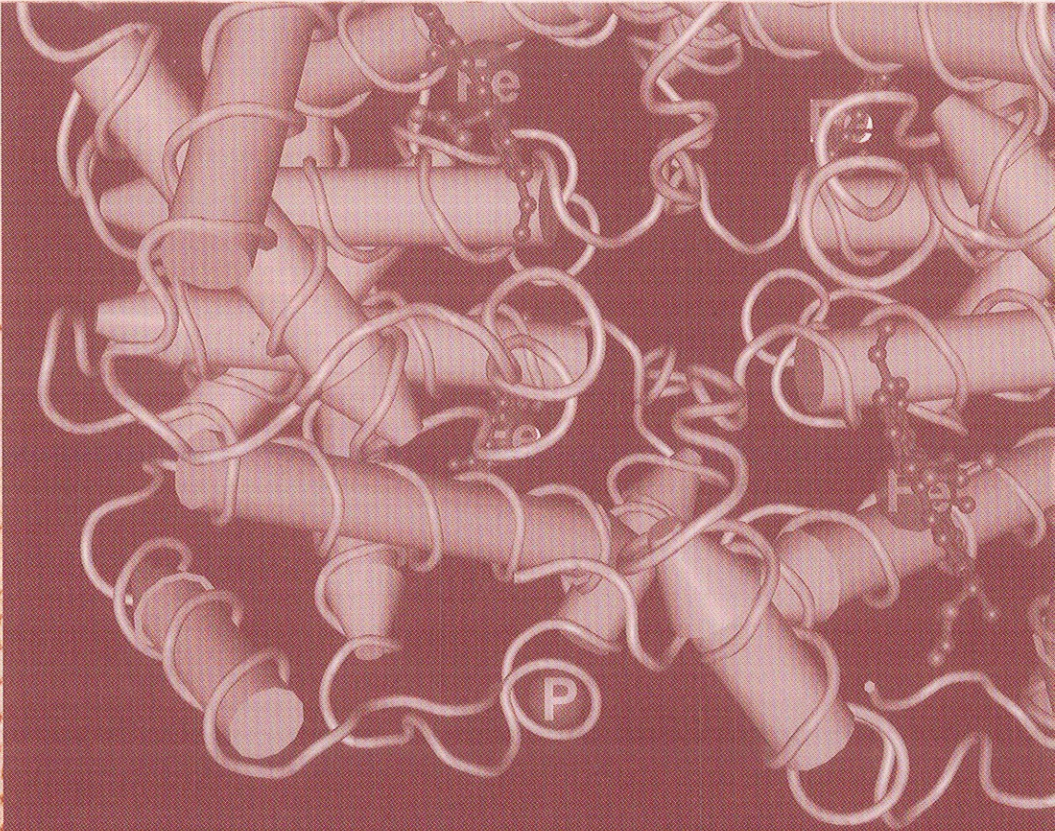


second edition

Bioinformatics and Functional Genomics



Jonathan Pevsner

 WILEY-BLACKWELL

WWW.
LINK AVAILABLE

Contents

Preface to the Second Edition, xxi

Preface to the First Edition, xxiii

Foreword, xxvii

PART I ANALYZING DNA, RNA, AND PROTEIN SEQUENCES IN DATABASES

1 Introduction, 3

- Organization of The Book, 4
- Bioinformatics: The Big Picture, 4
- A Consistent Example:
 - Hemoglobin, 8
- Organization of The Chapters, 9
- A Textbook for Courses on
 - Bioinformatics and Genomics, 9
- Key Bioinformatics Websites, 10
- Suggested Reading, 11
- References, 11

2 Access to Sequence Data and Literature Information, 13

- Introduction to Biological Databases, 13
- GenBank: Database of Most Known Nucleotide and Protein Sequences, 14
 - Amount of Sequence Data, 15
 - Organisms in GenBank, 16
 - Types of Data in GenBank, 18
 - Genomic DNA Databases, 19
 - cDNA Databases Corresponding to Expressed Genes, 19
 - Expressed Sequence Tags (ESTs), 19
 - ESTs and UniGene, 20
 - Sequence-Tagged Sites (STSs), 22

Genome Survey Sequences (GSSs), 22

High Throughput Genomic Sequence (HTGS), 23

Protein Databases, 23

National Center for Biotechnology Information, 23

Introduction to NCBI: Home Page, 23

PubMed, 23

Entrez, 24

BLAST, 25

OMIM, 25

Books, 25

Taxonomy, 25

Structure, 25

The European Bioinformatics Institute (EBI), 25

Access to Information: Accession Numbers to Label and Identify Sequences, 26

The Reference Sequence (RefSeq) Project, 27

The Consensus Coding Sequence (CCDS) Project, 29

Access to Information via Entrez Gene at NCBI, 29

Relationship of Entrez Gene, Entrez Nucleotide, and Entrez Protein, 32

Comparison of Entrez Gene and UniGene, 32

Entrez Gene and HomoloGene, 33

Access to Information: Protein Databases, 33

UniProt, 33

The Sequence Retrieval System at ExPASy, 34

Access to Information: The Three Main Genome Browsers, 35

The Map Viewer at NCBI, 35

The University of California, Santa Cruz (UCSC) Genome Browser, 35
 The Ensembl Genome Browser, 35
 Examples of How to Access Sequence Data, 36
 HIV *pol*, 36
 Histones, 38
 Access to Biomedical Literature, 38
 PubMed Central and Movement toward Free Journal Access, 39
 Example of PubMed Search: RBP, 40
 Perspective, 42
 Pitfalls, 42
 Web Resources, 42
 Discussion Questions, 42
 Problems, 42
 Self-Test Quiz, 43
 Suggested Reading, 44
 References, 44

3 **Pairwise Sequence Alignment, 47**

Introduction, 47
 Protein Alignment: Often More Informative Than DNA Alignment, 47
 Definitions: Homology, Similarity, Identity, 48
 Gaps, 55
 Pairwise Alignment, Homology, and Evolution of Life, 55
 Scoring Matrices, 57
 Dayhoff Model: Accepted Point Mutations, 58
 PAM1 Matrix, 63
 PAM250 and Other PAM Matrices, 65
 From a Mutation Probability Matrix to a Log-Odds Scoring Matrix, 69
 Practical Usefulness of PAM Matrices in Pairwise Alignment, 70
 Important Alternative to PAM: BLOSUM Scoring Matrices, 70
 Pairwise Alignment and Limits of Detection: The "Twilight Zone", 74
 Alignment Algorithms: Global and Local, 75
 Global Sequence Alignment: Algorithm of Needleman and Wunsch, 76

Step 1: Setting Up a Matrix, 76
 Step 2: Scoring the Matrix, 77
 Step 3: Identifying the Optimal Alignment, 79
 Local Sequence Alignment: Smith and Waterman Algorithm, 82
 Rapid, Heuristic Versions of Smith–Waterman: FASTA and BLAST, 84
 Pairwise Alignment with Dot Plots, 85
 The Statistical Significance of Pairwise Alignments, 86
 Statistical Significance of Global Alignments, 87
 Statistical Significance of Local Alignments, 89
 Percent Identity and Relative Entropy, 90
 Perspective, 91
 Pitfalls, 94
 Web Resources, 94
 Discussion Questions, 94
 Problems/Computer Lab, 95
 Self-Test Quiz, 95
 Suggested Reading, 96
 References, 97

4 **Basic Local Alignment Search Tool (BLAST), 101**

Introduction, 101
 BLAST Search Steps, 103
 Step 1: Specifying Sequence of Interest, 103
 Step 2: Selecting BLAST Program, 104
 Step 3: Selecting a Database, 106
 Step 4a: Selecting Optional Search Parameters, 106
 1. Query, 107
 2. Limit by Entrez Query, 107
 3. Short Queries, 107
 4. Expect Threshold, 107
 5. Word Size, 108
 6. Matrix, 110
 7. Gap Penalties, 110
 8. Composition-Based Statistics, 110
 9. Filtering and Masking, 111
 Step 4b: Selecting Formatting Parameters, 112
 BLAST Algorithm Uses Local Alignment Search Strategy, 115

BLAST Algorithm Parts: List, Scan, Extend, 115

BLAST Algorithm: Local Alignment Search Statistics and *E* Value, 118

Making Sense of Raw Scores with Bit Scores, 121

BLAST Algorithm: Relation between *E* and *p* Values, 121

Parameters of a BLAST Search, 123

BLAST Search Strategies, 123

General Concepts, 123

Principles of BLAST Searching, 123

How to Evaluate Significance of Your Results, 123

How to Handle Too Many Results, 128

How to Handle Too Few Results, 128

BLAST Searching With Multidomain Protein: HIV-1 pol, 129

Perspective, 134

Pitfalls, 134

Web Resources, 135

Discussion Questions, 135

Computer Lab/Problems, 135

Self-Test Quiz, 136

Suggested Reading, 137

References, 137

5 **Advanced Database Searching, 141**

Introduction, 141

Specialized BLAST Sites, 142

Organism-Specific BLAST Sites, 142

Ensembl BLAST, 142

Wellcome Trust Sanger Institute, 143

Specialized BLAST-Related Algorithms, 143

WU BLAST 2.0, 144

European Bioinformatics Institute (EBI), 144

Specialized NCBI BLAST Sites, 144

Finding Distantly Related Proteins: Position-Specific Iterated BLAST (PSI-BLAST), 145

Assessing Performance of PSI-BLAST, 150

PSI-BLAST Errors: The Problem of Corruption, 152

Reverse Position-Specific BLAST, 152

Pattern-Hit Initiated BLAST (PHI-BLAST), 153

Profile Searches: Hidden Markov Models, 155

BLAST-Like Alignment Tools to Search Genomic DNA Rapidly, 161

Benchmarking to Assess Genomic Alignment Performance, 162

PatternHunter, 162

BLASTZ, 163

MegaBLAST and Discontiguous MegaBLAST, 164

BLAT, 166

LAGAN, 168

SSAHA, 168

SIM4, 169

Using BLAST for Gene Discovery, 169

Perspective, 173

Pitfalls, 173

Web Resources, 174

Discussion Questions, 174

Problems/Computer Lab, 174

Self-Test Quiz, 175

Suggested Reading, 176

References, 176

6 **Multiple Sequence Alignment, 179**

Introduction, 179

Definition of Multiple Sequence Alignment, 180

Typical Uses and Practical Strategies of Multiple Sequence Alignment, 181

Benchmarking: Assessment of Multiple Sequence Alignment Algorithms, 182

Five Main Approaches to Multiple Sequence Alignment, 184

Exact Approaches to Multiple Sequence Alignment, 184

Progressive Sequence Alignment, 185

Iterative Approaches, 190

Consistency-Based Approaches, 192

Structure-Based Methods, 194

Conclusions from Benchmarking Studies, 196

Databases of Multiple Sequence
 Alignments, 197
 Pfam: Protein Family Database of
 Profile HMMs, 197
 Smart, 199
 Conserved Domain Database, 199
 Prints, 201
 Integrated Multiple Sequence
 Alignment Resources: InterPro
 and iProClass, 201
 PopSet, 202
 Multiple Sequence Alignment
 Database Curation: Manual
 versus Automated, 202
 Multiple Sequence Alignments of
 Genomic Regions, 203
 Perspective, 206
 Pitfalls, 207
 Web Resources, 207
 Discussion Questions, 207
 Problems/Computer Lab, 208
 Self-Test Quiz, 208
 Suggested Reading, 209
 References, 210

Molecular Phylogeny and Evolution, 215

Introduction to Molecular
 Evolution, 215
 Goals of Molecular
 Phylogeny, 216
 Historical Background, 217
 Molecular Clock
 Hypothesis, 221
 Positive and Negative
 Selection, 227
 Neutral Theory of Molecular
 Evolution, 230
 Molecular Phylogeny: Properties of
 Trees, 231
 Tree Roots, 233
 Enumerating Trees and
 Selecting Search
 Strategies, 234
 Type of Trees, 238
 Species Trees versus Gene/Protein
 Trees, 238
 DNA, RNA, or Protein-Based
 Trees, 240
 Five Stages of Phylogenetic
 Analysis, 243
 Stage 1: Sequence
 Acquisition, 243
 Stage 2: Multiple Sequence
 Alignment, 244

Stage 3: Models of DNA
 and Amino Acid
 Substitution, 246
 Stage 4: Tree-Building
 Methods, 254
 Phylogenetic Methods, 255
 Distance, 255
 The UPGMA Distance-Based
 Method, 256
 Making Trees by Distance-
 Based Methods: Neighbor
 Joining, 259
 Phylogenetic Inference: Maximum
 Parsimony, 260
 Model-Based Phylogenetic
 Inference: Maximum
 Likelihood, 262
 Tree Inference: Bayesian
 Methods, 264
 Stage 5: Evaluating Trees, 266
 Perspective, 268
 Pitfalls, 268
 Web Resources, 269
 Discussion Questions, 269
 Problems/Computer Lab, 269
 Self-Test Quiz, 271
 Suggested Reading, 272
 References, 272

PART II GENOMEWIDE ANALYSIS OF RNA AND PROTEIN

**Bioinformatic Approaches
 to Ribonucleic Acid (RNA), 279**
 Introduction to RNA, 279
 Noncoding RNA, 282
 Noncoding RNAs in the Rfam
 Database, 283
 Transfer RNA, 283
 Ribosomal RNA, 288
 Small Nuclear RNA, 291
 Small Nucleolar RNA, 292
 MicroRNA, 293
 Short Interfering RNA, 294
 Noncoding RNAs in the UCSC
 Genome and Table
 Browser, 294
 Introduction to Messenger RNA, 296
 mRNA: Subject of Gene
 Expression Studies, 300
 Analysis of Gene Expression in
 cDNA Libraries, 302
 Pitfalls in Interpreting Expression
 Data from cDNA Libraries, 308

- Full-Length cDNA Projects, 308
- Serial Analysis of Gene Expression (SAGE), 309
- Microarrays: Genomewide
 - Measurement of Gene Expression, 312
 - Stage 1: Experimental Design for Microarrays, 314
 - Stage 2: RNA Preparation and Probe Preparation, 316
 - Stage 3: Hybridization of Labeled Samples to DNA Microarrays, 317
 - Stage 4: Image Analysis, 317
 - Stage 5: Data Analysis, 318
 - Stage 6: Biological Confirmation, 320
 - Microarray Databases, 320
 - Further Analyses, 320
- Interpretation of RNA Analyses, 320
 - The Relationship of DNA, mRNA, and Protein Levels, 320
 - The Pervasive Nature of Transcription, 321
- Perspective, 322
- Pitfalls, 323
- Web Resources, 323
- Discussion Questions, 323
- Problems, 324
- Self-Test Quiz, 324
- Suggested Reading, 325
- References, 325

9 Gene Expression: Microarray Data Analysis, 331

- Introduction, 331
 - Microarray Data Analysis Software and Data Sets, 334
 - Reproducibility of Microarray Experiments, 335
- Microarray Data Analysis:
 - Preprocessing, 337
 - Scatter Plots and MA Plots, 338
 - Global and Local Normalization, 343
 - Accuracy and Precision, 344
 - Robust Multiarray Analysis (RMA), 345
- Microarray Data Analysis: Inferential Statistics, 346
 - Expression Ratios, 346
 - Hypothesis Testing, 347

- Corrections for Multiple Comparisons, 351
- Significance Analysis of Microarrays (SAM), 351
- From *t*-Test to ANOVA, 353
- Microarray Data Analysis: Descriptive Statistics, 354
 - Hierarchical Cluster Analysis of Microarray Data, 355
 - Partitioning Methods for Clustering: *k*-Means Clustering, 363
 - Clustering Strategies: Self-Organizing Maps, 363
 - Principal Components Analysis: Visualizing Microarray Data, 364
 - Supervised Data Analysis for Classification of Genes or Samples, 367
- Functional Annotation of Microarray Data, 368
- Perspective, 369
- Pitfalls, 370
- Discussion Questions, 370
- Problems/Computer Lab, 371
- Self-Test Quiz, 372
- Suggested Reading, 373
- References, 373

10 Protein Analysis and Proteomics, 379

- Introduction, 379
 - Protein Databases, 380
 - Community Standards for Proteomics Research, 381
- Techniques to Identify Proteins, 381
 - Direct Protein Sequencing, 381
 - Gel Electrophoresis, 382
 - Mass Spectrometry, 385
- Four Perspectives on Proteins, 388
- Perspective 1. Protein Domains and Motifs: Modular Nature of Proteins, 389
 - Added Complexity of Multidomain Proteins, 394
 - Protein Patterns: Motifs or Fingerprints Characteristic of Proteins, 394
- Perspective 2. Physical Properties of Proteins, 397
 - Accuracy of Prediction Programs, 399
 - Proteomic Approaches to Phosphorylation, 401

- Proteomic Approaches to Transmembrane Domains, 401
 - Introduction to Perspectives 3 and 4: Gene Ontology Consortium, 402
 - Perspective 3: Protein Localization, 406
 - Perspective 4: Protein Function, 407
 - Perspective, 411
 - Pitfalls, 411
 - Web Resources, 412
 - Discussion Questions, 414
 - Problems/Computer Lab, 415
 - Self-Test Quiz, 415
 - Suggested Reading, 416
 - References, 416
- 11 **Protein Structure, 421**
- Overview of Protein Structure, 421
 - Protein Sequence and Structure, 422
 - Biological Questions Addressed by Structural Biology: Globins, 423
 - Principles of Protein Structure, 423
 - Primary Structure, 424
 - Secondary Structure, 425
 - Tertiary Protein Structure: Protein-Folding Problem, 430
 - Target Selection and Acquisition of Three-Dimensional Protein Structures, 432
 - Structural Genomics and the Protein Structure Initiative, 432
 - The Protein Data Bank, 434
 - Accessing PDB Entries at the NCBI Website, 437
 - Integrated Views of the Universe of Protein Folds, 441
 - Taxonomic System for Protein Structures: The SCOP Database, 441
 - The CATH Database, 443
 - The Dali Domain Dictionary, 445
 - Comparison of Resources, 446
 - Protein Structure Prediction, 447
 - Homology Modeling (Comparative Modeling), 448
 - Fold Recognition (Threading), 450
 - Ab Initio Prediction (Template-Free Modeling), 450
 - A Competition to Assess Progress in Structure Prediction, 451
 - Intrinsically Disordered Proteins, 453
 - Protein Structure and Disease, 453
 - Perspective, 454
 - Pitfalls, 455
 - Discussion Questions, 455
 - Problems/Computer Lab, 455
 - Self-Test Quiz, 456
 - Suggested Reading, 457
 - References, 457
- 12 **Functional Genomics, 461**
- Introduction to Functional Genomics, 461
 - The Relationship of Genotype and Phenotype, 463
 - Eight Model Organisms for Functional Genomics, 465
 - The Bacterium *Escherichia coli*, 466
 - The Yeast *Saccharomyces cerevisiae*, 466
 - The Plant *Arabidopsis thaliana*, 470
 - The Nematode *Caenorhabditis elegans*, 470
 - The Fruitfly *Drosophila melanogaster*, 471
 - The Zebrafish *Danio rerio*, 471
 - The Mouse *Mus musculus*, 472
 - Homo sapiens*: Variation in Humans, 473
 - Functional Genomics Using Reverse Genetics and Forward Genetics, 473
 - Reverse Genetics: Mouse Knockouts and the β -Globin Gene, 475
 - Reverse Genetics: Knocking Out Genes in Yeast Using Molecular Barcodes, 480
 - Reverse Genetics: Random Insertional Mutagenesis (Gene Trapping), 483
 - Reverse Genetics: Insertional Mutagenesis in Yeast, 486
 - Reverse Genetics: Gene Silencing by Disrupting RNA, 489

Forward Genetics: Chemical Mutagenesis, 491

Functional Genomics and the Central Dogma, 492

Functional Genomics and DNA: The ENCODE Project, 492

Functional Genomics and RNA, 492

Functional Genomics and Protein, 493

Proteomics Approaches to Functional Genomics, 493

Protein-Protein Interactions, 495

The Yeast Two-Hybrid System, 496

Protein Complexes: Affinity Chromatography and Mass Spectrometry, 498

The Rosetta Stone Approach, 500

Protein-Protein Interaction Databases, 501

Protein Networks, 502

Perspective, 507

Pitfalls, 508

Discussion Questions, 508

Problems/Computer Lab, 509

Self-Test Quiz, 509

Suggested Reading, 510

References, 510

PART III GENOME ANALYSIS

13

Completed Genomes, 517

Introduction, 517

Five Perspectives on Genomics, 519

Brief History of Systematics, 520

History of Life on Earth, 521

Molecular Sequences as the Basis of the Tree of Life, 523

Role of Bioinformatics in Taxonomy, 524

Genome-Sequencing Projects: Overview, 525

Four Prominent Web Resources, 525

Brief Chronology, 526

First Bacteriophage and Viral Genomes (1976–1978), 527

First Eukaryotic Organellar Genome (1981), 527

First Chloroplast Genomes (1986), 528

First Eukaryotic Chromosome (1992), 529

Complete Genome of Free-Living Organism (1995), 530

First Eukaryotic Genome (1996), 532

Escherichia coli (1997), 532

First Genome of Multicellular Organism (1998), 532

Human Chromosome (1999), 533

Fly, Plant, and Human Chromosome 21 (2000), 534

Draft Sequences of Human Genome (2001), 535

Continuing Rise in Completed Genomes (2002), 535

Expansion of Genome Projects (2003–2009), 536

Genome Analysis Projects, 537

Criteria for Selection of Genomes for Sequencing, 538

Genome Size, 539

Cost, 540

Relevance to Human Disease, 541

Relevance to Basic Biological Questions, 541

Relevance to Agriculture, 541

Should an Individual from a Species, Several Individuals, or Many Individuals Be Sequenced, 541

Resequencing Projects, 542

Ancient DNA Projects, 542

Metagenomics Projects, 543

DNA Sequencing Technologies, 544

Sanger Sequencing, 544

Pyrosequencing, 545

Cyclic Reversible Termination: Solexa, 547

The Process of Genome Sequencing, 547

Genome-Sequencing Centers, 547

Sequencing and Assembling Genomes: Strategies, 548

Genomic Sequence Data: From Unfinished to Finished, 549

- Finishing: When Has a Genome Been Fully Sequenced, 551
- Repository for Genome Sequence Data, 552
- Role of Comparative Genomics, 552
- Genome Annotation: Features of Genomic DNA, 555
 - Annotation of Genes in Prokaryotes, 556
 - Annotation of Genes in Eukaryotes, 558
 - Summary: Questions from Genome-Sequencing Projects, 558
- Perspective, 559
- Pitfalls, 559
- Discussion Questions, 560
- Problems/Computer Lab, 560
- Self-Test Quiz, 560
- Suggested Reading, 561
- References, 561

14

Completed Genomes:

- Viruses, 567**
 - Introduction, 567
 - Classification of Viruses, 568
 - Diversity and Evolution of Viruses, 571
 - Metagenomics and Virus Diversity, 573
 - Bioinformatics Approaches to Problems in Virology, 574
 - Influenza Virus, 574
 - Herpesvirus: From Phylogeny to Gene Expression, 578
 - Human Immunodeficiency Virus, 583
 - Bioinformatic Approaches to HIV-1, 585
 - Measles Virus, 588
 - Perspectives, 591
 - Pitfalls, 591
 - Web Resources, 591
 - Discussion Questions, 592
 - Problems/Computer Lab, 592
 - Self-Test Quiz, 593
 - Suggested Reading, 593
 - References, 593

15

Completed Genomes: Bacteria and Archaea, 597

- Introduction, 598
- Classification of Bacteria and Archaea, 598

- Classification of Bacteria by Morphological Criteria, 599
- Classification of Bacteria and Archaea Based on Genome Size and Geometry, 602
- Classification of Bacteria and Archaea Based on Lifestyle, 607
- Classification of Bacteria Based on Human Disease Relevance, 610
- Classification of Bacteria and Archaea Based on Ribosomal RNA Sequences, 611
- Classification of Bacteria and Archaea Based on Other Molecular Sequences, 612
- Analysis of Prokaryotic Genomes, 615
 - Nucleotide Composition, 615
 - Finding Genes, 617
 - Lateral Gene Transfer, 620
 - Functional Annotation: COGs, 622
- Comparison of Prokaryotic Genomes, 625
 - TaxPlot, 626
 - MUMmer, 628
- Perspective, 629
- Pitfalls, 630
- Web Resources, 630
- Discussion Questions, 630
- Problems/Computer Lab, 631
- Self-Test Quiz, 631
- Suggested Reading, 632
- References, 632

16

The Eukaryotic Chromosome, 639

- Introduction, 640
 - Major Differences between Eukaryotes and Prokaryotes, 641
- General Features of Eukaryotic Genomes and Chromosomes, 643
- C Value Paradox: Why Eukaryotic Genome Sizes Vary So Greatly, 643
- Organization of Eukaryotic Genomes into Chromosomes, 644
- Analysis of Chromosomes Using Genome Browsers, 645

- Analysis of Chromosomes by the ENCODE Project, 647
 - Repetitive DNA Content of Eukaryotic Chromosomes, 650
 - Eukaryotic Genomes Include Noncoding and Repetitive DNA Sequences, 650
 1. Interspersed Repeats (Transposon-Derived Repeats), 652
 2. Processed Pseudogenes, 653
 3. Simple Sequence Repeats, 657
 4. Segmental Duplications, 658
 5. Blocks of Tandemly Repeated Sequences Such as Are Found at Telomeres, Centromeres, and Ribosomal Gene Clusters, 660
 - Gene Content of Eukaryotic Chromosomes, 662
 - Definition of Gene, 662
 - Finding Genes in Eukaryotic Genomes, 663
 - EGASP Competition and JIGSAW, 666
 - Protein-Coding Genes in Eukaryotes: New Paradox, 668
 - Regulatory Regions of Eukaryotic Chromosomes, 669
 - Transcription Factor Databases and Other Genomic DNA Databases, 669
 - Ultraconserved Elements, 672
 - Nonconserved Elements, 673
 - Comparison of Eukaryotic DNA, 673
 - Variation in Chromosomal DNA, 674
 - Dynamic Nature of Chromosomes: Whole Genome Duplication, 675
 - Chromosomal Variation in Individual Genomes, 676
 - Chromosomal Variation in Individual Genomes: Inversions, 678
 - Models for Creating Gene Families, 678
 - Mechanisms of Creating Duplications, Deletions, and Inversions, 680
 - Techniques to Measure Chromosomal Change, 682
 - Array Comparative Genomic Hybridization, 682
 - Single Nucleotide Polymorphism (SNP) Microarrays, 683
 - Perspective, 687
 - Pitfalls, 687
 - Web Resources, 688
 - Discussion Questions, 688
 - Problems/Computer Lab, 688
 - Self-Test Quiz, 689
 - Suggested Reading, 690
 - References, 690
- 11 **Eukaryotic Genomes: Fungi, 697**
- Introduction, 697
 - Description and Classification of Fungi, 698
 - Introduction to Budding Yeast *Saccharomyces cerevisiae*, 700
 - Sequencing the Yeast Genome, 701
 - Features of the Budding Yeast Genome, 701
 - Exploring a Typical Yeast Chromosome, 704
 - Gene Duplication and Genome Duplication of *S. cerevisiae*, 708
 - Comparative Analyses of Hemiascomycetes, 712
 - Analysis of Whole Genome Duplication, 712
 - Identification of Functional Elements, 714
 - Analysis of Fungal Genomes, 715
 - Aspergillus*, 715
 - Candida albicans*, 718
 - Cryptococcus neoformans*: Model Fungal Pathogen, 719
 - Atypical Fungus: Microsporidial Parasite *Encephalitozoon cuniculi*, 719
 - Neurospora crassa*, 719
 - First Basidiomycete: *Phanerochaete chrysosporium*, 720
 - Fission Yeast *Schizosaccharomyces pombe*, 721
 - Perspective, 721
 - Pitfalls, 722
 - Web Resources, 722
 - Discussion Questions, 722
 - Problems/Computer Lab, 723

Self-Test Quiz, 723
Suggested Reading, 724
References, 724

18

Eukaryotic Genomes: From Parasites to Primates, 729

Introduction, 729
Protozoans at the Base of the Tree Lacking Mitochondria, 732
Trichomonas, 732
Giardia lamblia: A Human Intestinal Parasite, 733
Genomes of Unicellular Pathogens: *Trypanosomes* and *Leishmania*, 735
Trypanosomes, 735
Leishmania, 736
The Chromalveolates, 738
Malaria Parasite *Plasmodium falciparum* and Other Apicomplexans, 738
Astonishing Ciliophora: *Paramecium* and *Tetrahymena*, 742
Nucleomorphs, 745
Kingdom Stramenopila, 746
Plant Genomes, 748
Overview, 748
Green Algae (*Chlorophyta*), 748
Arabidopsis thaliana Genome, 751
The Second Plant Genome: Rice, 753
The Third Plant Genome: Poplar, 755
The Fourth Plant Genome: Grapevine, 755
Moss, 756
Slime and Fruiting Bodies at the Feet of Metazoans, 756
Social Slime Mold *Dictyostelium discoideum*, 756
Metazoans, 758
Introduction to Metazoans, 758
Analysis of a Simple Animal: The Nematode *Caenorhabditis elegans*, 759
The First Insect Genome: *Drosophila melanogaster*, 761
The Second Insect Genome: *Anopheles gambiae*, 764
Silkworm, 765
Honeybee, 765

The Road to Chordates: The Sea Urchin, 766
750 Million Years Ago: *Ciona intestinalis* and the Road to Vertebrates, 767
450 Million Years Ago: Vertebrate Genomes of Fish, 768
310 Million Years Ago: Dinosaurs and the Chicken Genome, 771
180 Million Years Ago: The Opposum Genome, 772
100 Million Years Ago: Mammalian Radiation from Dog to Cow, 773
80 Million Years Ago: The Mouse and Rat, 774
5 to 50 Million Years Ago: Primate Genomes, 778
Perspective, 781
Pitfalls, 781
Web Resources, 782
Discussion Questions, 782
Problems/Computer Lab, 782
Self-Test Quiz, 783
Suggested Reading, 783
References, 784

19

Human Genome, 791

Introduction, 791
Main Conclusions of Human Genome Project, 792
The ENCODE Project, 793
Gateways to Access the Human Genome, 794
NCBI, 794
Ensembl, 794
University of California at Santa Cruz Human Genome Browser, 798
NHGRI, 800
The Wellcome Trust Sanger Institute, 800
The Human Genome Project, 800
Background of the Human Genome Project, 800
Strategic Issues: Hierarchical Shotgun Sequencing to Generate Draft Sequence, 802
Features of the Genome Sequence, 805
The Broad Genomic Landscape, 806

- Long-Range Variation in GC Content, 806
- CpG Islands, 807
- Comparison of Genetic and Physical Distance, 807
- Repeat Content of the Human Genome, 808
 - Transposon-Derived Repeats, 809
 - Simple Sequence Repeats, 811
 - Segmental Duplications, 811
- Gene Content of the Human Genome, 811
 - Noncoding RNAs, 812
 - Protein-Coding Genes, 812
 - Comparative Proteome Analysis, 814
 - Complexity of Human Proteome, 814
- 24 Human Chromosomes, 816
 - Group A (Chromosomes 1, 2, 3), 818
 - Group B (Chromosomes 4, 5), 822
 - Group C (Chromosomes 6 to 12, X), 823
 - Group D (Chromosomes 13 to 15), 823
 - Group E (Chromosomes 16 to 18), 824
 - Group F (Chromosomes 19, 20), 824
 - Group G (Chromosomes 21, 22, Y), 824
 - The Mitochondrial Genome, 825
- Variation: Sequencing Individual Genomes, 825
- Variation: SNPs to Copy Number Variants, 827
- Perspective, 831
- Pitfalls, 831
- Discussion Questions, 832
- Problems/Computer Lab, 832
- Self-Test Quiz, 833
- Suggested Reading, 833
- References, 834
- Garrod's View of Disease, 842
- Classification of Disease, 843
- NIH Disease Classification: MeSH Terms, 845
- Four Categories of Disease, 846
 - Monogenic Disorders, 847
 - Complex Disorders, 851
 - Genomic Disorders, 852
 - Environmentally Caused Disease, 855
 - Other Categories of Disease, 857
- Disease Databases, 859
 - OMIM: Central Bioinformatics Resource for Human Disease, 859
 - Locus-Specific Mutation Databases, 862
 - The PhenCode Project, 865
- Four Approaches to Identifying Disease-Associated Genes, 866
 - Linkage Analysis, 866
 - Genome-Wide Association Studies, 867
 - Identification of Chromosomal Abnormalities, 868
 - Genomic DNA Sequencing, 869
- Human Disease Genes in Model Organisms, 870
 - Human Disease Orthologs in Nonvertebrate Species, 870
 - Human Disease Orthologs in Rodents, 876
 - Human Disease Orthologs in Primates, 878
 - Human Disease Genes and Substitution Rates, 878
- Functional Classification of Disease Genes, 880
 - Perspective, 882
 - Pitfalls, 882
 - Web Resources, 882
 - Discussion Questions, 884
 - Problems, 884
 - Self-Test Quiz, 885
 - Suggested Reading, 885
 - References, 886
- Human Disease, 839**
 - Human Genetic Disease: A Consequence of DNA Variation, 839
 - A Bioinformatics Perspective on Human Disease, 841
- Glossary, 891**
- Answers to Self-Test Quizzes, 909**
- Author Index, 911**
- Subject Index, 913**