

# Contents

Preface	xxiii
How to Use This Book	xxv
Courses of Different Lengths	xxix
Acknowledgments	xxxi
<b>1</b>	<b>Introduction to Protein Structure and NMR</b> 1
1.1	Protein Structure 1
1.2	Structure Determination of Proteins with NMR Spectroscopy 2
<b>2</b>	<b>Basic Principles of NMR</b> 7
2.1	Overview of NMR 7
2.2	The Physical Basis of NMR Spectroscopy 8
2.3	Chemical Shifts 10
2.4	Introduction to NMR Experiments 11
<b>3</b>	<b>Proteins and NMR Structural Biology</b> 15
3.1	COSY 15
3.2	$^3J_{\text{HNH}\alpha}$ 15
3.3	$\text{H}^{\text{N}}\text{-}^{15}\text{N}$ HSQC 17
3.4	$^{15}\text{N}$ TOCSY 17
3.5	NOESY 18
3.6	RDC 19
<b>4</b>	<b>MBM, SVD, PCA, and RDCs</b> 23
4.1	MBM 23
4.2	SVD 24
4.2.1	Definition 24
4.2.2	Properties 24
4.3	PCA 25
4.3.1	Calculating PCA by SVD 25
4.4	RDCs 25

<b>5</b>	<b>Principal Components Analysis, Residual Dipolar Couplings, and Their Relationship in NMR Structural Biology</b>	<b>27</b>
	<b>Antony K. Yan and Bruce R. Donald</b>	
5.1	Introduction	27
5.2	Introduction to PCA	28
5.3	Residual Dipolar Couplings in Structural Biochemistry	34
5.4	RDCs and PCA	38
5.5	Conclusions and Future Work	43
<b>6</b>	<b>Orientalional Sampling of Interatomic Vectors</b>	<b>47</b>
6.1	Introduction	47
6.2	Theory	47
6.2.1	Sampling Tensor	47
6.2.2	Generalized Sampling Parameter	48
6.2.3	Average Constant	48
6.2.4	Generalized Quality Factor	48
6.2.5	Geometric Representation	49
6.3	Results	49
6.4	Applications	51
<b>7</b>	<b>Solution Structures of Native and Denatured Proteins Using RDCs</b>	<b>53</b>
7.1	Determining Native Protein Structure	53
7.1.1	Theoretical Background	53
7.1.2	The Algorithm	54
7.1.3	Results	55
7.2	Determination of Denatured or Disordered Proteins	55
7.2.1	A Probabilistic Interpretation of Restraints in the Denatured State	56
7.2.2	The Algorithm	56
7.2.3	Applications to Biological Systems	58
<b>8</b>	<b>JIGSAW and NMR</b>	<b>59</b>
8.1	Overview of JIGSAW	59
8.2	NMR Spectra Used in JIGSAW	59
8.3	Graph Representation of Atom Interactions in NOESY Spectra	60
8.3.1	Graph Representation	60
8.3.2	Graph Constraints for Identifying Secondary Structure	61
8.4	Secondary Structure Pattern Discovery	61
8.5	Assignment by Alignment of Side-Chain Fingerprints	64
8.5.1	Experimental Results	65
<b>9</b>	<b>Peptide Design</b>	<b>67</b>
9.1	Peptides	67
9.2	Peptide Backbone Reconstruction	68
9.2.1	Problem Statement	68
9.2.2	Motivation	68
9.2.3	Algorithm	69
9.2.4	Results	69

9.3	Peptides That Target Transmembrane Helices	70
9.3.1	Algorithm	70
9.3.2	Results	71
9.4	Foldamers	71
9.4.1	Types of Monomer Frameworks	72
9.4.2	Foldamer Structure	72
9.4.3	Foldamer Function	72
9.4.4	Foldamer Benefits	73
<b>10</b>	<b>Protein Interface and Active Site Redesign</b>	<b>77</b>
10.1	Minimalist Active Site Redesign	77
10.1.1	Subtilisin	78
10.1.2	Interconverting Homologous Enzymes	79
10.1.3	Introduction of Catalytic Machinery	80
10.1.4	Removal of Catalytic Nucleophiles	81
10.1.5	Partitioning of Reaction Intermediates	81
10.1.6	Controlling Stereo- and Regiochemistry	81
10.1.7	Improving Promiscuity	82
10.2	Protein Domain Interface Redesign via Directed Evolution	83
<b>11</b>	<b>Computational Protein Design</b>	<b>87</b>
11.1	Introduction	87
11.2	Overview of Methodology	87
11.3	Algorithm Design	88
11.4	Intuition: Dead-End Elimination	91
11.5	Complexity Analysis	92
11.6	Experimental Validation: Interplay of Computational Protein Design and NMR	92
<b>12</b>	<b>Nonribosomal Code and <math>K^*</math> Algorithms for Ensemble-Based Protein Design</b>	<b>97</b>
12.1	Nonribosomal Peptide Synthetase (NRPS) Enzymes	97
12.2	$K^*$ Algorithm Basics	98
12.3	Energy Functions	102
12.4	Redesigning Enzymes with $K^*$	105
12.5	Minimized Dead-End Elimination (minDEE)	106
12.5.1	$A^*$ Search and minDEE	106
12.6	Backbone Flexibility in DEE for Protein Design	107
12.6.1	Continuous Backbone Flexibility DEE	107
12.6.2	Backrub DEE	108
12.7	Application to Negative Design	109
12.8	Discussion	110
<b>13</b>	<b>RDCs in NMR Structural Biology</b>	<b>115</b>
13.1	Residual Dipolar Couplings	115
13.2	Computational Topics Related to RDCs	116
13.2.1	Assignment Problem	116
13.2.2	Structure Determination Problem	116

13.2.3	Estimation of Alignment Tensor Without Assignments	117
13.2.4	Structural Homology Detection	117
<b>14</b>	<b>Nuclear Vector Replacement</b>	<b>119</b>
14.1	Experimental Input	119
14.2	Nuclear Vector Replacement	119
14.2.1	Tensor Estimation	120
14.2.2	Resonance Assignment	121
14.3	An Expectation/Maximization NVR Algorithm	122
14.4	3D Structural Homology Detection via NVR	123
14.5	Matching Modulo a Group, and Clustering Modulo a Group	123
<b>15–</b>	<b>Short Course: Automated NMR Assignment and Protein Structure Determination</b>	
<b>18</b>	<b>Using Sparse Residual Dipolar Couplings</b>	<b>127</b>
	<b>Bruce R. Donald and Jeffrey Martin</b>	
15.1	Introduction	127
15.1.1	Motivation	127
15.1.2	Glossary of Abbreviations	128
15.1.3	Background	129
16.1	The Power of Exact Solutions	134
16.1.1	Computing the Globally Optimal Solution	142
16.1.2	Limitations and Extensions	144
17.1	NMR Structure Determination Algorithms Using Sparse RDCs	145
17.2	Nuclear Vector Replacement for Automated NMR Assignment and Structure Determination	149
17.3	Protein Fold Determination via <i>Unassigned</i> Residual Dipolar Couplings	153
17.4	Automated NOE Assignment Using a Rotamer Library Ensemble and RDCs	155
17.5	NMR Structure Determination of Symmetric Homo-Oligomers	158
17.6	Applications and Connections to Other Biophysical Methods	160
18.1	Looking Under the Hood: How the Algorithms Work, and Outlook for Future Developments	160
18.1.1	Exact Solutions for Computing Backbone Dihedral Angles from RDCs	161
18.1.2	Nuclear Vector Replacement and Fold Recognition Using Unassigned RDCs	166
18.1.3	Automated NOE Assignment	171
18.1.4	NMR Structure Determination of Symmetric Oligomers	172
<b>19</b>	<b>Proteomic Disease Classification Algorithm</b>	<b>187</b>
19.1	Proteomic Disease Classification	187
19.1.1	Methods	187
19.1.2	Q5: An MSCA Algorithm	188
19.2	Results and Discussion	189
<b>20</b>	<b>Protein Flexibility: Introduction to Inverse Kinematics and the Loop Closure Problem</b>	<b>191</b>
20.1	Loop Closure Problem and Exact Inverse Kinematics	191
20.1.1	Protein Backbone Representations	191
20.1.2	Loop Closure Problem	191
20.1.3	Denavit-Hartenberg Local Frames	192

20.2	Probik	193
20.2.1	Overview	193
20.2.2	Algorithm Description	193
20.2.3	Exploring Control Parameters Based on Principal Component Analysis	193
20.3	ChainTweak	193
20.3.1	Overview	193
20.3.2	Algorithm Description	194
20.4	Comparisons Between Probik and ChainTweak	195
<b>21</b>	<b>Normal Mode Analysis (NMA) and Rigidity Theory</b>	<b>197</b>
21.1	Normal Mode Analysis	197
21.1.1	Introduction	197
21.1.2	Different Normal Modes	199
21.2	Protein Flexibility Predictions Using Graph Theory	200
21.2.1	Overview of FIRST	200
21.2.2	Rigidity Theory	200
21.2.3	Pebble Game Analysis	201
<b>22</b>	<b>ROCK and FRODA for Protein Flexibility</b>	<b>205</b>
22.1	The ROCK Algorithm	205
22.1.1	Terminology	205
22.1.2	Overview	205
22.1.3	Conformation Sampling in Single-Ring Closure	206
22.1.4	Conformation Sampling in Multiple-Ring Closure	206
22.1.5	Conformation Sampling in Side Branches	207
22.1.6	Hydrophobic Interactions and Ramachandran Checks	207
22.2	Application of ROCK in Flexible Docking	208
22.3	FRODA	208
22.3.1	Overview	208
22.3.2	The FRODA Algorithm	208
22.3.3	Comparisons Between ROCK and FRODA	210
<b>23</b>	<b>Applications of NMA to Protein-Protein and Ligand-Protein Binding</b>	<b>213</b>
23.1	Structure Changes for Protein Binding in the Unbound State	213
23.1.1	Classical Models for Protein-Protein Interactions	213
23.1.2	Gaussian Network Model (GNM)	213
23.1.3	Anisotropic Network Model (ANM)	214
23.2	Receptor Flexibility Representation Through Relevant Normal Modes	215
23.2.1	Methodology Overview	215
23.2.2	Determination of the Relevant Normal Mode	215
23.2.3	Generation of MRCs	216
23.2.4	Side-Chain Optimization	216
23.2.5	Small-Scale Virtual Screening Using RED	216

<b>24</b>	<b>Modeling Equilibrium Fluctuations in Proteins</b>	<b>219</b>
24.1	Missing Loops and Protein Flexibility	219
24.2	Materials and Methods	220
24.2.1	Fragment Ensemble Method (FEM)	220
24.2.2	Protein Ensemble Method (PEM)	221
24.3	Results	226
<b>25</b>	<b>Generalized Belief Propagation, Free Energy Approximations, and Protein Design</b>	<b>227</b>
25.1	Free Energy	227
25.2	Graphical Models	228
25.2.1	Bayesian Networks	228
25.2.2	Pairwise Markov Random Fields	229
25.2.3	Factor Graphs	229
25.3	Belief Propagation (BP)	229
25.4	The Connection Between Belief Propagation and Free Energy	230
25.5	Generalized Belief Propagation (GBP)	231
25.6	An Application of GBP: Estimating the Free Energy of Protein Structures	231
25.6.1	Results and Discussion	232
25.7	Application: Graphical Models for Protein Design	233
25.7.1	Protein Design Problem	235
25.7.2	Graphical Models and Belief Propagation for Protein Design	236
25.7.3	Multiple Low-Energy Sequences Through BP	237
25.7.4	Graphical Models for Probabilistic Protein Design	238
25.7.5	Discussion and Future Directions	240
<b>26</b>	<b>Ligand Configurational Entropy</b>	<b>245</b>
26.1	Experimental Input	245
26.2	Entropy	245
26.3	Entropy in Ligand Binding	246
26.3.1	Conformational Entropy	246
26.3.2	Vibrational Entropy	246
26.4	Entropy and Amprenavir	246
26.5	Implications for Design	247
<b>27</b>	<b>Carrier Protein Structure and Recognition in Peptide Biosynthesis</b>	<b>249</b>
27.1	Carrier Proteins	249
<b>28</b>	<b>Kinetic Studies of the Initial Module PheATE of Gramicidin S Synthetase</b>	<b>253</b>
28.1	Background	253
28.2	Binding of the Amino Acid Substrate to the A Domain of GrsA	254
28.3	Aminoacyl-AMP Formation Catalyzed by the A Domain	254
28.3.1	The Steady-State Assays	254
28.3.2	The Pre-Steady-State Assay	255
28.4	Loading of the Amino Acid Substrate to the T Domain	255
28.5	Epimerization of the L-Form Substrate-Enzyme Complex to D-Form by the E Domain	256
28.6	Free Energy Profiles for HoloPheATE Catalysis	256

- 29 Protein-Ligand NOE Matching 259**
  - 29.1 Background 259
  - 29.2 Methods 260
  - 29.3 Results and Discussion 262
  
- 30 Side-Chain and Backbone Flexibility in Protein Core Design 265**
  - 30.1 Protein Modeling with Fixed or Flexible Backbone 265
  - 30.2 SoftROC 266
    - 30.2.1 Initializing Backbone Population 266
    - 30.2.2 Optimization with Genetic Algorithm 266
    - 30.2.3 Refining the Model with Monte Carlo Sampling 268
    - 30.2.4 Final Model 268
  - 30.3 Issues on Energy Calculations 268
  - 30.4 Results: Comparison to ROC Variants 269
    - 30.4.1 ROC Settings 269
    - 30.4.2 Experiments on 434 cro 269
    - 30.4.3 Experiments on T4 Lysozyme 270
  
- 31 Distance Geometry 273**
  - 31.1 The Molecule Problem 273
  - 31.2 Divide and Conquer 274
  - 31.3 Conditions for Unique Realizability 274
  - 31.4 Graph Partitioning 275
  - 31.5 Realizing Subgraphs 276
  - 31.6 Conclusion 277
  
- 32 Distance Geometry: NP-Hard, NP-Hard to Approximate 279**
  - 32.1 Introduction 279
    - 32.1.1 Review: Reductions 279
    - 32.1.2 NP-Hard Problems 280
  - 32.2 Reduction from Partition to 1-Embeddability 280
  - 32.3 Reduction from 3SAT to {1,2} 1-Embeddability 280
  - 32.4 Reduction from 3SAT to Integer 1-Embeddability 282
  - 32.5 Adding Dimensions 282
  - 32.6 Approximation 282
    - 32.6.1 Definition of  $\varepsilon$ -Approximate  $k$ -Embeddability 282
  
- 33 A Topology-Constrained Network Algorithm for NOESY Data Interpretation 285**
  - 33.1 Algorithms 285
  - 33.2 Results 291
  
- 34 MARS: An Algorithm for Backbone Resonance Assignment 293**
  - 34.1 MARS—Backbone Assignment of Proteins 293
    - 34.1.1 Backbone Resonance Assignment 293
    - 34.1.2 Method 293
    - 34.1.3 Results and Discussion 296

34.2	Backbone Assignment with Known Structure Using RDCs	296
34.2.1	Method	298
34.2.2	Results and Discussion	298
<b>35</b>	<b>Errors in Structure Determination by NMR Spectroscopy</b>	<b>301</b>
35.1	Errors in Published Protein Folds	301
35.2	Case Study: Dynein Light Chain	301
35.3	Identifying the Problems: Problems in Identifiers	303
<b>36</b>	<b>SemiDefinite Programming and Distance Geometry with Orientation Constraints</b>	<b>307</b>
36.1	SemiDefinite Programming and Two Applications	307
36.1.1	Overview of SemiDefinite Programming	307
36.1.2	Application in the Side-Chain Positioning Problem	307
36.1.3	Application in the Sensor Network Localization Problem	309
36.2	Distance Geometry with Orientation Constraints	310
36.2.1	Graph Embedding with Angle Information	310
36.2.2	Protein Structure Determination from RDCs	311
<b>37</b>	<b>Graph Cuts for Energy Minimization and Assignment Problems</b>	<b>315</b>
37.1	Construction of the Energy Function	315
37.2	Optimizing the Energy Function by Graph Cuts	316
37.2.1	Graph Construction	316
37.2.2	The MultiWay Cut Formulation	317
37.2.3	The MultiWay Cut Algorithm	317
37.3	Graph Cuts for Computing Visual Correspondence with Occlusions	318
37.3.1	Notation	318
37.3.2	Energy Function	318
37.3.3	The $\alpha$ -Expansion Move Algorithm	319
<b>38</b>	<b>Classifying the Power of Graph Cuts for Energy Minimization</b>	<b>323</b>
38.1	Feature Space Clustering	323
38.2	Energy Minimization Framework for Feature Space Clustering	323
38.2.1	The Pixel Labeling Problem	324
38.2.2	An EM-Style Energy Minimization Algorithm	324
38.3	Approaches to Incorporating Spatial Coherence	326
38.4	Classifying Energy Functions that Can Be Minimized Efficiently Using Graph Cuts	326
38.4.1	Using Graph Cuts in Energy Minimization	327
38.5	Representation of Energy Functions by Graphs	327
38.6	The Class $\mathcal{F}^2$	328
38.6.1	Graph Construction for $\mathcal{F}^2$	328
38.6.2	NP-Hardness of General $E^2$ Functions	329
38.7	The Class $\mathcal{F}^3$	330
38.7.1	Graph Construction for $\mathcal{F}^3$	330
38.8	Comments	332



<b>39</b>	<b>Protein Unfolding by Using Residual Dipolar Couplings</b>	<b>333</b>
39.1	Motivation and Overview	333
39.2	Ensemble Computation Using Only Local Sampling	333
39.3	Ensemble Computation with Both Local Sampling and Long-Range Order	335
39.4	An Unfolded Protein Structure Model from RDCs and Small-Angle X-Ray Scattering (SAXS) Data	337
39.4.1	Generation of the Conformation Ensemble	337
39.4.2	RDC Computation from the Conformational Ensemble	337
39.4.3	Prediction of SAXS Data from the Conformational Ensemble	337
<b>40</b>	<b>Structure-Based Protein-Ligand Binding</b>	<b>341</b>
40.1	Uncertainty in Experimentally Derived Structures	341
40.1.1	Uncertainty in X-Ray Structures	341
40.1.2	Uncertainty in NMR Structures	342
40.2	Protein Dynamics	342
40.3	Probabilistic Representations of Uncertainty and Dynamics	343
40.4	Representation of Protein Flexibility: Ensemble Docking	343
40.5	FDS: Flexible Ligand and Receptor Docking with a Continuum Solvent Model and Soft-Core Energy Function	344
<b>41</b>	<b>Flexible Ligand-Protein Docking</b>	<b>345</b>
41.1	Predicting Binding Energetics from Structure	345
41.2	Flexible Docking in Solution Using Metadynamics	346
41.2.1	Overview of Metadynamics	346
41.2.2	Application of Metadynamics in Flexible Docking	347
41.2.3	Results	349
<b>42</b>	<b>Analyzing Protein Structures Using an Ensemble Representation</b>	<b>351</b>
42.1	Mathematical Results	351
42.1.1	Terminology	351
42.1.2	Results	352
42.1.3	Brief Proof	352
42.2	Biological Significance	353
<b>43</b>	<b>NMR Resonance Assignment Assisted by Mass Spectrometry</b>	<b>355</b>
43.1	Motivation	355
43.2	Mass Spectrometry-Assisted NMR Assignment	355
43.2.1	Principle of the Approach	355
43.2.2	Extracting HX Rates by HSQC	355
43.2.3	Extracting HX Rates by MS	357
43.2.4	Correlating HX Rates Between NMR and MS	357
43.2.5	MS-Assisted Assignment	357
43.3	MS-Assisted NMR Assignment in Reductively <sup>13</sup> C-Methylated Proteins	357

<b>44</b>	<b>Autolink: An Algorithm for Automated NMR Resonance Assignment</b>	<b>363</b>
44.1	Algorithm Overview	363
44.2	Spin System Pair Scoring	365
44.2.1	Spin Density Bias	365
44.2.2	Assigned Spin Bias	365
44.2.3	Offset Bias	367
44.2.4	Atomic Assignment Bias	367
44.2.5	Overall Spin System Pair Scoring	367
44.3	Hypothesis Evaluation/Reevaluation Cycles	367
44.3.1	Calculation of the Base Priority Prime List	367
44.3.2	Calculation of the Relative Priority Prime List	369
<b>45</b>	<b>CS-Rosetta: Protein Structure Generation from NMR Chemical Shift Data</b>	<b>371</b>
45.1	Introduction	371
45.1.1	Rosetta	372
45.1.2	CS-Rosetta	373
45.2	Results	374
<b>46</b>	<b>Enzyme Redesign by SVM</b>	<b>377</b>
46.1	Overview	377
46.2	Data Representation	377
46.3	The Support Vector Machine (SVM) Approach	378
46.4	Results	381
<b>47</b>	<b>Cross-Rotation Analysis Algorithm</b>	<b>383</b>
47.1	CRANS	383
47.1.1	Methods	384
47.1.2	Complexity	384
<b>48</b>	<b>Molecular Replacement and NCS in X-ray Crystallography</b>	<b>387</b>
48.1	Background	387
48.1.1	The Phase Problem	387
48.1.2	Molecular Replacement	387
48.2	NMA in Molecular Replacement	388
48.2.1	Objectives	388
48.2.2	Normal Modes and Elastic Network Models (ENM)	389
48.2.3	Summary	390
48.3	NCS-Constrained Exhaustive Search Using Oligomeric Models	390
48.3.1	Methods	391
48.3.2	Examples	391
<b>49</b>	<b>Optimization of Surface Charge-Charge Interactions</b>	<b>393</b>
49.1	Algorithm Input	393
49.2	Genetic Algorithm	393
49.2.1	Chromosome Scoring	394
49.2.2	Parental Chromosome Selection and Crossover	395
49.2.3	Child Chromosome Mutation	395

49.3	Computational and Experimental Validations	395
49.3.1	Computational Validations	397
49.3.2	Experimental Validation	397
<b>50</b>	<b>Computational Topology and Protein Structure</b>	<b>399</b>
50.1	Topology	399
50.2	Homology	400
50.3	Simplicial Complexes	401
50.4	Homology Type Is Effectively Computable	402
50.4.1	Complexity	403
50.4.2	Applications	403
50.4.3	Foundations	404
50.5	Computing Homology Groups	404
50.5.1	Simplicial Homology	404
50.5.2	Computing the Homology Groups	405
50.5.3	The Algorithm for Homology Group Computation	407
50.6	Alpha Shapes ( $\alpha$ -Shapes) and Applications to Protein Structure	410
50.7	Conclusions and Future Work	411
	Index	415