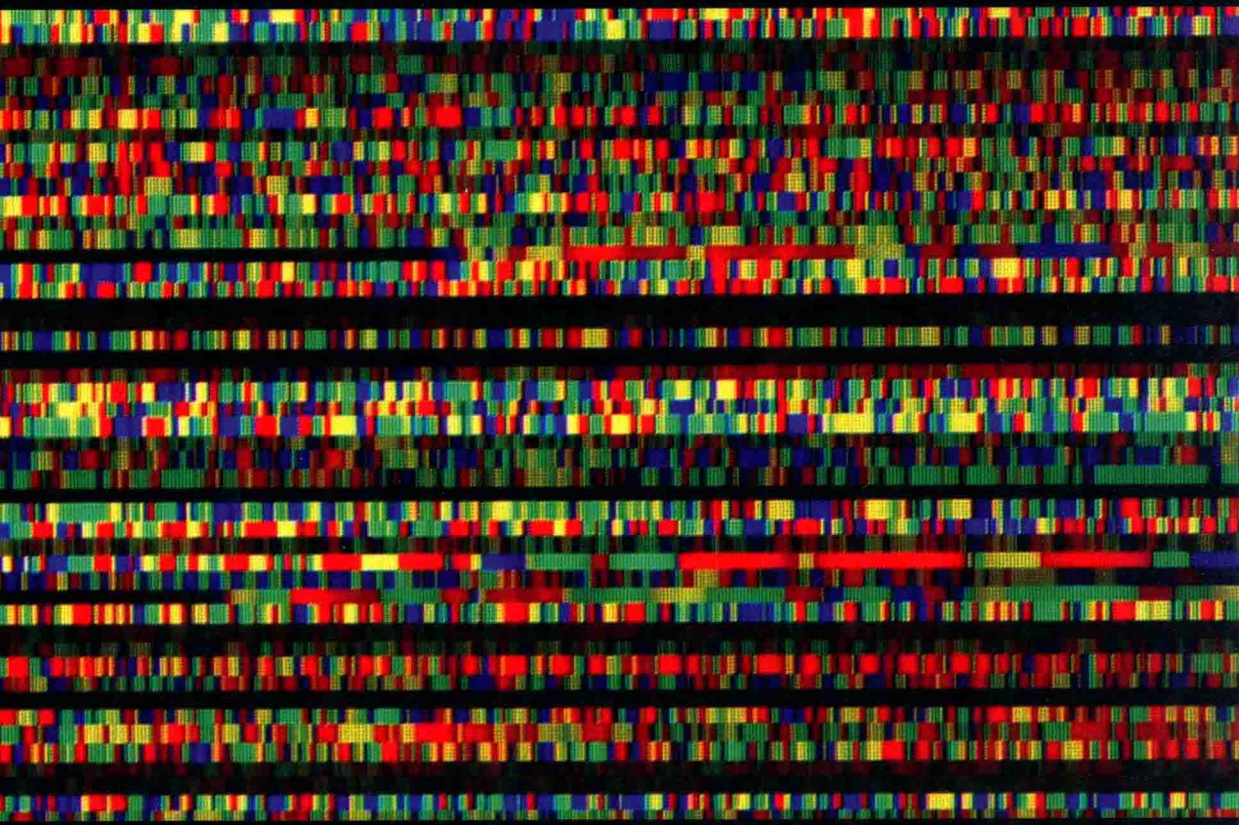




# Exploring Bioinformatics

A PROJECT-BASED APPROACH



Caroline St. Clair | Jonathan Visick

# Exploring Bioinformatics

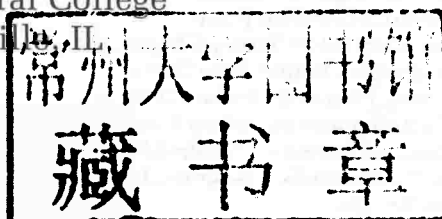
## A Project-Based Approach

**Caroline St. Clair, PhD**

Department of Computer Science  
North Central College  
Naperville, IL

**Jonathan Visick, PhD**

Department of Biology  
North Central College  
Naperville, IL



**JONES AND BARTLETT PUBLISHERS**

*Sudbury, Massachusetts*

BOSTON

TORONTO

LONDON

SINGAPORE

### *World Headquarters*

Jones and Bartlett Publishers  
40 Tall Pine Drive  
Sudbury, MA 01776  
978-443-5000  
info@jbpub.com  
www.jbpub.com

Jones and Bartlett Publishers  
Canada  
6339 Ormindale Way  
Mississauga, Ontario L5V 1J2  
Canada

Jones and Bartlett Publishers  
International  
Barb House, Barb Mews  
London W6 7PA  
United Kingdom

Jones and Bartlett's books and products are available through most bookstores and online booksellers. To contact Jones and Bartlett Publishers directly, call 800-832-0034, fax 978-443-8000, or visit our website, [www.jbpub.com](http://www.jbpub.com).

Substantial discounts on bulk quantities of Jones and Bartlett's publications are available to corporations, professional associations, and other qualified organizations. For details and specific discount information, contact the special sales department at Jones and Bartlett via the above contact information or send an email to [specialsales@jbpub.com](mailto:specialsales@jbpub.com).

Copyright © 2010 by Jones and Bartlett Publishers, LLC

All rights reserved. No part of the material protected by this copyright may be reproduced or utilized in any form, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without written permission from the copyright owner.

### **Production Credits**

Publisher, Higher Education: Cathleen Sether  
Production Director: Amy Rose  
Acquisitions Editor: Molly Steinbach  
Managing Editor: Dean W. DeChambeau  
Associate Editor: Megan R. Turner  
Editorial Assistant: Caroline Perry  
Senior Production Editor: Tracey Chapman  
Associate Production Editor: Kate Stein  
Senior Marketing Manager: Andrea DeFronzo  
V. P., Manufacturing and Inventory Control: Therese Connell  
Manufacturing and Inventory Control Supervisor: Amy Bacus  
Composition: Northeast Compositors, Inc.  
Art: Accurate Art, Inc.  
Senior Photo Researcher and Photographer: Christine McKeen  
Cover Design: Scott Moden  
Cover Image: Courtesy of DOE Joint Genome Institute  
Printing and Binding: Malloy, Inc.  
Cover Printing: Malloy, Inc.

### **Library of Congress Cataloging-in-Publication Data**

St. Clair, Caroline.

Exploring bioinformatics : a project-based approach / Caroline St. Clair and Jonathan Visick.  
p. cm.

ISBN-13: 978-0-7637-5829-5 (pbk.)

ISBN-10: 0-7637-5829-9 (pbk.)

1. Bioinformatics. I. Visick, Jonathan. II. Title.

QH324.2.S72 2009

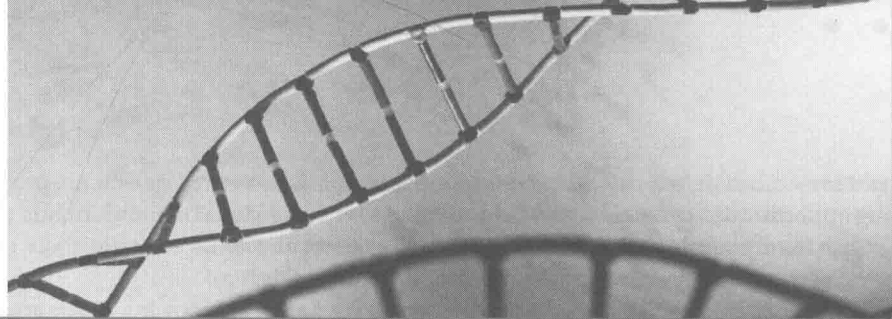
572.80285—dc22

2008040814

6048

Printed in the United States of America

13 12 11 10 09 10 9 8 7 6 5 4 3 2 1



# Preface

---

## ● Why We Wrote This Book

---

Bioinformatics is a rapidly growing field of increasing importance to both biologists and computer scientists. Although DNA sequence data, and later whole genome sequence data, obtained by molecular geneticists served as the driving force in the rapid development of bioinformatics, molecular biology is far from the only field in which these techniques are now employed. Ecologists, zoologists, botanists, evolutionary biologists, physiologists, developmental biologists, biochemists, and others are using bioinformatics to manipulate DNA and protein data to gain new insights. Bioinformatics has also found a wealth of applications in medicine, pharmacology, agriculture, and many other industries. With so much interest in the use of computers to sift through complex and voluminous biological data, it is a sure bet that biologists trained in computational analysis and computer scientists with a biological background will be in demand for years to come.

We sought to develop a course that would introduce both biologists and computer scientists to bioinformatics. Although we wanted our course to explore key concepts in some depth and to delve into the algorithms that make bioinformatic programs tick, we also wanted to keep it on an introductory level: A first course in bioinformatics for computer science students who have had an introductory course in molecular biology, or for biology or biochemistry majors who have had an introductory programming course. We wanted the course to be hands-on, investigative,

and research-rich, exposing students to current and unanswered questions, providing opportunities to tackle real-world problems with the kinds of bioinformatic programs used by actual researchers, and giving them the tools to write their own programs or modify existing ones to meet specific needs.

This book developed from that course. In each chapter, we cover a major area in which bioinformatics provides the means of attacking biological problems of current relevance. The topics are presented as a series of projects where students can learn both biological concepts and computational techniques as they explore questions using data and software freely available on the World Wide Web and then move on to an understanding of the underlying algorithms and finally development of their own code. This book is not intended to be a comprehensive reference on either molecular biology or computer programming; instead, we have endeavored to provide “just enough” instruction in each chapter so that students from both sides of the spectrum can master the ideas they need to investigate bioinformatic solutions to important problems. The exercises and questions found throughout the book are intended not merely to test students’ knowledge but to *develop* it, as it is our experience that the best learning takes place when students are actively engaged in finding their own solutions.

This book is adaptable for use in different kinds of courses. We approach our own course from a combination of biological and computational perspectives, and our students learn both the use of web-based tools and Perl programming. A biologist with more limited programming background might choose to have students explore how the algorithms work in the *Guided Project* sections but not expect them to write their own programs; complete Perl programs downloaded from the companion website could be used in a course of this nature to complete many of the exercises (identified by a disk icon) and to see how the algorithms work in practice. A mathematically oriented course might take a similar approach, emphasizing the mathematics underlying the algorithms without concern for their implementation. We also hope that we have provided “just enough” background material that a biologist with a real interest in programming or a computer scientist with a serious interest in biology could use the text effectively even in the absence of strong foundational knowledge in the complementary area.

We welcome your comments and ideas as you explore bioinformatics.

---

## ● How to Use This Book

---

The goal of this book is to engage you in learning bioinformatics by actually *using* bioinformatic tools—both existing web-based programs and your own software solutions—to approach real-world problems. This section will help you get acquainted with the layout of the book and the various resources it provides.

The first chapter serves as a general introduction to bioinformatics and its potential, and to the idea of biological problem-solving using computational techniques. Each of the remaining chapters introduces a current biological problem; genetic dis-



ease, antibiotic resistance, drug design, and so on. The *Understanding the Problem* sections in each chapter will help you recognize the importance of the problem to today's life scientists and the limitations of traditional techniques in exploring the topic. The next section, *Bioinformatics Solutions*, discusses the value of computational approaches in shedding light on the problem.

With this relatively brief introduction to the problem, two kinds of hands-on projects give you the opportunity to try using and writing relevant bioinformatic tools. First, a *Guided Project* will take you step-by-step through a structured approach to bioinformatic tools and algorithms. This project is divided into two parts: In the *Web Exploration* section, you will encounter data sources and analytical programs commonly used by biological researchers and freely available on the World Wide Web, while the *Programming Project* lets you look "under the hood" as bioinformatic algorithms implemented in the Perl programming language are presented, dissected, and used to explore data. The experience gained from the *Guided Project* will enable you to tackle the *On-Your-Own Project*, where programming concepts relevant to investigating a related but distinct problem are presented, but no actual program code is provided, allowing you to develop your own algorithms or implementations.

This book is deliberately not comprehensive or encyclopedic in its approach; instead, it emphasizes a limited number of key bioinformatic principles and encourages active learning through hands-on exploration. For students with less coursework in biological science, *BioBackground* sections provide additional background for understanding the projects. The programs in this text are written in Perl, an easy-to-use language with excellent text-processing features suitable to the manipulation of DNA and protein data. *Perl: Need to Know* sections throughout the text introduce Perl syntax, commands, and functions, while two appendices assist the less-experienced programmer in mastering the use of this language. In addition, *More to Explore* sections suggest possibilities for further study and *Connections* sections look forward to the future of bioinformatics.

Three kinds of exercises are included in each chapter to allow you to test and apply your knowledge and to provide opportunities for instructors to evaluate their students' performance. *BioConcept Questions* help you check your understanding of the biological ideas underlying the chapter's problem and determine whether you need to use the *BioBackground* sections to fill gaps in your knowledge. *Web Exploration Questions* ask you to develop your skills in using existing bioinformatic tools. *Putting Your Skills into Practice* exercises allow you to write your own Perl code to modify and extend the utility of the programs given in the *Guided Project* and then see how your programs work in dealing with real-world data. You will notice that all of the exercises in each chapter are numbered consecutively (even though there may be two or three groups of exercises) to make it easier for instructors to assign specific exercises from throughout the chapter.

Finally, a companion website, <http://biology.jbpub.com/bioinformatics>, provides additional resources. All of the code discussed in the *Guided Projects* can be downloaded to eliminate the need for retyping. DNA and protein sequences needed

to complete the projects and exercises are also found on the website, as are test data. For instructors, complete solutions to the programming exercises can be downloaded from <http://www.jbpub.com/biology/bioinformatics>, facilitating use of the software in courses that do not require programming.

---

● **Acknowledgments**

---

We would like to express our sincere thanks to the students of BIO/CSC 242 for their helpful suggestions. We also wish to thank North Central College for supporting this project. Finally, we want to thank our families for their encouragement throughout this process.

## **Jones and Bartlett Titles in Bioinformatics**

*Biomedical Informatics: A Data User's Guide*

Jules J. Berman

*Medical Informatics 20/20: Quality and Electronic Health Records  
through Collaboration, Open Solutions, and Innovation*

Douglas Goldstein, Peter J. Groen, Suniti Ponkshe, and Marc Wine

*Perl Programming for Medicine and Biology*

Jules J. Berman

*Python for Bioinformatics*

Jason Kinser

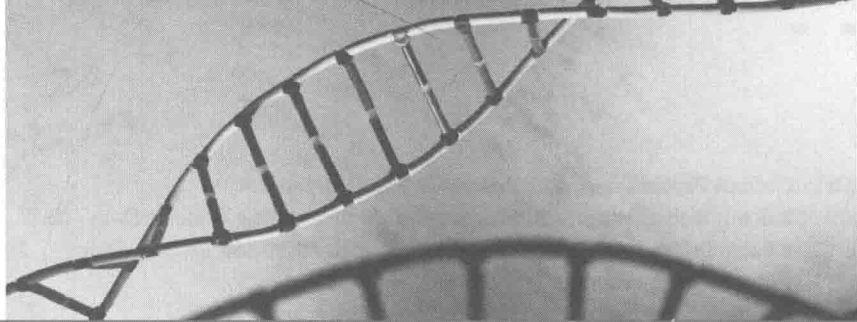
*R for Medicine and Biology*

Paul D. Lewis

*Ruby Programming for Medicine and Biology*

Jules J. Berman





# Contents

**Preface ix**

**How to Use This Book x**

**Acknowledgments xii**

## **Chapter 1 Exploring Bioinformatics 1**

- 1.1 Understanding the Problem: The Promise of Bioinformatics 2
- 1.2 Bioinformatics Solutions: Computational Approaches to Biological Problems 5
- 1.3 Web Exploration: Nova Scotia Niemann-Pick Disease 7
- 1.4 Programming Project: Developing an Algorithm 9

## **Chapter 2 The Central Dogma: Exploring Genetic Disease 15**

- 2.1 Understanding the Problem: Identifying Mutations 16
  - 2.1.1 BioConcept Questions 17
- 2.2 Bioinformatics Solutions: Computational Approaches to Genes 18
  - 2.2.1 Objectives 19
- 2.3 BioBackground: Will Mary's Child Have Sickle-Cell Disease? 19
- 2.4 BioBackground: The Role of Genes 21

- 2.5 Guided Project: Genetic Disease 25
  - 2.5.1 Web Exploration: Finding and Manipulating Genetic Data 25
  - 2.5.2 Programming Project: Working with DNA and Protein Strings 29
  - Programming Exercise 1: Transcription 30
  - Programming Exercise 2: Translation 34
  - Programming Exercise 3: Comparing Equal-Length Strings 38
  - 2.5.3 Perl: Need to Know 44
- 2.6 On-Your-Own Project: Gene Comparisons and Cystic Fibrosis 46
  - 2.6.1 Understanding the Problem 46
  - 2.6.2 Programming the Solution 50
  - 2.6.3 Running the Program 50
- 2.7 Connections: The Future of Genetics 51

### **Chapter 3 Gene Alignments: Investigating Antibiotic Resistance 53**

- 3.1 Understanding the Problem: Antibiotic Resistance 54
  - 3.1.1 BioConcept Questions 55
- 3.2 Bioinformatics Solutions: Aligning Genes 55
  - 3.2.1 Objectives 56
- 3.3 BioBackground: Antibiotic Resistance 57
- 3.4 BioBackground: Gene Transfer and Similarity 59
- 3.5 Guided Project: Antibiotic Resistance 60
  - 3.5.1 Web Exploration: Evidence for Horizontal Gene Transfer 60
  - 3.5.2 Programming Project: Global Alignment 67
  - Programming Part 1: Building the Scoring Matrix 74
  - Programming Part 2: Creating Directional Strings 76
  - Programming Part 3: Creating Optimal Alignments 78
  - 3.5.3 Perl: Need to Know 81
- 3.6 On-Your-Own Project: Local Alignment Algorithms 82
  - 3.6.1 Understanding the Problem: Local Alignments 83
  - 3.6.2 Programming the Solution 84
  - 3.6.3 Running the Program 85
- 3.7 Connections: Fighting Resistance 85

### **Chapter 4 Protein Alignments: Clues to Protein Function 87**

- 4.1 Understanding the Problem: Protein Function 88
  - 4.1.1 The Problem of Protein Function in Cystic Fibrosis 88
  - 4.1.2 BioConcept Questions 90
- 4.2 Bioinformatics Solutions: Protein Alignments 90
  - 4.2.1 Objectives 92
- 4.3 BioBackground: Protein Structure and Function 92
- 4.4 BioBackground: Protein Similarity 93

- 4.5 Guided Project: Protein Alignment and Protein Function 96
  - 4.5.1 Web Exploration: The Function of the CFTR Protein 96
  - 4.5.2 Programming Project: Protein Alignment with Substitution Matrices 102
  - Programming Exercise 1: Aligning Proteins Using Hydrophobicity 102
  - Programming Exercise 2: Aligning Proteins Using a Substitution Matrix 107
  - 4.5.3 Perl: Need to Know 112
- 4.6 On-Your-Own Project: Building a Substitution Matrix 114
  - 4.6.1 Understanding the Problem 115
  - 4.6.2 Programming the Solution 117
  - 4.6.3 Running the Program 118
- 4.7 Connections: Annotating Genomes 119

## **Chapter 5 Sequence Assembly: The Human Genome 121**

- 5.1 Understanding the Problem: Assembling the Human Genome 122
  - 5.1.1 Shotgun Sequencing and Assembly 123
  - 5.1.2 BioConcept Questions 124
- 5.2 Bioinformatics Solutions: Sequence Assembly 124
  - 5.2.1 Objectives 125
- 5.3 BioBackground: Sequencing DNA 126
- 5.4 BioBackground: Sequencing Complex Genomes 128
- 5.5 Guided Project: Shotgun Sequencing and Assembly 129
  - 5.5.1 Web Exploration: Reading and Assembling DNA Sequence Data 129
  - 5.5.2 Programming Project: Sequence Assembly 135
  - Programming Exercise 1: Generating Fragments 136
  - Programming Exercise 2: Determining Overlap 140
  - Programming Exercise 3: Sequence Validation 144
  - 5.5.3 Perl: Need to Know 150
- 5.6 On-Your-Own Project: A Mini-Assembly Program 151
  - 5.6.1 Understanding the Problem 151
  - 5.6.2 Programming the Solution 154
  - 5.6.3 Running the Program 156
- 5.7 Connections: The Future of Genome Sequencing 156

## **Chapter 6 Gene Prediction: Finding Genes in the Human Genome 159**

- 6.1 Understanding the Problem: Gene Discovery 159
  - 6.1.1 BioConcept Questions 162
- 6.2 Bioinformatics Solutions: Gene Prediction 162
  - 6.2.1 Objectives 164
- 6.3 BioBackground: ORFs and Genes 164
- 6.4 BioBackground: Coming to a Consensus 167

- 6.5 Guided Project: Gene Discovery in Prokaryotes 168
  - 6.5.1 Web Exploration: Tools for Exploring Prokaryotic Genes 168
  - 6.5.2 Programming Project: Sequence-Based Gene Prediction in Prokaryotes 174
  - Programming Exercise 1: Finding Long ORFs 175
  - Programming Exercise 2: Finding Transcriptional Units 181
  - 6.5.3 Perl: Need to Know 186
- 6.6 On-Your-Own Project: Gene Discovery in Eukaryotes 189
  - 6.6.1 Sequence-Based Pattern Matching in Eukaryotes 189
  - 6.6.2 Content-Based Gene Prediction in Eukaryotes 191
- 6.7 Connections: The Ongoing Need for Gene Discovery 194

## Chapter 7 RNA and Protein Structure: Structure Prediction 197

- 7.1 Understanding the Problem: Structure Prediction 198
  - 7.1.1 BioConcept Questions 201
- 7.2 Bioinformatics Solutions: RNA and Protein Structure 202
  - 7.2.1 Objectives 204
- 7.3 BioBackground: Protein Structure 204
- 7.4 BioBackground: Nucleic Acid Folding 208
- 7.5 Guided Project: Protein Structure Prediction 209
  - 7.5.1 Web Exploration: Protein Structure Modeling and Drug Design 209
  - 7.5.2 Programming Project: Structure Prediction with the Chou-Fasman Algorithm 216
- 7.6 On-Your-Own Project: RNA Structure Prediction with the Nussinov-Jacobson Algorithm 222
  - 7.6.1 Understanding the Problem 222
  - 7.6.2 Programming the Solution 224
  - 7.6.3 Running the Program 227
  - 7.6.4 Perl: Need to Know 228
- 7.7 Connections: Modeling Protein Folding 229

## Chapter 8 Molecular Phylogenetics: Measuring Evolution 231

- 8.1 Understanding the Problem: Molecular Evolution 232
  - 8.1.1 BioConcept Questions 236
- 8.2 Bioinformatics Solutions: Molecular Phylogeny 236
  - 8.2.1 Objectives 239
- 8.3 BioBackground: Measuring Evolution 239
- 8.4 BioBackground: Substitution Rates 241
- 8.5 Guided Project: Molecular Evolution and Phylogenetic Trees 244
  - 8.5.1 Web Exploration: RecQ as a Molecular Clock 244
  - 8.5.2 Programming Project 1: Using Distance Metrics to Measure Similarity 249
  - Programming Exercise 1: Jukes-Cantor Model 250
  - Programming Exercise 2: Kimura's Two-Parameter Substitution Model 251

- 8.5.3 Programming Project 2: Phylogenetic Trees Using Agglomerative Clustering 253
- 8.5.4 Perl: Need to Know 264
- 8.6 On-Your-Own Project: The Neighbor-Joining Method 265
  - 8.6.1 Understanding the Problem 266
  - 8.6.2 Programming the Solution 270
  - 8.6.3 Running the Program 271
- 8.7 Connections: What Is a Species? 271

## **Chapter 9 Microarrays and Gene Expression 273**

- 9.1 Understanding the Problem: Gene Expression 274
  - 9.1.1 BioConcept Questions 275
- 9.2 Bioinformatics Solutions: Microarrays 276
  - 9.2.1 Objectives 280
- 9.3 Guided Project: Analyzing Microarray Data 280
  - 9.3.1 Web Exploration: Using MAGIC Tool to Analyze Two-Channel Microarray Data 280
- 9.4 On-Your-Own Project: Analyzing Data from Phytoremediation Experiments 292
  - 9.4.1 Understanding the Problem: Preprocessing 294
  - 9.4.2 Programming the Solution: Preprocessing 298
  - 9.4.3 Running the Program: Preprocessing 300
  - 9.4.4 Understanding the Problem: Analysis of Gene Function 301
  - 9.4.5 Programming the Solution: Analysis of Gene Function 301
  - 9.4.6 Running the Program: Analysis of Gene Function 301
- 9.5 Connections 302

## **Appendix A Installing Perl and Running a Perl Program 305**

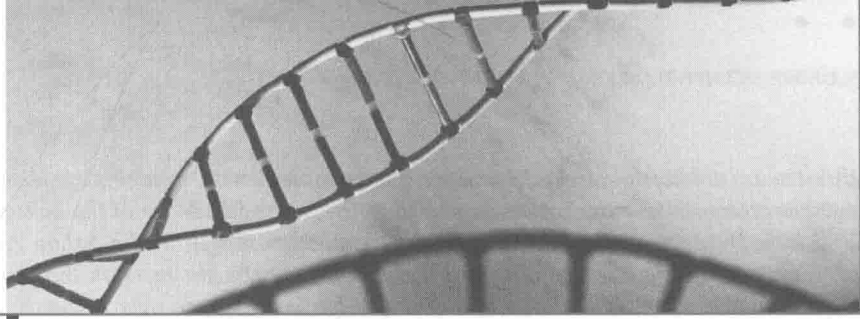
## **Appendix B Introduction to Programming and Perl Syntax 309**

### **Glossary 323**

### **References and Supplemental Reading 343**

### **Index 347**

### **Art Credits 359**

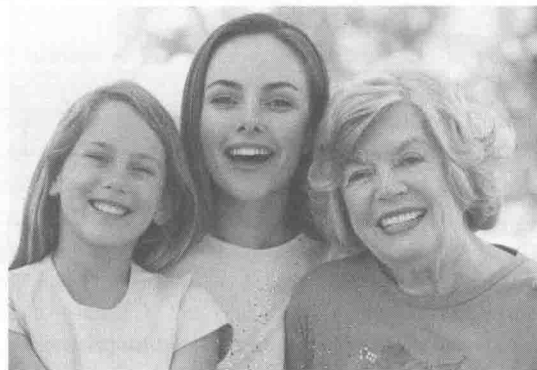


chapter

1

# Exploring Bioinformatics

*In a small community in western Nova Scotia, researchers discovered a rare genetic disease, now known as Nova Scotia Niemann-Pick disease (NS-NPD). This genetic disorder affects the nervous system and appears most commonly in children. A genealogical path was found connecting all known patients afflicted with this disease to a single seventeenth-century Acadian couple born and raised in western Nova Scotia. A single mutant gene is believed to be responsible for NS-NPD. It is presumed that inbreeding within the community resulted in the high frequency of*



**Figure 1.1** • The secrets of long life and health...are they in the genes?

this disease among descendants of the original couple. Today, residents are encouraged to marry outside the community to limit the future incidence of the disease.

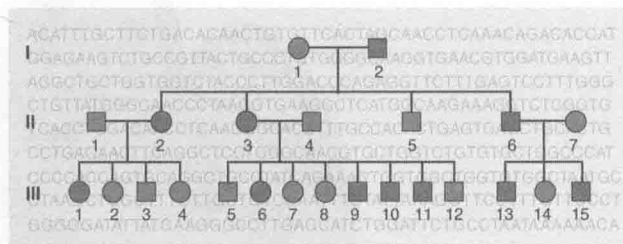
Across the ocean in Italy, Stoccareddo, a small town high in the Italian Alps, also has a high rate of inbreeding and has sparked geneticists' interest. But in this case, residents of the town are amazingly healthy. Even though over 97 percent of the residents are closely related, the rate of disease is extremely low and genetic disease is virtually nonexistent. Here, residents are being encouraged to marry inside the community, and researchers are seeking to identify specific genes that contribute to the extraordinary well-being of this population.

What happened over the 800-year history of Stoccareddo that left its modern residents so healthy? What "bad" genes have been eradicated by generations of natural selection? What "good" genes contribute to the startling good health of the community? Could modern technology bring similar benefits to others? And what of NS-NPD? Will the disease gene be "weeded out" over time? What can we do for those who have already inherited the disease?

## 1.1 Understanding the Problem: The Promise of Bioinformatics

In the late nineteenth century, Austrian monk Gregor Mendel quietly initiated a genetic revolution by providing science with its first tools for understanding how characteristics are inherited. The identification of **DNA** (deoxyribonucleic acid) as the genetic material in 1953 sent the revolution in a new direction, allowing researchers to directly correlate changes in genes with the workings of cells and organisms in health and disease. But the floodgates of genetic information truly opened in the mid-1990s, as large-scale DNA sequencing technology became capable of revealing the complete nucleotide sequence of entire genomes (see Figure 1.2), including the three billion base-pairs that make up the human genome (completed in 2003).

Today, molecular geneticists seeking to extract the secrets that lie within cells are awash in a vast and ever-increasing flood of genetic information. They have uncov-



**Figure 1.2 • Bioinformatics**—a field that has grown out of the information explosion resulting from the use of molecular tools to approach genetic problems.



ered the specific mutations that underlie sickle-cell anemia, cystic fibrosis, Tay-Sachs disease, and a host of other inherited diseases, and they have used that information for new therapeutic strategies. Scientists have developed corn and cotton that make their own insecticides and drugs made from proteins released in the milk of genetically engineered goats. Both the scientific literature and the popular press promise new and exciting innovations and applications, ranging from understanding the physiology of cells and organisms to curing genetic disease, improving food supplies, reversing environmental degradation, and providing individualized medicine. Truly, we live in the age of genetics.

Unfortunately, even with our tremendous technological advances and all the data we have amassed, the discovery process is slow, and many fascinating and important problems will undoubtedly remain unsolved for years to come. Part of the problem is finding ways to get meaningful information from the mountains of data available. Because of the enormity of genetic data, new analytical approaches must be developed in order to uncover new nuggets of information. **Bioinformatics** is the new science at the interface of molecular biology and computer science seeking to develop better ways to explore, analyze, and understand genetic data.

The goal of this book is to introduce several “hot topics” within the young field of bioinformatics and provide experience with key techniques used to investigate questions within those areas. Our project-based approach is intended to engage you in actually using and developing these techniques as you apply them to questions relevant to real-world research. As you work through the problems and exercises in each chapter, you will use existing web-based software, explore how bioinformatic algorithms work, and write your own software solutions in Perl.

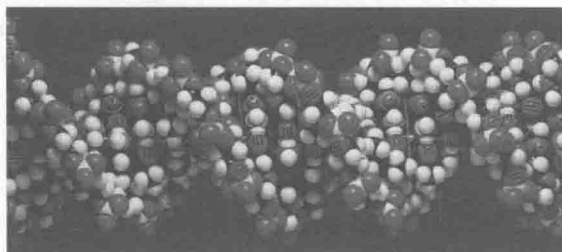
What kinds of questions and problems are investigated using bioinformatics? Bioinformatic tools are being used in both basic research and practical applications in fields as diverse as ecology, cancer biology, agriculture, pharmacology, and forensics. Here are just a few examples of biological questions whose solutions will include a key role for computer technology:

- Why don't all tumors of the same type respond to the same chemotherapies? Could we individualize treatment to increase the chances of successful therapy?
- What are the causes of the hundreds of genetic diseases? How do genes influence complex diseases such as diabetes, obesity, heart disease, cancer, and alcoholism? What can we do about these diseases?
- How should we define a species, and how can we determine whether two different organisms belong to the same species?
- In the human genome, how does a surprisingly small set of genes—probably fewer than 25,000—result in the synthesis of hundreds of thousands of proteins? And what functions remain to be discovered for the 99 percent of the genome that apparently does not encode proteins?

- How do bacteria and viruses cause disease, and how can better understanding of their physiology improve prevention and treatment?
- Can DNA evidence be relied upon in convicting—or exonerating—those accused of crimes?
- What is the risk that a child will inherit a genetic disease or susceptibility?
- What are the evolutionary pathways that have led to the development of the diversity of organisms that we see today?
- Are there organisms with genes that we could apply to breaking down oil spills or cleaning up toxic waste?
- How can drug design be improved so that specific drugs can be made to work on specific targets, rather than using a “shotgun” approach?
- What can the genes of bacteria, yeast, flies, worms, plants, and mice tell us about the functions of human genes?

We are able to ask most of these questions because of the advent of molecular biology, which got its start as scientists like Max Perutz and Linus Pauling began to dissect the structure of proteins, while others, including Frederic Griffith, Alfred Hershey, Martha Chase, James Watson, and Francis Crick found that DNA (see Figure 1.3) was the solution to the riddle of the genetic material. Initially, the goal of molecular biology was to understand the workings of genes and how they encode proteins. As the details of those processes became clearer, molecular biologists turned their attention to understanding key biological processes at the molecular level.

Today, the questions asked by molecular biologists span multiple areas of biology. Researchers using molecular biology techniques might actually be working in the field of development, trying to understand how genes are turned on and off in sequence to produce the changes that occur between the egg and the adult organism. Or, they might be comparing DNA sequences to study evolutionary relationships among species. A molecular biologist might work for a biotech company, altering the genetic makeup of a plant to improve its characteristics, or in a clinical setting seeking a permanent cure for a genetic disease. What unites these very different applications of molecular biology is their focus on DNA and the proteins it encodes. In addition, they all face the common problem of dealing with huge



**Figure 1.3** • DNA—the genetic material of all living things.