Contents

Preface xxiii
  How to Use This Book xxv
  Courses of Different Lengths xxix
Acknowledgments xxxi

1 Introduction to Protein Structure and NMR 1
  1.1 Protein Structure 1
  1.2 Structure Determination of Proteins with NMR Spectroscopy 2

2 Basic Principles of NMR 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

3 Proteins and NMR Structural Biology 15
  3.1 COSY 15
  3.2 $J_{HNHu}$ 15
  3.3 $HN^{15}$N HSQC 17
  3.4 $^{15}$N TOCSY 17
  3.5 NOESY 18
  3.6 RDC 19

4 MBM, SVD, PCA, and RDCs 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
5 Principal Components Analysis, Residual Dipolar Couplings, and Their Relationship in NMR Structural Biology 27
Antony K. Yan and Bruce R. Donald
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

6 Orientational Sampling of Interatomic Vectors 47
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

7 Solution Structures of Native and Denatured Proteins Using RDCs 53
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

8 JIGSAW and NMR 59
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

9 Peptide Design 67
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

10 Protein Interface and Active Site Redesign 77
  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

11 Computational Protein Design 87
  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

12 Nonribosomal Code and K* Algorithms for Ensemble-Based Protein Design 97
  12.1 Nonribosomal Peptide Synthetase (NRPS) Enzymes 97
  12.2 K-star (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

13 RDCs in NMR Structural Biology 115
  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
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

21 Normal Mode Analysis (NMA) and Rigidity Theory 197
  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

22 ROCK and FRODA for Protein Flexibility 205
  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

23 Applications of NMA to Protein-Protein and Ligand-Protein Binding 213
  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
24 Modeling Equilibrium Fluctuations in Proteins 219
   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

25 Generalized Belief Propagation, Free Energy Approximations, and Protein Design 227
   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

26 Ligand Configurational Entropy 245
   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

27 Carrier Protein Structure and Recognition in Peptide Biosynthesis 249
   27.1 Carrier Proteins 249

28 Kinetic Studies of the Initial Module PheATE of Gramicidin S Synthetase 253
   28.1 Background 253
   28.2 Binding of the Amino Acid Substrate to the A Domain of GrsA 254
   28.3 Aminoaeryl-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 Realizibility 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 $\epsilon$-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
39 Protein Unfolding by Using Residual Dipolar Couplings 333

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

40 Structure-Based Protein-Ligand Binding 341

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

41 Flexible Ligand-Protein Docking 345

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

42 Analyzing Protein Structures Using an Ensemble Representation 351

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

43 NMR Resonance Assignment Assisted by Mass Spectrometry 355

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 $^{13}$C-Methylated Proteins 357
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>Autolink: An Algorithm for Automated NMR Resonance Assignment</td>
<td>363</td>
</tr>
<tr>
<td>44.1</td>
<td>Algorithm Overview</td>
<td>363</td>
</tr>
<tr>
<td>44.2</td>
<td>Spin System Pair Scoring</td>
<td>365</td>
</tr>
<tr>
<td>44.2.1</td>
<td>Spin Density Bias</td>
<td>365</td>
</tr>
<tr>
<td>44.2.2</td>
<td>Assigned Spin Bias</td>
<td>365</td>
</tr>
<tr>
<td>44.2.3</td>
<td>Offset Bias</td>
<td>367</td>
</tr>
<tr>
<td>44.2.4</td>
<td>Atomic Assignment Bias</td>
<td>367</td>
</tr>
<tr>
<td>44.2.5</td>
<td>Overall Spin System Pair Scoring</td>
<td>367</td>
</tr>
<tr>
<td>44.3</td>
<td>Hypothesis Evaluation/Reevaluation Cycles</td>
<td>367</td>
</tr>
<tr>
<td>44.3.1</td>
<td>Calculation of the Base Priority Prime List</td>
<td>367</td>
</tr>
<tr>
<td>44.3.2</td>
<td>Calculation of the Relative Priority Prime List</td>
<td>369</td>
</tr>
<tr>
<td>45</td>
<td>CS-Rosetta: Protein Structure Generation from NMR Chemical Shift Data</td>
<td>371</td>
</tr>
<tr>
<td>45.1</td>
<td>Introduction</td>
<td>371</td>
</tr>
<tr>
<td>45.1.1</td>
<td>Rosetta</td>
<td>372</td>
</tr>
<tr>
<td>45.1.2</td>
<td>CS-Rosetta</td>
<td>373</td>
</tr>
<tr>
<td>45.2</td>
<td>Results</td>
<td>374</td>
</tr>
<tr>
<td>46</td>
<td>Enzyme Redesign by SVM</td>
<td>377</td>
</tr>
<tr>
<td>46.1</td>
<td>Overview</td>
<td>377</td>
</tr>
<tr>
<td>46.2</td>
<td>Data Representation</td>
<td>377</td>
</tr>
<tr>
<td>46.3</td>
<td>The Support Vector Machine (SVM) Approach</td>
<td>378</td>
</tr>
<tr>
<td>46.4</td>
<td>Results</td>
<td>381</td>
</tr>
<tr>
<td>47</td>
<td>Cross-Rotation Analysis Algorithm</td>
<td>383</td>
</tr>
<tr>
<td>47.1</td>
<td>CRANS</td>
<td>383</td>
</tr>
<tr>
<td>47.1.1</td>
<td>Methods</td>
<td>384</td>
</tr>
<tr>
<td>47.1.2</td>
<td>Complexity</td>
<td>384</td>
</tr>
<tr>
<td>48</td>
<td>Molecular Replacement and NCS in X-ray Crystallography</td>
<td>387</td>
</tr>
<tr>
<td>48.1</td>
<td>Background</td>
<td>387</td>
</tr>
<tr>
<td>48.1.1</td>
<td>The Phase Problem</td>
<td>387</td>
</tr>
<tr>
<td>48.1.2</td>
<td>Molecular Replacement</td>
<td>387</td>
</tr>
<tr>
<td>48.2</td>
<td>NMA in Molecular Replacement</td>
<td>388</td>
</tr>
<tr>
<td>48.2.1</td>
<td>Objectives</td>
<td>388</td>
</tr>
<tr>
<td>48.2.2</td>
<td>Normal Modes and Elastic Network Models (ENM)</td>
<td>389</td>
</tr>
<tr>
<td>48.2.3</td>
<td>Summary</td>
<td>390</td>
</tr>
<tr>
<td>48.3</td>
<td>NCS-Constrained Exhaustive Search Using Oligomeric Models</td>
<td>390</td>
</tr>
<tr>
<td>48.3.1</td>
<td>Methods</td>
<td>391</td>
</tr>
<tr>
<td>48.3.2</td>
<td>Examples</td>
<td>391</td>
</tr>
<tr>
<td>49</td>
<td>Optimization of Surface Charge-Charge Interactions</td>
<td>393</td>
</tr>
<tr>
<td>49.1</td>
<td>Algorithm Input</td>
<td>393</td>
</tr>
<tr>
<td>49.2</td>
<td>Genetic Algorithm</td>
<td>393</td>
</tr>
<tr>
<td>49.2.1</td>
<td>Chromosome Scoring</td>
<td>394</td>
</tr>
<tr>
<td>49.2.2</td>
<td>Parental Chromosome Selection and Crossover</td>
<td>395</td>
</tr>
<tr>
<td>49.2.3</td>
<td>Child Chromosome Mutation</td>
<td>395</td>
</tr>
</tbody>
</table>
49.3 Computational and Experimental Validations 395
   49.3.1 Computational Validations 397
   49.3.2 Experimental Validation 397

50 Computational Topology and Protein Structure 399
   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