BIOINFORMATICS AND FUNCTIONAL GENOMICS
For three generations of family: to my parents
Aihud and Lucille; to my wife Barbara; to my daughters Kim,
Ava, and Lillian; and to my niece Madeline
## Contents in Brief

### PART I  Analyzing DNA, RNA, and Protein Sequences

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>Access to Sequence Data and Related Information</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>Pairwise Sequence Alignment</td>
<td>69</td>
</tr>
<tr>
<td>4</td>
<td>Basic Local Alignment Search Tool (BLAST)</td>
<td>121</td>
</tr>
<tr>
<td>5</td>
<td>Advanced Database Searching</td>
<td>167</td>
</tr>
<tr>
<td>6</td>
<td>Multiple Sequence Alignment</td>
<td>205</td>
</tr>
<tr>
<td>7</td>
<td>Molecular Phylogeny and Evolution</td>
<td>245</td>
</tr>
</tbody>
</table>

### PART II  Genomewide Analysis of DNA, RNA, and Protein

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>DNA: The Eukaryotic Chromosome</td>
<td>307</td>
</tr>
<tr>
<td>9</td>
<td>Analysis of Next-Generation Sequence Data</td>
<td>377</td>
</tr>
<tr>
<td>10</td>
<td>Bioinformatic Approaches to Ribonucleic Acid (RNA)</td>
<td>433</td>
</tr>
<tr>
<td>11</td>
<td>Gene Expression: Microarray and RNA-seq Data Analysis</td>
<td>479</td>
</tr>
<tr>
<td>12</td>
<td>Protein Analysis and Proteomics</td>
<td>539</td>
</tr>
<tr>
<td>13</td>
<td>Protein Structure</td>
<td>589</td>
</tr>
<tr>
<td>14</td>
<td>Functional Genomics</td>
<td>635</td>
</tr>
</tbody>
</table>

### PART III  Genome Analysis

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>Genomes Across the Tree of Life</td>
<td>699</td>
</tr>
<tr>
<td>16</td>
<td>Completed Genomes: Viruses</td>
<td>755</td>
</tr>
<tr>
<td>17</td>
<td>Completed Genomes: Bacteria and Archaea</td>
<td>797</td>
</tr>
<tr>
<td>18</td>
<td>Eukaryotic Genomes: Fungi</td>
<td>847</td>
</tr>
<tr>
<td>19</td>
<td>Eukaryotic Genomes: From Parasites to Primates</td>
<td>887</td>
</tr>
<tr>
<td>20</td>
<td>Human Genome</td>
<td>957</td>
</tr>
<tr>
<td>21</td>
<td>Human Disease</td>
<td>1011</td>
</tr>
</tbody>
</table>

GLOSSARY, 1075

SELF-TEST QUIZ: SOLUTIONS, 1103

AUTHOR INDEX, 1105

SUBJECT INDEX, 1109
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
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
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
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