Genomic Perl

This introduction to computational molecular biology will help programmers and biologists learn the skills they need to start work in this important, expanding field. The author explains many of the basic computational problems and gives concise, working programs to solve them in the Perl programming language. With minimal prerequisites, the author explains the biological background for each problem, develops a model for the solution, and then introduces the Perl concepts needed to implement the solution.

The book covers pairwise and multiple sequence alignment, fast database searches for homologous sequences, protein motif identification, genome rearrangement, physical mapping, phylogeny reconstruction, satellite identification, sequence assembly, gene finding, and RNA secondary structure. The author focuses on one or two practical approaches for each problem rather than an exhaustive catalog of ideas. His concrete examples and step-by-step approach make it easy to grasp the computational and statistical methods, including dynamic programming, branch-and-bound optimization, greedy methods, maximum likelihood methods, substitution matrices, BLAST searching, and Karlin–Altschul statistics.

Rex A. Dwyer founded Genomic Perl Consultancy, Inc. in July 2001. He was formerly an Associate Professor of Computer Science at North Carolina State University, where he taught data structures, algorithms, and formal language theory and demonstrated his skill as both theoretician and practitioner. He has published more than a dozen papers in academic journals such as *Advances in Applied Probability, Algorithmica,* and *Discrete and Computational Geometry.* His accomplishments as a Perl software developer include a proprietary gene-finding system for the Novartis Agribusiness Biotechnology Research Institute (now part of Syngenta, Inc.) and a web-accessible student records database for faculty colleagues at NCSU.

Genomic Perl

From Bioinformatics Basics to Working Code

REX A. DWYER
Genomic Perl Consultancy, Inc.

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2002

Printed in the United States of America

Typeface Times 10/13 pt. *System* AMS-TEX [FH]

A catalog record for this book is available from the British Library.

Library of Congress Cataloging in Publication data available

ISBN 0 521 80177 X hardback

A mi chapinita y a mi chapinito.

Contents

Preface

This book is designed to be a concrete, digestible introduction to the area that has come to be known as "bioinformatics" or "computational molecular biology". My own teaching in this area has been directed toward a mixture of graduate and advanced undergraduate students in computer science and graduate students from the biological sciences, including biomathematics, genetics, forestry, and entomology. Although a number of books on this subject have appeared in the recent past – and one or two are quite well written – I have found none to be especially suitable for the widely varying backgrounds of this audience.

My experience with this audience has led me to conclude that its needs can be met effectively by a book with the following features.

- To meet the needs of computer scientists, the book must teach basic aspects of the structure of DNA, RNA, and proteins, and it must also explain the salient features of the laboratory procedures that give rise to the sorts of data processed by the algorithms selected for the book.
- To meet the needs of biologists, the book must (to some degree) teach programming and include *working programs* rather than abstract, high-level descriptions of algorithms – yet computer scientists must not become bored with material more appropriate for a basic course in computer programming.
- Justice to the field demands that its statistical aspects be addressed, but the background of the audience demands that these aspects be addressed in a concrete and relatively elementary fashion.

To meet these criteria, a typical chapter of this book focuses on a single problem that arises in the processing of biological sequence data: pairwise sequence alignment, multiple alignment, sequence database searching, phylogeny reconstruction, genome rearrangement, and so on. I outline both the biological origins of the input and the interpretation of the desired output; then I develop a single algorithmic approach to the problem. Finally, I show how to implement the algorithm as a working Perl program. Variations on the problem and/or improvements to the program are

presented in exercises at the end of each chapter. In a few cases, I develop a straightforward but inefficient algorithm in one chapter and devote the following chapter to a more sophisticated approach. Bibliographic notes in each chapter are limited to a half dozen of the most accessible references; I assume that a serious student can use these hooks into the literature to track down other relevant resources.

The choice of the Perl language as the medium for presenting algorithms might surprise some, but it has many advantages.

- Perl's built-in strings, lists, and hash tables make it possible to express some algorithms very concisely. (Chapter 12's 80-line program for constructing suffix trees is an outstanding example.)
- Perl is already widely used for server-side scripting (CGI) in web-based applications, and a large library of code (the bioPerl effort described at `www.bioperl.org`) is freely available to assist bioinformatic programmers.
- Perl falls outside the standard computer science curriculum. This means that attention to language details will be of some interest even to students with strong computer science backgrounds.
- Perl is portable; it runs on all major operating systems.
- Perl is available without charge as part of Linux distributions and from the World Wide Web at `http://www.perl.org`.
- Rare but legal Perl constructs like @{$_[$#_]||[ ]} inspire awe in the uninitiated; this awe sometimes rubs off on Perl programmers.

Perl also has a few disadvantages.

- Perl programs are not compiled to native-mode code, so they often run many times slower than equivalent C or C++ programs in computationally intensive applications. (In CGI applications dominated by disk-access or network-response times, this disadvantage evaporates.)
- Perl's built-in strings, lists, and hash tables sometimes hide potential performance problems that can be overcome only with nonintuitive tricks.
- Perl is profligate in its use of memory. This means that the input size that a Perl program can handle may be many times smaller than for an equivalent C or C++ program.

All in all, the advantages prevail for the purposes of this book, and – although using Perl makes many of the programs in this book teaching toys rather than production-grade tools – they do work. Unlike pseudocode descriptions, they can be modified to print traces of their operation or to experiment with possible improvements. They can also serve as prototypes for more efficient implementations in other languages.

In the interest of clarity, here are a few words about what this book is *not*.

- This book is not a consumer's guide to the many software packages available to assist the biologist with sequence assembly, database searching, phylogeny reconstruction, or other tasks. Such packages will be mentioned in passing from

time to time, but the focus of this book is on how these packages solve (or might solve) the problems presented to them and not on how problems must be presented to these packages.

- This book is not an encyclopedic compendium of computational aspects of molecular biology. The problems included are presumed to be significant to biologists, but other equally significant problems have doubtless been omitted.
- Likewise, this book is not a tutorial or reference manual on the Perl language. Features of Perl will be explained and used as needed, and diligent readers should come away with a good knowledge of Perl, but algorithms are the focus, and many widely used Perl features will be left unmentioned.

I embarked upon this writing project not to convey to others my own research contributions in the area (since I have made none) but rather to study the contributions of others and make them more accessible to students. It is certain that my own lack of expertise will show through now and again. Despite this, I hope that my concrete approach will permit a much larger audience to understand and to benefit from the contributions of the real experts.

Acknowledgments

Special thanks to Cambridge University Press and my editor, Lauren Cowles, for accepting a book proposal from a first-time author with no publications dealing directly with the subject matter of the book. Ms. Cowles's careful comments and other assistance in the preparation of the manuscript have been invaluable.

Phil Green of the University of Washington was kind enough to allow me to use the source of his PHRAP program to develop Chapter 13. Alejandro Schäffer and Steven Altshul of NCBI offered helpful feedback on chapters dealing with multiple alignment and BLAST.

While I was a faculty member at North Carolina State University, the students of CSC 695 in Spring 1997 and Spring 1999 and CSC 630 in Spring 2001 accepted and proofread an ever-growing collection of notes and programs upon which this book was eventually based. Among these, Barry Kesner deserves special mention for giving very helpful feedback on several chapters. Tim Lowman, systems administrator extraordinaire, was most helpful in answering Perl questions in the early weeks of this project.

Financially, this writing project was partially supported by NC State University in the form of sabbatical support. The laboratory now known as Syngenta Biotechnology, Inc. (in Research Triangle Park, NC) complemented this support on several occasions while permitting me to experiment with proprietary data.

Last but not least, Pam and Kevin Bobol, proprietors of the New World Cafe and Larry's Beans, provided the chairs in which a large fraction of this book was written. Staff members Amy, D.J., Heather, James, Jeramy, Kathy, Margaret, Marwa, Patty, Sam, Sue, and others served the hot liquids that kept my fingers moving.

As expected, any remaining errors are rightfully counted against my own account alone and not against that of my editor, colleagues, students, or barristas.