

# Contents

|                                                                |          |
|----------------------------------------------------------------|----------|
| <b>Chapter 1: Introduction .....</b>                           | <b>1</b> |
| 1.1 Prelude .....                                              | 1        |
| 1.2 What is Bioinformatics.....                                | 2        |
| 1.3 Book's Organization .....                                  | 3        |
| <b>Chapter 2: An Introduction to the Python Language .....</b> | <b>5</b> |
| 2.1 Features of the Python Language .....                      | 5        |
| 2.2 Variables and Pre-Defined Functions.....                   | 8        |
| 2.2.1 Variable Types.....                                      | 8        |
| 2.2.2 Assigning Values to Variables .....                      | 9        |
| 2.2.3 Numerical and Logical Variables.....                     | 10       |
| 2.2.4 Containers .....                                         | 13       |
| 2.2.5 Variable Comparison .....                                | 17       |
| 2.2.6 Type Conversion .....                                    | 18       |
| 2.3 Developing Python Code .....                               | 19       |
| 2.3.1 Indentation.....                                         | 19       |
| 2.3.2 User-Defined Functions.....                              | 21       |
| 2.3.3 Conditional Statements .....                             | 22       |
| 2.3.4 Conditional Loops.....                                   | 25       |
| 2.3.5 Iterative Loop Statements.....                           | 26       |
| 2.3.6 List Comprehensions.....                                 | 28       |
| 2.3.7 Help .....                                               | 29       |
| 2.4 Developing Python Programs .....                           | 29       |
| 2.4.1 Data Input and Output .....                              | 30       |
| 2.4.2 Reading and Writing From Files .....                     | 31       |
| 2.4.3 Handling Exceptions .....                                | 34       |
| 2.4.4 Modules.....                                             | 36       |
| 2.4.5 Putting It All Together .....                            | 37       |
| 2.5 Object-Oriented Programming .....                          | 39       |
| 2.5.1 Defining Classes and Creating Objects .....              | 39       |
| 2.5.2 Special Methods.....                                     | 42       |

|                                                                     |                                                                  |           |
|---------------------------------------------------------------------|------------------------------------------------------------------|-----------|
| 2.5.3                                                               | Inheritance .....                                                | 43        |
| 2.5.4                                                               | Modularity .....                                                 | 45        |
| 2.6                                                                 | Pre-Defined Classes and Methods .....                            | 45        |
| 2.6.1                                                               | Generic Methods for Containers .....                             | 45        |
| 2.6.2                                                               | Methods for Lists .....                                          | 48        |
| 2.6.3                                                               | Methods for Strings .....                                        | 50        |
| 2.6.4                                                               | Methods for Sets .....                                           | 52        |
| 2.6.5                                                               | Methods for Dictionaries .....                                   | 53        |
| 2.6.6                                                               | Assigning and Copying Variables .....                            | 54        |
|                                                                     | Bibliographical Notes and Further Reading .....                  | 55        |
|                                                                     | Exercises and Programming Projects .....                         | 56        |
|                                                                     | Exercises .....                                                  | 56        |
|                                                                     | Programming Projects .....                                       | 57        |
| <b>Chapter 3: Cellular and Molecular Biology Fundamentals .....</b> |                                                                  | <b>59</b> |
| 3.1                                                                 | The Cell: The Basic Unit of Life .....                           | 59        |
| 3.2                                                                 | Genetic Information: Nucleic Acids .....                         | 61        |
| 3.2.1                                                               | Transcription: RNA Synthesis .....                               | 62        |
| 3.2.2                                                               | Translation: Protein Synthesis .....                             | 63        |
| 3.3                                                                 | Genes: Discrete Units of Genetic Information .....               | 67        |
| 3.3.1                                                               | Gene Structure .....                                             | 67        |
| 3.3.2                                                               | Regulation of Gene Expression .....                              | 70        |
| 3.4                                                                 | Human Genome .....                                               | 71        |
| 3.5                                                                 | Biological Resources and Databases .....                         | 73        |
|                                                                     | Bibliographic References and Further Reading .....               | 77        |
|                                                                     | Exercises .....                                                  | 77        |
| <b>Chapter 4: Basic Processing of Biological Sequences .....</b>    |                                                                  | <b>79</b> |
| 4.1                                                                 | Biological Sequences: Representations and Basic Algorithms ..... | 79        |
| 4.2                                                                 | Transcription and Reverse Complement .....                       | 83        |
| 4.3                                                                 | Translation .....                                                | 84        |
| 4.4                                                                 | Seeking Putative Genes: Open Reading Frames .....                | 87        |
| 4.5                                                                 | Putting It All Together .....                                    | 90        |
| 4.6                                                                 | A Class for Biological Sequences .....                           | 91        |
| 4.7                                                                 | Processing Sequences With BioPython .....                        | 94        |
| 4.8                                                                 | Sequence Annotation Objects in BioPython .....                   | 98        |
|                                                                     | Exercises and Programming Projects .....                         | 104       |
|                                                                     | Exercises .....                                                  | 104       |
|                                                                     | Programming Projects .....                                       | 105       |

|                                                                         |            |
|-------------------------------------------------------------------------|------------|
| <b>Chapter 5: Finding Patterns in Sequences .....</b>                   | <b>107</b> |
| 5.1 Introduction: Importance of Pattern Finding in Bioinformatics ..... | 107        |
| 5.2 Naive Algorithm for Fixed Pattern Finding .....                     | 108        |
| 5.3 Heuristic Algorithm: Boyer-Moore .....                              | 110        |
| 5.4 Deterministic Finite Automata .....                                 | 113        |
| 5.5 Finding Flexible Patterns: Regular Expressions .....                | 117        |
| 5.5.1 Definitions and Regular Expressions in Python .....               | 117        |
| 5.5.2 Examples in Biological Sequence Analysis .....                    | 122        |
| 5.5.3 Finding Protein Motifs .....                                      | 125        |
| 5.5.4 An Application to Restriction Enzymes .....                       | 127        |
| Bibliographic Notes and Further Reading .....                           | 129        |
| Exercises and Programming Projects .....                                | 129        |
| Exercises .....                                                         | 129        |
| Programming Projects .....                                              | 130        |
| <b>Chapter 6: Pairwise Sequence Alignment .....</b>                     | <b>133</b> |
| 6.1 Introduction: Comparing Sequences and Sequence Alignment .....      | 133        |
| 6.2 Visual Alignments: Dot Plots .....                                  | 134        |
| 6.3 Sequence Alignment as an Optimization Problem .....                 | 138        |
| 6.3.1 Problem Definition and Complexity .....                           | 138        |
| 6.3.2 Objective Function: Substitution Matrices and Gap Penalties ..... | 139        |
| 6.3.3 Implementing the Calculation of the Objective Function .....      | 142        |
| 6.4 Dynamic Programming Algorithms for Global Alignment .....           | 146        |
| 6.4.1 The Needleman-Wunsch Algorithm .....                              | 146        |
| 6.4.2 Implementing the Needleman-Wunsch Algorithm .....                 | 149        |
| 6.5 Dynamic Programming Algorithms for Local Alignment .....            | 152        |
| 6.5.1 The Smith-Waterman Algorithm .....                                | 152        |
| 6.5.2 Implementing the Smith-Waterman Algorithm .....                   | 154        |
| 6.6 Special Cases of Sequence Alignment .....                           | 157        |
| 6.7 Pairwise Sequence Alignment in BioPython .....                      | 159        |
| Bibliographical Notes and Further Reading .....                         | 160        |
| Exercises and Programming Projects .....                                | 161        |
| Exercises .....                                                         | 161        |
| Programming Projects .....                                              | 162        |
| <b>Chapter 7: Searching Similar Sequences in Databases .....</b>        | <b>163</b> |
| 7.1 Introduction .....                                                  | 163        |
| 7.2 BLAST Algorithm and Programs .....                                  | 165        |
| 7.2.1 Overview of the BLAST Algorithm .....                             | 165        |
| 7.2.2 BLAST Programs .....                                              | 166        |

|                                                     |                                                                          |            |
|-----------------------------------------------------|--------------------------------------------------------------------------|------------|
| 7.2.3                                               | Significance of the Alignments .....                                     | 167        |
| 7.3                                                 | Implementing Our Own BLAST .....                                         | 168        |
| 7.4                                                 | Using BLAST Through BioPython.....                                       | 171        |
|                                                     | Bibliographical Notes and Further Reading .....                          | 176        |
|                                                     | Exercises and Programming Projects .....                                 | 176        |
|                                                     | Exercises .....                                                          | 176        |
|                                                     | Programming Projects.....                                                | 177        |
| <b>Chapter 8: Multiple Sequence Alignment .....</b> |                                                                          | <b>179</b> |
| 8.1                                                 | Introduction: Problem Definition and Complexity.....                     | 179        |
| 8.2                                                 | Classes of Optimization Algorithms for Multiple Sequence Alignment ..... | 180        |
| 8.2.1                                               | Dynamic Programming .....                                                | 180        |
| 8.2.2                                               | Heuristic Algorithms.....                                                | 182        |
| 8.3                                                 | Implementing Progressive Alignments in Python .....                      | 186        |
| 8.3.1                                               | Representing Alignments: Class <i>MyAlign</i> .....                      | 186        |
| 8.3.2                                               | Pairwise Alignment: Class <i>AlignSeq</i> .....                          | 188        |
| 8.3.3                                               | Implementing Multiple Sequence Alignment: Class <i>MultipleAlign</i> ..  | 190        |
| 8.4                                                 | Handling Alignments in <i>BioPython</i> .....                            | 193        |
|                                                     | Bibliographical Notes and Further Reading .....                          | 195        |
|                                                     | Exercises and Programming Projects .....                                 | 195        |
|                                                     | Exercises .....                                                          | 195        |
|                                                     | Programming Projects.....                                                | 197        |
| <b>Chapter 9: Phylogenetic Analysis .....</b>       |                                                                          | <b>199</b> |
| 9.1                                                 | Introduction: Problem Definition and Relevance .....                     | 199        |
| 9.2                                                 | Classes of Algorithms for Phylogenetic Analysis .....                    | 201        |
| 9.2.1                                               | Distance-Based Methods .....                                             | 202        |
| 9.2.2                                               | Maximum Parsimony .....                                                  | 206        |
| 9.2.3                                               | Statistical Methods .....                                                | 207        |
| 9.3                                                 | Implementing Distance-Based Algorithms in Python .....                   | 207        |
| 9.3.1                                               | Implementing Binary Trees.....                                           | 208        |
| 9.3.2                                               | Implementing the UPGMA Algorithm.....                                    | 211        |
| 9.4                                                 | BioPython Functions for Phylogenetic Analysis .....                      | 216        |
|                                                     | Bibliographical Notes and Further Reading .....                          | 218        |
|                                                     | Exercises and Programming Projects .....                                 | 218        |
|                                                     | Exercises .....                                                          | 218        |
|                                                     | Programming Projects.....                                                | 220        |
| <b>Chapter 10: Motif Discovery Algorithms .....</b> |                                                                          | <b>221</b> |
| 10.1                                                | Introduction: Problem Definition and Relevance .....                     | 221        |

|                                                                         |            |
|-------------------------------------------------------------------------|------------|
| 10.2 Brute-Force Algorithms: Exhaustive Search .....                    | 226        |
| 10.3 Branch-and-Bound Algorithms .....                                  | 227        |
| 10.4 Heuristic Algorithms .....                                         | 232        |
| Bibliographic Notes and Further Reading .....                           | 235        |
| Exercises and Programming Projects .....                                | 235        |
| Exercises .....                                                         | 235        |
| Programming Projects .....                                              | 236        |
| <b>Chapter 11: Probabilistic Motifs and Stochastic Algorithms .....</b> | <b>237</b> |
| 11.1 Representing and Searching Probabilistic Motifs .....              | 237        |
| 11.2 Stochastic Algorithms: Expectation-Maximization .....              | 244        |
| 11.3 Gibbs Sampling for Motif Discovery .....                           | 247        |
| 11.4 Probabilistic Motifs in BioPython .....                            | 250        |
| Bibliographic Notes and Further Reading .....                           | 252        |
| Exercises and Programming Projects .....                                | 253        |
| Exercises .....                                                         | 253        |
| Programming Projects .....                                              | 253        |
| <b>Chapter 12: Hidden Markov Models .....</b>                           | <b>255</b> |
| 12.1 Introduction: What Are Hidden Markov Models? .....                 | 255        |
| 12.2 Algorithms and Python Implementation .....                         | 260        |
| 12.2.1 Joint Probability of an Observed Sequence and State Path .....   | 261        |
| 12.2.2 Probability of an Observed Sequence Over All State Paths .....   | 262        |
| 12.2.3 Probability of the Remainder of an Observed Sequence .....       | 264        |
| 12.2.4 Finding the Optimal State Path .....                             | 265        |
| 12.2.5 Learning the Parameters of an HMM Model .....                    | 267        |
| 12.3 HMMs for Database Search .....                                     | 271        |
| Bibliographic Notes and Further Reading .....                           | 272        |
| Exercises and Programming Projects .....                                | 273        |
| <b>Chapter 13: Graphs: Concepts and Algorithms .....</b>                | <b>275</b> |
| 13.1 Graphs: Definitions and Representations .....                      | 275        |
| 13.2 A Python Class for Graphs .....                                    | 277        |
| 13.3 Adjacent Nodes and Degrees .....                                   | 279        |
| 13.4 Paths, Searches, and Distances .....                               | 281        |
| 13.5 Cycles .....                                                       | 286        |
| Bibliographic Notes and Further Reading .....                           | 287        |
| Exercises and Programming Projects .....                                | 287        |
| Exercises .....                                                         | 287        |
| Programming Projects .....                                              | 288        |



|                                                                                |            |
|--------------------------------------------------------------------------------|------------|
| <b>Chapter 14: Graphs and Biological Networks.....</b>                         | <b>289</b> |
| 14.1 Introduction.....                                                         | 289        |
| 14.2 Representing Networks With Graphs.....                                    | 290        |
| 14.2.1 A Python Class for Metabolic Networks.....                              | 290        |
| 14.2.2 An Example Metabolic Network for a Real Organism.....                   | 296        |
| 14.3 Network Topological Analysis .....                                        | 297        |
| 14.3.1 Degree Distribution .....                                               | 298        |
| 14.3.2 Shortest Path Analysis .....                                            | 300        |
| 14.3.3 Clustering Coefficients.....                                            | 301        |
| 14.3.4 Hubs and Centrality Measures.....                                       | 304        |
| 14.4 Assessing the Metabolic Potential .....                                   | 307        |
| Bibliographic Notes and Further Reading .....                                  | 309        |
| Exercises and Programming Projects .....                                       | 310        |
| Exercises.....                                                                 | 310        |
| Programming Projects.....                                                      | 310        |
| <b>Chapter 15: Assembling Reads Into Genomes: Graph-Based Algorithms .....</b> | <b>313</b> |
| 15.1 Introduction to Genome Assembly and Related Challenges .....              | 313        |
| 15.2 Overlap Graphs and Hamiltonian Cycles .....                               | 314        |
| 15.2.1 Problem Definition and Exhaustive Search .....                          | 314        |
| 15.2.2 Overlap Graphs.....                                                     | 317        |
| 15.2.3 Hamiltonian Circuits .....                                              | 320        |
| 15.3 DeBruijn Graphs and Eulerian Paths .....                                  | 325        |
| 15.3.1 DeBruijn Graphs for Genome Assembly .....                               | 325        |
| 15.3.2 Eulerian Paths .....                                                    | 327        |
| 15.4 Genome Assembly in Practice.....                                          | 332        |
| Bibliographic Notes and Further Reading .....                                  | 334        |
| Exercises and Programming Projects .....                                       | 334        |
| Exercises.....                                                                 | 334        |
| Programming Projects.....                                                      | 335        |
| <b>Chapter 16: Matching Reads to Reference Sequences.....</b>                  | <b>337</b> |
| 16.1 Introduction: Problem Definition and Applications .....                   | 337        |
| 16.2 Pre-Processing the Patterns: Tries .....                                  | 337        |
| 16.2.1 Definitions and Algorithms .....                                        | 337        |
| 16.2.2 Implementing Tries in Python .....                                      | 340        |
| 16.3 Pre-Processing the Sequence: Suffix Trees .....                           | 344        |
| 16.3.1 Definitions and Algorithms .....                                        | 344        |
| 16.3.2 Implementing Suffix Trees in Python .....                               | 349        |
| 16.4 Burrows-Wheeler Transforms .....                                          | 352        |

|                                                         |            |
|---------------------------------------------------------|------------|
| 16.4.1 Definitions and Algorithms .....                 | 352        |
| 16.4.2 Implementation in Python .....                   | 357        |
| 16.4.3 Aligning References to Genomes in Practice ..... | 362        |
| <b>Bibliographic Notes and Further Reading .....</b>    | <b>362</b> |
| <b>Exercises and Programming Projects .....</b>         | <b>363</b> |
| Exercises .....                                         | 363        |
| Programming Projects .....                              | 363        |
| <b>Chapter 17: Further Reading and Resources .....</b>  | <b>365</b> |
| 17.1 Complementary Books .....                          | 365        |
| 17.2 Journals and Conferences .....                     | 366        |
| 17.3 Formal Education .....                             | 368        |
| 17.4 Online Resources .....                             | 369        |
| <b>Final Words .....</b>                                | <b>373</b> |
| <b>Bibliography .....</b>                               | <b>375</b> |
| <b>Index .....</b>                                      | <b>383</b> |