen Alers at the Smithsonian Institution Archives, Washington, for sending me the photograph of the Washington thylacines. Jacqui Ward of the Tasmanian Museum and Art Gallery provided the photograph of two thylacines in Beaumaris zoo. The image of the Darwin termite in the chapter about termites was obtained courtesy of Katja Schultz of the Tree of Life Project (http://tolweb.org) and Smithsonian Institution, National Museum of Natural History.

For the chapters on personal genomes I received help from a number of people. I'm grateful to Adam Siepel and Melissa Jane Hubisz, Department of Biological

Statistics and Computational Biology at Cornell University, for sharing their SNP data from a number of human individuals. I gratefully acknowledge help regarding the Bushmen data from Stephan Schuster and Webb Miller at the Center for Comparative Genomics and Bioinformatics, Penn State University. They also supplied information on the thylacine story. In addition, Stephan Schuster generously provided photographs of the Bushmen individuals. With regard to the chapter on the family quartet, I received comments from Jared Roach, Gustavo Glusman and David Galas at the Institute of Systems Biology, Seattle. In particular, I'm grateful to Gustavo Glusman, who produced simulated data for chromosome 4 and provided a lot of helpful information concerning his processing of genotype data.

A number of people at the University of Gothenburg and at Chalmers University of Technology read my draft manuscript and had very useful comments: Marina Axelson-Fisk, Per Elias, Graham Kemp and Ka-Wei Tang. In particular, I'm also grateful to Marcela Dávila López for her detailed comments and ideas for improvement. Katrina Halliday, Hans Zauner, Lynette Talbot, Jonathan Ratcliffe and other staff at Cambridge University Press were very positive and helpful during the compilation of this book. I also thank Gary Smith for careful copy-editing. In addition, I'm very grateful to the Hasselblad Foundation for awarding me a stipend to spend two months in Grez-sur-Loing in France to work on the book. Special thanks to Birgitta Bergenholtz at the Foundation and Bernadette Plissart at Hotel Chevillon in Grez. I take this opportunity to apologize to the students of a bioinformatics course that I notoriously neglected while I was in France. I also sincerely thank my three 'A's, Annika, Anders and Anna for their contributions, including a set of Lego pieces, but most of all I acknowledge their great support and patience during the time I put together this book.

Finally, as an important source of inspiration I would like to mention my mentor and former supervisor Ulf Lagerkvist, who tragically passed away in 2010. He was an inspiration to all his students, not only because of his scientific achievements and attitude towards science, but also because he authored a number of highly readable books in the areas of life sciences and scientific history.

DESIGN AND CONVENTIONS OF THIS BOOK

This book is designed such that it covers a number of biological topics, one in each chapter. The topics are arranged in the following major categories:

- introduction to genetic information (one chapter)
- gene technology (three chapters)
- human disease (three chapters)
- evolution (three chapters)
- gene function (three chapters)
- information resources (one chapter)
- gene identification methods (three chapters)
- personal genomes (three chapters).

In each of the chapters one or more specific problems are addressed in a bioinformatics section where Perl, Unix or other bioinformatics software are used. The Unix or Perl topics that are novel to the chapter are listed in a box at the beginning of the bioinformatics section. In the bioinformatics section of each chapter the following conventions are used. Some text is presented in a coloured `fixed-width font`. These are (1) Unix command lines; (2) Perl code; and (3) names of files, programs or Unix utilities. Whenever something is to be typed at the Unix command line, this is indicated with a `%` symbol, to represent the Unix command line prompt. Thus, a reader trying these commands at his/her computer should *not* type the `%` symbol. An example would be:

```
% uname
```

which means that by typing 'uname' the program `uname` (a Unix utility to print system information) will be executed.

Complete Perl scripts are present within specifically highlighted boxed areas. The Perl scripts are, to some extent, explained and commented within the actual scripts, but mainly in the text preceding or following the script. All files that are used in the examples of this book, including all Perl scripts, are available for download at the supplementary website (www.cambridge.org/samuelsson).

For more background and practical information on Unix and Perl, the reader is referred to Appendices I and III. Appendices I and III also contain suggestions for further reading. A selection of bioinformatics software that is used in a Unix environment is presented in Appendix II. The web resources provided with this book have more information, such as solutions to the Perl exercises of the book, Python examples and a listing of bioinformatics resources.

Some of the figures in this book were created with R, a free software environment for statistical computing and graphics (http://www.r-project.org). In such cases, the R scripts are available for downloading from the web resource for this book. The scripts are not explained in any detail, but a short R reference is provided in Appendix IV.

Throughout this book it is assumed that the reader has access to a computer running a Unix operating system and that Perl is installed on this system. For more general background and technical information about Unix and Perl, see Appendices I and III.

This book assumes the reader has a basic knowledge of molecular biology, biochemistry or cell biology. In case the reader needs more background information in these areas, the following are all examples of excellent textbooks:

Alberts, B. (2008). *Molecular Biology of the Cell: Reference Edition*. New York, Garland Science.

Barton, N. H., D. E. G. Briggs, J. A. Eisen, D. B. Goldstein and N. H. Patel (2007) *Evolution*. Cold Spring Harbor, NY, Cold Spring Harbor Laboratory Press.

Berg, J. M., J. L. Tymoczko and L. Stryer (2010). *Biochemistry*. New York, W. H. Freeman.

Lodish, H. F. (2008). *Molecular Cell Biology*. New York, W. H. Freeman.