

Table of Contents

Preface	xix
Acknowledgments	xxiii
Chapter 1 . The Study of Structural Bioinformatics	1
1.1 Motivation	1
1.2 Small Beginnings	4
1.3 Structural Bioinformatics and the Scientific Method	5
1.3.1 Three Realms: Nature, Science, and Computation	6
1.3.2 Hypothesis, Model, and Theory	8
1.3.3 Laws, Postulates, and Assumptions	12
1.3.4 Model Theory and Computational Theory	13
1.3.5 Different Assumptions for Different Models	14
1.4 A More Detailed Problem Analysis: Force Fields	15
1.4.1 Nature	16
1.4.2 Science	16
1.4.2.1 <i>Energy Terms for Bonded Atoms</i>	16
1.4.2.2 <i>Energy Terms for Nonbonded Atoms</i>	19
1.4.2.3 <i>Total Potential Energy</i>	21
1.4.3 Computation	21
1.5 Modeling Issues	25
1.5.1 Rashomon	26
1.5.2 Ockham	26
1.5.3 Bellman	27
1.5.4 Interpretability	28
1.5.5 Refutability	29
1.5.6 Complexity and Approximation	29
1.6 Sources of Error	32
1.7 Summary	33
1.8 Exercises	34
References	36
Chapter 2 . Introduction to Macromolecular Structure	37
2.1 Motivation	37
2.2 Overview of Protein Structure	38
2.2.1 Amino Acids and Primary Sequence	38
2.2.2 Secondary Structure	44
2.2.2.1 <i>Alpha Helices</i>	44
2.2.2.2 <i>Beta Strands</i>	47
2.2.2.3 <i>Loops</i>	52
2.2.3 Tertiary Structure	53
2.2.3.1 <i>What Is Tertiary Structure?</i>	54
2.2.3.2 <i>The Tertiary Structure of Myoglobin</i>	54

2.2.3.3 <i>Tertiary Structure Beyond the Binding Pocket</i>	58
2.2.4 Quaternary Structure	64
2.2.5 Protein Functionality	67
2.2.6 Protein Domains	68
2.3 An Overview of RNA Structure	70
2.3.1 Nucleotides and RNA Primary Sequence	71
2.3.2 RNA Secondary Structure	72
2.3.3 RNA Tertiary Structure	75
2.4 Exercises	78
References	80
Chapter 3 . Data Sources, Formats, and Applications	83
3.1 Motivation	83
3.2 Sources of Structural Data	84
3.2.1 PDB: The Protein Data Bank	84
3.2.2 PDBsum: The PDB Summary	86
3.2.3 SCOP: Structural Classification of Proteins	86
3.2.4 CATH: The CATH Hierarchy	88
3.2.5 PubChem	92
3.2.6 DrugBank	94
3.3 PDB File Format	95
3.4 Visualization of Molecular Data	98
3.4.1 Plug-In versus Stand-Alone	99
3.4.2 Change of Viewing Perspective	99
3.4.3 Graphical Representation	99
3.4.4 Visual Effects	101
3.4.5 Selection Abilities	101
3.4.6 Computational Tools	102
3.4.7 Extras	102
3.5 Software for Structural Bioinformatics	103
3.5.1 PyMOL	103
3.5.2 Eclipse	103
3.5.3 MarvinSketch	104
3.5.4 ACD/ChemSketch	104
3.5.5 JOELib2	105
3.5.6 Chemistry Development Kit (CDK)	105
3.5.7 BioPython	105
3.6 Exercises	106
References	109
Chapter 4 . Dynamic Programming	111
4.1 Motivation	111
4.2 Introduction	112
4.3 A DP example: The AI Gore rhythm for giving talks	112
4.3.1 Problem Statement	112

4.3.2 Terminology: Configurations and Scores	113
4.3.3 Analysis of Our Given Problem	113
4.4 A Recipe for Dynamic Programming	116
4.5 Longest Common Subsequence	116
4.5.1 Problem Statement	117
4.5.2 Prefixes	118
4.5.3 Relations Among Subproblems	118
4.5.4 A Recurrence for the LCS	119
4.6 Exercises	123
Chapter 5 . RNA Secondary Structure Prediction	125
5.1 Motivation	126
5.2 Introduction to the Problem	128
5.2.1 NATURE	129
5.2.1.1 <i>Where Do Hydrogen Bonds Form?</i>	129
5.2.1.2 <i>Thermodynamic Issues</i>	130
5.2.1.3 <i>Consensus Sequence Patterns</i>	132
5.2.1.4 <i>Complications</i>	133
5.2.2 SCIENCE	133
5.2.2.1 <i>Modeling Secondary Structure</i>	133
5.2.2.2 <i>Single Base Pairs</i>	134
5.2.2.3 <i>Stacking Energy Models</i>	134
5.2.3 COMPUTATION	138
5.2.3.1 <i>Display of Secondary Structure</i>	139
5.2.4 Restating the Problem	145
5.3 The Nussinov dynamic programming algorithm	146
5.3.1 Execution Time	155
5.4 The MFOLD algorithm: terminology	155
5.4.1 The MFOLD Algorithm: Recursion	160
5.4.2 MFOLD Extensions	162
5.4.3 MFOLD Execution Time	162
5.5 Exercises	163
References	164
Chapter 6 . Protein Sequence Alignment	167
6.1 Protein Homology	167
6.1.1 NATURE	168
6.1.2 SCIENCE	170
6.1.2.1 <i>Partial Matches</i>	172
6.1.2.2 <i>Building a BLOSUM Matrix</i>	173
6.1.2.3 <i>Gaps</i>	179
6.1.2.4 <i>Summary</i>	180
6.1.3 COMPUTATION	180
6.1.3.1 <i>Subproblem Specification</i>	181
6.1.3.2 <i>Scoring Alignments</i>	181

6.1.3.3 Suitability of the Subproblem	182
6.1.3.4 A Global Alignment Example	186
6.2 Variations in the Global Alignment Algorithm	186
6.3 The Significance of a Global Alignment	187
6.3.1 Computer-Assisted Comparison	188
6.3.2 Percentage Identity Comparison	189
6.4 Local Alignment	190
6.5 Exercises	193
References	195
Chapter 7 . Protein Geometry	197
7.1 Motivation	197
7.2 Introduction	198
7.3 Calculations related to protein geometry	198
7.3.1 Inter-atomic Distance	198
7.3.2 Bond Angle	198
7.3.3 Dihedral Angles	199
7.3.3.1 Defining Dihedral Angles	199
7.3.3.2 Computation of a Normal	201
7.3.3.3 Calculating the Phi Dihedral Angle	204
7.3.3.4 Sign of the Dihedral Angle	204
7.3.3.5 Calculating the Psi Dihedral Angle	206
7.4 Ramachandran plots	206
7.5 Inertial Axes	212
7.6 Exercises	216
References	220
Chapter 8 . Coordinate Transformations	223
8.1 Motivation	223
8.2 Introduction	224
8.3 Translation Transformations	224
8.3.1 Translation to Find Centroid at the Origin	224
8.4 Rotation Transformations	225
8.4.1 Rotation Transformations in the Plane	226
8.4.2 Rotations in 3-D Space	227
8.5 Isometric transformations	231
8.5.1 Our Setting Is a Euclidean Vector Space	232
8.5.2 Orthogonality of A Implies Isometry of T	232
8.5.3 Isometry of T Implies Orthogonality of A	233
8.5.4 Preservation of Angles	234
8.5.5 More Isometries	234
8.5.6 Back to Rotations in the Plane	235
8.5.7 Rotations in the 3-D Space: A Summary	238
8.6 Exercises	238
References	239

Chapter 9 . Structure Comparison, Alignment, and Superposition	241
9.1 Motivation	242
9.2 Introduction	245
9.2.1 Specifying the Problem	245
9.3 Techniques for Structural Comparison	246
9.4 Scoring Similarities and Optimizing Scores	247
9.5 Superposition Algorithms	247
9.5.1 Overview	247
9.5.2 Characterizing the Superposition Algorithm	249
9.5.3 Formal Problem Description	249
9.5.4 Computations to Achieve Maximal Overlap	251
9.5.5 Summary	257
9.5.6 Measuring Overlap	259
9.5.6.1 <i>Calculation of the Root Mean Square Deviation (RMSD)</i>	259
9.5.6.2 <i>RMSD Issues</i>	259
9.5.7 Dealing with Weaker Sequence Similarity	260
9.5.8 Strategies Based on a Distance Matrix	261
9.6 Algorithms comparing relationships within proteins	263
9.6.1 <i>Dali</i>	263
9.6.2 SSAP	267
9.6.2.1 <i>Motivation</i>	267
9.6.2.2 <i>Introduction to SSAP</i>	269
9.6.2.3 <i>Overview of SSAP</i>	271
9.6.2.4 <i>Calculating the Views</i>	272
9.6.2.5 <i>Building the Consensus Matrix</i>	272
9.6.2.6 <i>Compute the Optimal Path in the Consensus Matrix</i>	278
9.7 Exercises	279
References	282
Chapter 10 . Machine Learning	285
10.1 Motivation	285
10.2 Issues of Complexity	287
10.2.1 Computational Scalability	287
10.2.2 Intrinsic Complexity	287
10.2.3 Inadequate Knowledge	288
10.3 Prediction via Machine Learning	289
10.3.1 Training and Testing	291
10.4 Types of learning	292
10.4.1 Types of Supervised Learning	293
10.4.2 Supervised Learning: Notation and Formal Definitions	293
10.5 Objectives of the Learning Algorithm	294
10.6 Linear Regression	295
10.7 Ridge Regression	297
10.7.1 Predictors and Data Recording	299
10.7.2 Underfitting and Overfitting	300
10.8 Preamble for Kernel Methods	300

10.9 Kernel Functions	303
10.9.1 The “Kernel Trick”	304
10.9.2 Design Issues	305
10.9.3 Validation Data Sets	306
10.9.3.1 Holdout Validation	307
10.9.3.2 N-Fold Cross Validation	307
10.10 Classification	308
10.10.1 Classification as Machine Learning	309
10.10.2 <i>Ad Hoc</i> Classification	310
10.11 Heuristics for Classification	311
10.11.1 Feature Weighting	311
10.12 Nearest Neighbour Classification	312
10.12.1 Delaunay and Voronoi	313
10.12.2 Nearest Neighbour Time and Space Issues	315
10.13 Support Vector Machines	315
10.13.1 Linear Discrimination	315
10.13.2 Margin of Separation	318
10.13.3 Support Vectors	319
10.13.4 The SVM as an Optimization Problem	320
10.13.5 The Karush–Kuhn–Tucker Condition	322
10.13.6 Evaluation of w_0	322
10.14 Linearly Non-separable Data	323
10.14.1 Parameter Values	326
10.14.2 Evaluation of w_0 (Soft Margin Case)	327
10.14.3 Classification with Soft Margin	327
10.15 Support Vector Machines and Kernels	328
10.16 Expected Test Error	328
10.17 Transparency	329
10.18 Exercises	331
References	334
APPENDICES	337
INDEX	385