生物信息学中的计算机技术

（影印版）

Developing

Bioinformatics Computer Skills

O'REILLY®
科学出版社

Cynthia Gibas & Per Jambeck

Developing Bioinformatics Computer Skills
生物信息学中的计算机技术

Cynthia Gibas and Per Jambeck

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Paris • Sebastopol • Taipei • Tokyo

O'Reilly & Associates, Inc. 授权科学出版社出版

科学出版社

图书在版编目（CIP）数据

生物信息学中的计算机技术：/（美）吉巴斯（Gibas, C.）、（美）杰贝克（Jambeck, J.）著—影印版．北京：科学出版社，2002. 5
书名原文：Developing Bioinformatics Computer Skills
ISBN 7-03-010421-8

Ⅰ.生… Ⅱ.①吉… ②杰… Ⅲ.计算机应用-生物信息论-英文 Ⅳ.Q811.4-39
中国版本图书馆 CIP 数据核字（2002）第 029659 号

北京市版权局著作权合同登记
图字：01-2002-1098 号

书　　名/ 生物信息学中的计算机技术
书　　号/ ISBN 7-03-010421-8/Q.1171
责任编辑/ 马学海
封面设计/ Ellie Volckhausen，张健
出版发行/ 科学出版社（www.sciencep.com）
地　　址/ 北京东黄城根北街 16 号（邮政编码 100717）
经　　销/ 各地新华书店
印　　刷/ 北京艺辉印刷有限公司
开　　本/ 787 毫米 × 1092 毫米　16 开本　28 印张
版　　次/ 2002 年 5 月第一版　2002 年 5 月第一次印刷
印　　数/ 0001-6000 册
定　　价/ 48.00 元（册）

Preface

Computers and the World Wide Web are rapidly and dramatically changing the face of biological research. These days, the term "paradigm shift" is used to describe everything from new business trends to new flavors of cola, but biological science is in the midst of a paradigm shift in the classical sense. Theoretical and computational biology have existed for decades on the "fringe" of biological science. But within just a few short years, the flood of new biological data produced by genomics efforts and, by necessity, the application of computers to the analysis of this genomic data, has begun to affect every aspect of the biological sciences. Research that used to start in the laboratory now starts at the computer, as scientists search databases for information that might suggest new hypotheses.

In the last two decades, both personal computers and supercomputers have become accessible to scientists across all disciplines. Personal computers have developed from expensive novelties with little real computing power into machines that are as powerful as the supercomputers of 10 years ago. Just as they've replaced the author's typewriter and the accountant's ledger, computers have taken their place in controlling and collecting data from lab equipment. They have the potential to completely replace laboratory notebooks and files as a means of storing data. The power of computer databases allows much easier access to stored data than nonelectronic forms of recording. Beyond their usefulness for the storage, analysis, and visualization of data, however, computers are powerful devices for understanding any system that can be described in a mathematical way, giving rise to the disciplines of computational biology and, more recently, bioinformatics.

Bioinformatics is the application of information technology to the management of biological data. It's a rapidly evolving scientific discipline. In the last two decades, storage of biological data in public databases has become increasingly common,

and these databases have grown exponentially. The biological literature is growing exponentially as well. It's impossible for even the most zealous researcher to stay on top of necessary information in the field without the aid of computer-based tools, and the Web has made it possible for users at any location to interact with programs and databases at any other site—provided they know how to build the right tools.

Bioinformatics is first and foremost a biological science. It's often less about developing perfectly elegant algorithms than it is about answering practical questions. Bioinformaticians (or bioinformaticists, if you prefer) are the tool-builders, and it's critical that they understand biological problems as well as computational solutions in order to produce useful tools. Bioinformatics algorithms need to encompass complex scientific assumptions that can complicate programming and data modeling in unique ways.

Research in bioinformatics and computational biology can encompass anything from the abstraction of the properties of a biological system into a mathematical or physical model, to the implementation of new algorithms for data analysis, to the development of databases and web tools to access them. To engage in computational research, a biologist must be comfortable using software tools that run on a variety of operating systems. This book introduces and explains many of the most popular tools used in bioinformatics research. We've included lots of additional information and background material to help you understand how the tools are best used and why they are important. We hope that it will help you through the first steps of using computers productively in your research.

Audience for This Book

Most biological science students and researchers are starting to use computers as more than word-processing or data-collection and plotting devices. Many don't have backgrounds in computer science or computational theory, and to them, the fields of computational biology and bioinformatics may seem hopelessly large and complex. This book, motivated by our interactions with our students and colleagues, is by no means a comprehensive bible on all aspects of bioinformatics. It is, however, a thoughtful introduction to some of the most important topics in bioinformatics. We introduce standard computational techniques for finding information in biological sequence, genome, and molecular structure databases; we talk about how to identify genes and detect characteristic patterns that identify gene families; and we discuss the modeling of phylogenetic relationships, molecular structures, and biochemical properties. We also discuss ways you can use your computer as a tool to organize data, to think systematically about data-analysis processes, and to begin thinking about automation of data handling.

Bioinformatics is a fairly advanced topic, so even an introductory book like this one assumes certain levels of background knowledge. To get the most out of this book you should have some coursework or experience in molecular biology, chemistry, and mathematics. An undergraduate course or two in computer programming would also be helpful.

Structure of This Book

We've arranged the material in this book to allow you to read it from start to finish or to skip around, digesting later sections before previous ones. It's divided into four parts:

Part 1, *Introduction*

Chapter 1, *Biology in the Computer Age*, defines bioinformatics as a discipline, delves into a bit of history, and provides a brief tour of what the book covers and why.

Chapter 2, *Computational Approaches to Biological Questions*, introduces the core concepts of bioinformatics and molecular biology and the technologies and research initiatives that have made increasing amounts of biological data available. It also covers the ever-growing list of basic computer procedures every biologist should know.

Part II, *The Bioinformatics Workstation*

Chapter 3, *Setting Up Your Workstation*, introduces Unix, then moves on to the basics of installing Linux on a PC and getting software up and running.

Chapter 4, *Files and Directories in Unix*, covers the ins and outs of moving around a Unix filesystem, including file hierarchies, naming schemes, commonly used directory commands, and working in a multiuser environment.

Chapter 5, *Working on a Unix System*, explains many Unix commands users will encounter on a daily basis, including commands for viewing, editing, and extracting information from files; regular expressions; shell scripts; and communicating with other computers.

Part III, *Tools for Bioinformatics*

Chapter 6, *Biological Research on the Web*, is about the art of finding biological information on the Web. The chapter covers search engines and searching, where to find scientific articles and software, how to use the online information sources, and the public biological databases.

Chapter 7, *Sequence Analysis, Pairwise Alignment, and Database Searching*, begins with a review of molecular evolution and then moves on to cover the

basics of pairwise sequence-analysis techniques such as predicting gene location, global and local alignment, and local alignment-based searching against databases using BLAST and FASTA. The chapter concludes with coverage of multifunctional tools for sequence analysis.

Chapter 8, *Multiple Sequence Alignments, Trees, and Profiles,* moves on to study groups of related genes or proteins. It covers strategies for multiple sequence alignment with tools such as ClustalW and Jalview, then discusses tools for phylogenetic analysis, and constructing profiles and motifs.

Chapter 9, *Visualizing Protein Structures and Computing Structural Properties,* covers 3D analysis of proteins and the tools used to compute their structural properties. The chapter begins with a review of protein chemistry and quickly moves to a discussion of web-based protein structure tools; structure classification, alignment, and analysis; solvent accessibility and solvent interactions; and computing physicochemical properties of proteins. The chapter concludes with structure optimization and a tour through protein resource databases.

Chapter 10, *Predicting Protein Structure and Function from Sequence,* covers the tools that determine the structures of proteins from their sequences. The chapter discusses feature detection in protein sequences, secondary structure prediction, predicting 3D structure. It concludes with an example project in protein modeling.

Chapter 11, *Tools for Genomics and Proteomics,* puts it all together. Up to now we've covered tools and techniques for analyzing single sequences or structures, and for comparing multiple sequences of single-gene length. This chapter discusses some of the datatypes and tools that are becoming available for studying the integrated function of all the genes in a genome, including sequencing an entire genome, accessing genome information on the Web, annotating and analyzing whole genome sequences, and emerging technologies and proteomics.

Part IV, *Databases and Visualization*

Chapter 12, *Automating Data Analysis with Perl,* shows you how a programming language such as Perl can help you sift through mountains of data to extract just the information you require. It won't teach you to program in Perl, but the chapter gives you a brief introduction to the language and includes examples to start you on your way toward learning to program.

Chapter 13, *Building Biological Databases,* is an introduction to database concepts. It covers the types of databases used in biological research, the database software that builds them, database languages (in particular, the SQL language), and developing web-based software that interacts with databases

Chapter 14, *Visualization and Data Mining,* covers the computational tools and techniques that allow you to make sense of your results. The first part of the

chapter introduces programs that are used to visualize data arising from bioinformatics research. They range from general-purpose plotting and statistical packages for numerical data, such as Grace and *gnuplot*, to programs such as TEXshade that are dedicated to presenting sequence and structural information in an interpretable form. The second part of the chapter presents tools for data mining—the process of finding, interpreting, and evaluating patterns in large sets of data—in the context of applications in bioinformatics.

Our Approach to Bioinformatics

We confess, we're structural biologists (biophysicists, actually). We have a hard time thinking about genes without thinking about their protein products. DNA sequences, to us, aren't just sequences. To a structural biologist, genes (with a few exceptions) imply 3D structures, molecular shapes and conformational changes, active sites, chemical reactions, and detailed intermolecular interactions. Our focus in this book is on using sequence information as structural biologists and biochemists tend to use it—to understand the chemical basis of biological function. We've probably neglected some applications of sequence analysis that are dear to the hearts of molecular biologists and geneticists, so feel free send us your comments.

URLs Referenced in This Book

For more information on the URLs we reference in this book and for additional material about bioinformatics, see the web page for this book, which is listed in the "Comments and Questions" section.

Conventions Used in This Book

The following conventions are used in this book:

Italic
: Used for commands, filenames, directory names, variables, URLs, and for the first use of a term

`Constant width`
: Used in code examples and to show the output of commands

`Constant width italic`
: Used in "Usage" phrases to denote variables.

The owl icon designates a note, which is an important aside to the nearby text.

The turkey icon designates a warning relating to the nearby text.

Comments and Questions

Please address comments and questions concerning this book to the publisher:

O'Reilly & Associates, Inc.
101 Morris Street
Sebastopol, CA 95472
(800) 998-9938 (in the United States or Canada)
(707) 829-0515 (international or local)
(707) 829-0104 (fax)

We have a web page for this book, where we list errata, examples, or any additional information. You can access this page at:

http://www.oreilly.com/catalog/bioskills/

To comment or ask technical questions about this book, send email to:

bookquestions@oreilly.com

For more information about our books, conferences, software, Resource Centers, and the O'Reilly Network, see our web site at:

http://www.oreilly.com

Acknowledgments

From Cynthia: I'd like to thank all of the people who have restrained themselves from laughing when they heard me say, for the thousandth time during the last year, "We're almost finished with the book." Thanks to my family and friends, for putting up with extremely infrequent phone calls and updates during the last few months; the students in my Fall 2000 Bioinformatics course, for acting as guinea pigs in my first bioinformatics teaching experiment and helping me identify topics

that needed to be explained more thoroughly; my colleagues at Virginia Tech, for a year's worth of interesting discussions of what bioinformatics means and what bioinformatics students need to know; and our friend and colleague Jim Fenton for his contributions early in the development of the book; and my thesis advisor Shankar Subramaniam. I'd also like to thank our technical reviewers, Sean Eddy, Peter Leopold, Andrew Odewahn, Clay Shirky, and Jim Tisdall, for their helpful comments and excellent advice. And finally, thanks goes to the staff of O'Reilly, and our editor, Lorrie LeJeune, for infinite patience and moral support during the writing process.

From Per: First, I am deeply grateful to my advisor, Professor Shankar Subramaniam, who has been a continuous source of inspiration and a mainstay of our lab's congenial working environment at UCSD. My thanks also go to two of my mentors, Professor Charles Elkan of the University of California, San Diego, and Professor Michael R. Brent, now of Washington University, whose wise guidance has shaped my understanding of computational problems. Sanna Herrgard and Markus Herrgard read early versions of this book and provided valuable comments and moral support. The book has also benefited from feedback and helpful conversations with Ewan Birney, Phil Bourne, Jim Fenton, Mike Farnum, Brian Saunders, and Winny Tan. Thanks to Joe Johnston of O'Reilly for providing Perl advice and code in Chapter 12. Our technical reviewers made indispensable suggestions and contributions, and I owe special thanks to Sean Eddy, Peter Leopold, Andrew Odewahn, Clay Shirky, and Jim Tisdall for their careful attention to detail. It has been a pleasure to work with the staff at O'Reilly, and in particular with our editor Lorrie LeJeune, who patiently and cheerfully guided us through the project. Finally, my part of this book would not have been possible without the support and encouragement of my family.

Table of Contents

I

Introduction

- Chapter 1, *Biology in the Computer Age*
- Chapter 2, *Computational Approaches to Biological Questions*

1

Biology in the Computer Age

From the interaction of species and populations, to the function of tissues and cells within an individual organism, biology is defined as the study of living things. In the course of that study, biologists collect and interpret data. Now, at the beginning of the 21st century, we use sophisticated laboratory technology that allows us to collect data faster than we can interpret it. We have vast volumes of DNA sequence data at our fingertips. But how do we figure out which parts of that DNA control the various chemical processes of life? We know the function and structure of some proteins, but how do we determine the function of new proteins? And how do we predict what a protein will look like, based on knowledge of its sequence? We understand the relatively simple code that translates DNA into protein. But how do we find meaningful new words in the code and add them to the DNA-protein dictionary?

Bioinformatics is the science of using information to understand biology; it's the tool we can use to help us answer these questions and many others like them. Unfortunately, with all the hype about mapping the human genome, bioinformatics has achieved buzzword status; the term is being used in a number of ways, depending on who is using it. Strictly speaking, bioinformatics is a subset of the larger field of *computational biology*, the application of quantitative analytical techniques in modeling biological systems. In this book, we stray from bioinformatics into computational biology and back again. The distinctions between the two aren't important for our purpose here, which is to cover a range of tools and techniques we believe are critical for molecular biologists who want to understand and apply the basic computational tools that are available today.

The field of bioinformatics relies heavily on work by experts in statistical methods and pattern recognition. Researchers come to bioinformatics from many fields, including mathematics, computer science, and linguistics. Unfortunately, biology is