

CONTENTS

Preface to the fifth edition	xv
Plan of the book	xxi
Introduction to bioinformatics on the web	xxii
Acknowledgements	xxiii
1 Introduction	1
Life in space and time	4
Phenotype = genotype + environment + life history + epigenetics	4
Evolution is the change over time in the world of living things	5
Biological classification and nomenclature	6
Dogmas: central and peripheral	9
The structure of DNA	9
Transcription and translation	12
The structures of proteins	12
Statics and dynamics	17
Systems biology	17
The human genome	19
Variation in human genome sequences	20
The human genome and medicine	21
Databases in molecular biology	28
Observables and data archives	29
A database without effective modes of access is merely a data graveyard	29
Information flow in bioinformatics	31
Curation, annotation, and quality control	32
The World Wide Web	33
Electronic publication	34
Computers and computer science	34
Programming	35
<i>Après moi, le déluge?</i> Sorry—too late!	38
How much sequencing power is there in the world?	41
How does the amount of data in bioinformatics compare with other large scientific information archives?	41
Recommended reading	42
Exercises and Problems	43
2 From genetics to genomes	48
The classical genetics background	49
DNA embodies genes	50

Maps and tour guides	51
Linkage maps	51
Linkage	51
Chromosome banding	53
High-resolution maps, based directly on DNA sequences	55
Restriction maps	56
DNA sequencing	57
Frederick Sanger and the development of DNA sequencing	57
DNA sequencing by termination of chain replication	58
Automation of DNA sequencing	60
Next-generation sequencing	61
Paired-end reads	66
Life in the fast lane	67
Assembly—computational aspects	68
Pattern matching	68
Suffix trees	68
Fragment assembly	70
Genomics in personal identification	73
DNA 'fingerprinting'	74
Personal identification by amplification of specific regions has superseded the RFLP approach	75
Mitochondrial DNA	76
Analysis of non-human DNA sequences	77
Parentage testing	78
Ethical, legal, and social issues	80
Databases containing human DNA sequence information	80
Use of DNA sequencing in research on human subjects	82
Recommended reading	83
Exercises and Problems	84
3 The panorama of life	88
Genomes, transcriptomes, and proteomes	89
Genes	89
Proteomics and transcriptomics	91
Eavesdropping on the transmission of genetic information	93
Genome-sequencing projects	93
Genomes of prokaryotes	94
The genome of the bacterium <i>Escherichia coli</i>	95
The genome of the archaeon <i>Methanocaldococcus jannaschii</i>	97
The genome of one of the simplest organisms: <i>Mycoplasma genitalium</i>	98
Metagenomics: the collection of genomes in a coherent environmental sample	99
The human microbiome	101

Genomes of eukarya	102
Gene families	103
The genome of <i>Saccharomyces cerevisiae</i> (Baker's yeast)	103
The genome of <i>Caenorhabditis elegans</i>	105
The genome of <i>Drosophila melanogaster</i>	105
The genome of <i>Arabidopsis thaliana</i>	107
The genome of <i>Homo sapiens</i> (the human genome)	108
Protein-coding genes	109
Repeat sequences	109
RNA	110
Single-nucleotide polymorphisms and haplotypes	110
Systematic measurements and collections of single-nucleotide polymorphisms	113
Genetic diversity in anthropology	114
DNA sequences and languages	116
Evolution of genomes	116
Please pass the genes: horizontal gene transfer	118
Comparative genomics of eukarya	119
Recommended reading	120
Exercises and Problems	121
4 Alignments and phylogenetic trees	123
Introduction to sequence alignment	124
Dotplots and sequence alignments	130
Measures of sequence similarity	132
Scoring schemes	132
Derivation of substitution matrices: PAM matrices	133
Computing the alignment of two sequences	135
Variations and generalizations	135
Approximate methods for quick screening of databases	135
The dynamic programming algorithm for optimal pairwise sequence alignment	137
Significance of alignments	141
Multiple sequence alignment	143
Applications of multiple sequence alignments to database searching	143
Profiles	146
PSI-BLAST	147
Complete pairwise sequence alignment of human PAX-6 protein and <i>Drosophila melanogaster</i> <i>eyeless</i>	151
Hidden Markov Models	152
Phylogeny	154
Determination of taxonomic relationships from molecular properties	155
Use of sequences to determine phylogenetic relationships	159
Use of SINES and LINES to derive phylogenetic relationships	161

Phylogenetic trees	162
Clustering methods	164
The maximum-likelihood method	165
Reconstruction of ancestral sequences	165
Pyruvate decarboxylase: synthesis, activity, and crystal structure of predicted ancestor	167
The problem of varying rates of evolution	168
Bayesian methods	169
Are trees the correct way to present phylogenetic relationships?	169
Computational considerations	170
Putting it all together	171
Recommended reading	171
Exercises and Problems	172
5 Structural bioinformatics and drug discovery	177
Introduction	178
Protein stability and folding	180
The Sasisekharan–Ramakrishnan–Ramachandran plot	
describes allowed mainchain conformations	180
The sidechains	181
Protein stability and denaturation	183
Protein folding as a process	185
Applications of hydrophobicity	187
Coiled-coiled proteins	187
Description of the variety of protein structures	190
Superposition of structures, and structural alignments	192
Evolution of protein structures	197
Classifications of protein structures	199
SCOP	199
Protein structure prediction and modelling	201
<i>A priori</i> and empirical methods	202
Critical Assessment of Structure Prediction	203
Secondary structure prediction	204
Homology modelling	205
Fold recognition	205
Conformational energy calculations and molecular dynamics	207
ROSETTA	209
Protein structure prediction from contact maps derived from correlated mutations	
in multiple sequence alignments	210
Design of novel proteins	213
Drug discovery and development	215
The lead compound	216
Improving on the lead compound: quantitative structure–activity relationships	217
Bioinformatics in drug discovery and development	218
Molecular modelling in drug discovery	219

Recommended reading	225
Exercises and Problems	228
6 Scientific publications and archives: media, content, access, and presentation	233
The scientific literature	234
Access to scholarly publications	235
Open access	236
The Public Library of Science	237
Traditional and digital libraries	237
How to populate a digital library	238
The information explosion	239
The web: higher dimensions	239
New media: video, sound	240
Searching the scientific literature	240
Bibliography management	241
Databases	242
Database contents	242
Database quality control	243
The literature as a database	244
Database organization	244
Annotation	246
Markup languages	248
Database access	250
Links	250
Database interoperability	251
Data mining	251
Programming languages and tools for database construction and access	255
Traditional programming languages	255
Scripting languages	256
Program libraries specialized for molecular biology	256
Java—computing over the web	256
Natural language processing	257
Natural language processing in mining the biomedical literature	258
Biomedical applications of text mining	260
Hypothesis generation	264
A glaucoma-related network derived by text mining	265
Recommended reading	268
Exercises and Problems	269
7 Artificial intelligence and machine learning	271
What are artificial intelligence and machine learning?	272
Classification and clustering	273
Binary classifier	276
Receiver Operating Characteristic (ROC) curves	277

Artificial neural networks	279
Self-organizing maps	281
Decision trees	281
Support vector machines (SVMs)	286
Kernel methods	286
Clustering	287
Clustering by graph spectral theory	291
Recommended reading	293
Exercises and Problems	293
8 Introduction to systems biology	296
Introduction	297
Networks and graphs	298
Connectivity in networks	299
Dynamics, stability, and robustness	301
Some sources of ideas for systems biology	302
Complexity of sequences	302
Shannon's definition of entropy	303
Complexity of sequences	304
The relationship between complexity, randomness, and compressibility	305
The Burrows-Wheeler Transform	305
Inverting the Burrows-Wheeler Transform	306
The Burrows-Wheeler Transform brings repeats together, facilitating compression	306
Use of the Burrows-Wheeler transform for searching for patterns in strings	306
Complexity of other types of biological data	308
Static and dynamic complexity	308
Predictability and chaos	309
Analysis and comparison of networks	310
Analysis of graphs by matrix algebra	311
Graph isomorphism	312
Recommended reading	314
Exercises and Problems	314
9 Metabolic pathways	317
Introduction	318
Classification of protein function	320
The Enzyme Commission	320
The Gene Ontology TM Consortium protein function classification	320
Prediction of protein function	321
Catalysis by enzymes	324
Active sites	325
Cofactors	325

Protein–ligand binding equilibria	326
Enzyme kinetics	327
Measures of effectiveness of enzymes	328
How do enzymes evolve new functions?	329
Control over enzyme activity	329
Structural mechanisms of evolution of altered or novel protein functions	329
Pathways and limits in the divergence of sequence, structure, and function	334
Evolution by gene duplication	335
Databases of metabolic pathways	337
The Kyoto Encyclopedia of Genes and Genomes (KEGG)	339
Evolution and phylogeny of metabolic pathways	341
Pathway comparison	341
Alignment of metabolic pathways	343
Comparing linear metabolic pathways	343
Comparing non-linear metabolic pathways: The pentose phosphate pathway and the Calvin-Benson cycle	346
Dynamics of metabolic networks	347
Robustness of metabolic networks	347
Dynamic modelling of metabolism	347
Simulation of metabolic pathways in <i>Plasmodium falciparum</i>	351
The Human Metabolome Database supports clinical applications to the study of inborn errors of metabolism, and to cancer	352
Recommended reading	353
Exercises and Problems	353
10 Control of organization and organization of control	355
Transcriptomics	356
The ENCODE Project	357
Determination of RNA sequences	358
RNAseq v. microarrays	358
DNA microarrays	359
RNAseq	363
The Genotype-Tissue Expression (GTEx) project	366
Expression patterns in different physiological states	367
Variation of expression patterns during the life cycle of <i>Drosophila melanogaster</i>	368
Different life stages make different demands on different genes	370
Protein complexes and aggregates	373
Properties of protein–protein complexes	373
Protein interaction networks	375
Components of the primosome assembly in <i>Bacillus subtilis</i>	378

Regulatory networks	380
Signal transduction and transcriptional control	380
Structural biology of regulatory networks	382
Examples of relatively simple regulatory control networks	383
Regulation of the lactose operon in <i>E. coli</i>	383
The genetic switch of bacteriophage λ	385
The diauxic shift in <i>Saccharomyces cerevisiae</i>	389
Logical structure of regulatory networks	391
The transcriptional regulatory network of <i>E. coli</i>	391
The transcriptional regulatory network of <i>Saccharomyces cerevisiae</i>	392
Adaptability of the yeast regulatory network	393
Recommended reading	396
Exercises and Problems	396
Conclusions	399
Index	400