

Chapman & Hall/CRC
Mathematical and Computational Biology Series

Introduction to Bioinformatics



Anna Tramontano

Chapman & Hall/CRC
Taylor & Francis Group

Chapman & Hall/CRC Mathematical and Computational Biology Series

Introduction to Bioinformatics

Anna Tramontano



Chapman & Hall/CRC
Taylor & Francis Group

Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

Bioinformatica. Authorized translation from the Italian language edition published by Zanichelli.

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2007 by Zanichelli
Chapman & Hall/CRC is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works
Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 1-58488-569-6 (Softcover)
International Standard Book Number-13: 978-1-58488-569-6 (Softcover)

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC) 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Library of Congress Cataloging-in-Publication Data

Tramontano, Anna.
Introduction to Bioinformatics / by Anna Tramontano.
p. cm. -- (Mathematical and computational biology series)
Includes bibliographical references and index.
ISBN 1-58488-569-6
1. Bioinformatics. I. Title.

QH324.2.T73 2006
572.80285--dc22

2006049140

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>
and the CRC Press Web site at
<http://www.crcpress.com>

Chapman & Hall/CRC Mathematical and Computational Biology Series

Introduction to Bioinformatics

CHAPMAN & HALL/CRC

Mathematical and Computational Biology Series

Aims and scope:

This series aims to capture new developments and summarize what is known over the whole spectrum of mathematical and computational biology and medicine. It seeks to encourage the integration of mathematical, statistical and computational methods into biology by publishing a broad range of textbooks, reference works and handbooks. The titles included in the series are meant to appeal to students, researchers and professionals in the mathematical, statistical and computational sciences, fundamental biology and bioengineering, as well as interdisciplinary researchers involved in the field. The inclusion of concrete examples and applications, and programming techniques and examples, is highly encouraged.

Series Editors

Alison M. Etheridge
Department of Statistics
University of Oxford

Louis J. Gross
Department of Ecology and Evolutionary Biology
University of Tennessee

Suzanne Lenhart
Department of Mathematics
University of Tennessee

Philip K. Maini
Mathematical Institute
University of Oxford

Shoba Ranganathan
Research Institute of Biotechnology
Macquarie University

Hershel M. Safer
Weizmann Institute of Science
Bioinformatics & Bio Computing

Eberhard O. Voit
The Wallace H. Coulter Department of Biomedical Engineering
Georgia Tech and Emory University

Proposals for the series should be submitted to one of the series editors above or directly to:

CRC Press, Taylor & Francis Group

24-25 Blades Court
Deodar Road
London SW15 2NU
UK

Published Titles

Cancer Modelling and Simulation

Luigi Preziosi

Computational Biology: A Statistical Mechanics Perspective

Ralf Blossey

Computational Neuroscience: A Comprehensive Approach

Jianfeng Feng

Data Analysis Tools for DNA Microarrays

Sorin Draghici

Differential Equations and Mathematical Biology

D.S. Jones and B.D. Sleeman

Exactly Solvable Models of Biological Invasion

Sergei V. Petrovskii and Lian-Bai Li

Introduction to Bioinformatics

Anna Tramontano

An Introduction to Systems Biology: Design Principles of Biological Circuits

Uri Alon

Knowledge Discovery in Proteomics

Igor Jurisica and Dennis Wigle

Modeling and Simulation of Capsules and Biological Cells

C. Pozrikidis

Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems

Qiang Cui and Ivet Bahar

Stochastic Modelling for Systems Biology

Darren J. Wilkinson

The Ten Most Wanted Solutions in Protein Bioinformatics

Anna Tramontano

Dedication

*To my students and to my friend and mentor,
Maurizio Brunori, with unwavering respect and affection*

Preface

Bioinformatics is a relatively recent discipline. The term first appeared in scientific papers at the beginning of the 1990s, but this fact can be misleading. Already in the 1960s, when research laboratories were able to afford computers with good graphical performance, the scientific community had started to use them to analyze biological data. From the point of view of informatics and biology, substantial progress has been made since then.

Hardware has become more efficient; the speed and graphical performance of personal computers now are astonishing compared with those of just 10 years ago. The process of developing software is made easier and more straightforward every day. Another important aspect has been the development and spread of the World Wide Web, an instrument that has changed our conception of communications and made a deep impact on the scientific community. Furthermore, biological data have been exponentially accumulating, thanks to new, powerful techniques available in every laboratory.

Bioinformatics has synergistically exploited new technologies, giving rise to a new scientific discipline, with its own history and even some revolutions. We can define bioinformatics as the science that uses the instruments of informatics to analyze biological data in order to formulate hypotheses about life. However, a scientific definition needs to be operational rather than semantic, and the aim of this book is to describe bioinformatics methods and tools in terms of their ability to help us solve biological problems.

Nevertheless, before starting we must address frequently asked questions: “Who is a bioinformatician?” “What is his or her cultural background?” “Should a bioinformatician be proficient in programming or should he or she know biology in detail?” The confusion arises from the fact that we use the word bioinformatician to indicate expert users of the available tools as well as developers of new and more powerful methods. Of course the background in the two cases could and should not be the same.

In the first case, it is important to have a good biological knowledge to be able to understand the results of bioinformatics analyses and, at the same time, to write simple programs. In the second, expertise in statistical methods, algorithms, and programming and a basic biological knowledge are required. In both cases, though, it is essential to understand biological problems and methods and the rational basis of available bioinformatics tools. This is a must if we wish to use them correctly or improve them.

Therefore, the aim of this book is to describe the rationale and the limitations of the methods and tools available to the biological community at large. It is directed to students who want to have an idea about what bioinformatics is before deciding whether it is worth getting deeper into the subject and to those who, having decided

to pursue a career in experimental biology, want to have a grasp of the methods they will undoubtedly need during their research.

The questions that we will address concern ways of storing and (more importantly) retrieving the enormous amount of biological data produced every day (Chapter 1) and the methods to decrypt the information encoded by a genome (Chapter 2), to detect and exploit the evolutionary and functional relationships among biological elements (Chapters 3, 4, and 5), and to predict the three-dimensional structure of a protein (Chapters 6, 7, 8, and 9).

Chapter 10 offers a window to what the future holds, although in such a young and quickly evolving field as bioinformatics, we have learned that it is hard to predict what will come next, even in the very near future. This is proving to be even more difficult than predicting the structure and the function of a biological macromolecule!

The future will challenge us with new methodologies for tackling new and old problems, but some fundamental aspects will not change. We will always need to apply new methods to the same types of biological data, to implement them efficiently, and, most of all, to be aware of the power and limitations of these methods, in order to evaluate the meaningfulness of their results and extract information useful to solving biological problems.

Note: At the end of most chapters is a list of problems. Some of the input data for the problems can be downloaded in electronic format from the publisher's Web site, www.crcpress.com.

Author

Anna Tramontano studied physics at the University of Naples, Italy. She continued her research at the University of California San Francisco and became a staff scientist in the biocomputing programme of the European Molecular Biology Laboratory (EMBL) in Heidelberg. In 1990, Dr. Tramontano returned to Italy to work at the Merck Research Laboratories near Rome. In 2001, she returned to academia as Chair and Professor of Biochemistry at La Sapienza University, Rome where she continues today to pursue research in protein structure prediction and analysis.

Dr. Tramontano is a member of the European Molecular Biology Organization, the Scientific Council of Institute Pasteur-Fondazione Cenci Bolognetti, and serves on the organizing committee of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) initiative. She is the director of two master's programs in bioinformatics, teaches at several universities, and is the coordinator in the European Permanent School in Bioinformatics.

Acknowledgments

While writing this book, I asked the advice of and help from many colleagues and friends. Special thanks go to Henriette Molinari, Arthur Lesk, Angelo Sironi, Armin Lahm, Sergio Ammendola, Renzo Bazzo, Valentina Cappello, Andrea Sbardellati, Domenico Cozzetto, and Adriana Miele, whose suggestions have been fundamental to the creation of the book. When there are errors, it is my fault; when there are none, it is thanks to them. Special thanks go to Domenico Raimondo and Alejandro Giorgetti, who preferred the challenge of the pixels and formats of several of the pictures in this book to the breathtaking beauty of the beaches on the island of Sardinia!

Table of Contents

Chapter 1	The Data: Storage and Retrieval.....	1
Glossary		1
1.1	Basic Principles.....	2
1.2	The Data.....	3
1.2.1	The Postgenomic Era.....	3
1.2.2	Nucleic Acid Data Banks.....	6
1.2.3	Protein Sequence Data Banks.....	7
1.2.4	Protein Structure Databases.....	14
1.2.5	Protein Interaction Databases.....	16
1.2.6	Derived Data Banks.....	17
1.2.7	Integration of Databases.....	18
1.3	Data Quality.....	18
1.4	Data Representation.....	18
1.4.1	Protein Architecture.....	20
References.....		24
Problems		26
Chapter 2	Genome Sequence Analysis.....	29
Glossary		29
2.1	Basic Concepts.....	30
2.2	Genome Sequencing.....	31
2.3	Finding the Genes.....	32
2.4	Statistical Methods to Search for Genes.....	34
2.4.1	Site-Specific Scoring Matrices.....	34
2.4.2	Artificial Neural Networks.....	37
2.4.3	Markov Models and Hidden Markov Models.....	40
2.4.4	Levels of Reliability.....	45
2.5	Comparative Genomics.....	47
2.6	A Virtual Window on Genomes: The World Wide Web.....	48
References.....		49
Problems		50
Chapter 3	Protein Evolution.....	53
Glossary		53
3.1	Basic Concepts.....	53
3.2	Molecular Evolution.....	54

3.3	How to Align Two Similar Sequences	55
3.4	Similarity Matrices.....	58
3.4.1	PAM Matrices	59
3.4.2	BLOSUM Matrices.....	60
3.5	Penalties for Insertions and Deletions.....	61
3.6	The Alignment Algorithm.....	61
3.7	Multiple Alignments	66
3.7.1	How to Add Sequences to a Pre-Existing Alignment.....	70
3.8	Phylogenetic Trees	71
	References.....	73
	Problems	74

Chapter 4 Similarity Searches in Databases..... 77

	Glossary	77
4.1	Basic Principles.....	77
4.2	The Methods	78
4.2.1	FASTA.....	79
4.2.2	BLAST	83
4.2.3	Profile Searches.....	86
4.2.4	PSI-BLAST	87
	References.....	89
	Problems	90

Chapter 5 Amino Acid Sequence Analysis 93

	Glossary	93
5.1	Basic Principles.....	93
5.1.1	Is It Really an Orphan Sequence?.....	94
5.2	Search for Sequence Patterns	95
5.3	Feature Extraction.....	96
5.4	Secondary Structure: Part One	96
	References.....	98
	Problems	99

Chapter 6 Prediction of the Three-Dimensional Structure of a Protein 103

	Glossary	103
6.1	Basic Principles.....	103
6.2	The CASP Experiment.....	105
6.3	Secondary Structure Prediction: Part Two.....	107
6.4	Long-Range Contact Prediction.....	109
6.5	Predicting Molecular Complexes: Docking Methods	110
	References.....	111
	Problems	113

Chapter 7	Homology Modeling	115
Glossary		115
7.1	Basic Principles	115
7.2	The Steps of Comparative Modeling	117
7.2.1	Template Selection	118
7.2.2	Sequence Alignment	119
7.2.3	Loops	121
7.2.4	Side Chain Modeling	124
7.2.5	Model Optimization	125
7.3	Accuracy of Homology Models	126
7.4	Manual versus Automatic Models	126
7.5	Practical Notes	126
7.6	Summing Up...	129
References		131
Problems		134
Chapter 8	Fold Recognition Methods	137
Glossary		137
8.1	Basic Principles	137
8.2	Profile-Based Methods	138
8.3	Threading Methods	140
8.4	The Fold Library	142
8.5	How Well Do These Methods Work?	142
References		143
Problems		143
Chapter 9	New Fold Modeling	147
Glossary		147
9.1	Basic Principles	147
9.2	Estimating the Energy of a Protein Conformation	148
9.3	Energy Minimization	150
9.4	Molecular Dynamics	151
9.4.1	The Monte Carlo Method	153
9.4.2	Genetic Algorithms	154
9.4.3	Combined Methodologies	156
References		156
Problems		157
Chapter 10	The “Omics” Universe	159
10.1	Basic Principles	159
10.2	Transcriptomics	159

10.3 Proteomics..... 162

10.4 Interactomics 165

10.5 Structural Genomics..... 166

10.6 Pharmacogenomics..... 166

10.7 But This Is Not All... 168

Useful Web Sites 169

Index 171

1 The Data: Storage and Retrieval

GLOSSARY

cDNA: complementary DNA; a DNA molecule obtained by retrotranscribing an mRNA molecule into DNA

Constraints: restrictions of the possible values taken by a parameter, such as a distance, an angle, or a solid angle

Data bank/database: collection of information stored in a systematic way that can be accessed electronically and searched by various parameters

DNA (RNA) polymerase: enzyme responsible for catalyzing the synthesis of a new molecule of DNA (RNA) using a pre-existing template filament

Electrophoresis: a technique that allows the separation of charged compounds in an electric field

Entry: element of a database

EST: expressed sequence tags; DNA sequence obtained from the (partial) sequencing of a cDNA molecule

Hybridization: the process by which two complementary single-strand oligonucleotides associate

Nitrogenous bases: nitrogenous compounds (purines or pyrimidines) found in nucleosides, nucleotides, and nucleic acids

NMR: nuclear magnetic resonance; a technique that uses the interactions of nuclei with an external magnetic field to reconstruct their position in space, hence the structure of the molecule

Primer: RNA or DNA fragment complementary to a portion of the DNA region to be synthesized by the DNA polymerase

Ramachandran plot: plot that shows the theoretically allowed or experimentally observed combinations of ϕ and ψ angles in a polypeptide chain

Resolution, R factor, Rfree factor: parameters to evaluate the accuracy of the reconstruction of a macromolecular structure starting from x-ray diffraction data

Sequence pattern: a pattern of amino acids deemed to have a functional significance

SNP: single nucleotide polymorphism; naturally occurring variants that affect a single nucleotide (A, T, C, or G) in a genome

X-ray crystallography: technique that uses x-ray diffraction for the reconstruction of the three-dimensional positions of atoms inside molecular crystals

1.1 BASIC PRINCIPLES

Bioinformatics is a relatively young discipline that deals with the storage, retrieval, and analysis of biological data with informatics tools. Many branches of science use computers, databases, and algorithms, from weather forecasts to economics, from physics to linguistics. Each of them treats the data in different ways dictated by the nature of the data. From this perspective, we could define bioinformatics as the science that analyzes biological data with computer tools in order to formulate hypotheses on the processes underlying life.

Despite still being a more qualitative than quantitative science, modern biology has given bioinformatics a powerful push. Thanks to the development of new revolutionary experimental techniques, biological data have accumulated (and keep doing so) at an impressive pace. For example, we have available sequences of hundreds of genomes and data on the expression of thousands of genes in many cell types, and structural genomics projects are producing thousands of three-dimensional structures of proteins every year.

In first approximation, we can divide biological data into three main categories: sequence data, structural data, and functional data. The nature of the data and how their peculiar characteristics influence the organization of the databases where they are collected are discussed in this chapter.

A data bank should store data as they have originally been deposited so that they can be analyzed or reanalyzed with new or improved techniques at any time. These databases are usually called primary databases.

It is often useful to compute some properties of frequently used data and store them in different “derived” databases. This avoids the problem of repeating the same analysis, but it also implies that the derived database needs to be updated every time the primary data are updated. Ideally, this should happen in real time.

Biological databases contain different types of data, but, by and large, they refer to the same biological entities (genes or proteins). Therefore, it is essential to have instruments to connect and integrate the information contained in all biological databases and to allow the user to navigate easily from one to the other.

In this chapter, we will briefly discuss primary and derived databases as well as systems to integrate their contents. However, these are not static systems—they change to fit novel needs or include new types of data when they become available. Therefore, we will describe the main databases, which are not expected to change too much (at least in their basic principles). However, the reader should frequently consult the many available Internet resources that list databases and their developments, such as the home pages of the NCBI (National Center for Biological Information) and of the EBI (European Bioinformatics Institute).

For the same reason, the proposed solutions for the exercises at the end of this and other chapters can be used today to solve the problems, but they are not necessarily unique and we cannot guarantee that they remain the fastest route to the answer even in the near future.