

Computational Biology

Editors-in-Chief

Andreas Dress
University of Bielefeld (Germany)

Martin Vingron
Max Planck Institute for Molecular Genetics (Germany)

Editorial Board

Gene Myers, Janelia Farm Research Campus, Howard Hughes Medical Institute (USA)
Robert Giegerich, University of Bielefeld (Germany)
Walter Fitch, University of California, Irvine (USA)
Pavel A. Pevzner, University of California, San Diego (USA)

Advisory Board

Gordon Grippen, University of Michigan (USA)
Joe Felsenstein, University of Washington (USA)
Dan Gusfield, University of California, Davis (USA)
Sorin Istrail, Brown University, Providence (USA)
Samuel Karlin, Stanford University (USA)
Thomas Lengauer, Max Planck Institut Informatik (Germany)
Marcella McClure, Montana State University (USA)
Martin Nowak, Harvard University (USA)
David Sankoff, University of Ottawa (Canada)
Ron Shamir, Tel Aviv University (Israel)
Mike Steel, University of Canterbury (New Zealand)
Gary Stormo, Washington University Medical School (USA)
Simon Tavaré, University of Southern California (USA)
Tandy Warnow, University of Texas, Austin (USA)

The Computational Biology series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

Author guidelines: [springer.com](http://www.springer.com) > Authors > Author Guidelines

For other titles published in this series, go to
<http://www.springer.com/series/5769>

Jeremy J. Ramsden

Bioinformatics

An Introduction

Second edition

 Springer

Jeremy J. Ramsden
Cranfield University
School of Applied Sciences
Bedfordshire, UK
j.ramsden@cranfield.ac.uk

Computational Biology Series ISSN 1568-2684
ISBN 978-1-84800-256-2 e-ISBN 978-1-84800-257-9
DOI 10.1007/978-1-84800-257-9

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2009920488

© Springer-Verlag London Limited 2009

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers.

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

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Printed on acid-free paper

Springer Science+Business Media
springer.com

*Mi a tudvágyat szakhoz nem kötők,
Átpillantását vágyuk az egésznek.*

IMRE MADÁCH

Preface to the Second Edition

Overview and aims. This book is intended as a self-contained guide to the entire field of bioinformatics, interpreted as the application of information science to biology. There is strong underlying belief that information is a profound concept underlying biology, and familiarity with the concepts of information should make it possible to gain many important new insights into biology. In other words, the vision underpinning this book goes beyond the narrow interpretation of bioinformatics sometimes encountered, which may confine itself to specific tasks such as the attempted identification of genes in a DNA sequence.

Organization and features. The chapters are grouped into three parts, covering the relevant fundamentals of information science, overviewing all of biology, and surveying applications. Thus Part I (fundamentals) carefully explains what information is, and discusses attributes such as value and quality, and its multiple significations of accuracy, meaning, and effect. The transmission of information through channels is described. Brief summaries of the necessary elements of set theory, combinatorics, probability, likelihood, clustering, and pattern recognition are given. Concepts such as randomness, complexity, systems and networks, needed for the understanding of biological organization, are also discussed. Part II (biology) covers both organismal (ontogeny and phylogeny, as well as genome structure) and molecular aspects. Part III (applications) is devoted to the most important practical applications of bioinformatics, notably gene identification, transcriptomics, proteomics, interactomics (dealing with networks of interactions), and metabolomics. These chapters start with a discussion of the experimental aspects (such as DNA sequencing in the genomics chapter), and then move on to a thorough discussion of how the data is analysed. Specifically medical applications are grouped in a separate chapter. A number of problems are suggested, many of which are open-ended and intended to stimulate further thinking. The bibliography points to specialized monographs and review articles expanding on material in the text, and includes guide references to very recently reported research not yet to be found in reviews.

Target audiences. This book is primarily intended as a textbook for undergraduates, for whom it aims to be a complete study companion. As such, it will also be useful to the beginning graduate student.

A secondary audience are physical scientists seeking a comprehensive but succinct guide to biology, and biological scientists wishing to better acquaint them-

selves with some of the physicochemical and mathematical aspects that underpin the applications.

It is hoped that all readers will find that even familiar material is presented with fresh insight, and will be inspired to new thoughts.

The author takes this opportunity to thank all those who gave him their comments on the first edition.

May 2008

Preface

This little book attempts to give a self-contained account of bioinformatics, so that the newcomer to the field may, whatever his point of departure, gain a rather complete overview. At the same time it makes no claim to be comprehensive: The field is already too vast—and let it be remembered that although its recognition as a distinct discipline (i.e., one after which departments and university chairs are named) is recent, its roots go back a long time.

Given that many of the newcomers arrive from either biology or informatics, it was an obvious consideration that for the book to achieve its aim of completeness, large portions would have to deal with matter already known to those with backgrounds in either of those two fields; that is, in the particular chapters dealing with them, the book would provide no information for them. Since such chapters could hardly be omitted, I have tried to consider such matter in the light of bioinformatics as a whole, so that even the student ostensibly familiar with it could benefit from a fresh viewpoint.

In one regard especially, this book cannot be comprehensive. The field is developing extraordinarily rapidly and it would have been artificial and arbitrary to take a snapshot of the details of contemporary research. Hence I have tried to focus on a thorough grounding of concepts, which will enable the student not only to understand contemporary work but should also serve as a springboard for his or her own discoveries. Much of the raw material of bioinformatics is open and accessible to all via the internet, powerful computing facilities are ubiquitous, and we may be confident that vast tracts of the field lie yet uncultivated. This accessibility extends to the literature: Research papers on any topic can usually be found rapidly by an internet search and, therefore, I have not aimed at providing a comprehensive bibliography.

In bioinformatics, so much is to be done, the raw material to hand is already so vast and vastly increasing, and the problems to be solved are so important (perhaps the most important of any science at present), we may be entering an era comparable to the great flowering of quantum mechanics in the first three decades of the twentieth century, during which there were periods when practically every doctoral thesis was a major breakthrough. If this book is able to inspire the student to take up some of the challenges, then it will have accomplished a large part of what it sets out to do.

Indeed, I would go further to remark that I believe that there are still comparatively simple things to be discovered and that many of the present directions of work in the field may turn out not to be right. Hence, at this stage in its development the most important thing is to illuminate that viewpoint that will facilitate new discoveries. This belief also underlies the somewhat more detailed coverage of the biological processes in which information processing in nature is embodied than might be considered customary.

A work of this nature depends on a long history of interactions, discussions, and correspondence with many present and erstwhile friends and colleagues, some of whom, sadly, are no longer alive. I have tried to reflect some of this debt in the citations. Furthermore, many scientific subjects and methods other than those mentioned in the text had to be explored before the ones best suited to the purpose of this work could be selected, and my thanks are due to all those who helped in these preliminary studies. I should like to add an especial word of thanks to Victoria Kechehmadze for having so ably drawn the figures.

January 2004

Contents

1 Introduction	1
1.1 What is Bioinformatics?	2
1.2 What Can Bioinformatics Do?	3

Part I Information

2 The Nature of Information	9
2.1 Structure and Quantity	15
2.1.1 The Generation of Information	15
2.1.2 Conditional and Unconditional Information	15
2.1.3 Experiments and Observations	16
2.2 Constraint	17
2.2.1 The Value of Information	22
2.2.2 The Quality of Information	23
2.3 Accuracy, Meaning, and Effect	23
2.3.1 Accuracy	23
2.3.2 Meaning	24
2.3.3 Effect	27
2.3.4 Significs	28
2.4 Further Remarks on Information Generation	28
2.5 Summary	29
3 The Transmission of Information	31
3.1 The Capacity of a Channel	33
3.2 Coding	35
3.3 Decoding	37
3.4 Compression	38
3.4.1 Use of Compression to Measure Distance	41
3.4.2 Ergodicity	41
3.5 Noise	42

3.6	Error Correction	44
3.7	Summary	46
4	Sets and Combinatorics	47
4.1	The Notion of Set	47
4.2	Combinatorics	47
4.2.1	Ordered Sampling With Replacement	48
4.2.2	Ordered Sampling Without Replacement	48
4.2.3	Unordered Sampling Without Replacement	49
4.2.4	Unordered Sampling With Replacement	51
4.3	The Binomial Theorem	51
5	Probability and Likelihood	53
5.1	The Notion of Probability	53
5.2	Fundamentals	54
5.2.1	Generalized Union	56
5.2.2	Conditional Probability	57
5.2.3	Bernoulli Trials	59
5.3	Moments of Distributions	61
5.3.1	Runs	62
5.3.2	The Hypergeometric Distribution	63
5.3.3	Multiplicative Processes	64
5.4	Likelihood	65
5.5	The Maximum Entropy Method	68
6	Randomness and Complexity	69
6.1	Random Processes	72
6.2	Markov Chains	73
6.3	Random Walks	75
6.4	Noise	77
6.5	Complexity	78
7	Systems, Networks, and Circuits	83
7.1	General Systems Theory	84
7.1.1	Automata	86
7.1.2	Cellular Automata	88
7.1.3	Percolation	88
7.2	Networks (graphs)	89
7.2.1	Trees	91
7.2.2	Complexity Parameters	92
7.2.3	Dynamical Properties	92
7.3	Synergetics	93
7.3.1	Some Examples	94
7.3.2	Reception and Generation of Information	96
7.4	Evolutionary Systems	96

8 Algorithms 99

8.1 Evolutionary Computing 100

8.2 Pattern Recognition 101

8.3 Botryology 102

8.3.1 Clustering 103

8.3.2 Principal Component and Linear
Discriminant Analyses 105

8.3.3 Wavelets 106

8.4 Multidimensional Scaling and Seriation 107

8.5 Visualization 110

Part II Biology

9 Introduction to Part II 115

9.1 Genotype, Phenotype, and Species 115

9.2 Adaptation 117

9.3 Timescales of Adaptation 118

9.3.1 The Rôle of Memory 119

9.3.2 The Integrating Rôle of Directive Correlation 119

9.4 Regulation 120

9.5 The Concept of Machine 121

9.6 The Architecture of Functional Systems 122

10 The Nature of Living Things 123

10.1 The Cell 123

10.1.1 The Structure of a Cell 125

10.1.2 Observational Overview 125

10.2 Metabolism 127

10.3 The Cell Cycle 128

10.3.1 The Chromosome 130

10.3.2 The Structure of Genome and Genes 133

10.3.3 The C-Value Paradox 136

10.3.4 The Structure of the Chromosome 139

10.4 The Immune System 140

10.5 Molecular Mechanisms 141

10.5.1 Replication 141

10.5.2 Proofreading and Repair 142

10.5.3 Recombination 143

10.5.4 Summary of Sources of Genome Variation 145

10.6 Gene Expression 145

10.6.1 Transcription 146

10.6.2 Regulation of Transcription 146

10.6.3 Prokaryotic Transcriptional Regulation 147

10.6.4 Eukaryotic Transcriptional Regulation 147

10.6.5	mRNA Processing	149
10.6.6	Translation	150
10.7	Ontogeny (Development)	151
10.7.1	Stem Cells	152
10.7.2	Epigenesis	153
10.7.3	r and K Selection	154
10.7.4	Homeotic Genes	155
10.8	Phylogeny and Evolution	155
10.8.1	Models of Evolution	158
10.8.2	Sources of Genome Variation	160
10.8.3	The Origin of Proteins	160
10.8.4	Geological Eras and Taxonomy	161
11	The Molecules of Life	163
11.1	Molecules and Supramolecular Structure	163
11.2	Water	165
11.3	DNA	166
11.4	RNA	171
11.5	Proteins	172
11.5.1	Amino Acids	173
11.5.2	Protein Folding and Interaction	175
11.5.3	Experimental Techniques for Protein Structure Determination	178
11.5.4	Protein Structure Overview	179
11.6	Polysaccharides	179
11.7	Lipids	180
 Part III Applications		
12	Introduction to Part III	185
13	Genomics	189
13.1	DNA Sequencing	190
13.1.1	Extraction of Nucleic Acids	190
13.1.2	The Polymerase Chain Reaction	191
13.1.3	Sequencing	191
13.1.4	Expressed Sequence Tags	192
13.2	DNA Methylation Profiling	193
13.3	Gene Identification	193
13.4	Extrinsic Methods	194
13.4.1	Database Reliability	194
13.4.2	Sequence Comparison and Alignment	194
13.4.3	Dynamic Programming Algorithms	196

13.5	Intrinsic Methods	197
13.5.1	Signals	198
13.5.2	Hidden Markov Models	199
13.6	Beyond Sequence	199
13.7	Minimalist Approaches	200
13.7.1	Nucleotide Frequencies	200
13.7.2	Word Occurrences	201
13.8	Phylogenies	202
14	Proteomics	205
14.1	Transcriptomics	206
14.1.1	Limitations	210
14.2	Proteomics	211
14.2.1	Two-Dimensional Gel Electrophoresis	212
14.2.2	Column Chromatography	213
14.2.3	Other Kinds of Electrophoresis	214
14.3	Protein Identification	214
14.4	Isotope-Coded Affinity Tags	215
14.5	Protein Microarrays	216
14.6	Protein Expression Patterns	217
14.7	The Kinome	218
15	Interactomics: Interactions and Regulatory Networks	221
15.1	Inference of Regulatory Networks	225
15.2	The Physical Chemistry of Interactions	226
15.3	Intermolecular Interactions	228
15.3.1	Time-Dependent Rate “Constants”	229
15.3.2	Specificity	230
15.3.3	Nonspecific Interactions	230
15.3.4	Cooperative Binding	230
15.3.5	Sustained Activation	231
15.4	<i>In vivo</i> Experimental Methods	232
15.4.1	The Yeast Two-Hybrid Assay	232
15.4.2	Crosslinking	233
15.4.3	Correlated Expression	233
15.4.4	Other Methods	234
15.5	<i>In vitro</i> Experimental Methods	234
15.5.1	Chromatography	235
15.5.2	Direct Affinity Measurement	235
15.5.3	Protein Chips	237
15.6	Interactions from Sequence	237
15.7	Global Statistics of Interactions	238

16	Metabolomics and Metabonomics	239
16.1	Data Collection	240
16.2	Data Analysis	241
16.3	Metabolic Regulation	242
	16.3.1 Metabolic Control Analysis	242
	16.3.2 The Metabolic Code	243
16.4	Metabolic Networks	243
17	Medical Applications	245
17.1	The Genetic Basis of Disease	246
17.2	Cancer	247
17.3	Toward Automated Diagnosis	249
17.4	Drug Discovery and Testing	249
17.5	Personalized Medicine	251
18	The Organization of Knowledge	253
18.1	Ontology	254
18.2	Knowledge Representation	255
18.3	The Problem of Bacterial Identification	256
18.4	Text Mining	257
	Bibliography	259
	Index	267