

Genomics and Bioinformatics

An Introduction to Programming
Tools for Life Scientists



TORE SAMUELSSON

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

© Tore Samuelsson 2012

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

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

Cambridge University Press has no responsibility for the persistence or
accuracy of URLs for external or third-party internet websites referred to
in this publication, and does not guarantee that any content on such
websites is, or will remain, accurate or appropriate.

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

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.

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.

CONTENTS

Preface | xi

Acknowledgements | xiv

Design and conventions of this book | xvi

1 Introduction: working with the molecules of life in the computer | 1

Life on Earth and evolution | 2

The machinery of genetic information: more about DNA | 4

Genes and genomes | 7

Genes at work: transcription and translation | 8

Organization of the human genome | 11

Inferring products of DNA replication | 12

Inferring RNA products of transcription | 14

Inferring protein products of translation | 14

Exercises | 17

2 Gene technology: cutting DNA | 19

Early days of restriction enzymes | 20

Properties of restriction enzymes | 21

Pattern matching | 23

Identifying restriction enzyme sites with Perl | 24

Exercises | 29

3 Gene technology: knocking genes down | 31

Interfering with gene expression | 31

Small silencing RNAs | 32

RNAi: functions and applications | 34

Silencing RNAs and design principles | 35

Identifying siRNA candidates | 37

Exercises | 42

4 Gene technology: amplifying DNA | 44

What is PCR? | 44

PCR applications | 46

Primer design | 46

	Reverse translation		48
	Good manners during Perl programming		51
	Exercises		54
5	Human disease: when DNA sequences are toxic		55
	Inherited disease and changes in DNA		55
	Huntington's disease and CAG repeat expansion		57
	Identifying mRNAs with CAG repeats		59
	Exercises		65
6	Human disease: iron imbalance and the iron responsive element		66
	An inherited disease affecting the iron-binding protein ferritin		66
	Many proteins of iron metabolism are regulated at the level of translation		66
	Structure of the iron responsive element		68
	Identifying the iron responsive element		69
	Exercises		72
7	Human disease: cancer as a result of aberrant proteins		74
	Cancer as a genetic disease		74
	Cancer and DNA repair		75
	Chromosome rearrangements and the Philadelphia chromosome		76
	Dotplots and alignments		77
	BLAST		81
	Using BLAST to examine the BCR-ABL fusion protein		83
	Exercises		90
8	Evolution: what makes us human?		92
	Genetic differences between humans and chimpanzees		92
	A protein related to human speech		93
	FOXP2 in other animals		94
	Comparing FOXP2 in different animals		95
	Changes in FOXP2 specific to humans		101
	Exercises		102
9	Evolution: resolving a criminal case		105
	A fatal injection		105
	Methods of molecular phylogeny		106
	Examination of the criminal case		113
	Exercises		120

- 10 Evolution: the sad case of the Tasmanian tiger | 121
 - Extinction | 121
 - The thylacine | 122
 - Tiger history | 123
 - Recent DNA analysis | 125
 - Inferring the phylogeny of marsupials | 126
 - Examining taxonomy | 129
 - Exercises | 135
- 11 A function to every gene: termites, metagenomics and learning about the function of a sequence | 137
 - Assigning function based on sequence similarity | 137
 - Metagenomics | 138
 - The other genome | 139
 - Termites and cellulose digestion | 140
 - Assigning function to termite sequences | 141
 - Exercises | 147
- 12 A function to every gene: royal blood and order in the sequence universe | 150
 - Royal disease | 150
 - Blood-clotting pathways | 151
 - Protein domain architecture | 153
 - Bioinformatics of protein domains | 154
 - Bioinformatic analysis of blood-clotting proteins | 156
 - Evolution of blood clotting | 160
 - Exercises | 162
- 13 A function to every gene: a slimy molecule | 164
 - Extensive sugar decoration | 164
 - Mucins and repeats | 165
 - Computational identification of mucin domains | 168
 - Exercises | 171
- 14 Information resources: learning about flu viruses | 173
 - Short history of sequence databases | 173
 - Features of nucleotide sequence databases | 175
 - Comparative genomics | 177
 - Protein sequence and structure data | 178
 - Exploring databases at the NCBI | 180

	NCBI databases in eUtils		181
	NCBI query syntax		182
	The Entrez Programming Utilities		185
	Parameters supplied to eUtils: scripts and construction of URLs		185
	EFetch		187
	ESearch		188
	Further analysis of influenza viruses: extracting and filtering information with Perl and Unix tools		192
	ELink		195
	Exercises		196
	NCBI documentation resources		196
15	Finding genes: going ashore at CpG islands	 	198
	Transcription and its regulation		198
	DNA sequences that influence transcription		199
	CpG islands		200
	Finding CpG islands		202
	Exercises		206
16	Finding genes: in the world of snurps	 	208
	Methods of gene prediction		208
	The splicing machinery		209
	Constructing a PSSM		213
	Scoring with a PSSM		216
	Exercises		219
17	Finding genes: hunting for the distant RNA relatives	 	222
	The RNA world		222
	Properties of RNA and computational RNA finding		223
	An RNA involved in protein transport		227
	The organelles and their evolution		227
	The quest for chloroplast RNAs		229
	Automating tasks with Unix and Perl		237
	Exercises		238
18	Personal genomes: the differences between you and me	 	241
	Personal genomes		241
	Individual variation and SNPs		242
	Counting SNPs		244
	Exercises		250

19	Personal genomes: what's in my genome?	252
	Human roots	252
	An SNP dataset of South African individuals	253
	What SNPs are unique to the Bushmen?	256
	What SNPs are in coding regions?	257
	A bitter taste	260
	Constructing your own modules	263
	Exercises	264
20	Personal genomes: details of family genetics	266
	Basic principles of genetic inheritance	266
	Analysis of a family quartet	269
	Where are the crossing-over sites?	272
	Exercises	277
	Appendix I: brief Unix reference	278
	Appendix II: a selection of biological sequence analysis software	289
	Appendix III: a short Perl reference	300
	Appendix IV: a brief introduction to R	323
	Index	330