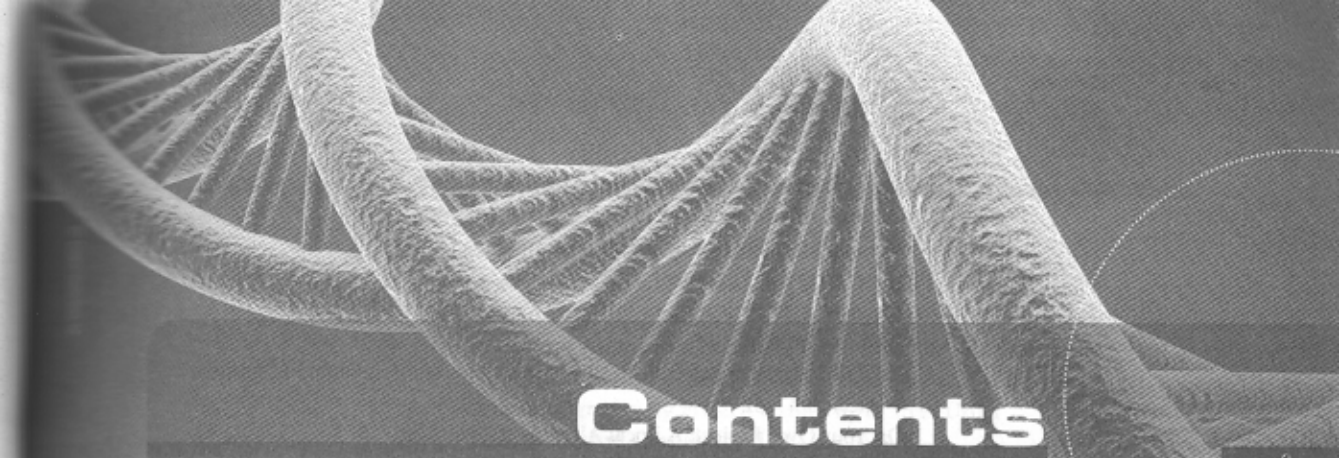




Contents

Preface vii

Chapter 1 Bioinformatics and Genomic Data: Investigating a Complex Genetic Disease 1

- Understanding the Problem: Parkinson Disease, A Complex Genetic Disorder 1
- Bioinformatics Solutions: Databases and Data Mining 3
- Understanding Genomic Databases 4
 - Primary Databases 4
 - Metadatabases 5
 - Database Searching 5
 - Genome Browsers 7
- Chapter Project: Genomic Regions Associated with Parkinson Disease 10
 - Web Exploration: Data Mining in the Genome Databases 10
 - On Your Own Project: Clues to a Genetic Disease 15
- References and Supplemental Reading 20

Chapter 2 Computational Manipulation of DNA: Genetic Screening for Disease Alleles 21

- Understanding the Problem: Genetic Screening and the Inheritance of Cystic Fibrosis 21
- Bioinformatics Solutions: Computational Approaches to Genes 23
- Understanding the Algorithm: Decoding DNA 24
 - A DNA Manipulation Algorithm 25
 - A Transcription Algorithm 26
 - A Translation Algorithm 26
 - A Mutation Detection Algorithm 26
- Chapter Project: Genetic Screening for Carriers of CF Mutations 28
 - Web Exploration 29
 - Guided Programming Project: Working with DNA and Protein Strings 32
 - On Your Own Project: Identifying Insertions and Deletions 35
- References and Supplemental Reading 38

Chapter 3 Sequence Alignment: Investigating an Influenza Outbreak 39

- Understanding the Problem: The 2009 H₁N₁ Influenza Pandemic 39
- Bioinformatics Solutions: Sequence Alignment and Sequence Comparison 41

Understanding the Algorithm: Global Alignment	43
Optimal Alignment and Scoring	43
Needleman-Wunsch Algorithm	44
Generating the Alignment	46
Chapter Project: Investigation of Influenza Virus Strains	47
Web Exploration	48
Guided Programming Project: The Needleman-Wunsch Global Alignment Algorithm	52
On Your Own Project: A Local Alignment Algorithm	57
References and Supplemental Reading	62

Chapter 4 Database Searching and Multiple Alignment: Investigating Antibiotic Resistance 63

Understanding the Problem: Antibiotic Resistance	63
Bioinformatics Solutions: Advanced Sequence Comparison Algorithms	64
Understanding the Algorithm: Database Searching and Multiple Alignment	65
BLAST: A Heuristic Approach to Database Searching	65
ClustalW: Multiple Sequence Alignment	67
Chapter Project: Horizontal Gene Transfer of Antibiotic Resistance	69
Searching for Erythromycin Resistance Genes with BLAST	71
Multiple Sequence Alignment with ClustalW	74
References and Supplemental Reading	77

Chapter 5 Substitution Matrices and Protein Alignments: Virulence Factors in *E. coli* 79

Understanding the Problem: Virulence Factors in <i>E. coli</i> Outbreaks	79
Bioinformatics Solutions: Protein Alignment and Clues to Function	81
Understanding the Algorithm: Protein Alignment and Substitution Matrices	83
Substitution Matrices	83
Protein Alignment Algorithm: Scoring with the Substitution Matrix	86
Chapter Project: Using Protein Alignment to Investigate Functions of Virulence Factors	88
Web Exploration	89
Guided Programming Project: Using a Substitution Matrix in a Protein Alignment Program	92
On Your Own Project: Building Your Own Substitution Matrix	96
References and Supplemental Reading	104

Chapter 6 Distance Measurement in Molecular Phylogenetics: Evolution of Mammals 105

Understanding the Problem: Mammalian Evolution	105
Bioinformatics Solutions: Molecular Phylogenetics and Distance Measurement	106
Understanding the Algorithm: Measuring Distance	109
Jukes-Cantor Model	110
Kimura's Two-Parameter Substitution Model	110
Tamura's Three-Parameter Model	110
Other Models	111
Chapter Project: Evolution of the Whale	112
Web Exploration: Evolution of Marine Mammals	112
Guided Programming Project: Using Distance Metrics to Measure Similarity	116
On Your Own Project: Alignment and Evolutionary Distance	117
References and Supplemental Reading	121

Chapter 7 Tree-Building in Molecular Phylogenetics: Three Domains of Life 122

- Understanding the Problem: Rooting the Tree 122
- Bioinformatics Solutions: Tree-Building 124
- Understanding the Algorithm: Clustering Algorithms 126
- Chapter Project: Placing the Archaea in the Tree of Life 130
 - Web Exploration: Molecular Clocks and the Archaea 130
 - Guided Programming Project: Phylogenetic Trees Using Agglomerative Clustering 134
 - On Your Own Project: The Neighbor-Joining Method 137
- References and Supplemental Reading 143

Chapter 8 DNA Sequencing: Identification of Novel Viral Pathogens 144

- Understanding the Problem: Deep Sequencing of Clinical Samples 144
- Bioinformatics Solutions: Assembly and Mapping of Short Sequence Reads 146
- Understanding the Algorithm: Determining Overlap in Sequence Assembly 148
- Chapter Project: Identifying Viruses Through Metagenomic Analysis of Clinical Samples 150
 - Web Exploration: Analysis of Virus Sequences in the Human Metagenome 152
 - Guided Programming Project: Sequencing and Assembly 157
 - On Your Own Project: A Mini-Assembly Program 162
- References and Supplemental Reading 171

Chapter 9 Sequence-Based Gene Prediction: Annotation of a Resistance Plasmid 173

- Understanding the Problem: Gene Discovery 173
- Bioinformatics Solutions: Gene Prediction 175
- Understanding the Algorithm: Pattern Matching in Sequence-Based Gene Prediction 176
- Chapter Project: Gene Discovery in a Resistance Plasmid 179
 - Web Exploration: Prokaryotic Gene Prediction and Annotation 179
 - Guided Programming Project: Pattern Matching for Sequence-Based Gene Prediction 184
 - On Your Own Project: Sequence-Based Gene Discovery in Eukaryotes 187
- References and Supplemental Reading 192

Chapter 10 Advanced Gene Prediction: Identification of an Influenza Resistance Gene 193

- Understanding the Problem: Exon Prediction 193
- Bioinformatics Solutions: Content- and Probability-Based Gene Prediction 195
- Understanding the Algorithm: Codon Usage, Frequency Matching, HMMs, and Neural Networks 196
 - Using Codon Frequencies in Gene Prediction 196
 - Prediction of CpG Islands 197
 - HMMs for Gene Prediction 198
 - Neural Network Modeling 201
- Chapter Project: Identifying an Influenza Resistance Gene 203
 - Web Exploration: Finding Genes in a Eukaryotic Genome Sequence 204
 - Guided Programming Project: Predicting CpG Islands 209
 - On Your Own Project: Hidden Markov Modeling in Gene Prediction 211
- References and Supplemental Reading 215

Chapter 11 Protein Structure Prediction and Analysis: Rational Drug Design 217

- Understanding the Problem: Structure Prediction 217
- Bioinformatics Solutions: Predicting and Modeling Protein Structure 219
- Understanding the Algorithm: The Chou-Fasman Algorithm for Secondary Structure Prediction 222
- Chapter Project: Protein Structure Prediction 224
 - Web Exploration: Protein Structure Modeling and Drug Design 225
 - Guided Programming Project: Structure Prediction with the Chou-Fasman Algorithm 234
 - On Your Own Project: A Complete Chou-Fasman Program 236
- References and Supplemental Reading 241

Chapter 12 Nucleic Acid Structure Prediction: Polymerase Chain Reaction and RNA Interference 242

- Understanding the Problem: Nucleic Acid Structure Prediction 242
- Bioinformatics Solutions: Secondary Structure Prediction 244
- Understanding the Algorithm: The Nussinov-Jacobson Algorithm 245
- Chapter Project: Nucleic Acid Structure Prediction 249
 - Web Exploration: Applications of Nucleic Acid Structure Prediction 250
 - Guided Programming Project: Structure Prediction with the Nussinov-Jacobson Algorithm 254
 - On Your Own Project: PCR Primer Analysis with Nussinov-Jacobson 257
- References and Supplemental Reading 262

Appendix Introduction to Programming 263

- Programming Basics 263
- Five Main Programming Constructs 263
 - Assignment Statements 264
 - Input Statements 265
 - Output Statements 265
 - Decision Statements 266
 - Loop Statements 266
- Data Structures 267
 - Arrays 268
 - Hash Tables 268
- Functions 268
- A Sample Program 270

Glossary 271

Index 285