

*"An excellent book for beginners and occasional practitioners"*  
Reviewer, *Journal of the American Medical Association*

# Bioinformatics

## FOR DUMMIES®

2nd Edition

**A Reference  
for the  
Rest of Us!®**

FREE eTips at [dummies.com](http://dummies.com)®

**Jean-Michel Claverie, PhD**

Research Director, France's Centre National  
de la Recherche Scientifique (CNRS)

**Cedric Notredame, PhD**

Professor of Bioinformatics, Switzerland's  
Lausanne University and the CNRS

Updated to cover  
multiple new  
genomes and  
databases



***Bioinformatics***  
FOR  
**DUMMIES®**  
2ND EDITION

**by Jean-Michel Claverie, PhD  
and Cedric Notredame, PhD**



Wiley Publishing, Inc.



***Bioinformatics***

FOR

**DUMMIES<sup>®</sup>**

2ND EDITION



***Bioinformatics***  
FOR  
**DUMMIES®**  
2ND EDITION

**by Jean-Michel Claverie, PhD  
and Cedric Notredame, PhD**



Wiley Publishing, Inc.

## Bioinformatics For Dummies®, 2nd Edition

Published by  
Wiley Publishing, Inc.  
111 River Street  
Hoboken, NJ 07030-5774

www.wiley.com

Copyright © 2007 by Wiley Publishing, Inc., Indianapolis, Indiana

Published by Wiley Publishing, Inc., Indianapolis, Indiana

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Legal Department, Wiley Publishing, Inc., 10475 Crosspoint Blvd., Indianapolis, IN 46256, (317) 572-3447, fax (317) 572-4355, or online at <http://www.wiley.com/go/permissions>.

**Trademarks:** Wiley, the Wiley Publishing logo, For Dummies, the Dummies Man logo, A Reference for the Rest of Us!, The Dummies Way, Dummies Daily, The Fun and Easy Way, Dummies.com, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. Wiley Publishing, Inc., is not associated with any product or vendor mentioned in this book.

**LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.**

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 800-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002.

For technical support, please visit [www.wiley.com/techsupport](http://www.wiley.com/techsupport).

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Control Number: 2006934844

ISBN13: 978-0-470-08985-9

ISBN10: 0-470-08985-7

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

1B/SX/RR/QW/IN



# *About the Authors*

**Jean-Michel Claverie** is Professor of Medical Bioinformatics at the School of Medicine of the Université de la Méditerranée, and a consultant in genomics and bioinformatics. He is the founder and current head of the Structural & Genomic Information Laboratory, located in Marseilles, a sunny city on the Mediterranean coast of France. Using science as a pretext to travel, Jean-Michel has held positions in Paris (France), Sherbrooke (PQ, Canada), the Salk Institute (La Jolla, CA), the Pasteur Institute (Paris), Incyte pharmaceutical (Palo Alto, CA); and the National Center for Biotechnology Information (Bethesda, MD). He has used computers in biology since the early days — his Ph.D. work involved modeling biochemical reactions by programming an 8K Honeywell 516 computer right from the console switches! Although he has no clear recollection of it, he has been credited with introducing the French word “bioinformatique” in the late eighties, before involuntarily coining the catchy “bioinformatics” by mistranslating it while giving a talk in English!

Jean-Michel’s current research interests are in microbial and structural genomics, and in the development of bioinformatic methods for the prediction of gene function. He is the author or coauthor of more than 150 scientific publications, and a member of numerous international review panels and scientific councils. In his spare time, he enjoys the relaxed pace of life in Marseilles, with his wife Chantal and their two sons, Nicholas and Raphael.

**Cedric Notredame** is a researcher at the French National Centre for Scientific Research. Cedric has used and abused the facilities offered by science to wander around Europe. After a Ph.D. at EMBL (Heidelberg, Germany) and at the European Bioinformatics Institute (Cambridge, UK) under the supervision of Des Higgins (yes, the ClustalW guy), Cedric did a post-doc at the National Institute of Medical Research (London, UK), in the lab of Willie Taylor and under the supervision of Jaap Heringa. He then did a post-doc in Lausanne (Switzerland) with Phillip Bucher, and remained involved with the Swiss Institute of Bioinformatics for several years. Having had his share of rain, snow, and wind, Cedric has finally settled in Marseilles, where the sun and the sea are simply warmer than any other place he has lived in.

Cedric dedicates most of his research to the multiple sequence alignment problem and its many applications in biology. His friends claim that his entire life (past, present, future) is somehow stuffed into the T-Coffee multiple-sequence alignment package. When he is not busy dismantling T-Coffee and brewing new sequences, Cedric enjoys life in the company of his wife, Marita.





# Dedication

This is for my parents Monique and Jack, for keeping me in school, and for Chantal, for keeping me happy — in and out of the lab. It's also for my daughter Vanessa, and my sons Nicholas and Raphael, for reminding me that not *everything* in life is scientific.

— J-MC

This is for my wife Marita, my daughter Lina, my mother Marie and in memory of my grandparents, Simone and Louis.

— CN

# Authors' Acknowledgments

The entire Wiley staff did a great job pulling together to publish this book on tight deadlines. We'd especially like to thank our tireless project editor, Paul Levesque, and Barry Childs-Helton, who did a great job copyediting a text full of obscure biochemical words.

We'd also like to thank Amey Godse, our technical editor. Amey nailed down major and minor inaccuracies alike. His many suggestions did much to improve the book.

We also have to thank the bioinformatics community for creating the many great Web resources that we describe in this book and for making them available for free over the Internet. We personally know a number of the folks who keep these sites up and running — and salute all of them for their hard work, enthusiasm, and dedication. Topping this list are the staff members of the Swiss Bioinformatics Institute, who run the ExPASy and the Swiss EMBnet Web server. They always went out of their way to answer any query regarding their site. The NCBI folks have also been very helpful, and we thank them for that.

We also want to pat each other on the back for making the writing of this book great fun!

Finally, we'd like to thank our families and friends, who put up with missed dinners, extra child care, changing deadlines, late nights, and the many other demands of a project like this. We really appreciate their patience — and promise that we won't do another one . . . at least not anytime soon!

**Acquisitions, Editorial, and  
Media Development**

**Project Editor:** Paul Levesque

**Acquisitions Editor:** Melody Layne

**Senior Copy Editor:** Barry Childs-Helton

**Technical Editor:** Amey Godse

**Editorial Manager:** Leah Cameron

**Media Development Specialists:** Angela Denny,  
Kate Jenkins, Steven Kudirka, Kit Malone

**Media Development Coordinator:**  
Laura Atkinson

**Media Project Supervisor:** Laura Moss

**Media Development Manager:**  
Laura VanWinkle

**Editorial Assistant:** Amanda Foxworth

**Sr. Editorial Assistant:** Cherie Case

Cartoons: Rich Tennant  
([www.the5thwave.com](http://www.the5thwave.com))

**Composition Services**

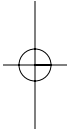
**Project Coordinator:** Jennifer Theriot

**Layout and Graphics:** Carl Byers,  
Lavonne Cook, Barbara Moore,  
Shelley Norris, Barry Offringa,  
Laura Pence

**Proofreaders:** Susan Moritz, Charles Spencer,  
Rob Springer, Techbooks

**Indexer:** Techbooks

**Anniversary Logo Design:** Richard Pacifico



---

**Publishing and Editorial for Technology Dummies**

**Richard Swadley**, Vice President and Executive Group Publisher

**Andy Cummings**, Vice President and Publisher

**Mary Bednarek**, Executive Acquisitions Director

**Mary C. Corder**, Editorial Director

**Publishing for Consumer Dummies**

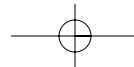
**Diane Graves Steele**, Vice President and Publisher

**Joyce Pepple**, Acquisitions Director

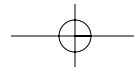
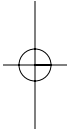
**Composition Services**

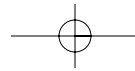
**Gerry Fahey**, Vice President of Production Services

**Debbie Stailey**, Director of Composition Services



<i>Introduction</i> .....	1
<b><i>Part I: Getting Started in Bioinformatics</i></b> .....	<b>7</b>
Chapter 1: Finding Out What Bioinformatics Can Do for You.....	9
Chapter 2: How Most People Use Bioinformatics .....	29
<b><i>Part II: A Survival Guide to Bioinformatics</i></b> .....	<b>67</b>
Chapter 3: Using Nucleotide Sequence Databases.....	69
Chapter 4: Using Protein and Specialized Sequence Databases.....	105
Chapter 5: Working with a Single DNA Sequence .....	129
Chapter 6: Working with a Single Protein Sequence .....	159
<b><i>Part III: Becoming a Pro in Sequence Analysis</i></b> .....	<b>197</b>
Chapter 7: Similarity Searches on Sequence Databases .....	199
Chapter 8: Comparing Two Sequences.....	235
Chapter 9: Building a Multiple Sequence Alignment.....	265
Chapter 10: Editing and Publishing Alignments .....	303
<b><i>Part IV: Becoming a Specialist: Advanced Bioinformatics Techniques</i></b> .....	<b>327</b>
Chapter 11: Working with Protein 3-D Structures .....	329
Chapter 12: Working with RNA .....	353
Chapter 13: Building Phylogenetic Trees .....	371
<b><i>Part V: The Part of Tens</i></b> .....	<b>403</b>
Chapter 14: The Ten (Okay, Twelve) Commandments for Using Servers .....	405
Chapter 15: Some Useful Bioinformatics Resources.....	411
<b><i>Index</i></b> .....	<b>417</b>



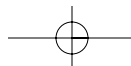
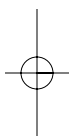


<i>Introduction</i> .....	1
What This Book Does for You.....	1
Foolish Assumptions .....	2
How This Book Is Organized.....	2
Part I: Getting Started in Bioinformatics .....	3
Part II: A Survival Guide to Bioinformatics .....	3
Part III: Becoming a Pro in Sequence Analysis .....	3
Part IV: Becoming a Specialist: Advanced Bioinformatics Techniques.....	3
Part V: The Part of Tens.....	4
Icons Used in This Book.....	4
Where to Go from Here.....	4

***Part 1: Getting Started in Bioinformatics* ..... 7**

**Chapter 1: Finding Out What Bioinformatics Can Do for You ..... 9**

What Is Bioinformatics? .....	9
Analyzing Protein Sequences .....	10
A brief history of sequence analysis.....	12
Reading protein sequences from N to C.....	13
Working with protein 3-D structures.....	14
Protein bioinformatics covered in this book.....	16
Analyzing DNA Sequences .....	17
Reading DNA sequences the right way.....	17
The two sides of a DNA sequence.....	18
Palindromes in DNA sequences.....	20
Analyzing RNA Sequences.....	21
RNA structures: Playing with sticky strands .....	22
More on nucleic acid nomenclature .....	23
DNA Coding Regions: Pretending to Work with Protein Sequences .....	23
Turning DNA into proteins: The genetic code .....	24
More with coding DNA sequences .....	25
DNA/RNA bioinformatics covered in this book.....	26
Working with Entire Genomes .....	26
Genomics: Getting all the genes at once .....	27
Genome bioinformatics covered in this book .....	28



Searching PubMed using limits .....	38
A few more tips about PubMed .....	41
Retrieving Protein Sequences.....	42
ExPASy: A prime Internet site for protein information .....	42
More advanced ways to retrieve protein sequences.....	45
Retrieving a list of related protein sequences .....	48
Retrieving DNA Sequences.....	51
Not all DNA is coding for protein .....	51
Going from protein sequences to DNA sequences.....	52
Retrieving the DNA sequence relevant to my protein .....	53
Using BLAST to Compare My Protein Sequence to Other Protein Sequences .....	57
Making a Multiple Protein Sequence Alignment with ClustalW .....	62

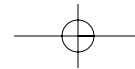
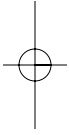
## ***Part II: A Survival Guide to Bioinformatics.....67***

### **Chapter 3: Using Nucleotide Sequence Databases .....69**

Reading into Genes and Genomes .....	70
Prokaryotes: Small bugs, simple genes .....	70
Eukaryotes: Bigger bugs, complex genes .....	72
Making Use (and Sense) of GenBank .....	73
Making sense of the GenBank entry of a prokaryotic gene .....	73
Making sense of the GenBank entry of an eukaryotic mRNA .....	78
Making sense of a GenBank eukaryotic genomic entry.....	79
Working with related GenBank entries .....	84
Retrieving GenBank entries without accession numbers .....	85
Using a Gene-Centric Database .....	86
Working with Whole-Genome Databases .....	88
Working with complete viral genomes .....	89
Working with complete bacterial genomes.....	92
More bacterial genomics at TIGR.....	94
Microbes from the environment at DoE .....	96
Exploring the Human Genome.....	97
Finding out about the Ensembl project .....	98

### **Chapter 4: Using Protein and Specialized Sequence Databases . . .105**

From Translated ORFs to Mature Proteins .....	107
ORFs: What you see is NOT what you get.....	107
A personal final destination for each protein.....	109
A combinatorial diversity of folds and functions.....	109



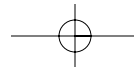
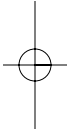
The Cross-References .....	116
The Keywords .....	118
The Features .....	119
Finally, the sequence itself .....	123
Finding Out More about Your Protein .....	123
Finding out more about “modified amino acids” .....	124
Some advanced biochemistry sites .....	125
Finding out more about biochemical pathways .....	125
Finding out more about protein structures .....	126
Finding out more about major protein families .....	127

**Chapter 5: Working with a Single DNA Sequence ..... 129**

Catching Errors Before It’s Too Late.....	130
Removing vector sequences .....	130
Cases when you shouldn’t discard your sequence.....	133
Computing/Verifying a Restriction Map .....	134
Designing PCR Primers .....	135
Analyzing DNA Composition.....	138
Establishing the G+C content of your sequence .....	138
Counting words in DNA sequences.....	139
Counting long words in DNA sequences .....	140
Experimenting with other DNA composition analyses.....	142
Finding internal repeats in your sequence .....	142
Identifying genome-specific repeats in your sequence .....	145
Finding Protein-Coding Regions .....	145
ORFing your DNA sequence.....	146
Analyzing your DNA sequence with GeneMark .....	148
Finding internal exons in vertebrate genomic sequences .....	149
Complete gene parsing for eukaryotic genomes.....	151
Analyzing your sequence with GenomeScan .....	151
Assembling Sequence Fragments.....	153
Managing large sequencing projects with public software.....	154
Assembling your sequences with CAP3 .....	155
Beyond This Chapter.....	157

**Chapter 6: Working with a Single Protein Sequence ..... 159**

Doing Biochemistry on a Computer .....	160
Predicting the main physico-chemical properties of a protein....	161
Interpreting ProtParam results.....	164
Digesting a protein in a computer .....	166





Finding Known Domains in Your Protein .....	180
Choosing the right collection of domains .....	182
Finding domains with InterProScan .....	183
Interpreting InterProScan results .....	185
Finding domains with the CD server .....	187
Interpreting and understanding CD server results .....	189
Finding domains with Motif Scan .....	190
Discovering New Domains in Your Proteins .....	194
More Protein Analysis for Free over the Internet .....	194

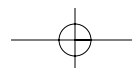
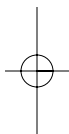
## ***Part III: Becoming a Pro in Sequence Analysis* ..... 197**

### **Chapter 7: Similarity Searches on Sequence Databases ..... 199**

Understanding the Importance of Similarity .....	200
The Most Popular Data-Mining Tool Ever: BLAST .....	201
BLASTing protein sequences .....	201
Understanding your BLAST output.....	209
BLASTing DNA sequences .....	216
The BLAST way of doing things.....	218
Controlling BLAST: Choosing the Right Parameters .....	219
Controlling the sequence masking.....	220
Changing the BLAST alignment parameters .....	223
Controlling the BLAST output.....	224
Making BLAST Iterative with PSI-BLAST .....	226
PSI-BLASTing protein sequences.....	226
Avoiding mistakes when running PSI-BLAST .....	228
Discovering and using protein domains	
with BLAST and PSI-BLAST.....	230
Similarity Searches for Free over the Internet .....	231

### **Chapter 8: Comparing Two Sequences ..... 235**

Making Sure You Have the Right Sequences and the Right Methods....	236
Choosing the right sequences .....	236
Choosing the right method .....	237
Making a Dot Plot .....	239
Choosing the right dot-plot flavor.....	240
Using Dotlet over the Internet .....	241
Doing biological analysis with a dot plot .....	249



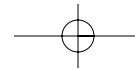
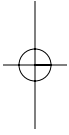
Aligning Proteins and DNA.....	262
Free Pairwise Sequence Comparisons over the Internet .....	262

**Chapter 9: Building a Multiple Sequence Alignment . . . . . 265**

Finding Out if a Multiple Sequence Alignment Can Help You.....	266
Identifying situations where multiple alignments do not help.....	267
Helping your research with multiple sequence alignments .....	267
Choosing the Right Sequences .....	270
The kinds of sequences you're looking for .....	271
Gathering your sequences with online BLAST servers .....	275
Choosing the Right Method of Multiple Sequence Alignment.....	281
Using ClustalW.....	282
Aligning sequences and structures with Tcoffee .....	287
Crunching large datasets with MUSCLE .....	291
Interpreting Your Multiple Sequence Alignment.....	291
Recognizing the good parts in a protein alignment .....	292
Taking your multiple alignment further .....	294
Comparing Sequences That You Can't Align .....	297
Making multiple local alignments with the Gibbs sampler.....	298
Searching conserved patterns.....	299
Internet Resources for Doing Multiple Sequence Comparisons .....	299
Making multiple alignments with ClustalW around the clock.....	300
Finding your favorite alignment method.....	300
Searching for motifs or patterns .....	301

**Chapter 10: Editing and Publishing Alignments . . . . . 303**

Getting Your Multiple Alignment in the Right Format .....	305
Recognizing the main formats .....	307
Working with the right format .....	307
Converting formats .....	309
Watching out for lost data.....	312
Using Jalview to Edit Your Multiple Alignment Online.....	313
Starting Jalview.....	314
Editing a group of sequences.....	316
Useful features of Jalview .....	318
Saving your alignment in Jalview .....	318
Preparing Your Multiple Alignment for Publication .....	319
Using Boxshade .....	319
Logos.....	322

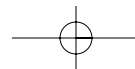
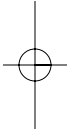


**Chapter 11: Working with Protein 3-D Structures ..... 329**

From Primary to Secondary Structures .....330  
    Predicting the secondary structure of a protein sequence .....330  
    Predicting additional structural features .....334  
From the Primary Structure to the 3-D Structure .....336  
    Retrieving and displaying a 3-D structure from a PDB site .....337  
    Guessing the 3-D structure of your protein .....340  
    Looking at sequence features in 3-D .....343  
Beyond This Chapter .....350  
    Finding proteins with similar shapes .....350  
    Finding other PDB viewers .....350  
    Classifying your PDB structure .....351  
    Doing homology modeling .....351  
    Folding proteins in a computer .....351  
    Threading sequences onto PDB structures .....351  
    Looking at structures in movement .....352  
    Predicting interactions .....352

**Chapter 12: Working with RNA ..... 353**

Predicting, Modeling and Drawing RNA Secondary Structures .....354  
Using Mfold .....355  
    Interpreting mfold results .....359  
    Forcing interaction in mfold .....361  
Searching Databases and Genomes for RNA Sequences .....362  
    Finding tRNAs in a genome .....363  
    Using PatScan to look for RNA patterns .....363  
Finding the “New” RNAs: miRNAs and siRNAs .....367  
Doing RNA Analysis for Free over the Internet .....368  
    Studying evolution with ribosomal RNA .....369  
    Finding the small, non-coding RNA you need .....369  
    Generic RNA resources .....370



Building the Kind of Tree You Need.....	383
Computing your tree.....	383
Knowing what's what in your tree.....	398
Displaying your phylogenetic tree.....	399
Doing Phylogeny for Free over the Internet.....	400
Finding online resources.....	400
Finding generic resources.....	401
Collections of orthologous genes.....	402

***Part V: The Part of Tens*.....403**

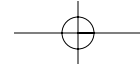
**Chapter 14: The Ten (Okay, Twelve) Commandments  
for Using Servers .....405**

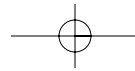
Keep in Mind: Your Data Is Never Secure on the Web.....	406
Remember the Server, the Database, and the Program Version You Used.....	406
Write Down the Sequence-Identification Numbers.....	407
Write Down the Program Parameters.....	407
Save Your Internet Results the Right Way.....	407
Use E-Values.....	408
Make Sure You Can Trust Your Alignments.....	408
Use Different Programs to Check Borderline Results.....	409
Stay Away from Unpublished Methods!.....	409
Databases Are Not Like Good Wine.....	409
Just Because It Looks Free Doesn't Mean It Is Free . . .	410
Biting the Bullet at the Right Time.....	410

**Chapter 15: Some Useful Bioinformatics Resources .....411**

Ten Major Databases.....	411
Ten Major Bioinformatics Software Programs.....	412
Ten Major Resource Locators.....	414
Some Places to Find Out What's Really Going On.....	415

***Index*.....417**





In the first edition, we presented bioinformatics as a brand new discipline on the rise. How right we were! Since then, it has become so prominent that anybody with an interest in biology, biotechnology, modern medicine, or (for that matter) genetically engineered food or drugs simply cannot afford to remain ignorant about the topic. With this book, you've come to the right place to quickly learn the basics.

But wait — if you expect something complicated, you're in for a (good or bad) surprise: Bioinformatics is nothing but good, sound, regular biology, appropriately dressed so it can fit into a computer.

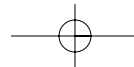
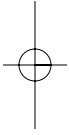
Bioinformatics is about searching biological databases, comparing sequences, looking at protein structures, and (more generally) asking biological and biomedical questions with a computer. The bioinformatics we show you in this book can save you months of work in the lab at the minute cost of a few hours' work with your computer.

Although you'll find standard biological terms throughout, don't look here for long equations and computer-geek gibberish. The purpose of this book is to show you quickly and plainly how to use the bioinformatics programs that you need to get your work done. On every page, we give you tricks and treats to get the most out of existing tools. If you didn't know that you can use the most sophisticated programs for free over the Internet — and that you can do this (sometimes) without installing anything on your own computer — then stay tuned: You're in for many more good surprises.

## *What This Book Does for You*

This book is here to help you get things done. For every standard bioinformatics task you may want to undertake, you'll find detailed steps that you can use to quickly produce the result you need.

To use most of the tools we describe in this book, you don't need to install any program on your computer. Everything we show you here runs over the Internet via your Internet browser.



At the end of most chapters you'll find a convenient "Doing It for Free over the Internet" section, where we list a few carefully chosen Web sites that are similar to those we describe in the rest of the chapter. Treat this information as a spare wheel! If the main site is down, this section probably lists a convenient replacement.

## *Foolish Assumptions*

Putting a project's assumptions right up front is just good policy. While writing this book, we have assumed that

- ✓ You have a PC running Microsoft Windows.
- ✓ You have an Internet connection (a fast one if possible, but not necessarily).
- ✓ You likely have a background in molecular biology. If you don't — or if you need to brush up on your molecular biology — Chapter 1 gives you a brief overview of the basics.
- ✓ You know how to use an Internet browser but not much more about computers.
- ✓ You don't want to become a bioinformatics guru; you simply want to use the right tools for your problem and not spend days finding out about things you don't need!
- ✓ Most private biotech companies consider it unsafe to send data over the Internet. We assume here that the data you want to analyze over the Internet is *not* very confidential. Also, some of the "public" databases and services listed in this book require commercial users to enter into a license agreement.



## *How This Book Is Organized*

Bioinformatics is a broad field, with many nooks and crannies, hills and dales, and other charming features. Rather than present the whole vast discipline in one fell swoop, we've divided our discussion into five (more manageable) parts.

you of just those bits of molecular biology that you'll need to know when you do sequence analysis. We show you here how to run the main bioinformatics tools so that you know what's in store for you.

## ***Part II: A Survival Guide to Bioinformatics***

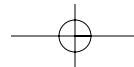
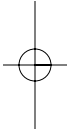
If you want to find out everything that's ever been published on your sequence, this part is for you. It shows you how you can deal with the bioinformaticist's bread and butter: *DNA or protein sequences and their databases*. Here we tell you where you can find all the available sequences, and how to find the one you really need among zillions of irrelevant others. We also show you how to gather everything that's known in the universe about this special sequence that interests you so much (at least all of it that's available online).

## ***Part III: Becoming a Pro in Sequence Analysis***

If you want to compare sequences, this is the part for you. Here we show you how to search databases for sequences that are similar to yours, as well as show you how to compare two or more sequences. This part also tells you how to gather hints about the function of a gene, through sequence comparisons. Finally, we give you pointers on how to produce, edit, and beautify your multiple sequence alignments so you can show them in presentations and publications.

## ***Part IV: Becoming a Specialist: Advanced Bioinformatics Techniques***

To take full advantage of this part, you should have a pretty good idea of what you're looking for. Heavy stuff is going on here: how to predict a protein structure, how to predict an RNA structure, and how to do phylogenetic analysis. These are complicated subjects; it's simply amazing what you can do with a simple PC, thanks to the Internet resources we describe in this part.





every student and his or her cousins putting semester reports online, finding exactly what you need with a simple keyword search can be a daunting task. In the Part of Tens, we give you a list of central resources that you can use as a starting point. Chances are that the program or server you're looking for is only one or two clicks away. In this part, we also give you ten important pieces of advice to make sure that your lab work can safely depend on your Internet work.

## Icons Used in This Book

Always eager to please, we've decided to use a series of icons in the margins of this book as a way to help you key in on important information. We came up with four, which seemed like a nice, round number.



Some particularly technoid information is coming up. You can skip it and nothing terrible will happen. Yet, if you want to be in full control of what you're doing, reading this may help! Your call. . . .



This icon shows you something simple, or smart, or a cute shortcut. In any case, it's something that can save you time and trouble.



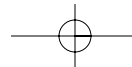
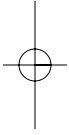
There are many booby traps around when you use Internet servers. This icon warns you when some ambiguity surrounds what the server you're using is up to — or when disaster is only one (wrong) mouse click away. Treat the Warning icon with respect — especially in a steps list!



This icon indicates something you should remember. It can be one of the few important principles that you need to know, or it can be a very special tip — the kind that can save you three days of work (or drive you nuts if you forget it). You may assume that the head of your institute/company got to the top by discovering and applying one or more of pearls of wisdom in these very special tips!

## Where to Go from Here

If you know nothing about bioinformatics, this book is here to reassure you. Bioinformatics is a much simpler subject than you ever thought possible. For most people new to this field, the main difficulty is finding out the kind of



logical knowledge — and you can do this with the most sophisticated tools ever developed by mankind. And how much is this going to cost you? Nothing!

If you do molecular biology, this is the equivalent of having an entire lab with expensive, state-of-the-art equipment and staffed by an army of post-docs who can go fetch anything you need any time you need it. The only difference is that you cannot set this lab on fire (even if you try very hard).

If you think of it, it is quite incredible to realize that all this is right here, at your fingertips, one or two mouse clicks away! The Web is borderless; it is colorblind and unimpressed by wealth! Whether you come from a rich or a poor country, whether you're a first-year student, a scientist, or a Nobel Prize winner, you have access — for free — to the same high-quality information. No other scientific discipline has ever been so democratically widespread.

This book isn't a textbook but a cookbook! And we take pride in this! It contains many recipes that colleagues showed us over the years or that we discovered ourselves. Accommodating and serving biological data is something very personal — and we're sure that you'll gradually find your own way to do it. In the meantime, if you need a quick fix, you can always use some of the off-the-shelf solutions that we provide here.

No discipline in science has benefited as much as biology from the “global village” phenomenon of the Internet. Whatever your question, whatever you want to do, starting on the Internet is the proper thing to do. Nonetheless, remember that the best *and* the worst appear online these days. Do as you do in real life — and trust only those sites or institutions that you know well.



This book is as up-to-date as we can make it, but the world doesn't stand still right after we finish correcting the last galley proofs and send *Bioinformatics For Dummies* into the bookstores. For those of you who want up-to-date info on the growing field of bioinformatics (including lists of our favorite bioinformatics links) and don't want to wait until the next edition, check out the Web site associated with this title at [www.dummies.com/extras](http://www.dummies.com/extras).

Sometimes browsing the Internet gives one the depressing feeling that everything has been done by others and that it's all over. This may be true. Now that the whole world talks together, it's clear that there's a finite number of interesting questions to ask. That's the bad news. The good news is that there are many more answers than there are questions! Never exclude the hypothesis that your answer may be the best in the universe (at least for a few days. . . .)!

