

CONTENTS

Preface to the first edition	xiv
Preface to the second edition	xvii
Preface to the third edition	xviii
Preface to the fourth edition	xix
Plan of the book	xx
Introduction to bioinformatics on the web	xxi
Acknowledgements	xxii
1 Introduction	1
Life in space and time	3
Phenotype = genotype + environment + life history + epigenetics	4
Evolution is the change over time in the world of living things	4
Dogmas: central and peripheral	6
Statics and dynamics	9
Networks	10
Observables and data archives	10
A database without effective modes of access is merely a data graveyard	12
Information flow in bioinformatics	12
Curation, annotation, and quality control	14
The world-wide web	15
Electronic publication	16
Computers and computer sciences	16
Programming	17
Biological classification and nomenclature	21
Use of sequences to determine phylogenetic relationships	23
Use of SINES and LINES to derive phylogenetic relationships	29
Searching for similar sequences in databases: PSI-BLAST	30
Introduction to protein structure	38
The hierarchical nature of protein architecture	40
Classification of protein structures	41
Protein structure prediction and engineering	46
Critical Assessment of Structure Prediction	47
Protein engineering	48
Proteomics and transcriptomics	48
DNA microarrays	49
Transcriptomics and RNA sequencing	50
Mass spectrometry	50

	Systems biology	50
	Clinical implications	50
	The future	53
	Recommended reading	53
	Exercises and problems	55
2	Genome organization and evolution	59
	Genomes, transcriptomes, and proteomes	59
	Genes	60
	Proteomics and transcriptomics	62
	Eavesdropping on the transmission of genetic information	64
	Identification of genes associated with inherited diseases	64
	Mappings between the maps	66
	High-resolution maps	68
	Genome-wide association studies	69
	Picking out genes in genomes	70
	Genome-sequencing projects	71
	Genomes of prokaryotes	72
	The genome of the bacterium <i>Escherichia coli</i>	73
	The genome of the archaeon <i>Methanococcus jannaschii</i>	75
	The genome of one of the simplest organisms: <i>Mycoplasma genitalium</i>	76
	Metagenomics: the collection of genomes in a coherent environmental sample	76
	The human microbiome	79
	Genomes of eukarya	79
	Gene families	82
	The genome of <i>Saccharomyces cerevisiae</i> (baker's yeast)	82
	The genome of <i>Caenorhabditis elegans</i>	84
	The genome of <i>Drosophila melanogaster</i>	85
	The genome of <i>Arabidopsis thaliana</i>	86
	The genome of <i>Homo sapiens</i> (the human genome)	88
	Protein-coding genes	88
	Repeat sequences	89
	RNA	90
	Single-nucleotide polymorphisms and haplotypes	90
	Systematic measurements and collections of single-nucleotide polymorphisms	92
	Ethical, legal, and social issues	94
	Genetic diversity in anthropology	95
	DNA sequences and languages	96
	Genetic diversity and personal identification	97
	Evolution of genomes	98
	Please pass the genes: horizontal gene transfer	100
	Comparative genomics of eukarya	101
	Recommended Reading	102

Exercises and problems	103
3 Scientific publications and archives: media, content, and access	107
The scientific literature	107
Economic factors governing access to scholarly publications	109
Open access	110
The Public Library of Science	111
Traditional and digital libraries	111
How to populate a digital library	112
The information explosion	113
The web: higher dimensions	113
New media: video, sound	113
Searching the literature	114
Bibliography management	114
Databases	115
Database contents	116
The literature as a database	116
Database organization	116
Annotation	119
Database quality control	120
Database access	122
Links	123
Database interoperability	125
Data mining	127
Programming languages and tools	128
Traditional programming languages	130
Scripting languages	130
Program libraries specialized for molecular biology	130
Java: computing over the web	131
Markup languages	131
Natural language processing	133
Natural language processing and mining the biomedical literature	134
Applications of text mining	136
Recommended reading	141
Exercises and problems	142
4 Archives and information retrieval	144
Database indexing and specification of search terms	144
Follow-up questions	145
Analysis and processing of retrieved data	146
The archives	146
Nucleic acid sequence databases	147
Genome databases and genome browsers	148
Protein sequence databases	149
Databases of protein families	152

Databases of structures	153
Classifications of protein structures	157
Accuracy and precision of protein structure determinations	157
Specialized, or 'boutique', databases	158
Expression and proteomics databases	159
Bibliographic databases	160
Surveys of molecular biology databases and servers	161
Gateways to archives	161
Access to databases in molecular biology	162
ENTREZ	162
The Protein Identification Resource	171
ExPASy: Expert Protein Analysis System	172
Where do we go from here?	173
Recommended reading	173
Exercises and problems	174
5 Alignments and phylogenetic trees	175
Introduction to sequence alignment	175
The dotplot	176
Dotplots and sequence alignments	181
Measures of sequence similarity	182
Scoring schemes	184
Derivation of substitution matrices: PAM and BLOSUM matrices	184
Computing the alignment of two sequences	187
Variations and generalizations	187
Approximate methods for quick screening of databases	188
The dynamic-programming algorithm for optimal pairwise sequence alignment	188
Significance of alignments	194
Multiple sequence alignment	196
Applications of multiple sequence alignments and database searching	198
Profiles	198
PSI-BLAST	200
Hidden Markov models	201
Phylogeny	203
Determination of taxonomic relationships from molecular properties	205
Phylogenetic trees	207
Clustering methods	209
Cladistic methods	210
Reconstruction of ancestral sequences	211
The problem of varying rates of evolution	213
Are trees the correct way to present phylogenetic relationships?	213
Computational considerations	214
Putting it all together	215

Recommended reading	215
Exercises and problems	216
6 Structural bioinformatics and drug discovery	222
Introduction	222
Protein stability and folding	224
The Sasisekharan–Ramakrishnan–Ramachandran plot describes allowed mainchain conformations	225
The sidechains	226
Protein stability and denaturation	227
Protein folding	229
Applications of hydrophobicity	230
Coiled-coiled proteins	230
Superposition of structures, and structural alignments	235
DALI and MUSTANG	237
Evolution of protein structures	238
Classifications of protein structures	240
Protein structure prediction and modelling	241
A priori and empirical methods	241
Critical Assessment of Structure Prediction	243
Secondary structure prediction	246
Homology modelling	249
Fold recognition	250
Conformational energy calculations and molecular dynamics	255
Assignment of protein structures to genomes	260
Prediction of protein function	261
Divergence of function: orthologues and paralogues	261
Drug discovery and development	264
The lead compound	265
Improving on the lead compound: quantitative structure-activity relationships	266
Bioinformatics in drug discovery and development	267
Molecular modelling in drug discovery	268
Recommended reading	274
Exercises and problems	276
7 Introduction to systems biology	282
Introduction	282
Networks and graphs –	284
Connectivity in networks	285
Dynamics, stability, and robustness	286
Some sources of ideas for systems biology	288
Complexity of sequences	288
Computational complexity	291

Static and dynamic complexity	291
Chaos and predictability	293
Recommended reading	294
Exercises and problems	294
8 Metabolic pathways	297
Classification and assignment of protein function	298
The Enzyme Commission	298
The Gene Ontology Consortium protein function classification	299
Catalysis by enzymes	301
Active sites	303
Cofactors	303
Protein–ligand binding equilibria	304
Enzyme kinetics	305
Measures of effectiveness of enzymes	306
How do proteins evolve new functions?	307
Control over enzyme activity	308
Structural mechanisms of evolution of altered or novel protein functions	308
Protein evolution at the level of domain assembly	311
Databases of metabolic pathways	312
EcoCyc	313
The Kyoto Encyclopedia of Genes and Genomes	313
Evolution and phylogeny of metabolic pathways	316
Pathway comparison	316
Alignment of metabolic pathways	320
Comparing linear metabolic pathways	320
Comparing nonlinear metabolic pathways: the pentose phosphate pathway and the Calvin–Benson cycle	321
Dynamics of metabolic networks	322
Robustness of metabolic networks	323
Dynamic modelling of metabolism	324
Recommended reading	326
Exercises and problems	326
9 Gene expression and regulation	328
DNA microarrays	329
Microarray data are quantitative but imprecise	330
Analysis of microarray data	330
Mass spectrometry	335
Identification of components of a complex mixture	335
Protein sequencing by mass spectrometry	337
Measuring deuterium exchange in proteins	338
Genome sequence analysis by mass spectrometry	339

Protein complexes and aggregates	342
Properties of protein-protein complexes	343
Protein interaction networks	345
Regulatory networks	348
Signal transduction and transcriptional control	349
Structures of regulatory networks	350
Structural biology of regulatory networks	350
The genetic switch of bacteriophage λ	352
What are the characteristics of the switch that must be implemented by DNA-protein interactions?	353
The materials	354
How to 'throw' the switch	355
The genetic regulatory network of <i>Saccharomyces cerevisiae</i>	356
Adaptability of the yeast regulatory network	358
Recommended reading	360
Exercises and problems	360
Conclusion	363
Index	365