Contents

Preface ix
    Book Overview xi
    Exercises xii
    Errors xii
    Acknowledgments xiii

1 Basic Concepts of Molecular Biology 1
    1.1 Life 1
    1.2 Proteins 2
    1.3 Nucleic Acids 5
        1.3.1 DNA 5
        1.3.2 RNA 8
    1.4 The Mechanisms of Molecular Genetics 9
        1.4.1 Genes and the Genetic Code 9
        1.4.2 Transcription, Translation, and Protein Synthesis 10
        1.4.3 Junk DNA and Reading Frames 12
        1.4.4 Chromosomes 13
        1.4.5 Is the Genome like a Computer Program? 15
    1.5 How the Genome Is Studied 15
        1.5.1 Maps and Sequences 16
        1.5.2 Specific Techniques 17
    1.6 The Human Genome Project 21
    1.7 Sequence Databases 23
    Exercises 30
    Bibliographic Notes 30

2 Strings, Graphs, and Algorithms 33
    2.1 Strings 33
    2.2 Graphs 35
    2.3 Algorithms 38
    Exercises 43
    Bibliographic Notes 45
3 Sequence Comparison and Database Search

3.1 Biological Background 47
3.2 Comparing Two Sequences 49
  3.2.1 Global Comparison — The Basic Algorithm 49
  3.2.2 Local Comparison 55
  3.2.3 Semiglobal Comparison 56
3.3 Extensions to the Basic Algorithms 58
  3.3.1 Saving Space 58
  3.3.2 General Gap Penalty Functions 60
  3.3.3 Affine Gap Penalty Functions 64
  3.3.4 Comparing Similar Sequences 66
3.4 Comparing Multiple Sequences 69
  3.4.1 The SP Measure 70
  3.4.2 Star Alignments 76
  3.4.3 Tree Alignments 79
3.5 Database Search 80
  3.5.1 PAM Matrices 80
  3.5.2 BLAST 84
  3.5.3 FAST 87
3.6 Other Issues 89
  * 3.6.1 Similarity and Distance 89
  3.6.2 Parameter Choice in Sequence Comparison 96
  3.6.3 String Matching and Exact Sequence Comparison 98
Summary 100
Exercises 101
Bibliographic Notes 103

4 Fragment Assembly of DNA 105

4.1 Biological Background 105
  4.1.1 The Ideal Case 106
  4.1.2 Complications 107
  4.1.3 Alternative Methods for DNA Sequencing 113
4.2 Models 114
  4.2.1 Shortest Common Superstring 114
  4.2.2 Reconstruction 116
  4.2.3 Multicontig 117
* 4.3 Algorithms 119
  4.3.1 Representing Overlaps 119
  4.3.2 Paths Originating Superstrings 120
  4.3.3 Shortest Superstrings as Paths 122
  4.3.4 The Greedy Algorithm 124
  4.3.5 Acyclic Subgraphs 126
4.4 Heuristics 132
  4.4.1 Finding Overlaps 134
  4.4.2 Ordering Fragments 134
  4.4.3 Alignment and Consensus 137
Summary 139
5 Physical Mapping of DNA  
5.1 Biological Background  
5.1.1 Restriction Site Mapping  
5.1.2 Hybridization Mapping  
5.2 Models  
5.2.1 Restriction Site Models  
5.2.2 Interval Graph Models  
5.2.3 The Consecutive Ones Property  
5.2.4 Algorithmic Implications  
5.3 An Algorithm for the C1P Problem  
5.4 An Approximation for Hybridization Mapping with Errors  
5.4.1 A Graph Model  
5.4.2 A Guarantee  
5.4.3 Computational Practice  
5.5 Heuristics for Hybridization Mapping  
5.5.1 Screening Chimeric Clones  
5.5.2 Obtaining a Good Probe Ordering  
5.6 Summary  
Exercises  
Bibliographic Notes  

6 Phylogenetic Trees  
6.1 Character States and the Perfect Phylogeny Problem  
6.2 Binary Character States  
6.3 Two Characters  
6.4 Parsimony and Compatibility in Phylogenies  
6.5 Algorithms for Distance Matrices  
6.5.1 Reconstructing Additive Trees  
6.5.2 Reconstructing Ultrametric Trees  
6.6 Agreement Between Phylogenies  
6.7 Summary  
Exercises  
Bibliographic Notes  

7 Genome Rearrangements  
7.1 Biological Background  
7.2 Oriented Blocks  
7.2.1 Definitions  
7.2.2 Breakpoints  
7.2.3 The Diagram of Reality and Desire  
7.2.4 Interleaving Graph  
7.2.5 Bad Components  
7.2.6 Algorithm  
7.3 Unoriented Blocks

<table>
<thead>
<tr>
<th>Exercise</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exercises</td>
<td>139</td>
</tr>
<tr>
<td>Bibliographic Notes</td>
<td>141</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Physical Mapping of DNA</td>
<td>143</td>
</tr>
<tr>
<td>5.1 Biological Background</td>
<td>143</td>
</tr>
<tr>
<td>5.1.1 Restriction Site Mapping</td>
<td>145</td>
</tr>
<tr>
<td>5.1.2 Hybridization Mapping</td>
<td>146</td>
</tr>
<tr>
<td>5.2 Models</td>
<td>147</td>
</tr>
<tr>
<td>5.2.1 Restriction Site Models</td>
<td>147</td>
</tr>
<tr>
<td>5.2.2 Interval Graph Models</td>
<td>149</td>
</tr>
<tr>
<td>5.2.3 The Consecutive Ones Property</td>
<td>150</td>
</tr>
<tr>
<td>5.2.4 Algorithmic Implications</td>
<td>152</td>
</tr>
<tr>
<td>5.3 An Algorithm for the C1P Problem</td>
<td>153</td>
</tr>
<tr>
<td>5.4 An Approximation for Hybridization Mapping with Errors</td>
<td>160</td>
</tr>
<tr>
<td>5.4.1 A Graph Model</td>
<td>160</td>
</tr>
<tr>
<td>5.4.2 A Guarantee</td>
<td>162</td>
</tr>
<tr>
<td>5.4.3 Computational Practice</td>
<td>164</td>
</tr>
<tr>
<td>5.5 Heuristics for Hybridization Mapping</td>
<td>167</td>
</tr>
<tr>
<td>5.5.1 Screening Chimeric Clones</td>
<td>167</td>
</tr>
<tr>
<td>5.5.2 Obtaining a Good Probe Ordering</td>
<td>168</td>
</tr>
<tr>
<td>5.6 Summary</td>
<td>169</td>
</tr>
<tr>
<td>Exercises</td>
<td>170</td>
</tr>
<tr>
<td>Bibliographic Notes</td>
<td>172</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 Phylogenetic Trees</td>
<td>175</td>
</tr>
<tr>
<td>6.1 Character States and the Perfect Phylogeny Problem</td>
<td>177</td>
</tr>
<tr>
<td>6.2 Binary Character States</td>
<td>182</td>
</tr>
<tr>
<td>6.3 Two Characters</td>
<td>186</td>
</tr>
<tr>
<td>6.4 Parsimony and Compatibility in Phylogenies</td>
<td>190</td>
</tr>
<tr>
<td>6.5 Algorithms for Distance Matrices</td>
<td>192</td>
</tr>
<tr>
<td>6.5.1 Reconstructing Additive Trees</td>
<td>193</td>
</tr>
<tr>
<td>6.5.2 Reconstructing Ultrametric Trees</td>
<td>196</td>
</tr>
<tr>
<td>6.6 Agreement Between Phylogenies</td>
<td>204</td>
</tr>
<tr>
<td>6.7 Summary</td>
<td>209</td>
</tr>
<tr>
<td>Exercises</td>
<td>209</td>
</tr>
<tr>
<td>Bibliographic Notes</td>
<td>211</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 Genome Rearrangements</td>
<td>215</td>
</tr>
<tr>
<td>7.1 Biological Background</td>
<td>215</td>
</tr>
<tr>
<td>7.2 Oriented Blocks</td>
<td>217</td>
</tr>
<tr>
<td>7.2.1 Definitions</td>
<td>219</td>
</tr>
<tr>
<td>7.2.2 Breakpoints</td>
<td>221</td>
</tr>
<tr>
<td>7.2.3 The Diagram of Reality and Desire</td>
<td>222</td>
</tr>
<tr>
<td>7.2.4 Interleaving Graph</td>
<td>228</td>
</tr>
<tr>
<td>7.2.5 Bad Components</td>
<td>231</td>
</tr>
<tr>
<td>7.2.6 Algorithm</td>
<td>234</td>
</tr>
<tr>
<td>7.3 Unoriented Blocks</td>
<td>236</td>
</tr>
</tbody>
</table>
## CONTENTS

7.3.1 Strips 238  
7.3.2 Algorithm 241  
Summary 242  
Exercises 243  
Bibliographic Notes 244  

8 Molecular Structure Prediction 245  
8.1 RNA Secondary Structure Prediction 246  
8.2 The Protein Folding Problem 252  
8.3 Protein Threading 254  
Summary 259  
Exercises 259  
Bibliographic Notes 260  

9 Epilogue: Computing with DNA 261  
9.1 The Hamiltonian Path Problem 261  
9.2 Satisfiability 264  
9.3 Problems and Promises 267  
Exercises 268  
Bibliographic Notes and Further Sources 268  

Answers to Selected Exercises 271  

References 277  

Index 289