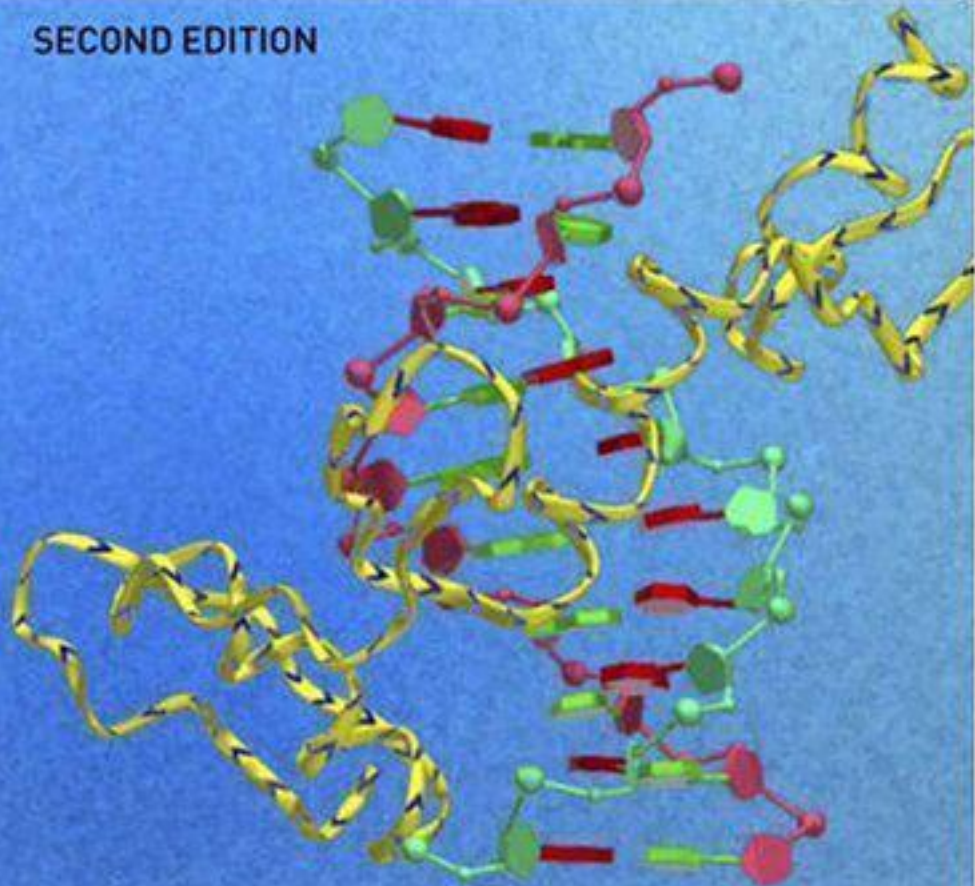


OXFORD

INTRODUCTION TO BIOINFORMATICS

SECOND EDITION



ARTHUR M. LESK

Introduction to **Bioinformatics**

SECOND EDITION

Arthur M. Lesk

The Pennsylvania State University

In nature's infinite book of secrecy
A little I can read.
- Antony and Cleopatra

OXFORD
UNIVERSITY PRESS

CHALMERS TEKNISKA
HÖGSKOLAS BIBLIOTEK

2005

OXFORD

UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in
Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan South Korea Poland Portugal
Singapore Switzerland Thailand Turkey Ukraine Vietnam

Published in the United States
by Oxford University Press Inc., New York

© Arthur M. Lesk 2005

The moral rights of the author have been asserted

Database right Oxford University Press (maker)

First published 2005

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other **binding** or cover
and you must impose this same condition on **any acquirer**

British Library Cataloguing in Publication Data

Data available

Library of Congress Cataloging in Publication Data

Data available

ISBN 0 19 9277877 (hbk)

10 9 8 7 6 5 4 3 2 1

Typeset by Newgen Imaging Systems (P) Ltd., Chennai, India

Printed in Great Britain

on acid-free paper by Ashford Colour Press Limited, Gosport, Hampshire

Dedicated to Eda, with whom I have merged my genes.

Contents

Plan of the book *xix*

1 Introduction 1

Life in space and time 3

Evolution is the change over time in the world of living things 4

Dogmas: central and peripheral 6

Observables and data archives 9

Information flow in bioinformatics 12

Curation, annotation, and quality control 13

The World Wide Web 14

Electronic publication 15

Computers and computer science 16

Programming 17

Biological classification and nomenclature 21

Use of sequences to determine phylogenetic relationships 24

Use of SINES and LINES to derive phylogenetic relationships 30

Searching for similar sequences in databases: PSI-BLAST 32

Introduction to protein structure 40

The hierarchical nature of protein architecture 41

Classification of protein structures 44

Protein structure prediction and engineering 51

Critical Assessment of Structure Prediction (CASP) 52

Protein engineering 52

Proteomics 52

DNA microarrays 53

Mass spectrometry 54

Systems biology 54

Clinical implications 55

The future 57

Recommended reading 57

Exercises, Problems, and Weblems 59

2 Genome organization and evolution 67

Genomes and proteomes 68

Genes 69

Proteomes 71

Eavesdropping on the transmission of genetic information 72

Mappings between the maps 77

High-resolution maps 78

Picking out genes in genomes 80**Genomes of prokaryotes** 81The genome of the bacterium *Escherichia coli* 82The genome of the archaeon *Methanococcus jannaschii* 85The genome of one of the simplest organisms: *Mycoplasma genitalium* 86**Genomes of eukaryotes** 87The genome of *Saccharomyces cerevisiae* (baker's yeast) 89The genome of *Caenorhabditis elegans* 93The genome of *Drosophila melanogaster* 94The genome of *Arabidopsis thaliana* 95**The genome of *Homo sapiens* (the human genome)** 96

Protein coding genes 97

Repeat sequences 99

RNA 100

Single-nucleotide polymorphisms (SNPs) 101**Genetic diversity in anthropology** 102

Genetic diversity and personal identification 103

Genetic analysis of cattle domestication 104

Evolution of genomes 104

Please pass the genes: horizontal gene transfer 108

Comparative genomics of eukaryotes 109

Recommended reading 111*Exercises, Problems, and Weblems* 112**3 Archives and information retrieval** 117**Introduction** 118

Database indexing and specification of search terms 118

Follow-up questions 120

Analysis of retrieved data 121

The archives 121

Nucleic acid sequence databases 122

Genome databases 124

Protein sequence databases 124

Databases of structures 128

Specialized, or 'boutique' databases 135

Expression and proteomics databases 136

Databases of metabolic pathways 138

Bibliographic databases 139

Surveys of molecular biology databases and servers 139

Gateways to archives 140

Access to databases in molecular biology 141

ENTREZ 141
The Sequence Retrieval System (SRS) 148
The Protein Identification Resource (PIR) 149
ExPASy—Expert Protein Analysis System 150
Ensembl 151

Where do we go from here? 152

Recommended reading 152

Exercises, Problems, and Weblems 153

4 Alignments and phylogenetic trees 157

Introduction to sequence alignment 158

The dotplot 160

Dotplots and sequence alignments 165

Measures of sequence similarity 171

Scoring schemes 171

Computing the alignment of two sequences 175

Variations and generalizations 175

Approximate methods for quick screening of databases 176

The dynamic programming algorithm for optimal pairwise sequence alignment 176

Significance of alignments 182

Multiple sequence alignment 186

Applications of multiple sequence alignments to database searching 188

Profiles 189

PSI-BLAST 191

Hidden Markov Models 193

Phylogeny 198

Phylogenetic trees 203

Clustering methods 205

Cladistic methods 206

The problem of varying rates of evolution 207

Computational considerations 208

Recommended reading 209

Exercises, Problems, and Weblems 210

5 Protein structure and drug discovery 219

Introduction 220

Protein stability and folding 223

The Sasisekharan-Ramakrishnan-Ramachandran plot describes
allowed mainchain conformations 223

The sidechains 225

Protein stability and denaturation 225

Protein folding 228

Applications of hydrophobicity 229

Superposition of structures, and structural alignments 233

DALI (Distance-matrix ALignment) 235

- Evolution of protein structures** 236
- Classifications of protein structures** 238
 - SCOP 239
- Protein structure prediction and modelling** 240
 - Critical Assessment of Structure Prediction (CASP) 242
 - Secondary structure prediction 244
 - Homology modelling 250
 - Fold recognition 252
 - Conformational energy calculations and molecular dynamics 255
 - ROSETTA 259
 - LINUS 259
- Assignment of protein structures to genomes** 263
- Prediction of protein function** 265
 - Divergence of function: orthologues and paralogues 266
- Drug discovery and development** 269
 - The lead compound 271
 - Bioinformatics in drug discovery and development 273
- Recommended reading* 284
- Exercises, Problems, and Weblems* 285

6 Proteomics and systems biology 291

- DNA microarrays** 293
 - Analysis of microarray data 295
- Mass spectrometry** 301
 - Identification of components of a complex mixture 301
 - Protein sequencing by mass spectrometry 304
 - Genome sequence analysis by mass spectrometry 306
- Systems biology** 311
- Networks and graphs** 313
 - Network structure and dynamics 318
- Protein complexes and aggregates** 320
 - Properties of protein-protein complexes 321
- Protein interaction networks** 324
- Regulatory networks** 329
 - Structures of regulatory networks 330
 - Structural biology of regulatory networks 336
- Recommended reading* 339
- Exercises, Problems, and Weblems* 339

Conclusions 345

Answers to Exercises 347

Glossary 353

Index 357

Colour plates

Plan of the book

- ◆ Chapter 1 sets the stage and introduces all of the major players: DNA and protein sequences and structures, genomes and proteomes, databases and information retrieval, the World Wide Web, computer programming. Before developing individual topics in detail it is important to see the framework of their interactions.
- ◆ Chapter 2 presents the nature of individual genomes, including the Human Genome, and the relationships among them, from the biological point of view.
- ◆ Chapter 3 imparts basic skills in using the Web in bioinformatics. It describes archival databanks, and leads the reader through sample sessions involving information retrieval from some of the major archival databases in molecular biology.
- ◆ Chapter 4 treats the analysis of relationships among sequences—alignments and phylogenetic trees. These methods underlie some of the major computational challenges of bioinformatics: detecting distant relatives, understanding relationships among genomes of different organisms, and tracing the course of evolution at the species and molecular levels.
- ◆ Chapter 5 moves into three dimensions, treating protein structure and folding. Sequence and structure must be seen as full partners, with bioinformatics developing methods for moving back and forth between them as fluently as possible. Understanding protein structures in detail is essential for determining their mechanisms of action, and for clinical and pharmacological applications.
- ◆ Chapter 6 treats proteomics and systems biology, including new high-throughput sources of information about the expression and distribution of proteins in cells, and attempts to synthesize the information to reveal patterns of organization.