Genes, Genomes, Molecular Evolution, Databases and Analytical Tools

Supratim Choudhuri

With contribution from Dr Michael Kotewicz on the Optical Mapping of DNA

Center for Food Safety and Applied Nutrition, FDA, College Park, Maryland
## Contents

Preface ix  
Acknowledgment xi  

### Fundamentals of Genes and Genomes

1.1 Biological Macromolecules, Genomics, and Bioinformatics 2  
1.2 DNA as the Universal Genetic Material 2  
1.3 DNA Double Helix 2  
1.4 Conformations of DNA 5  
1.5 Typical Eukaryotic Gene Structure 5  
1.6 Mutations in the DNA Sequence 12  
1.7 Some Features of RNA 12  
1.8 Coding Versus Noncoding RNA 14  
1.9 Protein Structure and Function 15  
1.10 Genome Structure and Organization 18  
References 25  

### Fundamentals of Molecular Evolution

2.1 Bioinformatics, Molecular Evolution, and Phylogenetics 27  
2.2 Biological Evolution and Basic Premises of Darwinism 28  
2.3 Molecular Basis of Heritable Genetic Variations—The Raw Materials for Evolution 30  
2.4 Factors that Affect Gene Frequency in a Population 41  
2.5 The Neutral Theory of Evolution 47  
2.6 Molecular Clock Hypothesis in Molecular Evolution 49  
2.7 Molecular Phylogenetics 49  
References 52  

### Genomic Technologies

3.1 Advances in Genomics 55  
3.2 From Sanger Sequencing to Pyrosequencing 55  
3.3 Pyrosequencing, Mutation Detection, and SNP Genotyping 56  
3.4 Next-Generation Sequencing Platforms 57  
3.5 Next-Next-Generation Sequencing Technology 61  
3.6 High-Density Oligonucleotide-Probe-Based Array to Investigate Genome Expression 62  
3.7 Genome-Wide Mutagenesis, Genome Editing, and Interference of Genome Expression 64  
3.8 Special Topic: Optical Mapping of DNA 67  
References 72  

### Data, Databases, Data Format, Database Search, Data Retrieval Systems, and Genome Browsers

5.1 Genomic Data 78  
5.2 Sequence Data Formats 78  
5.3 Conversion of Sequence Formats Using Readseq 79  
5.4 Primary Sequence Databases—GenBank, EMBL-Bank, and DDBJ 79  
5.5 Secondary Databases 97  
5.6 Some Examples of Publicly Available Secondary and Specialized Databases 98  
5.7 Data Retrieval 101  
5.8 An Example of Retrieval of mRNA/Gene Information 103  
5.9 Data Visualization in Genome Browsers 117  
5.10 Using Map Viewer to Search the Genome 127  
5.11 A Note on the State of the Sequence-Assembly Data in Different Databases 130  
References 131  

### Sequence Alignment and Similarity Searching in Genomic Databases: BLAST and FASTA

6.1 Evolutionary Basis of Sequence Alignment 133  
6.2 Three Terms—Sequence Identity, Sequence Similarity, and Sequence Homology—and Their Proper Usage 134  
6.3 Sequence Identity and Sequence Similarity 135  
6.4 Global Versus Local Alignment 135  
6.5 Pairwise and Multiple Alignment 139  
6.6 Alignment Algorithms, Gaps, and Gap Penalties 140  
6.7 Scoring Matrix, Alignment Score, and Statistical Significance of Sequence Alignment 144  
6.8 Database Searching with the Heuristic Versions of the Smith–Waterman Algorithm—BLAST and FASTA 149  
6.9 Sequence Comparison, Synteny, and Molecular Evolution 155  
References 155
7. Additional Bioinformatic Analyses Involving Nucleic-Acid Sequences

7.1 Genome Sequencing 157
7.2 Sequence Assembly 159
7.3 Genome Annotation 160
7.4 Prediction of Promoters, Transcription-Factor-Binding Sites, Translation Initiation Sites, and the ORF 167
7.5 Restriction-Site Mapping of the Input Sequence 169
7.6 RNA Secondary-Structure Prediction 169
7.7 Microarray Analysis 173
7.8 Detection of Sequence Polymorphism and the SNP Database 176
References 181

Additional Bioinformatic Analyses Involving Protein Sequences

8.1 Protein Structure 183
8.2 Peptide Bond, Peptide Plane, Bond Rotation, Dihedral Angles, and Ramachandran Plot 185
8.3 Prediction of Physicochemical Properties of a Protein 186
8.4 Prediction of Protease Digestibility 186
8.5 Hydrophobicity, Hydrophilicity, and Antigenicity Prediction, and the Hydropathy Plot 186
8.6 Prediction of Post-Translational Modification and Sorting 189
8.7 Secondary-Structure Prediction 190
8.8 Prediction of Domains and Motifs 193
8.9 Viewing the 3D Structure of Proteins (and Other Biological Macromolecules) 197
8.10 Allergenic Protein Databases and Protein-Allergenicity Prediction 198
8.11 Intrinsically Disordered Protein Analysis 203
References 206

Phylogenetic Analysis

9.1 Phylogenetics and the Widespread Use of the Phylogenetic Tree 209
9.2 Phylogenetic Trees 210
9.3 Phylogenetic Analysis Tools 211
9.4 Principles of Phylogenetic-Tree Construction 211
9.5 Monophyly, Polyphyly, and Paraphyly 217
9.6 Species Trees Versus Gene Trees 217
References 218

Index 219