discovering

GENOMICS, PROTEOMICS, & BIOINFORMATICS

A. MALCOLM CAMPBELL & LAURIE J. HEYER

DISCOVERING GENOMICS, PROTEOMICS, AND BIOINFORMATICS

A. Malcolm Campbell
Davidson College

Laurie J. Heyer
Davidson College

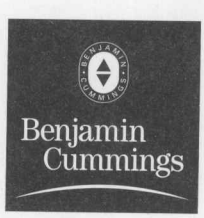
San Francisco Boston New York
Cape Town Hong Kong London Madrid Mexico City
Montreal Munich Paris Singapore Sydney Tokyo Toronto

Senior Project Manager: Peggy Williams
Publishing Assistant: Michael McArdle
Production Editors: Larry Lazopoulos, Jamie Sue Brooks
Production and Composition Services: The Left Coast Group
Text Designer: Andrew Ogus
Cover Designer: Yvo Reizebos
Copyeditor: Gary Morris
Illustrations: Steve McEntee and Mary Ann Tenorio
Proofreader: Martha Ghent
Photo Research: Kathleen Olson, Maureen Spuhler
Manufacturing Supervisor: Vivian McDougal
Marketing Manager: Josh Frost
Cover Printer: Phoenix Color
Printer: Courier Book Companies

Library of Congress Cataloging-in-Publication Data
Campbell, A. Malcolm
Discovering genomics, proteomics, and bioinformatics / A. Malcolm Campbell, Laurie J. Heyer.
p. cm.
Includes index.
ISBN 0-8053-4722-4
1. Genomics. 2. Proteomics. 3. Bioinformatics. I. Heyer, Laurie, J. II. Title.

QH447 .C35 2002
572.8'6—dc21 2002067456

Text, photography, and illustration credits appear following the Glossary.



ISBN 0-8053-4722-4

2 3 4 5 6 7 8 9 10 — CRW — 06 05 04 03

www.aw.com/bc

FOREWORD

Stan Fields
University of Washington

New fields of biology often open up when new tools become available to analyze living creatures and their constituent parts. Developments in microscopy and cell fractionation led to the rise of cell biology; the availability of X-ray sources and later Nuclear Magnetic Resonance machines brought about the three-dimensional images of structural biology, and recombinant DNA technologies advanced the already spreading techniques of molecular biology. As these disciplines matured, their approaches began to affect the thinking that surrounds major questions in biology; at the same time, these new branches retained some of their distinctive styles and adherents.

The advent of the Human Genome Project captured the imagination of both scientists and the general public. On 26 June 2000, President Bill Clinton, at a press conference with the heads of the National Human Genome Research Institute and Celera Genomics Corporation, declared "We are here to celebrate the completion of the first survey of the entire human genome. Without a doubt, this is the most important, most wondrous map ever produced by humankind." For biologists, the Human Genome Project meant the arrival of billions of base pairs of DNA sequence, not just of humans but of model organisms such as *E. coli,* yeast, the nematode worm, the fruit fly, and the mustard weed. In conjunction with these sequences came a flurry of technologies to interpret the DNA sequence, the RNAs that are copied from it, and the proteins that are synthesized from these RNAs.

Discovering Genomics, Proteomics, and Bioinformatics gives students a detailed view of the revolutionary methodologies of the last few years and their impact on biological thinking. It provides a lucid explanation of technologies used for such tasks as DNA sequencing, detecting polymorphisms, arraying genes, making gene knockouts, and identifying protein interactions. The application of these technologies to specific biological problems provides a model for how to approach many other questions in biology—from development to human disease. The text also illustrates how the acquisition of large data sets has led to the generation of networks of genes and proteins, as well as models of cellular behavior. A series of *Discovery Questions* in each chapter guides the reader through scientific questions as well as ethical, legal, and social issues.

The disciplines of genomics and proteomics cannot be uncoupled from bioinformatics, the tools to handle and analyze the prodigious amounts of data that continue to emerge from large-scale DNA, RNA, and protein projects. The book draws on numerous databases and computational algorithms to make this coupling clear. The inclusion of *Math Minutes* provides the statistical and probability calculations that add rigor to the interpretations of these data sets.

Have these powerful approaches led to a new way of thinking? At one level, the answer is clearly "yes," as syntheses of genomic and proteomic data make possible inferences not previously apparent. Yet some might argue that biologists have always studied "systems biology," and what has changed is not the thinking but the experimental tools. When it was possible to detect only a single transcript, an RNA might be analyzed under various parameters to measure single gene responses. Now that it is feasible to employ DNA microarrays, scientists can detect virtually all the transcripts of an organism. Perhaps students beginning their studies now will enter a discipline in which these experimental and computational tools will constitute not "systems biology" but simply "biology."

PREFACE

The term "genomics" was derived from the term "genome," which means the complete (haploid) DNA content of an organism; genomics is the field of genome studies. Once genome and genomics became popular terms, a flurry of new terms that ended in "-omes" and "-omics" began to appear in publications. *Discovering Genomics, Proteomics, and Bioinformatics* is more than a "tome of -omes" because the field has expanded beyond a narrow definition of genomics. Genomics, as presented in this book, includes the interaction of molecules inside cells, including DNA, protein, lipids, and carbohydrates. In the spirit of discovery, we will explore the tools and questions behind the revolution that is changing the way biology is studied.

Discovering Genomics, Proteomics, and Bioinformatics is based on two pedagogical principles that have been successful for many teachers; teach in the context of an interesting question and on a need-to-know basis. This is how everyone learns new information, and it is the best way to help students learn so they will be motivated and more likely to retain the information. *Discovering Genomics, Proteomics, and Bioinformatics* is built on "stories" or case studies taken from scientific publications. In answering the many questions raised by these case studies, we will explore scientific content and process. The content includes all the major areas such as sequences (whole genomes and variations), microarrays, and proteomics.

This book is designed as an interactive resource to use when exploring topics in genomics and proteomics. The figures provide real data that you can mine to extract more information than is initially apparent. The online databases engage you in real-time discoveries using the same databases investigators are using for their own research. Discovery Questions focus your attention on critical information and urge you to think for yourself using the tools and information presented in text and figures. Traditional textbooks supply you with facts and details that you are inclined to memorize for tests. Genomics requires you to analyze, hypothesize, think, and formulate models; this book was designed to develop your critical thinking skills.

Studying biology in the twenty-first century offers you unique opportunities. Therefore, this textbook must also look and "feel" different. You will need to use the computer a lot to access the latest information. To fully understand genomics, proteomics, and bioinformatics, you will read about real and compelling cases that challenge and encourage you to learn. So, immerse yourself in the case studies and discover what genomics is all about.

Writing Style

The text is written in a style that is easy to read and comprehend. It avoids unnecessary jargon, yet new terms are included when essential to help you understand the material.

Discovery Questions

The process of critical thinking is enhanced by Discovery Questions which are imbedded within the case studies rather than saved for the end of each chapter. Discovery Questions focus your attention on key concepts as well as experimental design, interpretation of data, and the need to support your

opinion with data. Analyzing real data reproduced from peer-reviewed publications will allow you to reach your own conclusions and the text will help guide you through the data. To answer some Discovery Questions (see the following examples), you will use online public databases, many of which are regularly updated.

DISCOVERY QUESTIONS

16. Go to *S. cerevisiae* Genomic View and find the location (base 1,236,754 of chromosome 4) of the NORF identified by this research. Find the annotated gene that overlaps with the insertion site.
17. Go to SGD Gene/Sequence Resource page, retrieve the nucleotides on chromosome 4 ranging from 1,236,454–1,237,054, and click on "6-Frame translation." The mTn insertion happened at base 300 of these 600 bases. How many potential proteins are located at this site?

All the Discovery Questions are available via the companion web site, to facilitate your interaction with online resources and permit you to submit your answers to your instructor via email.

Media Menu

Throughout every chapter are Media Menus to alert you to the resources contained on the book's companion web site. From this site, you can read in-depth descriptions of methods, access sequence information, view a 3D structure, and link to related web sites. These media tools allow you to participate in the interactive process of discovery that is at the core of genomics.

METHODS
QuickPDB
STRUCTURES
Cyclooxygenase
SEQUENCES
Uncharacterized Protein
LINKS
Conserved Domain
PDB
PREDATOR

Math Minutes

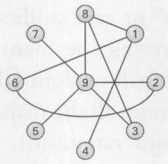
Most biologists in the cell/molecular field do not use much math in their work, but genomics, proteomics, and bioinformatics are changing this reality; these fields rely heavily on mathematics. To facilitate appreciation of how the data were analyzed and the role mathematics plays in understanding biology, we have produced Math Minutes as enrichment for those who want to discover the interaction of math and biology. Math Minutes (see sample below) use the case studies as foundations for concise lessons in statistical analysis, probability and computational methods.

MATH MINUTE 6.2 • IS Sup35 A CENTRAL PROTEIN IN THE NETWORK?

In Figure 6.16, Sup35 appears to be central to the network of interacting proteins. However, looks can be deceiving, particularly in such a complex network. We would like to quantify whether Sup35 interacts with an unusually large number of proteins, compared to other proteins in the network.

A mathematician would call the network of interacting proteins in Figure 6.16 a *graph*. In graph theory, the lines between nodes are called *arcs* or *edges*. A *directed* graph has arrows on the edges to indicate the direction of information flow between two nodes. In Figure 6.16 where the connections or relationships between nodes are important but there is no directional information, the arrows are left off and the graph is *undirected*. The number of edges touching a node is called the *degree* of the node. In graph theory terms, the question at hand is whether the Sup35 node has a significantly greater degree than is "typical" for this graph. If so, investigators are led to believe that Sup35 plays a central role in the function of the entire network of proteins.

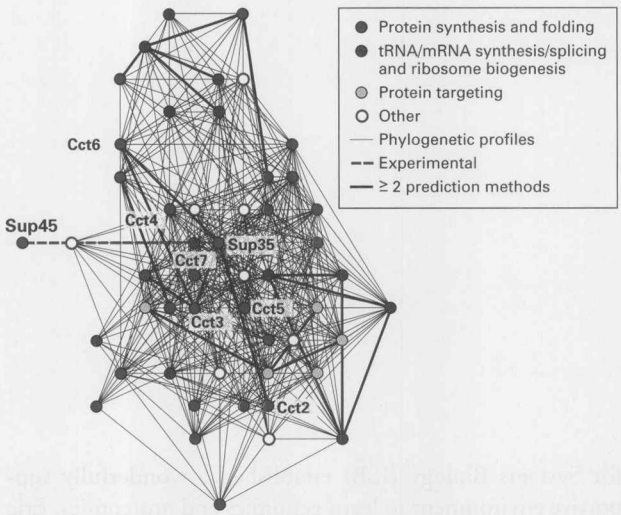
One way to approach questions about graphs like that in Figure 6.16 is to build a probabilistic model of the graph, called a *random graph*. Under this model, an edge is drawn between each pair of nodes with a certain probability, each edge independent of the others. Advanced probability theory and graph theory provide precise ways to determine the probability that the maximum degree in an arbitrarily large graph exceeds a given number. However, under the simplifying assumption that the degree numbers are approximately normally distributed (see Math Minute 8.1), you can evaluate whether the degree of the Sup35 node is significantly greater than expected using the mean and standard deviation of the degree numbers. Since Figure 6.16 is too complex to illustrate this process, consider the following smaller network:



In this graph, nodes 4, 5, and 7 have degree 1; nodes 2, 3, and 6 have degree 2; nodes 1 and 8 have degree 3; and node 9 has degree 5. The mean degree is the average of the nine degree numbers (2.22). The standard deviation of the nine degree numbers (1.23) is used to determine if the degree of node 9 is unusual. Specifically, if node 9's degree is more than two standard deviations larger than the mean degree, it can be considered unusually large. Since $5 > 4.68 = 2.22 + 2 \times 1.23$, we conclude that node 9 has an unusually large degree. This same procedure could be applied to the graph in Figure 6.16 to help quantify whether Sup35 plays a central role in this network of proteins.

Art Program

Detailed and abundant illustrations are reproductions of original data and expand on the basic information provided in the text, as shown in the example below. Your understanding will be enhanced by analysis of the figures from which you can extract additional information.



Figures on CD-ROM

All the illustrations in the text appear on the CD-ROM that accompanies every copy of the book. Some figures, which are best seen in full color and large format do not appear in the text but are available exclusively on the CD-ROM. These illustrations are designated in the text as follows:

FIGURE 4.19 • Stress-induced expression of 900 genes clustered into the ESR.
Go to the CD-ROM to view this figure.

These figures may be projected in class to enhance classroom lectures and discussions.

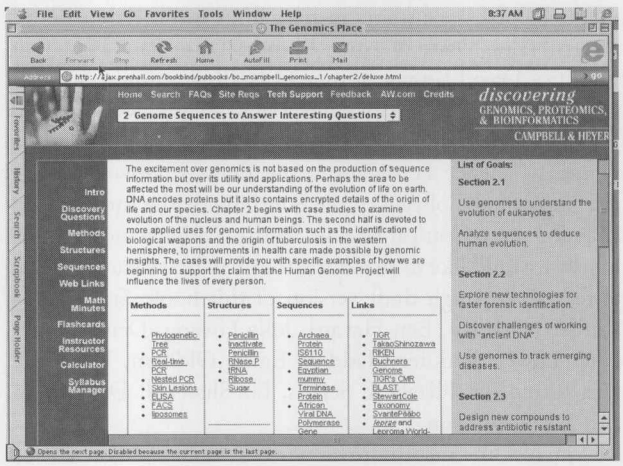
Transition from Genetics to Genomics

Typically, people think of genomics as high-volume genetics. However, a genome is more than the sum of its parts and you need to approach it with a new mindset. The last unit, Unit Four, contains three chapters to help ease your transition into genomics. Specifically, these three chapters confront the common misconception of "one gene, one protein, one phenotype." The case studies contain real data for you to interpret and discover the interconnectivity of the "cell web." The discovery approach to learning will foster scientific skills of analyzing data and formulating models to explain the data.

Media

The companion web site for *Discovering Genomics, Proteomics, and Bioinformatics* (www.geneticsplace.com) is a tool to enhance your study of genomics.

- *Methods:* These web pages explain how molecular and genomic methods are conducted and what type of data they produce. They are intended to supplement the textbook and provide background information if you have not learned them in previous courses.
- *Sequences:* In several Discovery Questions, you will use online bioinformatics tools to analyze protein or DNA sequences. To save you time and the potential problem of typos, all sequences are supplied in web pages (see the illustration below) that you can copy and paste for analysis.
- *Structures:* A significant aspect of any protein is its 3D structure. Periodically, web pages with chime tutorials have been created to illustrate structural features that are best understood when you interact with them.
- *Links:* Two types of links have been collected for each chapter. The first provides direct access to online databases and bioinformatics analysis tools such as the National Center for Biotechnology Information and the Protein Data Bank. The second facilitates easy access to investigators' laboratory web pages when you are particularly interested in a case study or area of research.



For the Instructor

Instructor's Guide ISBN 0-8053-4726-7

The printed version of the Instructor's Guide provides tips on presenting the material and answers to all the Discovery Questions. In addition to the written answers, the electronic version of the Instructor's Guide (www.geneticsplace.com) provides color figures to accompany some answers. For example, if students are asked to draw a graph or construct a circuit diagram, your electronic version will illustrate one possible answer. These illustrations, along with all illustrations on the CD-ROM, may be used for lecture presentation such as PowerPoint or web pages.

ACKNOWLEDGMENTS

Davidson College is home to many supportive people. The leadership and backing of Clark Ross, Bobby Vagt, and Verna Case provided the foundation for much of this work. Colleagues in many departments helped the authors in numerous ways: Karen Bernd, Karen Hales, Dave Wessner, Barbara Lom, Pam Hay, Julio Ramirez, David Brown, Jeanne O'Neill, Suzanne Churchill, Chris Paradise, Mur Muchane, Peggy Maiorano, and Betty Hartsell. The Waksman Foundation for Microbiology, the National Science Foundation, the Duke Endowment, and the Associated Colleges of the South made the funding possible for my sabbatical year in Seattle. Perhaps most importantly, I would like to thank the fourteen brave students who tested out a rough draft version of this book: Sean Burke, Amber Hartman, Ben Havard, Julie Hwang, Dennis Jones, Peter Lowry, Jennifer Madden, Emily Oldham, La Powell, Lisa Robinson, Elizabeth Sellars, Liz Shafer, J. D. Willson, and Marisa Wilson.

Much of the first draft was written while I was on sabbatical in Seattle. Mary Claire King, Maynard Olson, and Leroy Hood were very supportive hosts during this time. Several members of the University of Washington Genome Center provided technical and moral support. People at the Institute for Systems Biology (ISB) established a wonderfully supportive environment to learn genomics and proteomics. Eric Davidson's presentation at the ISB annual retreat was inspirational in many ways. And thanks to all the investigators around the world who freely shared their work and experience, and answered questions. It was heartwarming to be a part of a community where knowledge is shared and education valued. Personally, our friends and family in Seattle provided a wonderful home away from home, especially the Hope-Young family, the Hill family, Ginger Armbrust and Susan Francis, the Walker family, and the 20 families in the Meadowbrook Co-op Preschool.

Any major project requires a large number of people to create it. This book began as a vague idea and was quickly shaped into a reality thanks to the nurturing guidance of my friend Michele Sordi. Without her, it never would have come to fruition. Producing the actual book was largely due to the diligence and sweat equity of Peggy Williams. Her labors were critical to the successful invention of a new book in a new field. In addition, wisdom and editorial collaboration at Cold Spring Harbor Labratory Press was shepherded by John Inglis, who leads a substantial team of dedicated professionals. At Benjamin Cummings, many people played

substantial roles in shaping this book, including Steve McEntee and Mary Ann Tenorio whose artistic skills were able to make the figures true to their original form while also enhancing their pedagogical value. Jamie Sue Brooks and Larry Lazopoulos at Benjamin Cummings, and the staff at The Left Coast Group admirably shaped the production work. Thanks to Andrew Ogus for creating the look of the book, and to Yvo Rezebios for the cover design. Finally, Michael McArdle as publishing assistant kept everyone in touch with each other.

Finally, the most important support came from those closest to home. Almost overnight a wonderful collaboration was born between a mathematician who knows a lot of biology and a biologist who tries his best with math; thanks, Laurie. Members of the Genome Consortium for Active Teaching (GCAT) provided a collegial support system that encouraged me when I needed it. Susan, Paulina, and Celeste crossed the country twice and provided me with the stability needed to produce this book. My extended family also supplied needed support along the way without whom I would still be making floor tiles in the factory.

—A. Malcolm Campbell

Colleagues in the mathematics department, especially chair Stephen Davis and mentor Donna Molinek, offered endless encouragement and support to me. Our students are a continuing source of energy and inspiration. I am particularly indebted to Computational Biology students Frank Chemotti, Amber Hartman, Soren Johnson, Jennifer Kawwass, Peter Leese, Rachel Patton McCord, Cecilia Mendiondo, Emily Oldham, La Powell, Talbot Presley, Stanley Prybe, Lang Robertson, Lisa Robinson, and Megan Shafer. This interdisciplinary team produced first-class Kyte Doolittle and hierarchical clustering web pages. Students in both Genomics and Computational Biology battle-tested numerous Math Minutes; special thanks to those whose questions led to Math Minute 6.1.

I continue to be inspired by the former teachers and mentors who led me to and through this field: Mike Waterman, Simon Tavaré, Gary Stormo, and John Williamson. They, along with friends and my family, made my contribution to this book a possibility. Malcolm, who was willing to share his vision for genomics education with a new colleague, made it a reality. Finally, I thank my husband Bill, whose patience and love never wear thin. He is my greatest support in all things.

—Laurie J. Heyer

REVIEWERS

Ian Boussy
Loyola University

Steven Brenner
University of California, Berkeley

Michael Buratovich
Spring Arbor University

Scott Cooper
University of Wisconsin, La Crosse

Todd Eckdahl
Missouri Western State College

Laura Hoopes
Pomona College

Bob Ivarie
University of Georgia

Mitrick Johns
Northern Illinois University

David Kass
Eastern Michigan University

Lee Kozar
Stanford University

Carissa Krane
University of Dayton

Ken Kubo
American River College

Mark Lubkowitz
Saint Michael's College

Albert MacKrell
Bradley University

John Merriam
University of California, Los Angeles

Peggy A. Redshaw
Austin College

Frank Rosenzweig
University of Florida College of Medicine

Carl J. Schmidt
University of Delaware

James A. Shapiro
University of Chicago

Elizabeth Vallen
Swarthmore College

Manuel F. Varela
Eastern New Mexico University

Quinn Vega
Montclair State University

Denise Wallack
Muhlenberg State College

CONTENTS

UNIT ONE

Genome Sequences

CHAPTER 1

Genome Sequence Acquisition and Analysis

Goals for Chapter 1

1.1 Defining Genomes

Define the field of genomics.
Learn how genomes are sequenced.
Understand the utility of short DNA segments.
Utilize online tools to analyze genome sequences.

1.2 What Have We Learned from the Human Genome Draft Sequences?

Survey human genome.
Verify genome annotations with online tools.
Recognize alternative forms of genes.
Explore epigenetic regulation of genome function.