

William W. Cohen · Charles K. Cohen

---

A Computer  
Scientist's  
Guide to  
**Cell Biology**

---

*Second Edition*

 Springer

# A Computer Scientist's Guide to Cell Biology

William W. Cohen • Charles K. Cohen

# A Computer Scientist's Guide to Cell Biology

Second Edition

 Springer

William W. Cohen  
Machine Learning Department  
Carnegie Mellon University  
Pittsburgh, PA, USA

Charles K. Cohen  
Johns Hopkins University  
Baltimore, MD, USA

ISBN 978-3-031-55906-8      ISBN 978-3-031-55907-5 (eBook)  
<https://doi.org/10.1007/978-3-031-55907-5>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2007, 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Heiti Paves / Alamy Stock Photo

If disposing of this product, please recycle the paper.

# Acknowledgments

*Charlie:*

This book was a family affair.

The first edition of this volume was almost entirely the work of my father and co-author, William. None of this would exist if not for him—while I was struggling to survive high school, he's the one who made the time to do the research, organize the information, drew the original figures, and write the original words. This was, and remains, his book, and I am forever grateful that he gave me the opportunity to work with him on this project and run roughshod over the text he spent so long writing.

My mother, Susan Cohen, proofread and indexed both editions (my late grandfather, William Daniel Kundin, also proofread the first edition). I'm sorry we put you through such a painful ordeal. Hopefully it was worth it.

My sister, Cassandra Cohen, updated a number of the figures, despite not knowing what she was working on half the time. If something's wrong, don't blame her.

Dr. Helen Rich gave us very helpful comments on an early draft of the second edition, including discussing important advances in biology in recent years.

As for me, I provided the words in this edition and not much else. The book already existed; all I did was go back and repeat it in a funny voice. Feel free to blame me if something's wrong.

On a personal note, I would never have finished the task if not for the tireless support of my partner, Ru—you're my home and my harbor, and without you I would have washed away long before finishing.

Finally, a big shout-out to the librarians at the Hazelwood Public Library for letting me camp out in front of the public computers for hours on end. Sorry!

*William:*

Writing the first edition of this book was, like many writing tasks, slow going for most of the way, although worthwhile in the end. I worked hard in organizing and curating the content of the first edition, but while that edition had many good points, I don't think any of the readers found it actually fun to read.

Writing the second edition with Charlie was rewarding on a totally different level. If there is anything more enjoyable than watching someone else take your precise but somewhat stilted prose, and make it engaging, it must be having one of your offspring do that. Charlie's modest comments in the acknowledgments above definitely undersell his contributions: he brought a new perspective to the book, as (unlike myself) he has actually spent years not just studying biology but doing real biology in real labs. Working with him was very rewarding technically, not just personally, and the end product of the collaboration is not only more up-to-date and readable, but a work that is a better bridge between how biologists and computer scientists might think about the field. It is also, in more than one place, more correct.

I'd like to thank my wife Susan for all her support over the years, for her standout job indexing and proofreading this book—and, of course, for her patience with Charlie and me for getting her drafts later than promised. I'm also grateful to my daughter Cassie for her work adding figures for the new material and updating the existing figures.

Finally, I'd like to thank the many readers of the first edition who approached me with questions, corrections, and occasionally thanks and encouragement for that work.

# Contents

<b>Introduction</b>	1
<b>How Cells Work: The Basics</b>	5
What Life Is Made Of	5
DNA	5
Proteins	6
Lipids and Membranes	7
Types of Life	8
Prokaryotes	8
Eukaryotes	9
Multicellular Life, Tissues, and Signaling	11
Viruses	12
Plasmids	13
Prions	13
Cellular Activity	14
The Central Dogma	14
Cellular Signaling	16
Cell Division	17
<b>Why Is Biology Hard?</b>	21
Proteins Interact in Complexes and Pathways	21
Individual Interactions Can Be Complicated	23
Enzymes Control Reaction Rates	23
Reaction Rates Can Be Highly Nonlinear	25

Enzymatic Pathways Are Complicated	29
Cellular Energy	29
Enzymatic Pathways Have Many Steps	30
Amplification and Pathways	30
Modularity and Locality Is Limited	32
How Things That Interact Find Each Other	32
Membranes and Locality	33
Biological Processes Can Cross Membranes	36
Wrap-Up	40
<b>Looking at Very Small Things</b>	43
Limitations of Optical Microscopes	43
Fluorescent Microscopes	45
Confocal Microscopes	46
Electron Microscopes	46
<b>Manipulation of the Very Small</b>	49
Taking Small Things Apart	49
Sorting Small Things	51
Centrifugation: Separation by Weight	51
Chromatography: Separation by Charge or Other Properties	51
Electrophoresis: Separation by Size or Shape	52
The Many Approaches to Sorting and Selection	54
Measuring Proteins at Scale	55
Parallelism, Automation, and Reuse in Biology	58
Classifying Things by Their Pieces	60
Histograms and Peptide Maps	60
Mass Spectrometry	61
DNA Fingerprinting	61
Wrap-Up on Classifying Things by Their Parts	62
<b>Reprogramming Cells</b>	63
Our Friends, the Microorganisms	63
Restriction Enzymes and Restriction-Methylase Systems	63
CRISPR/Cas9	66
Inserting Foreign DNA into a Cell	67
Plasmids as Insertion Vectors	68
Phages as Insertion Vectors	70
Using Genomic DNA Libraries	71



Creating Novel Proteins: Tagging and Phage Display 72  
 Yeast Two-Hybrid Assays Using Fusion Proteins 73

**Other Ways to Use Biology for Biological Experiments 77**

Replicating DNA in a Test Tube 77  
   DNA Replication: The Basics 77  
   DNA Replication: Not So Basic 78  
   DNA Replication in a Tube: PCR 79  
 Sequencing DNA by Partial Replication and Sorting 80  
   Sanger Sequencing 80  
   Massively Parallel Sequencing 82  
   How Fast, Cheap Sequencing Has Changed Biology 83  
 Other In Vitro Systems: Translation and Reverse Transcription 84  
   Translation in a Tube 84  
   Reverse Transcription 84  
 Antibodies: Exploiting the Natural Defenses of a Cell 85  
   Immunofluorescence 86  
   Antibodies in Action: COVID Tests 86  
   mRNA Vaccines 88  
 RNA Interference 90  
 Serial Analysis of Gene Expression 91

**Bioinformatics 95**

DNA Sequence Analysis 95  
   Levenshtein Distance 96  
   Smith-Waterman Similarity 97  
   Multiple Sequence Alignment 101  
   Analyzing Similarities of Species 102  
   Molecular Clocks 103  
   Data Mining DNA Sequence Databases 103  
 High-Throughput Experiments 104  
 Search Engines 105

**Where to Go From Here? 107**

**Sources 109**

**Index 111**



# Introduction

As the amount of biological data grows, the task of understanding existing data becomes increasingly important, and this is largely a task best undertaken by computer science. This book is for the many curious souls who are coming into biology from backgrounds in computer science, especially the fields of information retrieval, natural language processing, and/or machine learning.

One major difference between biology and computer science is that in computer science, the world we explore is in large part our own creation, and a large part of what we do is make our creation understandable by finding useful abstractions, and then building more complex things by combining these abstractions together. For example, a deterministic finite state machine is a useful abstraction for computations that process discrete inputs sequentially with limited memory—we study this, and study stack data structures, and then study the result of combining them to make a push-down automaton. These abstractions might be compromised when we optimize our systems for performance, but they are rarely abandoned completely, because comprehensibility, elegance, and simplicity are practically important for systems that must be maintained and improved by humans.

In contrast, biology doesn't lend itself to clean and comprehensible abstract models: evolution relentlessly marches toward improved performance without worrying much about simplicity. Even the "simplest" forms of life are seemingly endless in their unique complexities, and almost every general statement about how organisms function comes with an asterisk. And unlike in computer science, the details that underlie the complexity of the real

systems are not something we can or should ignore, hoping they will be cleaned up in the next version—instead, the awkward details are, collectively, the real subject of the science of biology.

For the purposes of this book, we have broken the field down into three parts:

- **Biological mechanics** are the actual nitty-gritty details of how things work at the cellular level—protein pathways, chloroplasts, and so on. This is the typical focus of introductory biology classes and textbooks, and you would correctly suppose this is the essence of what biologists actually study. However, it's a surprisingly small part of what biologists write and talk about.
- **Experimental methods**, on the other hand, *are* what biologists spend most of their time talking about. If you pick up a typical biology paper, the actual *conclusions*, the newly discovered details about how these systems function, are compact enough to be laid out in the abstract.

As you read through this book, you'll find that it's mostly about methods. Biologists spend most of their word count talking about how they conducted their experiments, how cells were cultured, and what assays were run and a host of other details. The results, in isolation, tell you very little—the only way to tell the difference between good research and bad research is to examine how data was initially gathered. But to an outsider, that can be far from a simple task. The language of biology is rich, detailed, and almost impenetrable to the average layperson; learning its intricacies is as important as learning about biological mechanisms or experimental techniques.

- **Language and nomenclature** can be considered a “part” of biology in its own right. Without spending at least a little bit of time learning how to speak it, this book would be pretty useless.

If you like, you can think of biology as a journey to some strange, exotic land. The inhabitants speak a strange and often incomprehensible language, the customs and practices may be like nothing seen before, and even the most basic of tasks appear completely alien. With that in mind, our goal is to provide a short introduction to the three core aspects of cell biology—a travel guide, to continue the previous metaphor, focusing on high-level principles, and relating as much as possible to familiar concepts from computer science.

Consequently, in this book, we will gloss over some concepts and oversimplify others, setting aside many otherwise-fascinating theories and details. Biology is fractal; no matter how deep you look, there is always another layer of complexity. For a more comprehensive background on biology, there are many excellent textbooks, written by people far more qualified—the last chapter of this book will introduce several of our favorites.