Bioinformatics for Evolutionary Biologists

Bernhard Haubold · Angelika Börsch-Haubold

Bioinformatics for Evolutionary Biologists

A Problems Approach

Springer

Bernhard Haubold
Department of Evolutionary Genetics
Max-Planck-Institute for Evolutionary Biology
Plön, Schleswig-Holstein
Germany

Angelika Börsch-Haubold
Plön, Schleswig-Holstein
Germany

ISBN 978-3-319-67394-3 ISBN 978-3-319-67395-0 (eBook)
https://doi.org/10.1007/978-3-319-67395-0

Library of Congress Control Number: 2017955660

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Evolutionary biologists have two types of ancestors: naturalists such as Charles Darwin (1809–1892) and theoreticians such as Ronald A. Fisher (1890–1962). The intellectual descendants of these two scientists have traditionally formed quite separate tribes. However, the distinction between naturalists and theoreticians is rapidly fading these days: Many naturalists spend most of their time in front of computers analyzing their data, and quite a few theoreticians are starting to collect their own data. The reason for this coalescence between theory and experiment is that two hitherto expensive technologies have become so cheap, they are now essentially free: computing and sequencing. Computing became affordable in the early 1980s with the advent of the PC. More recently, next generation sequencing has allowed everyone to sequence the genomes of their favorite organisms. However, analyzing this data remains difficult.

The difficulties are twofold: conceptual, which method should I use, and practical, how do I carry out a certain computation. The aim of this book is to help the reader overcome both difficulties. We do this by posing a series of problems. These come in two forms, paper and pencil problems, and computer problems. Our choice of concepts is centered on the analysis of sequences in an evolutionary context. The aim here is to give the reader a look under the hood of the programs applied in the computer problems. The computer problems are solved in the same environment used for decades by scientists, the UNIX command line, also known as the shell. This is available on all three major desktop operating systems, Windows, Linux, and OS-X. Like any skill worth learning, using the shell takes practice. The computer problems are designed to give the reader plenty of opportunity for that.

In Chap. 1, we introduce the command line. After explaining how to get started, we deal with plain text files, which serve as input and output of most UNIX operations. Many of these operations are themselves text files containing commands to be executed on some input. Such command files are called scripts, and their treatment concludes Chap. 1.

In Chap. 2, the newly acquired UNIX skills are used to explore a central concept in Bioinformatics: sequence alignment. A sequence alignment represents an evolutionary hypothesis about which residues have a recent common ancestor. This is determined using optimal alignment methods that extract the best out of a very large number of possible alignments. However, this optimal approach consumes a lot of time and memory.

The computation of exact matches, the topic of Chap. 3, is less resource intensive than the computation of alignments. Taken by themselves, exact matches are also less useful than alignments, because exact matches cannot take into account mutations. Nevertheless, exact matching is central to many of the most popular methods for inexact matching. We begin with methods for exact matching in time proportional to the length of the sequence investigated. Then we concentrate on methods for exact matching in time independent of the text length. This is achieved by indexing the input sequence through the construction of suffix trees and suffix arrays.

In Chap. 4, we show how to combine alignment with exact matching to obtain very fast programs. The most famous example of these is BLAST, which is routinely used to find similarities between sequences. Up to now we have only looked at pairwise alignment. At the end of Chap. 4, we generalize this to multiple sequence alignment.

In Chap. 5, multiple sequence alignments are used to construct phylogenies. These are hypotheses about the evolution of a set of species. If we zoom in from evolution between species to evolution within a particular species, we enter the field of population genetics, the topic of Chap. 6. Here, we concentrate on modeling evolution by following the descent of a sample of genes back in time to their most recent common ancestor. These lines of descent form a tree known as the coalescent, the topic of much of modern population genetics.

We conclude in Chap. 7 by introducing two miscellaneous topics: statistics and relational databases. Both would deserve books in their own right, and we restrict ourselves to showing how they fit in with the UNIX command line.

Our course is sequence-centric, because sequence data permeates modern biology. In addition, these data have attracted a rich set of computer methods for data analysis and modeling. The sequences we analyze can be downloaded from the companion website for this book:

http://guanine.evolbio.mpg.de/problemsBook/

To these sequences, we apply generic tools provided by the UNIX environment, published bioinformatics software, and programs written for this course. The latter are designed to allow readers to analyze a particular computational method. The programs are also available from the companion site.

At the back of the book, we give complete solutions to all the problems. The solutions are an integral part of the course. We recommend you attempt each problem in the order in which they are posed. If you find a solution, compare it to ours. If you cannot find a solution, read ours and try again. If our solution is unclear or you have a better one, please drop us a line at problemsbook@evolbio.mpg.de

The tongue-in-cheek Algorithm 1 summarizes these recommendations.

Algorithm 1 Using the Solutions

```
while problem unsolved do
    solve problem
    read solution
    if solution unclear or your solution is better than ours then
        drop us a line
    end if
end while
```

This book has been in the works since 2003 when BH started teaching Bioinformatics at the University of Applied Sciences, Weihenstephan. We thank all the students who gave us feedback on this material as it evolved over the years. We would also like to thank a few individuals who contributed in more specific ways to the gestation of this book: Mike Travisano (University of Minnesota) gave us encouragement at a critical time. Nicola Gaedeke and Peter Pfaffelhuber (University of Freiburg) commented on an early draft, and our students Linda Krause, Xiangyi Li, Katharina Dannenberg, and Lina Urban read large parts of the manuscript in one of the many guises it has taken over the years. We are grateful to all of them.

Plön, Germany
July 2017

Bernhard Haubold
Angelika Börsch-Haubold

The original version of the book backmatter was revised: For detailed information please see Erratum. The erratum to this chapter is available at https://doi.org/10.1007/978-3-319-67395-0_9

Contents