

# Contents

Preface to the Third Edition, xxxi

About the Companion Website, xxxiii

## **PART I** ANALYZING DNA, RNA, AND PROTEIN SEQUENCES

### **1** Introduction, 3

Organization of the Book, 4

Bioinformatics: The Big Picture, 5

    A Consistent Example: Globins, 6

Organization of the Chapters, 8

Suggestions For Students and Teachers: Web Exercises, Find-a-Gene, and Characterize-a-Genome, 9

Bioinformatics Software: Two Cultures, 10

    Web-Based Software, 11

    Command-Line Software, 11

    Bridging the Two Cultures, 12

    New Paradigms for Learning Programming for Bioinformatics, 13

    Reproducible Research in Bioinformatics, 14

Bioinformatics and Other Informatics Disciplines, 15

Advice for Students, 15

    Suggested Reading, 15

    References, 16

### **2** Access to Sequence Data and Related Information, 19

Introduction to Biological Databases, 19

Centralized Databases Store DNA Sequences, 20

Contents of DNA, RNA, and Protein Databases, 24

    Organisms in GenBank/EMBL-Bank/DDBJ, 24

    Types of Data in GenBank/EMBL-Bank/DDBJ, 26

    Genomic DNA Databases, 27

        DNA-Level Data: Sequence-Tagged Sites (STSs), 27

        DNA-Level Data: Genome Survey Sequences (GSSs), 27

        DNA-Level Data: High-Throughput Genomic Sequence (HTGS), 27

    RNA data, 27

        RNA-Level Data: cDNA Databases Corresponding to Expressed Genes, 27

        RNA-Level Data: Expressed Sequence Tags (ESTs), 28

        RNA-Level Data: UniGene, 28

- Access to Information: Protein Databases, 29
  - UniProt, 31
- Central Bioinformatics Resources: NCBI and EBI, 31
  - Introduction to NCBI, 31
  - The European Bioinformatics Institute (EBI), 32
- Ensembl, 34
- Access to Information: Accession Numbers to Label and Identify Sequences, 34
  - The Reference Sequence (RefSeq) Project, 36
  - RefSeqGene and the Locus Reference Genomic Project, 37
  - The Consensus Coding Sequence CCDS Project, 37
  - The Vertebrate Genome Annotation (VEGA) Project, 37
- Access to Information via Gene Resource at NCBI, 38
  - Relationship Between NCBI Gene, Nucleotide, and Protein Resources, 41
  - Comparison of NCBI's Gene and UniGene, 41
  - NCBI's Gene and HomoloGene, 42
- Command-Line Access to Data at NCBI, 42
  - Using Command-Line Software, 42
  - Accessing NCBI Databases with EDirect, 45
    - EDirect Example 1, 46
    - EDirect Example 2, 46
    - EDirect Example 3, 46
    - EDirect Example 4, 47
    - EDirect Example 5, 48
    - EDirect Example 6, 48
    - EDirect Example 7, 48
- Access to Information: Genome Browsers, 49
  - Genome Builds, 49
  - The University of California, Santa Cruz (UCSC) Genome Browser, 50
  - The Ensembl Genome Browser, 50
  - The Map Viewer at NCBI, 52
- Examples of How to Access Sequence Data: Individual Genes/Proteins, 52
  - Histones, 52
  - HIV-1 pol, 53
- How to Access Sets of Data: Large-Scale Queries of Regions and Features, 54
  - Thinking About One Gene (or Element) Versus Many Genes (Elements), 54
  - The BioMart Project, 54
  - Using the UCSC Table Browser, 54
  - Custom Tracks: Versatility of the BED File, 56
  - Galaxy: Reproducible, Web-Based, High-Throughput Research, 57
- Access to Biomedical Literature, 58
  - Example of PubMed Search, 59
- Perspective, 59
- Pitfalls, 60
- Advice for Students, 60

Web Resources, 60

Discussion Questions, 61

Problems/Computer Lab, 61

Self-Test Quiz, 63

Suggested Reading, 64

References, 64

### 3 Pairwise Sequence Alignment, 69

Introduction, 69

Protein Alignment: Often More Informative than DNA Alignment, 70

Definitions: Homology, Similarity, Identity, 70

Gaps, 78

Pairwise Alignment, Homology, and Evolution of Life, 78

Scoring Matrices, 79

Dayhoff Model Step 1 (of 7): Accepted Point Mutations, 79

Dayhoff Model Step 2 (of 7): Frequency of Amino Acids, 79

Dayhoff Model Step 3 (of 7): Relative Mutability of Amino Acids, 80

Dayhoff Model Step 4 (of 7): Mutation Probability Matrix for the  
Evolutionary Distance of 1 PAM, 82

Dayhoff Model Step 5 (of 7): PAM250 and Other PAM Matrices, 84

Dayhoff Model Step 6 (of 7): From a Mutation Probability Matrix to a  
Relatedness Odds Matrix, 88

Dayhoff Model Step 7 (of 7): Log-Odds Scoring Matrix, 89

Practical Usefulness of PAM Matrices in Pairwise Alignment, 91

Important Alternative to PAM: BLOSUM Scoring Matrices, 91

Pairwise Alignment and Limits of Detection: The "Twilight Zone", 94

Alignment Algorithms: Global and Local, 96

Global Sequence Alignment: Algorithm of Needleman and Wunsch, 96

Step 1: Setting Up a Matrix, 96

Step 2: Scoring the Matrix, 97

Step 3: Identifying the Optimal Alignment, 99

Local Sequence Alignment: Smith and Waterman Algorithm, 101

Rapid, Heuristic Versions of Smith–Waterman: FASTA and BLAST, 103

Basic Local Alignment Search Tool (BLAST), 104

Pairwise Alignment with Dotplots, 104

The Statistical Significance of Pairwise Alignments, 106

Statistical Significance of Global Alignments, 106

Statistical Significance of Local Alignments, 108

Percent Identity and Relative Entropy, 108

Perspective, 110

Pitfalls, 112

Advice for Students, 112

Web Resources, 112

Discussion Questions, 113

Problems/Computer Lab, 113

Self-Test Quiz, 114  
Suggested Reading, 115  
References, 116

- 4 Basic Local Alignment Search Tool (BLAST) , 121**  
Introduction, 121  
BLAST Search Steps, 124  
    Step 1: Specifying Sequence of Interest, 124  
    Step 2: Selecting BLAST Program, 124  
    Step 3: Selecting a Database, 126  
    Step 4a: Selecting Optional Search Parameters, 127  
    Step 4b: Selecting Formatting Parameters, 132  
    Stand-Alone BLAST, 135  
BLAST Algorithm Uses Local Alignment Search Strategy, 138  
    BLAST Algorithm Parts: List, Scan, Extend, 138  
    BLAST Algorithm: Local Alignment Search Statistics and *E* Value, 141  
    Making Sense of Raw Scores with Bit Scores, 143  
    BLAST Algorithm: Relation Between *E* and *p* Values, 143  
BLAST Search Strategies, 145  
    General Concepts, 145  
    Principles of BLAST Searching, 146  
        How to Evaluate the Significance of Results, 146  
        How to Handle Too Many Results, 150  
        How to Handle Too Few Results, 150  
    BLAST Searching with Multidomain Protein: HIV-1 Pol, 151  
Using Blast For Gene Discovery: Find-a-Gene, 155  
Perspective, 159  
Pitfalls, 160  
Advice for Students, 160  
Web Resources, 160  
    Discussion Questions, 160  
    Problems/Computer Lab, 160  
    Self-Test Quiz, 161  
    Suggested Reading, 162  
    References, 163

- 5 Advanced Database Searching , 167**  
Introduction, 167  
Specialized BLAST Sites, 168  
    Organism-Specific BLAST Sites, 168  
        Ensembl BLAST, 168  
        Wellcome Trust Sanger Institute, 170  
Specialized BLAST-Related Algorithms, 170  
    WU BLAST 2.0, 170  
    European Bioinformatics Institute (EBI), 170

- Specialized NCBI BLAST Sites, 170
- BLAST of Next-Generation Sequence Data, 170
- Finding Distantly Related Proteins: Position-Specific Iterated BLAST (PSI-BLAST) and DELTA-BLAST, 171
  - PSI-BLAST Errors: Problem of Corruption, 177
  - Reverse Position-Specific BLAST, 177
  - Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST), 177
  - Assessing Performance of PSI-BLAST and DELTA-BLAST, 179
  - Pattern-Hit Initiated BLAST (PHI-BLAST), 179
- Profile Searches: Hidden Markov Models, 181
  - HMMER Software: Command-Line and Web-Based, 184
- BLAST-Like Alignment Tools to Search Genomic DNA Rapidly, 186
  - Benchmarking to Assess Genomic Alignment Performance, 187
  - PatternHunter: Nonconsecutive Seeds Boost Sensitivity, 188
  - BLASTZ, 188
  - Enredo and Pecan, 191
  - MegaBLAST and Discontinuous MegaBLAST, 191
  - BLAST-Like Tool (BLAT), 192
  - LAGAN, 192
  - SSAHA2, 194
- Aligning Next-Generation Sequence (NGS) Reads to a Reference Genome, 194
  - Alignment Based on Hash Tables, 194
  - Alignment Based on the Burrows–Wheeler Transform, 196
- Perspective, 197
- Pitfalls, 197
- Advice For Students, 198
- Web Resources, 198
  - Discussion Questions, 198
  - Problems/Computer Lab, 198
  - Self-Test Quiz, 199
  - Suggested Reading, 200
  - References, 201

## 6 Multiple Sequence Alignment, 205

- Introduction, 205
  - Definition of Multiple Sequence Alignment, 206
  - Typical Uses and Practical Strategies of Multiple Sequence Alignment, 207
  - Benchmarking: Assessment of Multiple Sequence Alignment Algorithms, 207
- Five Main Approaches to Multiple Sequence Alignment, 208
  - Exact Approaches to Multiple Sequence Alignment, 208
  - Progressive Sequence Alignment, 208
  - Iterative Approaches, 214
  - Consistency-Based approaches, 218
  - Structure-Based Methods, 220
- Benchmarking Studies: Approaches, Findings, Challenges, 221

- Databases of Multiple Sequence Alignments, 222
  - Pfam: Protein Family Database of Profile HMMs, 223
  - SMART, 224
  - Conserved Domain Database, 226
  - Integrated Multiple Sequence Alignment Resources: InterPro and iProClass, 226
  - Multiple Sequence Alignment Database Curation: Manual Versus Automated, 227
- Multiple Sequence Alignments of Genomic Regions, 227
  - Analyzing Genomic DNA Alignments via UCSC, 229
  - Analyzing Genomic DNA Alignments via Galaxy, 229
  - Analyzing Genomic DNA Alignments via Ensembl, 231
  - Alignathon Competition to Assess Whole-Genome Alignment Methods, 231
- Perspective, 234
- Pitfalls, 234
- Advice for Students, 235
  - Discussion Questions, 235
  - Problems/Computer Lab, 235
  - Self-Test Quiz, 237
  - Suggested Reading, 238
  - References, 239

## 7 Molecular Phylogeny and Evolution, 245

- Introduction to Molecular Evolution, 245
- Principles of Molecular Phylogeny and Evolution, 246
  - Goals of Molecular Phylogeny, 246
  - Historical Background, 247
  - Molecular Clock Hypothesis, 250
  - Positive and Negative Selection, 254
  - Neutral Theory of Molecular Evolution, 258
- Molecular Phylogeny: Properties of Trees, 259
  - Topologies and Branch Lengths of Trees, 259
  - Tree Roots, 262
  - Enumerating Trees and Selecting Search Strategies, 263
- Type of Trees, 266
  - Species Trees versus Gene/Protein Trees, 266
  - DNA, RNA, or Protein-Based Trees, 268
- Five Stages of Phylogenetic Analysis, 270
  - Stage 1: Sequence Acquisition, 270
  - Stage 2: Multiple Sequence Alignment, 271
  - Stage 3: Models of DNA and Amino Acid Substitution, 272
  - Stage 4: Tree-Building Methods, 281
    - Distance-Based, 282
    - Phylogenetic Inference: Maximum Parsimony, 287

- Model-Based Phylogenetic Inference: Maximum Likelihood, 289
- Tree Inference: Bayesian Methods, 290
- Stage 5: Evaluating Trees, 293
- Perspective, 295
- Pitfalls, 295
- Advice for Students, 296
- Web Resources, 297
  - Discussion Questions, 297
  - Problems/Computer Lab, 297
  - Self-Test Quiz, 298
  - Suggested Reading, 298
  - References, 299

## 8 DNA: The Eukaryotic Chromosome, 307

- Introduction, 308
  - Major Differences between Eukaryotes and Bacteria and Archaea, 308
- General Features of Eukaryotic Genomes and Chromosomes, 310
  - C Value Paradox: Why Eukaryotic Genome Sizes Vary So Greatly, 312
  - Organization of Eukaryotic Genomes into Chromosomes, 310
  - Analysis of Chromosomes Using Genome Browsers, 314
  - Analysis of Chromosomes Using BioMart and biomaRt, 314
    - Example 1, 317
    - Example 2, 319
    - Example 3, 319
    - Example 4, 319
    - Example 5, 320
  - Analysis of Chromosomes by the ENCODE Project, 320
  - Critiques of ENCODE: the C Value Paradox Revisited and the Definition of Function, 322
- Repetitive DNA Content of Eukaryotic Chromosomes, 323
  - Eukaryotic Genomes Include Noncoding and Repetitive DNA Sequences, 323
    - Interspersed Repeats (Transposon-Derived Repeats), 325
    - Processed Pseudogenes, 326
    - Simple Sequence Repeats, 331
    - Segmental Duplications, 331
    - Blocks of Tandemly Repeated Sequences, 333
- Gene Content of Eukaryotic Chromosomes, 334
  - Definition of Gene, 334
  - Finding Genes in Eukaryotic Genomes, 336
  - Finding Genes in Eukaryotic Genomes: EGASP Competition, 339
  - Three Resources for Studying Protein-Coding Genes: RefSeq, UCSC Genes, GENCODE, 340
  - Protein-Coding Genes in Eukaryotes: New Paradox, 342

- Regulatory Regions of Eukaryotic Chromosomes, 342
  - Databases of Genomic Regulatory Factors, 342
  - Ultraconserved Elements, 345
  - Nonconserved Elements, 345
- Comparison of Eukaryotic DNA, 346
- Variation in Chromosomal DNA, 347
  - Dynamic Nature of Chromosomes: Whole-Genome Duplication, 347
  - Chromosomal Variation in Individual Genomes, 349
  - Structural Variation: Six Types, 351
    - Inversions, 351
    - Mechanisms of Creating Duplications, Deletions, and Inversions, 351
    - Models for Creating Gene Families, 353
  - Chromosomal Variation in Individual Genomes: SNPs, 354
- Techniques to Measure Chromosomal Change, 355
  - Array Comparative Genomic Hybridization, 356
  - SNP Microarrays, 356
  - Next-Generation Sequencing, 359
- Perspective, 359
- Pitfalls, 359
- Advice to Students, 360
- Web Resources, 360
  - Discussion Questions, 361
  - Problems/Computer Lab, 361
  - Self-Test Quiz, 364
  - Suggested Reading, 365
  - References, 366
  
- 9 Analysis of Next-Generation Sequence Data, 377**
  - Introduction, 378
  - DNA Sequencing Technologies, 377
    - Sanger Sequencing, 379
    - Next-Generation Sequencing, 379
    - Cyclic Reversible Termination: Illumina, 382
    - Pyrosequencing, 384
    - Sequencing by Ligation: Color Space with ABI SOLiD, 385
    - Ion Torrent: Genome Sequencing by Measuring pH, 387
    - Pacific Biosciences: Single-Molecule Sequencing with Long Read Lengths, 387
    - Complete Genomics: Self-Assembling DNA Nanoarrays, 387
  - Analysis of Next-Generation Sequencing of Genomic DNA, 387
    - Overview of Next-Generation Sequencing Data Analysis, 387
    - Topic 1: Experimental Design and Sample Preparation, 389
    - Topic 2: From Generating Sequence Data to FASTQ, 390
      - Finding and Viewing FASTQ files, 392
      - Quality Assessment of FASTQ data, 393
      - FASTG: A Richer Format than FASTQ, 394

- Topic 3: Genome Assembly, 394
  - Competitions and Critical Evaluations of the Performance of Genome Assemblers, 396
  - The End of Assembly: Standards for Completion, 398
- Topic 4: Sequence Alignment, 399
  - Alignment of Repetitive DNA, 400
  - Genome Analysis Toolkit (GATK) Workflow: Alignment with BWA, 401
- Topic 5: The SAM/BAM Format and SAMtools, 402
  - Calculating Read Depth, 405
  - Finding and Viewing BAM/SAM files, 405
  - Compressed Alignments: CRAM File Format, 406
- Topic 6: Variant Calling: Single-Nucleotide Variants and Indels, 408
- Topic 7: Variant Calling: Structural Variants, 409
- Topic 8: Summarizing Variation: The VCF Format and VCFtools, 410
  - Finding and Viewing VCF files, 413
- Topic 9: Visualizing and Tabulating Next-Generation Sequence Data, 413
- Topic 10: Interpreting the Biological Significance of Variants, 417
- Topic 11: Storing Data in Repositories, 421
- Specialized Applications of Next-Generation Sequencing, 421
- Perspective, 422
- Pitfalls, 423
- Advice for Students, 423
- Web Resources, 424
  - Discussion Questions, 424
  - Problems/Computer Lab, 424
  - Self-Test Quiz, 425
  - Suggested Reading, 425
  - References, 425

## 10 Bioinformatic Approaches to Ribonucleic Acid (RNA), 433

- Introduction to RNA, 433
- Noncoding RNA, 436
  - Noncoding RNAs in the Rfam Database, 436
  - Transfer RNA, 438
  - Ribosomal RNA, 441
  - Small Nuclear RNA, 445
  - Small Nucleolar RNA, 445
  - MicroRNA, 445
  - Short Interfering RNA, 447
  - Long Noncoding RNA (lncRNA), 447
  - Other Noncoding RNA, 448
  - Noncoding RNAs in the UCSC Genome and Table Browser, 448
- Introduction to Messenger RNA, 450
  - mRNA: Subject of Gene Expression Studies, 450
  - Low- and High-Throughput Technologies to Study mRNAs, 452

- Analysis of Gene Expression in cDNA Libraries, 455
  - Full-Length cDNA Projects, 459
  - BodyMap2 and GTEx: Measuring Gene Expression Across the Body, 459
  - Microarrays and RNA-Seq: Genome-Wide Measurement of Gene Expression, 460
    - Stage 1: Experimental Design for Microarrays and RNA-seq, 461
    - Stage 2: RNA Preparation and Probe Preparation, 461
    - Stage 3: Data Acquisition, 464
      - Hybridization of Labeled Samples to DNA Microarrays, 464
      - Data acquisition for RNA-seq, 465
    - Stage 4: Data Analysis, 465
    - Stage 5: Biological Confirmation, 465
    - Microarray and RNA-seq Databases, 465
    - Further Analyses, 465
  - Interpretation of RNA Analyses, 466
    - The Relationship between DNA, mRNA, and Protein Levels, 466
    - The Pervasive Nature of Transcription, 467
    - eQTLs: Understanding the Genetic Basis of Variation in Gene Expression through Combined RNA-seq and DNA-seq, 468
  - Perspective, 469
  - Pitfalls, 470
  - Advice to Students, 470
  - Web Resources, 470
    - Discussion Questions, 471
    - Problems/Computer Lab, 471
    - Self-Test Quiz, 471
    - Suggested Reading, 472
    - References, 473
- 11 Gene Expression: Microarray and RNA-seq Data Analysis, 479**
- Introduction, 479
  - Microarray Analysis Method 1: GEO2R at NCBI, 482
    - GEO2R Executes a Series of R Scripts, 482
    - GEO2R Identifies the Chromosomal Origin of Regulated Transcripts, 485
    - GEO2R Normalizes Data, 486
    - GEO2R uses RMA Normalization for Accuracy and Precision, 488
    - Fold Change (Expression Ratios), 490
    - GEO2R Performs >22,000 Statistical Tests, 490
    - GEO2R Offers Corrections for Multiple Comparisons, 494
  - Microarray Analysis Method 2: Partek, 495
    - Importing Data, 496
    - Quality Control, 496
    - Adding Sample Information, 497
    - Sample Histogram, 498
    - Scatter Plots and MA Plots, 498

Working with Log<sub>2</sub> Transformed Microarray Data, 498

Exploratory Data Analysis with Principal Components  
Analysis (PCA), 498

Performing ANOVA in Partek, 501

From *t*-test to ANOVA, 503

Microarray Analysis Method 3: Analyzing a GEO Dataset with R, 504

Setting up the Analyses, 504

Reading CEL Files and Normalizing with RMA, 506

Identifying Differentially Expressed Genes (Limma), 508

Microarray Analysis and Reproducibility, 510

Microarray Data Analysis: Descriptive Statistics, 511

Hierarchical Cluster Analysis of Microarray Data, 511

Partitioning Methods for Clustering: k-Means Clustering, 516

Multidimensional Scaling Compared to Principal Components  
Analysis, 517

Clustering Strategies: Self-Organizing Maps, 517

Classification of Genes or Samples, 517

RNA-Seq, 519

Setting up a TopHat and CuffLinks Sample Protocol, 523

TopHat to Map Reads to a Reference Genome, 524

Cufflinks to Assemble Transcripts, 525

Cuffdiff to Determine Differential Expression, 525

CummeRbund to Visualize RNA-seq Results, 526

RNA-seq Genome Annotation Assessment Project  
(RGASP), 527

Functional Annotation of Microarray Data, 528

Perspective, 529

Pitfalls, 530

Advice for Students, 531

Suggested Reading, 531

Problems/Computer Lab, 532

Self-Test Quiz, 532

Suggested Reading, 533

References, 534

## 12 Protein Analysis and Proteomics, 539

Introduction, 539

Protein Databases, 540

Community Standards for Proteomics Research, 542

Evaluating the State-of-the-Art: ABRF analytic  
challenges, 542

Techniques for Identifying Proteins, 543

Direct Protein Sequencing, 543

Gel Electrophoresis, 543

Mass Spectrometry, 547

#### Four Perspectives on Proteins, 551

Perspective 1: Protein Domains and Motifs: Modular Nature of Proteins, 552

Added Complexity of Multidomain Proteins, 557

Protein Patterns: Motifs or Fingerprints Characteristic of Proteins, 557

Perspective 2: Physical Properties of Proteins, 559

Accuracy of Prediction Programs, 561

Proteomic Approaches to Phosphorylation, 563

Proteomic Approaches to Transmembrane Regions, 565

Introduction to Perspectives 3 and 4: Gene Ontology Consortium, 567

Perspective 3: Protein Localization, 568

Perspective 4: Protein Function, 570

Perspective, 575

Pitfalls, 575

Advice for Students, 575

Web Resources, 576

Discussion Questions, 578

Problems/Computer Lab, 578

Self-Test Quiz, 579

Suggested Reading, 580

References, 580

## 13 Protein Structure, 589

Overview of Protein Structure, 589

Protein Sequence and Structure, 590

Biological Questions Addressed by Structural Biology: Globins, 591

Principles of Protein Structure, 591

Primary Structure, 591

Secondary Structure, 594

Tertiary Protein Structure: Protein-Folding Problem, 598

Structural Genomics, the Protein Structure Initiative, and Target Selection, 600

Protein Data Bank, 602

Accessing PDB Entries at NCBI Website, 606

Integrated Views of Universe of Protein Folds, 609

Taxonomic System for Protein Structures: SCOP Database, 610

CATH Database, 613

Dali Domain Dictionary, 615

Comparison of Resources, 617

Protein Structure Prediction, 617

Homology Modeling (Comparative Modeling), 618

Fold Recognition (Threading), 619

*Ab Initio* Prediction (Template-Free Modeling), 621

A Competition to Assess Progress in Structure Prediction, 621

Intrinsically Disordered Proteins, 622

Protein Structure and Disease, 622

Perspective, 625

Pitfalls, 625

Advice for Students, 625

Discussion Questions, 625

Problems/Computer Lab, 626

Self-Test Quiz, 627

Suggested Reading, 628

References, 628

## 14 Functional Genomics, 635

Introduction to Functional Genomics, 635

The Relationship Between Genotype and Phenotype, 637

Eight-Model Organisms For Functional Genomics, 638

1. The Bacterium *Escherichia coli*, 639

2. The Yeast *Saccharomyces cerevisiae*, 640

3. The Plant *Arabidopsis thaliana*, 643

4. The Nematode *Caenorhabditis elegans*, 643

5. The Fruit Fly *Drosophila melanogaster*, 645

6. The Zebrafish *Danio rerio*, 645

7. The Mouse *Mus musculus*, 646

8. *Homo sapiens*: Variation in Humans, 647

Functional Genomics Using Reverse and Forward Genetics, 648

Reverse Genetics: Mouse Knockouts and the  $\beta$ -Globin Gene, 650

Reverse Genetics: Knocking Out Genes in Yeast Using Molecular Barcodes, 653

Reverse Genetics: Random Insertional Mutagenesis (Gene Trapping), 657

Reverse Genetics: Insertional Mutagenesis in Yeast, 660

Reverse Genetics: Gene Silencing by Disrupting RNA, 662

Forward Genetics: Chemical Mutagenesis, 665

Comparison of Reverse and Forward Genetics, 665

Functional Genomics and the Central Dogma, 666

Approaches to Function and Definitions of Function, 646

Functional Genomics and DNA: Integrating Information, 668

Functional Genomics and RNA, 668

Functional Genomics and Protein, 670

Proteomics Approaches to Functional Genomics, 670

Functional Genomics and Protein: Critical Assessment of Protein Function Annotation, 672

Protein-Protein Interactions, 672

Yeast Two-Hybrid System, 673

Protein Complexes: Affinity Chromatography and Mass Spectrometry, 675

Protein-Protein Interaction Databases, 676

From Pairwise Interactions to Protein Networks, 678

Assessment of Accuracy, 680

Choice of Data, 680

- Experimental Organism, 680
- Variation in Pathways, 681
- Categories of Maps, 681
- Pathways, Networks, and Integration: Bioinformatics Resources, 682
- Perspective, 685
- Pitfalls, 686
- Advice for Students, 686
- Web Resources, 686
  - Discussion Questions, 686
  - Problems/Computer Lab, 686
  - Self-Test Quiz, 687
  - Suggested Reading, 688
  - References, 688

## PART III GENOME ANALYSIS

### 15 Genomes Across the Tree of Life, 699

- Introduction, 700
  - Five Perspectives on Genomics, 701
  - Brief History of Systematics, 701
  - History of Life on Earth, 705
  - Molecular Sequences as the Basis of the Tree of Life, 705
  - Role of Bioinformatics in Taxonomy, 709
- Prominent Web Resources, 710
  - Ensembl Genomes, 710
  - NCBI Genome, 710
  - Genome Portal of DOE JGI and the Integrated Microbial Genomes, 710
  - Genomes On Line Database (GOLD), 710
  - UCSC, 710
- Genome-Sequencing Projects: Chronology, 711
  - Brief Chronology, 711
  - 1976–1978: First Bacteriophage and Viral Genomes, 711
  - 1981: First Eukaryotic Organellar Genome, 712
  - 1986: First Chloroplast Genomes, 714
  - 1992: First Eukaryotic Chromosome, 715
  - 1995: Complete Genome of Free-Living Organism, 715
  - 1996: First Eukaryotic Genome, 715
  - 1997: *Escherichia coli*, 715
  - 1998: First Genome of Multicellular Organism, 716
  - 1999: Human Chromosome, 716
  - 2000: Fly, Plant, and Human Chromosome 21, 716
  - 2001: Draft Sequences of Human Genome, 716
  - 2002: Continuing Rise in Completed Genomes, 717
  - 2003: HapMap, 717
  - 2004: Chicken, Rat, and Finished Human Sequences, 717

- 2005: Chimpanzee, Dog, Phase I HapMap, 718
- 2006: Sea Urchin, Honeybee, dbGaP, 718
- 2007: Rhesus Macaque, First Individual Human Genome, ENCODE Pilot, 718
- 2008: Platypus, First Cancer Genome, First Personal Genome Using NGS, 718
- 2009: Bovine, First Human Methlyome Map, 718
- 2010: 1000 Genomes Pilot, Neandertal , Exome Sequencing to Find Disease Genes, 719
- 2011: A Vision for the Future of Genomics, 719
- 2012: Denisovan Genome, Bonobo, and 1000 Genomes Project, 719
- 2013: The Simplest Animal and a 700,000-Year-Old Horse, 719
- 2014: Mouse ENCODE, Primates, Plants, and Ancient Hominids, 719
- 2015: Diversity in Africa, 720
- Genome Analysis Projects: Introduction, 720
  - Large-Scale Genomics Projects, 721
  - Criteria for Selection of Genomes for Sequencing, 722
    - Genome Size, 722
    - Cost, 722
    - Relevance to Human Disease, 723
    - Relevance to Basic Biological Questions, 724
    - Relevance to Agriculture, 724
    - Sequencing of One Versus Many Individuals from a Species, 724
  - Role of Comparative Genomics, 724
  - Resequencing Projects, 725
  - Ancient DNA Projects, 725
  - Metagenomics Projects, 725
- Genome Analysis Projects: Sequencing, 728
  - Genome-Sequencing Centers, 728
  - Trace Archive: Repository for Genome Sequence Data, 728
  - HTGS Archive: Repository for Unfinished Genome Sequence Data, 730
- Genome Analysis Projects: Assembly, 730
  - Four Approaches to Genome Assembly, 730
  - Genome Assembly: From FASTQ to Contigs with Velvet, 733
  - Comparative Genome Assembly: Mapping Contigs to Known Genomes, 734
  - Finishing: When Has a Genome Been Fully Sequenced?, 735
  - Genome Assembly: Measures of Success, 735
  - Genome Assembly: Challenges, 735
- Genome Analysis Projects: Annotation, 737
  - Annotation of Genes in Eukaryotes: Ensembl Pipeline, 738
  - Annotation of Genes in Eukaryotes: NCBI Pipeline, 739
  - Core Eukaryotic Genes Mapping Approach (CEGMA), 739
  - Assemblies from the Genome Reference Consortium, 741
  - Assembly Hubs and Transfers at UCSC, Ensembl, and NCBI, 741
  - Annotation of Genes in Bacteria and Archaea, 741
  - Genome Annotation Standards, 741
  - Perspective, 742

Pitfalls, 742

Advice for Students, 743

Discussion Questions, 743

Problems/Computer Lab, 743

Self-Test Quiz, 745

Suggested Reading, 743

References, 745

## 16 Completed Genomes: Viruses, 755

Introduction, 755

International Committee on Taxonomy of Viruses (ICTV) and  
Virus Species, 756

Classification of Viruses, 758

Classification of Viruses Based on Morphology, 758

Classification of Viruses Based on Nucleic Acid Composition, 758

Classification of Viruses Based on Genome Size, 758

Classification of Viruses Based on Disease Relevance, 760

Diversity and Evolution of Viruses, 762

Metagenomics and Virus Diversity, 764

Bioinformatics Approaches to Problems in Virology, 765

Human Immunodeficiency Virus (HIV), 766

NCBI and LANL resources for HIV-1, 766

Influenza Virus, 771

Measles Virus, 774

Ebola Virus, 775

Herpesvirus: From Phylogeny to Gene Expression, 776

The Pairwise Sequence Comparison (PASC) Tool, 780

Giant Viruses, 782

Comparing genomes with MUMmer, 783

Perspectives, 785

Pitfalls, 786

Advice for Students, 786

Web Resources, 786

Discussion Questions, 787

Problems/Computer Lab, 787

Self-Test Quiz, 788

Suggested Reading, 789

References, 789

## 17 Completed Genomes: Bacteria and Archaea, 797

Introduction, 797

Classification of Bacteria and Archaea, 798

Classification of Bacteria by Morphological Criteria, 800

Classification of Bacteria and Archaea Based on Genome Size and  
Geometry, 801

- Classification of Bacteria and Archaea Based on Lifestyle, 805
- Classification of Bacteria Based on Human Disease Relevance, 808
- Classification of Bacteria and Archaea Based on Ribosomal RNA Sequences, 809
- Classification of Bacteria and Archaea Based on Other Molecular Sequences, 810
- The Human Microbiome, 811
- Analysis of Bacterial and Archaeal Genomes, 814
  - Nucleotide Composition, 817
  - Finding Genes, 819
    - Interpolated Context Model (ICM), 822
    - GLIMMER3, 824
  - Challenges of Bacterial and Archaeal Gene Prediction, 825
  - Gene Annotation, 825
  - Lateral Gene Transfer, 827
- Comparison of Bacterial Genomes, 830
  - TaxPlot, 830
  - MUMmer, 833
- Perspective, 834
- Pitfalls, 835
- Advice for Students, 835
- Web Resources, 835
  - Discussion Questions, 836
  - Problems/Computer Lab, 836
  - Self-Test Quiz, 836
  - Suggested Reading, 837
  - References, 837

## 18 Eukaryotic Genomes: Fungi, 847

- Introduction, 847
- Description and Classification of Fungi, 848
- Introduction to Budding Yeast *Saccharomyces Cerevisiae*, 849
  - Sequencing Yeast Genome, 851
  - Features of Budding Yeast Genome, 851
  - Exploring Typical Yeast Chromosome, 854
    - Web Resources for Analyzing a Chromosome, 854
    - Exploring Variation in a Chromosome with Command-Line Tools, 857
    - Finding Genes in a Chromosome with Command-Line Tools, 858
    - Properties of Yeast Chromosome XII, 860
- Gene Duplication and Genome Duplication of *S. cerevisiae*, 860
- Comparative Analyses of Hemiascomycetes, 865
  - Comparative Analyses of Whole-Genome Duplication, 866
  - Identification of Functional Elements, 868
- Analysis of Fungal Genomes, 869
  - Fungi in the Human Microbiome, 870

- Aspergillus, 871
- Candida albicans, 871
- Cryptococcus neoformans*: model fungal pathogen, 872
- Atypical Fungus: Microsporidial Parasite *Encephalitozoon cuniculi*, 873
- Neurospora crassa, 873
- First Basidiomycete: *Phanerochaete chrysosporium*, 875
- Fission Yeast *Schizosaccharomyces pombe*, 875
- Other Fungal Genomes, 876
- Ten Leading Fungal Plant Pathogens, 876
- Perspective, 876
- Pitfalls, 877
- Advice for Students, 877
- Web Resources, 877
  - Discussion Questions, 877
  - Problems/Computer Lab, 878
  - Self-Test Quiz, 879
  - Suggested Reading, 880
  - References, 880

## 19 Eukaryotic Genomes: From Parasites to Primates, 887

- Introduction, 887
- Protozoans at Base of Tree Lacking Mitochondria, 890
  - Trichomonas*, 890
  - Giardia lamblia*: A Human Intestinal Parasite, 891
- Genomes of Unicellular Pathogens: Trypanosomes and *Leishmania*, 890
  - Trypanosomes, 892
  - Leishmania*, 894
- The Chromalveolates, 895
  - Malaria Parasite *Plasmodium falciparum*, 895
  - More Apicomplexans, 898
  - Astonishing Ciliophora: *Paramecium* and *Tetrahymena*, 899
  - Nucleomorphs, 902
  - Kingdom Stramenopila, 904
- Plant Genomes, 906
  - Overview, 906
  - Green Algae (*Chlorophyta*), 908
  - Arabidopsis thaliana* Genome, 910
  - The Second Plant Genome: Rice, 913
  - Third Plant: Poplar, 914
  - Fourth Plant: Grapevine, 915
  - Giant and Tiny Plant Genomes, 915
  - Hundreds More Land Plant Genomes, 915
  - Moss, 916
- Slime and Fruiting Bodies at the Feet of Metazoans, 916
  - Social Slime Mold *Dictyostelium discoideum*, 916

## Metazoans, 917

Introduction to Metazoans, 917

900 MYA: the Simple Animal *Caenorhabditis elegans*, 918

900 MYA: *Drosophila melanogaster* (First Insect Genome), 919

900 MYA: *Anopheles gambiae* (Second Insect Genome), 921

900 MYA: Silkworm and Butterflies, 922

900 MYA: Honeybee, 923

900 MYA: A Swarm of Insect Genomes, 923

840 MYA: A Sea Urchin on the Path to Chordates, 924

800 MYA: *Ciona intestinalis* and the Path to Vertebrates, 925

450 MYA: Vertebrate Genomes of Fish, 926

350 MYA: Frogs, 929

320 MYA: Reptiles (Birds, Snakes, Turtles, Crocodiles), 929

180 MYA: The Platypus and Opposum Genomes, 931

100 MYA: Mammalian Radiation from Dog to Cow, 933

80 MYA: The Mouse and Rat, 934

5–50 MYA: Primate Genomes, 937

Perspective, 940

Pitfalls, 941

Advice for Students, 941

Web Resources, 942

Discussion Questions, 942

Problems/Computer Lab, 942

Self-Test Quiz, 943

Suggested Reading, 944

References, 944

## 20 Human Genome, 957

Introduction, 957

Main Conclusions of Human Genome Project, 958

Gateways to Access the Human Genome, 959

NCBI, 959

Ensembl, 959

University of California at Santa Cruz Human Genome Browser, 961

NHGRI, 961

Wellcome Trust Sanger Institute, 964

Human Genome Project, 964

Background of Human Genome Project, 964

Strategic Issues: Hierarchical Shotgun Sequencing to Generate Draft Sequence, 966

Human Genome Assemblies, 966

Broad Genomic Landscape, 968

Long-Range Variation in GC Content, 969

CpG Islands, 969

Comparison of Genetic and Physical Distance, 970

- Repeat Content of Human Genome, 971
  - Transposon-Derived Repeats, 972
  - Simple Sequence Repeats, 973
  - Segmental Duplications, 973
- Gene Content of Human Genome, 974
  - Noncoding RNAs, 975
  - Protein-Coding Genes, 975
  - Comparative Proteome Analysis, 975
  - Complexity of Human Proteome, 978
- 25 Human Chromosomes, 979
  - Group A (Chromosomes 1–3), 981
  - Group B (Chromosomes 4, 5), 982
  - Group C (Chromosomes 6–12, X), 983
  - Group D (Chromosomes 13–15), 983
  - Group E (Chromosomes 16–18), 984
  - Group F (Chromosomes 19, 20), 984
  - Group G (Chromosomes 21, 22, Y), 984
  - Mitochondrial Genome, 985
- Human Genome Variation, 986
  - SNPs, Haplotypes, and HapMap, 986
  - Viewing and Analyzing SNPs and Haplotypes, 988
    - HaploView, 988
    - HapMap Browser, 988
    - Integrative Genomics Browser (IGV), 988
    - NCBI dbSNP, 988
    - PLINK, 992
    - SNPduo, 990
  - Major Conclusions of HapMap Project, 994
  - The 1000 Genomes Project, 995
  - Variation: Sequencing Individual Genomes, 998
- Perspective, 999
- Pitfalls, 1000
- Advice for Students, 1001
  - Discussion Questions, 1001
  - Problems/Computer Lab, 1001
  - Self-Test Quiz, 1003
  - Suggested Reading, 1004
  - References, 1004
- 21 Human Disease, 1011**
  - Human Genetic Disease: A Consequence of DNA Variation, 1011
    - A Bioinformatics Perspective on Human Disease, 1012
    - Garrod's View of Disease, 1014
    - Classification of Disease, 1015
    - NIH Disease Classification: MeSH Terms, 1017

- Categories of Disease, 1020
  - Allele Frequencies and Effect Sizes, 1020
  - Monogenic Disorders, 1021
  - Complex Disorders, 1024
  - Genomic Disorders, 1025
  - Environmentally Caused Disease, 1029
  - Disease and Genetic Background, 1030
  - Mitochondrial Disease, 1030
  - Somatic Mosaic Disease, 1032
  - Cancer: A Somatic Mosaic Disease, 1033
- Disease Databases, 1036
  - OMIM: Central Bioinformatics Resource for Human Disease, 1036
  - Human Gene Mutation Database (HGMD), 1039
  - ClinVar and Databases of Clinically Relevant Variants, 1040
  - GeneCards, 1041
  - Integration of Disease Database Information at the UCSC Genome Browser, 1041
  - Locus-Specific Mutation Databases and LOVD, 1041
  - The PhenCode Project, 1044
  - Limitations of Disease Databases: The Growing Interpretive Gap, 1045
  - Human Disease Genes and Amino Acid Substitutions, 1045
- Approaches to Identifying Disease-Associated Genes and Loci, 1046
  - Linkage Analysis, 1047
  - Genome-Wide Association Studies, 1047
  - Identification of Chromosomal Abnormalities, 1050
  - Human Genome Sequencing, 1051
    - Genome Sequencing to Identify Monogenic Disorders, 1051
    - Genome Sequencing to Solve Complex Disorders, 1051
    - Research Versus Clinical Sequencing and Incidental Findings, 1052
    - Disease-causing Variants in Apparently Normal Individuals, 1054
- Human Disease Genes in Model Organisms, 1055
  - Human Disease Orthologs in Nonvertebrate Species, 1056
  - Human Disease Orthologs in Rodents, 1058
  - Human Disease Orthologs in Primates, 1059
- Functional Classification of Disease Genes, 1060
  - Perspective, 1063
  - Pitfalls, 1063
- Advice for Students, 1063
  - Discussion Questions, 1064
  - Problems/Computer Lab, 1062

Self-Test Quiz, 1065

Suggested Reading, 1066

References, 1066

GLOSSARY, 1075

SELF-TEST QUIZ: SOLUTIONS, 1103

AUTHOR INDEX, 1105

SUBJECT INDEX, 1109