Bioinformatics FOR DUMMIES®

2ND EDITION

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

BICENTENNIAL
1807 WILEY 2007

Wiley Publishing, Inc.

Bioinformatics FOR DUMMIES®

2ND EDITION

Bioinformatics FOR DUMMIES®

2ND EDITION

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

BICENTENNIAL
1807
WILEY
2007
BICENTENNIAL

Wiley Publishing, Inc.

Bioinformatics For Dummies®, 2nd Edition

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

www.wiley.com

Published by Wiley Publishing, Inc., Indianapolis, Indiana

Published simultaneously in Canada

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.

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

WILEY

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!

Publisher's Acknowledgments

We're proud of this book; please send us your comments through our online registration form located at www.dummies.com/register/.

Some of the people who helped bring this book to market include the following:

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)

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

Contents at a Glance

Table of Contents

Introduction

Welcome to the second edition of *Bioinformatics For Dummies!*

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.

If you know what you want to do — or at least know the task by name — going through the Table of Contents is the best strategy for finding exactly what you need. If you have an idea of what you want to do but you're not sure how to express it with words, Chapter 2 is here to help you decide which part of the book will suit your needs.

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!

WARNING!

- 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.

Part I: Getting Started in Bioinformatics

If you have less than an hour to find out what bioinformatics can do for you, Part I is the right place for you! It tells you everything you need to know in order to actually *do* something with bioinformatics. In Part I, we also remind 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.

Part V: The Part of Tens

Welcome to our bazaar! If you haven't found what you were looking for in the other parts, you're now in the right place. The wealth of online resources that exist in bioinformatics is extraordinary — and almost overwhelming. With 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.

TECHNICAL STUFF

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. . . .

TIP

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.

WARNING!

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!

REMEMBER

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 questions they can ask with these new tools. If you're a biologist, don't let the computer scare you; bioinformatics is nothing more than good, sound, regular biology hidden inside a computer.

The magic thing about bioinformatics is that, with a simple Internet connection, you can browse databases that contain the sum of our entire human biological 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.

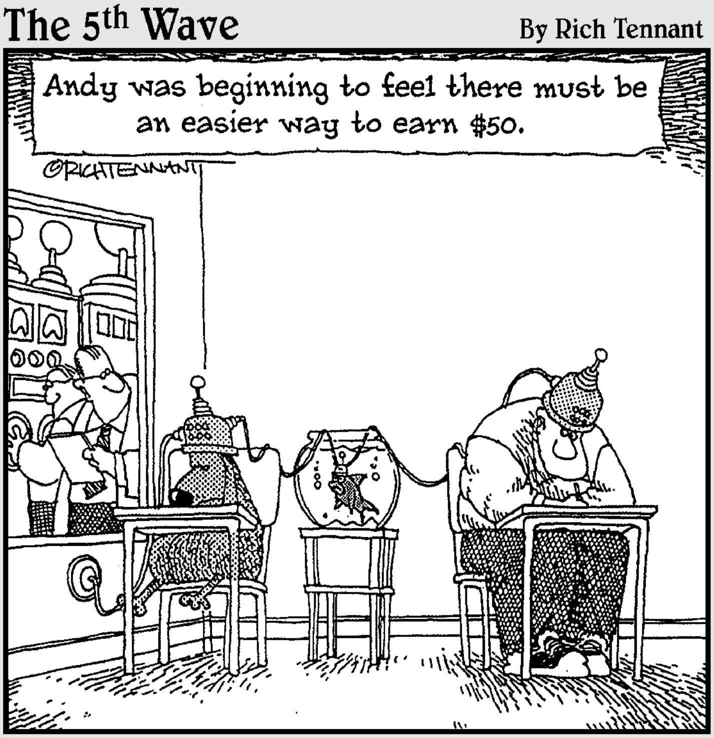
TIP

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`.

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. . . .)!

Part I

Getting Started in Bioinformatics


The 5th Wave
By Rich Tennant
Andy was beginning to feel there must be an easier way to earn $50.
©RICHTENNANT

In this part . . .

Bioinformatics is a new discipline, which means that nobody should feel ashamed if he or she doesn't have a clue what the excitement's all about. Don't worry; after finishing this book, you'll be speaking bioinformaticsspeak with the best of them.

We start you off in Part I with a quick reminder of what you need to know about DNA and proteins to make sense of this book. We also give you an overview of the main bioinformatics tools available on the Internet.

We don't give too many details here, but if all you need to know is which Internet page to open and which button to press, come on in, 'cuz we've got just what you need!

Chapter 1

Finding Out What Bioinformatics Can Do for You

In This Chapter

- Defining bioinformatics
- Understanding the links between modern biology, genomics, and bioinformatics
- Determining which biological questions bioinformatics can help you answer quickly

> *Organic chemistry is the chemistry of carbon compounds. Biochemistry is the study of carbon compounds that crawl.*
>
> — Mike Adam

It looks like *bio*logists are colonizing the dictionary with all these *bio*-words: we have bio-chemistry, bio-metrics, bio-physics, bio-technology, bio-hazards, and even bio-terrorism. Now what's up with the new entry in the bio-sweepstakes, bio-informatics?

What Is Bioinformatics?

In today's world, computers are as likely to be used by biologists as by any other highly trained professionals — bankers or flight controllers, for example. Many of the tasks performed by such professionals are common to most of us: We all tend to write lots of memos and send lots of e-mails; many of us use spreadsheets, and we all store immense amounts of never-to-be-seen-again data in complicated file systems.

However, besides these general tasks, biologists also use computers to address problems that are very specific to biologists, which are of no interest to bankers or flight controllers. These specialized tasks, taken together, make up the field of *bioinformatics.* More specifically, we can define bioinformatics as the computational branch of molecular biology.

Time for a little bit of history. Before the era of bioinformatics, only two ways of performing biological experiments were available: within a living organism (so-called *in vivo*) or in an artificial environment (so-called *in vitro,* from the Latin *in glass*). Taking the analogy further, we can say that bioinformatics is in fact *in silico* biology, from the silicon chips on which microprocessors are built.

This new way of doing biology has certainly become very trendy, but don't think that "trendy" translates into "lightweight" or "flash-in-the-pan." Bioinformatics goes way beyond trendy — it's at the center of the most recent developments in biology, such as the deciphering of the human genome (another buzzword), "system biology" (trying to look at the global picture), new biotechnologies, new legal and forensic techniques, as well as the personalized medicine of the future.

Because of the centrality of bioinformatics to cutting-edge developments in molecular biology, people from many different fields have been stumbling across the term in a variety of different contexts. If you're a biology, medical, or computer science student, a professional in the pharmaceutical industry, a lawyer or a policeman worrying about DNA testing, a consumer concerned about GMOs (Genetically Modified Organisms), or even a NASDAQ investor interested in start-up companies, you'll already have come across the word *bioinformatics.* If you're good at what you do, you'll want to know what all the fuss is about. This chapter, then, is for you.

Instead of a formal definition that would take hours to cover all the ins and outs of the topic, the best way to get a quick feel for what bioinformatics — or swimming, for that matter — is all about is to jump right into the water; that's what we do next. Go ahead and get your feet wet with some basic molecular biology concepts — and the relevant questions intimately connected with such concepts — that all together define bioinformatics.

Analyzing Protein Sequences

If you eat steak, you're intimately acquainted with proteins. (Your taste buds know them intimately anyway, even if your rational mind was too busy with dinner to master the concept.) For you non-steak lovers out there, you'll be pleased to know that proteins abound in fish and vegetables, too. Moreover, all these proteins are made up of the same basic building blocks, called *amino acids.* Amino acids are already quite complex organic molecules, made of carbon, hydrogen, oxygen, nitrogen, and sulfur atoms. So the overall recipe for a protein (the one your rational mind will appreciate, even if your taste buds won't) is something like $C_{1200}H_{2400}O_{600}N_{300}S_{100}$.