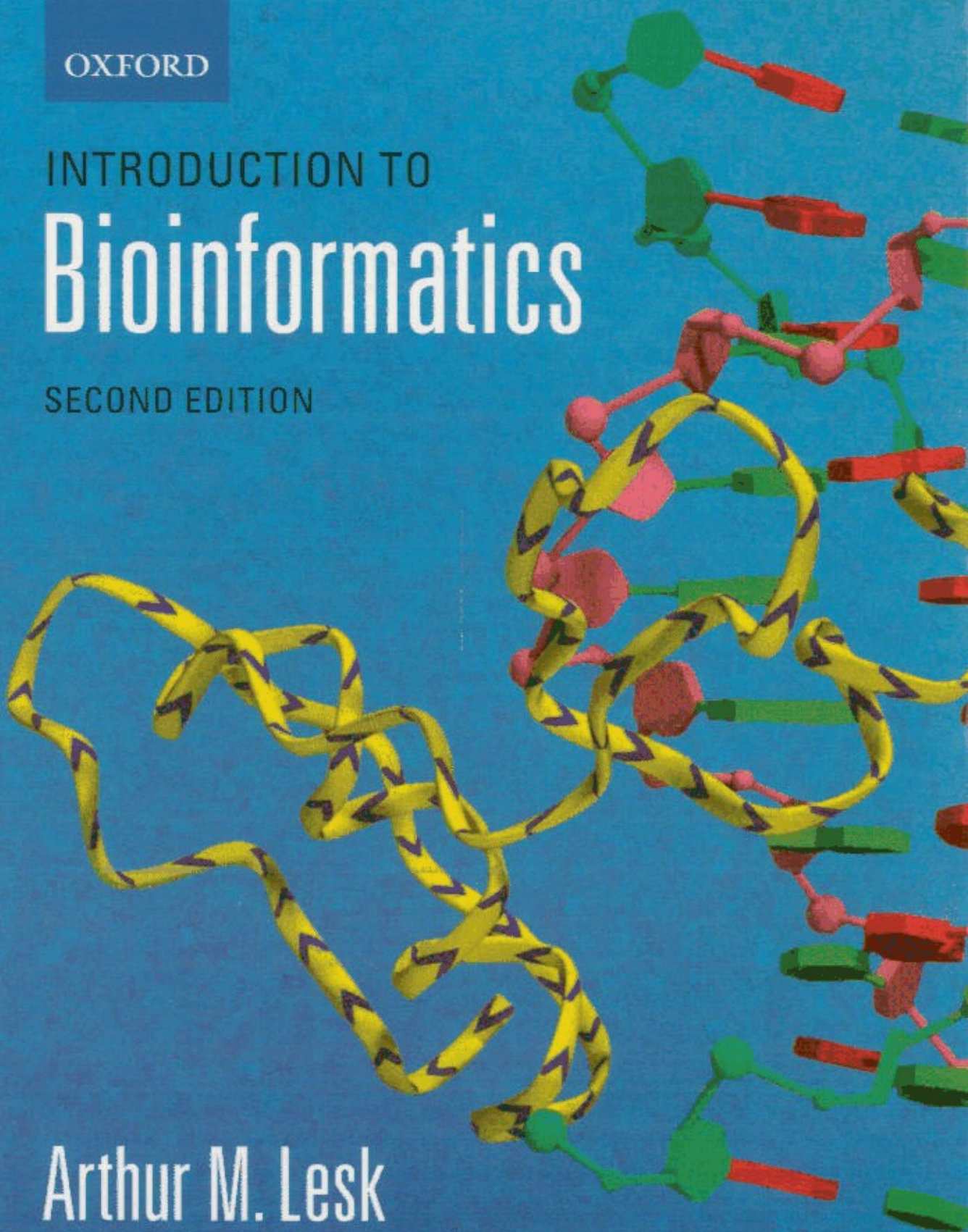


OXFORD

INTRODUCTION TO  
**Bioinformatics**

SECOND EDITION

Arthur M. Lesk



# 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 81**The 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 87**The 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

## CONTENTS

|   |     |
|---|-----|
| <b>Evolution of protein structures</b>                    | 236 |
| <b>Classifications of protein structures</b>              | 238 |
| SCOP  | 239 |
| <b>Protein structure prediction and modelling</b>         | 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 |
| <b>Assignment of protein structures to genomes</b>        | 263 |
| <b>Prediction of protein function</b>                     | 265 |
| Divergence of function: orthologues and paralogues        | 266 |
| <b>Drug discovery and development</b>                     | 269 |
| The lead compound   | 271 |
| Bioinformatics in drug discovery and development          | 273 |
| <i>Recommended reading</i>                                | 284 |
| <i>Exercises, Problems, and Weblems</i>                   | 285 |

## 6 Proteomics and systems biology 291

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

|                             |     |
|-----------------------------|-----|
| <b>Conclusions</b>          | 345 |
| <b>Answers to Exercises</b> | 347 |
| <b>Glossary</b>             | 353 |
| <b>Index</b>                | 357 |
| <b>Colour plates</b>        |     |