

# Contents

## 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 22

Use of SINES and LINES to derive phylogenetic relationships 31

Searching for similar sequences in databases: PSI-BLAST 33

Introduction to protein structure 42

The hierarchical nature of protein architecture 43

Classification of protein structures 46

Protein structure prediction and engineering 52

Critical Assessment of Structure Prediction (CASP) 53

Protein engineering 53

Proteomics 54

DNA microarrays 54

Systems biology 56

Clinical implications 56

The future 59

*Recommended reading* 59

*Exercises, Problems and Weblems* 61

## 2 Genome organization and evolution 69

Genomes and proteomes 70

Genes 70

Proteomics 73

Eavesdropping on the transmission of genetic information 74

Identification of genes associated with inherited diseases 76

Mappings between the maps 77

- High-resolution maps** 80
- Picking out genes in genomes** 82
- Genome sequencing projects** 83
  - Genomes on the Web 84
- Genomes of prokaryotes** 86
  - The genome of the bacterium *Escherichia coli* 87
  - The genome of the archaeon *Methanococcus jannaschii* 90
  - The genome of one of the simplest organisms: *Mycoplasma genitalium* 91
- Genomes of eukarya** 92
  - Gene families 95
  - The genome of *Saccharomyces cerevisiae* (Baker's yeast) 97
  - The genome of *Caenorhabditis elegans* 98
  - The genome of *Drosophila melanogaster* 101
  - The genome of *Arabidopsis thaliana* 101
- The genome of *Homo sapiens* (The Human Genome)** 104
  - Protein-coding genes 104
  - Repeat sequences 107
  - RNA 108
  - Single-nucleotide polymorphisms (SNPs) and haplotypes 108
  - Systematic measurements and collections of SNPs 111
- Genetic diversity in anthropology** 113
  - Genetic diversity and personal identification 116
- Evolution of genomes** 116
  - Please pass the genes: horizontal gene transfer 121
- Comparative genomics of eukarya** 123
  - The ENCODE project 124
- Metagenomics: the collection of genomes in a coherent environmental sample** 128
  - Recommended reading* 131
  - Exercises, Problems and Weblems* 132

### **3 Scientific publications and archives: media, content and access** 137

- The scientific literature** 138
- Economic factors governing access to scholarly publications** 139
  - Open access 141
  - The Public Library of Science (PLOS) 142
- Traditional and digital libraries** 143
  - How to populate a digital library 144
- The information explosion** 145
  - The Web—higher dimensions 145
  - New media—video, sound 146
  - Searching the literature 146
  - Bibliography management 147

**Databases** 148

- Database contents 148
- The literature as a database 149
- Organization 149
- Annotation 152
- Database quality control 153
- Database access 156
- Links 157
- Database interoperability 159
- Data mining 162

**Programming languages and tools** 165

- Traditional programming languages 166
- Scripting languages 166
- Program libraries specialized for molecular biology 167
- Java—computing over the Web 167
- Markup languages 168

**Natural language processing** 170

- Identifying keywords and combinations of keywords 171
- Knowledge extraction: protein–protein interactions 172
- Applications of text mining 174

*Recommended reading* 181

*Exercises, Problems and Weblems* 182

**4 Archives and information retrieval** 186**Database indexing and specification of search terms** 187**The archives** 190

- Nucleic acid sequence databases 190
- Genome databases and genome browsers 192
- Protein sequence databases 193
- Databases of protein families 198
- Databases of structures 199
- Classifications of protein structures 204
- Accuracy and precision of protein structure determinations 204

**Classification and assignment of protein function** 206

- The Enzyme Commission 206
- The Gene Ontology<sup>TM</sup> Consortium protein function classification 207
- Specialized, or 'boutique' databases 210
- Expression and proteomics databases 211
- Databases of metabolic pathways 213
- Bibliographic databases 218
- Surveys of molecular biology databases and servers 218

**Gateways to archives** 219

- Access to databases in molecular biology 220
- ENTREZ 220
- The Sequence Retrieval System (SRS) 233

The Protein Identification Resource (PIR) 234

ExPASy—Expert Protein Analysis System 236

**Where do we go from here?** 237

*Recommended reading* 237

*Exercises, Problems and Weblems* 237

## **5 Alignments and phylogenetic trees** 242

**Introduction to sequence alignment** 243

**The dotplot** 244

**Dotplots and sequence alignments** 249

**Measures of sequence similarity** 254

Scoring schemes 255

Derivation of substitution matrices: PAM matrices 256

**Computing the alignment of two sequences** 258

Variations and generalizations 259

Approximate methods for quick screening of databases 259

**The dynamic-programming algorithm for optimal pairwise sequence alignment** 261

**Significance of alignments** 267

**Multiple sequence alignment** 271

**Applications of multiple sequence alignments to database searching** 273

Profiles 273

PSI-BLAST 276

Hidden Markov Models 278

**Phylogeny** 281

Determination of taxonomic relationships from molecular properties 284

**Phylogenetic trees** 286

Clustering methods 288

Cladistic methods 290

Reconstruction of ancestral sequences 291

**The problem of varying rates of evolution** 293

Are trees the correct way to present phylogenetic relationships? 294

Computational considerations 296

**Putting it all together** 296

*Recommended reading* 297

*Exercises, Problems and Weblems* 298

## **6 Structural bioinformatics and drug discovery** 307

**Introduction** 308

**Protein stability and folding** 310

The Sasisekharan–Ramakrishnan–Ramachandran plot describes allowed mainchain conformations 310

The sidechains 312

Protein stability and denaturation 314

**Protein folding** 316

- Applications of hydrophobicity** 317
  - Coiled-coil proteins 321
- Superposition of structures, and structural alignments** 324
  - DALI and MUSTANG 326
- Evolution of protein structures** 327
  - Classifications of protein structures 330
- Protein structure prediction and modelling** 333
  - A priori* and empirical methods 334
  - Critical Assessment of Structure Prediction (CASP) 337
  - Secondary structure prediction 338
  - Homology modelling 343
  - Fold recognition 345
  - Conformational energy calculations and molecular dynamics 351
- Assignment of protein structures to genomes** 358
- Prediction of protein function** 360
  - Divergence of function: orthologues and paralogues 362
- Drug discovery and development** 364
  - The lead compound 367
  - Improving on the lead compound: quantitative structure-activity relationships (QSAR) 367
  - Bioinformatics in drug discovery and development 369
  - Molecular modelling in drug discovery 370
- Recommended reading* 379
- Exercises, Problems and Weblems* 380

## 7 Proteomics and systems biology 389

- DNA microarrays** 390
  - Analysis of microarray data 392
- Mass spectrometry** 398
  - Identification of components of a complex mixture 399
  - Protein sequencing by mass spectrometry 402
  - Measuring deuterium exchange in proteins 404
  - Genome sequence analysis by mass spectrometry 404
- Systems biology** 409
  - Protein complexes and aggregates** 411
    - Properties of protein-protein complexes 412
  - Protein interaction networks** 415
  - Networks and graphs** 420
  - Dynamics, stability and robustness** 424
  - The background of systems biology: sources of ideas** 426
    - Complexity of sequences 427
    - Computational complexity 430
    - Static and dynamic complexity 432
    - Chaos and predictability 433



## CONTENTS

**Alignment of metabolic pathways** 434

**Regulatory networks** 438

Signal transduction and transcriptional control 441

Structures of regulatory networks 441

Structural biology of regulatory networks 441

The genetic switch of bacteriophage  $\lambda$  443

**The genetic regulatory network of *Saccharomyces cerevisiae*** 451

Adaptability of the yeast regulatory network 454

*Recommended reading* 456

*Exercises, Problems and Weblems* 456

**Conclusions** 465

**Index** 467

**Colour Plates** 475