Protein Non-Folding: A Molecular Simulation Study of the Structure and Self-Aggregation of Elastin

by

Sarah Rauscher

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy

Department of Biochemistry
University of Toronto

© Copyright by Sarah Rauscher 2012
Protein Non-Folding: A Molecular Simulation Study of the Structure and Self-Aggregation of Elastin

Sarah Rauscher

Doctor of Philosophy

Department of Biochemistry
University of Toronto

2012

Abstract

Elastin is the protein responsible for the elastic recoil of skin, arteries, and lungs. In order to function, elastin, as well as other self-assembling elastomeric proteins, must remain disordered even when aggregated, and must therefore avoid the formation of amyloid fibrils. Elastin undergoes phase separation and self-organization into a fibrillar structure upon increasing temperature. In order to investigate the molecular basis for elastin self-aggregation, we performed extensive atomistic molecular dynamics simulations of both monomers and aggregates. To obtain statistical convergence of average structural properties, it was necessary to use massive sampling in combination with simulated tempering distributed replica sampling. The development and testing of this method is an important contribution of this thesis.

The first key finding of this thesis is that the sequences of all known rubber-like elastomeric proteins are separated from amyloidogenic sequences by a threshold in the combined content of proline and glycine. Our results suggest that avoiding the formation of a water-
excluding core involving extensive self-interactions is not only a fundamental requirement, but may very well constitute the single most essential design principle of self-assembling elastomeric proteins. The conformational ensemble of an elastin-like aggregate resembles that of a polymer melt. In the aggregate, individual chains retain significant hydration and do not form a water-excluding hydrophobic core, but are also completely entangled with each other and form significant intermolecular interactions. As a result, the overall chain dimensions approach the expected dimensions of chains in an ideal solvent, a state in which chain entropy is maximized. This is the prediction of the Flory theorem for generic polymer chains within a polymer melt, but has never, to our knowledge, been observed before for an aggregate of polypeptide chains in atomistic detail. Finally, we note that our results are not consistent with the current model of elastin self-aggregation, which involves a conformational transition of a monomer towards a more ordered, aggregation-prone state. Instead, we propose a model in which both the hydrophobic effect and the enhanced chain entropy afforded by the interactions with other peptides within the aggregate favour elastin self-aggregation.
Acknowledgements

Because of the exceptional group of people acknowledged here, the past seven years have been an incredible learning experience.

First, I am grateful to my advisor, Dr. Régis Pomès, for teaching me the fundamentals of scientific “detective work,” for sharing his enthusiasm for biomolecular simulation, and for encouraging me to explore without limits from the very beginning. His guidance, mentorship, and attention to detail strengthened this work immeasurably.

Thanks to the other members of my supervisory committee, Dr. Hue Sun Chan, Dr. Fred Keeley, and Dr. Christopher Yip, for many valuable suggestions and insightful discussions.

From the first day that I started on this project, I have benefited tremendously from a theory-experiment collaboration that led to the results presented in Chapter 4. A special thanks to Dr. Fred Keeley, Dr. Ming Miao, Dr. Lisa Muiznieks, and the entire Keeley lab, who have patiently explained the essentials of coacervation to me far too many times to count, and collaborated to directly connect our theoretical and experimental findings.

My sincere gratitude to all the members of the Pomès laboratory (past and present) for creating an encouraging, fun, and supportive research environment. Thanks to David Caplan, Dr. Nilu Chakrabarti, Rowan Henry, Dr. John Holyoake, Dr. Loan Huynh, Kethika Kulleperuma, and Dr. Marty Kurylowicz for their kind advice and assistance over the years. Many thanks to Grace Li, Dr. Chris Madill, and Dr. Ching-Hsing Yu for managing our in-house data storage and cluster, without which this project would not have been possible. Additional thanks to Grace and Dr. Ana Nikolic for collaborating on related studies of other disordered peptides. I am grateful to Dr. Tom Rodinger for developing distributed replica sampling, and taking the time to explain it to me; this method to asynchronously couple replicas proved essential when we began studying larger systems. Thanks to Chris Neale for consistently
pushing the boundaries of modern MD simulations, and encouraging all of us to do the same, as well as for many useful discussions on free energy calculations and enhanced sampling that led in part to Chapter 5. I am grateful to Zhuyi Xue and Aditi Ramesh for asking the tough questions recently that led me to revisit many of the concepts I took for granted – I wish them the best in their continued investigations of elastin. Many thanks to Dr. Stéphanie Baud and Dr. Elisa Fadda for “showing me the ropes” in the early days of graduate school. Thanks also to Stéphanie for collaborating on the protein aggregation study in Chapter 4.

Thanks to the anonymous reviewers of the published material included in this thesis. Their careful reading and suggestions certainly improved the work. Additional thanks to Dr. Rohit Pappu for useful discussions on polymer physics and disordered proteins.

Although two related studies are not included in this thesis, I am grateful for both of these collaborations because they informed the research presented here in many ways. First, I am grateful to Dr. Julie Forman-Kay, Dr. Claudiu Gradinaru, Amir Mazoughi, and Dr. Mickaël Krzeminski for a collaboration to study the unfolded state. I am also grateful to Dr. Peter Tieleman and Dr. Mikyung Seo for a collaboration to parameterize a coarse-grained force field using our atomistic simulation data.

I would also like to acknowledge the resources that enabled this research. First, thanks to Compute Canada and SciNet for the unprecedented computational resources, and technical support that made our simulations in the last three chapters of the thesis possible. Thanks to SHARCNET for dedicated resource allocations and data storage. Additional thanks to the High Performance Compute Facility (HPF), operated by the Centre for Computational Biology (CCB) for CPU time. This work was partly supported by the award of a Canada Graduate Scholarship from the National Science and Education Research Council of Canada, as well as a fellowship from the Research Training Center at the Hospital for Sick Children.
Lastly, I offer my deepest gratitude to my family. Thanks to my parents, Josef and Darlene Rauscher, and to my sister and brother-in-law, Johanna and Jason McNulty, for their endless encouragement and unwavering support of everything I do. For these reasons, and many others, this thesis is dedicated to them.
## Contents

List of Tables ......................................................................................................................... xiv
List of Figures .......................................................................................................................... xv
List of Appendices .................................................................................................................... xix
List of Acronyms and Symbols ................................................................................................. xx

### CHAPTER 1  Preface and Overview of the Thesis ................................................................. 1

1.1 Protein Disorder ................................................................................................................... 2
1.2 Molecular Simulations of Protein Disorder ........................................................................ 2
1.3 Elastin and Other Rubber-like Elastomeric Proteins .......................................................... 4
1.4 Two Types of Protein Self-Aggregation: Elastomeric and Amyloidogenic ....................... 5
1.5 Enhanced Conformational Sampling is Essential ............................................................... 6
1.6 Massive Sampling Facilitated by the Rapid Evolution of Supercomputing ....................... 8
1.7 Thesis Contributions and Organization ........................................................................... 10
References .............................................................................................................................. 12

### CHAPTER 2  Introduction to the Study of Protein Disorder: Bioinformatics, Experimental Approaches, and Molecular Simulations ................................................................. 14

2.1 Summary .......................................................................................................................... 15
2.2 Introduction ...................................................................................................................... 15
2.3 Defining Disorder ........................................................................................................... 17
2.4 Biological Relevance of Protein Disorder ......................................................................... 20
2.5 Sequence Features of Disordered Proteins: Bioinformatics Approaches ......................... 24
2.6 Identifying and Characterizing Disordered Proteins: Experimental Approaches ............ 29
2.7 Molecular Simulations ...................................................................................................... 32
  2.7.1 Overview of Simulation Approaches ............................................................................ 32
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.7.2 Hybrid Approaches: Using Experimental Restraints in Simulations</td>
<td>32</td>
</tr>
<tr>
<td>2.8 De Novo Simulations</td>
<td>37</td>
</tr>
<tr>
<td>2.8.1 Challenges and Considerations</td>
<td>37</td>
</tr>
<tr>
<td>2.8.2 Conformational Sampling Methods</td>
<td>37</td>
</tr>
<tr>
<td>2.8.3 Simulations of Entropic Chains</td>
<td>40</td>
</tr>
<tr>
<td>2.8.4 Simulation and Experiment: Interdisciplinary Studies</td>
<td>41</td>
</tr>
<tr>
<td>2.9 Coarse-grained Simulations</td>
<td>44</td>
</tr>
<tr>
<td>2.10 Conclusions and Outlook</td>
<td>46</td>
</tr>
<tr>
<td>References</td>
<td>48</td>
</tr>
</tbody>
</table>

CHAPTER 3 Structural Disorder and Protein Elasticity                      | 60   |
| 3.1 Summary                                                             | 61   |
| 3.2 Introduction                                                        | 61   |
| 3.3 Elasticity and Elastic Mechanical Properties                       | 63   |
| 3.3.1 Disorder and Elasticity                                          | 63   |
| 3.3.2 Rubber-like Elasticity                                           | 64   |
| 3.3.3 Measures of Elasticity                                           | 65   |
| 3.4 Rubber-like Elastomeric Proteins                                   | 67   |
| 3.4.1 Elastin                                                          | 69   |
| 3.4.2 Resilin                                                          | 72   |
| 3.4.3 Spider Silk                                                      | 75   |
| 3.4.4 Abductin                                                         | 77   |
| 3.4.5 ColP from Byssal Threads                                         | 78   |
| 3.4.6 Other Rubber-like Elastomeric Proteins                           | 79   |
| 3.5 Sequence Features of Disordered Elastomeric Proteins               | 80   |
| 3.6 Conclusion and Perspectives                                        | 81   |
| References                                                             | 84   |
CHAPTER 7 The Effects of Increasing Temperature on the Conformational Ensembles of Monomeric Elastin-like Peptides

7.1 Summary ..........................................................196
7.2 Introduction ..........................................................196
  7.2.1 Elastin Undergoes Phase Separation upon Increasing Temperature ........................................196
  7.2.2 Phase Diagrams of Elastin-like Peptides and Other Proteins ..................................................197
  7.2.3 Coacervation is Thought to be an Inverse Temperature Transition .........................................200
7.3 Investigations of the Molecular Basis for Coacervation Using Molecular Simulation ..................200
7.4 Motivation and Rationale .............................................202
7.5 Methods ......................................................................203
7.6 Results and Discussion ..................................................204
  7.6.1 The Octapeptide Becomes More “Open” With Increasing Temperature ..................................204
  7.6.2 The Radius of Gyration of (GVPGV)7 and the Importance of Removing the “Equilibration Phase” in Generalized-Ensemble Simulations ..................................................210
  7.6.3 The Radius of Gyration of (GVPGV)7, Intra-Chain Distance Scaling, and Solvent Quality ........212
  7.6.4 The Effect of Temperature on Nonpolar Interactions in (GVPGV)7 ........................................ 214
  7.6.5 The Effect of Temperature on Hydrogen-Bonded Interactions in (GVPGV)7 ..............................216
  7.6.6 The Effect of Peptide Length on Structural Properties: Comparing GVGVPVG and (GVPGV)7 ..................................................................................................................218
7.7 Conclusions ...................................................................222
References......................................................................................................................................224

CHAPTER 8  Protein Non-Folding: An Elastin-like Aggregate Resembles a Polymer Melt........227

8.1 Summary................................................................................................................................228
8.2 A Seventy Year Old Controversy: What is the Structure of an Elastin Aggregate? ........228
  8.2.1 The Random Network Model ..........................................................................................228
  8.2.2 The Two Phase Model......................................................................................................230
8.3 The Flory Theorem: Ideal Chains in a Polymer Melt.........................................................231
8.4 Methods ..................................................................................................................................232
8.5 Results and Discussion ..........................................................................................................232
  8.5.1 ELP Chains in the Aggregate are Highly Entangled ......................................................232
  8.5.2 “Disentangling” ELP Chains in the Aggregate .................................................................233
  8.5.3 To What Extent Do ELP Chains Within the Aggregate Conform to the Flory
       Theorem? ..............................................................................................................................234
  8.5.4 Backbone Hydration Precludes the Formation of An Ideal Polymer Melt..................236
  8.5.5 Comparison of the Structural Properties of ELP Chains in Water and in the
       Aggregate ............................................................................................................................237
8.6 Conclusion: Towards a Unified Model for the Structure of an Elastin-like Aggregate ......241
References......................................................................................................................................243

CHAPTER 9  Significant Contributions and Future Directions .................................................246

9.1 Summary of the Significance and Contributions of the Thesis........................................247
9.2 Future Avenues of Research .................................................................................................249
References......................................................................................................................................254

APPENDIX A  Analysis of Proline and Glycine Spacing in Human Elastin.............................255
A.1 Proline and Glycine Spacing in Human Elastin ..................................................................256
References.............................................................................................................................259

APPENDIX B Supplementary Information for Chapter 5: Simulated Tempering Distributed Replica Sampling, Virtual Replica Exchange, and Other Generalized-Ensemble Methods for Conformational Sampling ...........................................................................................................260

B.1 Supporting Tables and Figures for Chapter 5 ................................................................261
B.2 An Example Calculation of the Distributed Replica Potential Energy (DRPE) ..........265
B.3 Computing Error in Exchange Probabilities for ST and SREM ..................................269
References.............................................................................................................................271

APPENDIX C Structural Properties of Disordered States of Proteins: Definitions and Criteria ...272

C.1 Structural Properties of Disordered States of Proteins .................................................273
C.2 Radial Distribution Functions .........................................................................................274
C.3 Hydrogen Bonds ...........................................................................................................275
  C.3.1 Peptide-Peptide Hydrogen Bonds ...........................................................................275
  C.3.2 Analysis of the Energy of Peptide-Peptide Hydrogen Bonds ..................................279
  C.3.3 Peptide-Water Hydrogen Bonds ............................................................................283
  C.3.4 Water-Water Hydrogen Bonds ...............................................................................288
C.4 Other Chain-Chain Interactions ......................................................................................292
  C.4.1 Contacts Between Residues Based on α-carbons (Cα-contacts) ............................292
  C.4.2 Contacts Between Residues Based on Contacts Between Heavy Atoms (HA-contacts) 293
  C.4.3 Contacts Between Residues Based on Contacts Between Nonpolar Carbons (NP-contacts) ........................................................................................................294
C.5 Other Chain-Water Interactions ......................................................................................295
  C.5.1 Hydration Shell Defined Using Heavy Atoms (HA hydration shell) .......................295
  C.5.2 Hydration Shell Defined Using Nonpolar Carbon Atoms (NP hydration shell) ....296
C.6 Radius of Gyration ........................................................................................................296
References.............................................................................................................................298
# List of Tables

Table 3.1  Elastomeric proteins: sequence, structure and function ......................................................68

Table 3.2  Elastomeric proteins: measurements of elastic mechanical properties ..........................69

Table 5.1  Practical advantages and disadvantages of generalized-ensemble algorithms ..............141

Table 5.2  Evaluating temperature diffusion ......................................................................................145

Table 6.1  Evaluating the diffusion of replicas in temperature space .................................................182

Table 8.1  Comparison of chain-chain and chain-water interactions in the monomer and the aggregate .................................................................................................................................238

Table B.1  List of temperatures ............................................................................................................261

Table C.1  Criteria for chain-chain, chain-water and water-water interactions ...............................273

Table C.2  Average energy of hydrogen bonds ..................................................................................290
### List of Figures

Figure 1.1  A disordered ensemble of configurations of an elastin-like peptide ........................................3
Figure 1.2  A conventional MD simulation of a disordered peptide .................................................................7
Figure 2.1  Sequence features of disordered proteins ........................................................................................27
Figure 3.1  Entropy-driven elastic recoil ........................................................................................................65
Figure 3.2  Measurements of elastic mechanical properties .................................................................................66
Figure 4.1  Sequence of model peptides ...........................................................................................................94
Figure 4.2  Representative conformations of monomers and aggregates .............................................................96
Figure 4.3  Average structural properties ...........................................................................................................99
Figure 4.4  Proline and glycine composition of elastomeric and amyloidogenic peptides ..........................101
Figure 4.5  Thioflavin-T binding assay for amyloid fibril formation .................................................................104
Figure 4.6  Role of entropy in polypeptide self-organization ............................................................................106
Figure 4.7  Assignment of local structure ........................................................................................................110
Figure 5.1  The conformational landscape and hydrogen-bonding contact map of GVGVPGVG .................................134
Figure 5.2  Assessing convergence of weight factors and PEDFs .................................................................139
Figure 5.3  Assessing the accuracy of the end-to-end distance (EED) distribution ........................................148
Figure 5.4  EED distributions at different temperatures ..................................................................................149
Figure 5.5  Assessing structural convergence using multiple criteria ..........................................................152
Figure 5.6  Correlation between structural convergence and temperature diffusion ....................................155
Figure 5.7  Comparing STDR and conventional MD for GVGVPGVG .......................................................158
Figure 5.8  Deviation from sampling homogeneity for simulations of (GVPGV)7 ........................................161
Figure 5.9  Comparing STDR and conventional MD for (GVPGV)$_7$ ..................................................163
Figure 6.1  Implementation of the STDR algorithm .................................................................180
Figure 6.2  Node time required for each step in STDR .................................................................181
Figure 6.3  A two-dimensional plot of the radius of gyration of the peptide vs. temperature .........183
Figure 6.4  Number of intramolecular peptide hydrogen bonds (HBs) observed with increasing simulation time for STDR (green and purple) and canonical MD (red) ........................................185
Figure 6.5  Configurations of the (GVPGV)$_7$ peptide from canonical MD and STDR simulations at 300 K ..................................................................................................................................186
Figure 6.6  Statistical convergence of structural properties ..........................................................188
Figure 6.7  Probability distribution of the radius of gyration, $R_g$ ................................................189
Figure 6.8  Estimated standard error obtained using “blocking” vs. the number of configurations in each block ........................................................................................................................................191
Figure 7.1  Phase diagrams for protein liquid-liquid phase separation (LLPS) .........................198
Figure 7.2  Structure of a VPGV $\beta$-turn ..................................................................................201
Figure 7.3  The effect of temperature on the end-to-end distance, $R_{EED}$, of the octapeptide .........205
Figure 7.4  The effect of temperature on the end-to-end distance, $R_{EED}$, of the octapeptide fragments within (GVPGV)$_7$ .................................................................................................................206
Figure 7.5  The effect of temperature on hydrogen bonds in the octapeptide (GVGVPVG) .....208
Figure 7.6  The effect of temperature on nonpolar interactions in the octapeptide ..................209
Figure 7.7  The effect of temperature on the radius of gyration, $R_g$, of (GVPGV)$_7$ ..............211
Figure 7.8  The effect of temperature on the average intrachain distance scaling ...................213
Figure 7.9  The effect of temperature on nonpolar interactions in (GVPGV)$_7$ .........................215
Figure 7.10 The effect of temperature on hydrogen bonds in (GVPGV)$_7$ ..............................217
Figure 7.11 Local and nonlocal interactions in (GVPGV)$_7$ ....................................................218
Figure 7.12  The effect of temperature on nonpolar interactions in (GVPGV)$_7$ and GVGVPGVG .................................................................220
Figure 7.13  The effect of temperature on hydrogen bonds in (GVPGV)$_7$ and GVGVPGVG ..........221
Figure 8.1  Representative configurations of the aggregate .........................................................233
Figure 8.2  The aggregate of (GVPGV)$_7$ chains “disentangled” .........................................234
Figure 8.3  Distribution of the radius of gyration, $R_g$, in water and in the aggregate ...............235
Figure 8.4  The effect of temperature and aggregation on average intrachain distance scaling ......236
Figure 8.5  An aggregate of elastin-like peptides retains significant backbone hydration ...........237
Figure 8.6  Hydrogen bonding and nonpolar contact maps ..........................................................240
Figure 9.1  A large aggregate of sixty four ELP chains .................................................................250
Figure 9.2  Configurational ensembles of the folded state and unfolded state of drkN SH3 .........253
Figure A.1  Inter-proline spacing in human elastin .................................................................256
Figure A.2  Inter-glycine spacing in human elastin .................................................................257
Figure B.1  Deviation from sampling homogeneity .................................................................262
Figure B.2  Structural convergence at multiple temperatures, bar graphs .................................263
Figure B.3  Structural convergence at multiple temperatures, 2D plot .....................................264
Figure C.1  The geometry of peptide-peptide hydrogen bonds ...............................................276
Figure C.2  RDF for carbonyl oxygen atoms and amide nitrogen atoms ......................................277
Figure C.3  RDF for carbonyl oxygen atoms and amide hydrogen atoms ..................................278
Figure C.4  Probability distribution for the hydrogen-nitrogen-oxygen angle, $\theta$ .....................279
Figure C.5  Atoms used in the calculation of the non-bonded energy of a hydrogen bond .........281
Figure C.6  2D energy distributions for peptide-peptide hydrogen bonds ....................................282
Figure C.7  Geometry of peptide-water hydrogen bonds ..........................................................283
Figure C.8  RDFs used to determine distance criteria for peptide-water hydrogen bonds ..........284
Figure C.9  Atoms used to compute the non-bonded energy of peptide-water hydrogen bonds.....285
Figure C.10  2D energy distributions for N-H --- water hydrogen bonds.................................286
Figure C.11  2D energy distributions for C=O --- water hydrogen bonds ...............................287
Figure C.12  RDFs for water oxygen atoms and water hydrogen atoms...............................288
Figure C.13  2D energy distributions for water --- water hydrogen bonds............................289
Figure C.14  Probability distributions of \( E_{\text{non-bonded}} \) for peptide-peptide, peptide-water and water-water hydrogen bonds ..........................................................................................................................290
Figure C.15  RDF for \( \alpha \)-carbon atoms ..................................................................................293
Figure C.16  RDF for heavy atoms .......................................................................................294
Figure C.17  RDF for nonpolar carbon atoms ........................................................................295
Figure C.18  RDFs used to define the hydration shell and nonpolar hydration shell ..........296
List of Appendices

APPENDIX A  Analysis of Proline and Glycine Spacing in Human Elastin.................................255

APPENDIX B  Supplementary Information for Chapter 5: Simulated Tempering Distributed
Replica Sampling, Virtual Replica Exchange, and Other Generalized-Ensemble Methods for
Conformational Sampling...........................................................................................................260

APPENDIX C  Structural Properties of Disordered States of Proteins: Definitions and Criteria ...272
## List of Acronyms and Symbols

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2D</td>
<td>two-dimensional</td>
</tr>
<tr>
<td>$\varepsilon_0$</td>
<td>permittivity of free space</td>
</tr>
<tr>
<td>Å</td>
<td>Ångstroms</td>
</tr>
<tr>
<td>A</td>
<td>alanine</td>
</tr>
<tr>
<td>Aβ</td>
<td>amyloid β peptide</td>
</tr>
<tr>
<td>CD</td>
<td>circular dichroism</td>
</tr>
<tr>
<td>CPU</td>
<td>central processing unit</td>
</tr>
<tr>
<td>DisProt</td>
<td>database of protein disorder</td>
</tr>
<tr>
<td>DNA</td>
<td>deoxyribonucleic acid</td>
</tr>
<tr>
<td>DSSP</td>
<td>dictionary of protein secondary structure</td>
</tr>
<tr>
<td>DR</td>
<td>distributed replica sampling</td>
</tr>
<tr>
<td>DRPE</td>
<td>distributed replica potential energy</td>
</tr>
<tr>
<td>ELP</td>
<td>elastin-like peptide</td>
</tr>
<tr>
<td>G</td>
<td>glycine</td>
</tr>
<tr>
<td>GROMACS</td>
<td>Groningen machine for chemical simulations</td>
</tr>
<tr>
<td>HA</td>
<td>heavy atom</td>
</tr>
<tr>
<td>HB</td>
<td>hydrogen bond</td>
</tr>
<tr>
<td>IDP</td>
<td>intrinsically disordered protein</td>
</tr>
<tr>
<td>IDR</td>
<td>intrinsically disordered region</td>
</tr>
<tr>
<td>J</td>
<td>Joules</td>
</tr>
<tr>
<td>K</td>
<td>Kelvin (temperature scale)</td>
</tr>
<tr>
<td>kcal</td>
<td>kilocalories</td>
</tr>
<tr>
<td>LLPS</td>
<td>liquid-liquid phase separation</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>MD</td>
<td>molecular dynamics</td>
</tr>
<tr>
<td>mol</td>
<td>mole (1 mole = Avogadro’s number of particles = 6.022 x 10^{23})</td>
</tr>
<tr>
<td>nm</td>
<td>nanometers</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>NMR</td>
<td>nuclear magnetic resonance</td>
</tr>
<tr>
<td>NOE</td>
<td>nuclear Overhauser effect</td>
</tr>
<tr>
<td>NP</td>
<td>nonpolar</td>
</tr>
<tr>
<td>OPLS</td>
<td>optimized potentials for liquid simulations</td>
</tr>
<tr>
<td>OPLS-AA/L</td>
<td>optimized potentials for liquid simulations – all atom / LMP2 calculations</td>
</tr>
<tr>
<td>P</td>
<td>proline</td>
</tr>
<tr>
<td>PDB</td>
<td>protein data bank</td>
</tr>
<tr>
<td>PPII</td>
<td>polyproline II</td>
</tr>
<tr>
<td>PRE</td>
<td>paramagnetic relaxation enhancement</td>
</tr>
<tr>
<td>RE</td>
<td>replica exchange</td>
</tr>
<tr>
<td>RMSD</td>
<td>root-mean-square deviation</td>
</tr>
<tr>
<td>SASA</td>
<td>solvent-accessible surface area</td>
</tr>
<tr>
<td>SAXS</td>
<td>small angle x-ray scattering</td>
</tr>
<tr>
<td>SH3</td>
<td>Src-homology-3</td>
</tr>
<tr>
<td>SREM</td>
<td>serial replica exchange method</td>
</tr>
<tr>
<td>STDR</td>
<td>simulated tempering distributed replica sampling</td>
</tr>
<tr>
<td>ST</td>
<td>simulated tempering</td>
</tr>
<tr>
<td>VREX</td>
<td>virtual replica exchange</td>
</tr>
<tr>
<td>T</td>
<td>temperature</td>
</tr>
<tr>
<td>TB</td>
<td>terabyte</td>
</tr>
<tr>
<td>TF</td>
<td>transcription factor</td>
</tr>
<tr>
<td>TIP3P</td>
<td>transferable intermolecular potential function, three point model</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
<tr>
<td>V</td>
<td>valine</td>
</tr>
<tr>
<td>WHAM</td>
<td>weighted histogram analysis method</td>
</tr>
</tbody>
</table>
CHAPTER 1

Preface and Overview of the Thesis
1.1 Protein Disorder

Disordered proteins are abundant in all kingdoms of life: more than one-third of eukaryotic proteins and more than 75% of cancer-associated proteins are predicted to contain disordered regions.\textsuperscript{1-3} Despite their biological importance, intrinsically disordered proteins (IDPs) are poorly understood relative to the wealth of structural data available for folded proteins in the Protein Data Bank (PDB). IDPs are notoriously difficult to study using experimental approaches.\textsuperscript{4,5} Their tendency to aggregate in solution presents significant challenges. However, more fundamentally, disordered proteins and disordered regions of proteins have many energetically-accessible conformational states. To illustrate this conformational heterogeneity with a specific example, we show a small subset of the configurations obtained in a simulation of a disordered peptide in Figure 1.1.

1.2 Molecular Simulations of Protein Disorder

By definition, disordered proteins and disordered regions of proteins populate many conformational states (thousands or millions of states? – at present, we do not have an adequate description of their conformational landscape to describe the phenomenon quantitatively). Site-specific measurements obtained using nuclear magnetic resonance (NMR) or fluorescence spectroscopy represent ensemble averages population-weighted over many states. Even with hundreds of experimental observables (for the best-characterized IDPs), it is not possible to obtain an unambiguously-determined ensemble of conformations. Thus, the structural characterization of disordered proteins is an inherently under-determined problem: a small number of restraints are insufficient to uniquely define the conformations of a system with thousands of degrees of freedom.\textsuperscript{6} In principle, molecular simulations, with their empirical force fields,\textsuperscript{7,8} can offer the additional information required to obtain conformational ensembles for disordered states of proteins.
Figure 1.1  A disordered ensemble of configurations of an elastin-like peptide
To illustrate the structural heterogeneity of an elastin-like peptide, (GVPGV)$_7$, we show representative structures from the conformational ensemble generated by molecular dynamics (MD) simulations with the simulated tempering distributed replica sampling (STDR) algorithm. Chains are shown in cartoon representation (yellow) with a surface representation of all protein atoms (in green). Refer to Chapters 6 and 7 for a detailed description of this simulation.

In Chapter 2, we review what is currently known about protein disorder. In particular, we describe the experimental and bioinformatics approaches used to study protein disorder. We also provide a survey of the computational studies of IDPs with an emphasis on the interdisciplinary nature of these studies. A comprehensive description of what is currently understood about protein disorder, as well as the experimental and theoretical approaches that have already been applied to characterize disordered states, provides relevant background for the present work.
1.3 Elastin and Other Rubber-like Elastomeric Proteins

Elastomeric proteins were the first class of proteins proposed to be disordered. In fact, it was suggested that elastomeric proteins are the only class of proteins with a biological function that necessitates disorder in the native state:

“The elastic structure proteins are the only group of proteins in which the molecular configuration is not uniquely prescribed in the native state. … In fulfillment of their functions, the chains of these proteins occur as random coils even in the native state.”
– Paul J. Flory (1969)

More than forty years later, we now know that disorder is abundant, and fulfills a myriad of biological functions. However, we still do not have a complete understanding of the structural properties of elastomeric proteins. Elastomeric proteins provide the high efficiency elastic recoil necessary to undergo reversible deformation to biological machinery as diverse as the mammalian arterial wall, the capture spiral of spider webs, and the hinge of scallop shells.

The class of rubber-like elastomeric proteins includes several types of spider silk, resilin, abductin, and ColP, which is a protein found in mussel byssus thread. Of particular interest is elastin, which, together with other structural proteins, forms the fabric of extensible tissues, including skin, blood vessels, and elastic ligaments, and provides the elasticity required for proper physiological function. Elastin and elastin-derived peptides self-aggregate upon heating to form an organized fibrillar structure in a process known as coacervation. Remarkable durability and intrinsic capacity for self-organization make elastin an ideal biomimetic model in the development of synthetic biomaterials. The sequence of tropoelastin, the monomeric precursor of elastin, is composed of alternating cross-linking and hydrophobic domains. The covalent cross-linking of elastin monomers imparts strength and stability to the polymeric matrix, while the hydrophobic domains are thought to confer the propensities for self-aggregation and extensibility. Elastin-like peptides modeled after the hydrophobic domains of elastin are the focus of this thesis.
In Chapter 3, we provide an overview of what is currently known about rubber-like elastomeric proteins with a particular focus on elastin. While the studies contained in the present work are concerned with elastin-like peptides, the structural insights derived from our simulations may also be applicable to other rubber-like proteins. For this reason, we provide the reader with a general introduction to this class of proteins. The aim of the introductory chapters of the thesis (Chapters 2 and 3) is to place the current study of elastin-like peptides within the context of the study of other intrinsically disordered proteins and elastomeric proteins. Future studies of other disordered proteins may benefit from some of the methodology developed in the present study of elastin-like peptides. Thus, Chapters 2 and 3 together constitute the necessary background for the thesis.

1.4 Two Types of Protein Self-Aggregation: Elastomeric and Amyloidogenic

In contrast to elastin fibres, which are essential in many extensible tissues, amyloid fibrillar deposits are associated with numerous tissue-degenerative pathologies, including Alzheimer’s and Parkinson’s diseases. Using electron microscopy and solid state NMR, the molecular structure underlying amyloid fibrils has been shown to consist of cross-β sheets extending along the main axis of the fibril. Both elastin-like and amyloid-like materials result from the self-assembly of proteins into fibrils. The goal of the study presented in Chapter 4 is to investigate the molecular basis for the differing physical properties of these two classes of protein self-aggregation.

Towards this goal, we performed molecular dynamics (MD) simulations of both monomers and aggregates of a model set of peptides based on sequence motifs derived from the hydrophobic domains of elastin. (Note that throughout the thesis, the term “monomer” is used to describe a single polypeptide chain, and not a monomeric unit of a polymer. We adopt the conventional use of the term “monomer” from the field of protein science, and not that of polymer physics.) The results of these simulations reveal that elastin-like peptides, unlike amyloid-like peptides, are unable to form extended β-sheet secondary structure. These results are consistent with the fact that, in order to function, elastin and other rubber-like elastomeric proteins must remain disordered even when aggregated. These proteins are “entropic chains”
with a function that necessitates disorder.\textsuperscript{14} Their high chain entropy underlies their propensity for elastic recoil. Importantly, we find that peptide self-organization into elastomeric or amyloid-like fibrils is modulated by sequence. Accordingly, the next step in this study was to perform an analysis of the sequence composition of proteins that self-assemble into elastomeric or amyloid-like aggregates. We find that all known rubber-like elastomeric proteins, including spider silk, elastin, mussel byssus thread, abductin, wheat gluten and resilin, exhibit a high combined content of proline and glycine. In particular, there is a threshold in the combined content of proline and glycine separating elastomeric and amyloidogenic sequences. Our findings support a unified model of protein aggregation in which hydration and conformational disorder are fundamental requirements for elastomeric function. Importantly, the conformational disorder of elastin-like peptides is a consequence of their intrinsic inability to form a compact, water-excluding core involving extensive backbone self-interactions.\textsuperscript{14}

### 1.5 Enhanced Conformational Sampling is Essential

The MD simulations described in Chapter 4 are relatively short by current standards (tens of nanoseconds) and do not make use of any enhanced sampling approach. As a result, the disordered peptides in these simulations explore only a very limited region of their energetically-accessible conformational space. To illustrate this point, Figure 1.2 B shows the time evolution the radius of gyration, $R_g$, for one of the elastin-like peptides in this study ($R_g$ is defined in section C.6 of Appendix C). The simulation begins with the peptide in a more extended configuration; after about 10 ns, the peptide reaches a collapsed state. The configurations obtained from 10 ns to 60 ns differ only slightly from each other (Figure 1.2 A). Following its initial collapse, the peptide undergoes very little conformational change during the remainder of the simulation. This observation of a collapse to a single conformation is typical of conventional MD simulations of disordered peptides. For this reason, a single MD simulation offers a very limited view of the conformational landscape of a disordered peptide: it provides only one out of the many conformations of the ensemble.
Figure 1.2  A Conventional MD simulation of a disordered peptide
(A) Sequential configurations from a time trajectory of a conventional MD simulation of the elastin-like peptide (PGV)$_{12}$ are shown. The colour scheme indicates the time sequence of the configurations (the sequence from red to white to blue corresponds to the time series from the beginning (0 ns) to the end (60 ns) of the simulation). (B) The time series of the radius of gyration, $R_g$, for the same simulation as (A) is shown. After the first 10 ns, the peptide has collapsed, and remains collapsed for the duration of the simulation. The details of this simulation are described in Chapter 4.

The results of Chapter 4 are based on statistically-meaningful differences in the average structural properties of elastin-like and amyloid-like peptides. The fact that average properties are sufficiently statistically converged to make this separation even though essentially only a single state is observed is itself an interesting result. It indicates that the conformations of these peptides resemble each other in terms of global properties, even though no two conformations are the same. However, in order to characterize the structural properties of a disordered state (that is, one that by definition can populate many conformational states), multiple
conformational states must be observed. We are therefore interested in obtaining ensembles of configurations that reflect the intrinsic structural heterogeneity of disordered states. In order to address the sampling limitations of conventional MD for disordered states of proteins, we turned to generalized-ensemble algorithms.

In Chapter 5, we present a novel generalized-ensemble algorithm, simulated tempering distributed replica sampling (STDR), which is a combination of simulated tempering and distributed replica sampling. STDR makes use of a random walk in temperature to surmount barriers in the energy landscape in order to sample many conformational states. It avoids the limitations of conventional MD illustrated in Figure 1.2, and provides heterogeneous ensembles of configurations (for example, the set of configurations shown in Figure 1.1). The focus of Chapter 6 is a detailed description of a large scale implementation of STDR, as well as a comparison to conventional MD simulations.

1.6 Massive Sampling Facilitated by the Rapid Evolution of Supercomputing

The work described in this thesis began seven years ago, in late 2004; our original aim was to investigate the molecular basis for the temperature-induced self-aggregation of elastin using atomistic MD simulations. As it turned out, this goal was far more ambitious than we realized at first. In fact, we required several orders of magnitude more computing power than was available in 2004 before we could even begin to address our original scientific question.

In this case, however, technology caught up to our needs. Because of the dramatic advances in both the speed and availability of supercomputing in the past few years, we were able to obtain the sampling required for statistical convergence of important structural properties. One of the major challenges in this work was developing new methods to manage both the computers (up to 10000 at a time) and the resulting flood of data (tens of terabytes). While the evolution in our computing capabilities is not emphasized in the thesis, it is an underlying thread, and certainly had a major impact on the scope of this work.
In general, what is considered “possible” in the field of biomolecular simulations is partly governed by the current state-of-the-art in supercomputing. That is, the scientific questions that one can address using biomolecular MD simulation are in large part determined by a combination of the speed and availability of computing resources. Accordingly, advances in supercomputing had a significant impact on the studies described in this thesis. The first study (Chapter 4) was conducted using our in-house 88 node computing cluster in 2005. These simulations required the use of twenty CPUs (central processing units) for more than six months to obtain 1 µs of total simulation time. At that time, these simulations were considered state-of-the-art. The study described in Chapter 5 required about fifty times as many CPU hours compared to Chapter 4; the studies described in Chapters 6 to 8 required about 3500 times as many CPU hours, and nearly 2000 times as much hard disk storage compared to Chapter 4. In total, the studies in Chapters 6 to 8 required nearly 15 million CPU hours of computing time, and the capacity to store and analyze millions of configurations (about 35 TB (terabytes) of hard disk storage). The comparison of these studies illustrates how rapidly the field of biomolecular simulation is advancing; while these numbers seem large today (2011), it is reasonable to expect that they will seem small in the near future.

In Chapters 7 and 8, we use the STDR algorithm coupled with massive sampling to elucidate the effect of temperature on the structural properties of a monomer and an aggregate of eight chains of a model elastin-like peptide. All prior studies of the effect of temperature on elastin-like peptides were limited in system size and timescale by the availability of supercomputing resources. They necessarily focused on monomeric systems, even though elastin self-assembly is by definition a process involving multiple polypeptide chains. The present work directly addresses the sampling limitations of prior studies, and, for the first time, a statistically meaningful investigation of the structure of an elastin-like aggregate is presented. Our results show temperature-induced changes in the structural propensities of the monomer. However, these effects cannot offer a complete explanation of the molecular basis for coacervation. In particular, in contrast to prior studies, we do not observe an increase in conformational order induced by increasing temperature. We propose that elastin self-aggregation is not an “inverse temperature transition” involving an increase in conformational order of the polypeptide chains.
Rather, it is a process driven by the overall increase in system entropy, with contributions from both the hydrophobic effect and the increase of chain entropy within the aggregate. In this new model for elastin self-aggregation, the aggregate has structural properties resembling that of a “polymer melt”, and, accordingly, one of the major driving forces for self-aggregation is the enhanced chain entropy within the “melt”.

1.7 Thesis Contributions and Organization

The main research contributions of this thesis are:

- the separation of elastin-like and amyloid-like peptides on the basis of their structural properties, and the elucidation of the sequence determinants of these properties,
- the development and testing of an enhanced sampling algorithm (simulated tempering distributed replica sampling, STDR) that efficiently explores the energy landscape of disordered peptides,
- the use of STDR in combination with massive sampling to obtain disordered ensembles of configurations with statistically-converged structural properties,
- a detailed investigation of the effect of temperature on the conformational ensembles of monomers of elastin-like peptides, and
- the elucidation of the molecular driving forces of elastin self-aggregation through a detailed comparison of the configurational ensembles of a monomer in water and a monomer within an aggregate.

The central aim of this dissertation is to elucidate the structural properties and self-aggregation of elastin-like peptides at the atomistic level. These topics are the subjects of Chapters 4, 7, and 8. Towards this goal, we utilize an enhanced sampling method that relies on a random walk in temperature to explore configurational space; the details of this method are described in Chapters 5 and 6. While the particular focus of this thesis is the protein elastin, the methods developed here are, in principle, applicable to study other disordered proteins. As background, a review of the application of molecular simulations to the study of protein disorder
(Chapter 2), and a discussion of the connection between rubber-like protein elasticity and structural disorder (Chapter 3) are provided. Finally, for the interested reader, we provide an appendix (Appendix C) that describes the structural properties used to characterize disordered states in Chapters 7 and 8. We also provide analysis of the proline and glycine spacing in human elastin (Appendix A), and an example of how the distributed replica potential energy is computed (Appendix B), which are supplementary material related to Chapters 4 and 5, respectively.
References


CHAPTER 2

Introduction to the Study of Protein Disorder: Bioinformatics, Experimental Approaches, and Molecular Simulations

The contents of this chapter were adapted with permission from a review article published in *Biochemistry and Cell Biology*.

*Reference:*

*Contributions:*
S.R. wrote the manuscript. R.P. provided editorial input and guidance.
2.1 Summary

Protein disorder is abundant in proteomes throughout all kingdoms of life and serves many biologically important roles. Disordered states of proteins are challenging to study experimentally due to their structural heterogeneity and tendency to aggregate. Computer simulations, which are not impeded by these properties, have recently emerged as a useful tool to characterize the conformational ensembles of intrinsically disordered proteins. In this chapter, we provide a survey of computational studies of protein disorder with an emphasis on the interdisciplinary nature of these studies. The application of simulation techniques to the study of disordered states is described in the context of experimental and bioinformatics approaches. Experimental data can be incorporated into simulations, and simulations can provide predictions for experiment. In this way, simulations have been integrated into the existing methodologies for the study of disordered state ensembles. We provide recent examples of simulations of disordered states, and emphasize important predictions and biophysical understanding made possible through the use of simulations. This chapter is intended as both an overview and a guide for structural biologists and theoretical biophysicists seeking accurate, atomic-level descriptions of disordered state ensembles.

2.2 Introduction

Proteins can simplistically be classified into two categories: those that fold (folders) and those that do not (non-folders). Historically, the structure-function paradigm has defined our understanding of proteins. It was thought that a protein’s function is encoded in the structure of its folded, and therefore biologically active, state.1 This view was supported by numerous experiments and theories, including Fischer’s lock-and-key model.2 Anfinsen’s classic experiment demonstrated that the information defining the secondary and tertiary structure of proteins resides in the amino acid sequence.3 A protein’s loss of function in the denatured state was associated with a loss of structure.4 Structural biologists therefore focused their studies on proteins with well-defined folded states, and proteins capable of functioning without a unique folded state were thought to be rare.1,5 The connection between sequence, structure, and function is essential to the understanding of enzymes and transport proteins. In contrast, the functions of many proteins involved in cell signalling, molecular recognition and transcriptional
regulation rely on the presence of unstructured regions. Most experimental techniques are suited for characterizing ordered structure, that is, for characterizing the folded state. However, the native state of many proteins \textit{in vivo} may not be folded, but may instead consist of an ensemble of interconverting structures, none of which correspond to a folded state. Intrinsically disordered proteins (IDPs) lack a unique, folded conformation under physiological conditions.\footnote{1} In the past decade, the structure-function paradigm has been revisited in light of the prevalence of IDPs.\footnote{9,10}

Disordered proteins are both abundant in nature and diverse in their functions,\footnote{11} underscoring the need for detailed studies. Bioinformatics approaches predict that disordered proteins are prevalent in all proteomes, especially in eukaryotes, for which long disordered regions (> 30 residues) are predicted to be present in 33 % of all proteins.\footnote{7} In mammals, 75 % of signalling proteins are predicted to contain long disordered regions.\footnote{6} The biological importance of disorder has driven the development of computational and experimental approaches to characterize IDPs, which can be significantly more challenging than studying proteins with well-defined folded states.\footnote{12} Biophysical studies of disordered states are complicated by structural heterogeneity, and consequently the need to describe a dynamic ensemble rather than a singular structure.\footnote{13} Using relatively few experimental measurements to determine a disordered ensemble for a system with many thousands of degrees of freedom is an inherently underdetermined problem. Molecular simulations, with their empirical potentials,\footnote{14} provide additional information that can be used to characterize ensembles of IDPs. Different levels of simulation detail range from coarse-grained models with implicit solvent to all-atom representations with explicit solvent.

In this introductory chapter, we begin by providing essential background on the nomenclature of protein disorder. We discuss the widespread biological relevance of disorder, including specific examples of functional roles for disorder and diseases associated with IDPs. To complement this background, we then summarize some of the main bioinformatics and structural techniques used to identify and characterize disorder. Bioinformatics has been instrumental in revealing the importance and prevalence of protein disorder with several
prediction algorithms. Disordered proteins can also be identified and characterized using an array of experimental approaches, nuclear magnetic resonance (NMR) being the most commonly used. Finally, the remainder of this chapter provides an overview of the application of molecular simulation to study disordered proteins. A significant fraction of studies using molecular simulations to characterize disordered ensembles have been hybrid experimental-theoretical approaches. We provide selected examples of these studies, as well as studies relying exclusively on simulation. Different conformational sampling approaches are described, along with their advantages and disadvantages. Throughout this chapter, we emphasize important predictions and biophysical understanding made possible through the use of simulations.

2.3 Defining Disorder

We first describe the nomenclature of the study of disordered proteins, including definitions of the terms intrinsically disordered proteins, intrinsically disordered regions, configuration, conformation, molten globule, pre-molten globule, random coil, and folded state. There is a lack of consensus in the field about specific terminology. For the purpose of this thesis, we will utilize the following definitions. Intrinsically disordered proteins (IDPs) are proteins that lack a well-defined, ordered structure. Segments that are unstructured both in isolation and in the context of the full length protein to which they belong are referred to as intrinsically disordered regions (IDRs). Proteins populating many dissimilar conformations may be thought of as the “liquid state” of proteins.

An important distinction has been proposed between configurations and conformations. The configuration of a peptide is defined by the three-dimensional coordinates of each of its atoms. Configuration space is, in principle, infinite and continuous. A conformation is the set of configurations which are related to each other based on a selected measure of structural similarity. Conformational space is therefore discrete, and is defined by the criterion used to distinguish between conformations. Similarity between configurations can be quantified using a variety of measures, including root-mean-square deviation (RMSD) of atomic positions, dihedral angles, secondary structure, or another relevant structural property. Based on similarity measures, configurations can be grouped into conformations using various clustering
algorithms.\textsuperscript{16} Clustering is therefore useful for rigorously defining conformational states.\textsuperscript{16-18} A thermodynamically-accessible conformation is a conformation which is populated under a particular set of conditions, such as temperature and concentration.\textsuperscript{16} In order to fully describe a disordered state ensemble, in principle, one should specify each conformation along with its population.\textsuperscript{12,19}

It has been suggested that globular proteins can access at least four unique states: native, molten globule, pre-molten globule, and unfolded.\textsuperscript{9} The \textit{folded state} corresponds to a conformation that is both highly populated and structured.\textsuperscript{20} It may also be referred to as the \textit{native state} if it is the biologically relevant and experimentally-observable conformation.\textsuperscript{16} However, a fundamental distinction must be made between folded and native states. All folded proteins exist in equilibrium with their unfolded states.\textsuperscript{21} Relative to the folded state, the \textit{molten globule} exhibits an increase in hydrodynamic volume of no more than 50 \%.\textsuperscript{9} Its secondary structure and folding pattern are native-like,\textsuperscript{9} with loosened or ‘molten’ tertiary interactions.\textsuperscript{1,22} The \textit{pre-molten globule} state is proposed to have no well-defined tertiary structure, approximately 50 \% native-like secondary structure, and a hydrodynamic volume about three times larger than that of the folded state.\textsuperscript{9}

The central tenet of the Protein Trinity, consisting of the folded state, the molten globule and the random coil;\textsuperscript{23} and of the Protein Quartet, which includes the pre-molten globule as a unique thermodynamic state,\textsuperscript{9} is that any of these states may be the native state, that is, the state relevant to a protein’s biological function. Transitions between any of these states may be functionally important.\textsuperscript{9,23} There is also an important distinction between the unfolded state and the denatured state, which is a chemically-stabilized or temperature-induced unfolded state.\textsuperscript{24} The unfolded state is itself a disordered ensemble of conformations. Finally, in the \textit{random coil} state, the backbone conformation of each residue is independent of the backbone torsion angles, $\phi$ and $\psi$, of neighbouring residues.\textsuperscript{25} However, it should be noted that polypeptides cannot be described as true random coils since each ($\phi$, $\psi$) pair is not independent of the conformation of neighbouring residues, resulting in preferred and sterically-disfavoured conformations.\textsuperscript{25,26}
There is currently a lack of consensus in the field about specific definitions for disordered states.\textsuperscript{27,28} IDPs have been referred to as natively unfolded, indicating that the native state corresponds to an ensemble of unfolded structures,\textsuperscript{29} as well as rheomorphic, natively denatured, flexible, and intrinsically unstructured.\textsuperscript{27,28} Further, “intrinsically disordered” has also been used to describe any state that is incompletely folded, and “intrinsically unstructured” to describe random-coil-like or pre-molten globule-like states.\textsuperscript{27} It has also been proposed that IDPs should be subdivided into two classes: intrinsic coils and pre-molten globules.\textsuperscript{9} This classification was motivated by hydrodynamic and far-UV circular dichroism (CD) measurements, which suggested that intrinsic coils are IDPs with larger hydrodynamic volume and less secondary structure compared to pre-molten globules.\textsuperscript{9} While it is useful to conceptualize in terms of these specific classifications, it is important to remember that there is in fact a continuum of levels of disorder, with varying degrees of compactness and varying amounts of secondary and tertiary structure.\textsuperscript{30-32}

Finally, it is necessary to distinguish proteins that are capable of folding under certain conditions (folders) from those that do not have a folded or ordered state under any known conditions (non-folders). We classify IDPs that are competent to fold upon binding a partner as folders even though their unbound state is disordered. Many IDPs/IDRs fold upon binding to their targets.\textsuperscript{33} Molecular recognition is analogous to protein folding, since both processes involve a thermodynamically-stable folded state and an unfolded state of higher conformational entropy.\textsuperscript{34} Thus, the cellular activities of disordered proteins can be separated into two main functional classes: recognition and entropic chains.\textsuperscript{8,35} Recognition generally involves IDPs (folders) that become ordered or partly ordered upon binding to their target. By contrast, entropic chains are necessarily non-folders since their functions rely on their high conformational entropy. They derive their function by populating many accessible conformations, with no unique folded structure. In this way, the structure-function paradigm can be reformulated to include both folded proteins and disordered proteins that function with an ensemble of conformations.
2.4 Biological Relevance of Protein Disorder

To motivate the importance of studying IDPs/IDRs, we provide a few select examples of their biological relevance, noting that there are several excellent reviews on the topic.\(^1,6,8,27,31,33,35\)

There are relatively few well-characterized examples of disordered proteins, and bioinformatics has been useful in demonstrating that these few examples are part of a large set of IDPs found in genomes throughout all kingdoms of life.\(^27\) Disordered sequences are predicted to be more prevalent in the genomes of eukaryotes compared to prokaryotes,\(^6\) likely due to the increased complexity of eukaryotic signalling and regulatory networks.\(^36\) The biological roles of disordered proteins are in general complementary to those of proteins with well-defined three-dimensional structures. By analysing the functional annotation of sequences predicted to be disordered, it was shown that sequences enriched in disorder have roles in regulation, transcription, and development. In contrast, sequences predicted to be depleted in disorder are more likely to have annotations for transport, catalytic activity and membrane localization.\(^37\) In some cases, disordered proteins are able to fill the same roles as folded proteins while offering additional functionally important attributes. For example, both folded proteins and IDPs can participate in protein-protein interactions. However, IDPs provide intermolecular interface areas that are generally much larger per residue.\(^38\) In order for a folded protein to present an interface of a similar size and remain stable as a monomer, it would have to be 2-3 times larger than an IDP. An increase in protein size would necessarily entail a higher energy cost to produce a folded protein fulfilling the same function. Since there are many protein-protein interactions utilizing IDPs, either molecular crowding or cell size would increase without protein disorder. IDPs therefore provide the necessary interaction interfaces while minimizing both protein and cell size.\(^38\)

In order to recognize their binding partners, IDPs/IDRs often undergo induced folding. Possible binding partners include other proteins, nucleic acids, membranes, and small molecules.\(^39\) Transcription factors (TFs) take part in both protein-DNA and protein-protein interactions. The function of TFs is to recognize specific DNA sequences and to recruit the proteins necessary for transcription. More than 80 % of TFs are predicted to contain extended disordered regions.\(^40\) Disorder-to-order transitions are important to both the DNA-binding
domains and transcriptional activation domains of eukaryotic TFs. There are several examples of DNA-binding domains with IDRs that undergo local folding transitions upon binding to DNA, resulting in sequence-specific recognition. It is also common for transcriptional activation domains to be unstructured, with induced folding only in the presence of their targets. Similarly, proteins involved in RNA recognition, including ribosomal proteins, become structured in the presence of their cognate RNA partners.

There are several advantages of binding involving a disordered partner versus the binding of two ordered partners. First, the absence of a unique structure facilitates the possibility of binding to multiple targets. Conformational flexibility may confer a functionally important binding promiscuity, with different conformations suited to different binding targets. This is referred to as one-to-many signalling. Conversely, many different disordered regions may bind to a common partner in many-to-one signalling. Proteins with multiple binding partners are hubs in protein interaction networks. Disordered proteins frequently function either as hubs or as interaction partners of hubs due to their intrinsic ability to participate in one-to-many and many-to-one signalling. Compared to proteins with a single interaction partner, eukaryotic hub proteins are significantly enriched in disorder.

Another advantage of binding a disordered partner is high specificity with modest affinity. The loss in conformational entropy due to the disorder-to-order transition is compensated for in part by favourable interactions at the binding interface. The result is an interaction with high complementarity and low binding affinity, allowing the complex to readily dissociate and terminate the signalling event. Protein-protein interactions are important drug targets, and disordered protein interactions may be easier to block with competitors compared to a similar interaction of two structured proteins due to their reduced affinity.

In addition, there may also be functionally important disorder in the bound state, in so-called fuzzy complexes. Sic1, an inhibitor of cyclin-dependent kinases involved in regulation of the yeast cell cycle, is an example of an IDP that is largely unstructured while performing its biological function. Its flexibility enables transient interactions in a dynamic complex, with
each binding site interacting in a dynamic equilibrium. Sic1 does not exhibit a disorder-to-order transition. Rather, its structure resembles a molten or pre-molten globule, with some secondary and tertiary structure.

At the other end of the disorder spectrum, entropic chains represent a subset of IDPs that lack well-defined structure under physiological conditions and do not undergo a disorder-to-order transition. Their functions are therefore completely outside the realm of folded proteins; elastin is only one such example. Flexible linkers or spacers are entropic chains that connect domains and allow them to move with respect to each other. In Shaker potassium channels, a flexible linker connects the channel with a domain that interacts with the open channel to inactivate it. This example demonstrates that disorder also plays a role in transport proteins. Another role of entropic chains is to function as “entropic brushes.” The thermal motion of polypeptide chains (acting as “bristles” of the brush) excludes volume, preventing the motion of large molecules, while allowing water and small solutes to pass through. One important example of the entropic brush mechanism is in the selectivity of the nuclear pore complex (NPC). Selective transport between the nucleus and cytoplasm occurs through the NPC. The size-dependent selectivity of the NPC is due to the phenylalanine-glycine nucleoporins (FG-nups), which are intrinsically disordered. FG-nups coating a nanopore of the same dimensions as the NPC are sufficient to reproduce the NPC’s selectivity. Their thermal motion creates an entropic barrier to the passage of large molecules. In Chapter 3, we describe how entropic chains, including elastin, resilin, and spider silk, provide rubber-like elasticity in biological systems.

Unstructured regions are often found in proteins that are targeted for degradation. Their turnover facilitates rapid cellular response to changing conditions. Proteolytic degradation is used to regulate concentrations of certain proteins in the cell. In general, the half-lives of disordered proteins are shorter than those of folded proteins, but can be extended by modifications like phosphorylation or binding to partners. The cell cycle must be sensitive to environmental conditions. It is thought that the rapid turnover of IDPs by the cell’s proteolytic machinery may play a role in this sensitivity and in the necessary short response times.
abundance and residence times of IDPs are tightly regulated in the cell and their overexpression or underexpression are associated with several diseases including cancer, Alzheimer’s disease, and Parkinson’s disease.\textsuperscript{52}

Disorder has been demonstrated to correlate with disease-related proteins (\textit{disorder in disorders} – the D\textsuperscript{2} concept)\textsuperscript{39}. Diseases in which disordered proteins are implicated include cancer, neurodegenerative diseases, cardiovascular disease, and diabetes.\textsuperscript{29,39} Disordered proteins therefore are an emerging class of drug targets.\textsuperscript{27} Proteins associated with cell signalling and cancer in humans are predicted to be enriched in protein disorder.\textsuperscript{36} In particular, 79 \% of cancer-associated proteins\textsuperscript{27} and 60 \% of proteins associated with cardiovascular disease are predicted to contain contiguous regions of disorder longer than 30 residues.\textsuperscript{29} A specific example of the link between disorder and cancer is the involvement of intrinsically disordered oncoproteins in cervical cancer.\textsuperscript{53} Human papilloma viruses (HPVs) have proteomes with only eight proteins, two of which are the oncoproteins E6 and E7. Different types of HPVs cause benign warts and carcinomas, including cervical cancer. HPV types associated with an increased risk of cancer have increased disorder in the E6 and E7 proteins.\textsuperscript{29,53}

IDPs are implicated in several amyloid diseases, including Alzheimer’s disease (amyloid β-protein, Aβ), Parkinson’s disease (α-synuclein), type II diabetes (amylin), and Huntington’s disease (polyglutamine repeats).\textsuperscript{54} Amyloid fibrils are protein aggregates characterized by a cross-β quaternary structure, with β-strands running perpendicular to the main axis of the fibril.\textsuperscript{55,56} The deposition of amyloid fibrils is associated with numerous tissue-degenerative pathologies,\textsuperscript{57} but amyloid fibrils are not necessarily toxic. There are many well-characterized examples of amyloids serving biologically-important functions.\textsuperscript{56} Small oligomeric species, which may be intermediates in amyloid formation, are thought to be responsible for toxicity and tissue damage.\textsuperscript{56} In general, the formation of amyloid requires a partially or fully unfolded precursor.\textsuperscript{54,58} The compact native state of folded proteins must be at least partially unfolded to form states that are competent for aggregation, with the main chain and hydrophobic groups exposed.\textsuperscript{58} However, IDPs do not need to undergo such conformational rearrangements to form aggregation-prone states.\textsuperscript{59} Aggregation of IDPs is also a matter of practical interest due to the
challenges of applying biophysical approaches to insoluble aggregates in vitro.\textsuperscript{59} Due to their tendency to self-aggregate, IDPs represent a significant portion of known amyloidogenic proteins.\textsuperscript{54} There are, however, several sequence features that protect IDPs from amyloid-like aggregation, including high net charge, low hydrophobicity, and the presence of proline residues.\textsuperscript{60} Elastomeric proteins represent a unique class of IDPs that must aggregate to function. In addition, in order to have the intrinsic disorder necessary for elastic recoil, they cannot aggregate to form highly ordered amyloid fibrils. Elastomeric proteins avoid the formation of amyloid with a high combined content of proline and glycine;\textsuperscript{61} this is one of the key findings of this thesis. In general, the amino acid content of disordered proteins is significantly different from that of globular proteins. In the next section, we summarize several sequence features that allow disordered sequences to be identified.

2.5 Sequence Features of Disordered Proteins: Bioinformatics Approaches

The widespread biological importance of disorder has prompted several groups to develop prediction techniques based exclusively on amino acid sequence. Just as the primary sequence encodes the folded state of proteins, it similarly defines the free energy landscape, including the unfolded state. For proteins with no folded state, the amino acid sequence precludes the possibility of folding. IDPs should therefore have some common sequence features that result in an inability to fold. The fact that predictors have been developed that can discriminate between ordered and disordered proteins based on sequence alone supports this idea.\textsuperscript{6,29,62} There are currently more than 50 disorder predictors, and the accuracy of disorder prediction continues to increase.\textsuperscript{27,28,63} Refer to He et al. for a recent review of disorder prediction.\textsuperscript{28}

There are two general approaches to disorder prediction. In the first approach, sequence features are identified in databases of disordered proteins using sequences of folded proteins as a control. One such collection of disordered sequences is the Database of Protein Disorder (DisProt), which includes hundreds of IDPs/IDRs identified as disordered using various experimental approaches.\textsuperscript{64} The majority of disorder prediction algorithms were developed using sequence characteristics identified from datasets of disordered proteins, including PONDR\textsuperscript{®} VL-
XT, GlobPlot, DISOPRED, TOP-IDP, and FoldIndex. The second approach includes algorithms that predict the ability of a sequence to fold without being ‘trained’ on databases of IDPs. Instead, predictors like IUPred, FoldUnfold, and UCon rely on statistical potentials developed using databases of folded proteins.

The IUPred disorder predictor computes a pairwise interaction energy based on amino acid composition as a test of a sequence’s ‘foldability’. It uses a statistical potential for the energetic contribution of each inter-residue interaction based on the observed frequency of the pair’s interaction in a database of globular protein structures. The parameterization of IUPred’s 20 x 20 interaction matrix does not rely on databases of disordered proteins. Using this predictor, IDPs/IDRs are distinguished from globular proteins on an energetic basis due to their reduced potential to form stable intra-chain contacts. A total pairwise energy below a certain threshold indicates that the sequence is capable of forming an ordered structure. It is important to note that random sequences with low predicted pairwise energy content do not necessarily fold. IUPred predicts coil-like disordered sequences to be energetically neutral, with nearly balanced attractive and repulsive interactions, while molten globule-like IDPs have a net energetic stabilization. IUPred was tested on a dataset of disordered sequences, yielding a true positive rate of 76.33 % (with a false positive rate of 5.33 %). The success of IUPred as a disorder predictor suggests that the compositions of disordered proteins result in fewer stabilizing interactions compared to globular proteins.

IDPs have biased amino acid compositions, with enrichment or depletion of certain amino acids relative to sequences of folded proteins. The compositional bias of IDPs/IDRs is illustrated in Figure 2.1 B. The average composition of disordered proteins in the current version of the DisProt database was computed relative to a set of folded proteins from the Protein Data Bank (PDB). This non-redundant set represents ordered proteins, which are known to be ordered by virtue of the fact that their structures have been solved using NMR or x-ray crystallography. The amino acids that are enriched in disordered sequences are: P, Q, S, E, G, K, D, R and A; inversely, there is a decrease in the content of T, N, M, H, V, F, L, Y, W, C and I. Overall, charged residues are enriched, while large hydrophobic and aromatic residues are
depleted in disordered sequences. This finding is consistent with a previous analysis of IDP/IDR amino acid compositions, in which it was found that W, C, F, I, Y, V, L, and N are depleted while A, R, G, Q, S, P, E, and K are enriched.\textsuperscript{74} It is also qualitatively consistent with the ranking of order-promoting residues in the TOP-IDP disorder predictor (W, F, Y, I, M, L, V, N, C, T, A, G, R, D, H, Q, K, S, E, P).\textsuperscript{68} A disorder predictor incorporating amino acid composition alone can recognize disordered sequences with an accuracy of 87\%.\textsuperscript{76} Such a high accuracy suggests that composition is a very important determinant of disorder. Accordingly, a high combined composition of proline and glycine is a necessary requirement for elastomeric proteins, which represent an important class of disordered proteins.\textsuperscript{61} As a consequence, conservation of sequence composition may be more important for disordered regions than conservation of exact sequence motifs.\textsuperscript{6} Sequence composition alone, however, is in general insufficient to determine conclusively if a sequence is disordered.\textsuperscript{77}

Due to the biased amino acid composition of disordered polypeptide chains, high net charge and low hydrophobicity are common sequence attributes of IDPs and IDR\texteds. Disordered proteins and globular proteins have been distinguished by correlating mean hydrophobicity with net charge in a charge-hydrophobicity (CH) plot\textsuperscript{78} (and similarly, a charge-hydropathy plot\textsuperscript{29}). A set of natively-unfolded proteins were found to have lower mean hydrophobicity and higher net charge compared to a set of small, monomeric globular proteins.\textsuperscript{78} Disordered and ordered sequences occupy different regions of the CH plot due to a combination of two effects. First, increased charge repulsion compared to globular proteins favours unfolding. Second, there are fewer hydrophobic groups, and therefore less driving force for hydrophobic collapse.\textsuperscript{78} CH plots are used by several disorder predictors, including FoldIndex.\textsuperscript{69}

Since the separation of disordered and ordered sequences in the CH plot was first made, there have been many more disordered sequences confirmed experimentally. In Figure 2.1 A, we provide an up-to-date CH plot using the database of disordered proteins (DisProt)\textsuperscript{64} in its present state (version 4.9). Hydrophobicity versus net charge is plotted for disordered regions from DisProt, and a set of ordered proteins from the PDB. The average hydrophobicity is lower and the average net charge is higher for disordered sequences compared to ordered sequences.
Figure 2.1 Sequence features of disordered proteins

A set of proteins from the PDB with less than 20 % sequence identity was obtained using the PISCES server. The set of disordered regions was obtained from the current release of DisProt (version 4.9). Only regions with more than five contiguous residues were included in the data set. Amino acid composition was computed individually for each sequence and averaged over the data set. In (A), net charge is plotted versus hydrophobicity. Hydrophobicity was computed using the normalized scale of Sweet and Eisenberg. Net charge was computed by taking aspartic acid and glutamic acid to have a charge of -1 and lysine and arginine to have a charge of +1, then normalizing by the chain length. The average hydrophobicity and average net charge are statistically different for the two data sets (based on a t-test with P-values < 0.0001). In (B), the average composition of DisProt minus the average composition (C) of the PDB, normalized by the average composition of the PDB [(C_{DISPROT} - C_{PDB})/ C_{PDB}] is shown. The amino acids that are enriched in DisProt relative to the PDB are P, Q, S, E, G, K, D, R and A, while T, N, M, H, V, F, L, Y, W, C and I are depleted (statistically significant with a P-value < 0.0001).
However, it is not possible to unambiguously classify a sequence as disordered based exclusively on its position in a CH plot. Our analysis is consistent with a recent study showing that ordered and disordered sequences cannot be clearly separated based solely on net charge and mean hydrophobicity. In fact, the CH plot is only a two-dimensional projection of the twenty-dimensional amino acid composition space in which disordered and ordered sequences overlap significantly.\textsuperscript{77}

The so-called predictors of natural disordered regions (PONDR), including PONDR VL-XT, incorporate sequence composition, hydropathy, net charge, and complexity into their prediction algorithms.\textsuperscript{65} Low sequence complexity is a common attribute of IDPs/IDRs, whose primary structure is significantly less complex than the sequences of natively folded proteins. Since IDPs do not need to maintain a specific folded structure to function, low complexity may be a consequence of having fewer sequence constraints. However, it is important to note that there is no observed upper limit on the complexity of IDPs, since a high sequence complexity does not guarantee that a protein will fold.\textsuperscript{65} Inversely, low sequence complexity is not a sufficient condition for disorder, as demonstrated by low complexity sequences with repetitive ordered structures, such as fibrous proteins, collagen and coiled coils.\textsuperscript{65,81} In fact, nearly 20\% of proteins in the human genome contain multiple repeats of 30-40 residues. Many of these are ankyrin repeat proteins, which are stable and cooperatively folded.\textsuperscript{82}

A significant limitation to the prediction of protein disorder is the presence of ordered sequences misclassified in databases of disordered proteins.\textsuperscript{65} When databases of ordered and disordered proteins are used as training sets for disorder predictors, false positives in both data sets limit the accuracy of the resulting predictors.\textsuperscript{6} Noise in the disordered protein data set is partly due to inconsistencies in the definition of structural disorder. Different methods have different limitations with regard to identifying disorder.\textsuperscript{62} For example, CD may misclassify a structured loop as being disordered. Crystal packing in x-ray crystallography sometimes results in the ordering of segments that are disordered in solution. Poor signal dispersion in NMR spectroscopy is often used to classify a protein as disordered, which is not sufficient to distinguish between molten globule and coil states.\textsuperscript{62} Fortunately, noise in the disordered
protein databases does not affect the accuracy of predictors that are not trained on disordered sequences, like IUPred.\textsuperscript{62}

Currently, no disorder predictor is completely accurate on its own, but accuracy can be improved by combining multiple approaches.\textsuperscript{32,83,84} Disordered regions can be predicted using a metaserver (DISMETA) that simultaneously reports the results of multiple disorder predictors.\textsuperscript{85} Alternatively, the recently developed META-Disorder predictor directly combines multiple unrelated disorder predictors. It achieves greater accuracy than any of its constituent methods.\textsuperscript{84} One reason for the success of these approaches is that different disorder predictors have been developed to detect different aspects of disorder, depending on the set of proteins on which they are trained. For this reason, there is disagreement between predictors, with some predictors best suited to detect specific types of disorder (e.g. linkers, short IDRs, long IDRs, etc).\textsuperscript{84} Consequently, estimates of disorder in proteomes vary significantly. For example, more than 20% of eukaryotic proteins are predicted to have a majority of residues in disordered regions,\textsuperscript{86} while around 25% of all proteins are predicted to have some disorder.\textsuperscript{10} Given the huge number of IDPs and IDRs predicted by bioinformatics, an ongoing challenge is to confirm and characterize their structural ensembles using both experiment and simulation.

\section*{2.6 Identifying and Characterizing Disordered Proteins: Experimental Approaches}

Several experimental techniques can be used to identify protein disorder. NMR spectroscopy has so far proven to be the most effective method for obtaining site-specific structural information for both IDPs and IDRs.\textsuperscript{5,31} Studying disordered states using NMR is challenging because the chemical environments of atomic nuclei in a disordered ensemble are similar. The resulting lack of resonance dispersion makes resonance assignment problematic, but this limitation can be overcome using high magnetic field spectrometers and isotope labeling.\textsuperscript{22} Heteronuclear multidimensional NMR is useful for studying disordered states.\textsuperscript{9} Once resonances are assigned, NMR provides both structural and dynamic information. For example, NMR can provide information on residual secondary structure in disordered ensembles, which is usually transient and confined to short segments.\textsuperscript{5} Secondary chemical shifts provide site-specific information quantifying fractional secondary structure propensity.\textsuperscript{87} Paramagnetic relaxation
enhancement (PRE) is useful for observing long range contacts. PRE measurements provide ensemble-averaged measurements of distances using spin labels. The distance range of the PRE method (∼20 Å) exceeds that of nuclear Overhauser effect (NOE) distance measurements (∼5 Å) and is therefore well-suited to detect the long range, transient contacts commonly found in disordered ensembles. Pulsed-field gradient diffusion methods are used to measure the diffusion coefficient of a disordered state, which is a population-weighted ensemble average. The diffusion coefficient provides a measure of hydrodynamic radius by approximating the protein as a sphere. Excellent reviews of the application of NMR spectroscopy to disordered and unfolded states are available.

Other spectroscopic methods, such as CD, can be used to identify and characterize protein disorder. IDPs have far-UV CD spectra with characteristic shapes. They can be identified by the presence of a minimum around 200 nm and an ellipticity near zero around 222 nm. In the near-UV region (250-350 nm), the spectra of IDPs are simplified compared to spectra of folded proteins due to their reduced secondary and tertiary structure content. Binding-induced folding of IDPs can be observed using fluorescence spectroscopy. Single molecule Förster resonance energy transfer (FRET) can also be used to obtain a distribution of distances between two residues with fluorophores (accessible distances in the range of 10 to 80 Å). These distance distributions are easily incorporated as restraints in the determination of conformational ensembles, or compared to distances from simulations. Similarly, small angle x-ray scattering (SAXS) provides information about pairwise distance and a measure of radius of gyration.

Missing or poorly-defined electron density in a structure determined by x-ray crystallography is often interpreted as evidence of a disordered region. Flexibility leads to incoherent x-ray scattering, resulting in missing electron density in a crystal structure. Obtaining crystals of proteins with some conformational heterogeneity may be problematic. If crystals do form and a structure is obtained, it is only representative of a single conformation, not necessarily of the ensemble in solution. For these reasons, x-ray crystallography is less suitable than NMR spectroscopy for the study of disordered states. Other techniques used to
identify disordered regions include Fourier transform infrared spectroscopy (FTIR), measurements of susceptibility to proteolytic degradation, electrospray ionization mass-spectrometry, and amide proton-deuterium exchange methods.\textsuperscript{31,63,91} Currently, no single experimental technique suffices for the assignment of residues as ordered or disordered. Ideally, several methods are used in combination. However, since different experimental techniques are suited to detect different aspects and types of conformational disorder, they occasionally disagree.\textsuperscript{84} Furthermore, it is still not completely clear whether proteins that are identified as disordered \emph{in vitro} remain unstructured \emph{in vivo},\textsuperscript{11} as some IDPs may become more structured in the crowded environment of the cell.\textsuperscript{33}

Both bioinformatics and experimental approaches have proven useful for the identification of IDPs/IDRs. Once a protein is identified as disordered, the next step is the characterization of its conformational ensemble. Relatively few IDPs/IDRs have been structurally characterized in detail. Biophysical studies of structurally heterogeneous states are challenging and often involve the use of multiple approaches, such as NMR, CD and SAXS.\textsuperscript{92} The goal of these studies is to obtain a comprehensive description of the ensemble, including residual secondary structure, mobility, and transient long range contacts.\textsuperscript{5} Experimental measurements of disordered states typically represent averages over a broad ensemble of unrelated structures. On their own, ensemble averages are insufficient to uniquely determine the underlying conformational distribution. Even with the most extensive set of experimental restraints currently attainable, disordered ensembles cannot be characterized with experiment alone.\textsuperscript{93} It is an inherently underdetermined problem because the number of degrees of freedom in the system far exceeds the number of available restraints.\textsuperscript{94} Experimental approaches can provide quantitative information for simulation in the form of restraints to determine conformational ensembles consistent with the data.\textsuperscript{30} In this way, sparse experimental data can be complemented by the information contained in molecular mechanics force fields.\textsuperscript{12}
2.7 Molecular Simulations

2.7.1 Overview of Simulation Approaches

Molecular simulations, with their empirical potentials,\textsuperscript{14} provide information that can be used to characterize ensembles of IDPs. Simulations rely on force fields that include terms to describe the interactions between particles (bonds, angles, dihedrals, van der Waals interactions, and electrostatics). Solvent can be represented implicitly or explicitly. Different levels of simulation detail are possible, from all-atom representations with explicit solvent to coarse-grained models, in which a residue is represented by a single pseudo-particle and the solvent is treated implicitly. There are also different approaches to conformational sampling, including molecular dynamics (MD), and Monte Carlo (MC) simulations. The sampling in MD involves the propagation of Newton’s equations of motion in time. As a result, MD simulations provide time trajectories containing information on the dynamics of all particles in the system. By contrast, in MC, sampling is stochastic, with changes in configuration accepted or rejected according to an energetic criterion. Since biomolecular systems are ergodic, ensemble averages can be calculated from simulation trajectories and compared to experimental measurements. Molecular simulations are therefore readily combined with other biophysical approaches in interdisciplinary studies. Further details about molecular simulation techniques can be found in useful reviews and books.\textsuperscript{14,95,96}

2.7.2 Hybrid Approaches: Using Experimental Restraints in Simulations

Molecular simulations are used to some extent by nearly all structural biologists in the process of structure determination for folded proteins. Structure refinement involves optimizing the agreement with both diffraction data, and \textit{a priori} knowledge about bond lengths, bond angles and dihedral angles.\textsuperscript{97} In the Crystallography and NMR System (CNS)\textsuperscript{98} and Xplor-NIH\textsuperscript{99} softwares, simulated annealing and energy minimization may be used at different stages of refinement in the determination of x-ray crystal structures or NMR structures. The use of simulated annealing significantly enhances the efficiency of refinement.\textsuperscript{97} Energy terms corresponding to observables, such as interproton distances from NOE measurements, may be used in combination with energy terms from empirical force fields, including van der Waals and dihedral angle potentials, in the process of structure determination.\textsuperscript{99} Even proteins with well-
defined folded states exhibit significant mobility, sampling many conformations at room temperature.\textsuperscript{100} Simulated annealing MD simulations using restraints from x-ray studies were performed to characterize the extent of conformational disorder in ribonuclease A.\textsuperscript{100} Likewise, disordered states of IDPs/IDRs can be studied using analogous experimental-simulation approaches. However, studying disordered ensembles is significantly more complicated because there are generally fewer experimental restraints available. Instead, experimentally-derived restraints can be directly combined with molecular simulations to produce disordered state ensembles that are consistent with input data.\textsuperscript{12}

Several hybrid experimental-simulation approaches have been developed to impose restraints on ensembles of conformations. One of the most straightforward approaches involves the generation of a set of random structures that are population-weighted to create an ensemble consistent with a set of restraints.\textsuperscript{19,93,101} An example of such an approach is the ENSEMBLE algorithm, which makes use of a wide variety of experimental restraints such as chemical shifts, NOEs, PREs, residual dipolar couplings, hydrogen exchange protection factors, solvent-accessible surface area and hydrodynamic radius.\textsuperscript{93,101} ENSEMBLE generates the simplest ensembles of conformations that are consistent with the input data in order to minimize the possibility of overfitting.\textsuperscript{93} An initial pool of sterically-plausible unfolded conformers can be generated using the program TraDES.\textsuperscript{102} In principle, any method of conformational sampling can be used to generate a conformer pool, including MD simulations.\textsuperscript{101} Conformers are generated, selected and population-weighted to match experimental restraints in an iterative process.\textsuperscript{30,93,103} ENSEMBLE has been used to characterize the highly populated unfolded state of the N-terminal Src-homology-3 (SH3) domain of \textit{Drosophila} drk, which exists in equilibrium with the folded state.\textsuperscript{93,103} Its disordered ensemble was shown to be highly heterogeneous, with significant non-native secondary and tertiary structure.\textsuperscript{93}

An analogous approach is the ensemble optimization method (EOM).\textsuperscript{104} In this approach, a conformer pool is generated with the flexible-meccano algorithm, which randomly selects dihedral angles for each residue from a library of coil conformations.\textsuperscript{105} An x-ray scattering curve is then calculated for each conformer, and sets of conformers that fit the
experimental SAXS distribution are selected. Different EOM runs produce different conformer pools, all of which match the same experimental SAXS profile. This type of underfitting due to insufficient data is a common issue in the determination of disordered ensembles. Completely different conformational ensembles can be consistent with a given set of experimental data. Conformational ensembles produced by EOM can be further refined and validated using information from complementary experimental approaches, such as FRET, CD and NMR.

Chemical shift information from NMR experiments can be either directly incorporated as restraints, or used as a criterion to evaluate the accuracy of a conformational ensemble. For instance, the partially unfolded signalling state of photoactive yellow protein was modeled using MD simulations with restraints based on NMR chemical shifts. This approach is effective when the partially unfolded state represents a relatively small perturbation from the folded state, and the structure of the folded state is known. Another approach that makes use of chemical shift information is the energy-minima mapping and weighting (EMW) method. The EMW method consists of three steps: conformational sampling, ensemble generation, and validation. This method is motivated by the idea that a disordered ensemble consists of a set of energetically favourable conformations. To ensure that a structurally heterogeneous set of initial conformations is generated, EMW uses a combination of high temperature MD and end-to-end distance restraints, followed by simulation at low temperature and energy minimization. It is assumed that configurations generated using this approach represent populated conformations. Using the library of initial conformations, EMW generates many candidate ensembles, each of which is independently consistent with a set of \(^{13}\text{C} \alpha\) chemical shifts. Validation of the ensembles is performed using \(^{13}\text{CO}, ^{15}\text{N}, \text{and } ^{1}\text{H} \alpha\) chemical shifts and scalar J-couplings. The EMW method has been used to study the conformational ensemble of a C-terminal fragment of p21, an IDP capable of recognising and binding to at least 25 different substrates. Its structural plasticity allows it to adopt both \(\alpha\)-helical and extended structures. Interestingly, the bound conformations of p21 were found to exist in the conformational ensemble of its unbound state, supporting a conformational selection mechanism for p21’s binding promiscuity.
Experimentally-derived restraints can be imposed in a simulation using time-averaging. In this approach, the ensemble of configurations from the complete simulation trajectory satisfies all restraints, while no single configuration is required to do so.\textsuperscript{109} Time-averaged restraints minimize the disturbance to the force field due to the addition of an artificial term. Inter-proton distances from rotating frame nuclear Overhauser effect intensities and dihedral angles from scalar couplings were used as time-averaged restraints in MD simulations of the XAO peptide (Ac-XX(A)-OO-NH\textsubscript{2}, where X is diaminobutyric acid and O is ornithine).\textsuperscript{110} Using this methodology, the structure of the XAO peptide was found to be an interconverting ensemble, with no extended polyproline II (PPII) structure. Although individual residues of the XAO peptide sample the PPII conformation, the PPII helix is not the unique conformation of the XAO peptide. Importantly, the properties of the ensemble were verified by agreement with the radius of gyration from SAXS measurements.\textsuperscript{110} In another example of the time-dependent use of restraints, $^3$J coupling constants were adaptively restrained by keeping track of configurations sampled throughout the simulation.\textsuperscript{111}

Molecular simulations can either use time-averaged restraints, or apply restraints simultaneously to multiple simulations (replicas).\textsuperscript{88,109,112} The Monte Carlo Replica Sampling (MCRES) method enforces PRE-derived distance restraints on an ensemble of replicas simulated in parallel.\textsuperscript{88} Restraints are imposed using an energetic penalty on the ensemble-average of distances rather than on any individual replica. The MCRES method was used to study the denatured state of bovine acyl-coenzyme A binding protein (ACBP) using a C\alpha-representation of the protein.\textsuperscript{88} For this system, twenty replicas were required to capture the broad ensemble of structures in the denatured state. This study of ACBP was subsequently extended by using an all-atom representation of the protein with implicit solvent, thereby incorporating structural information from a molecular mechanics force field to supplement the PRE-derived distance restraints.\textsuperscript{113} A similar approach was used to model the disordered ensemble of $\alpha$-synuclein.\textsuperscript{13} Using all-atom MD simulations with PRE-derived distance restraints, the unfolded ensemble of $\alpha$-synuclein was found to contain no significant secondary structure and to be more compact than a random coil. The long range contacts between the hydrophobic region and the C-terminus favour the collapsed state of $\alpha$-synuclein and may offer some protection against
aggregation. Importantly, the hydrodynamic radius obtained using the distance-restrained MD simulations was consistent with the experimentally-determined value, supporting the validity of the approach.\textsuperscript{13}

Applying experimentally derived restraints in MD simulations ensures that the simulations produce ensembles consistent with what is known experimentally. However, a few cautionary notes must be kept in mind when using experimental-theoretical hybrid approaches. First, the experimental methods used to obtain the restraints may perturb the ensembles they are characterizing. For example, in order to obtain PRE measurements of distances, nitroxide spin labels are coupled to cysteine residues in the protein. Both the introduction of spin labels and the mutation of residues to cysteine may result in perturbations in the conformations in the ensemble, as well as their relative populations.\textsuperscript{30} Using experimental restraints obtained under one set of thermodynamic conditions (e.g. temperature, pH, ionic strength and pressure) in an MD simulation with different conditions will produce an unphysical conformational ensemble.\textsuperscript{94} For this reason, the simulation system must capture the \textit{in vitro} conditions as closely as possible. The same considerations apply when comparing simulation data with experimental observables. It is also recommended that only primary (directly observable) experimental data be used as restraints, if possible. Secondary data that is computed using approximations and assumptions based on primary data, such as $S^2$ order parameters, may result in unexpected artifacts.\textsuperscript{94} In order to validate the ensembles generated by hybrid approaches, experimental data not used as restraints in the initial calculation can be used.\textsuperscript{12,30} Cross-validation is also possible by using a subset of the restraints to generate the ensemble, then verifying that the remaining restraints are reproduced.\textsuperscript{12} In principle, however, molecular simulations of disordered states of proteins do not require any experimental data as input, and therefore can be used independently to predict experimental measurements or to provide fundamental biophysical insight. Such studies are reviewed in the next section.
2.8 De Novo Simulations

2.8.1 Challenges and Considerations

Molecular simulations of IDPs, with or without experimental restraints, are complicated by the fact that not one but possibly many thousands of conformations must be sampled in order to characterize a structurally heterogeneous ensemble. Energetic barriers to conformational transitions involving significant structural rearrangements are often larger than the available thermal energy, and consequently transitions between conformations in the ensemble are statistically rare events. Continuous MD simulations are often insufficient to achieve complete Boltzmann sampling of all important conformational states. Without achieving statistical convergence, it is not possible to draw meaningful conclusions about the conformational equilibrium from the simulations. This is an important issue since computing capabilities limit the possible timescales of MD simulations.

Not only is it necessary for a simulation to sample all relevant conformations, the model used to represent the polypeptide must also be sufficiently accurate to capture the relevant physical properties of the system. In general, simulations utilizing more detailed molecular representations are more costly. In particular, simulations with atomistic detail and an explicit representation of the solvent are very computationally intensive. They are currently limited to the nanosecond to microsecond timescale for continuous MD simulations, depending on the size of the system and the available computational resources. For example, α-synuclein was recently studied using atomistic MD simulations as a monomer in solution and as an aggregate with a membrane (for a total of more than 200,000 atoms) for several nanoseconds.

2.8.2 Conformational Sampling Methods

Similar challenges are encountered in simulations of the unfolded state or partially-unfolded state of globular proteins due to essential similarities between disordered ensembles and unfolded or denatured ensembles. It is therefore instructive to consider the extensive work on simulations of unfolded states and denatured states of proteins whose native state is folded. One approach to achieving sufficient conformational sampling is to run thousands of independent MD simulations in parallel for tens of nanoseconds each using a distributed
In this “brute force” approach, many conformations of the unfolded ensemble are simulated simultaneously starting from the fully extended state, thereby overcoming the problem of conformational transitions. The main drawback is the significant computational expense of running many thousands of simulations. The unfolded ensemble of villin headpiece, the tryptophan zipper, and BBA5 were each studied using this approach with implicit solvent, for a total simulation time of nearly 800 µs. Collectively, these simulations resulted in the formulation of the “mean structure hypothesis”: the unfolded ensemble has the same mean structure as the folded state, only with a much higher structural diversity. This hypothesis is supported by the observation that the average distance matrix of Cα pairs of the unfolded ensemble is very similar to that of the folded state for all three proteins. The results of this study underscore the utility of studying average properties of disordered ensembles. Another approach to studying unfolded ensembles has been the use of biased molecular dynamics to preferentially sample conformations with different radii of gyration.

Starting from the native state of α-lactalbumin, a biasing force was used to generate conformers of larger radii of gyration, which were subsequently simulated using unbiased MD in order to model the partially-denatured, molten globule-like state.

The two most straightforward approaches to achieving adequate sampling in MD simulations involve running many short simulations (MS) or sampling a few long trajectories (FL). These two approaches were rigorously compared with explicit solvent simulations of the RN24 peptide (totaling more than 800 µs). Although structural properties obtained with both FL and MS simulations agreed qualitatively, MS simulations resulted in greater precision. Some of the short timescale conformational transitions observed in MS simulations were not observed in the FL simulations. Conversely, a few transitions occurring over long timescales were observed in the FL simulations, and not the MS simulations, leading the authors to conclude that the FL and MS approaches are complementary. However, these conclusions are system-dependent and the sampling efficiency in general depends on the ruggedness of the energy landscape. An approach intermediate between MS and FL is also possible (i.e. several simulations of intermediate length). For example, the intrinsically disordered transactivation domain of p53 was simulated in explicit solvent for six simulations totaling nearly 0.5 µs. The
resulting conformational ensemble was found to be dominated by compact states with significant amounts of secondary structure, largely due to the high composition of leucine.\textsuperscript{123} MS simulations can involve more complex methods, including Markov State Models (MSMs).\textsuperscript{124,125} An MSM is constructed from a large set of MD simulation trajectories. Configurations from the trajectories are clustered into discrete conformations in order to compute a transition probability matrix between all possible conformations. Relying on the assumption that transitions between conformations are Markov processes, the MSM approach effectively translates a large number of continuous MD trajectories into a transition matrix. An MSM provides both thermodynamic and kinetic information, such as the mean first passage time between the folded and unfolded state.\textsuperscript{124}

MS implementations may incorporate coupling or communication between simulations. Generalized-ensemble algorithms in which each replica executes a random walk in temperature are an example of this approach.\textsuperscript{114,126} Compared to constant temperature MC or MD simulations, generalized-ensemble algorithms improve sampling of the rugged energy landscapes underlying biomolecular motion. Their enhanced efficiency relies on the fact that the free energy surface becomes less rugged at high temperature, increasing the frequency of interconversion between conformational states.\textsuperscript{127} Replica exchange (RE) is the most commonly used generalized-ensemble method.\textsuperscript{126,128,129} An RE simulation consists of identical copies of the system (replicas) which sample canonical ensembles at different temperatures using MD or MC. Exchanges are performed between replicas at neighbouring temperatures according to a Metropolis criterion. In order for an exchange to occur, replicas at neighbouring temperatures must stop their respective simulations. This requirement for synchronization means that RE requires a large, dedicated and homogeneous computing cluster to function efficiently when applied to complex systems. RE has been successfully used to characterize the conformational ensemble of polyalanine, quantifying the relative populations of $\alpha$-helical, PPII, and disordered conformations as a function of temperature.\textsuperscript{130} We have reviewed and performed a thorough comparison of generalized-ensemble algorithms, specifically evaluating these methods for their performance in conformational sampling of IDPs,\textsuperscript{114} which is the focus of Chapter 5.
2.8.3 Simulations of Entropic Chains

There is currently little high resolution structural information available from experimental approaches for disordered ensembles of entropic chains, including elastin-like peptides. *De novo* MD simulations are therefore very useful for the study of these proteins. In this section, we review the efforts of other groups to apply molecular simulations to study entropic chains to place our simulations of elastin-like peptides within the context of the larger field.

The simplest entropic chains are poly-amino acids, such as polyglutamine and polyglycine. Pappu and co-workers have extensively characterized the conformational ensemble of polyglutamine, which is implicated in Huntington’s disease.\textsuperscript{131-134} Using MD simulations, polyglutamine was shown to adopt a heterogeneous ensemble of collapsed structures. Polyglutamine is disordered as a monomer because the glutamine side chains form hydrogen bonds to the backbone, essentially competing with water as a solvent.\textsuperscript{131} Individual residues have significant populations in the PPII and α-helical regions of the Ramachandran plot, but there is no extended PPII helix or α-helix. The ensemble of monomeric polyglutamine has a high configurational entropy that must be overcome in order for aggregation and β-sheet formation to occur. While statistical fluctuations lead to transient sampling of β-hairpin-like states,\textsuperscript{131} conformations with a high content of β-type structures are thermodynamically-disfavoured for polyglutamine monomers.\textsuperscript{133} The results from the MD simulations are consistent with the hydrodynamic properties of polyglutamine measured using fluorescence correlation spectroscopy.\textsuperscript{132} Chain size, as measured by hydrodynamic radius, is related to chain length by a power law. Polyglutamine in aqueous solution behaves like a polymer in a poor solvent (in which chain-chain interactions are preferred to chain-solvent interactions).\textsuperscript{132} This behaviour is not unique to polyglutamine. In fact, water is also a poor solvent for polyglycine and poly(glycine-serine). In 8M urea, both polyglycine and poly(glycine-serine) have conformational ensembles that are more swollen than the collapsed ensembles in water.\textsuperscript{135} Thus, despite the lack of a hydrophobic core, polar but uncharged sequences collapse in water, suggesting that water is a poor solvent for the polypeptide backbone.\textsuperscript{6,132,135} These observations have direct implications
for the aggregation of polyglutamine, since phase separation (aggregation) only occurs in a poor solvent, and increasing chain length increases the driving force for aggregation.\textsuperscript{136}

Another class of entropic chains are the nucleoporins, which are responsible for the selectivity of the nuclear pore complex. An FG-nucleoporin domain from yeast and a mutant with all ten phenylalanines substituted with alanines (the F\textrightarrow{}A mutant) were studied using all-atom MD simulations with implicit solvent.\textsuperscript{18,50} Results from these simulations indicated that the conformational ensemble of the FG domain is best described as a native pre-molten globule, while the F\textrightarrow{}A mutant is more coil-like. The compactness of the FG domain and the F\textrightarrow{}A mutant was characterized using diffusion coefficients measured with NMR. Both simulation and experiment characterized the ensemble of the FG domain as significantly more compact than that of the F\textrightarrow{}A mutant, providing a validation of the conformational ensemble generated using molecular simulation.\textsuperscript{50}

2.8.4 Simulation and Experiment: Interdisciplinary Studies

We have already seen several examples of powerful hybrid theoretical-experimental studies of IDPs that directly incorporate experimentally-derived restraints into molecular simulations. It is also common for IDPs to be studied using simulations without restraints in interdisciplinary studies. \textit{De novo} simulations can provide biophysical explanations for experimental observations, or can be validated using experimental approaches. For example, MD simulations can be used to complement structural information from NMR spectroscopy in the study of binding complexes of IDPs. This approach is particularly useful if the ligand remains partially disordered in the bound state. The interaction of a 15-residue peptide from mouse guanine-nucleotide exchange factor Sos2 and the SH3 domain of mouse Grb2 was characterized using NMR.\textsuperscript{137} A structure was obtained for the complex, but complete assignment of the peptide resonances was not possible. Consequently, it could not be concluded if the peptide adopts a PPII helix conformation, similar to that of the bound state of other SH3 ligands.\textsuperscript{137} NMR provided sufficient data to determine the orientation of the peptide and the location of only two residues. Starting from this limited data, MD simulations provided a conformational ensemble describing the bound state of the peptide. The peptide was found to
adopt a predominantly extended conformation in contact with the SH3 domain, stabilized by sidechain-sidechain and sidechain-backbone hydrogen bonds. The peptide retains significant flexibility in the bound state. The ideal PPII helix proposed in the NMR study as the structure of the peptide is not compatible with the conformational ensemble of the MD simulation.\textsuperscript{138}

In another interdisciplinary study of protein-protein interactions, MC simulations were used to characterize the formation of a complex between the p27\textsuperscript{Kip1} protein and cyclin A-cyclin-dependent kinase 2 (Cdk2).\textsuperscript{34} The p27\textsuperscript{Kip1} protein is intrinsically disordered in its unbound state and becomes ordered upon binding to the cyclin A-Cdk2 complex. Simulations were initiated from the crystal structure of the bound state in order to model the transition state ensemble, with the cyclin A-Cdk2 complex held fixed. A large ensemble of unfolding-unbinding trajectories were generated at high temperature. A significant amount of native-like topology was found in the transition state. It is, however, problematic to infer information about the binding free energy landscape at physiological temperature from simulations at high temperature. Subsequent simulations of the order-disorder transition of the p27\textsuperscript{Kip1} protein with the cyclin A-Cdk2 complex were performed using simulated annealing.\textsuperscript{139} When p27\textsuperscript{Kip1} is bound to the cyclin A-Cdk2 complex, its C-terminal region remains disordered. The conformational ensemble of the C-terminal region was characterized using both MD simulations and SAXS measurements to be extended and flexible in the bound state.\textsuperscript{140}

It is essential to validate theoretical approaches used to simulate IDPs using appropriate experimental data. One example of such a rigorous validation is a recent study in which the conformational ensembles of Aβ40 and Aβ42 were studied using MD simulations in explicit water on the microsecond timescale, with RE used to enhance conformational sampling.\textsuperscript{141} A comparison was made between different force fields for the Aβ42 peptide. The results of simulations using the OPLS-AA/L force field\textsuperscript{142} were found to be in quantitative agreement with scalar couplings determined from NMR experiments. This work validated the conformational ensembles generated using RE simulations in addition to providing a useful ranking of force fields based on a comparison to experimental data.\textsuperscript{141}
In another interdisciplinary study, different solvent representations were compared using a set of disordered peptides as test systems. This comparison was validated using experimental data from triplet-triplet energy transfer (TTET) experiments. TTET can be used to monitor loop formation in unfolded polypeptides. Since contact formation between residues occurs on a timescale accessible to MD simulation, a direct comparison of the kinetics obtained using simulation and experiment was performed. Using this approach, polyserine and poly(glycine-serine) peptides were found to have little persistent structure and non-Gaussian end-to-end distance distributions, indicative of rugged energy landscapes. The use of an explicit solvent (TIP4P) resulted in better agreement with TTET experiments compared to implicit solvent representations, which overestimated compactness, secondary structure content, and the number of peptide-peptide hydrogen bonds.

Another evaluation of the accuracy of an implicit representation of water was performed using NMR experiments for validation. RE simulations of the GB1 peptide were performed and chemical shifts and scalar couplings were calculated for the ensembles at each temperature. The ensemble of structures that most closely matched the NMR data obtained at 278 K corresponded to temperatures near 400 K. Like the TTET experiments, comparison to NMR data suggested that implicit solvent models result in overly-structured ensembles of IDPs at low temperature. While the use of implicit solvent provides advantages over explicit solvent in terms of computational efficiency, these two studies suggest that the conformational ensembles of IDPs in implicit solvent are not in agreement with experimental observations.

A promising new implicit solvent model, ABSINTH (self-assembly of biomolecules studied by an implicit, novel and tunable Hamiltonian), was recently developed for use in MC simulations. It is calibrated primarily for simulating the conformational ensembles of IDPs. ABSINTH is designed to maximize computational efficiency, requiring only 2.5 to 5 times the expense of simulations in vacuo. In order to better understand how sequence dictates a disordered conformational ensemble, high-throughput studies of many IDPs are required. Reaching statistical convergence for such ensembles is computationally intensive. ABSINTH is well-suited for such high-throughput simulation studies and is capable of providing a semi-
quantitative level of accuracy. Implicit solvent representations represent a significant simplification compared to explicit solvent. Similarly, simplified models of the polypeptide chain are possible in coarse-grained simulations.

2.9 Coarse-grained Simulations

Simulations of proteins may include different levels of detail, ranging from lattice models to all-atom representations. So-called ‘minimalist models’ offer several advantages, and have already proven useful in studies of protein folding. First, the computational cost is dramatically reduced, thereby facilitating simulations of significantly larger systems and/or longer timescales. Comparing simulations of coarse-grained models to experiment also offers the possibility of determining the specific properties of proteins that result in experimentally observed properties. Even if the model does not reproduce the experimental data, it is possible to learn from the result and refine the model. Validation of minimalist models with experiment is essential to determine if the elements of the model are sufficient. Simplified representations of polypeptide chains also offer the possibility of making more general observations about the behaviour of IDPs, rather than focusing on specific systems.

One of the simplest models used to represent a polypeptide chain is the worm-like chain (WLC) model. The polypeptide is modeled as a continuous cylinder with a randomly-directed radius of curvature using only two parameters, contour length, $l_c$, and persistence length, $l_p$. For a polypeptide, $l_c$ is the number of residues multiplied by the distance per residue (3.8 Å) and $l_p$ is the chain length beyond which the direction of the tangent vector to the cylinder is uncorrelated. Flexible linkers connecting independently folded protein domains have successfully been modeled as worm-like chains with an $l_p$ of 4.0 Å, while shorter loops were modeled with an $l_p$ of 3.04 Å. The use of polymer models to represent unfolded states and flexible linkers has been reviewed.

A Cα-model was recently developed specifically for the simulation of unfolded and disordered ensembles. The energy parameters were optimized through an iterative process incorporating PRE measurements of the unfolded ensemble of the Δ131Δ fragment of
Staphylococcal nuclease. An initial guess of the energy parameters was used to generate an unfolded ensemble from which PRE measurements were back-calculated. An iterative optimization scheme was used to obtain new energy parameters in better agreement with the experimental data. However, a limitation of this approach is that in order to obtain transferable energy parameters, PRE measurements from other disordered ensembles are required. Other types of experimental data can also be readily incorporated in this model.\textsuperscript{153}

Coarse-grained simulations have helped to elucidate the biophysical underpinnings of the sequence characteristics of disordered proteins uncovered by bioinformatics.\textsuperscript{77,154} For example, it is known that the compositions of IDPs/IDRs tend to be depleted in hydrophobic residues and enriched in charged residues.\textsuperscript{23} Based on a small dataset of natively-unfolded and folded proteins, it was suggested that a combined criterion of net charge and hydrophobicity defines a border between disordered and ordered sequences.\textsuperscript{78} However, it was shown using a larger dataset that the border in the charge-hydrophobicity plot is not well-defined, with significant overlap between ordered and disordered sequences\textsuperscript{77} (also shown in Figure 2.1 A). This so-called “twilight zone” between order and disorder was investigated using a simplified model of a protein.\textsuperscript{77} The polypeptide chain was represented using an HP model with hydrophobic (H) and polar (P) residues, and an HPN model with hydrophobic (H), positive (P) and negative (N) residue. Each residue occupies a point on a two-dimensional lattice. Conformations of peptides of varying length and composition were generated. Using this simple model, it was found that composition is insufficient to unambiguously assign a sequence as ordered or disordered, and that sequence context is important. However, as chain length increases, sequence context becomes less important. For long polypeptides, sequence composition is a sufficient criterion for classification as ordered or disordered. Remarkably, the results of this study are in agreement with a bioinformatics analysis of the compositions of disordered and ordered sequences. Sequence composition was found to be a better determinant of disorder for longer sequences. Importantly, the results of this study suggest that disorder predictors based on composition alone are not adequate for short IDPs/IDRs.\textsuperscript{77}
Coarse-grained simulations using a 3D-lattice model have been used to study eukaryotic linear motifs. A linear motif is a pattern common to a group of sequences associated with a given function, such as a protein interaction site or a cellular-compartment targeting signal. For example, SH3 domains recognize sequences with a PXXP linear motif. Linear motifs are short (3-8 residues) and often found within IDRs. A possible role for disordered regions flanking hydrophobic linear binding motifs was recently elucidated using coarse-grained simulations. In Abeln and Frenkel’s model of a polypeptide, each residue occupies a point on a cubic lattice and interacts with other residues via a pairwise interaction energy, using a Monte Carlo algorithm for conformational sampling. Using this simple representation, it was shown that flexible hydrophobic binding motifs that are not flanked by disordered regions are prone to aggregation. Without the disordered flanks, hydrophobic linear motifs are suggested to be toxic. The function of the disordered regions flanking the binding motif is to impede aggregate formation without obstructing substrate binding. In this study, a simple model was able to provide fundamental biophysical insight regarding the biological role of IDRs.

2.10 Conclusions and Outlook

Although molecular simulations have only recently been applied to the study of protein disorder, they have already provided structural biologists with an unprecedented view of the conformational ensembles of IDPs/IDRs. As evidenced by the studies reviewed here, interdisciplinary approaches incorporating both experiment and simulation are particularly valuable for elucidating the complexities of disordered states. Experimental data, usually in the form of ensemble averages, is insufficient to completely characterize a disordered ensemble consisting of many conformations. Molecular simulations using empirical force fields are well-suited to make use of restraints based on experimental data, such as PRE measurements, chemical shifts, and SAXS data. In addition, unrestrained simulations have been used to characterize disordered ensembles, providing insight for experimental observations. In turn, experimental data provides an important validation of empirical force fields, and an understanding of their limitations. De novo MD simulations afford fundamental biophysical insight into disordered ensembles, especially in the case of entropic chains, for which high resolution experimental data is not available. Continuous increases in the capability of high
performance computing facilitates simulations of ever longer timescales and larger systems. However, the development of enhanced sampling algorithms has already made it possible to achieve statistical convergence for disordered ensembles. Molecular simulation represents a promising tool for the study of protein disorder, which has so far been under-utilized.
References


CHAPTER 3

Structural Disorder and Protein Elasticity

The contents of this chapter were adapted from a published book chapter with kind permission of Landes Biosciences and Springer Science and Business Media.

Reference:

Contributions:
S.R. wrote the manuscript. R.P. provided editorial input and guidance.
3.1 Summary

An emerging class of disordered proteins underlies the elasticity of many biological tissues. Elastomeric proteins are essential to the function of biological machinery as diverse as the human arterial wall, the capture spiral of spider webs, and the jumping mechanism of fleas. In this chapter, we review what is known about the molecular basis and the functional role of structural disorder in protein elasticity. In general, the elastic recoil of proteins is due to a combination of internal energy and entropy. In rubber-like elastomeric proteins, the dominant driving force is the increased entropy of the relaxed state relative to the stretched state. We focus our discussion on the sequence, structure, and function of five rubber-like elastomeric proteins, elastin, resilin, spider silk, abductin, and ColP. Although we group these disordered elastomers together into one class of proteins, they exhibit a broad range of sequence motifs, mechanical properties, and biological functions.

3.2 Introduction

Elasticity is the intrinsic ability of a material to return to its original shape after being deformed by an external force. Elastic recoil is a property of many different materials, familiar examples of which include steel, rubber, silicon, and lycra. Flexible elastomers, such as rubber, are stretched or compressed with minimal force. Extensible elastomers, including lycra, can be stretched significantly before rupturing. Resilient elastic materials are often used as components in mechanical devices because of their ability to undergo stretching and relaxation reversibly, dissipating minimal elastic energy as heat. Thus, different elastomers are suited to different functions because of their unique set of elastic properties.

Accordingly, elastomeric proteins fulfill essential roles in species throughout the animal kingdom. In vertebrates, elastin is responsible for the elastic recoil of arteries, skin, lung alveoli, and uterine tissue. Elastin’s mechanical properties are remarkably similar to those of resilin, an elastic insect protein. Diverse insect tissues use resilin deposits to store elastic energy, including the wing joints of dragonflies and the jumping mechanism of fleas. Also found in arthropods, spider silks are a class of elastomeric materials with a wide range of elastic properties; rigid and strong silks are used for encasing eggs and restraining trapped prey, while flexible silks are used to
Molluscs (scallops and mussels) have two well-characterized elastomeric proteins: abductin and ColP. Abductin forms the flexible hinge of the scallop’s shell. When the scallop is “swimming”, the opening and closing motion of the shell propels the scallop through water. Other molluscs, including mussels, are stationary and require elastomeric threads to tether their shells to underwater surfaces. The elastomeric protein ColP permits these threads to be stretched by the force of tides without breaking. Taken together, these examples demonstrate that the biological roles of elastomeric proteins are remarkably diverse. Accordingly, elastomeric proteins exhibit a wide range of elastic properties: dragline spider silk is one of the toughest materials ever discovered, resilin is more resilient than the best synthetic rubbers, and elastin exhibits remarkable durability.

The wide range of elastic mechanical properties of biological tissues is dictated by the molecular structure of their constituent elastomeric proteins. Because of their desirable mechanical properties, the study of elastomeric proteins is motivated by their potential use in biomedical engineering and materials science. A detailed understanding of the sequence, structure, and function of these proteins provides a framework for the rational design of novel biomaterials.

Importantly, some elastomeric proteins have well-defined secondary and tertiary structures, while others are intrinsically disordered: this dichotomy in structural tendencies is embodied by collagen and elastin. Although both are elastomeric proteins, they have very different structural properties. Collagen is the protein responsible for the strength and elasticity of tendon. In fact, collagen has more than 10 times the elastic energy storage capacity of steel. Collagen has been shown by x-ray crystallography to adopt a highly-ordered, triple helix structure. By contrast, elastin’s structure is characterized by a high degree of conformational disorder, which makes it flexible and easily stretched. Due to intrinsic differences in the degree of structural order of elastin and collagen, very little force is required to stretch skin compared to the force required to stretch tendon. Other examples of elastomeric proteins with well-defined molecular structures include spectrin, keratin, and a protein recently discovered in the egg capsule of the marine snail. Structurally-ordered elastomeric proteins have been
reviewed in detail elsewhere,\textsuperscript{19,20} and are beyond the scope of the present discussion. Here, we focus on intrinsically disordered elastomeric proteins.

The purpose of this chapter is to review what is currently known about the molecular basis for the elastic properties of rubber-like elastomeric proteins. We introduce the relationship between intrinsic disorder and elasticity in section 3.3, with a brief background on rubber-like elasticity and its associated mechanical properties. We then provide a detailed description of elastomeric proteins that require structural disorder to function, including elastin, resilin, spider silk, abductin, and ColP (section 3.4). Finally, we discuss the essential sequence features of these proteins (section 3.5).

\textbf{3.3 Elasticity and Elastic Mechanical Properties}

\textbf{3.3.1 Disorder and Elasticity}

Elastic materials exhibit a broad spectrum of mechanical properties due to fundamental differences in their molecular mechanisms of elasticity. The driving force of elastic recoil, \( f \), is the sum of two contributions: an entropic component, \( f_s \), and an internal energy component, \( f_e \):\textsuperscript{3,22}

\[ f = f_s + f_e. \quad (3.1) \]

Changes in internal energy occur when an applied force distorts the material’s underlying molecular structure. In this case, the driving force for elastic recoil arises from the tendency of the molecular structure to return to the state of lowest potential energy upon removal of the external force. Stiff materials, like steel, store elastic energy in changes in internal energy (i.e. \( f_e > f_s \)). In contrast, entropic elastomers, like rubber, store elastic energy in the difference in entropy between the stretched and relaxed states (i.e. \( f_s > f_e \)).\textsuperscript{3} Entropic elastomers have a disordered molecular structure (Figure 3.1). Because there are many more ways of arranging a recoiled polymer chain than a stretched one, stretching a disordered polymer lowers the chain entropy, which is restored upon release of the strain. Compressing an entropic elastomer has the same effect as stretching: in both cases, the entropy of the relaxed state is higher than that of the
deformed state, resulting in elastic recoil. In summary, there are two general mechanisms of elastic recoil: (1) due to internal energy in a structurally-ordered elastomer, and (2) due to entropy, in a structurally-disordered elastomer.

3.3.2 Rubber-like Elasticity

The molecular driving force of rubber-like elasticity is the increased entropy of the relaxed state relative to the stretched state (Figure 3.1).\(^1\) Due to their entropy-driven elastic restoring force, rubber-like materials exhibit near-perfect recovery of stored elastic energy following deformation.\(^2,3\) The term rubber-like refers only to the elastic properties of a material, and does not imply that the chemical composition is similar to that of natural rubber.\(^6\) However, it is important to note that there are three molecular characteristics common to rubber-like materials: (1) sufficient polymer chain length, (2) high chain flexibility, and (3) the presence of inter-chain cross-links.\(^2,23\) Chain length and flexibility are necessary for elastic recoil: long and flexible polymer chains have many energetically-accessible spatial configurations, and the vast majority of these configurations are compact, resulting in the difference in entropy between the stretched and relaxed states. In turn, covalent or non-covalent cross-links maintain the structural integrity of the network by preventing the polymer chains from being pulled apart during extension.\(^2,23,24\) Thus, in order for proteins to exhibit rubber-like elasticity, their amino acid sequence must encode a sufficiently flexible structure to give rise to entropy-driven elastic recoil, and must contain amino acid residues capable of forming inter-chain cross-links. The mechanical properties of rubber-like elastomeric proteins are modulated by the flexibility of the disordered regions, the nature of the cross-links, and the spacing between cross-links.

Experimentally, rubber-like materials are identified by their unique thermoelastic behaviour: when held at a constant force, a rubber-like elastomer shrinks with increasing temperature.\(^2\) Similarly, if held at constant extension, its elastic restoring force increases in proportion to the temperature.\(^25\) Underlying both of these thermoelastic properties is the enhancement of molecular motion at higher temperature, which results in an increased tendency of the system to populate states of higher entropy.\(^3,23\) Raising the temperature increases the contribution of entropy to the free energy of the system. In addition to being entropically-
favourable, the relaxed state of an elastomer has a lower potential energy than the stretched state due to stabilizing interactions between polymer chains, such as hydrogen bonds. The relative contributions of entropy ($f_s$) and potential energy ($f_e$) to the elastic restoring force can be determined using thermoelastic (force-temperature) measurements. In natural rubber, 82% of the elastic restoring force is due to entropy, and 18% is due to energetic stabilization of the relaxed state. Similarly, the internal energy component of the elastic force is 13% for the synthetic rubber polybutadiene (used in tires).

**Figure 3.1 Entropy-driven elastic recoil**

The driving force for elastic recoil in a rubber-like elastomer is the increased entropy of the relaxed state relative to the stretched state. The relaxed state has higher entropy because there are many more ways of arranging a collapsed polymer chain than a stretched polymer chain. The effect of an external force is to decrease the entropy, which is recovered when the force is removed and the material recoils to its relaxed state. Cross-links are essential to this mechanism of elastic recoil because they prevent the chains from sliding past each other during stretching.

### 3.3.3 Measures of Elasticity

In order to fully understand the biological role of elastomeric proteins, and furthermore, to use them effectively in novel biomaterials, measurements of their elastic mechanical properties are required. These measurements are typically performed on a small biomaterial sample using a specialized mechanical testing apparatus. Mechanical (force-deformation) tests produce stress-strain curves. Stress is the applied force normalized by the cross-sectional area of the material (in units of pascals, Pa). An applied stress induces a strain, which is the change in length of the material normalized by the initial length. A typical stress-strain curve for a rubber-like elastomeric protein is shown in Figure 3.2, along with illustrations of several elastic mechanical properties.
Figure 3.2 Measurements of elastic mechanical properties

Shown here are stress-strain curves for a rubber-like elastomer in two types of mechanical tests. (A) In the first type of experiment, an elastomer is stretched past its breaking point. This measures strength, which is the stress at the rupture point, and extensibility, which is the strain at the rupture point. Toughness is the work required to rupture the material, which is the area below the stress-strain curve. (B) In the second type of experiment, the material is allowed to return to its relaxed state (without being stretched to its rupture point). The elastic modulus is the slope in the linear regime of the stress-strain curve. The work done to stretch the material is the area under the top curve (W). When the material is allowed to relax, it follows a stress-strain curve below the original curve. The area between the stretch curve and the relax curve is the elastic energy lost to heat (H). Resilience is the difference between W and H, normalized by W.

Two types of measurements are performed to measure mechanical properties: (1) extending the elastomer until it ruptures (Figure 3.2 A), and (2) allowing the elastomer to relax before it reaches its breaking point (Figure 3.2 B). The stress and strain at the point of rupture are a measure of the material’s strength and extensibility, respectively. Integrating the area under the stress-strain curve is a measure of the work needed to perform a given deformation. The work required to rupture the material is a measure of its toughness. If the elastomer is stretched and allowed to return to its relaxed state, the hysteresis between the two stress-strain curves is a measure of the elastic energy lost to heat. The corresponding mechanical property is resilience, which is the initial work minus the lost heat, normalized by the initial work done to strain the material. An elastomer’s stiffness is a measure of how easily it is deformed; stiffness is quantified by the elastic modulus, which is the slope in the linear regime of the stress-strain curve. Taken together, the strength, extensibility, toughness, resilience and elastic modulus provide a description of the elastic behaviour of a material.
3.4 Rubber-like Elastomeric Proteins

Here, we describe in detail what is currently known about the sequence features, structural characteristics, mechanical properties, and biological roles of five rubber-like elastomeric proteins: elastin, resilin, spider silk, abductin, and ColP. In addition, we summarize recent advances in other elastomeric proteins. The biological roles, sequence features, and structural properties are summarized in Table 3.1.

Thermoelastic measurements have been performed on elastin, resilin, abductin, and hydrated major ampullate spider silk. In particular, the internal energy components of the recoil force of elastin and major ampullate spider silk are 26 %\textsuperscript{3,18,26} and 14 %\textsuperscript{27} respectively. On the basis of these measurements, the thermoelastic behaviour of rubber-like elastomeric proteins is consistent with entropic elasticity\textsuperscript{9,27,28}. It is essential to note that elastomeric proteins are rubber-like only when hydrated or in polar solvents\textsuperscript{28}. Water is thought to act as a ‘plasticizer’ by forming direct interactions with the polypeptidic backbone, resulting in elastic mechanical properties\textsuperscript{16}.

Measurements of mechanical properties for the elastomeric proteins described in this chapter are provided in Table 3.2. Elastomeric proteins exhibit a remarkable diversity of elastic properties, which may differ by orders of magnitude. Different biological roles demand different combinations of extensibility, resilience, strength, and stiffness, which are all ultimately determined by structural properties. The elastic properties of proteins are modulated by the primary sequence, the domain organization, and the spacing between adjacent cross-links; these features are reviewed below.
<table>
<thead>
<tr>
<th>Protein</th>
<th>Biological Role</th>
<th>Sequence Features</th>
<th>Structural Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abductin</td>
<td>- the hinge ligament of the shells of bivalve molluscs is composed of abductin - it facilitates swimming - when compressed, abductin stores elastic energy, which when released is used to open the shell</td>
<td>- cross-links not completely characterized (^{29}) - possibly 3,3' methylene bistyrosine (^{30}) or disulfide cross-links (^{30,31}) - may have cross-linking involving Y, K residues (^{29})</td>
<td>GGFGGMGGX (^{29}) - PPII, (\beta)-turns, unordered (^{31})</td>
</tr>
<tr>
<td>ColP</td>
<td>- the elastic domains of preColP confer extensibility on the threads connecting mussels to surfaces</td>
<td>- cross-links may be formed by histidines bound to transition metals (^{32}) - DOPA lysines (^{31})</td>
<td>GPGGG (^{31}) GXGPG, XGGPG, GGXPG ((X=I,F \text{ or } A)^{12}) - no high-resolution structural data - disordered based on x-ray fibre diffraction studies of intact byssal threads (^{33})</td>
</tr>
<tr>
<td>Elastin</td>
<td>- responsible for the elasticity of various vertebrate tissues (skin, aorta, lung alveoli, elastic ligaments, uterine tissue) (^{4})</td>
<td>- desmosine, isodesmosine or lysinonorleucine cross-links, all formed from Lys residues (^{30,31})</td>
<td>VGVAPG (^{31,34}) VGVP (^{31,34}) VPG (^{31}) - the hydrophobic domains of elastin contain (\beta)-turns and PPII structure, and are intrinsically disordered (^{16,17}) - alanine-rich cross-linking domains have significant (\alpha)-helix content (^{35})</td>
</tr>
<tr>
<td>Flagelliform Silk</td>
<td>- capture spiral of orb webs (^{36})</td>
<td>- non-covalent cross-links (^{31})</td>
<td>GPGSGPGGY (^{31}) GPGGX (^{37}) GGX (^{37}) - random coil/unordered by CD (^{37}) - absence of (\beta)-sheet structure (^{38})</td>
</tr>
<tr>
<td>Major Ampullate Silk</td>
<td>- dragline silk and radial fibres of orb webs - rubber-like properties have no known biological role</td>
<td>- noncovalent cross-links composed of polyalanine intermolecular (\beta)-sheets</td>
<td>GPGXX (^{39}) - in the supercontracted state, (f/f = 14%) - structure consists of alternating (\beta)-sheet crystalline domains and amorphous domains</td>
</tr>
<tr>
<td>Resilin</td>
<td>- dragonfly wing hinge ligament (^{6}) - cicada tymbal (essential for sound production) (^{40}) - flea cuticle (stores energy for jumping) (^{7}) - tick cytoskeleton (^{41})</td>
<td>- no unique cross-linking domains (^{42}) - di-tyrosine and tri-tyrosine cross-links (^{31,42})</td>
<td>AQTPSSQYGAP (^{43}) GGRPDSYGAPGGGN (^{13,31}) GYSGRPGGQDLG (^{31,42}) - PPII, (\beta)-turns, unordered (^{44})</td>
</tr>
</tbody>
</table>
Table 3.2  Elastomeric proteins: measurements of elastic mechanical properties

<table>
<thead>
<tr>
<th>Protein</th>
<th>Elastic Mechanical Properties</th>
<th>Modulus</th>
<th>Resilience</th>
<th>Strength</th>
<th>Extensibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abductin</td>
<td></td>
<td>4.7 MPa&lt;sup&gt;9&lt;/sup&gt;</td>
<td>82-97%&lt;sup&gt;9,25&lt;/sup&gt;</td>
<td>91%&lt;sup&gt;9&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>ColP</td>
<td></td>
<td>50 MPa&lt;sup&gt;12&lt;/sup&gt;</td>
<td>53%&lt;sup&gt;12&lt;/sup&gt;</td>
<td>12.4-34.6 MPa&lt;sup&gt;45&lt;/sup&gt;</td>
<td>1.6 X&lt;sup&gt;12&lt;/sup&gt;</td>
</tr>
<tr>
<td>Elastin</td>
<td></td>
<td>1.1 MPa&lt;sup&gt;46&lt;/sup&gt;</td>
<td>90%&lt;sup&gt;46&lt;/sup&gt;</td>
<td>2 MPa&lt;sup&gt;46&lt;/sup&gt;</td>
<td>1.5 X&lt;sup&gt;46&lt;/sup&gt;</td>
</tr>
<tr>
<td>Flagelliform Silk</td>
<td></td>
<td>3 MPa&lt;sup&gt;36&lt;/sup&gt;</td>
<td>35%&lt;sup&gt;36&lt;/sup&gt;</td>
<td>0.5 GPa&lt;sup&gt;36&lt;/sup&gt;</td>
<td>2.7 X&lt;sup&gt;36&lt;/sup&gt;</td>
</tr>
<tr>
<td>Major Ampullate Silk</td>
<td></td>
<td>(10 GPa when dry)&lt;sup&gt;1&lt;/sup&gt;</td>
<td>35%&lt;sup&gt;36&lt;/sup&gt;</td>
<td>1.1 GPa&lt;sup&gt;36&lt;/sup&gt;</td>
<td>0.3 X&lt;sup&gt;36&lt;/sup&gt;</td>
</tr>
<tr>
<td>Resilin</td>
<td></td>
<td>2 MPa&lt;sup&gt;1&lt;/sup&gt;</td>
<td>92%&lt;sup&gt;1&lt;/sup&gt;</td>
<td>4 MPa&lt;sup&gt;1&lt;/sup&gt;</td>
<td>1.9 X&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

3.4.1 Elastin

Human life is entirely dependent on the elastic properties of elastin. Together with other structural proteins, elastin forms the fabric of extensible tissues, including skin, blood vessels, and elastic ligaments, and provides the elasticity required for proper physiological function.<sup>47</sup> Elastin is a major component of large arteries; bovine aorta is composed of nearly 50% elastin.<sup>48</sup> The aorta expands when the heart contracts (during systole), and recoils elastically when the heart refills with blood (during diastole).<sup>49,50</sup> In the walls of the aorta, elastin functions in tandem with collagen to produce a “J-shaped” stress-strain curve.<sup>50</sup> Elastin is responsible for the initial low stiffness region of the curve, while collagen confers increased stiffness at higher strains. Thus, collagen provides the strength required to prevent rupture due to high blood pressure, while the resilience and extensibility of the aorta imparted by elastin minimize the energetic demands on the heart and ensure smooth blood flow to tissues throughout the body.<sup>49,50</sup>

In addition to extensibility and resilience, elastin possesses remarkable durability: once laid down in tissue during development, elastin does not turn over at an appreciable rate.<sup>14</sup> In order to sustain a lifetime of breaths and heartbeats, elastin must therefore undergo millions of stretching-relaxation cycles without damage or permanent deformation. Unlike elastin in blood vessels and lungs, elastin in the uterus is degraded and replaced during adulthood.<sup>51</sup> In order to accommodate the rapid growth and motion of the fetus, the uterus requires significant
extensibility. Accordingly, during pregnancy, uterine elastin content increases by more than 500%, the majority of which is quickly degraded post partum. As a result of elastin’s impressive diversity of biological roles and exceptional mechanical properties, it is the best-characterized rubber-like elastomeric protein.

It is now possible to mimic the elastic properties and self-assembly of elastin using smaller recombinant polypeptides, which can be used to fabricate materials suitable for stress-strain measurements. Both elastin and elastin-derived peptides self-aggregate upon heating to form an organized fibrillar structure in a process known as coacervation. Remarkable durability and intrinsic capacity for self-organization make elastin an ideal biomimetic model in the development of synthetic biomaterials. Biomaterials composed of either elastin or elastin-derived peptides have desirable elastic mechanical properties: low stiffness, high resilience, and high extensibility (refer to Table 3.1 for measurements).

Elastin’s mechanical properties are encoded in its amino acid sequence, which fulfills the essential requirements for rubber-like elasticity. The sequence of tropoelastin, the monomeric precursor of elastin, is composed of alternating cross-linking and hydrophobic domains. The covalent cross-linking of elastin monomers imparts strength and stability to the polymeric matrix, while the hydrophobic domains are thought to confer the propensities for self-aggregation and extensibility. More than 80% of the sequence of the hydrophobic domains consists of proline, glycine, valine, and alanine. The hydrophobic domains of elastin have a pseudo-periodic, low complexity sequence, with repeat motifs PGVGVA, PGV, and PGVGV. There are several excellent reviews of elastin’s biochemistry and structural properties. Here, we briefly review what is currently known about the structural features of elastin’s hydrophobic domains.

For several decades, the elastin field was plagued by controversy surrounding the structure of elastin, and, correspondingly, the molecular mechanism of its elastic recoil. Models of elastin structure and function were simplistic and largely incompatible with each other. Urry postulated that elastin’s repetitive sequence must encode a perfectly repetitive structure: the
In this model, an ordered spiral consists of consecutive type-II $\beta$-turns with PG motifs forming the corners of the turn. Elasticity was thought to arise from the “librational motions” of the $\beta$-spiral. The $\beta$-spiral was also postulated as the structure other elastomeric proteins, including wheat gluten, spider silk, and resilin. In contrast to the ordered view of the $\beta$-spiral model, Flory and Gosline put forth models of elastin as a random, rubber-like structure. Thus, models of elastin structure ranged from ordered to predominantly random coil.

In support of rubber-like elasticity and a highly-disordered structure, force-temperature measurements on elastin indicate that internal energy contributes between 10% and 26% to the elastic restoring force, with entropy being the dominant molecular driving force. Consistent with these macroscopic measurements, solid state NMR has provided significant insight, suggesting the absence of $\alpha$-helix and $\beta$-sheet, and a high degree of dynamic disorder. In addition, $^{13}$C NMR studies demonstrated that the hydrated state of elastin has significant structural mobility, which decreases as water is removed. Taken together, the thermoelasticity and NMR evidence indicates that the polypeptide chains of elastin are highly mobile, and therefore possess high configurational entropy. These observations are inconsistent with models requiring conformationally-restricted structures, such as the $\beta$-spiral.

However, the random network model of elastin is too simplistic to account for experimental data consistent with the presence of $\beta$-turns and polyproline II (PPII) structure. Circular dichroism (CD) and Fourier transform infrared spectroscopy (FTIR) have provided limited structural data suggesting a high degree of conformational flexibility, together with a measurable propensity to adopt $\beta$-turns and PPII conformations. It should be noted that only qualitative interpretations of CD spectra of elastin are possible because the reference databases used by CD deconvolution programs consist primarily of globular proteins. Given these limitations, it is not possible to obtain information about equilibrium populations of either PPII or $\beta$-turn structures using CD.

Similarly to many other intrinsically disordered proteins (IDPs), the insolubility, conformational heterogeneity, and intrinsic flexibility of elastin have precluded the use of
conventional high-resolution structural determination methods, including x-ray crystallography and solution NMR. In contrast to experimental approaches, MD simulations are not hindered by conformational disorder and have therefore proven useful in obtaining atomic-level descriptions of the conformational ensembles of elastin-like peptides. Molecular simulations provide information that can be used to characterize the ensemble of IDPs. This is because MD simulations provide time trajectories containing information on the dynamics of all particles in the system. Due to limited computing power, MD simulation studies of elastin were until recently restricted to short time scales (nanoseconds) or small oligopeptides (only eight residues).

However, the simulations of elastin-like peptides described in this thesis have dramatically extended both the timescale and system size compared to prior computational studies: it is now possible to reach the microsecond time scale for peptides of similar size to a hydrophobic domain (35 residues). Using all-atom MD simulations with explicit water, we obtained a disordered conformational ensemble for the elastin-like peptide (GVPGV) (see Chapters 7 and 8). In order to effectively sample many conformational states, we employed a novel enhanced sampling method (see Chapters 5 and 6). The structure of elastin-like peptides is flexible and disordered, which indicates that the underlying energy landscape is defined by conformations that are very similar in structure and energy, and these conformations exchange rapidly with one another. For this reason, it is relatively easy to obtain meaningful structural information and compute thermodynamic averages from molecular simulations. More detailed structural insights are the subject of the following chapters.

3.4.2 Resilin

Another rubber-like elastomeric protein with very similar mechanical properties to elastin is the insect cuticle protein, resilin. First discovered in the elastic tendon of dragonflies, resilin is found in many arthropod species and is important to insect flight, locomotion, and sound production. The tymbal mechanism of the cicada utilizes resilin as an energy storage device. When the tymbal is compressed, elastic energy is stored in resilin, and the subsequent release of this energy is accompanied by the cicada’s characteristic sound. Resilin in the cuticle of ticks
facilitates the dramatic expansion of the cytoskeleton during feeding, a property not found in most other insects that shed their cytoskeleton before significant growth can occur. Another biological role of resilin is as an energy storage device in the jumping mechanism of fleas: muscle contraction alone is incompatible with the timescale of energy release (less than one millisecond) and the necessary power output. Accordingly, the cuticle of fleas contains a resilin pad, the size of which correlates with jumping ability.

Unlike elastin, which is challenging to isolate and purify, the resilin pads and tendons in insects are isotropic and easily isolated, and therefore are convenient for experimental characterization. However, in order to produce sufficient amounts of resilin to manufacture biomaterials, several groups utilize recombinant expression systems. A resilin-like protein (rec1 resilin) was cloned and expressed in E. coli. Large quantities of soluble rec1 resilin were produced, and cast into rods and strips. Importantly, synthetic rec1 resilin materials have the same resilience (90-92%) as elastic tendon isolated from dragonfly wing (92%). Rec1 resilin strips have a resilience of 97% in solution, dissipating only 3% of elastic energy as heat. Thus, resilin’s resilience is greater than that of polybutadiene, a high-resilience rubber, and is unmatched by any other elastomeric protein. Furthermore, synthetic resilin can be stored in a dehydrated state, and recover the same resilience upon rehydration. Besides its high resilience, resilin also has a high extensibility. Synthetic resilin strips can be stretched to more than three times their original length without permanent deformation, and elastic tendons composed of resilin can be compressed by a factor of one-third.

In addition to these exceptional elastic properties, resilin has a high durability. It is deposited in the insect cuticle during the pupal stage, and remains in place throughout the adult lifetime. While insects generally have much shorter life spans than vertebrates, their resilin deposits have similar durability requirements as elastin. For example, the resilin deposit in the cicada’s tymbal is compressed hundreds of millions of times in sound production, necessitating a high durability. Taken together, resilin’s high extensibility, high resilience, high durability, and low stiffness and make it ideal for a wide variety of biomaterials applications. To this aim, a resilin-like sequence was recently combined with a cell binding domain in a designed
recombinant protein. The resulting biomaterial has a high extensibility (up to 200 %), and is capable of cell adhesion. This study represents an important first step towards tissue engineering because fibroblast cells were able to adhere, and proliferate on the extensible resilin-like scaffold.67

Based on their similar mechanical properties, it is not surprising that the sequences of elastin and resilin share a number of similar features. Both have a high content of proline and glycine, and are highly repetitive (see Table 3.1 for specific repeat motifs).16,42,44 However, unlike the predominantly non-polar sequence of elastin, the sequence of resilin is depleted of large hydrophobic residues like valine and isoleucine, and has a significantly enhanced content of hydrophilic residues.44 Towards the development of a resilin-inspired biomaterial, Tamburro and coworkers recently identified PGGGN as a putative minimal repeat motif for resilin-like properties.44 Peptides based on this motif were found to readily self-assemble into fibrillar structures.44 In contrast to the regular alternance of cross-linking and hydrophobic domains in elastin, the sequence of resilin has no specific cross-linking domains; instead, tyrosine residues are interspersed throughout the elastic repeat motifs,24 and form di- and tri-tyrosine cross-links. It is estimated that 20 % of tyrosine residues are covalently cross-linked,13 with a spacing of 40 to 60 residues between cross-links.68 Thus, the sequence of resilin is compatible with the requirements for rubber-like elasticity, with cross-links separated by flexible polymeric chains.

Resilin was the first elastomeric protein to be identified as rubber-like on the basis of thermoelasticity experiments.28 In agreement with an entropy-driven mechanism of elastic recoil, CD, NMR, and Raman spectra of a resilin-like protein are all consistent with a heterogeneous and dynamic structure.43 Vicinal coupling constants from NMR indicate an absence of either α-helix or β-sheet structure, while chemical shifts are consistent with “random coil” values.43 Sequential and medium range NOEs indicate the presence of β-turn conformations, with PG, GG and PS motifs forming the corners of the turn.44 Similarly, CD spectra are consistent with the presence of PPII structure and β-turns.31 X-ray diffraction measurements indicate that resilin does not attain any significant ordered structure, even when stretched to three times its length or when dried.68 Although the β-spiral structure has also been
proposed for resilin, all of the experimental evidence obtained to date is incompatible with this model. Instead, a common thread emerges: the polypeptide chain of resilin is highly flexible and intrinsically disordered.

3.4.3 Spider Silk

In order to construct their intricate webs, spiders utilize silks with diverse elastic mechanical properties. Using a complex spinning and extrusion process, spiders finely tune the elastic properties of the silk proteins (spidroins) produced in their abdomen. The best-characterized spidroins belong to spiders from the family Araneoidea. Known for their orb-shaped webs, araneid spiders include the common garden spider, *Araneus diadematus*, and the golden orb weaver, *Nephila clavipes*. Araneid spiders produce seven unique types of silk, which are named after the specialized abdominal glands in which they are synthesized: (1) major ampullate silk is spun into fibres forming both the dragline and the radial threads of the web; (2) minor ampullate silk reinforces the dragline and web frame; (3) flagelliform silk forms the capture spiral; (4) aggregate silk is an aqueous ‘glue’ coating the capture spiral; (5) the ‘cement-like’ silk from the piriform gland is used to attach the web to a surface; (6) aciniform silk is used to restrain captured prey; and (7) tubuliform/cylindriform silk is used to protect the egg sac.

While spiders make use of silk for a wide range of purposes, the most familiar use is in the construction of the spider web. A web is an incredibly efficient insect-catching device: a spider can produce a web covering nearly 1 m² using only 180 µg of protein. Although the threads of the web are very thin, they exhibit toughness greater than the best synthetic materials, including Kevlar®. It is for this reason that the web does not break upon impact with an incoming insect, or when a trapped insect scrambles to get free. Orb-web spiders construct their webs using a combination of two silks with complementary elastic mechanical properties. The threads forming the radial frame of the web are stiff and extremely tough; they are composed of both major and minor ampullate silk. The capture spiral connecting the radial threads is sticky, highly extensible and easily stretched; it is composed of flagelliform silk, with stickiness and hydration provided by “glue” silk from the aggregate gland. Both major ampullate silk and flagelliform silk have low resilience (roughly 30%). This is essential for two reasons: (1)
the elastic energy stored in the web by the impact of an incoming insect is dissipated as heat, preventing the insect from bouncing off the web, and (2) the dissipation of elastic energy contributes to the overall toughness of the web, preventing the strands from breaking.8

Here, we are particularly interested in the two types of spider silk possessing rubber-like elasticity: major ampullate silk and flagelliform silk. Both major ampullate and flagelliform silk contract when immersed in water, but the effect is much more pronounced for major ampullate silk, and is called “supercontraction”.36 Thermoelastic measurements indicate that major ampullate silk exhibits rubber-like elasticity, but only in the supercontracted, hydrated state. The internal energy component of the elastic force is only 14 %.27 When hydrated, major ampullate silk has an elastic modulus three orders of magnitude smaller than the dry state (107 Pa compared to 1010 Pa).8 A biological function has not yet been identified for the supercontracted, hydrated state of major ampullate silk.36 The structure of major ampullate silk is thought to consist of alternating crystalline and amorphous domains. The crystalline regions are short polyalanine stretches of 8 to 10 residues, which form β-sheets oriented parallel to the fibril axis.36,37 Because the β-sheets are intermolecular, they effectively act as noncovalent cross-links,36 while the conformational entropy of the amorphous domains results in elastic recoil.27 As a result, the structure of major ampullate silk resembles a rubber with “crystalline inclusions”.8

The mechanical properties of flagelliform silk are qualitatively similar to those of supercontracted major ampullate silk.27 In contrast to elastin and resilin, flagelliform silk has a relatively low resilience, which is essential to its functional role: if it returned the stored elastic energy efficiently, insects that fly into a spider web would immediately bounce off.36 In order to effectively trap insects in the web, flagelliform silk must be coated in a sticky mixture of hygroscopic peptides and glycoproteins, which are produced in the aggregate gland.72 Flagelliform silk is a very promising biomaterial by virtue of its unusually high strength (0.5 GPa), which is approximately ten times that of any other rubber-like elastomer.36 However, the cannibalistic and territorial nature of spiders precludes the possibility of directly harvesting sufficient quantities of silk to manufacture materials.72 An additional complication arises from the importance of the effect of spinning on mechanical properties. Even if silk can be
recombinantly-expressed, mimics of the spider’s spinnerets are essential to obtain strong and extensible fibres.

Compared to elastin and resilin, there is relatively little high-resolution structural information on flagelliform silk. A recent solid-state NMR study of the flagelliform-like sequence (GPGGA)$_n$G demonstrated that the motif GPGG has a high propensity to form a $\beta$-turn. In addition, $^{13}$C chemical shifts and Raman spectra are consistent with a “random-coil” structure. These results are inconsistent with the $\beta$-spiral structure, which was proposed for flagelliform silk because the sequence encodes repetitive pentapeptide motifs. While cross-links are necessary to explain the remarkable strength of flagelliform silk, the residues involved in cross-linking have not yet been identified.

3.4.4 Abductin

Abductin is an elastomeric protein that forms the hinge ligament of the bivalve mollusc shell. In molluscs of the family Pectinidae, the shell opens and closes three times per second to facilitate swimming. The name of abductin is derived from its biological role: the hinge composed of abductin acts as an abductor, an antagonist to the adductor muscle in the opening and closing motion of the shell. The adductor muscle stores elastic energy in the abductin hinge by compressing it as the shell closes. The shell opens when the adductor muscle relaxes and the energy stored in the abductin hinge is released to oppose the force of the surrounding water. In order to minimize the energy required by the mollusc for swimming, the energy stored in the abductin hinge must be recovered to open the shell. Thus, high resilience is essential to abductin’s biological function. Mechanical tests on abductin demonstrated its resilience to be between 82 and 96 %, with significant variation between mollusc species. Interestingly, an analysis of the amino acid composition of abductin from several species revealed that resilience is correlated with glycine content. In fact, the sequence of abductin is characterized by an unusually high glycine content (nearly 70 % in fast-swimming mollusc species). Glycine content is therefore a fundamental determinant of the rubber-like mechanical properties of abductin.
Since abductin’s biological function demands a high resilience, it is not surprising that its amino acid sequence and its structural properties are similar to those of both elastin and resilin. Thermoelasticity measurements on abductin place it among the rubber-like elastomeric proteins, with a primarily entropy-driven elastic restoring force. Like elastin, resilin, major ampullate silk, and flagelliform silk, abductin exhibits rubber-like elasticity only when hydrated. Accordingly, the abductin hinge ligament is composed of approximately 50% water. The amino acid sequence of abductin is highly repetitive, with a consensus repeat motif GGFGGMGGGX. Although the complete sequence of abductin from scallops has been determined, the residues involved in cross-linking have not been unambiguously identified. CD and NMR spectra of abductin-like sequences are consistent with the presence of both PPII structure and β-turns, but no high-resolution structural studies have been performed to date. Abductin’s high resilience and compressibility make it an interesting biomaterial worthy of further structural and mechanical studies.

3.4.5 ColP from Byssal Threads

Marine mussels use byssal threads to attach themselves to solid substrates, such as rocks and harbour walls. Since a secure attachment is vital to their survival, byssal threads require both strength and extensibility. In addition, the low resilience of byssal threads effectively dissipates elastic energy like a damped spring, preventing the mussel from hitting the hard surface to which it is attached. The mechanical properties of the byssal thread vary along its length: the distal end (near the point of attachment to the surface) is stiff and strong, while the proximal end (near the shell) has a high extensibility and low elastic modulus.

The byssal thread’s continuum of mechanical properties is the result of a protein gradient along its length. The distal end is primarily composed of the protein ColD, whereas the proximal end is primarily composed of ColP, and the intermediate region contains a mixture of both ColP and ColD. The protein ColP is itself a hybrid: it is the first known example of a block copolymer containing both elastic and collagen-like domains. The collagen-like domain is predicted to adopt a triple helix structure, while the elastic domains are enriched in hydrophobic residues and contain many instances of the PG motif. Based on sequence
similarity between ColP’s elastic domains and the repeat motifs of elastin and flagelliform silk, it has been proposed that ColP confers elasticity and extensibility on byssal threads.\(^\text{77}\) As a consequence of combining these two types of sequences, proximal byssus has mechanical properties intermediate between those of elastin and collagen. Compared to collagen, it has higher toughness and extensibility, at the cost of reduced strength.\(^\text{77}\) The extensibility of the proximal region is similar to that of elastin and resilin, with a significantly lower resilience and an elastic modulus that is an order of magnitude greater.\(^\text{1,45}\) The hybrid nature of the byssal thread endows it with the remarkable strength and flexibility to maintain surface attachment against powerful tides.\(^\text{33,78}\)

Very little is known about the structural properties of ColP. X-ray fibre diffraction studies of intact byssal threads indicate a gradual decrease in structural order along the length of the thread (in the distal to proximal direction). Diffraction patterns are consistent with the presence of ordered collagen-like structure in the distal region, gradually giving way to increased structural disorder in the proximal region.\(^\text{33}\) While structural studies on intact byssal threads are a useful first step, higher resolution structural studies, ideally on ColP or its elastic domains in isolation, are essential to elucidate the connection between structural disorder and elastic properties. Strictly speaking, ColP has not yet been demonstrated to be a rubber-like elastomer with thermoelasticity measurements, but here we group its elastic domains with the other rubber-like elastomeric proteins on the basis of high sequence similarity.

### 3.4.6 Other Rubber-like Elastomeric Proteins

Interestingly, some proteins that form rubber-like biomaterials do not require rubber-like elasticity to fulfill their biological role. The sequence signatures of disordered elastomeric proteins were recently found in a domain of the transcription factor Ultrabithorax (Ubx) from *Drosophila melanogaster*. The sequence of Ubx is enriched in glycine and contains multiple GGX and GXXP elastin-like motifs. In vivo, Ubx interacts directly with DNA to regulate transcription; it is not known to form aggregates as part of its biological function.\(^\text{79,80}\) Remarkably, however, Ubx was recently found to self-assemble into elastic materials with several morphologies, including films, fibres and sheets.\(^\text{80}\) Ubx ‘ropes’ were found to have an
extensibility approximately one-third that of elastin. Similar to the rubber-like elastomeric proteins, Ubx materials are only extensible when hydrated and become brittle when desiccated. The identification of Ubx as an elastomeric protein suggests an interesting research direction: the identification of sequences with similar features to known elastomeric proteins as possible novel rubber-like elastomers.

In addition to looking for new elastomeric proteins through sequence similarity to known elastomeric proteins, it is also essential to investigate the protein constituents of extensible and soft biological tissues. It is very likely that many more rubber-like elastomeric proteins exist that have not yet been discovered. For example, studies on octopus aorta revealed the presence of a rubber-like elastomeric protein.\textsuperscript{81} The “octopus arterial elastomer” (OAE) performs the same biological role as elastin in vertebrate arteries. Like the other rubber-like elastomeric proteins, its elastic recoil is predominantly entropy-driven. However, the amino acid composition of OAE is very different than that of elastin, resilin or abductin. More than 33% of the sequence consists of charged residues, and there is significantly less proline and glycine. OAE is stiffer, less extensible, and less resilient than elastin and abductin.\textsuperscript{81} Although both elastin and OAE are responsible for the elastic recoil of the aorta, their sequence features are quite different. Thus, in developing a toolkit of elastomeric proteins for incorporation in biomaterials, it is valuable to study proteins from the myriad of elastic tissues that exist in nature. As a first step, we require information on the sequence determinants of rubber-like elasticity.

### 3.5 Sequence Features of Disordered Elastomeric Proteins

The sequences of elastomeric proteins have very little sequence homology,\textsuperscript{19} and are instead characterized by a common ‘style’ of sequence. Elastomeric domains typically have low complexity sequences with repeat motifs (refer to examples in Table 3.1).\textsuperscript{24} These repeat motifs often contain PG and GG dipeptides, which preferentially form β-turns.\textsuperscript{4,82} For several decades, it was thought that the repetitive sequences of elastomeric proteins must encode a repetitive molecular structure (the β-spiral).\textsuperscript{24} However, a wealth of experimental evidence indicates the absence of a well-defined or repetitive structure, which is corroborated by computational results. The observation that a repetitive sequence leads to a disordered structure is not unexpected,
given the well-established connection between low sequence complexity and structural disorder. In general, tandem repeats are more common in the sequences of IDPs when compared to all sequences in the Swiss-Prot database. In this sense, elastomeric proteins are prototypical IDPS. Our studies have uncovered an important sequence feature, a high combined content of proline and glycine that is common to all rubber-like elastomeric proteins. We describe this sequence feature in detail in Chapter 4.

3.6 Conclusion and Perspectives

There is an emerging consensus that rubber-like elastomeric proteins are intrinsically disordered, and therefore exhibit entropy-driven elastic recoil. Although a fundamental requirement for elastomeric domains is to remain disordered even when aggregated, they are not “random coils”. Spectroscopic evidence and computational studies both point to the presence of significant amounts of transiently-populated hydrogen bonded turns and PPII structure. Importantly, the work reviewed in this chapter shows that the study of rubber-like elastomeric proteins benefits greatly from a synergy of theoretical and experimental approaches. Molecular simulations offer high resolution structural information that is complementary to data obtained using spectroscopic approaches, including CD, FTIR, and solid-state NMR. However, despite significant progress in the structural characterization of disordered states, there is at present very little insight into how the global effect of chain entropy relates to the fine balance of microscopic properties resulting in the distribution of conformations of the polypeptide chain and various side chains, their hydration, aggregation, and extension. Such detailed understanding is required to explain why the mechanical properties of rubber-like elastomers exhibit significant variation. Understanding the structural properties of rubber-like elastomeric proteins is a necessary prerequisite to their effective use in biomaterials applications, and, furthermore, to the rational design of novel elastomeric proteins.

Thus far, the majority of detailed studies on elastomeric proteins have focused on elastin, and, more recently, on spider silks. This is likely because biological science is driven to achieve advances that directly impact human health, which often narrows the focus of investigations to proteins related to specific human diseases. While there are advantages to such a focused
approach, namely, the wealth of structural data on elastin and elastin-derived peptides, we have demonstrated in this chapter that there are many similarities between elastin and the other rubber-like elastomeric proteins, all of which have tremendous potential in biomaterials development. For the most part, biomaterials scientists have not yet exploited the emerging knowledge of the myriad of elastomeric proteins that are adapted to various functional roles in nature. Thus, as a next step, it is essential to obtain more detailed structural and mechanical characterization of resilin, abductin, CoIP and other rubber-like elastomeric proteins. Simultaneous to fundamental structural investigations, hypotheses regarding essential sequence features can be tested by recombinantly-producing elastomeric materials. These studies will provide complementary information to our knowledge of elastin and spider silks. Our toolkit of rubber-like elastomers needs to be dramatically expanded. It is likely that many more rubber-like elastomeric proteins can be discovered by characterizing the proteins responsible for elastic recoil in biological tissues, which is the approach that led to the discovery of the octopus arterial elastomer. Once we have identified the essential sequence features of rubber-like elastomers, we can search genomic databases based on sequence similarity. It is likely that there are other proteins, like Ubx, that self-assemble to form elastomeric materials even though elastic properties have no known relevance to their role in vivo.

Many of the elastomeric biomaterials reviewed in this chapter exhibit a common design theme: the combination of elastomers with diverse mechanical properties in hybrid materials tuned to fulfill a particular biological role. Combinations of elastomers are found in byssal threads, which unite collagen-like and elastin-like sequences; the spider orb-web, which combines stiff and strong major ampullate silk with sticky and extensible flagelliform silk; and mammalian arterial walls, which achieve resilience and strength with a combination of elastin and collagen fibrils. In addition to designs incorporating elastomeric domains with differing mechanical properties, it is also possible to incorporate other biologically active domains, such as those facilitating cell-adhesion. Furthermore, even the sequence of elastin exhibits duality in its alternating cross-linking and hydrophobic domains. Only after we understand the sequence determinants of elasticity for elastomers in isolation will it be possible to effectively combine
different elastomers in hybrid materials, which represent a clear next step in biomimetic materials.

Through sequence, structure and mechanical studies of rubber-like elastomeric proteins, recent advances have led to the development of a framework for the rational design of self-assembling biomaterials. These studies will advance the development of biomimetic materials for a variety of applications, including vascular grafts, tissue replacements, and scaffolds for tissue regeneration.
References


CHAPTER 4

Proline and Glycine Control Protein Self-Organization into Elastomeric or Amyloid Fibrils

The contents of this chapter were adapted with permission from an article published in *Structure* as well as parts of the book chapter that formed the basis for Chapter 3.

Reference:

Contributions:
This study resulted from a collaboration between the Pomès and Keeley labs; the contributions of the authors are as follows. S.R. was the primary author of the manuscript. S.R. performed the simulations and analysis of monomeric systems and the analysis of elastomeric and amyloidogenic sequences included in Figure 4.4. S.B. performed the simulations and the analysis of the aggregated systems, contributed to the writing of the section describing the results for the aggregated systems (section 4.5), and prepared Figures 4.3 and 4.6. The experimental part of the study was performed by M.M, and M.M. and F.W.K. wrote the experimental methods and results sections (section 4.9, 4.12.3, and 4.12.4). R.P. provided significant editorial input and guidance.
4.1 Summary

Elastin provides extensible tissues, including arteries and skin, with the propensity for elastic recoil, whereas amyloid fibrils are associated with tissue-degenerative diseases, such as Alzheimer’s. Although both elastin-like and amyloid-like materials result from the self-organization of proteins into fibrils, the molecular basis of their differing physical properties is poorly understood. Using molecular simulations of monomeric and aggregated states, we demonstrate that elastin-like and amyloid-like peptides are separable on the basis of backbone hydration and peptide-peptide hydrogen bonding. The analysis of diverse sequences, including those of elastin, amyloids, spider silks, wheat gluten, and insect resilin, reveals a threshold in proline and glycine composition above which amyloid formation is impeded and elastomeric properties become apparent. The predictive capacity of this threshold is confirmed by the self-assembly of recombinant peptides into either amyloid or elastin-like fibrils. Our findings support a unified model of peptide aggregation in which hydration and conformational disorder are fundamental requirements for elastomeric function.

4.2 Introduction: Two types of protein self-organization: elastomeric and amyloid-like

Rubber-like elastomeric proteins require a high level of structural disorder for entropy-driven elastic recoil. In addition, they must self-aggregate in order to form an elastomeric network. Their sequence must therefore preclude the possibility of forming well-structured protein aggregates. In particular, elastomers must avoid the formation of amyloid fibrils. In contrast to elastin fibres, which are essential in many extensible tissues, amyloid fibrillar deposits are associated with numerous tissue-degenerative pathologies, including Alzheimer’s and Parkinson’s diseases. Using electron microscopy and solid state NMR, the molecular structure underlying amyloid fibrils has been shown to consist of a cross-β quaternary structure, with β-strands running perpendicular to the main axis of the fibril. While most studies of amyloidogenic proteins have understandably focused on specific sequences implicated in human disease, there is evidence for the wider relevance of self-assembly of polypeptide chains into amyloid fibres. An increasing number of proteins with no known associated pathology have been shown to populate the amyloid state under destabilizing conditions, or with the introduction of destabilizing mutations. These include the natively α-helical protein...
myoglobin⁶ and exon 30 of human elastin, which forms amyloid when removed from the context of the full tropoelastin sequence.⁷ Recently, functional amyloids have been identified: Pmel17, which is an intracellular amyloid acting as a template for melanin synthesis in mammalian melanosomes,⁸,⁹ and silk moth chorion protein, which acts as a protective barrier for silk moth eggs.¹⁰ It has been proposed that the amyloid fibril represents an inherent form of organization potentially accessible to all polypeptide chains under appropriate conditions.¹¹,¹² Highly hydrophobic sequences, such as that of elastin, are, in principle, susceptible to forming amyloid. Therefore, it is important to understand how the hydrophobic domains of tropoelastin, and indeed all self-associating elastomeric proteins, manage to avoid the amyloid fate.

### 4.3 Study of Model Peptides

In order to uncover the fundamental balance of forces underlying the assembly of polypeptides into amyloid-like or elastin-like protein aggregates, we designed a set of model sequences with physico-chemical properties compatible with those of the hydrophobic domains of elastin, as well as amyloid fibrils. Our design strategy is based on studies of recombinant elastin-like peptides composed of alternating hydrophobic and cross-linking domains (Figure 4.1 A). The sequence of the hydrophobic domains of elastin is simple and repetitive (Figure 4.1 B). Elastin-like peptides containing human exon 24 (which contains the repeat fragment PGVGVA), or chicken exon 24 (with repeat fragment PGVGV), have been shown to form biomaterials with self-aggregation and mechanical properties similar to native polymeric elastin.¹¹,¹² Mutations of the tandem PGVGVA repeats to GGVGVA, intended to increase the conformational flexibility of the polypeptide chains, promoted the formation of amyloid-like fibrils.¹² Interestingly, mutations of the PGVGV repeats to GGVGV resulted in a polypeptide that could either coacervate like elastin peptides or form amyloid, depending on solution conditions (Figure 4.1 B and C).¹² Poly(GGVGV) displayed a similar ambiguity in its aggregation – forming an elastin-like film or amyloid fibrils when deposited from methanol or aqueous solution, respectively.¹³ Based on these observations, we considered four repeat units, PGVGVA, GGVGVA, PGVGV, and GGVGV, which encode for materials spanning elastin-like and amyloid-like aggregates. Recognizing that these repeats are formed by the pairwise combination of four distinct fragments: PGV, GGV, GV and GVA, we completed our model set
with the six remaining combinations: PGVPGV, GGVPGV, GGVGGV, GGVV, GVAGVA and GVGVA. The first two of these are the repeat units found in chicken exon 20 and human exon 20 (Figure 4.1 B). All ten pairwise combinations may be represented as a simple matrix (Figure 4.1 D) in which each sequence motif differs from its nearest neighbors by a single residue, as does the ordered set (PGV, GGV, GV, GVA). This minimalist approach facilitates systematic comparison between peptides, in addition to the computational averaging provided by repetitive sequences.

To investigate how the ability of peptides to self-assemble into elastin-like or amyloid-like fibrils is modulated by sequence, we performed extensive molecular dynamics simulations of this model set of peptides in both monomeric and aggregated states. Our results show that elastin-like and amyloid-like peptides are separable on the basis of backbone hydration and conformational disorder, and that these properties are modulated by proline and glycine. Furthermore, there is a threshold in combined proline and glycine content above which elastomeric properties become apparent and below which the formation of amyloid fibrils is possible. Taken together, our findings reveal a direct quantitative relationship between the intrinsic structural properties imparted by the sequence of self-assembling proteins and their propensity to form either amyloid or elastomeric fibrils.
**Figure 4.1 Sequence of model peptides**

(A) The recombinant peptides EP20-X1-X2, with hydrophobic domains in blue and cross-linking domains in orange. The first hydrophobic domain is human exon 20 (HU20), while X1 and X2 are variable. (B) A variety of sequences which may be substituted for X1 and X2, including human exon 24 (HU24), either native or with mutations of prolines to glycines; chicken repeat sequence (CRS) or chicken exon 24, either native or with PG mutations; human exon 20 (HU20); chicken exon 20 (CH20); and a repeat of GVA. Elastin’s hydrophobic domains contain extended tracts composed mostly of four residues (P, G, V and A) which together comprise nearly 75% of the sequence of tropoelastin. To highlight the repetitive nature of these sequences, repeats of PGV, GGV, GV and GVA are indicated in red, blue, yellow and green, respectively. (C) Summary of experimental characterization of EP20-X1-X2 recombinant peptides. (D) The set of model peptides used in our computational study based on the sequences in (B) and constructed using repeats of pairwise combinations of PGV, GGV, GV and GVA. The length of these model periodic sequences (35 or 36 residues) is comparable to that of the hydrophobic domains of elastin.
4.4 Results for Monomers

The monomeric peptides adopt collapsed, water swollen conformations reminiscent of the unfolded ensemble of globular protein domains (Figure 4.2 D, E, and F). The structures exhibit a highly flexible polypeptide backbone, with exchanging conformations and overall structural disorder. Although these structures contain no classical α-helix or β-sheet, they are not random coils. Ordered structure is observed predominantly in the form of polyproline II (PPII) content (Figure 4.2 A) and hydrogen-bonded turns (Figure 4.2 B), both of which are local.

Due to the intrinsic conformational flexibility of the peptides, structural properties are best described in a probabilistic manner (Figure 4.3). Hydrogen-bonded turn content, $X_{\text{HB}}$, is anti-correlated to the degree of backbone hydration, $X_{\text{W}}$ (Figure 4.3 A). This anti-correlation reflects the local equilibrium of the peptide backbone: hydrogen bonding may be satisfied either with water or with other backbone groups in the form of hydrogen-bonded turns. The peptides separate into three clusters, ranked in order of increasing turn and decreasing hydration propensities: (1) proline-containing elastomers, (2) presumed “ambivalent” sequences including GGVGV and (3) presumed amyloids including GVAGGV. At around 10%, the low turn propensity of P-containing peptides may be rationalized in terms of conformational restrictions of the backbone induced by prolines. This restriction in dihedral angles is reflected in a higher propensity to adopt PPII conformations, which are characterized by high backbone hydration (Figures 4.2 A and 4.3 C). It has been established that PPII structure is present in significant amounts in the unfolded state of peptides and proteins. In agreement with studies of peptides modeling the unfolded state of proteins, the PPII structure observed is not extensive but is instead confined to one or two consecutive residues.
Figure 4.2  Representative conformations of monomers and aggregates
Significantly-populated structural elements: (A) polyproline II with nearby water; (B) hydrogen-bonded turn; and (C) β-sheet. Backbone representations of monomers and aggregates: (D) and (G) (GVPGV)_7; (E) and (H) (GGVGV)_7; and (F) and (I) (GVAGGV)_6. Chain color indicates residue type: G (yellow), P (red), V (green), A (blue). Proline-containing elastin-like repeats, including GVPGV, form amorphous and disordered aggregates. By contrast, sequences devoid of proline form amyloid-like structures with significant β-sheet content.

4.5  Results for Aggregates

Although it is established that amyloid fibrils contain cross-β sheets, the molecular organization of elastin aggregates is largely unknown. To probe the intrinsic ability of our model set of peptides to self-aggregate into extended β-sheets, four extended chains were placed parallel
and adjacent to each other in vacuo, so as to bias the starting conformation towards extended β-sheets, and allowed to relax with unconstrained MD simulations, first in vacuo and then in explicit water. After 20 ns of simulation in water, the chains devoid of proline retained extensive cross-strand β-sheets (Figure 4.2 H and I). In contrast, the P-containing sequences did not form extended β-sheets and instead remained highly disordered (Figure 4.2 G). This is consistent with solid-state NMR data indicating a lack of β-sheet in the aggregated state of bovine elastin. These results indicate that the potential for forming amyloid-like structures is restricted to the sequences without proline. Presumed ambivalent and presumed amyloid sequences cannot be separated on the basis of peptide-peptide hydrogen bonding propensity, reflecting the capacity of both groups to form inter-strand hydrogen bonding in extended β-sheets. This is consistent with the ability of ‘ambivalent’ sequences, like GGVGV, to form amyloid fibrils under appropriate conditions.

4.6 Comparison of Monomers and Aggregates

The structural tendencies of the tetrameric aggregates are largely consistent with those of hydrated single chains. In particular, both states display a similar anti-correlation between hydration and peptide-peptide hydrogen bonding (Figure 4.3 A and B). Significantly, elastin-like sequences retain higher backbone hydration and PPII propensities and form fewer peptide-peptide hydrogen bonds compared to the other model sequences (Figure 4.3 A to D). In support of these findings, high PPII content is observed for hydrophobic exons of elastin using circular dichroism. Qualitative agreement between monomers and aggregates is also seen in the amount of hydrophobic surface buried relative to the fully extended state, S_{HB} (Figure 4.3 F). Three clusters emerge in order of increasing S_{HB}: (1) presumed ambivalent sequences, (2) presumed amyloids, and (3) elastin-like peptides, with an overall increase in surface burial for aggregates relative to single chains. The buried hydrophobic surface of all ten peptides is considerably less than that of globular proteins of the same number of residues (cellobiohydrolase I, 19.6 nm$^2$; porcine peptide YY, 24.3 nm$^2$; villin headpiece, 24.3 nm$^2$, compared to an average of 5 ± 1 nm$^2$ for the elastin-like single chains). This observation is consistent with the lack of a compact folded state or buried hydrophobic core. It should be noted that the non-polar packing of amyloid-like aggregates in our simulations is likely to underestimate the extent of
hydrophobic burial in amyloid fibrils, in which layered β-sheets have been observed. However, the overall consistency of the properties of monomers and aggregates suggests that the ability of the polypeptide chains to self-organize into either elastin-like or amyloid-like materials is determined by the intrinsic backbone propensities imparted by the sequence.

4.7  Rigid Prolines, Plastic Glycines

In the division between elastin-like and amyloidogenic peptides, there appear to be two major sequence determinants: proline and glycine. The combination of these two residues is remarkable in light of their radically different backbone plasticity. Proline is the primary determinant: its conformationally-restricted main chain induces a significant propensity for PPII structure and an intrinsically reduced ability to form hydrogen-bonded turns and β-sheet. Both of these structural tendencies lead to increased backbone hydration. Glycine is the secondary determinant: in the absence of proline, it is the fraction of glycine that determines the extent of backbone hydration in monomers. Indeed, the propensity for the backbone to form hydrogen-bonded turns, $X_{\text{HB}}$, decreases linearly with increasing glycine composition, $X_G$, in water (Figure 4.3 E). This can be rationalized in terms of polypeptide chain entropy, as increasing glycine content increases the entropic cost of constraining the backbone in the formation of turns. Inversely, in the absence of competition for hydrogen bonding from the solvent, the backbone plasticity conferred by glycine helps better satisfy polar interactions through intramolecular hydrogen bonds. As a result, the probability to form hydrogen-bonded turns in vacuo increases with glycine content (Figure 4.3 E). These results expose the dual character of glycine, which favors either ordered or disordered structures, depending on the environment. This adaptive capacity could contribute to the ambivalent nature of glycine-rich peptides devoid of proline, such as poly-(GGVGV).
Figure 4.3 Average structural properties

Average structural properties computed from simulations of the ten model sequences: 1, (PGV)_{12}; 2, (GGVPGV)_{6}; 3, (GVPGV)_{7}; 4, (GVAPGV)_{6}; 5, (GGV)_{12}; 6, (GGVGV)_{7}; 7, (GV)_{18}; 8, (GVAGGV)_{6}; 9, (GVGVA); 10, (GVA)_{12}, which are grouped into three classes: (black) elastin-like, (red) presumed ambivalent; (blue) presumed amyloid. (A) Correlation between $X_{W}$, the number of water molecules bound per backbone hydrogen-bonding group, and $X_{HB}$, the hydrogen-bonded turn probability per residue, for monomers. (B) Correlation between $X_{W}$ and $X_{HB}$, the peptide-peptide hydrogen-bonding propensity, for homo-aggregates. (C) and (D). Correlation between $X_{W}$ and $X_{ppII}$, the fractional polyproline II content, for monomers and aggregates, respectively. (E) Correlations between $X_{HB}$ and $X_{G}$, the fraction of glycine content in the sequence, for explicit water (closed symbols) and in vacuo (open symbols) simulations. (F) Correlation between $S_{HB}$, the buried hydrophobic surface area relative to a fully extended chain, for hydrated multi-chain (MC) and single chain (SC) simulations. Error bars represent the standard deviation of the mean obtained by block averaging.
4.8 A PG-Composition Threshold

Based on the above results, both proline and glycine appear to play a central role in governing elastin-like properties in model sequences. The general significance of this finding is revealed in a two dimensional diagram relating the P and G contents of many natural elastomeric protein domains, as well as experimentally-confirmed amyloids (Figure 4.4). A clear separation between elastomeric and amyloidogenic sequences is apparent using a wide array of protein sequences from many different species. Although the hydrophobic domains of mammalian, amphibian, avian, and fish elastin span a wide region of the plot, the overwhelming majority of these sequences are found above a linear threshold of combined P and G content. The fact that approximately two glycines are equivalent to one proline at this threshold confirms the role of proline as the primary determinant of elastin’s properties. Remarkably, the compliance with a PG composition threshold is not limited to the hydrophobic domains of elastin, but is also observed for other elastomeric proteins. These include the elastic domains of preColP, abductin, and resilin.

Below the PG-threshold are many sequences that have been found to form amyloid fibrils in vivo or in vitro, some of which are linked to diseases and others which have no known associated pathology. Of particular interest are the peptide fragments of islet amyloid polypeptide (IAPP, associated with type II diabetes). Human IAPP contains the amyloidogenic fragment SNNFGAILSS, while rat IAPP has SNNLGPVLP at the same location in its sequence. The latter fragment, which has a combined PG composition well above the threshold, cannot form amyloid fibrils in vitro. It is thought that this sequence difference protects rats from acquiring type II diabetes.\textsuperscript{25,26} It has also been demonstrated that randomly scrambled variants of the prion domain of Ure2p can form amyloid fibrils in vivo and in vitro, emphasizing the importance of sequence composition over exact sequence in amyloid formation.\textsuperscript{27} The effects of P and G on the inhibition of amyloid fibrillization have separately been the object of previous studies.\textsuperscript{28,29} The present evidence for the existence of a quantitative PG threshold above which amyloid formation is impeded is compatible with the fact that even sequences with relatively high content of either P or G can form amyloid fibres.
Figure 4.4 Proline and glycine composition of elastomeric and amyloidogenic peptides
A two dimensional plot correlating proline and glycine content for a wide variety of peptides. The coexistence region (shaded in grey) contains P and G compositions consistent with both amyloidogenic and elastomeric properties. Elastomeric proteins, including the domains of elastin, major ampullate spidroin 2 (MaSp2), flagelliform silk, the elastic domains of mussel byssus thread, and abductin, appear above a composition threshold (upper dashed line). Amyloidogenic sequences are primarily found below the PG-threshold, along with rigid lizard egg shells, tubulliform silk (TuSp1), a protective silk for spider eggs, and aciniform silk (AcSp), used for wrapping prey. The coexistence region contains amyloid-like peptides as well as the elastomeric adhesive produced by the frog Notaden bennetti, the PEVK domains of titin, wheat glutenin protein, and the strongest spider silks, namely major ampullate spidroin 1 (MaSp1) and minor ampullate spidroin (MiSp). Our model peptides are also shown (triangles) with color coding defined in Figure 4.3. The sequences and corresponding references are published online.  

Sequences of the various classes of spider silks also appear to satisfy the PG threshold. The most elastic forms of spider silk, flagelliform silk and major ampullate spidroin 2 (MaSp2), are found above the threshold in the region populated by elastin’s hydrophobic domains, while the two forms of silk imparting strength to the dragline and the web (respectively, major ampullate spidroin 1, MaSp1, and minor ampullate spidroin, MiSp) appear below the
threshold. The rigid silks (aciniform, AcSp, used for wrapping prey, and tubuliform, TuSp1, used to encase eggs) are found in the amyloidogenic region. Similarly, flexible and rigid lizard egg shells are located respectively above and below the composition threshold, with the primary difference being a twofold increase in P content for the flexible-type shells.

The transition in composition space between elastomers and amyloids does not appear to be an abrupt one. Rather, a coexistence region includes sequences that are either elastomeric, amyloidogenic, or both. Elastomeric proteins in this region include glutenin (found in wheat gluten), the adhesive produced by the Australian frog Notaden bennetti, and the PEVK domains of muscle protein titin. There are also amyloidogenic sequences, including silk moth chorion protein and the ambivalent sequence poly-(GGVGV). The boundaries indicated in Figure 4.4 are not intended to be definite, and the delineation and characterization of the threshold will require further investigation. Accordingly, the PG diagram should not be taken to imply that random sequences satisfying the threshold in combined PG content would be elastomeric, or that the composition of proline and glycine is the sole measure of a protein’s ability to form amyloid fibrils. For a given composition, changes in primary sequence may influence hydration and peptide-peptide hydrogen bonding, and therefore aggregation tendencies. The effect of sequence variability may be especially important in the coexistence region. It should also be noted that the sequence of human elastin is not random with regard to the location of prolines and glycines (refer to Figures A.1 and A.2 in Appendix A for an analysis of the inter-proline and inter-glycine spacing of human elastin; the most common separations are five and one for proline and glycine, respectively). Additional factors, such as the compositions of other residues and solution conditions, are expected to contribute to the modulation of protein aggregation tendencies. However, it is striking that all known natural elastomeric sequences satisfy a composition threshold, which therefore appears to constitute a necessary condition for the onset of elastomeric properties.

In support of the role of proline as the primary determinant of elastomeric properties, proline content is the main sequence difference separating major ampullate silks with or without rubber-like elasticity. The major ampullate silks of orb-weavers Araneus diadematus and
*Nephilia clavipes* have nearly the same glycine content (40 % and 45 %, respectively) but significantly different proline content (16 % and 3.5 %, respectively). The composition of *Araneus* silk places it well above the PG-composition threshold, while *Nephilia* silk is near the lower boundary of the coexistence region. The differences in proline content between these two species result in opposite mechanisms of elastic recoil: *Araneus* silk has entropically-driven elastic recoil, while the elastic recoil of *Nephilia* silk is almost entirely due to internal energy. On the basis of both thermoelasticity and birefringence measurements, it is thought that the molecular structure of *Araneus* silk is highly disordered, while that of *Nephilia* silk contains energetically-stable secondary structure.39,40 In addition, based on a study of major ampullate silks from a wide range of species, increased proline content is strongly correlated with decreased stiffness, increased extensibility, and increased capacity to shrink (supercontract).41 Thus, proline content is an essential factor in determining the thermoelastic and mechanical properties of spider silk.

In support of the crucial role of glycine content in elasticity, Dicko *et al* observed that increasing glycine content correlates with increasing structural disorder for the various silks produced by *Nephilia edulis*.42 In this study, structural disorder was quantified by a folding index, defined as the ratio of CD ellipticities at 220 nm and 200 nm; the folding index is a measure of the ratio of folded to unfolded structures. Interestingly, glycine-rich silks, such as major ampullate and flagelliform, have a significantly lower folding index than glycine-poor silks, such as aciniform, piriform and cylindriform.42 These results are consistent with the view that increasing glycine content increases structural disorder, and therefore is correlated with the onset of elastomeric properties.

### 4.9 Experimental Cross-Validation

Further support for the above findings is provided by the cross-validation of independent observations made on the basis of (i) combined PG composition, (ii) structural properties obtained from atomistic simulations of model peptides, and (iii) in vitro studies of recombinant peptides containing the same repeat motifs. The structural properties computed for the model sequences (Figure 4.1 D) are consistent with the distribution of these sequences on the PG diagram (Figure 4.4). The elastin-like peptides are found above the threshold, while the
remaining sequences are found in the coexistence region or below the PG-threshold. Based on
the composition threshold in the PG diagram (Figure 4.4), poly-GVA is predicted to form
amyloid-like structures. Accordingly, in our simulations of both monomers and aggregates,
(GVA)_{12} exhibited structural tendencies similar to \((GVAGGV)_{6}\), an amyloidogenic sequence
motif,\(^{12}\) with low hydration and high peptide-peptide hydrogen bonding propensity. To test this
prediction, recombinant polypeptides that contain tandem GVA repeats were produced: EP20-
24-[GVA] and EP20-[GVA]-[GVA], where the sequence of the [GVA] domain is provided in
Figure 4.1 B. Both peptides formed fibrillar precipitates and were subjected to the thioflavin-T
binding assay for amyloid fibril detection (Figure 4.5).\(^{43}\) Thioflavin-T fluorescence for EP20-24-
24, the polypeptide with native human elastin sequence PGVGVA, was minimal and
independent of concentration. In contrast, EP20-24-24[P/G] in 0.5 M NaCl, which has been
shown previously to form amyloid-like fibrils by electron microscopy,\(^{12}\) exhibited a thioflavin-T
fluorescence which was strongly concentration dependent. Similarly, EP20-24-GVA and EP20-
GVA-GVA both exhibited strong concentration dependence for dye binding comparable to
EP20-24-24[P/G]. The confirmation that the GVA motif promotes amyloid formation
highlights the capacity of the PG diagram to accurately predict peptide aggregation tendencies.

![Figure 4.5 Thioflavin-T binding assay for amyloid fibril formation](image)

**Figure 4.5 Thioflavin-T binding assay for amyloid fibril formation**

Binding of Thioflavin-T to EP20-24-24 (♦), EP20-24-24[P/G] in 0.5M NaCl (■), EP20-24-
GVA (▲) and EP20-GVA-GVA (●). It has previously been demonstrated that EP20-24-
24[P/G] (in the presence of salt) did not coacervate, and instead formed amyloid-like fibril
structures.\(^{12}\) EP20-24-24[P/G] contains the GGVGVA tandem repeat, which is found below
the threshold on the PG diagram. The dependence of fluorescence on peptide concentration for
both GVA recombinant peptides is similar to that of this known amyloidogenic sequence.
4.10 Towards a Unified Model of Elastomeric Structure and Function

The existence of a threshold of combined proline and glycine composition observed by elastomeric proteins points to a remarkably simple and robust design principle, whereby the onset of elastomeric function is controlled by the intrinsic backbone properties conferred by just two amino acid residue types. The analysis of the results obtained from our minimalist set of peptide sequences indicates that both amino acids help keep the polypeptidic backbone of elastomers disordered and hydrated, though for opposite reasons: prolines, because they are too stiff to form secondary structure, and glycines, because in the presence of water they are too flexible to do so. This finding is consistent with the fact that these are the two residue types most likely to be found in loops, most of which are solvent exposed, rather than in the secondary structure of globular proteins.44

Furthermore, the interplay between hydration and disorder of the polypeptide chain is highlighted by the apparent threshold in backbone hydration separating elastin-forming aggregates from their amyloid-forming counterparts (indicated as a dotted line in Figure 4.3 A to D). In support of the existence of a hydration threshold, not only is elastin brittle when dry, but solid-state NMR has also shown that a relative mass of water greater than 30% is required for the onset of conformational flexibility.45 Converting the apparent threshold of $X_W = 0.7$ (the average number of water molecules bound to each backbone hydrogen-bonding group) into a percent mass of water leads to the following results for elastin-like sequences: PGVPGV, 27%; GGVPGV, 29%; GVPGV, 28%; GVAPGV, 32%. The agreement with the experimentally-determined hydration threshold suggests that the emergence of elastomeric properties of elastin is controlled by backbone hydration.

Taken together, the above findings have general implications for elastomeric self-assembly and function (Figure 4.6). Two major entropic forces are at play in the folding and aggregation of biopolymers: polypeptide chain entropy ($\Delta S_C$) and hydrophobic packing ($\Delta S_{HP}$). Chain entropy opposes both folding (Figure 4.6 C) and full extension of the polypeptide chain (Figure 4.6 A) since both events dramatically decrease the number of accessible conformations. Hydrophobic forces drive the emergence of collapsed states of polypeptide chains (Figure 4.6 B),
Figure 4.6 Role of entropy in polypeptide self-organization
Schematic diagram of possible protein states, highlighting the polypeptide backbone (black), non-polar side-chains (yellow), water molecules solvating the polypeptide backbone (blue), and peptide-peptide H bonds (red). (A) extended single chain; (B) disordered, water-swollen monomer representing the unfolded state of a globular protein; (C) folded globular protein; (D) extended aggregate; (E) hydrated amorphous aggregate characterized by an ensemble of many degenerate conformations; (F) β-sheet of amyloid fibres. Both protein folding and amyloid formation are favored by hydrophobic forces ($\Delta S_{HP}$) but opposed by the polypeptide chain entropy ($\Delta S_C$). Both forces oppose the extension of disordered hydrophobic chains. The present study suggests that the single most important feature of elastomeric chains is their inability to fold into a native structure or to form extended cross-β sheets, which allows their aggregates to remain amorphous and to readily undergo extension and elastic recoil.
which precede formation of the hydrophobic core of globular proteins⁴⁶ (Figure 4.6 C). Like the native state of globular proteins, the molecular structure underlying amyloid fibrils is characterized by an ordered, water-excluding core containing extensive secondary structure (Figure 4.6 F). The present work suggests that the reason elastomeric chains remain hydrated and disordered even after aggregation is that their backbone is inherently unable to form extensive self-interactions (Figure 4.6 E). Accordingly, the functional state of an elastomer is incompatible with the amyloid state. In turn, this supports a model of elastomeric materials in which, consistent with (a) rubber-like elasticity and (b) the existence of a hydration threshold, backbone entropy and hydration play a central role in their function. Because the aggregated state contains relatively few hydrogen-bonded self-interactions, the polypeptide chains can readily extend under strain (Figure 4.6 D). Subsequently, both chain entropy and hydrophobic packing contribute to elastic recoil.

4.11 Conclusion

We derive the following simple conclusions: (1) a fundamental requirement for elastomeric domains is to remain disordered, even when aggregated; (2) disorder is an indirect consequence of the inability of the polypeptide to form a compact, water-excluding core involving extensive backbone self-interactions; (3) most significantly, the fact that a minimum threshold of combined P, G content appears to be fulfilled by proteins forming such diverse biomaterials as human aorta, spider silk, and lizard egg shells suggests that maintaining a critical level of disorder is not only a fundamental requirement, but may very well constitute the single most essential design principle of self-assembling elastic proteins. These insights establish a framework for more detailed studies of the physical determinants of structure in elastomers and amyloids, as well as the modulation of elastomeric properties. Ultimately, this understanding will help advance the rational design of self-assembling biomimetic materials, such as artificial skin, and facilitate the development of therapeutic approaches to combat the increasingly prevalent amyloid diseases.
4.12 Experimental Procedures

4.12.1 Molecular Dynamics

Blocked single-chain and tetrameric assemblies of each of the ten periodic polypeptide sequences depicted in Figure 4.1 D were subjected to extensive MD simulations in explicit water at 300K using the OPLS/AA/L force field\(^{47,48}\) and the TIP3P model for water.\(^{49}\) Simulations of monomeric and aggregated systems were performed for 60 ns and 20 ns, respectively, for a total of 800 ns of simulation, using the GROMACS MD simulation package.\(^{50,51}\) The monomeric and aggregated peptides were solvated in a parallelepipedic box of 35 x 35 x 65 and 40 x 40 x 80 Å\(^3\), respectively, except for aggregates of (GVGVA)_7 and (GVAGVA)_6 (30 x 80 x 50 and 35 x 45 x 115 Å\(^3\), respectively). Periodic boundary conditions were applied and a 1.4 nm cutoff was used for Lennard-Jones interactions. Covalent bonds involving hydrogen atoms were constrained with the SHAKE algorithm.\(^{52}\) Long-range electrostatics interactions were handled using the Particle Mesh Ewald (PME) summation method\(^{53,54}\) with a Fourier spacing of 0.15 nm and a fourth order interpolation. All simulations were performed in the NVT ensemble at 300 K. Peptide and solvent were coupled to the same reference temperature bath with a time constant of 2 ps using the Nosé-Hoover method.\(^{55,56}\) An integration step size of 2 fs was used and coordinates were stored every 10 ps.

The initial conformation of the peptides consisted of fully extended chains in vacuo. Homo-aggregates were constructed by placing four polypeptide chains side-by-side and parallel to each other at 10 Å intervals on a plane. To reduce the size of the water box to a tractable size, prior to immersion in explicit water the peptides were subjected to preliminary simulations for a few nanoseconds using the CHARMM simulation package\(^{57}\) in implicit solvent and in vacuo for monomers and aggregates, respectively. By favouring intermolecular hydrogen bonds between peptide groups, this procedure biased the initial conformation of hydrated aggregates towards the formation of β-sheets.

4.12.2 Structural Properties

Structural properties were computed from the last 50ns of the simulations for the monomers and the last 10ns of the simulations for the aggregates. The analysis of the data
accumulated in the trajectory files was performed using an in-house script based on a modified version of the Dictionary of Secondary Structure in Proteins (DSSP). For each snapshot, possible backbone hydrogen bonds were evaluated using both (a) the energetic criterion of DSSP and (b) the following geometric criteria: (i) donor-acceptor and H-acceptor distances are less than 3.5 and 2.5 Å, respectively; and (ii) the value of the acceptor-donor-hydrogen angle is less than 60˚. Because the polypeptide chains studied are very flexible, these geometric criteria were necessary in order to avoid false positives in the identification of hydrogen bonds.

Based on the positions of hydrogen bonds, turns, bridges, α-helices and β-ladders were identified. We assign an n-turn at residues i and i+n if there is a hydrogen bond between carbonyl group i and amine group j = i+n (see Figure 4.2 B); we also considered reverse turns defined by a hydrogen bond between amine group i and carbonyl group j = i+n. There is a bridge between residue i and j if there are two hydrogen bonds characteristic of β-structure. α-helices and β-ladders are defined by at least two consecutive n-turns or bridges respectively. Because a given residue can simultaneously be found in different local structures, it is necessary to define a hierarchy. To avoid misassigning short turns as bridges, we inverted the order of these two types of structures relative to the DSSP hierarchy (see Figure 4.7). Refer to Appendix C for a detailed discussion of hydrogen bonding criteria.

Unlike the other types of structure, polyproline II (PPII) does not involve peptide-peptide hydrogen bonds and was therefore assigned solely on the basis of geometric criteria. Deviations of ±20˚ from the canonical (φ,ψ) values of a perfect ppII helix (-75˚,145˚) were allowed.

The degree of hydration of the backbone, $X_W$, was computed as the average number of hydrogen bonds between water molecules and each carbonyl or amine group of the peptides. The hydrophobic solvent accessible surface area was calculated using the program g_sas included in the GROMACS package. Since some of the sequences studied correspond to extensible peptides, the buried hydrophobic surface area, $A_{HP}$, was computed relative to the completely extended configuration of the peptide.
**Figure 4.7 Assignment of local structure**

(I) Example of hydrogen bond with the specification of the angle $\theta$. The determination of the position of (a) hydrogen bonds is crucial in the assignment of the positions of (b) the turns and the bridges which will form (c) $\alpha$-helices and $\beta$-ladders. (II) Hierarchy used to attribute the final local structure to a given residue.

### 4.12.3 Recombinant Polypeptides

Construction and production of elastin polypeptides EP20-24-24 and EP20-24-24[P/G] has been described elsewhere (Miao et al., 2003; Bellingham et al., 2001). EP20-24-GVA and EP20-GVA-GVA were constructed by replacing oligonucleotides coding for 11 repeats of GVA with the 7 repeats of PGVGVA in one or two exon24 of EP20-24-24, respectively. All polypeptides were confirmed by amino acid analysis and Q-TOF mass spectrometry using the facilities of the Advanced Protein Technology Centre, Hospital for Sick Children.

### 4.12.4 Thioflavin-T Assay

The assay for binding of thioflavin-T was performed according to the method of LeVine.$^{43}$ Briefly, polypeptides were dissolved in water or 0.5M NaCl at a concentration of 0.5 $\mu$g/$\mu$L and left at 4°C overnight. Then 10 $\mu$L of polypeptide was incrementally added to 2 ml of 3 $\mu$M thioflavin-T in potassium phosphate buffer (50 mM, pH 6.0). Fluorescence at 482 nm (excitation 450 nm) was monitored using a Hitachi F-2500 fluorescence spectrophotometer.
References


CHAPTER 5

Simulated Tempering Distributed Replica Sampling, Virtual Replica Exchange, and Other Generalized-Ensemble Methods for Conformational Sampling

The contents of this chapter were adapted with permission from an article published in the Journal of Chemical Theory and Computation.

Reference:

Contributions:
S.R. was the primary author of the manuscript, implemented all methods, and performed all simulations and analysis. C. N. invented the technique “virtual replica exchange,” provided specific guidance on its implementation, and performed a critical reading of the manuscript. R.P. provided significant editorial input and guidance throughout.
5.1 Summary

Generalized-ensemble algorithms in temperature space have become popular tools to enhance conformational sampling in biomolecular simulations. A random walk in temperature leads to a corresponding random walk in potential energy, which can be used to cross over energetic barriers and overcome the problem of quasi-nonergodicity. In this chapter, we introduce two novel methods: simulated tempering distributed replica sampling (STDR) and virtual replica exchange (VREX). These methods are designed to address the practical issues inherent in the replica exchange (RE), simulated tempering (ST), and serial replica exchange (SREM) algorithms. RE requires a large, dedicated and homogeneous cluster of CPUs to function efficiently when applied to complex systems. ST and SREM both have the drawback of requiring extensive initial simulations, possibly adaptive, for the calculation of weight factors or potential energy distribution functions. STDR and VREX alleviate the need for lengthy initial simulations, and for synchronization and extensive communication between replicas. Both methods are therefore suitable for distributed or heterogeneous computing platforms. We perform an objective comparison of all five algorithms in terms of both implementation issues and sampling efficiency. We use disordered peptides in explicit water as test systems, for a total simulation time of over 42 µs. Efficiency is defined in terms of both structural convergence and temperature diffusion, and we show that these definitions of efficiency are in fact correlated. Importantly, we find that ST-based methods exhibit faster temperature diffusion and correspondingly faster convergence of structural properties compared to RE-based methods. Within the RE-based methods, VREX is superior to both SREM and RE. Based on our observations, we conclude that ST is ideal for simple systems, while STDR is well-suited for complex systems.

5.2 Introduction

Achieving complete (or even adequate) conformational sampling is one of the key challenges in biomolecular simulations. The energy landscape of most biomolecules is “rugged” and the source of this ruggedness is two-fold. The energetic barriers separating accessible states are often larger than the available thermal energy, and there are typically a large number of states to be sampled. The timescales of many biomolecular processes, such as protein folding, are still
far beyond the reach of our current computational capability, which is generally limited to the $10^{-8}$-$10^{-7}$ s timescale for continuous simulations. For example, even the folding of small domains or secondary structure elements, such as $\beta$-hairpins and mini-proteins, occur on the 1-10 $\mu$s timescale. Consequently, conventional or “brute force” molecular dynamics (MD) alone is often insufficient to achieve complete Boltzmann sampling of the important states of many biologically relevant systems. For this reason, generalized-ensemble algorithms have become popular tools for conformational sampling.

A variety of generalized-ensemble algorithms have been developed with the common intention of overcoming energetic barriers in order to enhance sampling of conformational space. These methods use a generalized Hamiltonian for the purpose of achieving uniform sampling along a reaction coordinate of interest. Practically, one is faced with choosing the most appropriate method and reaction coordinate for a particular application. While the optimal reaction coordinate is not known a priori, it may be possible to make generalizations regarding the optimal methodology. To this end, we consider the following important question: given limited computational resources, which algorithm is most efficient at sampling a complex energy landscape? Some generalized-ensemble methods employ a random walk in potential energy, while others use different parameters which are relevant to the system of interest. In this chapter, we compare the efficiency of a set of algorithms which make use of a random walk in temperature to enhance conformational sampling of biomolecules. We focus on the following five methods: simulated tempering (ST), replica exchange (RE), the serial replica exchange method (SREM) and two novel methods: virtual replica exchange (VREX) and simulated tempering distributed replica sampling (STDR), which is a combination of ST and distributed replica sampling (DR).

The generalized-ensemble algorithms compared in this chapter all rely on the fact that the free energy surface becomes less rugged at high temperature, increasing the frequency of interconversion between conformational states. Simulations performed at low temperature often require a relatively long time to cross the energetic barriers between states and appear to be trapped. Transitions between regions separated by barriers may not be observed over timescales
accessible to simulation. In this case, multiple simulations initiated in different conformational basins may sample different subsets of phase space. The result is that an ergodic system appears nonergodic, a phenomenon known as quasi-nonergodicity. Utilizing generalized-ensemble algorithms that induce a random walk in temperature may alleviate this source of error.

The sampling enhancement of generalized-ensemble methods relative to canonical MD or Monte Carlo (MC) simulations has been demonstrated for several systems, including peptides. Conversely, there have also been studies that question the relative sampling efficiency of RE compared to brute force MD, highlighting the importance of a rigorous definition of efficiency which accounts for the total computer time required for all temperatures. It is important to note that data obtained at multiple temperatures in generalized-ensemble simulations may be of interest in some studies, such as protein folding. In general, however, the data at high temperature is not useful. Furthermore, the observed speed-up also strongly depends on the lowest temperature. It is essential to assess the convergence of both the conventional MD simulations as well as the generalized-ensemble simulations in order to perform a meaningful comparison, in addition to identifying a meaningful quantity on which to base the comparison. Any evaluation of sampling enhancement compared to single-temperature MD is also likely to depend heavily on the molecular system under study (depending on the number of basins in the landscape and the heights of barriers). It is therefore quite difficult to accurately quantify the sampling enhancement due to the introduction of a random walk in temperature.

We begin with a brief introduction of each of the generalized-ensemble methods, including the presentation of our two novel methods, STDR and VREX. We then perform a thorough comparison of the algorithms in terms of both practical implementation limitations and sampling efficiency for a disordered octapeptide in explicit water, a molecular system combining high relevance to protein folding and moderate complexity. In addition to providing a comparison between generalized-ensemble algorithms, we also provide a comparison to conventional MD. We discuss efficiency in terms of both convergence of structural properties and temperature diffusion, and we show that these definitions of efficiency are correlated.
Finally, we compare the efficiency of STDR and conventional MD for a 35-residue peptide with a complex conformational landscape.

5.3 Theory and Methods

5.3.1 Simulated Tempering (ST)

Simulated tempering was originally introduced to enhance sampling of a Random Field Ising Model. This system has a rough energy landscape for which spin flips from the state favored by the magnetic field to the opposite state are statistically rare events. ST facilitates exchanges between these states, whereas the MC algorithm remains trapped. ST has also been shown to be effective in exploring the energy landscapes of biomolecules, which similarly have multiple energy minima separated by barriers.

In the ST algorithm, temperature becomes a dynamic variable that can take on discrete values labeled by an index \( m \) (\( m = 1, \ldots, M \)). ST makes use of a generalized Hamiltonian, \( H(X,m) \), which depends on all configurational degrees of freedom \( (X) \), in addition to temperature:

\[
H(X,m) = \beta_m H(X) - a_m, \quad (5.1)
\]

where \( \beta_m \) is the inverse temperature, \( H(X) \) is the system’s original Hamiltonian, and \( a_m \) is a constant that depends on temperature. The generalized-ensemble has a corresponding generalized partition function, \( Z \), given by:

\[
Z = \sum_m \int dX \left[ e^{-H(X,m)} \right] = \sum_m \int dX \left[ e^{-\beta_m H(X) + a_m} \right] = \sum_m Z_m e^{a_m}, \quad (5.2)
\]

where \( Z_m \) is the partition function corresponding to the temperature \( T_m \). The partition function of the generalized-ensemble, \( Z \), is the weighted sum of the partition functions of the canonical ensembles at each temperature, \( Z_m \). We therefore refer to the constants, \( a_m \), as “weight factors.” The probability of sampling a given temperature, \( T_m \), is:
which depends on the generalized Hamiltonian, $H$, and therefore depends on the weight factor, $a_m$. The goal in ST is to perform a random walk in temperature such that all temperatures are visited uniformly, that is, to choose weight factors such that for any two temperatures (labeled $i$ and $j$):

$$Z_i e^{a_i} = Z_j e^{a_j}.$$  \hspace{1cm} (5.4)

Since the partition function in the canonical ensemble, $Z_m$, is related to the Helmholtz free energy, $A_m$, the optimal weight factors are dimensionless Helmholtz free energies (the Helmholtz free energy multiplied by the inverse temperature, $\beta$):\textsuperscript{50,31}

$$Z_m = e^{-\beta \cdot A_m} = e^{-a_m}.$$  \hspace{1cm} (5.5)

The use of accurate dimensionless Helmholtz free energies as weight factors leads to sampling all temperatures with equal probability. In principle, the weight factors may take any value without resulting in biased, non-Boltzmann sampling at the individual temperatures. However, inaccuracy in the weight factors leads to corresponding differences in the probabilities of sampling at each temperature.\textsuperscript{4,32}

An ST simulation consists of a short canonical MD (or MC) simulation at temperature $T_i$ followed by an exchange attempt to a neighboring temperature, $T_j$. The probability of this exchange occurring is given by:

$$p(T_i \rightarrow T_j) = \min \left\{ \frac{1}{e^{-(\beta_j - \beta_i)E_a(a_j - a_i)}}, 1 \right\},$$  \hspace{1cm} (5.6)
where $E$ is the potential energy of the system at the end of the previous simulation at temperature $T_i$, and $\beta_i$ and $\beta_j$ are the inverse temperatures. The weight factors need only be accurate up to an additive constant, since only differences in weight factors are required to determine the acceptance probability. Through many repetitions of these alternating simulation and exchange steps, a random walk in temperature is realized, corresponding to a random walk in potential energy and efficient exploration of the energy landscape. In fact, ST has been shown to be as effective as the multicanonical algorithm (MUCA), which employs a random walk in potential energy.

The underlying challenge in ST is accurately obtaining the dimensionless Helmholtz free energies, $a_m$. There have been two general approaches to their calculation. The first method involves making use of the Weighted Histogram Analysis Method (WHAM) to obtain the density of states and the weight factors. The second method, which we utilize throughout this thesis, was recently proposed as a fast and efficient scheme to obtain an accurate estimate of the weight factors based on average energies. The average potential energy at each temperature, $\langle E \rangle$, is obtained from initial simulations, and the differences in weight factors are calculated as follows:

$$a_{i+1} - a_i \approx (\beta_{i+1} - \beta_i) \left( \frac{\langle E \rangle_i + \langle E \rangle_{i+1}}{2} \right).$$

The weight factor for the lowest temperature can be set to zero since only differences in weight factors are needed in the exchange probability. The replica exchange simulated tempering method (REST) may also be used to obtain weight factors. In this method, an initial RE simulation is run for the purpose of obtaining accurate weight factors, which are then used in an ST simulation. REST may be used with either WHAM or the method outlined in equation 5.7. Weight factors may be updated throughout the ST simulation if required. Adaptive determination of weights using schemes like the one developed by Zhang and Ma based on the Wang-Landau algorithm were shown to be useful for a square lattice Ising model, a bead model of a protein and a Lennard-Jones fluid. Biomolecular systems containing thousands of atoms...
represent a completely different level of complexity. Convergence of weights is non-trivial\textsuperscript{14,30,31,39} and can be slow even with adaptive approaches.\textsuperscript{30} Calculating weight factors has been the main impediment to the widespread use of the simulated tempering method.\textsuperscript{30}

The accuracy of the weight factors (that is, how close the differences in weight factors are to accurate dimensionless Helmholtz free energy differences) can be assessed by computing the deviation from sampling all temperatures homogeneously in a sufficiently long ST simulation. In the extreme case for which all weight factors are equal and all differences in weight factors are zero, only the lowest temperature is significantly sampled. This is because the first term in the exponent of the exchange probability (equation 5.6) depends on the potential energy, which is generally a large, negative number for biomolecular systems. When multiplied by the difference in inverse temperatures, the resulting exchange probability dictates that moves to lower temperature are accepted, while moves to higher temperature are rejected. Conversely, if the differences in weight factors are equal to the differences in dimensionless Helmholtz free energies, the temperatures in the ST simulation are sampled uniformly, which is the optimal situation. In practice, weight factors obtained for ST result in temperature sampling inhomogeneity somewhere between these two extremes. Calculating the dimensionless Helmholtz free energies for a complex system such as a peptide in explicit water is computationally expensive since it requires an accurate estimate of the partition function. These calculations can require tens of nanoseconds per temperature or more and the computational expense increases with both system size and complexity.\textsuperscript{14}

5.3.2 Replica Exchange (RE)

Replica exchange has been the most widely used of the methods we discuss in this chapter to enhance sampling of biomolecular simulations. It can be thought of as a parallel version of ST, and it is also known as parallel tempering\textsuperscript{5} or multiple Markov chains.\textsuperscript{8} In fact, parallel tempering was applied to proteins even before ST.\textsuperscript{41} An RE simulation consists of $M$ identical copies of the system (replicas) which sample $M$ canonical ensembles at different temperatures. Exchanges are performed between neighboring temperatures, $T_i$ and $T_j$. The
probability of making an exchange depends on the potential energies, $E_i$ and $E_j$, and the inverse temperatures, $\beta_i$ and $\beta_j$:

$$P(T_i \leftrightarrow T_j) = \min \left\{ 1, e^{-\left(\beta_j - \beta_i\right)(E_i - E_j)} \right\}.$$  \hspace{1cm} (5.8)

RE is analogous to ST, but instead of using weight factors in the exchange probability, the upward move of one replica is coupled to the downward move of another. RE therefore has the critical advantage of not requiring any initial simulation for the calculation of weight factors. Importantly, it also satisfies detailed balance.\(^2\)

One drawback of the RE method is its significant computational requirements. There is a one-to-one correspondence between the number of replicas ($M$) and the number of temperatures ($M$). The number of replicas needed for an RE simulation is related to the number of degrees of freedom, $N$, as $O(N^{1/2})$.\(^2,\(^7,\(^42\) Systems with many particles therefore require many replicas. Although it is not a specific requirement of the RE algorithm, in its typical implementation, each replica is run on a dedicated central processing unit (CPU). This setup minimizes the amount of information that must be passed between nodes.\(^3\) Thus, $M$ CPUs are running simultaneously throughout the course of the RE simulation. The use of $M$ CPUs in RE can be overcome by running multiple replicas per CPU. However, using one CPU for multiple replicas does not effectively take advantage of the parallelization inherent in the RE method.

The RE algorithm requires the synchronization of attempted moves, which results in wasted CPU time if any replica waits for other replicas to perform exchanges. Inhomogeneity of CPU speeds affects the amount of wasted time, since the speed of the calculation depends on the speed of the slowest processor. Modified versions of RE have been developed in an effort to minimize wasted CPU time, including the multiplexed replica exchange method (MREM)\(^23\) and asynchronous replica exchange.\(^43\) MREM makes use of multiplexed layers of replicas ($n$ layers, each with $M$ temperatures), with exchanges occurring both within and between layers.\(^23\) MREM is even more computationally demanding than RE, using $n$ times as many processors. MREM
does not offer a significant advantage if there is a shortage of CPUs, but it does offer a way of using more CPUs without adding more temperatures. In asynchronous replica exchange, only the replicas undergoing exchange are synchronized, therefore increasing efficiency on heterogeneous computing platforms. More complex replica management schemes have also been proposed to increase the efficiency of RE. However, modified RE algorithms do not completely alleviate the need for synchronization and frequent communication between replicas. This is especially important to users of distributed computing, such as the massively parallel Folding@Home project, who must contend with inhomogeneity of processor speeds.

5.3.3 Serial Replica Exchange (SREM)

The serial replica exchange method was recently developed to address the main practical limitations inherent in the RE method, namely the need for synchronization and a large number of processors. The exchange probability in SREM has an identical form to that of RE (equation 5.8) for a replica at temperature $T_i$ attempting to move to a neighboring temperature $T_j$:

$$P(T_i \rightarrow T_j) = \min \left\{ \frac{1}{e^{(\beta_i - \beta_j)(E_j - E_i,\text{PEDF})}} \right\}.$$  \hspace{1cm} (5.9)

Unlike RE, the attempted move from $T_i$ to $T_j$ does not simultaneously involve another replica moving from $T_j$ to $T_i$. In SREM, the potential energy, $E_{i,\text{PEDF}}$, does not come from another replica at temperature $T_i$, but rather is selected at random from a potential energy distribution function (PEDF) for that temperature. The PEDFs are determined through initial simulations at each temperature, which may use either constant-temperature MD or RE. These initial simulations can be very computationally demanding for biomolecular systems. For example, to obtain converged PEDFs for a small RNA hairpin, approximately 100 ns per temperature were required. PEDFs may also need to be updated throughout the course of the SREM simulation. SREM also cannot be applied to temperature-dependent force fields.
In terms of practical implementation, SREM offers the same advantages as ST. In both methods, there is absolutely no communication required between independent simulations. Neither method requires a fixed number of CPUs and there is no wasted CPU time in the synchronization of attempted exchanges. In principle, both ST and SREM can be run on a single CPU. SREM also presents the same critical challenge as ST: an initial simulation is needed to determine PEDFs, the length of which is highly dependent on system complexity. The significant computational cost of calculating accurate PEDFs is a key drawback of SREM, since an SREM simulation can only be considered to be approximately correct (in terms of obeying detailed balance) if unconverged or incorrect PEDFs are used.\textsuperscript{10,14} In contrast, the weight factors of ST can deviate from the accurate dimensionless Helmholtz free energies and still yield correct results.\textsuperscript{3,10,14}

5.3.4 Virtual Replica Exchange (VREX)

The first novel method we propose, virtual replica exchange, is based on the principles of both RE and SREM. A replica at temperature $T_i$ attempts a move to temperature $T_j$, with the probability of exchange given by the following equation:

$$P(T_i \rightarrow T_j) = \min \left\{ \frac{1}{e^{-(\beta_j - \beta_i)(E_j - E_i)}}, 1 \right\}.$$  \hfill (5.10)

Here, the potential energy, $E_{j,\text{virtual}}$ comes from a list of stored energy values obtained at temperature $T_j$. This is analogous to exchanging with a potential energy value selected from a PEDF in SREM, or the current potential energy of a replica at temperature $T_j$ in RE. Like SREM, only a move from temperature $T_i$ to temperature $T_j$ occurs, with no simultaneous reverse move. In VREX, an energy value that occurred at temperature $T_j$ in the past is used, and following the attempted exchange, the occurrence of this energy value is removed from the potential energy list. This constitutes a “virtual exchange”.

VREX is derived to obey detailed balance in a similar manner to RE.\textsuperscript{2,14} Consider a state A in which a replica with configuration X is at temperature $T_i$ and a virtual replica with
configuration \(X_{\text{virt}}\) is at temperature \(T_i\). An exchange is attempted to a state \(B\) in which the replica with configuration \(X\) is at temperature \(T_j\) and the virtual replica is at temperature \(T_i\). The detailed balance criterion is written:

\[
P(A)P(A \rightarrow B) = P(B)P(B \rightarrow A). \tag{5.11}
\]

Transition probabilities that satisfy this criterion can then be derived as follows:

\[
\frac{P(A \rightarrow B)}{P(B \rightarrow A)} = \frac{\frac{e^{-\beta_j H(X)}}{Z_j}}{\frac{e^{-\beta_i H(X)}}{Z_i}} = e^{-(\beta_j - \beta_i)(E(X) - E(\text{sim}))}. \tag{5.12}
\]

This transition probability is equivalent to the Metropolis criterion in equation 5.10. It also resembles the transition probability for RE (equation 5.8). The VREX algorithm is completely analogous to RE, except that one replica undergoes a virtual reverse move. Similarly, SREM also involves a virtual reverse move, in this case by a replica whose energy distribution is represented by the PEDF. In contrast, the RE algorithm involves two real replicas undergoing temperature moves.

In practice, VREX requires very short initial simulations in order to generate a preliminary list of energies for each temperature. These lists are then updated as the simulation progresses, with values being added from each short MD simulation between exchange attempts, and values being removed as they are used in virtual exchanges. It is possible to run out of potential energy values in the primary lists if temperatures are sampled heterogeneously. In order to address this possibility, implementations of VREX may include the use of secondary lists, to which potential energies from the primary lists are moved after a single use. Potential energies from a secondary list may be used in the rare case that the primary list for that temperature is completely used. Further, recent values can be prioritized in the primary lists,
and relatively short and continually overwritten secondary lists can be maintained in order to reduce the likelihood of using pre-equilibration potential energies in post-equilibration virtual exchanges.

The main advantage of VREX is that it avoids the need to calculate converged PEDFs (like SREM) or weight factors (like ST), and only requires a short list of potential energies to begin sampling. It also addresses the main shortcoming of RE because it completely eliminates the synchronization between replicas, as well as the need for a fixed number of replicas. It is theoretically very similar to RE, with the addition of a variable time delay between the time when a potential energy is produced and when it is used for an exchange.

5.3.5 Distributed Replica Sampling (DR)

Distributed replica sampling\textsuperscript{11} is a general scheme for Boltzmann sampling of conformational space in which multiple replicas undergo a random walk in a reaction coordinate of interest. Individual replicas are coupled through a generalized Hamiltonian containing a potential energy term that depends on the distribution of all replicas, which acts to enforce a desired sampling distribution of the reaction coordinate. DR can therefore be used to enforce uniform sampling along a reaction coordinate of interest. This may be, for instance, a nonphysical spatial “fourth” dimension\textsuperscript{12} or a dihedral angle.\textsuperscript{13} We briefly summarize the DR algorithm with temperature as the coordinate.\textsuperscript{11} The implementation of DR in other coordinates has also been previously outlined.\textsuperscript{11-13}

The generalized Hamiltonian of DR in temperature contains a pseudo-energy term and depends on the current inverse temperature and current configuration ($q$) of all replicas:

$$H(q_1, \beta_1, q_2, \beta_2, \ldots, q_M, \beta_M) = \sum_{m=1}^{M} \beta_m E(q_m) + DRPE(\beta_1, \beta_2, \ldots, \beta_M),$$  \hspace{1cm} (5.13)

where $E$ is the potential energy. There are $M$ replicas in total, each labeled by an index $m=1,\ldots,M$. The distributed replica potential energy (DRPE) can take any functional form that
depends on the distribution of replicas and fulfills the purpose of enforcing homogeneous sampling of the temperature coordinate. Importantly, although the DRPE is a pseudo-energetic penalty, it is not a function of system complexity.\textsuperscript{11} The probability of a replica currently at temperature $T_i$ exchanging to a temperature $T_j$ is:

$$p(T_i \rightarrow T_j) = \min \left\{ \frac{1}{e^{(\beta_j - \beta_i)E(q_j) - (\text{DRPE}_j - \text{DRPE}_i)}}, \right\}$$

which depends on the difference between the DRPE with the replica at temperature $T_j$ ($\text{DRPE}_j$) and at temperature $T_i$ ($\text{DRPE}_i$).\textsuperscript{11} DR can be analogously used to achieve a random walk in a parameter of the Hamiltonian, $\xi$, with an exchange probability:

$$p(\xi_i \rightarrow \xi_j) = \min \left\{ \frac{1}{e^{(-\beta)((H(q,\xi_j) - H(q,\xi_i)) + (\text{DRPE}_j - \text{DRPE}_i))}}, \right\}$$

DR was designed specifically to suit shared or distributed computing platforms.\textsuperscript{11} In contrast to RE, in which pairwise exchanges of replicas are attempted, DR considers stochastic moves of individual replicas one at a time. The stochastic move of one replica is coupled to the distribution of all other replicas through the DRPE and no direct communication between replicas is required. In DR, synchronization of exchange attempts is therefore completely eliminated, which results in 100 % CPU utilization.\textsuperscript{11} The algorithm also readily accommodates fluctuations in CPU availability.\textsuperscript{11} DR in combination with thermodynamic integration (TI) was shown to sample conformational space more effectively than TI alone in the calculation of the binding free energy of benzene to T4 lysozyme, while simultaneously optimizing the use of available computational resources.\textsuperscript{12} This approach was also successfully employed to compute partial water occupancy in the pathway of proton uptake in cytochrome $c$ oxidase.\textsuperscript{50} In addition, DR has been combined with umbrella sampling (DRUS) to allow equilibrium exchange between different umbrella biasing potentials.\textsuperscript{13,50} When applied to alanine dipeptide, umbrella sampling alone exhibited quasi-nonergodic behaviour, while DRUS alleviated this systematic error.\textsuperscript{13}
Application of the DRPE restores sampling homogeneity of temperature only when the DRPE contribution is large enough to balance the preference for sampling the lowest temperature. When ST is conducted with all weight factors equal, all replicas migrate with a strong preference for the lowest temperature (as described above). In this case, a very strong DRPE is required to achieve sampling homogeneity (see equation 5.14). However, it has been demonstrated that as the energetic penalty of the DRPE becomes stronger, replica mobility (as measured by acceptance ratio) decreases,\textsuperscript{11} and therefore some modification to the DR exchange probability is necessary. This issue can be addressed by simply adding weight factors to the exchange probability, analogous to the weight factors, $f$, in the DRUS exchange probability:\textsuperscript{13}

$$p\left(\xi_i \rightarrow \xi_j\right) = \min \left\{ \frac{1}{e^{-\beta[H(q,\xi_j) - H(q,\xi_i)] - (f_j - f_i)\{DRPE_D}} - \left[DRPE_i\right]} \right\}.$$  

This form of the exchange probability results in good replica mobility and nearly perfect sampling homogeneity of the reaction coordinate.\textsuperscript{13} The efficiency and practical advantages of DR in other coordinates have been well established.\textsuperscript{11-13} It is therefore a central objective of this study to develop and test an implementation of DR which functions optimally in temperature space.

### 5.3.6 Simulated Tempering Distributed Replica Sampling (STDR)

Building on the success of the both the ST method\textsuperscript{3,4} and DR sampling,\textsuperscript{11-13} we have developed a new algorithm, STDR, which combines the two approaches. STDR is essentially DR implemented rigorously in temperature. The combination of these two methods was originally suggested when DR was developed.\textsuperscript{11} In STDR, approximately homogeneous sampling of a set of temperatures is enforced. The probability of accepting a move from a temperature $T_i$ to a neighboring temperature, $T_j$ is:
\[ p(T_i \rightarrow T_j) = \min \left\{ \frac{1}{e^{-(\beta_i - \beta_j)(E_{ji} + E_{ij})} - (\text{DRPE}_i - \text{DRPE}_j)} \right\} \] (5.17)

This is the same as the exchange probability from ST, with the addition of the difference in DRPE between the states for which the replica is at temperature \( T_i \) (\( \text{DRPE}_i \)) and temperature \( T_j \) (\( \text{DRPE}_j \)). The calculation of the DRPE is straightforward. Its functional form depends upon the current temperatures of all replicas as follows:

\[
\text{DRPE} = c_1 \sum_{m=1}^{M} \sum_{n=1}^{M} \left[ (\lambda_{m,\text{linear}} - \lambda_{n,\text{linear}}) - \omega(m-n) \right]^2 + c_2 \left[ \sum_{m=1}^{M} \lambda_{m,\text{linear}} - \omega \sum_{m=1}^{M} \right]^2 . \] (5.18)

Replicas are labeled by indices \( m \) and \( n \), where \( M \) is the number of replicas. The values of \( \lambda_{m,\text{linear}} \) refer to a linearly-spaced temperature coordinate. In this coordinate, the lowest temperature has \( \lambda_{m,\text{linear}} = 1 \), and the highest temperature has \( \lambda_{m,\text{linear}} \) equal to the number of temperatures. This procedure transforms the exponentially-spaced temperatures into a uniformly-spaced coordinate. The factor \( \omega \), which we introduce to the DRPE in this work, is the ratio of the number of temperatures to the number of replicas. This factor allows DR to be used with an arbitrary number of replicas. The first term in equation 5.18 introduces an energetic penalty for two replicas sampling the same temperature, while the second term introduces a penalty for an overall drift of the replicas towards high or low temperature. The second term is not essential when using DR in temperature. The constants \( c_1 \) and \( c_2 \) control the influence of the DRPE and can be tuned to enforce homogeneous sampling of temperatures as required. In the case of accurate weight factors, the influence of the DRPE only needs to be small such that values of \( c_1 \) and \( c_2 \) near zero can be used. With increasingly inaccurate weight factors, larger DRPE values are required to maintain homogeneous sampling of temperature, and this reduces the acceptance ratio to some degree. An example calculation of the DRPE using temperature as the reaction coordinate is provided in Appendix B.
If the weight factors, $a_m$, are inaccurate, ST results in uneven sampling of the temperature coordinate. As we will demonstrate, introducing the DRPE recovers homogeneous sampling. The STDR method is therefore more generally applicable than ST because it can make use of a poor estimate of the dimensionless Helmholtz free energies and still yield uniform sampling of the canonical ensembles at each temperature. STDR is equivalent to ST in the limit of one replica and is compatible with adaptive schemes for computing weight factors. Below, we show that STDR is the preferred method for systems with a complex energy landscape for which limitations on computational resources preclude obtaining sufficiently accurate estimates of Helmholtz free energies for an ST simulation.

### 5.4 Test System

For the purpose of comparing different generalized-ensemble methods, we use two related test systems, the peptides GVGVPGVG and (GVPGV)$_7$. These peptides are both based on the pentapeptide GVPGV, which is found as a repeat motif in the protein elastin. In our previous study of (GVPGV)$_7$ and other related elastin-like peptides, we observed that this peptide is intrinsically disordered, having many conformations and no extended secondary structure in the form of $\alpha$-helices or $\beta$-sheets. Understanding the structural heterogeneity of elastin-like peptides is required to elucidate the structure-function relationship of elastin, for which experimental characterization is notoriously difficult due to its flexibility and insolubility. The peptide GVGVPGVG has also been studied previously and was suggested to exhibit an “inverse temperature transition” with an increased probability of “closed” conformations (in which the N- and C-termini are closer than 8 Å) at higher temperatures. Motivated by this work, we select the octamer as a simple yet appropriate peptide to study in the aim of understanding the temperature-dependent behaviour of elastin. Because the main focus of this chapter is the thorough comparison of generalized-ensemble methods using these peptides as test systems, we do not elaborate fully on the structural details of either the octapeptide or the 35-residue peptide in this chapter. A description of the conformations of these peptides is the subject of Chapter 7. Both GVGVPGVG and (GVPGV)$_7$ are valuable test systems because of their structural complexity and the fact that they represent a real scientific problem in the sense that they are not well understood or characterized a priori. Simple test systems are often used for
comparison purposes, such as alanine dipeptide,\textsuperscript{10,13} although generalized-ensemble methods are typically applied to systems which are much larger and more complex. While simple test systems are useful for the sake of demonstration and for the elucidation of major problems, they are less likely to detect the subtleties and practical issues experienced when studying systems of biologically-relevant complexity.

The conformational landscape of the octapeptide is complex, with many energetically accessible states that must be sampled in order to accurately compute free energies. A representative selection of these conformations is shown in Figure 5.1 A, with “closed” states in which the N and C termini are in close proximity, “hairpin”-like states, and extended structures. Although it is a short peptide, GVGVPGVG represents a challenging sampling problem due to the large number of thermally-accessible conformations. In Figure 5.1 B, we show the hydrogen-bonding contact map for this peptide obtained using STDR. The only secondary structure consists of hydrogen-bonded turns, with no α-helix or β-sheet. The most populated turn is the VPGV β-turn, with a hydrogen bond between the C=O group of valine 4 and the N-H group of valine 7. Several other turns form with lower populations. As we will show, single-temperature MD, if run for a sufficiently long time, provides a converged description of the conformational landscape. This makes it an ideal test system because we can verify that the generalized-ensemble algorithms, given sufficient sampling, lead to correct Boltzmann-weighted sampling of conformational space, in addition to assessing their relative efficiency.

The 35-residue peptide, (GVPGV)$_7$, is used as a more complex test system to demonstrate the sampling enhancement provided by STDR for a landscape which not only has many populated states, but also has significant energetic barriers between those states. The larger system is only simulated using ST and STDR because of the extensive amount of computational resources required. Of the methods we consider, STDR is better suited to this particular application based on its performance for the octapeptide. It is as efficient and accurate as the other methods, while offering the most practical advantages for a large and complex system (see below).
Figure 5.1 The conformational landscape and hydrogen-bonding contact map of GVGVPGVG

(A) A selection of 35 random conformations from the STDR simulation at 280 K of GVGVPGVG, with glycine in red, valine in yellow and proline in blue. (B) The hydrogen-bonding contact map of GVGVPGVG at 280 K, with corresponding snapshots showing the presence of significantly populated contacts. N-H groups are on the horizontal axis and C=O groups are on the vertical axis. Each square in the matrix (i,j) corresponds to a contact between the N-H group of residue i and the C=O group of residue j. The color scheme of the legend indicates the relationship between color and probability of contact formation.
5.5 Simulation Details

For all five methods (ST, STDR, SREM, RE and VREX), the same exponentially-spaced temperature list was used. This list is provided as supplementary information in Appendix B, Table B.1. The simulation system consists of the GVGVPGVG octapeptide, capped with an acetyl group at the N-terminus and an NH$_2$ group at the C-terminus, in a 3 x 3 x 3 nm$^3$ box with 872 water molecules. The same fully-extended starting structure was used for all temperatures and all methods. Simulations were performed using the GROMACS MD simulation package, version 3.3.1,$^{54,55}$ with the OPLS-AA/L force field$^{56,57}$ for the solute and the TIP3P model for water.$^{58}$ Periodic boundary conditions were applied. The switch function of GROMACS was used for Lennard-Jones interactions, which corresponds to the usual Lennard-Jones function until 1.3 nm, after which it is switched to reach zero at 1.4 nm. Covalent bonds involving hydrogen atoms were constrained with the SHAKE algorithm.$^{59}$ Calculations of electrostatic forces utilized the Particle Mesh Ewald (PME) summation method$^{60,61}$ with a Fourier spacing of 0.15 nm and a fourth-order interpolation. The real-space Coulombic cutoff was 1.49 nm. All MD simulations were performed in the canonical ensemble. Peptide and solvent were coupled to the same reference temperature bath with a time constant of 2 ps using the Nosé-Hoover method.$^{62,63}$ An integration step size of 2 fs was used and coordinates were stored every 1 ps.

In order to compare the generalized-ensemble methods, the simulations were conducted as similarly as possible. To this end, the same total amount of simulation time (summed over all replicas) was performed. This amount was 4.75 µs, with an average of approximately 144 ns per replica. This time was used because it was sufficient for all methods to achieve statistical convergence, as shown in the results. Stochastic exchanges using the Metropolis Monte Carlo algorithm$^{64}$ were attempted every 25 ps. Exchange probabilities were calculated using equations 5.6, 5.8, 5.9, 5.10 and 5.17, as appropriate for the method. Details of the calculation of weight factors and PEDFs are discussed below. The constants $c_1$ and $c_2$ for the DRPE in equation 5.18 were both 0.005. These values were found to achieve an appropriate balance between homogeneity of temperature sampling and replica mobility.$^{11}$ The value of the factor $\omega$ was 1.0,
since the number of replicas equaled the number of temperatures. The generalized-ensemble algorithms were implemented using an in-house bash script.

The same simulation protocol was used for the simulation of \((\text{GVPGV})_7\), which was simulated in a 4.5 x 4.5 x 4.5 nm\(^3\) box with 2856 water molecules using both ST and STDR. Starting conformations and weight factors for each temperature were generated using canonical MD for 15 ns per temperature (storing 250 energy values per picosecond). Seventy temperatures were used for each generalized-ensemble simulation. The list of temperatures is provided in Appendix B, Table B.1. Temperatures were spaced more closely than those of the octapeptide. This is because it is a larger system, resulting in less overlap between potential energy distributions of adjacent temperatures for a given temperature separation. This system was simulated for a total of 8.2 µs (117.6 ns per replica on average) using the STDR algorithm. An ST simulation using the same weight factors was also performed for 420 ns (6 ns per replica on average). ST and STDR simulations were also performed using weight factors calculated based on the first 500 ps of continuous MD at each temperature for a total of 280 ns.

A conventional MD simulation of the 35-residue peptide system in the isothermal-isobaric ensemble was also performed using GROMACS version 4.0.2.\(^{65}\) In this simulation a 4 fs time step was used, and constraints on bonds and angles involving hydrogen were imposed using the LINCS algorithm.\(^{66}\) This simulation was run for 200 ns at 261 K, which corresponds to the lowest temperature in the STDR simulation. The pressure was kept constant at 1.0 bar using the Parrinello-Rahman algorithm.\(^{67}\)

The analysis of the data accumulated in the trajectories was performed using an in-house script based on a modified version of the Dictionary of Secondary Structure in Proteins (DSSP).\(^{68}\) For each snapshot, possible backbone hydrogen bonds were evaluated using both (a) the energetic criterion of DSSP and (b) the following geometric criteria: (i) donor-acceptor and hydrogen-acceptor distances are less than 3.5 and 2.5 Å, respectively; and (ii) the value of the acceptor-donor-hydrogen angle is less than 60°. Definitions of turns and bends are the same as those in DSSP.\(^{68}\) End-to-end distance (EED) is calculated as the distance between the \(\alpha\)-
carbons of the first and last residue. Root mean square deviation (RMSD) was calculated using the g_rms program in GROMACS.\(^{65}\). All molecular visualizations in this chapter were produced using VMD.\(^{59}\)

### 5.6 Calculations of weight factors for ST and STDR, PEDFs for SREM and potential energy lists for VREX

The calculation of weight factors for equations 5.6 and 5.17 required initial simulations of the octapeptide in the canonical ensemble for each of the temperatures. These simulations were performed using conventional MD for 19.5 ns (for a total simulation time of 643.5 ns). Although obtaining these accurate weight factors was resource-intensive, it involved a straightforward procedure. The weight factors were computed using the average potential energy at each temperature according to equation 5.7.\(^{30}\) The accuracy of these weight factors was assessed by using them in an ST simulation and observing the temperature sampling uniformity, as shown in results. Since all temperatures were sampled with nearly equal probability, as expected from equations 5.3 and 5.5 for accurate dimensionless Helmholtz free energies, these weight factors were deemed to be sufficiently converged and correct.

Using the same data from the conventional MD simulations, PEDFs were computed as described in the original SREM paper.\(^{10}\) The convergence of the PEDFs was assessed by calculating the \(\chi^2\) measure suggested by Hagen et al.\(^{10}\)

\[
\chi^2(t) = \sum_{n=1}^{N_{\text{bins}}} \left( P_n(t) - P_{n,\text{reference}} \right)^2. \tag{5.19}
\]

This measure computes the deviation of each bin in the current distribution, \(P_n(t)\), from a reference distribution, \(P_{n,\text{reference}}\). The current distribution is cumulative, using the data up to time \(t\). For the reference distributions, we used PEDFs computed using all of the data at each temperature. By this assessment, the PEDFs appeared to be stationary, as shown in Figure 5.2 A. When \(\chi^2\) was plotted individually for each temperature, we also observed that each PEDF was stationary. However, an initial SREM simulation using these PEDFs resulted in non-uniform
sampling of temperatures. We therefore proceeded to calculate the PEDFs using a different data set. We used the first 25 ns at each temperature of the RE simulation (for a total time of 825 ns), and these were the PEDFs used for the SREM simulation. While this procedure is similar to what would likely be done in practice with SREM, we emphasize that making this selection of PEDFs gave SREM somewhat of an advantage over ST, since more data was used in the initial simulation. The use of RE in the calculation of PEDFs is similar to the replica exchange simulated tempering method (REST). Although REST results in faster convergence of the weight factors compared to conventional MD, it may be difficult or impossible to obtain access to the required number of homogeneous and dedicated CPUs for the initial RE simulation. Thus, we did not use REST to obtain the weight factors for ST to better represent the general case where it may not be convenient to do so. In contrast, it was necessary to use RE to obtain PEDFs for SREM in a reasonable amount of time.

Figure 5.2 B shows the error in the exchange probability for both SREM and ST using the data from 19.5 ns of conventional MD at each temperature. The method for computing the error in exchange probabilities is provided in Appendix B. The weight factors of ST produce an average error in the exchange probability of less than 2 % after 19.5 ns per temperature. Using the same amount of data, the PEDFs produce a significantly higher error in the exchange probability (more than 5 %), which is why the weight factors used in ST from conventional MD produced more homogeneous sampling than the PEDFs. In Figure 5.2 C, the error in the exchange probability for both ST and SREM is shown using the data from the first 25 ns at each temperature of RE. This data set was used to calculate the PEDFs for the SREM simulation, producing an error in the exchange probability of less than 4%. The convergence of the PEDFs estimated using all of the data from RE is shown in Figure 5.2 D. The error in the exchange probability had only decreased to less than 2 % after approximately 60 ns per temperature. That is, SREM would have required preliminary simulations which were half as computationally expensive as the entire RE simulation in order to produce error in the exchange probability equivalent to that of ST. The slow convergence of PEDFs is likely why they have been updated throughout the course of the simulation in other studies. However, an SREM simulation is strictly correct only with accurate PEDFs.
Figure 5.2  Assessing convergence of weight factors and PEDFs

(A) The convergence of the PEDFs for SREM is quantified using the $\chi^2$ measure defined in equation 5.19. Using this measure, the PEDFs obtained using 19.5 ns of conventional MD at each temperature appear to be stationary. (B) Convergence of the PEDFs and weight factors using data from 19.5 ns of conventional MD per temperature, using the data from the complete RE simulation as reference. (C) Similarly, data from the replica exchange simulation, using only the first 25 ns per temperature, with the data from the complete RE simulation as reference. (D) Similarly, using all of the data from the RE simulation (4.75 µs). Error in the acceptance ratio is shown in (B), (C) and (D) for both ST in yellow (computed using equations B.5 and B.9) and SREM in purple (computed using equations B.4 and B.7); see Appendix B.

Figure 5.2 demonstrates that the error in the weight factors of ST leads to smaller error in the exchange probability than the error in the average energy of PEDFs. This finding is in qualitative agreement with a previous study comparing SREM and ST for a helical peptide. The PEDFs of SREM were observed to converge more slowly than the weight factors of ST when starting from a coil conformation, but not when both ST and SREM were started with a
helical conformation. In the original SREM paper, it was hypothesized, but not shown, that the calculation of PEDFs should be significantly easier than the calculation of weight factors for ST. In fact, we observe that the opposite is true for this system. The weight factors converge significantly faster than PEDFs and lead to more homogeneous sampling of temperature. The difference in errors is likely because the exchange probability in ST uses a difference in dimensionless free energies, whereas the absolute value of the potential energy is used in the exchange probability of SREM. Additionally, the method for computing the weight factors uses only the average potential energy at each temperature. It is harder to obtain convergence for a distribution of potential energies than the average of the distribution. Since weight factors are computed using average potential energies, they reach convergence more quickly than PEDFs. The accuracy of the potential energy value selected from the discrete PEDF in SREM is also affected by the number of bins and the bin width. The accuracy is decreased by having too few bins, whereas the convergence of the distribution is slower with a larger number of bins. These errors must then be balanced. Even if the PEDFs and weight factors converged at the same rate, ST has the advantage of convenience, since it entails storing a short list of weight factors rather than a distribution of energy values for each temperature.

We also tested the effects of using a poor estimate of the weight factors in ST. In order to generate suboptimal weight factors, we used the data from the first 750 ps of the RE simulation. This required a total of 24.75 ns summed over all temperatures, compared to 643.5 ns used to generate accurate weight factors. These weight factors produced inhomogeneous sampling of temperature, confirming that they were inaccurate estimates of the dimensionless Helmholtz free energies (demonstrated below). The purpose of this exercise was to emulate the more general case of a complex system for which one may not be able to accurately calculate weight factors due to the prohibitive computational cost. ST and STDR simulations carried out with these inaccurate weight factors are hereafter referred to as STb and STDRb, respectively.

Potential energy lists for the VREX simulation were also generated using the RE data. A list of 1000 energy values from the first 1 ns was used for each temperature. It is possible to run out of potential energies in the primary lists if temperatures are sampled heterogeneously. We
did not encounter this issue in the VREX simulation, and no secondary lists were used. However, secondary potential energy lists may be necessary in the application of this method to other systems. In summary, we highlight the varying costs of the initial simulations for each of the methods in terms of the simulations times: RE (0 ns), VREX (33 ns), SREM (825 ns), STDR (643.5 ns), STDRb (24.75 ns), ST (643.5 ns) and STb (24.75 ns).

5.7 Practical Implementation Issues

Before we begin a detailed comparison of the efficiency of the temperature-based generalized-ensemble methods, we briefly compare them with regard to the practical issues encountered in their implementation. A summary of this comparative discussion is provided in Table 5.1. Prior to beginning an enhanced sampling simulation, it is necessary to assess the available computational resources including the number of processors available, the heterogeneity of their speeds, and their failure rate (frequency of “crashes”).

<table>
<thead>
<tr>
<th>Implementation Issue</th>
<th>RE</th>
<th>SREM</th>
<th>VREX</th>
<th>STDR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalable to any number of CPU’s (even one CPU)?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algorithm readily accommodates a fluctuating number of CPU’s?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Efficiency impaired by inhomogeneity of CPU’s?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance severely affected by CPU failure?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Initial calculation of weight factors, potential energy distribution functions or potential energy lists required?</td>
<td>0</td>
<td>825</td>
<td>33</td>
<td>643.5 (STDR)</td>
<td>643.5 (ST)</td>
</tr>
<tr>
<td>Initial simulation times (ns)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 5.1** Practical advantages and disadvantages of generalized-ensemble algorithms

If a method is not affected by an implementation issue, the corresponding square is colored in green. Yellow indicates that the issue is somewhat of a concern, and red indicates that it is potentially a major pitfall. The only major issues for SREM and ST are the calculation of PEDFs and weight factors, respectively. STDR and VREX are not severely affected by any implementation issue. However, they do require very short initial simulations to obtain weight factors and potential energy lists, and ideally would not be run with a single replica. RE, in its typical implementation in which each replica is run on a dedicated CPU, is hindered by all of the issues listed, except that it does not require initial simulation.
In terms of the number of CPUs required, the RE algorithm specifies that the number of replicas equal the number of temperatures, which grows with system size. In the typical implementation of RE, the number of processors equals the number of replicas. If it is not possible to obtain access to the required number of processors, an alternative method or a more advanced RE implementation must be sought. Another possible scenario is that extra processors are available which could be utilized to speed up the calculation, but the RE algorithm does not allow the possibility of having more replicas than temperatures. This particular issue is addressed by the MREM algorithm, which utilizes multiple layers of replicas. However, there is no general mechanism to adapt RE to most efficiently use available resources. In contrast, both ST and SREM completely eliminate the need for a specific number of replicas. Multiple ST or SREM simulations can be run independently to take advantage of a computing cluster or distributed computing. The benefit of utilizing several processors simultaneously, each running an independent ST or SREM simulation, is simply reaching convergence more quickly in terms of wall clock time. Similarly, STDR and VREX algorithms do not require a fixed number of replicas. However, the aim of the DRPE is to enforce homogeneous sampling of temperatures for multiple replicas. Using only one replica is therefore not optimal, and ideally one would use a comparable number of replicas to the number of temperatures, though there is no specific requirement. A VREX simulation can in principle have any number of replicas. However, there is likely some benefit to having multiple replicas sampling different regions of conformational space in the updating of the potential energy lists (that is, running more than one replica at a time).

Of the generalized-ensemble methods we consider, only RE restricts the number of replicas from fluctuating during the course of the simulation. This may be a drawback in distributed computing platforms and shared computing clusters for which there is no way to predict the number of available processors in advance. Furthermore, the efficiency of RE is significantly affected by inhomogeneity of CPU speeds. Each exchange step can only occur when all the replicas have completed their MD calculation. Any inhomogeneity in the computing environment results in a waste of computational resources as some replicas must wait for the replica with the slowest processor to finish its calculation. This issue has been partly
addressed by the asynchronous replica exchange method, although some degree of synchronization is still required for the replicas undergoing exchange.\textsuperscript{43} Since none of the other methods require any direct communication between replicas, they do not suffer from this inefficiency. Another key drawback of the typical implementation of RE is its sensitivity to CPU failure.\textsuperscript{10,11} If one of the replicas is running on a processor that crashes, the entire RE simulation is stalled until this replica can be restarted on a functioning processor. The time wasted due to CPU failure depends on the failure rate of the cluster, and can be quite significant. Failure rates also rise with the number of replicas, and therefore the failure rate of RE is equal to the number of replicas times the failure rate of either SREM or ST.\textsuperscript{10}

In contrast to RE, the other four methods all have the advantage of not requiring a fixed and synchronized cluster of CPUs to function optimally. From a practical point of view, these methods are all superior to RE, except in one regard. Only RE does not require initial simulations at multiple temperatures to obtain weight factors, PEDFs, or potential energy lists. In particular, ST and SREM appear to only be suited to systems for which accurate weight factors or PEDFs can be calculated in a reasonable amount of simulation time. For the test system in the present study, accurate weight factors for ST were computed using single-temperature MD, whereas SREM required more simulation time and the use of RE in order to obtain sufficiently accurate PEDFs (see methods section). We will demonstrate that STDR can function with less accurate weight factors, and therefore requires less initial simulation time than ST. Similarly, VREX requires significantly less initial simulation than SREM, STDR or ST. Only short lists of potential energies at each temperature are needed to begin a VREX simulation.

An ideal temperature-based generalized-ensemble method would not require a significant initial simulation (as do ST and SREM), but also would not involve the use of a large cluster of homogeneous CPUs (as is common for RE). STDR and VREX address both of these issues, and are the most flexible algorithms in terms of practical concerns. These issues are particularly important if one is using a distributed computing platform with fluctuating numbers of
heterogeneous CPUs in many different locations, or a shared computing cluster which may present similar limitations.

5.8 Diffusion in Temperature

We characterize the efficiency of the temperature diffusion of each method using several different metrics, which are summarized in Table 5.2. First, we calculate the average acceptance ratio, which is a metric commonly reported for RE simulations. The methods separate into two categories based on their acceptance ratios: the RE-based methods (RE, VREX and SREM), and the ST-based methods (ST, STb, STDR and STDRb). ST has a higher acceptance ratio than RE for the same set of temperatures, in agreement with a previous comparison of the methods. Similarly, ST has a higher acceptance ratio than SREM. Zhang and Ma also observed that the rate of traversing temperatures is faster in ST, and that this effect becomes especially apparent if separations between adjacent temperatures are large, or if exchanges are attempted less frequently. Park proved that this is generally true for a given set of temperatures and concluded with a question as to whether the enhanced acceptance ratio affects the rate of sampling different microstates, and therefore structural convergence. We investigate whether the higher acceptance ratios in serial tempering algorithms (both ST and STDR) compared to parallel tempering (RE, VREX and SREM) do in fact lead to faster structural convergence in the next section. It should be noted that the DRPE in STDR decreases the acceptance ratio relative to ST, since it increases the probability of rejecting moves that result in inhomogeneous temperature sampling. The extent of this effect depends on the constants $c_1$ and $c_2$ in equation 5.18.

Next we consider a quantity which we call “replica speed”. Back exchanges can occur in which a replica accepts a move to an adjacent temperature, and at the next exchange returns to its previous position. These back exchanges contribute to the acceptance ratio, but they result in no net change in temperature, and typically no crossing of significant energetic barriers. In order to account for these “unproductive” moves, we calculate the replica speed as the average distance traveled after 50 exchange attempts. Values of replica speed are reported in Table 5.2. All of the methods have similar values for the replica speed, with SREM and VREX being slightly slower.
The higher acceptance ratios of the ST-based methods do not correspond to significantly faster replica speeds. That is to say, the higher acceptance ratios for the ST-based methods are partly due to an increased frequency of unproductive moves.

<table>
<thead>
<tr>
<th>Property</th>
<th>RE</th>
<th>VREX</th>
<th>SREM</th>
<th>STDR</th>
<th>STDRb</th>
<th>ST</th>
<th>STb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptance Ratio</td>
<td>0.237</td>
<td>0.241</td>
<td>0.248</td>
<td>0.378</td>
<td>0.376</td>
<td>0.463</td>
<td>0.404</td>
</tr>
<tr>
<td>Replica Speed (distance/time)</td>
<td>0.058</td>
<td>0.050</td>
<td>0.051</td>
<td>0.058</td>
<td>0.059</td>
<td>0.065</td>
<td>0.065</td>
</tr>
<tr>
<td>Mean Free Path</td>
<td>0.322</td>
<td>0.245</td>
<td>0.255</td>
<td>0.402</td>
<td>0.402</td>
<td>0.431</td>
<td>0.445</td>
</tr>
<tr>
<td>Diffusion Coefficient</td>
<td>0.208</td>
<td>0.181</td>
<td>0.156</td>
<td>0.195</td>
<td>0.196</td>
<td>0.246</td>
<td>0.249</td>
</tr>
<tr>
<td>Average Deviation from</td>
<td>0</td>
<td>6.62</td>
<td>12.61</td>
<td>2.50</td>
<td>2.98</td>
<td>3.81</td>
<td>17.40</td>
</tr>
<tr>
<td>Sampling Homogeneity (%)</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Composite Score (a normalized</td>
<td>0.79</td>
<td>0.70</td>
<td>0.68</td>
<td>0.87</td>
<td>0.88</td>
<td>0.98</td>
<td>0.94</td>
</tr>
<tr>
<td>linear combination)</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

**Table 5.2 Evaluating temperature diffusion**

The quality of the random walk in temperature for each generalized-ensemble method is assessed using five criteria, and an overall score is obtained by taking the normalized linear combination. Normalization is performed by dividing each value by the maximum value of that measure. The fifth measure was included in the normalized linear combination as \([1.0 - (\text{average deviation from sampling homogeneity})/100\%]\) for consistency with the other measures of temperature diffusion. The overall scores for each property are ranked from 1 to 7, representing decreasing performance. Scores are colored as follows: 1 and 2 (green), 3 to 5 (yellow), and 6 and 7 (red).

Making an analogy with the replicas traveling in temperature space as a type of diffusion in a one-dimensional coordinate, we calculate the mean free path and diffusion coefficient for each method. Mean free path is defined as the average distance traveled between successive rejected moves (“collisions”). The diffusion coefficient is defined as the rate of change of mean squared deviation of distance over time. We notice that ST, both with accurate and inaccurate weight factors, has the highest mean free path and diffusion coefficient. Both STDR simulations behave remarkably similarly, and are slightly slower at diffusion in temperature compared to ST.
RE has a higher diffusion coefficient than STDR, but a lower mean free path. It is also slightly more efficient at temperature diffusion than VREX or SREM.

Another important criterion is the deviation from sampling homogeneity, which indicates the amount of deviation from uniform sampling averaged over all the temperatures:

\[
\text{average deviation from homogeneity} = \frac{1}{M} \sum_{m=1}^{M} \frac{|N_m - \langle N_m \rangle|}{\langle N_m \rangle}, \quad (5.20)
\]

where the number of samples at temperature \( m \) is \( N_m \), the average number of samples per temperature is \( \langle N_m \rangle \), and \( M \) is the number of temperatures. We report the deviation from sampling homogeneity for each temperature in Appendix B, Figure B.1. The coupling of upward and downward moves in the RE algorithm results in perfectly uniform sampling of all temperatures. STDR produces nearly uniform sampling, with deviations from uniformity of 2.50 % and 2.98 % for accurate (STDR) and inaccurate (STDRb) weight factors, respectively. This is expected because the application of the DRPE favors uniform sampling of the temperature coordinate.\(^{11}\) Even with inaccurate weight factors, the sampling of temperature is still uniform, and the diffusion coefficient is still approximately the same. This indicates that STDR in the general case (i.e. with inaccurate weight factors) still successfully produces uniform sampling and good mobility in temperature. Our ST simulation also results in nearly uniform sampling, confirming the accuracy of the weight factors. ST with inaccurate weight factors (STb) produces the least uniform sampling, by design (17.40 %). We intentionally selected weight factors to produce uneven sampling to represent a more complex system for which calculating weight factors accurately would be computationally expensive. VREX produces relatively uniform sampling, with an average deviation of 6.62 %. Most of the inhomogeneous sampling in VREX occurs early in the simulation when the potential energy lists were based on a small amount of sampling, and the sampling is increasingly homogeneous with time. In contrast, SREM does not produce uniform sampling, with less sampling at the lowest temperatures and an average deviation of 12.61 %.
Temperature sampling efficiency is characterized by an overall score. The five measures of efficiency defined in this section are combined by averaging their normalized values. The overall score for each method is reported in the last row of Table 5.2. ST with accurate weight factors performs the best overall, and all of the ST-based methods perform better than the RE-based methods (especially SREM, which has the lowest overall temperature diffusion score).

5.9 Convergence of Structural Properties

The octamer GVGVPGVG is a disordered peptide with many thermally-accessible conformations, as shown in Figure 5.1. A useful descriptor of the conformation of such a short and flexible peptide is the end-to-end distance (EED). The EED probability distributions obtained using each of the generalized-ensemble methods at 280 K (the lowest temperature) are shown in Figure 5.3. Also shown is the average distribution, which is obtained by taking the average of all the methods. There is no systematic bias of either ST-based or RE-based methods towards sampling particular conformations. For example, for the peak at 5 Å, RE, STDRb and SREM are above the average while STDR, VREX and STb are below. Based on this observation and on the large amount of sampling in the combined data set of all seven methods (nearly 35 µs), we take the average to be the “gold standard” for comparison throughout the analysis of structural convergence (it is hereafter referred to as the “reference”). We quantify the deviation (\(\sigma_{eed}\)) of the EED distribution of each method (\(P_{eed}(n)\)) from the reference EED distribution (\(P_{eed,\text{reference}}(n)\)) by computing:

\[
\sigma_{eed} = \sum_{n=1}^{N_{\text{bins}}} (P_{eed}(n) - P_{eed,\text{reference}}(n))^2,
\]

(5.21)

where the index \(n\) labels bins, and there are \(N_{\text{bins}}\) in total. The values of \(\sigma_{eed}\) are reported in Figure 5.3. STDR exhibits the best agreement with the average distribution. In general, the ST-based methods have lower values for \(\sigma_{eed}\), corresponding to more accurate EED distributions than the RE-based methods.
Figure 5.3 Assessing the accuracy of the end-to-end distance (EED) distribution

The EED probability distribution is shown for each method with colors indicated in the legend using data from 280 K. The average distribution is computed as the average of all seven methods, and is shown in purple (dashed line). The error of the distribution of each generalized-ensemble algorithm, $\sigma_{\text{eed}}$, is shown next to the legend, and was computed using equation 5.21 with the average distribution as the reference.

In order to confirm that the ST-based methods produce more accurate EED distributions when compared to the RE-based methods, EED distributions for the lowest ten temperatures for each generalized-ensemble method are also computed and compared to the reference using equation 5.21. The EED distributions for each method and each temperature are displayed in Figure 5.4, along with the $\sigma_{\text{eed}}$ value which is the average of the ten temperatures. The ST-based methods produce EED distributions which are quantitatively more accurate than the RE-based methods at all temperatures. STDR shows the best overall agreement with the reference data set, with an average $\sigma_{\text{eed}}$ value of only 0.006, and distributions which clearly show the same temperature trend as the reference distributions.
Figure 5.4  EED distributions at different temperatures

The EED probability distributions are shown for the lowest ten temperatures for each generalized-ensemble method, as well as the average of all seven methods. The average error of the distributions of each generalized-ensemble algorithm, $\sigma$, is also shown. This was computed for each of the ten temperatures using equation 5.21 with the average distribution as the reference, and the average of these errors is shown on each plot. The RE-based methods are shown in the top row, and have larger errors than the ST-based methods, shown in the second and third rows. The average over all methods is the central plot.

For a systematic comparison of the generalized-ensemble methods, the convergence of several structural properties in addition to the EED distribution are considered. A useful ergodic measure is the 1,4 pair distance metric, which quantifies the convergence of the distance between 1,4 residue pairs (residues with indices $i$ and $i+3$) over time. We extend this metric to include all residue pairs, and therefore quantify the convergence of the $\alpha$-carbon distance matrix as follows:
\[
d_{\text{dC\alpha\ matrix}}(t) = \frac{1}{N_{\text{residues}}^2} \sum_{i=1}^{N_{\text{residues}}} \sum_{j=1}^{N_{\text{residues}}} \left( \bar{r}_{ij}(t) - \bar{r}_{ij,\text{reference}} \right)^2, \quad (5.22)
\]

where the indices \( i \) and \( j \) correspond to residue number and the number of residues is \( N_{\text{residues}} \).

The difference between each average pairwise distance \( \bar{r}_{ij} \) and the same average pairwise distance from the reference \( \alpha \)-carbon distance matrix (\( \bar{r}_{ij,\text{reference}} \)) is computed. In this equation, \( t \) refers to simulation time accumulated at the temperature considered, and \( \bar{r}_{ij}(t) \) is a cumulative average. As with EED, the average of all seven generalized-ensemble methods is used as the reference. We compute an analogous measure of convergence for the hydrogen-bonding contact map, depicted in Figure 5.1 B:

\[
d_{\text{contact map}}(t) = \frac{1}{N_{\text{residues}}^2} \sum_{i=1}^{N_{\text{residues}}} \sum_{j=1}^{N_{\text{residues}}} \left( P_{ij}(t) - P_{ij,\text{reference}} \right)^2, \quad (5.23)
\]

where \( P_{ij} \) is the probability of a hydrogen bond forming between the C=O group of residue \( i \) and the N-H group of residue \( j \), and \( P_{ij}(t) \) is a cumulative average of all the data. The elements of the reference contact map, \( P_{ij,\text{reference}} \), are computed using the data from all seven methods. We also directly compute the probability of forming certain turns (\( \gamma \)-, \( \beta \)- and \( \alpha \)-turns, defined by hydrogen bonds between residues \( i \) and \( i+2 \), \( i+3 \) and \( i+4 \), respectively) as well as the VPGV \( \beta \)-turn (shown in Figure 5.1, the most probable turn). In addition, the average probability of forming a hydrogen bond and a “bend” (as defined in the DSSP algorithm\(^6\)) on a per-residue basis are computed. The convergence of each of these structural properties is considered individually and compared to the reference data. Taken as a set, these structural properties provide a detailed description of the octapeptide’s complex conformational ensemble.

A representative example of how these structural properties measure convergence is shown in Figure 5.5 for ST at 280 K. The convergence of the \( \alpha \)-carbon distance matrix, the hydrogen-bonding contact map, and the EED distribution are displayed in Figure 5.5 A. The cumulative averages for the different types of turns, as well as hydrogen bonds and bends, are
shown in Figure 5.5 B. It is apparent from both of these plots that selecting a particular time at which the simulation is converged is ambiguous. Each structural property appears converged at a slightly different time. This ambiguity highlights the importance of considering multiple metrics when discussing the convergence of a simulation. In order to define convergence quantitatively, we consider the time taken to reach the reference value of the structural property of interest and remain within both one and two standard deviations, shown in Figure 5.5 C. Taking the average of these times provides a composite measure of when structural convergence is reached, and this average is a “structural convergence time”, $t_{sc}$. By comparing to the reference data, both convergence and accuracy are simultaneously assessed. The time at which each structural metric reaches the reference value is significantly different. For example, the EED distribution reaches the reference distribution faster than any of the other structural metrics, while the population of $\alpha$-turns requires nearly the entire ST simulation to reach the reference value to within one standard deviation. The structural convergence times are provided in Figure 5.5 D for each of the generalized-ensemble methods at 280 K. At this temperature, STDR converges fastest to the reference data, closely followed by ST and STDRb.
Figure 5.5  Assessing structural convergence using multiple criteria

The data for (A), (B) and (C) are from ST at 280 K. The trajectory is separated into fifty time intervals, and the quantities reported are calculated cumulatively. Time intervals are used to compare all methods fairly, since each method results in a different amount of sampling time at the lowest temperature. (A) Structural convergence is assessed using $\sigma_{\text{eed}}$ (equation 5.21), $d_{\text{contact map}}$ (equation 5.23) and $d_{\text{dCa matrix}}$ (equation 5.22, plotted on the secondary axis). (B) The probability per residue of a $\gamma$-turn, $\beta$-turn and $\alpha$-turn are shown, as well as the population of the VPGV $\beta$-turn. The probability of a hydrogen bond per residue and a bend per residue (plotted on the secondary axis) are also shown. (C) For each of the structural properties shown in (A) and (B), the time interval at which they reached and remained within one and two standard deviations of the reference data set are shown. The average of these times is also shown, corresponding to the average structural convergence time, $<t_{\text{sc}}>$. One standard deviation is calculated based on the values of each of the seven generalized-ensemble methods at the end of the simulation and their standard deviation from the reference value. (D) The average structural convergence times for one and two standard deviations are shown for all seven methods at 280 K in yellow and purple, respectively. These times are provided for temperatures 288 K, 296 K, 305 K, 314 K, 323 K and 332 K as supplementary information in Appendix B, Figure B.2.
For a systematic ranking of the structural convergence times, $t_{sc}$ is also calculated for the lowest seven temperatures. These times are provided as supplementary information in Appendix B, Figure B.2. While STDR converges faster than the other methods at 280 K, this is not a general trend for all temperatures. Each temperature has a different $t_{sc}$ for each method. The ranking of the methods varies between temperatures. For example, at 305 K, RE reaches convergence faster than STDR. This highlights the importance of evaluating more than the lowest temperature when comparing the performance of the methods, in addition to considering several structural metrics. It also suggests a way of quantifying the error in the measurement of $t_{sc}$. An average structural convergence time, $<t_{sc}>$, for each method is obtained by averaging $t_{sc}$ for the lowest seven temperatures, for both one and two standard deviations. The error in $<t_{sc}>$ is then the standard error of these measurements. Figure 5.6 A shows a two-dimensional plot of $<t_{sc}>$ to within two standard deviations versus $<t_{sc}>$ to within one standard deviation. Lower values for $<t_{sc}>$ indicate faster structural convergence. A clear trend emerges: ST-based methods reach structural convergence more quickly than RE-based methods. The method that reaches convergence the fastest is ST with accurate weight factors, while the method slowest to converge is SREM. It is not possible to conclusively rank the other methods due to error in $<t_{sc}>$. However, it is important to note that both VREX and RE converge faster than SREM. VREX is therefore not only a more convenient method for removing the synchronization from the RE algorithm than SREM, it is also faster at conformational sampling.

We can now answer a key question: does faster diffusion in temperature lead to a corresponding speed-up in conformational sampling? Figure 5.6 B demonstrates that this is in fact the case. The combined average structural convergence time, obtained by taking the sum of $<t_{sc}>$ for one and two standard deviations, is plotted versus the composite temperature diffusion score from Table 5.2. The ST-based methods, which have higher acceptance ratios and diffusion coefficients, also exhibit faster structural convergence. This key observation indicates that, in general, it is preferable to use an ST-based method because it provides enhanced efficiency in terms of conformational sampling. ST with accurate weight factors is clearly superior in both temperature diffusion and structural convergence, while SREM is the least efficient method in terms of both of these metrics. In the case of a simple system for which weight factors can be
obtained accurately with relatively little computational expense, ST is the method of choice. In the case of a more complex system for which sufficiently accurate weight factors might be expensive to obtain, the best choice would be to compute an initial estimate for the weight factors and use ST or STDR (corresponding to STb and STDRb here). Using the octapeptide as a test system, it is not possible to conclude which of these options is preferable. Inaccurate weight factors for this system yield comparable temperature diffusion and structural convergence for both STb and STDRb. To investigate this further, a more complex system, (GVPGV)₇, is also studied below.

Finally, another important question is whether inaccurate weight factors or PEDFs still lead to accurate, Boltzmann-weighted sampling at each temperature. It has been suggested that simulations with incorrect weight factors still yield correct statistics, only with suboptimal sampling of temperature. Analysis of the effect of suboptimal Helmholtz free energies on the accuracy of the data demonstrates that the resulting conformational populations are not biased by the use of inaccurate free energies (Figures 5.3 and 5.6 A). Both ST and STDR with inaccurate weight factors (STb and STDRb) converge to the reference data set, which indicates that they achieve accurate, Boltzmann-weighted conformational sampling. Figure 5.6 A shows that, even with inaccurate PEDFs, SREM still leads to Boltzmann-weighted sampling of conformational space (within one standard deviation) for this system. However, it converges more slowly than RE and all the other generalized-ensemble algorithms considered in this study.
Figure 5.6  Correlation between structural convergence and temperature diffusion

(A) Average structural convergence times, $<t_{sc}>$, obtained using the lowest seven temperatures are shown. The $<t_{sc}>$ to reach two standard deviations is plotted against the $<t_{sc}>$ to reach one standard deviation for each method. Error bars represent the standard error of $<t_{sc}>$ for the seven temperatures. Another version of this plot is provided as supplementary information in Appendix B, Figure B.3, with the $<t_{sc}>$ for each temperature shown. (B) The $<t_{sc}>$ times for one and two standard deviations from (A) are added together to create a structural convergence score, which is plotted against the temperature diffusion score from Table 5.2 for each method. A correlation is observed between structural convergence and temperature diffusion. ST-based methods (in yellow) have superior temperature diffusion, which leads to faster structural convergence compared to RE-based methods (in purple).
It has been pointed out that SREM is not rigorously correct when employing PEDFs that are not representative of the potential energies sampled during the simulation.\textsuperscript{10,14} In other words, SREM is correct only when stationary potential energy distributions are used, which may in the general case require adaptation of the PEDFs. In contrast, VREX is inherently adaptive due to the update of the potential energy lists. All RE-based methods, including VREX and SREM, require an initial equilibration phase to reach stationary potential energy distributions. Quasi-nonergodicity due to finite run length is a potential issue for any of the generalized-ensemble algorithms, but is most significant for SREM. Systematic error associated with selecting potential energies from unrepresentative conformations can be minimized in VREX by maintaining short, frequently updated potential energy lists. Reducing the size of the potential energy list in the VREX approach decreases the equilibration lag. Both VREX and RE use recent potential energy values for attempted moves. In addition, another inherent advantage of VREX over SREM is that in VREX, potential energy values utilized in the virtual temperature transitions are drawn from the same ensemble as those utilized in the evaluation of actual transitions, albeit with a stochastic time delay.

5.10 Comparison of STDR and Conventional MD

The relative sampling enhancement of RE compared to conventional MD has been the subject of significant controversy.\textsuperscript{20} For example, one study found that RE produced a speed up of 71.5 times at 275 K for a 21-residue helical peptide with implicit solvent, based on the autocorrelation function of helicity.\textsuperscript{19} In another work, an RE simulation of met-enkephalin in explicit solvent sampled five times more conformational space than a conventional MD simulation of the same duration.\textsuperscript{18} It has also been shown analytically that the expected speed up of RE is directly related to the activation enthalpy for two-state protein folding. The efficiency of RE is optimal when the maximum temperature is chosen just slightly above the temperature at which the folding activation enthalpy is zero.\textsuperscript{20} There are several key issues that emerge when reviewing comparative studies of RE and conventional MD. First, the observed sampling enhancement, or lack thereof, is heavily system dependent, as well as dependent on the structural or thermodynamic parameter on which the comparison is based. Second, evaluation of convergence for either the RE simulation or the MD simulation is often neglected.
Comparisons of other generalized ensemble methods, including ST and SREM, and conventional MD have also been performed.\textsuperscript{14}

Here we attempt to provide a rigorous comparison between STDR and conventional MD for the octapeptide (Figure 5.7). Figure 5.7 A and 5.7 B show a superposition of 200 structures obtained using STDR and MD respectively at 280 K. The amount of simulation time is the same for both methods (144 ns for conventional MD, and 144 ns in total for all temperatures for STDR, corresponding to 4.4 ns at 280 K). The root mean square deviations (RMSD) of these two collections of structures, 3.52 Å for STDR and 3.88 Å for conventional MD, are comparable. By this measure, both STDR and conventional MD produce a similar amount of conformational sampling using the same amount of CPU time.

We also show the convergence of the structural properties described in the previous section for both STDR (in Figure 5.7 C and 5.7 E) and conventional MD (in Figure 5.7 D and 5.7 F). STDR converges more quickly, approximately by a factor of 2-3 at 280 K. However, given that STDR requires sampling 33 temperatures for the same amount of time, it is much less computationally efficient. Specifically, the STDR simulation was 4.75 µs, compared to 200 ns for conventional MD. Overall, for this particular system there is no computational advantage in using STDR over conventional MD when the total cost of simulating all temperatures is considered.
Figure 5.7 Comparing STDR and conventional MD for GVGVPGVG

(A) Two hundred structures in ribbon representation obtained using the first 4.4 ns at 280 K for STDR are shown, and (B) for 144 ns of conventional MD, along with the corresponding RMSD. Glycine is in purple, proline is in yellow, and valine is in grey. (C) and (D) show $\sigma_{\text{total}}$ (equation 5.21), $d_{\text{contact map}}$ (equation 5.23) and $d_{\text{Ca matrix}}$ (equation 5.22, plotted on the secondary axis) for STDR and conventional MD respectively. The trajectories are separated into fifty time intervals, and the quantities reported are calculated cumulatively, as in Figure 5.5. (E) and (F) show the probability per residue of a $\gamma$-turn, $\beta$-turn and $\alpha$-turn are shown, as well as the population of the VPGV $\beta$-turn. The probability of a hydrogen bond per residue and a bend per residue (plotted on the secondary axis) are also shown.

However, in the present case, we are interested in the conformational ensemble at both low and high temperature because of the predicted temperature transition of the octapeptide GVGVPGVG.\textsuperscript{53,73,74} It is therefore still beneficial to use STDR because it enhances sampling at the individual temperatures. It is of key importance to note that we only know that conventional MD is able to satisfactorily reproduce the conformational ensemble of the octamer
by simultaneously using generalized-ensemble algorithms. It is only by comparing to STDR, as well as the combined data set of all the generalized-ensemble methods, that we are able to verify the convergence of the conventional MD simulation. Pseudo-convergence can be observed for a structural ensemble generated by conventional MD which is energetically trapped. In this way, it is possible to achieve convergence without simultaneously achieving accuracy. Using a generalized-ensemble method and allowing a random walk in temperature allows the system to overcome energetic barriers. Without knowledge of the energy landscape of the system of interest, it is hard to predict the expected sampling enhancement of a generalized-ensemble method. Similarly, it is hard to assess the accuracy of an apparently converged value, which is also expected to depend on the topology of the energy landscape.

5.11 Choosing Between ST and STDR

For relatively small and simple systems, such as the octapeptide used in this study or a short poly-alanine peptide in water, the calculation of dimensionless Helmholtz free energies is possible, although computationally intensive. For these cases, ST is an ideal method, since it alleviates the need for communication between processors in parallel tempering and the subsequent waste of computational resources. However, calculation of the Helmholtz free energies increases in difficulty as system size and complexity increase. When the system is sufficiently large and complex, as is often the case for biomolecular systems of interest, limited computational resources may preclude the calculation of sufficiently accurate weight factors. That is, it is only possible to obtain dimensionless Helmholtz free energies which result in an acceptable level of sampling uniformity with very extensive initial simulations. Even with near optimal weight factors updated throughout the simulation, Park and Pande still observed an average deviation from sampling homogeneity of 4.9 % for a short peptide (calculated based on the data in Table 1 of reference 30). With very inaccurate weight factors, sampling of temperatures may be far from uniform. That is, there may be too little sampling at certain temperatures to obtain a reasonable estimate of the weight factors to facilitate adaptation. ST may therefore not be an appropriate method, even with adaptation of the weight factors throughout the simulation. This may be the case for many biomolecular systems of interest which are larger than the small peptides or peptides in implicit solvent commonly used to test
generalized-ensemble methods. We now describe a complex system for which STDR is better suited than ST.

In addition to studying the octapeptide, GVGVPVG, we also study a longer peptide based on the same motif, (GVPGV)$_7$. Accurate weight factors for this system could not be obtained using a reasonable investment of computational resources (15 ns per temperature for 70 temperatures, for a total of 1.05 µs). Even with this large amount of data, the sampling of temperature in an ST simulation using these weight factors is heterogeneous. The average deviation from sampling homogeneity is 21.30 % (computed using equation 5.20). In contrast, using STDR and the same weight factors, the average deviation from sampling homogeneity is only 3.40 %. In Figure 5.8, the deviation from sampling homogeneity at each temperature is shown for both ST and STDR. In the ST simulation, sampling in the middle of the temperature range is less than both low and high temperatures, deviating from homogeneity by more than 60 %. Since it is unfavorable in this case to sample intermediate temperatures, diffusion from high to low temperature is impeded. In particular, the ST simulation using these weight factors experiences 56 % fewer transitions between 417 K and 454 K compared to the STDR simulation using the same weight factors. The sampling barrier in the intermediate temperature range impedes the random walk. The application of the DRPE results in a slight decrease of the acceptance ratio from 0.43 in the ST simulation to 0.38 in the STDR simulation. Importantly, the average replica speed in STDR is higher than that of the ST simulation (0.058 and 0.053 for STDR and ST, respectively). This indicates that replicas are able to efficiently explore temperature in the STDR simulation. In this case, the addition of the DRPE does not significantly impair replica mobility, and allows nearly uniform sampling of all temperatures. It is therefore a more suitable method than ST for this particular system.
Figure 5.8  Deviation from sampling homogeneity for simulations of (GVPGV)$_7$

For each temperature, the deviation from sampling homogeneity is computed:

\[
\text{% deviation from homogeneity} = \frac{\left( N_m - \langle N_m \rangle \right)}{\langle N_m \rangle} \times 100\%,
\]

(5.24)

where $N_m$ is the number of samples at temperature $T_m$, and $\langle N_m \rangle$ is the average number of samples per temperature. The ST simulation used the same weight factors as the STDR simulation, with 6 ns of sampling per replica (each started at a different temperature), for a total of 420 ns of simulation. There is a decrease in sampling in the middle of the temperature range, between 410 K and 460 K, which results in poor mobility of replicas between high and low temperature.

We performed another comparison between ST and STDR using weight factors obtained with only 500 ps of canonical MD at each temperature. These weight factors are more inaccurate than those based on 15 ns of simulation at each temperature, leading to an average sampling inhomogeneity of 96.88 % in an ST simulation. By applying the DRPE in an STDR simulation, the average sampling inhomogeneity is reduced to 8.84 %. The acceptance ratio for the STDR simulation with these inaccurate weight factors is 0.38. This is exactly the same as the acceptance ratio of the STDR simulation using the weight factors based on 15 ns at each temperature. This observation is in agreement with the results of the STDR and STDRb simulations of the octapeptide. In Table 2, both STDR and STDRb simulations exhibit nearly
identical temperature diffusion coefficients, acceptance ratios and mean free paths. Importantly, the results for both the octapeptide and 35-residue peptide demonstrate that replica mobility in STDR simulations is not significantly affected by inaccuracy of the weight factors. Further, more accurate weight factors do not improve temperature diffusion, or corresponding structural convergence. Thus, we recommend that when the STDR algorithm is used, the computational investment for the initial calculation of weight factors should be minimized.

5.12 Performance of STDR for More Complex Systems

To compare STDR with conventional MD, we also performed an MD simulation at the lowest temperature of the STDR simulation (261 K). A superposition of 200 structures, obtained every 1 ns from a 200 ns trajectory generated using conventional MD at 261 K is shown in Figure 5.9 A. These structures have an average RMSD of 1.66 Å, indicating that the peptide is trapped in one conformational basin and undergoes only small conformational changes. This set of structures contrasts with the set of 200 randomly-selected structures from the complete STDR simulation at 261 K (Figure 5.9 B), which represents completely different conformations with an average RMSD of 8.40 Å. For clarity, we also show six example structures in Figure 5.9 D to demonstrate the variety of conformations observed in the STDR simulation. To make a more direct comparison between conventional MD and STDR, Figure 5.9 C shows 200 structures from STDR using the same amount of simulation time as the conventional MD simulation (200 ns summed over all the temperatures, corresponding to approximately 3 ns at 261 K). The radius of gyration distributions for conventional MD, STDR and the first 3 ns of STDR are shown in Figure 5.9 E. Both distributions from STDR show several conformational states, while the conventional MD simulation is trapped in one state. Even using the same amount of computational resources, STDR produces a more heterogeneous ensemble of conformations. Single-temperature MD severely underestimates the heterogeneity of the conformational landscape and exhibits both pseudo-convergence and quasi-nonergodicity.

It is not possible for this system to quantitatively measure the speed-up of STDR versus conventional MD because limited computational resources preclude performing conventional MD simulations for the time required to achieve structural convergence. We observe that
conventional MD is trapped in one conformational basin for 200 ns. It is not possible to accurately predict how long it would take to sample all relevant states and reach convergence. Qualitatively, we observe a dramatic sampling enhancement due to STDR. Using the same amount of computational resources, STDR generates more unique conformations for this peptide. This indicates that the random walk in temperature does in fact lead to enhanced sampling, establishing the efficacy of the STDR method for a complex polypeptide.

**Figure 5.9 Comparing STDR and conventional MD for (GVPGV)$_7$**

Two hundred structures in ribbon representation along with their RMSD are shown for (A) the conventional MD simulation of length 200 ns at 261 K, (B) for the STDR simulation at 261 K using all of the data (120 ns at this temperature), (C) and for the STDR simulation at 261 K using the first 3 ns (this is the same simulation time summed over all replicas as (A)). Glycine is in purple, proline is in yellow, and valine is in grey. (D) A selection of six example structures is shown from the structures in (B) to illustrate the structural diversity obtained using STDR. The probability distributions of radius of gyration for the data sets described are shown in (E).

Before sampling the complete energy landscape of a system of interest, there is no way to confidently predict the height of the energy barriers, or the number of energetically-stable conformations (local minima of the energy surface). By coupling to simulations at higher
temperatures, high energetic barriers can be overcome. However, if one is not simultaneously interested in the behavior of the system at multiple temperatures, it may be less computationally expensive to run very long simulations, or a collection of simulations, at a single temperature.\textsuperscript{75} In the present study, conventional MD successfully produced the conformational ensemble of the octapeptide, but resulted in quasi-nonergodicity for the 35-residue peptide. In order to truly “enhance sampling” relative to single-temperature MD simulations, an enhanced sampling method must achieve convergence at a rate which is greater than the product of the number of replicas and the computer time for each replica.

The present study of (GVPGV)$_7$ shows that it is possible to observe pseudo-convergence using single-temperature MD (that is, to observe convergence of a quantity of interest without observing the true value of that quantity, Boltzmann-weighted by the populations of all possible conformations). Long-time MD simulations do not yield the appropriate conformational distribution and the system remains trapped in a local minimum of the energy landscape. In contrast, we observe that conventional MD is able to satisfactorily reproduce the conformational ensemble of the octapeptide at a significantly reduced computational cost compared to using a generalized-ensemble method. In light of this apparent contradiction, how are the averages of quantities obtained using MD simulations to be interpreted? Based on this work, it appears that using generalized-ensemble algorithms is a more prudent approach, even if in some cases it may be less efficient overall to do so (for increased confidence in the accuracy of the data). Several other examples have shown that the enhanced sampling provided by generalized-ensemble methods provides convergence that would not be feasible with single-temperature MD.\textsuperscript{6,14,18-24} These observations underscore the need not only for enhanced sampling methods, but also the shortcomings of techniques such as block averaging over simulations initiated in a single conformational basin in estimating the convergence of results. The challenge in simulating complex systems is that \textit{a priori} one does not know the efficiency of the generalized-ensemble approach relative to the “brute force” MD approach. It may be advisable to use a generalized-ensemble algorithm, especially if conformational sampling, and not dynamic information, is sought.
5.13 Conclusions

We now return to the original question: given limited computational resources, which generalized-ensemble algorithm is most efficient at sampling a complex conformational landscape? The first important distinction between methods is the separation between those based on ST and those based on RE. In this chapter, we demonstrate that ST-based methods result in both faster temperature diffusion and faster structural convergence. They are therefore preferable to RE-based methods. This is the most general conclusion of this chapter.

Within the family of RE-based methods, the computational efficiencies of the various algorithms are not equivalent. SREM should only be applied to systems for which PEDFs can be accurately obtained. Therefore, due to limited computational resources, SREM can only be applied to simple systems. Like SREM, RE is not well-suited to complex systems because of the need to synchronize simulations of a large number of replicas (and typically, a large number of processors). Although there is no theoretical limit on the number of replicas that one can use for a RE simulation, it is generally difficult in practice to obtain access to a large, dedicated and homogeneous computing cluster. Even if one does have access to such a computational resource, the wasted CPU time may also increase sharply with the number of replicas due to both CPU failure and inhomogeneity in CPU speeds.

VREX represents an attractive alternative to RE since it completely eliminates synchronization and communication between replicas. It produces more homogeneous sampling of temperature compared to SREM with much less initial simulation time. It is therefore preferable to both SREM and RE. Since RE-based methods suffer from slower structural convergence and temperature diffusion compared to ST-based methods, it is preferable to use an ST-based method in temperature. This disadvantage may become less significant for complex systems. VREX may also be a more suitable method in another reaction coordinate other than temperature for which weight factors are much more difficult to obtain. Moreover, initial simulations for weight factor calculations may also benefit from the use of VREX.
In the case of relatively simple systems for which weight factors can be accurately calculated using minimal computational resources, ST is the most appropriate method. We have shown that ST with accurate weight factors exhibits the fastest temperature diffusion, and correspondingly, the fastest structural convergence. However, for more complex systems, for which weight factors are costly to compute, STDR becomes the preferred method. Even in the limit of infinite resources, a long initial simulation to compute weight factors accurate enough to yield homogeneous sampling in ST may not be the most efficient use of computational resources. The feasibility of an ST simulation is determined by the accuracy of the weight factors, which can only be assessed by actually performing an ST simulation. Importantly, we have demonstrated that STDR can make use of inaccurate weight factors to achieve homogeneous sampling of temperature and consequently structural convergence. Replica mobility is only slightly impeded by the DRPE. By contrast, ST with inaccurate weight factors produces heterogeneous sampling of temperatures, which is also an impediment to the random walk in temperature. STDR is suitable for any computing cluster or distributed computing environment, since it requires no fixed number of CPUs or synchronization of exchanges. Complex systems can therefore benefit from the STDR method, which is increasingly advantageous as system size and complexity grow.
References


CHAPTER 6

Simulated Tempering Distributed Replica Sampling and Statistical Convergence of Disordered Ensembles: Methodological Details

Parts of this chapter were adapted from an article published in the Journal of Physics: Conference Proceedings. This chapter also contains unpublished results on statistical convergence of structural properties, a method to determine the “equilibration period”, and a description of error analysis using the “blocking” method.

Reference:

Contributions:
S.R. was the primary author of the manuscript. R.P. provided editorial input and guidance.
6.1 Summary

STDR is a generalized-ensemble method designed specifically for simulations of large molecular systems on shared and heterogeneous computing platforms. The STDR algorithm consists of an alternation of two steps: (1) a short MD simulation, and (2) a stochastic temperature jump. The essentials of the algorithm are quite simple; its implementation, however, is non-trivial when applied to complex biomolecular systems. The aim of this chapter is to describe the details of our STDR implementation, which is a highly-parallel algorithm designed to maximize computational efficiency while simultaneously minimizing network communication and data storage requirements. We quantify the computational overhead for performing STDR compared to conventional MD simulations, and we determine the amount of processor time required for each step of the algorithm. Using a 35-residue disordered peptide in explicit water as a test system, we characterize the efficiency of the STDR algorithm with respect to both diffusion in temperature space and statistical convergence of structural properties. Importantly, we show that STDR provides a dramatic enhancement of conformational sampling compared to a canonical MD simulation. We also carefully delineate the extent of the “equilibration” period for this simulation based on an analysis of several structural properties. It is essential to remove unequilibrated configurations in order to accurately compute average properties of disordered ensembles.

6.2 Introduction

In this chapter, we present our implementation of simulated tempering distributed replica sampling (STDR). This algorithm is described in section 5.3.6 of Chapter 5. We refer the reader to this section, and equations 5.18 and 5.19, which form the basis for this algorithm. We also refer the reader to Appendix C for the definitions of all structural properties (for example, hydrogen bonding, nonpolar contacts, and hydration) used in this chapter, and the following two chapters.

This chapter describes the methodology used to obtain the results described in Chapters 7 and 8. In those two chapters, we present the results of STDR simulations of the following three systems: (1) the octapeptide, GVGVPVG, (2) the 35-mer, (GVPGV), and (3) an
aggregate of eight chains of (GVPGV)₇. These peptides contain the elastin-like sequence motif “GVPGV”. Because peptides based on this motif have been the most commonly studied elastin-like peptides using both simulation and experiment, we have chosen these systems for the present study. The motif “GVPGV” is derived from exon 24 of chicken elastin.³ It is one of the elastin-like sequences in our model set in the study described in Chapter 4.⁴ The application of STDR to the peptide (GVPGV)₇ is also the subject of sections 5.11 and 5.12 in the previous chapter. However, the simulations described in this chapter involve an order of magnitude more sampling than those described in Chapter 5, which we demonstrate is necessary to obtain statistical convergence of average structural properties. In this chapter, we focus on the application of STDR to obtain a configurational ensemble for the 35-residue disordered peptide (GVPGV)₇. The STDR simulations of the octapeptide and the aggregate were performed using essentially the same method; the necessary differences in methodology are specified in section 6.3.

6.3 Implementation Details

6.3.1 Computational Platform

The general purpose cluster (GPC) at the University of Toronto’s SciNet consortium was used to perform large-scale STDR simulations of the (GVPGV)₇ peptide in water. Each replica was run on an individual node of the GPC. Each Nehalam node has eight 2.5 GHz CPUs and 16 GB of memory. Because the STDR algorithm requires only minimal communication between nodes, the portion of the GPC connected with gigabit Ethernet was utilized. In total, 106 nodes were used for the simulations described here: 105 nodes were used to simulate individual replicas, with one additional node designated as a server. In fact, the number of nodes running concurrently occasionally is slightly less than 106 due to rare node failure, and fluctuations in node availability (since the SciNet GPC is a shared computing resource). This is not problematic, however, since STDR is well-suited for such a shared computing platform.²⁵ In total, these simulations required 1.5 million CPU hours (approximately 106 nodes running for 75 days).
The simulation system consists of the (GVPGV)$_7$ peptide capped with an acetyl group at the N-terminus and an NH$_2$ group at the C-terminus in a rhombic dodecahedral box with 9663 water molecules. Simulations were performed with an accurate leap-frog stochastic dynamics integrator using GROMACS version 4.0.5. The OPLS-AA/L force field and the TIP3P model were used for the peptide and water, respectively. Periodic boundary conditions were applied. Calculation of electrostatic forces utilized the particle mesh Ewald summation method with a Fourier spacing of 1.2 Å and a fourth-order interpolation. The real-space Coulombic cutoff was 11 Å. Lennard Jones interactions were cut off at 14 Å. The LINCS algorithm was used to constrain covalent bonds involving hydrogen atoms, and the SETTLE algorithm was used for water, allowing the use of a 2 fs integration time step. The solvated system was initially equilibrated in the isothermal-isobaric ensemble at 300 K and 1 atm for 0.5 ns using the Parrinello-Rahman barostat. All STDR simulations were performed in the canonical ensemble using stochastic dynamics as a thermostat. In addition, a long-time canonical MD simulation was performed for 700 ns starting with a randomly-selected configuration from the STDR simulation at 300 K.

For the STDR simulation, 105 temperatures exponentially-spaced between 266 K and 749 K were used. At each of these temperatures, a preliminary canonical MD simulation was performed for 15 ns. The lists of potential energies obtained from these simulations were used to compute the dimensionless Helmholtz free energies in equation 5.7 using the method of Park and Pande. A set of 105 random starting configurations (one for each replica) was obtained from ten 10 ns simulations at 1000 K. At the beginning of the STDR simulation, each replica was started at one of the 105 temperatures. The DRPE constants $c_1$ and $c_2$ were set to 0.004 and 0.002, respectively. Each replica performed $-10^5$ 8 ps MD runs, and attempted temperature jumps according to equations 5.17 and 5.18. In total, this resulted in an accumulated simulation time of 84 µs for the entire STDR simulation, which corresponds to approximately 800 ns per replica. This is an order of magnitude more simulation than our the study described in Chapter 5. The implementation of STDR makes use of several GROMACS programs including g_energy, trjconv, trjcat, eneconv, and trjorder. These programs are used for analysis,
and data management, as described in detail in section 6.4. The definitions of all structural properties are described in Appendix C, and summarized in Table C.1.

6.3.3 MD Simulation Details for the Octapeptide

The simulation of the octapeptide used 53 temperatures exponentially-spaced between 266 K and 749 K. We used half as many temperatures compared to the simulation of (GVPGV)$_7$ because the system size was significantly smaller (a rhombic dodecahedral box with 2713 water molecules). Furthermore, we also used ST, not STDR, to improve the quality of the random walk and to decrease the simulation time required to reach structural convergence on the basis of our findings in Chapter 5. The resulting acceptance ratio for temperature jumps is reasonably high (38.1 %). A total of ~ 35000 8 ps MD runs were performed by each replica, yielding a total simulation time of 16 µs. All other simulation parameters were identical to the simulation of (GVPGV)$_7$ described in section 6.3.2.

6.3.4 MD Simulation Details for the Aggregate of Eight (GVPGV)$_7$ Chains

The STDR simulation of the aggregate of eight (GVPGV)$_7$ chains used 306 temperatures exponentially-spaced between 270 K and 749 K. Approximately three times as many temperatures were required compared to the simulation of (GVPGV)$_7$ because the system size was much larger (a rhombic dodecahedral box with 24211 water molecules and 276065 atoms in total). In order to achieve good overlap between potential energy probability distributions for neighbouring temperatures, the temperatures must be closely-spaced. For this system, the temperatures were separated by only 1 K at the lower end of the temperature range. The resulting acceptance ratio for temperature jumps is adequate (10 %). Approximately 135000 5 ps MD runs were performed by each replica, for a total simulation time of 207 µs. All other simulation parameters were identical to the simulation of (GVPGV)$_7$ described in section 6.3.2. This simulation was run over the course 18 months, in total requiring 13 million CPU hours. To the best of our knowledge, this is the largest generalized-ensemble simulation performed to date.
6.4 Implementation of the STDR Algorithm

6.4.1 Parallelization

Our STDR implementation incorporates two levels of parallelization:

(1) First, the STDR algorithm is inherently highly parallel: the system is simulated as a set of replicas, each occupying a single node. Each replica executes alternating MD simulations, and stochastic temperature jumps. Minimal information sharing between replicas is required: only the current temperature location of all replicas is required to compute the DRPE in equation 5.18. In the present implementation, a server node collects the temperature information, and the trajectories of all replicas.

(2) Parallelization also occurs at the level of the individual replica node. Each 8 ps MD simulation is run in parallel using all 8 cores of the node. The time required for 1 ns of simulation is 17.98 hours and 2.12 hours for one core and eight cores, respectively. Thus, the parallel efficiency is 106%. Superlinear scaling is achieved using GROMACS version 4.0, which has been highly optimized to scale well on parallel machines.7

6.4.2 The Replica Nodes

The specific steps executed by a single STDR replica are illustrated in the flow chart in Figure 6.1. The cycle of steps in a single temperature move begins with a short MD simulation (8 ps in length). The final potential energy from the last step of this MD simulation is obtained using the GROMACS program g_energy. This potential energy value, along with an up-to-date list of all replicas’ temperatures (received from the server node), is used to compute the probability of accepting a proposed temperature jump (equations 5.17 and 5.18). Information about the temperature jump is stored, including the potential energy, DRPE, probability, and temperature. In order to begin the next MD step, grompp generates a run input file. To minimize the amount of hard disk space required for storage, the water molecules in the MD trajectory are ordered by proximity to the protein using trjorder: only the nearest water molecules are saved in the final trajectory using trjconv. The replica’s trajectories are concatenated locally (on the node on which the replica is running). Trajectories corresponding
to the same temperature must also be concatenated; this occurs on the server node. The trajectory corresponding to each STDR step must therefore be sent to the server node over the network. The replica periodically backs up the data stored in memory to disk once every 4 hours. This checkpointing limits the amount of data lost in case of a node failure to a maximum of a few hours, which corresponds to a few hundred STDR steps. This approach was designed to minimize disk I/O, minimize the amount of long-term data storage required for trajectories, and maximize the fraction of compute time spent on MD (see Figure 6.2). Because GPFS, the high performance file system of the SciNet GPC, is not designed for accessing many small files, it is important that all trajectory files be concatenated while still in memory. Data locality is maximized by passing the smallest possible amount of data between the server node and the replica nodes. All of the processes on the nodes are controlled by bash scripts that call GROMACS programs, as well as an in-house Fortran program that computes the STDR temperature jump criterion. In Figure 6.2, we assess the efficiency of the STDR implementation for the (GVPGV)$_7$ test system by determining the amount of processor time required for each step of the algorithm. The majority of the STDR overhead results from managing the trajectory files (trjorder, trjconv, and concatenation). These steps are not part of the core STDR algorithm. However, the trajectories must be the minimum size possible and concatenated to facilitate structural analysis and long term data storage.
The cycle of steps involved in STDR is illustrated as a flow chart. Each cycle consists of two parts: (1) a brief MD simulation (8 ps in length), and (2) an attempted temperature jump. Associated with this temperature jump are several essential data processing steps. During the cycle, all data is stored in shared memory, and no disk access is required. The cycle is repeated thousands of times, with occasional interruptions to back up essential data to disk (every few hours), and concatenate trajectories (every two days).

### 6.4.3 The Server Node

The server node fulfills two key roles in our implementation of STDR:

1. It stores the current temperature of each replica. The server maintains a current list of the temperatures of all replicas. This temperature list is communicated frequently to each replica node, ensuring that an up-to-date temperature list is used in the calculation of the DRPE (equation 5.18). After every temperature jump, the replica communicates its current temperature to the server.

2. The server node also collects the short MD trajectories of all replicas. Trajectories must be stored according to temperature to facilitate further structural analysis (described in detail in section 6.6 below). In order to concatenate trajectories by temperature, the trajectories of...
individual replicas must be collected in a central location (the shared memory of the server node). When the memory is almost full, the trajectories are concatenated by temperature and moved to disk. By performing all trajectory concatenation in memory on the nodes, disk I/O is kept to an absolute minimum.

Figure 6.2  Node time required for each step in STDR
(A) The percent of replica node time spent on each step in the STDR algorithm is represented as a pie chart (colours correspond to the same steps illustrated in Figure 6.1). (B) Overall, 96.14 % of node time is spent performing MD, while 3.86 % is spent on all tasks associated with STDR, including management of trajectory files. Note that the mdrun step in Figure 6.1 is coloured blue to corresponding to the overhead for starting and stopping mdrun indicated in Fig. 6.2 A.
6.5 Evaluating Temperature Diffusion

The quality of the random walk in temperature is assessed using several different metrics, which we define in Chapter 5. These metrics are reported in Table 6.1. Importantly, the sampling of temperatures is very close to perfect homogeneity: the average deviation from homogeneity is only 1.3%. This indicates that the constants $c_1$ and $c_2$ in equation 5.18 have been appropriately selected to enforce homogeneous temperature sampling. Almost 30% of temperature jumps are accepted, which indicates good overlap of potential energy distributions of neighbouring temperatures, and therefore adequate temperature spacing. The majority of replicas have visited all temperatures at least once, and on average each replica has visited 94% of temperatures.

Table 6.1 Evaluating the diffusion of replicas in temperature space

<table>
<thead>
<tr>
<th>Temperature Diffusion Properties</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>average deviation from homogeneity $= \frac{1}{M} \sum_{m=1}^{M} \frac{</td>
<td>N_m - \langle N_m \rangle</td>
</tr>
<tr>
<td>acceptance ratio = number of accepted T jumps / total number of attempted jumps</td>
<td>0.292</td>
</tr>
<tr>
<td>mean free path = average distance travelled between rejected temperature jumps</td>
<td>0.280</td>
</tr>
<tr>
<td>replica speed = (average distance travelled in 50 steps) / 50</td>
<td>0.051</td>
</tr>
<tr>
<td>diffusion coefficient = 0.5 x slope in the linear region of the mean square displacement vs. steps plot</td>
<td>0.085</td>
</tr>
<tr>
<td>average fraction of temperatures visited per replica</td>
<td>0.94</td>
</tr>
</tbody>
</table>

$N_m$ is the number of samples at temperature $m$, and $M$ is the number of temperatures.

In Figure 6.3, the random walk in temperature is shown for one of the replicas. By varying the temperature, the STDR algorithm leads to exploration of conformational states with a wide range of radius of gyration, which is a measure of the spatial extension of the peptide. At
low temperature, the peptide samples more compact conformations; increasing temperature 
allows the peptide to sample more expanded conformations. The random walk in temperature 
facilitates the peptide’s expansion and collapse, preventing it from becoming trapped in any 
specific conformational state.

Figure 6.3  A two-dimensional plot of the radius of gyration of the peptide vs. temperature 
The complete STDR simulation of a single replica (750 ns in total) is shown in gray. Superimposed in black is the same trajectory including only every hundredth point with arrows for clarity.

6.6  Comparison of Canonical MD and STDR

In order to quantify the sampling enhancement provided by STDR, we compare the 
conformational ensemble at 300 K with that of an MD simulation in the canonical ensemble at 
300 K (Figures 6.4 and 6.5). First, we monitor the number of intramolecular peptide hydrogen 
bonds that have been observed with increasing simulation time, shown in Figure 6.4 (a hydrogen 
bond was counted as observed if it was formed in at least one of the configurations in the
trajectory). With the complete STDR trajectory at 300 K, more than 92% of all possible hydrogen bonds were observed in at least one configuration. In contrast, the canonical MD simulation at the same temperature only sampled 11.9% of all possible peptide-peptide hydrogen bonds. We also show the number of hydrogen bonds observed for the replica trajectory depicted in Figure 6.3. This replica has sampled 95% of all possible hydrogen bonds. The STDR simulation has not yet sampled all possible hydrogen bonds, which indicates that unexplored regions of conformational space remain to be sampled. It is also possible that some of the intramolecular hydrogen bonds that have not yet been observed cannot be made due to structural constraints. Nevertheless, the vast majority of intramolecular peptide hydrogen bonds have been sampled at least once.

The canonical MD simulation was 700 ns in length. Representative sets of configurations from both the MD simulation and the STDR simulation are shown in Figure 6.5, along with hydrogen-bonding contact maps for these ensembles. The canonical MD simulation exhibits quasi-nonergodicity:16 even though the system is ergodic, it appears non-ergodic because it samples essentially a single conformational basin throughout the entire simulation (over hundreds of nanoseconds). The configurations from the canonical MD simulation do not differ significantly from each other based on RMSD. In contrast, the configurations from STDR represent a small subset of the thousands of configurations generated by the simulations, and illustrate the heterogeneous set of structures that this peptide can adopt. By comparing the hydrogen-bonding contact maps obtained using canonical MD and STDR (Figure 6.5 C and 6.5 D), it is evident that canonical MD severely underestimates the structural heterogeneity of this disordered peptide, while STDR provides a dramatic enhancement in conformational sampling.
Figure 6.4  Number of intramolecular peptide hydrogen bonds (HBs) observed with increasing simulation time for STDR (green and purple) and canonical MD (red)

In purple, the number of HBs observed for the replica trajectory corresponding to Figure 6.3 is shown. The number of possible HBs is computed based on the number of possible contacts in the HB contact plot, Figure 6.5 A and 6.5 B (35 x 35 = 1225). From this value, we subtract the diagonal and the two off diagonals (35+34+34); these HBs cannot be formed due to conformational restrictions. Contacts that would involve the N-H of proline residues are also subtracted. Thus, there are 898 possible HBs. The definition of HBs here is the same as that described in Chapter 4.
Figure 6.5 Configurations of the (GVPGV)₇ peptide from canonical MD and STDR simulations at 300 K

Intramolecular hydrogen-bonding contact maps for the ensembles from canonical MD and STDR are shown in (A) and (B), respectively. The probability of a particular hydrogen bond between backbone C=O and N-H groups is indicated by the colour of the corresponding square. Note that the contact maps have different colour scales; some of the contacts in the canonical MD simulation are highly populated, whereas the STDR ensemble has no contacts populated in more than 12 % of structures. In (C) and (D), we show 100 randomly selected configurations from each ensemble and the average root mean-square deviation (RMSD) of backbone atoms compared to the starting configuration. Configurations are shown with a cartoon representation of the backbone using VMD ¹⁷ (glycine in orange, valine in purple, and proline in yellow). The definition of HBs here is the same as that described in Chapter 4.
6.7 Delineating the “Equilibration” Phase

The STDR simulation of each replica started from a random configuration generated at high temperature. Because these configurations are not representative of the true ensembles at each temperature, an initial equilibration phase occurs during which these random configurations collapse to more representative, compact structures. In order to delineate this equilibration phase of the STDR simulation, we consider the running average of five structural properties as a function of time: radius of gyration, $R_g$; number of peptide-peptide hydrogen bonds per group, $x_{HB}$; number of nonpolar contacts per residue, $x_{NP}$; number of heavy atom contacts per residue, $x_{HA}$; and number of water molecules in the hydration shell per residue, $x_{shell}$ (Figure 6.6). In Figure 6.6 F, we show each of these running averages normalized by the average over the last 60% of the trajectory, denoted by the symbol $X$. The length of the initial “equilibration” period is then the time at which all of these quantities fall within 5% of their final averaged value. Based on this analysis, the first 40% of the trajectory is discarded as “equilibration”. Configurations accumulated during the equilibration phase are discarded, and all other configurations are used for further structural analysis.

The delineation of the equilibration period based on running averages (Figure 6.6) is further supported by the behaviour of the probability distribution of $R_g$ for each time interval in Figure 6.7. The trajectory at 300 K is divided into 10 equally sized intervals, and the probability distribution of $R_g$ is computed for each interval. On the basis of this analysis, it is clear that the first four intervals (up to approximately 305 ns) contain a significant number of configurations with unrepresentative larger values of $R_g$. The probability distribution of $R_g$ appears to have reached statistical convergence as it does not change with additional simulation time after 305 ns. It is always possible that longer simulations will reveal longer timescale phenomena. Thus, it is not possible to be certain in declaring that a simulation or a structural property is converged.\(^\text{18}\)

Similarly, we note that it is important to investigate the convergence of several structural properties because they may reach statistical convergence with differing amounts of conformational sampling.\(^\text{5}\)
Figure 6.6  Statistical convergence of structural properties

Each graph shows the running average based on the trajectory obtained at 300 K divided into 10 equally sized intervals. Configurations in the trajectory are stored according to the time at which they were sampled (in real time). (A) The running average of the radius of gyration, $R_g$. (B) The running average of the number of hydrogen bonds per hydrogen bonding group (C=O or N-H), $x_{HB}$. (C) The running average of the number of nonpolar contacts per residue, $x_{NP}$. (D) The running average of the number of heavy atom contacts per residue, $x_{HA}$. (E) The running average of the number of water molecules in the hydration shell defined based on the heavy atoms per residue, $x_{shell}$. (F) Here, $X$ is the running averages of the structural properties shown in panels (A) to (E), normalized by the average of over the last 60% of the trajectory. In this case, a value of 1 indicates that $X$ is equal to the average over the last part of the trajectory.
Figure 6.7  Probability distribution of the radius of gyration, $R_g$

The probability distribution of $R_g$ is shown for the trajectory at 300 K divided into 10 equally sized intervals. The colour scheme varies from purple to red to correspond to the time intervals (from the first to the last interval). Note that the simulation time in this case does not indicate a time trajectory, only the accumulated simulation time at 300 K.

6.8 Estimating the standard error using the “blocking” method

For the analysis of structural properties presented in Chapters 7 and 8, we compute the estimated standard error of the mean, $\sigma_{est}$, using the “blocking” method first described by Flyvbjerg and Petersen.\textsuperscript{19} This algorithm was designed to estimate error for sets of data in which there is significant correlation. Thus, it is ideally suited to estimate error in averages obtained from MD simulations. Because sequential configurations obtained in an MD simulation are similar to each other, the measurements of a quantity of interest in a time series cannot be considered to be independent measurements of the quantity of interest. The “blocking” method has similarly been applied to compute errors for averages derived from MC simulations; a series of measurements from an MC simulation also exhibits correlation because each configuration in the series was obtained from the previous configuration by making a random move.\textsuperscript{19,20}

We describe the “blocking” method here in brief. Consider a series of measurements $x_1$, $x_2$, ..., $x_n$ of a quantity of interest. We divide this series into blocks. The standard error, $\sigma'$, corresponding to a given number of blocks, $n'$, is estimated:
\[
\sigma' = \sqrt{\frac{1}{n'\left(n'-1\right)} \sum_{k=1}^{n'} \left(x_k - \bar{x}\right)^2},
\]

where \( \bar{x} \) is the average computed using the complete series of measurements, and \( x_k \) is the average computed using only the measurements in the \( k \)th block.\(^{19}\) By plotting \( \sigma' \) versus the number of measurements per block, the standard error is estimated (\( \sigma_{\text{est}} \)); it corresponds to the plateau value of \( \sigma' \) (see for example Figure 6.8).\(^{21}\)

### 6.9 Using the “blocking” method to estimate error in average properties of STDR ensembles

Grossfield and Zuckerman have suggested that the “blocking method” cannot be used to estimate error in measurements from generalized-ensemble simulations.\(^{21}\) However, we propose that the “blocking method” is applicable to estimate error in the average structural properties of STDR ensembles. Similar to the series of configurations obtained from an MD or MC simulation, the configurations in our STDR ensembles exhibit significant correlation. The configurations in the ensemble are ordered in sequence by the clock time at which they were sampled. The result of this ordering is that configurations that are close in sequence often come from the same replica. Since many STDR temperature jumps are required for significant conformational change to occur, the configurations close in sequence resemble each other. The time correlation in our sequence of configurations is actually a correlation in wall clock time, not simulation time, as is normally the case for MD simulations. Because configurations close in sequence are correlated, the “blocking method” offers a meaningful estimate of the standard error.

The “blocking” method makes no assumptions concerning the nature or the underlying physical basis of the time correlation.\(^{19}\) Thus, we may apply it to obtain \( \sigma_{\text{est}} \) for the averages of quantities of interest in our STDR ensemble. To illustrate this point specifically, we present an analysis of standard error as a function of block size for the radius of gyration, \( R_g \) (Figure 6.8).
As the number of measurements of $R_g$ increases, the estimate of the standard error increases until it reaches a plateau value: this is the expected behaviour for a series of measurements with time correlation.\textsuperscript{21} The standard error as a function of block size for other structural properties of interest (hydrogen bonds, contacts, and the hydration shell) behaves the same way. Thus, the error bars for all analysis presented in Chapters 7 and 8 are obtained using “blocking”.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure6_8.png}
\caption{Estimated standard error obtained using “blocking” vs. the number of configurations in each block. \textsuperscript{20}

The estimated standard error, $\sigma'$, is shown as a function of the number of configurations in each block (axis shown on a logarithmic scale). As expected for a series of configurations with a significant time correlation, the estimated standard error reaches a plateau (at around 2000 configurations). Thus, averages from blocks containing ~2000 configurations can be considered independent measurements of $R_g$.}
\end{figure}

We note that there is currently a controversy in the literature on this point. “Blocking” is routinely used to compute error for structural properties of ensembles obtained using temperature replica exchange simulations.\textsuperscript{22} It has been suggested that the “blocking” method does not provide a reliable estimate of standard error in the case of generalized ensemble algorithms because the ensembles consist of configurations that are correlated, but not necessarily ordered in time.\textsuperscript{21} In the present study, a clear time correlation is present because of the method we use to store the configurations, which makes the “blocking” method applicable. However,
when the configurations obtained from a generalized ensemble simulation are ordered randomly, and not sequentially, the “blocking” method cannot provide a meaningful estimate of the standard error, as suggested by Grossfield and Zuckerman.²¹

6.10 Conclusions

We have presented our implementation of STDR on a large-scale computing cluster, demonstrating the computational efficiency of the approach with a detailed account of the overhead of each step in the algorithm. An atomic-level description of the disordered conformational ensemble of an elastin-like peptide was obtained. Our results push the boundaries of modern MD simulations with nearly 0.1 ms of sampling. By comparing the conformational ensembles obtained using STDR and canonical MD, the sampling enhancement provided by the random walk in temperature is clearly demonstrated. The detailed structural properties of the ensemble of conformations for this peptide are the subject of Chapter 7. We emphasize the importance of using multiple structural properties in determining the length of the “equilibration period” in STDR simulations. We also demonstrate that the conventional “blocking” method is useful in this case to obtain estimates of standard error. Both of these approaches may also be useful for other generalized-ensemble simulations for which the configurations are stored in trajectories according to the time at which they are sampled. Significantly, this study establishes a methodological framework for future applications of STDR to other disordered states of proteins.
References


CHAPTER 7

The Effects of Increasing Temperature on the Conformational Ensembles of Monomeric Elastin-like Peptides
7.1 Summary

Elastin and elastin-like peptides undergo phase separation (coacervation) upon increasing temperature. It has been proposed that the molecular basis for this phase separation involves a temperature-induced conformational change in monomers to a more ordered state. This phenomenon has been referred to as an “inverse temperature transition” because it is thought to involve a transition from a state of higher conformational entropy to a state of lower conformational entropy. We begin with a brief review of the simulation and experimental studies that suggest that conformational change in a monomer precedes elastin aggregation. This provides relevant background for the focus of this chapter, which is a detailed description of the effect of temperature on the structural properties of two monomeric systems: (1) an octapeptide, GVGVPGVG, and (2) a 35-residue peptide, (GVPGV)$_7$. Simulations of these systems were performed using STDR in combination with massive sampling; the details of the methodology of these simulations are described in Chapter 6. While we find statistically significant changes in the balance of chain-chain and chain-water interactions due to increasing temperature, neither the octapeptide nor (GVPGV)$_7$ exhibits an increase in conformational order. In fact, the opposite behaviour is seen for both peptides. With increasing temperature, the peptides preferentially populate more “expanded” conformations: we find an increase in the average end-to-end distance for the octapeptide, as well as the average radius of gyration for (GVPGV)$_7$. While water is a poor solvent for this ELP, the temperature-induced expansion of (GVPGV)$_7$ is consistent with an improvement in “solvent quality” with increasing temperature.

7.2 Introduction

7.2.1 Elastin Undergoes Phase Separation upon Increasing Temperature

In vitro, elastin and elastin-like peptides (ELPs) self-aggregate to form a second phase upon increasing temperature; this process is called coacervation, and is thought to be primarily due to self-association of hydrophobic domains.$^{1-4}$ Coacervation is a type of liquid-liquid phase separation (LLPS).$^{5-7}$ Above a specific temperature (at a given concentration), an aqueous solution of ELPs separates into two liquid phases: (1) a protein-rich phase, which is referred to as a coacervate, and (2) a water-rich phase.$^{1,3,6,7}$ Experimentally, coacervation is observed by monitoring the turbidity of the solution: when the temperature of a solution containing elastin-
like peptides is raised above a certain temperature (known as the coacervation temperature, $T_c$), a sharp rise in turbidity is observed (that is, the solution becomes “cloudy”).$^{1,6,7}$

### 7.2.2 Phase Diagrams of Elastin-like Peptides and Other Proteins

Temperature-induced phase separation is not unique to elastin – many other proteins have been shown to undergo LLPS.$^{5,8-10}$ In fact, this behaviour has important implications for x-ray crystallography, as crystal formation is thought to be enhanced in the coacervate phase.$^5$ By mapping the conditions (e.g. temperature, volume fraction) under which phase separation occurs, a phase diagram may be determined. The coexistence curve (binodal) in a phase diagram delineates the two-phase region from the single phase region. Only two phase diagrams of ELPs have been obtained to date: these studies utilized ELPs based on the elastin-like sequence motifs “VPGVG”$^7$ and “VPGG”$^6$ (Figure 7.1 A and B). Both of these phase diagrams show a two-phase region above the binodal, and a single phase region below.$^{6,7}$ Together, these phase diagrams show that ELPs undergo LLPS with increasing temperature. However, most proteins for which LLPS has been observed undergo phase separation with decreasing temperature.$^5$ For example, the protein $\gamma$-crystallin (a mammalian lens protein) forms a coacervate below $37^\circ$C (see Figure 7.1 C).$^8$ Unlike the phase diagrams of ELPs, the two phase region in the phase diagram of $\gamma$-crystallin is found below the binodal. Interestingly, similar to the phase separation behaviour of the ELPs, an increase in temperature has also been shown to cause LLPS in sickle cell hemoglobin; sickle cell anemia was the first disease linked to protein aggregation.$^9,10$ Protein aggregation has since been associated with many tissue degenerative pathologies, such as Alzheimer’s disease. Furthermore, it is becoming increasingly clear that an understanding of the mechanism of protein aggregation can benefit from viewing it as a type of phase separation.$^{11}$
Figure 7.1 Phase Diagrams for Protein Liquid-Liquid Phase Separation (LLPS)

In each of the protein phase diagrams in (A) to (D), the axes correspond to temperature (in °C) and volume fraction. Each figure was reproduced with permission.6-8,11 (A) The phase diagram for an ELP, (VPGVG)$_{251}$.7 The empty squares are points on the binodal line, and the black filled circles are points on the spinodal line (the two phase system is unstable in the region between the binodal and the spinodal lines).11 (B) The phase diagram of (VPGG)$_{300}$ (polydisperse).6 Black triangles and empty circles define the binodal and spinodal lines, respectively. (C) The phase diagram of γIII A-crystallin.8 (D) The phase diagram of a simplified heteropolymer model of a protein.12 The shape of the diagram is that of an “asymmetric closed loop”, and the two phase region is found within the interior of the loop. The shape of this loop is strongly affected by small variations in the fractional content of nonpolar residues, Φ.

A general theory for protein phase separation was developed by Fields et al.12 They utilized a simple lattice model that describes proteins as heteropolymers with two types of residues: hydrophobic and hydrophilic. Using this model, they predicted that the phase diagram for protein aggregation has the shape of an asymmetric closed loop (Figure 7.1 D).
Fields et al proposed that the conformations of polypeptide chains within the protein-rich phase have significant intermolecular hydrophobic interactions. Because the individual chains are thought to form extensive interactions with neighbouring chains, their conformations are primarily determined by local interactions, and contain few intramolecular nonlocal interactions. Furthermore, using the language of polymer physics to describe disordered states of proteins (see references 11 and 14 for excellent reviews), the polypeptide chains are predicted to reach a “theta-like” state when the protein-rich phase contains little to no water (that is, they behave as they would in a theta (θ) solvent, which is also sometimes referred to as an ideal solvent). It is in this state that polymer chains have maximal chain entropy. This is essentially a statement of the Flory Theorem.

The comparison of the protein phase diagrams of ELPs and other proteins in Figure 7.1 leads directly to several fundamental questions: (1) Why do elastin and ELPs self-aggregate and undergo phase separation with increasing temperature in contrast to the phase separation behaviour typical of most other proteins? (2) Do polypeptide chains within protein aggregates adopt the same conformational properties as they would in a θ solvent, as predicted by Fields et al?; and (3) To what extent do the predictions of the Flory theorem hold for real proteins, and to what extent do they hold for ELPs? The precise molecular mechanism of elastin and ELP phase separation is not well understood, though several models have been proposed. Understanding the coacervation of elastin is clearly a matter of practical interest: the capacity of elastin and ELPs to undergo temperature-controlled self-association makes them ideal candidates for incorporation into novel biomaterials. Knowledge of the molecular details underlying elastin self-aggregation may provide fundamental insight to facilitate rational biomaterial design. These questions and goals together form the motivation for the studies presented in Chapters 7 and 8. In the present work, it is not our objective to characterize the phase transition of elastin and elastin-like peptides, which would require simulations at much longer length and time scales than those currently accessible to atomistic MD simulations. Instead, it is our goal to characterize the conformations of an elastin-like peptide in aqueous solution and within an aggregate, which may be representative of the conformations accessible to the polypeptide chain in the phase-separated state.
7.2.3 Coacervation is Thought to be an Inverse Temperature Transition

Coacervation is generally thought to involve an increase in the order of elastin and ELPs with increasing temperature.19,25-28 Because this postulated ordering of elastin is in contrast to the usual effect of temperature (that is, to favour states of higher entropy), coacervation has been called an inverse temperature transition (ITT). The idea that elastin’s phase separation is a process of ordering was first introduced by Urry et al in 1969.25 They observed a temperature-dependent change in the CD spectrum of elastin, and interpreted this change to be the result of an increase in α-helical content upon raising the temperature.25 More generally, the ITT model put forth by Urry and others claims that increasing temperature induces an increase in the conformational order of elastin and ELPs on two different length scales: (1) at the mesoscopic level, an increase in order is observed in the formation of a fibrillar structure accompanying phase separation, and (2) at the molecular level, temperature induces an increase in order corresponding to the formation of secondary structure.18,25,29 Urry’s model was later revised to suggest that the increase in order is not due to increasing content of α-helix, but rather the formation of a structure called the β-spiral.30,31 The β-spiral consists of a repetition of β-turns formed by the residues VPGV. These turns are stabilized by a hydrogen bond between the C=O group of the first valine and the N-H group of the fourth valine (Figure 7.2). Urry et al suggested that the population of the VPGV β-turn increases with temperature, and that this change in structural preferences is related to coacervation.29 The molecular mechanism for coacervation in the ITT model involves a conformational change in a monomer to an aggregation-prone state,28,32 and the entropy of the entire system is thought to increase because the increase in entropy of the water molecules expelled upon aggregation is greater than the loss of entropy due to the ordering of the polypeptide chain.29

7.3 Investigations of the Molecular Basis for Coacervation Using Molecular Simulation

The β-spiral model for elastin’s structure is no longer widely accepted. This is in part due to Li and Daggett’s MD simulation study of an ELP monomer, (VPGVG)18, which used the β-spiral structure as a starting configuration.19 In less than 10 ns, the β-spiral structure is lost, and the peptide adopts a collapsed, amorphous structure.19 A more recent study of the ELP GVG(VPGVG)3 confirmed Li and Daggett’s observation: they also observed that the β-spiral
structure rapidly changes to an amorphous structure.\textsuperscript{25} Based on these two studies, the β-spiral is an unstable conformation, even for a perfectly repetitive ELP based on the motif “VPGV”. These results are consistent with the current view of elastin as an intrinsically disordered protein, which is reviewed in Chapter 3. Li and Daggett also found that the (VPGVG)\textsubscript{18} peptide adopts significantly more collapsed structures with more β-turns at temperatures above 20–40 °C; they suggest that these observations may be consistent with ITT-like behaviour.\textsuperscript{19}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig72}
\caption{Structure of a VPGV β-turn}
\end{figure}

A configuration of the peptide (GVPGV)\textsubscript{7} is shown with a hydrogen-bonded β-turn. The residues involved are V, P, G, and V, with a hydrogen bond formed between the C=O group of valine (residue $i$) and the N-H group of valine (residue $i+3$). Only the residues near the turn are shown. This configuration was randomly selected from the trajectory at 300 K of the STDR simulation described in Chapter 6.

The smallest ELP proposed to exhibit an ITT is the octapeptide, GVGVPGVG.\textsuperscript{28} Based on changes in the CD spectrum with temperature, Reiersen \textit{et al} suggested that GVGVPGVG undergoes a β-sheet to β-turn transition, and that this transition constitutes an ITT.\textsuperscript{28} Because of its small size, the octapeptide represents the most tractable system to study using molecular simulations. For this reason, almost all statistically-meaningful MD simulations probing the effect of temperature on ELPs have used GVGVPGVG as their model system.\textsuperscript{20-22,33} Taken together, the results of these studies suggest that GVGVPGVG undergoes a structural transition
in the temperature range of 20 to 40 °C. Below approximately 20 °C, the octapeptide preferentially populates a more open conformation; above 40 °C, a closed conformation becomes increasingly populated. Principal component analysis showed that the eigenvector with the largest eigenvalue corresponds to an opening/closing motion. The proline residue in position 5 acts as a hinge governing this opening/closing motion.

Molecular simulations of ELPs have, for the most part, investigated monomers (except for the simulations of aggregates described in Chapter 4). Even though coacervation by definition involves protein aggregation, there have been three reasons why simulations probing the molecular basis for coacervation have been limited to monomeric systems. First, as is the case for all molecular simulations, increasing system size rapidly increases computational costs. Second, the effects of temperature on intramolecular interactions in a monomer have been hypothesized to "predict" the effect of temperature on intermolecular interactions in an aggregate. Thus, simulations of monomers have been thought to be sufficient in the case of elastin to understand the process of coacervation. Third, the experimental studies of Urry et al., Reiersen et al., and Nicolini et al. suggest that a monomer of elastin undergoes a specific conformational change at or near the coacervation temperature, and that this conformational change is related to aggregation. These studies motivated several groups to search for a conformational transition in a monomer. ITT-like behaviour has been ascribed to many different structural changes in monomers of ELPs: (1) the formation of β-turns and more collapsed states, (2) changes in the first principle component (an opening/closing motion) of an octapeptide, and (3) a change from a rigid to a flexible state. In summary, many different models have been proposed to explain elastin’s temperature-induced self-aggregation, but no conclusive molecular mechanism has been established.

7.4 Motivation and Rationale

In order to investigate the molecular driving forces for the self-aggregation of elastin and ELPs, we use atomistic molecular simulations to determine the effects of temperature on the structural properties of ELPs (monomers in Chapter 7, and an aggregate in Chapter 8). In particular, we address the following questions:
1. How are the structural properties (e.g. chain dimensions, chain-chain, and chain-water interactions) of ELP monomers affected by increasing temperatures?

2. How are the structural properties of ELPs affected by aggregation? That is, how does the conformational ensemble of ELP monomers within the aggregate differ from that of monomers in water?

Based on the results of prior studies, we began with the hypothesis that a monomer of elastin undergoes a temperature-induced change in structural properties. Using large-scale STDR simulations of two ELPs (GVGVPGVG and (GVPGV)\textsubscript{7}), we tested this hypothesis. Our results suggest that both of these ELPs exhibit the opposite behaviour compared to an ITT: they undergo temperature-induced expansion, and not collapse. These results suggest that increasing temperature effectively improves the “quality” of water as a solvent, and that models for coacervation involving a temperature-induced collapse in monomers should be revisited. We provide a detailed characterization of the effect of temperature on chain-chain (hydrogen bonding and nonpolar) interactions and chain-water interactions. Finally, we note that the structural properties of monomers of ELPs presented in this chapter should be placed in context with the results of Chapter 8 in which we present the results of a large scale STDR simulation of an elastin-like aggregate.

7.5 Methods

In this chapter, we present the results of STDR simulations of two systems: (1) the octapeptide, GVGVPGVG, and (2) the 35-residue peptide, (GVPGV)\textsubscript{7}. All simulation details are described in Chapter 6 and Appendix C. We have chosen to study systems based on the sequence motif “GVPGV”, which is derived from exon 24 of chicken elastin,\textsuperscript{35} because peptides based on this motif have been the most extensively characterized using both experiment and simulation. In addition, an ELP containing the GVPGV motif is one of only two ELPs for which the temperature-volume fraction phase diagram has been measured (Figure 7.1 A).\textsuperscript{7} We began the present study of the effect of temperatures on monomers using GVGVPGVG as a test system because prior studies had reported meaningful temperature-induced structural changes.\textsuperscript{20-}
We also performed simulations of (GVPGV)$_7$ because it is comparable in length to a typical hydrophobic domain, and because a direct comparison of the two monomeric systems affords insights into the length dependence of various structural properties.

7.6 Results and Discussion

7.6.1 The Octapeptide Becomes More “Open” With Increasing Temperature

In Figure 7.3 A, we show the temperature dependence of the ensemble average of the end-to-end distance, $<R_{eed}>$, for the octapeptide. Increasing temperature results in a slight overall increase in the average end-to-end distance, which is due to the fact that closed states (that is, states for which $R_{eed}$ are less than 7 Å) become less populated at higher temperatures, as seen in the probability distribution of $R_{eed}$ (Figure 7.3 B). A temperature-dependent increase in the average end-to-end distance of the octapeptide is consistent with the results of a prior simulation study. These results are not consistent with the octapeptide becoming more closed or ordered with increasing temperature, as predicted based on CD data and earlier simulation studies.

It has also been proposed that the structural properties (and the effect of temperature on these properties) of an octapeptide in water will resemble those of octapeptide fragments within a longer peptide. We test this hypothesis by comparing the probability distribution of $R_{eed}$ for the octapeptide in water (Figure 7.3 B) with that of octapeptide fragments within (GVPGV)$_7$ (Figure 7.4 B). There are five instances of the octapeptide sequence GVGVPGVG within the sequence of the peptide (GVPGV)$_7$. The average over these fragments is used to obtain the results shown in Figure 7.4. At lower temperatures, closed states are more populated for the octapeptide in water than the octapeptide fragments in (GVPGV)$_7$. For example, at 266 K, states with $R_{eed} < 7$ Å have populations of 26.6 % and 13.1 % for the octapeptide and the octapeptide fragments within (GVPGV)$_7$, respectively. Temperature has no statistically significant effect on $<R_{eed}>$ for octapeptide fragments within (GVPGV)$_7$ (Figure 7.4 A). These results suggest that the conformational preferences of the octapeptide are not the same as those of octapeptide fragments within a longer peptide.
Figure 7.3  The effect of temperature on the end-to-end distance, $R_{EED}$, of the octapeptide

(A) The ensemble average of the end-to-end distance, $<R_{EED}>$, is shown as a function of temperature. Error bars represent the estimated standard error using “blocking” (see Chapter 6, section 6.9). Error bars in all subsequent figures were also computed using “blocking”. Increasing temperature results in a slight overall increase in $<R_{EED}>$.  

(B) The probability distribution of $R_{EED}$ is shown for temperatures between 266 K and 429 K according to the colour scheme on the right. With increasing temperature, more “open” states of the octapeptide are favoured over “closed” states (states for which $R_{EED} < 7$ Å).
Figure 7.4  The effect of temperature on the end-to-end distance, $R_{EED}$, of the octapeptide fragments within (GVPGV)$_7$

Here, we show the same analysis as Figure 7.3 for the “average octapeptide” within (GVPGV)$_7$ (that is, based on the average value of $R_{EED}$ of the five instances of the sequence GVGVPVGVG within the sequence (GVPGV)$_7$). (A) The ensemble average of end-to-end distance, $<R_{EED}>$, is shown as a function of temperature. There is no statistically significant change in $<R_{EED}>$ with increasing temperature. (B) The probability distribution of $R_{EED}$ is shown for temperatures between 266 K and 429 K according to the colour scheme on the right. Temperature has no significant effect on this distribution. Furthermore, the shape of the distribution differs from that of Figure 7.3 B, with a decreased population of “closed” states.
In addition to its effect on the end-to-end distance, increasing temperature also modulates the populations of the different types of hydrogen bonds formed by the backbone. (Note, the definitions of hydrogen bonds and hydrogen-bonded turns are provided in Appendix C, Table C.1.) First, the total number of peptide-peptide hydrogen bonds per backbone hydrogen bonding group, $x_{\text{HB}}$, undergoes no statistically significant changes with increasing temperature (Figure 7.5 A). The reason why $x_{\text{HB}}$ is unaffected by temperature is because the changes in the populations of different types of hydrogen-bonded turns effectively “cancel out” (Figure 7.5 B). While the population of $\gamma$-turns increases linearly with temperature, the population of $\beta$-turns and other turns decrease. This “trade” of one type of hydrogen-bonded turn for another with increasing temperature may be rationalized on the basis of chain entropy. A $\gamma$-turn is stabilized by a hydrogen bond between the C=O group of residue $i$ and the N-H group of residue $i+2$, and it therefore limits the conformation of three residues. On the other hand, a $\beta$-turn constrains four residues in order to form a hydrogen bond between residues $i$ and $i+3$. Both types of turns result in the formation of one hydrogen bond, and thus offer essentially equivalent energetic stabilization. However, the formation of a $\gamma$-turn has a less significant effect on chain entropy than the formation of a $\beta$-turn, and therefore becomes preferentially populated (lower in free energy) with increasing temperature. The existence of an equilibrium between $\gamma$-turns and $\beta$-turns is consistent with a study of peptides containing the VPGV fragment using NMR, CD, and IR spectroscopy. \(^{36}\) In addition to its effect on peptide-peptide hydrogen bonds, increasing temperature also causes a linear decrease in backbone-water hydrogen bonds (Figure 7.5 C), which is consistent with other studies of the octapeptide. \(^{22,33}\) Glaves et al observed a gradual displacement of the amide chemical shifts with increasing temperature in the range 288 K to 328 K. \(^{33}\) Since the amide protons of each residue exhibit a similar temperature-dependent change, they attributed this shift to be due to backbone-water interactions, and not due to intramolecular hydrogen bonding. \(^{33}\)
Figure 7.5  The effect of temperature on hydrogen bonds in the octapeptide (GVGVPGVG)
(A) The average number of peptide-peptide hydrogen bonds per backbone hydrogen bonding
group, $x_{HB}$, vs. temperature, $T$.  (B) The probability ($x_{\text{turn}}$) of a hydrogen bonding group
participating in different types of turns (β-turns, in red; γ-turns, in blue; and all other turns, in
purple) vs. $T$.  (C) The number of peptide-water hydrogen bonds per hydrogen bonding group,$x_{HB,\text{H2O}}$, is shown as a function of temperature, with a linear fit (red line).  The equation of this
line and the corresponding $R^2$ coefficient are also shown.
Figure 7.6  The effect of temperature on nonpolar interactions in the octapeptide
(A) The number of nonpolar contacts per residue, $x_{NP}$, is shown as a function of temperature, T. (B) The number of water molecules in the nonpolar hydration shell per residue, $x_{NP,\text{shell}}$, vs. T. Nonpolar contacts and the nonpolar hydration shell are defined in Appendix C.

We also examine the effect of temperature on chain-chain nonpolar interactions and the interactions of nonpolar groups with water (definitions of these structural properties are provided in Appendix C, sections C.4.3 and C.5.2). Increasing temperature causes no significant change in the number of nonpolar contacts per residue, $x_{NP}$ (Figure 7.6 A). The number of water molecules in the NP hydration shell, $x_{NP,\text{shell}}$, decreases by about 2 % over the temperature range from 266 K to 350 K, after which there is no change. A decrease of 2 % in $x_{NP,\text{shell}}$ corresponds to a loss of only 2 water molecules on average: this is a very small effect (Figure 7.6 B). Even though the majority of the residues in the octapeptide are nonpolar, it is too small to be able to
bury significant hydrophobic surface or form a hydrophobic core. As a result, it undergoes no significant changes in nonpolar interactions or chain dimensions with increasing temperature.

Next, we return to one of our original questions: does GVGVPVG undergo an ITT? Our results concerning the effect of temperature on structural properties are not consistent with ITT-like behaviour. We find no structural property that undergoes a dramatic or significant change at a particular temperature, or any parameter that might indicate an increase in conformational order with temperature. The temperature-dependence of chain-chain and chain-water interactions of the octapeptide presented in this chapter is representative of that of other elastin-like and amyloid-like octapeptides, such as GVPGVPGV, GVGVGGVG, GVGVAGVG, Q₈, and A₈ (Rauscher & Pomès, unpublished results). In particular, these octapeptides exhibit a similar loss in β-turns and increase in γ-turns as shown in Figure 7.5 B, and a similar linear decrease in peptide-water hydrogen bonds with temperature (Figure 7.5 C). The fact that the same temperature-induced changes in structural properties occur in both elastin-like and amyloid-like octapeptides suggests that these structural changes cannot provide an adequate explanation for the molecular basis of elastin coacervation. Taken together, these results strongly suggest that the octapeptide does not undergo an ITT or become ordered. One possible reason for the disagreement between the present study and earlier studies may be due to differences in statistical sampling. With less sampling, we saw similar effects of temperature on structural properties, but these effects disappear when adequate statistical sampling is obtained.

7.6.2 The Radius of Gyration of (GVPGV)₇ and the Importance of Removing the “Equilibration Phase” in Generalized-Ensemble Simulations

Like the octapeptide, (GVPGV)₇ also adopts more “expanded” configurations with increasing temperature. The ensemble average of the radius of gyration, <R_g>, increases with increasing temperature (Figure 7.7 A), which is due to a gradual and slight shift in the probability distribution of R_g towards larger values (Figure 7.7 B).
Figure 7.7  The effect of temperature on the radius of gyration, $R_g$, of (GVPGV)$_7$

(A) Average radius of gyration, $\langle R_g \rangle$, vs. temperature, $T$.  (B) Probability distribution of $R_g$ for ten temperatures between 266 K and 429 K (colour scheme indicated on the right).

Importantly, this effect is only seen when the “equilibration” period is removed from the initial part of the STDR simulation (Figure 6.6). Surprisingly, this required that the first 40 % of the simulation be discarded in order for the probability distribution of the radius of gyration distribution to become stationary (Figure 6.7). Without removing the “equilibration period”, the plot of $\langle R_g \rangle$ vs. temperature shows an initial decrease, followed by a plateau, then an increase. This result would imply that (GVPGV)$_7$ undergoes a temperature-induced collapse,
but it is actually due to inadequate sampling at the lower temperatures. Significantly more simulation time is required for the ensembles at low temperature to reach their equilibrium size, and when this effect is not accounted for, a temperature-induced collapse is observed. It is possible that a similar effect occurred in a prior study of the effect of temperature on the structure of the ELP (VPGVG)\textsubscript{18}.\textsuperscript{19} With 9 ns of simulation at each temperature, and a starting structure modeled after the β-spiral, \(R_g\) was found to be significantly lower for temperatures above 40 °C. In the present work, we find that three orders of magnitude more sampling (combined with enhanced sampling) is necessary to achieve convergence of average properties.

7.6.3 The Radius of Gyration of (GVPGV)\textsubscript{7}, Intra-Chain Distance Scaling, and Solvent Quality

With increasing temperature, water acts as a “better” solvent of the polypeptide chain, the result of which is an overall expansion of the chain (Figure 7.7). This observation is in agreement with the view that temperature acts as a kind of “denaturant” of polypeptide chains.\textsuperscript{14} Solvent quality is described by the interaction parameter, \(\chi\), which represents the free energy of interaction between a monomer of the chain and the solvent (divided by RT such that it is dimensionless).\textsuperscript{14} For homopolymers, \(\chi\) is equal to \(\frac{1}{2}\) for an “ideal” (“theta”) solvent, less than \(\frac{1}{2}\) for a “poor” solvent, and greater than \(\frac{1}{2}\) for “good” solvent. “Good” solvents and “poor” solvents have chain-solvent interactions that are more favourable and less favourable than those of a “theta” solvent, respectively.\textsuperscript{14} An average \(\chi\) parameter for nonpolar residues is predicted to reach a maximum near room temperature.\textsuperscript{37} Using the temperature dependence of \(\chi\), a study modeling polypeptides as simplified heteropolymers predicted that polypeptide chains undergo a temperature-induced expansion, which is consistent with the role of temperature as a kind of “denaturant”.\textsuperscript{37} The results of our atomistic simulations are qualitatively consistent with the predictions of this simplified model. In particular, a monomeric (GVPGV)\textsubscript{7} chain in water undergoes a temperature-induced increase in the average radius of gyration (Figure 7.7 A) because the populations of more expanded states increase (refer to the probability distribution of the radius of gyration, Figure 7.7 B). The intrachain distance scaling (Figure 7.8) also suggests a temperature-induced increase in solvent quality. Consistent with the observation of a temperature-induced expansion, there is an upward trend in the scaling of the average internal distances (Figure 7.8).
Figure 7.8  The effect of temperature on the average intrachain distance scaling

For two residues $i$ and $j$, the ensemble average of the separation distance, $<R_{ij}>$, is:

$$<R_{ij}> = \frac{1}{n_{ij}} \sum_{k \in i} \sum_{l \in j} (\bar{r}_{ik} - \bar{r}_{jl})$$

(7.1)

where $R_{ij}$ is the average separation distance of all heavy atoms pairs in residues $i$ and $j$, and $n_{ij}$ is the number of unique pairwise distances. The indices $k$ and $l$ are for the heavy atoms in residue $i$ and $j$, respectively. $|i - j|$ is the number of residues separating residues $i$ and $j$. The black curve represents the scaling expected for a $\theta$ solvent ($<R_{ij}> = R_o |i - j|^{1/2}$ where $R_o = 4.7 \text{ Å}$).

For polymer chains in a $\theta$ solvent, the average distance between two residues $i$ and $j$ is predicted to be proportional to $|i - j|^{1/2}$. In contrast, for collapsed conformational ensembles in a “poor” solvent, the average intrachain distances are expected to be invariant to the sequence separation $|i - j|$. At low temperatures, the intrachain distance scaling of (GVPGV)$_7$ is consistent with that expected for a collapsed state. Increasing temperature causes a gradual “improvement” in solvent quality: the intrachain distance scaling approaches that expected for a chain in an ideal solvent. However, water is still acting as a “poor” solvent, even up to the highest temperature in our STDR temperature scale, 749 K. That is, the scaling of internal distances even at this high temperature has not yet reached the expected scaling for a polymer chain in an ideal ($\theta$) solvent, corresponding to the black line in Figure 7.8.
7.6.4 The Effect of Temperature on Nonpolar Interactions in (GVPGV)$_2$

Increasing temperature has a significant effect on the balance of chain-chain and chain-water interactions for the (GVPGV)$_7$ peptide. Chain-chain nonpolar interactions ($x_{NP}$) increase slightly (by about 2%) between 266 K and 310 K (Figure 7.9 A). However, given the estimated standard error, this increase is not statistically significant. Above 310 K, there is a monotonic decrease in $x_{NP}$ with increasing temperature. The effect of temperature on the number of water molecules in the nonpolar hydration shell per residue, $x_{NP\ shell}$, is exactly the opposite of that of $x_{NP}$ (comparing Figures 7.9 A and B). Below 310 K, there is a loss of about 2% of the water molecules in the nonpolar hydration shell, followed by an increase (Figure 7.9 B). The effect of temperature on $x_{NP}$ and $x_{NP\ shell}$ is qualitatively consistent with the temperature dependence of the “solvophobicity” of nonpolar residues in water. In Figure 7.9 C, we show a two-dimensional (2D) plot relating $x_{NP\ shell}$ and $R_g$. At low temperatures, there is essentially no change in the overall size of the structures in the ensemble, and there is a very small expulsion of water molecules. Similar to the plots in Figure 7.9 A and B, the “turning point” in the $x_{NP\ shell}$ vs. $R_g$ plot occurs at around 310 K. Above this temperature, the chain begins to expand, and this causes an increase in $x_{NP\ shell}$ as the nonpolar groups of the sidechains become more exposed to water. The same effect causes the decrease in chain-chain nonpolar interactions above 310 K (Figure 7.9 A). Below 310 K, the peptide maintains a constant $R_g$, but undergoes slight conformational rearrangements that facilitate a decrease in the number of water molecules associated with nonpolar groups, with a concomitant increase in chain-chain nonpolar contacts. There are two opposing “forces” at work here: the hydrophobic effect favours collapse, while chain entropy strongly opposes further collapse. The strength of both effects increases with increasing temperature. The fact that (GVPGV)$_7$ (a nonpolar polypeptide) expands with increasing temperature suggests that the effect of chain entropy is dominant. This highlights the important role of chain entropy in governing the ensemble of configurations: while it is preferable to minimize exposure of nonpolar groups, the (GVPGV)$_7$ peptide is unable to undergo collapse.
Figure 7.9  The effect of temperature on nonpolar interactions in (GVPGV)$_7$

(A) The average number of nonpolar contacts per residue, $x_{NP}$, vs. temperature, $T$.  (B) The average number of water molecules in the nonpolar hydration shell per residue, $x_{NP\ shell}$, vs. $T$.  (C) $x_{NP\ shell}$ vs. $R_g$.  The semi-transparent squares around each data point indicate the estimated standard error in each dimension (this is also the case for the 2D plots in Figures 7.12 and 7.13).  The data points represent the average values of $x_{NP\ shell}$ and $R_g$ for the ensembles at each temperature.  Adjacent temperatures are connected by lines.  Detailed definitions of nonpolar contacts and the nonpolar hydration shell are provided in Appendix C.
7.6.5 The Effect of Temperature on Hydrogen-Bonded Interactions in (GVPGV)$_2$

Increasing temperature causes an overall decrease in both peptide-peptide and peptide-water hydrogen bonds, but the rates at which they decrease differ significantly from each other. At low temperatures (up to about 310 K), $x_{\text{HB}}$ remains constant (Figure 7.10 A). Above 310 K, $x_{\text{HB}}$ decreases. The shape of the $x_{\text{HB}}$ vs. $T$ curve is “concave down”. In contrast, the rate at which peptide-water hydrogen bonds are lost with increasing temperature is initially higher, then decreases (Figure 7.10 B). Accordingly, the shape of the $x_{\text{HB, H2O}}$ vs. $T$ curve is “concave up”. While these two types of hydrogen bonds exhibit supralinear and superlinear decreases with temperature, the total number of hydrogen bonds formed by each backbone group ($x_{\text{HB, total}} = x_{\text{HB}} + x_{\text{HB, H2O}}$) decreases linearly with temperature (Figure 7.10 C). The loss of hydrogen bonds with temperature is consistent with a similar effect of temperature on the populations of hydrogen bonds in water.\textsuperscript{39} Interestingly, this linear loss of hydrogen bonds occurs because the probability of peptide-water hydrogen bonds decreases at a faster rate in the low temperature range (where $x_{\text{HB}}$ is constant). The reason why the total number of peptide-peptide hydrogen bonds remains constant up to about 310 K is because of a “trade” of nonlocal hydrogen bonds for local hydrogen bonds (Figure 7.11 A). The increase in local hydrogen bonds and decrease in nonlocal hydrogen bonds is analogous to the effect of temperature on hydrogen-bonded turns in the octapeptide (Figure 7.5 B): $\gamma$-turns are favoured over $\beta$-turns because fewer residues must be constrained to achieve the energetic stabilization of a hydrogen bond. A similar argument based on chain entropy may be applied here: local interactions are increasingly favoured over nonlocal interactions at higher temperatures because they effectively constrain fewer residues. Similarly, there is a “trade off” of nonlocal nonpolar interactions for local nonpolar interactions with increasing temperature (Figure 7.11 B).
Figure 7.10  The effect of temperature on hydrogen bonds in (GVPGV)$_7$

Overall, increasing temperature results in a net loss of hydrogen bonds. (A) The average number of peptide-peptide hydrogen bonds formed by each backbone hydrogen bonding group, $x_{HB}$, versus temperature, $T$. (B) The average number of peptide-water hydrogen bonds per group, $x_{HB, H2O}$, versus $T$. (C) The total number of hydrogen bonds per group ($x_{HB, total} = x_{HB} + x_{HB, H2O}$) versus $T$. In each case, the curve is fit by a line, and the equation of the line and corresponding $R^2$ coefficient are shown.
Overall, increasing temperature increases the population of local interactions and decreases the population of nonlocal interactions. Local interactions involve residues separated by fewer than 8 residues, and nonlocal interactions involve residues separated by 8 or more residues. (A) Average number of peptide-peptide hydrogen bonds per group, $x_{HB}$, versus temperature, T. Local $x_{HB}$ is shown in red, and nonlocal $x_{HB}$ is shown in blue. (B) Average number nonpolar contacts per residue, $x_{NP}$, vs. T. Local and nonlocal interactions are shown in red and blue, respectively.

7.6.6 The Effect of Peptide Length on Structural Properties: Comparing GVGVPVG and (GVPGV)$_7$

We now perform a direct comparison between the two monomeric systems (the octapeptide and (GVPGV)$_7$) with respect to the effect of temperature on chain-chain and chain water interactions. First, we consider the nonpolar interactions (Figure 7.12). For the octapeptide, increasing temperature from 266 K to 438 K results in no statistically significant change in $x_{NP}$ and a very slight decrease in $x_{NP, shell}$ (Figure 7.12 A). Specifically, the average
number of water molecules in the nonpolar hydration shell is 98 ± 1 at 266 K, and 95.7 ± 0.2 at 438 K. In contrast, for (GVPGV)$_7$, increasing temperature results in a small but significant change in both $x_{NP}$ and $x_{NP, \text{shell}}$ (Figure 7.12 B). The average number of water molecules in the nonpolar hydration shell decreases from 253 ± 2 (at 266 K) to 245.4 ± 1 (at 309 K); it then increases to 251 ± 0.6 (at 434 K). Clearly, this effect is a very small, but is nevertheless statistically significant. As described earlier (in section 7.5.4), because of chain entropy, the chain is unable to collapse with increasing temperature, and thus is able to “expel” only a small number of water molecules that are in contact with the nonpolar side chains. In Figure 7.12 C, we show the 2D $x_{NP}$ vs. $x_{NP, \text{shell}}$ plots of the octapeptide and (GVPGV)$_7$ on the same scale for comparison. Because of its larger size, (GVPGV)$_7$ is able to form significantly more nonpolar self-interactions than GVGVPGVG. For example, at 300 K, $x_{NP}$ is 2.05 ± 0.02 and 0.30 ± 0.02 for (GVPGV)$_7$ and the octapeptide, respectively. Because many of the nonpolar groups are involved in chain-chain interactions, (GVPGV)$_7$ has significantly fewer water molecules in the nonpolar hydration shell per residue than the octapeptide (at 300 K, $x_{NP, \text{shell}}$ is 7.06 ± 0.01 and 12.20 ± 0.06 for (GVPGV)$_7$ and the octapeptide, respectively).

Next, we consider the hydrogen bonding interactions (Figure 7.13). Both the octapeptide (Figure 7.13 A) and (GVPGV)$_7$ (Figure 7.13 B) undergo a statistically significant loss in backbone hydration ($x_{HB, H2O}$) with increasing temperature. Between 266 K and 310 K, $x_{HB, H2O}$ decreases by 0.05 ± 0.01 and 0.03 ± 0.01 for (GVPGV)$_7$ and the octapeptide, respectively. However, the octapeptide continues to lose backbone hydration as temperature is increased above 310 K, while $x_{HB, H2O}$ remains essentially constant above 310 K for (GVPGV)$_7$. This “hook” shape in the plot of $x_{HB}$ vs. $x_{HB, H2O}$ for (GVPGV)$_7$ is related to the “concave down” and “concave up” temperature trends of $x_{HB}$ and $x_{HB, H2O}$ (Figure 7.10 A and B). At low temperature, peptide-water hydrogen bonds decrease in population, while at high temperature, peptide-peptide hydrogen bonds decrease in population; around 310 K is the temperature at which the trend changes (thus, 310 K is the corner of the “hook”). It is also at 310 K that the (GVPGV)$_7$ peptide begins to expand (Figure 7.7 A). Taken together, these results suggest the following: (1) the hydrophobic effect favours collapse and self-interactions, (2) below 310 K, the
Figure 7.12  The effect of temperature on nonpolar interactions in \((\text{GVPGV})_7\) and \((\text{GVGVPGVG})\) 
(A), (B) and (C) each show 2D plots of \(x_{\text{NP}}\) vs. \(x_{\text{NP}, \text{shell}}\). Each point in the plot corresponds to the average value of \(x_{\text{NP}}\) and \(x_{\text{NP}, \text{shell}}\) for the ensemble of configurations at a given temperature. Points corresponding to adjacent temperatures are connected by lines, and the squares surrounding the points depict the estimated standard error in both dimensions. (A) \(x_{\text{NP}}\) vs. \(x_{\text{NP}, \text{shell}}\) for the octapeptide. (B) \(x_{\text{NP}}\) vs. \(x_{\text{NP}, \text{shell}}\) for \((\text{GVPGV})_7\). (C) \(x_{\text{NP}}\) vs. \(x_{\text{NP}, \text{shell}}\) for both \((\text{GVGVPGVG})\) and \((\text{GVPGV})_7\) (plots on the same scale for comparison).
Figure 7.13  The effect of temperature on hydrogen bonds in (GVPGV)$_7$ and GVGVPGVG
(A), (B) and (C) each show 2D plots of $x_{\text{HB}}$ vs. $x_{\text{HB, H}_2\text{O}}$. Each point in the plot corresponds to
the average value of $x_{\text{HB}}$ and $x_{\text{HB, H}_2\text{O}}$ for the ensemble of configurations at a given temperature.
Points corresponding to adjacent temperatures are connected by lines, and the squares
surrounding the points depict the estimated standard error in both dimensions. (A) $x_{\text{HB}}$ vs. $x_{\text{HB, H}_2\text{O}}$ for the octapeptide. (B) $x_{\text{HB}}$ vs. $x_{\text{HB, H}_2\text{O}}$ for (GVPGV)$_7$. (C) $x_{\text{HB}}$ vs. $x_{\text{HB, H}_2\text{O}}$ for both
GVGVPGVG and (GVPGV)$_7$ (plots on the same scale for comparison).
peptide remains collapsed and chain-chain interactions increase or remain constant, and (3) above 310 K the peptide expands, and chain-water interactions increase. For these reasons, the hydrophobic effect appears to influence the balance of chain-chain and chain-water hydrogen bonds formed by the polypeptide backbone of (GVPGV)$_7$ (Figure 7.13 B).

7.7 Conclusions

In summation, we find that the octapeptide is not well-suited to elucidate the molecular basis for elastin’s temperature-induced self-aggregation. It is too small for its conformational equilibrium to be significantly affected by the hydrophobic effect (it cannot bury nonpolar groups or undergo collapse). It therefore exhibits significantly different temperature-dependent behaviour compared to (GVPGV)$_7$ (a peptide of similar length to elastin’s hydrophobic domains) with respect to both hydrogen-bonding interactions and nonpolar interactions. Furthermore, our results are not consistent with the view that this peptide undergoes an ITT.

Similarly, we find no evidence for an ITT in the longer peptide, (GVPGV)$_7$. Increasing temperature causes an overall expansion of the chain. Importantly, we only see this effect of temperature when we carefully account for equilibration. It is often assumed that the use of an enhanced sampling or generalized-ensemble algorithm (most commonly replica exchange) is sufficient to guarantee adequate conformational sampling from the beginning of the simulation. The results of the current study suggest that the delineation of “equilibration” is essential for simulations utilizing generalized-ensemble algorithms. Just as in conventional MD simulations, the starting configurations in simulations utilizing generalized-ensembles are not necessarily representative of the final, equilibrated ensembles.

We also find that increasing temperature has significant effects on the relative populations of different types of hydrogen bonds. In the octapeptide, increasing temperature favours the formation of γ-turns over β-turns. In the (GVPGV)$_7$ peptide, increasing temperature favours local interactions over nonlocal interactions. These effects are analogous, and both may be rationalized on the basis of chain entropy: γ-turns effectively constrain fewer residues than β-
turns, and the same is true for local and nonlocal interactions. Thus, increasing temperature causes the peptide to preferentially form $\gamma$-turns and local hydrogen bonds.

Finally, our observation of the temperature-induced expansion of a monomer is consistent with a hypothesis offered in one of the earliest studies of elastin coacervation:

“The high solubility in cold water which is in fact observed may be due to a structure in which the non-polar parts of the molecule are free to form intramolecular associations, thus reducing their effective surface. A rise in temperature would tend to reduce this type of association, and by causing the molecule to unfold would favour the formation of the intermolecular bonds which ultimately lead to the separation of a second phase.”
S. M. Partridge et al (1955)\(^{40}\)

In the next chapter, we investigate this possibility with an STDR simulation of an aggregate of eight chains in which the unfolded polypeptides are able to form intermolecular interactions.
References


CHAPTER 8

Protein Non-Folding: An Elastin-like Aggregate Resembles a Polymer Melt
8.1 Summary

There is a long-standing controversy surrounding the structure of elastin aggregates. As background for the current study, we briefly outline the two prevailing models for elastin’s structure: the random network model, and the two phase model. In order to investigate the structural properties of an aggregate of ELPs, we performed a large-scale STDR simulation of an aggregate with eight (GVPGV)$_7$ chains, which is the same system that we studied as a monomer in Chapter 7. We find that extended states are dramatically more favourable for monomers within the aggregate than for monomers in water, which is consistent with neighbouring chains in the aggregate acting to “solvate” each other. The aggregated state of an ELP resembles a polymer melt, and the behaviour of the polypeptide chains within the aggregate is consistent with the predictions of the Flory theorem. However, the elastin-like aggregate is not a perfect solvent-excluding polymer melt because the polypeptide backbone retains significant hydrogen bonds to water. These results provide further support for the model of elastomeric aggregates as highly disordered and hydrated, as proposed in Chapter 4. Taken together, our results suggest that the two models for elastin structure that were previously thought to be contradictory may be unified as follows: it is only through significant chain-chain nonpolar interactions (as in the two phase model) that chain entropy nearly reaches its maximum (a key feature of the random network model).

8.2 A Seventy Year Old Controversy: What is the Structure of an Elastin Aggregate?

Despite its biological importance, a comprehensive understanding of the structural properties and self-aggregation of elastin has remained elusive since the first biophysical studies of this protein were performed more than seventy years ago. Two prevailing models have been proposed to describe the structure of elastin aggregates, and the molecular basis for their capacity for elastic recoil: (1) the random network model, and (2) the two phase model.

8.2.1 The Random Network Model

The first model proposed to described the aggregated state of elastin is the random network model, which was introduced by Hoeve and Flory in 1958. They performed thermoelasticity measurements on elastin samples, and found that entropy provides the
dominant contribution to the elastic restoring force (that is, \( f \approx f_s \) in equation 3.1). On the basis of these results, they concluded that elastin behaves as an ideal rubber-like elastomer, and that its structure resembles that of a random polymer network.\(^2_5\) In terms of elastin’s structure, this implies that the average end-to-end distance squared of the polypeptide chains is independent of temperature, and that the different configurations of the chain are similar in energy.\(^2\) The random network model of elastin’s structure has also been referred to as the single phase model because the polypeptide chains are thought to exist in an isotropic environment with water molecules distributed uniformly throughout the material.\(^6\) Independent support for the random network model came from x-ray diffraction experiments, which suggested that elastin adopts a disordered structure.\(^7\)

Hoeve and Flory’s model of elastin as a random network relies in part on assumptions that were necessary to carry out their thermoelasticity measurements. In order to separate the elastic restoring force into energetic and entropic terms, thermoelasticity experiments must be performed on samples that do not undergo changes in volume. This requirement represents a major experimental challenge in the case of elastin because elastin samples swell (absorb water) with decreasing temperature. In order to maintain swelling equilibrium, Hoeve and Flory performed their experiments in a 30:70 mixture of ethylene glycol and water.\(^2\) Other thermoelasticity measurements on elastin performed in triethylene glycol confirmed that the elastic recoil force, \( f \), is dominated by the entropic contribution, \( f_s \).\(^4\) However, this later study found that the energetic contribution, \( f_e \), is not negligible; \( f_e/f \) is approximately \( 26 \pm 9 \% \).\(^4\) An underlying assumption in both of these studies is that the addition of a diluent (ethylene glycol or triethylene glycol) will not significantly alter structural properties; that is, the polypeptide chains will have equally high chain entropy in aqueous solution as they do in the presence of alcohol diluents. The random network model rests in part on this assumption.

Despite the evidence for elastin’s structural disorder from thermoelasticity measurements and x-ray diffraction, the random network model is not universally accepted. In particular, Ellis and Packer questioned the assumption that ethylene glycol exerts negligible effects on elastin’s structure.\(^8\) There is evidence to suggest that the use of alcohols as diluents perturbs elastin’s
structure dramatically. Specifically, alkaline hydrolysis of elastin is enhanced in the presence of alcohols, and this effect is even more pronounced for long chain alcohols than short chain alcohols.\textsuperscript{9,10} Thus, alcohols may interact with the nonpolar side chains of elastin, resulting in the peptides becoming more solvent-exposed, and therefore susceptible to hydrolysis.\textsuperscript{10} Furthermore, these observations suggest that alcohols or alcohol/water mixtures may be better solvents of elastin and elastin-like peptides than water alone. For this reason, Gosline,\textsuperscript{11,12} Weis-Fogh and Anderson,\textsuperscript{13} and Ellis and Packer\textsuperscript{8} questioned the validity of the random network model for elastin, and instead proposed a two phase model.

8.2.2 The Two Phase Model

In the second model for elastin’s aggregated state, the two phase model, the polypeptide chains are thought to be arranged such that there is significant burial of hydrophobic groups,\textsuperscript{13,14} but the precise details of how this burial is achieved are not well understood. Various models have been proposed, including the “liquid drop model”,\textsuperscript{13} in which individual elastin chains are collapsed upon themselves in globules to minimize exposure of hydrophobic groups to water, and covalent cross-links to other chains are found on the surface of the droplets. The β-spiral model is another variant of the “two phase” model (refer to Chapter 7 for a more detailed discussion of this model).\textsuperscript{6,15,16} Furthermore, Gosline,\textsuperscript{11,12} Weis-Fogh,\textsuperscript{13} and Gray\textsuperscript{14} suggested that the primary contribution to elastic recoil may be due to the hydrophobic effect, and not entirely due to chain entropy. That is, elastic recoil may result in part from the increased exposure of nonpolar side chains to water in the stretched state, and the burial of these groups in the collapsed state.\textsuperscript{11,13,14}

The main arguments against the two phase model are based on the assumption that polypeptide chains in the hydrophobic aggregate are not sufficiently disordered or mobile to facilitate entropy-driven elastic recoil.\textsuperscript{3,17} “This view is summarized as follows:

“Since the compact chains in the two phase model are predicted to be much less mobile than the random coil chains in a rubber-like network, information on the mobility of elastin chains is useful for distinguishing between the two types of models.”

Lyerla and Torchia (1975)\textsuperscript{17}
There appears to be an essential contradiction between the random network and two phase models: they offer two distinct views of the structure of elastin, which lead to two different dominant contributions to the elastic restoring force (chain entropy vs. the hydrophobic effect).

Both the random network model and the two phase model of the structure of elastin in its aggregated state are inherently qualitative in nature. Recall from the discussion in Chapter 2 that conformational ensembles for disordered states of proteins cannot be obtained using experimental data alone. That is, the structure determination problem in the case of IDPs, including elastin, is underdetermined. Molecular simulations have already proven useful in providing quantitative structural information for other IDPs.\(^18\) In this study, our aim is to directly address the controversy surrounding the structural properties of elastin aggregates using molecular simulations as a tool to obtain, for the first time, a statistically-meaningful view of the structure of an ELP as an aggregate in atomistic detail.

### 8.3 The Flory Theorem: Ideal Chains in a Polymer Melt

In the introduction of Chapter 7, we introduced the hypothesis that polypeptide chains may reach a “theta-like” state in protein aggregates that are devoid of water.\(^19\)-\(^22\) To describe this amorphous, “theta-like” state of polymers, the term “polymer melt” is often used. Within a polymer melt, polymer chains are entangled with each other, forming extensive intermolecular interactions. Paul Flory was the first to reason that polymer chains may act as an ideal solvent of themselves. The consequence of this is the Flory Theorem: polymer chains within an amorphous polymer melt are predicted to have the same average dimensions as chains in an ideal (\(\theta\)) solvent.\(^23\),\(^24\) In an ideal solvent, chain-chain interactions are as probable as chain-solvent interactions. Within a polymer melt, intrachain interactions are as favourable as interchain interactions; this leads to chains in a polymer melt having dimensions matching those of chains in a \(\theta\) solvent. This is a surprising yet intuitive and elegant theory. It has been validated using neutron scattering for homopolymers.\(^25\),\(^26\) However, the extent to which it holds for polypeptide chains, which are heteropolymers, is at present unclear. In this chapter, we investigate the extent to which an elastin-like aggregate resembles a polymer melt.
8.4 Methods

The results presented in this chapter build upon the understanding of the effects of temperature on monomers established in Chapter 7. Here, we investigate the structural properties of an aggregate of eight chains. We performed an STDR simulation of an aggregate of eight \((\text{GVPGV})_7\) chains in explicit water, which required 306 temperatures (exponentially-spaced between 270 K and 749 K), and a cumulative simulation time of over 0.2 ms. An analysis of the temperature dependence of structural properties is ongoing work. All methodological details for this STDR simulation are provided in Chapter 6, and definitions of structural properties are provided in Appendix C (refer to Table C.1 for a summary).

8.5 Results and Discussion

8.5.1 ELP Chains in the Aggregate are Highly Entangled

The STDR simulation produces a large set of configurations for the ELP aggregate (about 100,000 configurations at each temperature). In Figure 8.1, we provide a random, but representative, selection of four of these configurations to illustrate that the polypeptide chains are entangled with each other, suggestive of a polymer melt-like state. Importantly, it is only because we utilize an enhanced sampling algorithm that we are able to observe significant chain entanglement. Results of a related study suggest that entanglement occurs over much longer time scales than those accessible to conventional MD simulations (Rauscher and Pomès, unpublished results). In particular, MD simulations of over 1 µs in length of an aggregate of 64 \((\text{GVPGV})_7\) peptides show that ELP chains do not significantly entangle with each other on the microsecond time scale (refer to Chapter 9, Figure 9.1).
Figure 8.1 Representative configurations of the aggregate
Four randomly selected configurations of the aggregate at 300 K are shown. Monomers are coloured individually and shown in “cartoon” representation for clarity.

8.5.2 “Disentangling” ELP Chains in the Aggregate

Next, we consider the properties of individual chains within the aggregate. In Figure 8.2, a single configuration of the aggregate is shown, along with the configuration of each of the eight chains. As expected for a disordered state, the chains exhibit a heterogeneous array of configurations: some chains are relatively compact, while others are almost fully extended. For example, the dark purple monomer in Figure 8.2 “threads” through the center of the aggregate and is nearly completely extended. Such an extended configuration is very improbable in aqueous solution because it leaves the majority of the peptide’s nonpolar side chains exposed to water. Extended states are accessible for chains within the aggregate because of the possibility of forming nonpolar interactions with other monomers.
Figure 8.2 The aggregate of (GVPGV)$_7$ chains “disentangled”
A randomly-selected configuration of the aggregate of eight (GVPGV)$_7$ chains with each of the eight monomers coloured differently and shown outside of the aggregate for clarity. The monomers exhibit a range of spatial extensions.

8.5.3 To What Extent Do ELP Chains Within the Aggregate Conform to the Flory Theorem?

In Chapter 7, we found that increasing temperature favours less compact states of the (GVPGV)$_7$ monomer in water, which is analogous to the effect of a denaturant. In particular, increasing temperature causes a gradual shift in the probability distribution of the radius of gyration, $R_g$, towards more extended states (Figure 7.7 B), and a trend towards increased average distances between residue pairs (Figure 7.8). We perform a similar analysis here to compare the average dimensions of chains within the aggregate to the average dimensions of the single chain in water (Figure 8.3). First, we find that the overall dimensions of chains within the aggregate at 300 K are dramatically expanded compared to the chain in water at 300 K, which is highly compact because water is a poor solvent. Remarkably, the probability distribution of $R_g$ for a
monomer within the aggregate is similar to that for a monomer in water at 749 K, which is the
highest temperature in our STDR temperature ladder.

![Figure 8.3 Distribution of the radius of gyration, $R_g$, in water and in the aggregate](image)

**Figure 8.3 Distribution of the radius of gyration, $R_g$, in water and in the aggregate**
The probability distribution of the radius of gyration is shown for a single (GVPGV)$_7$ chain in
the following three conditions: (1) in water at 300 K (purple), (2) in water at 749 K (red), and
(3) within the aggregate at 300 K (blue). Note that we are simply using the ensemble obtained
at 749 K for comparison, and do not mean to imply that this ensemble is physically “realistic”.

The similarity of the probability distributions of $R_g$ for monomers within the aggregate
and monomers at 749 K (Figure 8.3) suggests that the polypeptide chains within the aggregate
may behave similarly to chains in a $\theta$ solvent, at least with regard to their overall dimensions.
This leads directly to two questions: (1) to what extent do chains in an ELP aggregate resemble
ideal chains?; and (2) to what extent does the Flory Theorem hold for ELPs? We address both
of these questions with an analysis of the intrachain distance scaling. Recall from Chapter 7
(section 7.5.3 and Figure 7.8) that the average inter-residue distance for ideal chains scales as the
sequence separation to the exponent $\frac{1}{2}$, $\langle R_{ij} \rangle = R_o|i-j|^{1/2}$. This scaling corresponds to the
black curve in Figure 8.4. The inter-residue distance scaling of chains within the aggregate at
300 K is similar to that of a monomer in water at 749 K. However, the curve does not coincide
with the expected scaling for an ideal chain. Thus, while monomers within the aggregate of
eight chains exhibit inter-residue distance scaling that is closer to ideality that monomers in
water, they do not behave as ideal chains: they are still in the “poor” solvent regime. Next, we investigate why monomers in the aggregate do not attain the intrachain distance scaling typical of ideal chains.

\[ \langle R_{ij} \rangle = R_0 |i - j|^{1/2} \]

Figure 8.4 The effect of temperature and aggregation on average intrachain distance scaling
For two residues \( i \) and \( j \), the ensemble average of the separation distance, \( \langle R_{ij} \rangle \), is given by equation 7.1.\(^{27} \) \( \langle R_{ij} \rangle \) is the average separation distance of all heavy atoms pairs in residues \( i \) and \( j \), and \( n_{ij} \) is the number of unique pairwise distances. The indices \( k \) and \( l \) are for the heavy atoms in residue \( i \) and \( j \), respectively. \( |i - j| \) is the number of residues separating residues \( i \) and \( j \). The black curve represents the scaling expected for an “ideal” solvent (\( \langle R_{ij} \rangle = R_0 |i - j|^{1/2} \) where \( R_0 = 4.7 \text{ Å} \)).\(^{27} \) Here, we compare the average inter-residue distances for three conditions to the ideal chain: (1) a monomer in water at 300 K (purple), (2) a monomer within the aggregate at 300 K (red), and (3) a monomer in water at 749 K (yellow).

8.5.4 Backbone Hydration Precludes the Formation of An Ideal Polymer Melt

The inter-residue distance scaling for chains within the aggregate does not match the scaling expected for ideal chains (Figure 8.4). Thus, the environment for polypeptide chains does not mimic that of a \( \theta \) solvent, and as such the aggregate does not behave as “perfect” polymer melt. The reason for this deviation from ideality is quite simple: water molecules are present throughout the aggregate (Figure 8.5). It is not possible to expel all the water molecules from the aggregate because the backbone is disordered, and unable to form a tightly packed hydrophobic core; this property of elastomeric aggregates is actually one of the key postulates of our unified model for protein self-aggregation presented in Chapter 4. Furthermore, we have
shown in Chapter 7 that water molecules are a poor solvent of ELPs (see for example Figures 7.7 and 7.8). For this reason, the water molecules present within the aggregate lead to a deviation from ideal chain behaviour.

Figure 8.5 An aggregate of elastin-like peptides retains significant backbone hydration
A single configuration of the elastin-like aggregate is shown with the closest 400 water molecules. Chains are shown in “cartoon” representation and coloured individually, while water molecules are shown in CPK representation and coloured blue. Hydrogen bonds are black dotted lines.

8.5.5 Comparison of the Structural Properties of ELP Chains in Water and in the Aggregate

In Table 8.1, we provide a summary of the chain-chain and chain-water interactions of the monomer in water at 300 K, and the monomer within the aggregate at 300 K. Chain-chain interactions include both hydrogen bonds and nonpolar interactions, each of which is separated into local, nonlocal, intramolecular, and intermolecular contributions. Chain-water interactions include both hydrogen bonds between the polypeptide backbone and water molecules, as well as interactions between the nonpolar groups and water molecules.
Table 8.1  Comparison of chain-chain and chain-water interactions for the monomer in water and monomers in the aggregate

For the STDR ensemble of configurations of the monomer in water, and the average monomer within the aggregate, we provide the chain-chain interactions (both peptide-peptide hydrogen bonds and nonpolar contacts), and chain-water interactions (hydrogen bonds between the backbone and water, the number of bound waters and bridging waters, as well as the number of water molecules in the nonpolar hydration shell). While hydrogen bonding is affected only slightly by aggregation, there is a dramatic difference in nonpolar contacts and the number of water molecules in the nonpolar hydration shell. Importantly, the amount of local interactions change very little upon aggregation, and it is primarily the nonlocal interactions that are replaced by intermolecular interactions. All errors are obtained by the “blocking” method.

With regard to chain-chain interactions, we find that monomers within the aggregate have a slightly enhanced probability of peptide-peptide interactions compared to the monomer in water. While the probability of peptide-peptide hydrogen bonds increases only slightly, there is a dramatic increase in the probability of nonpolar peptide-peptide interactions in the aggregate. The loss of nonlocal nonpolar interactions is more than compensated for by the dramatic increase in intermolecular nonpolar contacts. These effects are further illustrated by the hydrogen bonding and nonpolar contact maps shown in Figure 8.6. Taken together, these results suggest that the contribution of nonpolar interactions in stabilizing the aggregated state is
significantly more important than peptide-peptide hydrogen bonding. Interestingly, the probability of local interactions (both hydrogen bonding and nonpolar interactions) is essentially the same for monomers in the aggregate compared to monomers in water. This is because the probabilities of various hydrogen-bonded turns (such as the VPGV turn, Figure 7.2) are not significantly affected by aggregation. By comparing the off-diagonal elements in the hydrogen bonding contact maps of the monomer in water and the monomer in the aggregate at 300 K (Figure 8.6), it is clear that the populations of local hydrogen bonds remain nearly unchanged upon aggregation.

With regard to chain-water interactions, there is a dramatic loss of water molecules solvating nonpolar groups in the aggregate compared to the monomer: the number of water molecules per residue in the nonpolar hydration shell of the aggregate is $127 \pm 1$ compared to $247 \pm 1$ for the monomer. Aggregation has a much less significant effect on hydrogen bonding interactions between the backbone and water molecules: the number of hydrogen bonds to water per residue is $1.37 \pm 0.02$ and $1.13 \pm 0.01$ for the monomer and aggregate, respectively. This observation is consistent with the model for an elastomeric aggregate that we proposed in Chapter 4: backbone hydration is an essential ingredient in conformational disorder. The chains are unable to form $\beta$-sheet structures with extensive backbone-backbone hydrogen bonding. Nevertheless, it is energetically favourable for the backbone groups to form hydrogen bonds. Since all backbone hydrogen-bonding groups cannot be satisfied through self-interactions, hydrogen bonds to water must also occur. As a result of this backbone hydration, the elastin-like aggregate is unable to form a water-excluding hydrophobic core. Based on an analysis of the coordination numbers of water molecules in the hydration shell (refer to Appendix C), we find that a small number of water molecules are trapped within the aggregate with no interacting water molecules. These waters are found within the largely nonpolar environment of the aggregate in order to satisfy the hydrogen-bonding requirements of backbone groups.
Figure 8.6 Hydrogen bonding and nonpolar contact maps
Hydrogen bonding contact maps (left) and nonpolar contact maps (right) are shown for (GVPGV)$_7$ chains under three conditions: (1) water at 300 K, (2) within aggregate at 300 K, and (3) water at 749 K. The brightness of each square indicates the probability of contact formation between a pair of residues. For the monomer within the aggregate (middle row), intermolecular contacts are shown below the diagonal.
8.6 Conclusion: Towards a Unified Model for the Structure of an Elastin-like Aggregate

We now relate the structural properties of the elastin-like aggregate obtained in our
STDR simulation to the previous models for elastin’s structure and function.

1. Relating Our Model for the ELP Aggregate to the Random Network Model

In agreement with Hoeve and Flory’s prediction based on thermoelasticity
measurements, we find that the polypeptide chains within the aggregate have high (nearly
maximal) chain entropy. This is only possible when they are self-aggregated, which mimics θ
solvent conditions. Based on the inter-residue distance scaling, chains within the aggregate
resemble ideal chains (but, importantly, are not exactly like ideal chains, since they are still in the
“poor” solvent regime). These results are the first atomistic observation of Flory’s theorem for
polypeptide chains. In the case of ELPs, deviation from ideal chain behaviour occurs because of
significant water penetration into the aggregate. This may have implications for elastin’s
mechanism of elastic recoil: perfectly ideal chains have maximal chain entropy, and therefore
would contribute to the entropic term in the elastic restoring force \( f_s \) in equation 3.1.
Deviations from ideal behaviour may contribute to the fact that the energetic term in the
restoring force \( f_e \) in equation 3.1) is nonzero.

2. Relating Our Model for the ELP Aggregate to the “Two Phase” Model

Our observation of the important role of the hydrophobic effect in elastin self-
aggregation is consistent with a hypothesis put forth by Chalmers, Gosline, and Lillie.\(^{12}\) They
performed an analysis of the sequence of various species of elastin, and proposed that
hydrophobicity has an evolutionary advantage that is not correlated with mechanical properties,
and therefore might be related to self-assembly.\(^{12}\) We find that the most dramatic differences
between the structural properties of monomers and aggregates are the increase in peptide-peptide
nonpolar contacts and the related decrease in the number of water molecules in the nonpolar
hydration shell upon aggregation (Table 8.1). Our results suggest that the hydrophobic effect
plays a dominant role in promoting the aggregation of ELPs.
3. Relating Our Model for the ELP Aggregate to the \( \beta \)-Spiral Model

Evidence from both simulation\(^6,28-30\) and experiment\(^2,3,31,32\) strongly suggests that elastin and elastin-like peptides are intrinsically disordered, and do not adopt a unique structure, such as the \( \beta \)-spiral. However, it is worth noting that elements of the \( \beta \)-spiral model are consistent with the results reported here. Specifically, we find that the VPGV \( \beta \)-turns (Figure 7.2) are highly populated (they are the second most populated turn; the most populated are the GVGV \( \beta \)-turns). No single configuration in our configurational ensemble of (GVPGV)\(^7\) as either a monomer or aggregate resembles a \( \beta \)-spiral. However, on average, the contact map of monomers within the aggregate does resemble that of a \( \beta \)-spiral-type structure, albeit with low populations for each of the turns (Figure 8.6). This is not to say that the \( \beta \)-spiral model for elastin should be resurrected. We merely point out that the repetitive sequence of this peptide does lead to a repetitive and regular structure, but this is only true when viewed as an average over all configurations in the ensemble.

In conclusion, our results are consistent with elements of both the random network model and the two phase model. We propose a simple resolution to the random network vs. two phase model controversy: it is only when ELPs are in an aggregated state with extensive intermolecular nonpolar interactions that they are able to approach the “theta”-like state in which their chain entropy is nearly maximal. In this way, the random network model and the two phase model may be unified.
References


CHAPTER 9

Significant Contributions and Future Directions
9.1 Summary of the Significance and Contributions of the Thesis

The elasticity of skin, lungs, blood vessels, and uterine tissue is imparted by the protein elastin. Human life is entirely dependent on the elasticity and resilience of this remarkable protein. Increasing temperature causes elastin to undergo phase separation in a process known as coacervation. Because of its exceptional elastic mechanical properties and capacity for self-organization, elastin is well-suited for incorporation in artificial skin to treat burn victims, and vascular grafts for heart patients. Despite its biological and biomedical relevance, a comprehensive understanding of the structural properties and self-aggregation of elastin has remained elusive since the very first biophysical studies of this protein were performed more than seventy years ago. In order to directly probe the molecular basis for elastin self-aggregation, we performed large-scale simulations of elastin-like peptides. These simulations involved the use of thousands of computers simultaneously, and the use of a novel enhanced sampling algorithm.

In contrast to elastin, amyloid fibrils are protein deposits that accumulate in a variety of organs, and are associated with several tissue-degenerative pathologies, including Alzheimer's disease, Parkinson's disease, and type II diabetes. To investigate the molecular determinants underlying these two classes of protein self-organization, we performed a systematic simulation study of monomeric systems, as well as sequence analysis of elastomeric and amyloidogenic proteins. We find that a high combined content of proline and glycine separates the sequences of all known elastomeric proteins (including elastin, spider silk, insect resilin, wheat gluten, and scallop abductin) from the sequences of amyloid-forming proteins. Thus, a high combined PG content seems to be a necessary condition for the structural disorder necessary for entropy-driven elasticity in proteins. We find that conventional MD simulations are sufficient to separate elastin-like and amyloid-like peptides on the basis of their hydrogen-bonding propensities as monomers and aggregates. However, we show that single MD time trajectories of highly nonpolar disordered peptides result in collapse to what is essentially a single conformation. Thus, conventional MD simulations are not well-suited to describe the heterogeneous conformational ensembles of disordered proteins.
This observation illustrates one of the major challenges in biomolecular MD simulation. Biomolecular systems, such as proteins, are challenging to study because of their “rugged” energy landscapes. That is, important states are separated by high energy barriers. For this reason, the timescales of many important biomolecular processes, such as protein folding, are inaccessible to conventional MD simulations. To address this issue, we developed, tested, and implemented an algorithm (Simulated Tempering Distributed Replica Sampling, STDR) that uses temperature to allow the system to “jump” over energy barriers. STDR is a combination of two previously existing generalized-ensemble methods: simulated tempering (ST), and distributed replica sampling (DR). We have carefully delineated the circumstances under which it is more practical and efficient to use STDR than ST. In particular, we find that STDR is ideally suited to study complex systems for which it is not possible to compute accurate Helmholtz free energies. We have also shown that ST and STDR out-perform replica exchange based methods in two respects: both ST and STDR exhibit faster “diffusion” of replicas in temperature space, and, correspondingly, faster statistical convergence of structural properties. STDR represents a significant practical and methodological advance in enhanced sampling: it has opened the door to statistically-meaningful atomistic descriptions of disordered states of proteins.

In order to investigate the molecular basis for elastin’s temperature-induced self-aggregation, we performed extensive STDR simulations of a both a monomer and an aggregate of eight chains of the elastin-like peptide (GVPGV)$_7$. We find that the polypeptide backbone of both monomeric and aggregated chains remains highly hydrated. The configurational ensemble of the elastin-like aggregate resembles that of an “imperfect” polymer melt. Chains within the aggregate are entangled with each other, but are unable to completely expel water, and therefore do not behave as perfectly ideal chains. Our results are not consistent with the current model of elastin self-aggregation, which involves a conformational transition of a monomer towards a more ordered state. Instead, a new model emerges: both the hydrophobic effect and the enhanced chain entropy afforded by the interactions with other peptides within the aggregate promote self-aggregation. Finally, this view unifies the random network model and the two phase model for the structure of elastin aggregates as follows. It is only because of the significant nonpolar interactions between chains in the aggregate (as proposed in the two phase model) that
they are able to achieve a nearly “theta-like” state that affords the high chain entropy required in
the random network model. Because this work represents the first atomistic simulation, to our
knowledge, demonstrating that polypeptide chains can form entangled polymer melt-like states,
it contributes to an improved understanding of both elastin coacervation, and the more general
phenomenon of protein aggregation.

9.2 Future Avenues of Research

The body of work described in this thesis forms the foundation for future avenues of
research related to both elastin, as well as other disordered states of proteins. Disordered and
unfolded states of proteins are, at present, poorly understood relative to the wealth of
information available for folded states of proteins. The methods developed in the course of this
study may prove useful in advancing our general understanding of “protein non-folding.” Here,
I describe what are, in my opinion, important directions and key unanswered questions in the
field. Some of these research directions are currently being pursued, while others are at present
hypotheses and questions.

To what extent are the structural properties of an aggregate of eight elastin-like chains
representative of larger aggregates?

The study described in Chapter 8 is the most computationally intensive atomistic
simulation of a protein aggregate reported to date. The system size (8 chains) was chosen
because it represents the largest system for which we might hope to achieve statistical
convergence using a generalized-ensemble algorithm and current computational resources.
However, it is not clear exactly how the structural properties of ELP chains in an aggregate of
eight peptides may differ from the behaviour in even larger aggregates. To address this question,
I have been concurrently performing simulations of an aggregate of 64 (GVPGV)7 chains (Figure
9.1) in explicit water (for a total system size of nearly one million atoms) at selected temperatures
for more than 1 µs. A detailed analysis of the balance of chain-chain and chain-water
interactions in this aggregate in relation to the behaviour of monomeric chains and chains in the
aggregate of eight peptides will shed light on the behaviour of chains in the interior of the
aggregate.
Figure 9.1 A large aggregate of sixty four ELP chains
Here, we show a single configuration obtained from a one microsecond simulation of an aggregate of sixty four \((\text{GVPGV})_7\) chains in explicit water at 300 K. The comparison of the structural properties of chains within this large aggregate and those of chains in the aggregate of eight peptides described in Chapter 8 will provide a deeper understanding of the effect of aggregate size on the balance of chain-chain and chain-water interactions.

How does the observation of a polymer melt-like state relate to the structure of elastin-like aggregates at much larger length scales?

One of the major limitations of the studies presented in this thesis is that they have exclusively focused on sequences modeled after the hydrophobic domains of elastin. It is known that covalent crosslinking involving lysine residues in the cross-linking domains is crucial to the mechanical properties of elastin and elastin-like peptides. Furthermore, the sequences of tropoelastin from various species is hundreds of amino acids in length, but the peptides in this study were only 35 or 36 residues in length. Current computational capabilities are inadequate to conduct statistically-meaningful simulations of elastin-like peptides with both hydrophobic domains and cross-linking domains in atomistic detail. We therefore turn to coarse-graining to probe elastin's properties at larger length scales and longer time scales than those currently accessible to atomistic simulations. Towards this goal, the conformational ensembles from our atomistic simulations of elastin-like and amyloid-like sequences have been used to incorporate
dihedral potentials into the MARTINI coarse-grained force field (in collaboration with Dr. Peter Tieleman and Dr. Mikyung Seo). Using this modified MARTINI force field, it will be possible to perform simulations of large aggregates of elastin. In Chapter 8, we show that chains within the aggregate have nonspecific nonlocal and intermolecular interactions, and relatively specific local interactions. These observations suggest that the structural propensities of ELP chains in the aggregated state readily lend themselves to representation using coarse grained models.

What is the θ solvent for elastin-like peptides?

In the introduction to Chapter 8, we noted that Hoeve and Flory’s original paper in which they proposed that the structure of elastin resembles that of a random polymer network utilized ethylene glycol as a diluent to prevent the elastin samples from undergoing volume changes.¹ Ellis and Packer previously suggested that ethylene glycol may significantly perturb the structure of elastin.² These concerns were echoed by Gosline,³,⁴ whose work also supported the role of hydrophobicity in the mechanism of elastic recoil of elastin. These studies, in combination with the present work, lead directly to the question: what is the θ solvent of ELPs such as (GVPGV)₇? In such a solvent, the chains are expected to take on similar dimensions and structural properties as they do within the “melt”. Because the present STDR simulation of an elastin aggregate took over a year and a half, it is not feasible at the present time to repeat this study for other ELP sequences. If an “ideal” solvent can be found, it represents somewhat of a short cut to systematically characterize the structural properties of ELPs with a variety of sequences. Based on the work of Hoeve and Flory¹ and Andrady and Mark,⁵ ethylene glycol and triethylene glycol are possible candidates for the θ solvent of ELPs. This avenue of research – the search for the “ideal” solvent – is currently being pursued by Zhuyi Xue.

Why do the phase diagrams of ELPs have the opposite behaviour compared to other proteins?

In the introduction to Chapter 7, we posed the question: “Why do elastin and ELPs self-aggregate and undergo phase separation with increasing temperature in contrast to the phase separation behaviour typical of most other proteins?” The body of work in this thesis is not sufficient to completely address this question, but we may speculate on why this is the case. If the phase diagram of ELPs is an asymmetric closed loop similar to the phase diagram in the
general theory for protein aggregation developed by Fields et al., then it may be that the two characterizations of ELP phase diagrams obtained thus far\textsuperscript{6,7} are actually observations of the lower coexistence curve (binodal) of the phase diagram. Given the relevance of a deeper understanding of elastin self-assembly, further characterization of ELP phase diagrams seems to be a particularly important, but relatively unexplored, research direction.

What are the relative contributions of chain entropy and the hydrophobic effect to the elastic restoring force of ELPs and elastin?

While the results of Chapter 8 lead to a new model unifying the random network model and the two phase model, we still do not have a quantitative prediction for the relative contributions of chain entropy, and the burial of nonpolar groups to the elastic restoring force. This question is certainly one of the most fundamental questions in the elastin field, and has remained controversial for over sixty years. One possible, but likely challenging, direction may be to perform thermoelasticity measurements without the use of diluents.

What does the unfolded state look like? How big is it?

One research direction that arose as a direct result of the methodology developed in this thesis is a study of the unfolded state of a globular protein. We performed large scale STDR simulations to investigate the structural properties of the unfolded state of the N-terminal SH3 domain of Drosophila drk (Figure 9.2). This globular domain is ideal as a model system because it exists in equilibrium between folded and unfolded states under non-denaturing conditions, and is the best characterized disordered protein state ever studied experimentally (in the laboratory of Dr. Julie Forman-Kay).\textsuperscript{8,9} In addition to simulations of the wild type protein, we have also performed an STDR simulation of a mutant, T22G, which is known to fully populate the folded state. Grace Li has also performed large scale STDR simulations of the WT protein in the presence of 2M GdnCl. With a total simulation time of 1.4 milliseconds, these simulations represent one of the most extensive computational descriptions of the unfolded state and denatured state of any protein domain reported to date. Our simulations of this system were made possible by the award of a national resource allocation on the SciNet supercomputer, which provided an unprecedented opportunity for large-scale sampling. One key finding of this
study is a significant disagreement between the overall size (as measured by the radius of gyration) of the configurational ensemble generated by STDR and the expected size based on pulse field gradient NMR measurements, and fluorescence correlation spectroscopy measurements (as measured by the hydrodynamic radius). These results suggest that the molecular mechanics force fields we use may lead to an overestimate of the stability of compact states. One hypothesis for this effect involving the energetics of hydrogen bonding is briefly discussed in Appendix C. We are currently investigating in detail the source of this discrepancy between simulation and experiment. Importantly, this work will lead to suggestions for force field improvement towards the goal of obtaining more accurate descriptions of disordered and unfolded states.

**Figure 9.2 Configurational ensembles of the folded state and unfolded state of drkN SH3**

On the left is an ensemble of configurations obtained from a 100 ns simulation starting with one of the configurations in the NMR structure of the wild type drkN SH3. On the right is a random selection of configurations from the STDR simulation at 300 K. Because the two ensembles are shown at the same scale, the relative size is depicted. In particular, we find that the ensemble average of the radius of gyration of the STDR ensemble is only 10 % larger than that of the folded state.
References


APPENDIX A

Analysis of Proline and Glycine Spacing in Human Elastin

The analysis presented in this appendix was originally published online as supplementary material for our article in Structure.

Reference:

Contributions:
S.R. performed this analysis and made the figures.
A.1 Proline and Glycine Spacing in Human Elastin

An analysis of the spacing between consecutive proline and glycine residues in human elastin reveals that the spacing of these residues is non-random (Figure A.1 and A.2 for proline and glycine, respectively).

**Figure A.1 Inter-proline spacing in human elastin**

The sequence of human elastin is not random with regard to the positioning of prolines and glycines. (A) An analysis of the number of residues between consecutive prolines reveals clear patterns which result from the repetitive nature of elastin’s sequence. Larger spacings between prolines occur due to cross-linking domains, many of which contain long proline-free stretches. (B) Converting the inter-proline spacing data to a probability distribution indicates that it is mostly likely to find 2, 3, 4 or 5 residues between consecutive prolines.
Figure A.2 Inter-glycine spacing in human elastin

(A) Analysis of the number of residues between consecutive glycines demonstrates the presence of repetition in the sequence of human elastin. (B) Inter-glycine spacings of 0, 1 or 2 residues are most preferable. The closer spacing between glycines compared to prolines is consistent with our earlier proposition that proline is the primary determinant of elastomeric properties, while glycine is secondary. Prolines are able to ‘exert’ their effect over longer distances along the sequence, while the distribution of inter-G spacing is sharply peaked at low spacing. The clearly non-random inter-P and inter-G spacing present in elastin’s sequence may play a role in the maintenance of conformational disorder and hydration necessary in the avoidance of amyloid formation and the onset of elastomeric properties.
The most common spacing between proline residues is five residues, while that of glycine is one residue. In general, glycine residues are separated by fewer residues, consistent with the fact that the glycine composition is higher than that of the proline composition. Based on this analysis, we hypothesize that the spacing between proline and glycine residues may play a role in determining whether peptides self-assemble into elastomeric or amyloid-like aggregates. That is, the sequence determinants of protein self-organization are more complex than a simple threshold in proline and glycine content (although obeying this threshold appears to be a necessary condition for rubber-like elastomeric properties). A recent study performed in the Keeley lab supports the importance of close inter-proline spacing in modulating the self-assembly of elastin-like peptides. Furthermore, Muiznieks and Keeley found that proline-poor elastin-like peptides formed smaller droplets upon self-association, and their aggregates contained significantly more β-sheet (though they did not bind the dye Thioflavin-T, which is a characteristic of amyloid-like fibrils).
References

APPENDIX B

Supplementary Information for Chapter 5: Simulated Tempering Distributed Replica Sampling, Virtual Replica Exchange, and Other Generalized-Ensemble Methods for Conformational Sampling

The analysis presented in this appendix was originally published online as supplementary material for our article in the Journal of Chemical Theory and Computation.

Reference:

Contributions:
S.R. wrote and performed all analysis in this appendix.
B.1 Supporting Tables and Figures for Chapter 5

In this appendix, we provide detailed information related to the material presented in Chapter 5.

Temperature list for the octapeptide (33 temperatures):


Temperature list for the 35-residue peptide (70 temperatures):


**Table B.1 List of temperatures**

A list of the 33 exponentially spaced temperatures for generalized-ensemble simulations of GVGVPVG, and the 70 temperatures used in the STDR simulation of (GVPGV)\textsuperscript{2}. 
Figure B.1 Deviation from sampling homogeneity

Deviation from homogeneity is computed:

\[
\% \text{ deviation from homogeneity} = \frac{|N_m - \langle N_m \rangle|}{\langle N_m \rangle} \times 100\% ,
\]

where \( N_m \) is the number of samples at temperature \( T_m \), and \( \langle N_m \rangle \) is the average number of samples per temperature. RE is not shown since it samples temperatures uniformly by definition. The average deviations from sampling homogeneity reported in Table 5.2 are computed as averages over temperature using the data in this figure. STb was created to have inhomogeneous sampling, and it clearly achieves this. SREM also exhibits significant deviation from homogeneity.
Figure B.2  Structural convergence at multiple temperatures, bar graphs
The average structural convergence times for one and two standard deviations are shown for all seven methods in yellow and purple, respectively. These times are provided for temperatures 280 K, 288 K, 296 K, 305 K, 314 K, 323 K and 332 K. Below each plot, a ranking of the methods is provided, from the fastest to reach convergence to the slowest.
Figure B.3  Structural convergence at multiple temperatures, 2D plot
Average structural convergence times, $<t_{sc}>$, obtained using the lowest seven temperatures are shown. The $<t_{sc}>$ to reach two standard deviations is plotted against the $<t_{sc}>$ to reach one standard deviation for each method in circles. Squares represent the average over the 7 temperatures and error bars represent the standard error of the $<t_{sc}>$ for the seven temperatures. In (A) the methods are colored red for ST-based and blue for RE-based, while in (B) they are colored individually.
B.2 An Example Calculation of the Distributed Replica Potential Energy (DRPE)

There are five possible temperatures in this case ($\lambda_i$) which are exponentially spaced. There are five corresponding values for $\lambda_{i,\text{unit}}$ which are uniformly spaced:

<table>
<thead>
<tr>
<th>$\lambda_{i,\text{unit}}$ (uniform spacing)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_i$ (temperatures, K)</td>
<td>280</td>
<td>288</td>
<td>296</td>
<td>305</td>
<td>314</td>
</tr>
</tbody>
</table>

There are five replicas of the system, and their temperatures currently are: 305 K, 296 K, 288 K, 280 K, and 288 K. The move which will be attempted will be for replica 1 to change its temperature from 305 K to 296 K. First we compute the DRPE for the current state:

**STEP 1**: Sort the current temperatures into ascending order.

305 → 296 → 288 → 280 → 288

**STEP 2**: Transform this list into the corresponding uniformly spaced values.

1 → 2 → 2 → 3 → 4

**STEP 3**: Compute the DRPE

$$DRPE = c_1 \sum_{m=1}^{\omega} \sum_{n=1}^{\omega} \left[ (\lambda_{m,\text{unit}} - \lambda_{n,\text{unit}}) - \omega(m-n) \right]^2 + c_2 \left[ \sum_{m=1}^{\omega} \lambda_{m,\text{unit}} - \omega \sum_{m=1}^{\omega} m \right]^2$$

(B.2)

In this case, $\omega=1$ (the number of temperatures equals the number of replicas).
We use $c_1=0.02$ and $c_2=0.05$ (these are different than the values in the paper), and we find the DRPE is:

\[
DRPE = c_1 \left\{ \left[ (2-2)-(3-3) \right]^2 + \left[ (2-3)-(3-4) \right]^2 + \left[ (2-4)-(3-5) \right]^2 + \left[ (3-1)-(4-1) \right]^2 + \left[ (3-2)-(4-2) \right]^2 + \left[ (3-2)-(4-3) \right]^2 + \left[ (3-3)-(4-4) \right]^2 + \left[ (3-4)-(5-1) \right]^2 + \left[ (4-2)-(5-2) \right]^2 + \left[ (4-3)-(5-3) \right]^2 + \left[ (4-3)-(5-4) \right]^2 + \left[ (4-4)-(5-5) \right]^2 \right\} + c_2 \left[ 1+2+2+3+4 \right] \left( 1+2+3+4+5 \right) \]
STEP 1: Sort the temperatures into ascending order.

STEP 2: Transform this list into the corresponding uniformly spaced values.

STEP 3: Compute the DRPE

$$DRPE = c_1 \sum_{m=1}^{M} \sum_{n=1}^{M} \left[ (\lambda_{m,\text{unit}} - \lambda_{n,\text{unit}}) - \omega (m-n) \right]^2 + c_2 \left[ \sum_{m=1}^{M} \lambda_{m,\text{unit}} - \omega \sum_{m=1}^{M} m \right]^2$$
The difference in the DRPEs is:

\[ \text{DRPE(\text{proposed move})} - \text{DRPE(\text{current})} = 1.36 - 0.69 = 0.67 \]

Since the acceptance ratio in STDR has the form:

\[
p(T_i \rightarrow T_j) = \min \left\{ 1, e^{-\left(\beta_j - \beta_i\right) + \left(\mathbf{a}_j - \mathbf{a}_i\right) - (\text{DRPE}_j - \text{DRPE}_i)} \right\}
\]

(B.3)

the addition of the DRPE is the same as multiplying the acceptance ratio of ST by a factor of 
\[ e^{-0.67}, \] decreasing the likelihood of the move by approximately 51 %, which is the same as applying a penalty of 0.67 kcal/mol. This move is unfavourable because the replicas are less
spread out. The constants $c_1$ and $c_2$ control the degree of energetic penalty imposed for these unfavourable moves.

**B.3 Computing Error in Exchange Probabilities for ST and SREM**

The PEDFs of the octapeptide are nearly perfect Gaussian distributions, as expected due to the large number of degrees of freedom of the system, and the central limit theorem.\(^1\) Assuming that the PEDFs are Gaussian is in general a valid assumption for biomolecular systems.\(^2\) As an estimate of the error in the PEDFs, we considered the average deviation of the average energy of each PEDF, $<E_n>$, from the average energy of a reference PEDF, $<E_n>_{\text{reference}}$, as follows:

$$
\sigma_{\text{PEDFs}} \approx \frac{1}{N_{\text{temps}}} \sum_{n=1}^{N_{\text{temps}}} \left|<E_n>-<E_n>_{\text{reference}}\right|.
$$

(B.4)

where $N_{\text{temps}}$ is the number of temperatures. For the reference PEDFs, we used potential energy distribution functions calculated based on all of the data from the RE simulation (a total of 4.75 $\mu$s for all temperatures). We computed the average error in the differences of weight factors in an analogous way, also using the RE simulation as reference data:

$$
\sigma_{\text{weight factors}} \approx \frac{1}{N_{\text{temps}}-1} \sum_{n=1}^{N_{\text{temps}}-1} \left|(a_{n+1}-a_n)-(a_{n+1}-a_n)_{\text{reference}}\right|.
$$

(B.5)

The selection of the RE simulation as a reference was made because it was the only generalized-ensemble method that we tested that did not make use of any initial simulation.

In order to make a fair comparison between the errors in the weight factors used in ST and the PEDFs used in SREM, it is important to consider the error in not only the potential energy distribution functions and dimensionless Helmholtz free energies, but also the error in the resulting exchange probabilities. The error in the exchange probability of SREM (equation 5.9) was computed as follows:
\[
\sigma_{P_y} = \sqrt{\left(\frac{\partial P_{ij}}{\partial E_{ij,\text{PEDF}}} \right)^2 \sigma_{E_{ij,\text{PEDF}}}^2} = e^{-(\beta_j - \beta_i)(E_i - E_{ij,\text{PEDF}})} (\beta_j - \beta_i) \sigma_{E_{ij,\text{PEDF}}}. \tag{B.6}
\]

We estimate this error by using the estimate for the error in the PEDFs obtained in equation B.4, and the average acceptance ratio and average difference in inverse temperatures:

\[
\sigma_{P_y,\text{estimate}} \approx \left( e^{-(\beta_j - \beta_i)(E_i - E_{ij,\text{PEDF}})} \right) (\beta_j - \beta_i) \sigma_{\text{PEDFs}}. \tag{B.7}
\]

Similarly, the error in the exchange probability for ST (given by equation 5.6) is:

\[
\sigma_{P_y} = \sqrt{\left(\frac{\partial P_{ij}}{\partial (a_j - a_i)} \right)^2 \sigma_{(a_j - a_i)}^2} = e^{(\beta_j - \beta_i)(a_j - a_i)} (\beta_j - \beta_i) \sigma_{(a_j - a_i)}. \tag{B.8}
\]

and this error is estimated using the average error in the weight factor differences from equation B.5 and the average acceptance ratio:

\[
\sigma_{P_y,\text{estimate}} \approx \left( e^{(\beta_j - \beta_i)(a_j - a_i)} \right) \sigma_{\text{weight factors}}. \tag{B.9}
\]
References


APPENDIX C

Structural Properties of Disordered States of Proteins: Definitions and Criteria
C.1 Structural Properties of Disordered States of Proteins

In this appendix, we define and describe the structural properties used to characterize the conformational ensembles in Chapters 6 to 8. Table C.1 serves as a reference guide to these properties. In the following sections, we provide motivation for our definitions.

Table C.1 Criteria for chain-chain, chain-water, and water-water interactions

<table>
<thead>
<tr>
<th>Chain-chain interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptide-peptide hydrogen bond</td>
</tr>
<tr>
<td>( C_\alpha - C_\alpha ) contact between residues</td>
</tr>
<tr>
<td>Heavy atom contact between residues</td>
</tr>
<tr>
<td>Nonpolar contact between residues</td>
</tr>
<tr>
<td>Local interaction</td>
</tr>
<tr>
<td>Non-local interaction</td>
</tr>
<tr>
<td>“Forward” turns</td>
</tr>
<tr>
<td>( \gamma )-turn</td>
</tr>
<tr>
<td>( \beta )-turn</td>
</tr>
<tr>
<td>( \alpha )-turn</td>
</tr>
<tr>
<td>5-turn</td>
</tr>
<tr>
<td>6-turn</td>
</tr>
<tr>
<td>7-turn</td>
</tr>
</tbody>
</table>

| “Reverse” turns                               | hydrogen bond between N-H of residue \( i \) and C=O of residue \( i+2 \) |
| \( \gamma \)-turn                              | hydrogen bond between N-H of residue \( i \) and C=O of residue \( i+3 \) |
| \( \beta \)-turn                               | hydrogen bond between N-H of residue \( i \) and C=O of residue \( i+4 \) |
| \( \alpha \)-turn                              | hydrogen bond between N-H of residue \( i \) and C=O of residue \( i+5 \) |
| 5-turn                                       | hydrogen bond between N-H of residue \( i \) and C=O of residue \( i+6 \) |
| 6-turn                                       | hydrogen bond between N-H of residue \( i \) and C=O of residue \( i+7 \) |

<table>
<thead>
<tr>
<th>Chain-water interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptide-water hydrogen bond (N-H---O-H)</td>
</tr>
<tr>
<td>Peptide-water hydrogen bond (C=O---H-O)</td>
</tr>
<tr>
<td>Heavy atom – water oxygen contact</td>
</tr>
<tr>
<td>Nonpolar carbon – water oxygen contact</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Water-water interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>water coordination</td>
</tr>
<tr>
<td>water-water hydrogen bond</td>
</tr>
</tbody>
</table>

Disordered states of proteins may be described in terms of the balance between chain-chain and chain-solvent interactions. In order to define these interactions systematically, we employ a set of geometric criteria. We follow a similar approach to choosing these criteria as
Jorgensen et al. used in the development and testing of the optimized potentials for liquid simulations (OPLS) united atom and all atom force fields; this is done for consistency because the OPLS-AA/L force field was employed for all simulations in our study. Specifically, we consider two particles to be involved in an “interaction” if their separation distance is less than a cutoff distance. This cutoff corresponds to the first minimum in the radial distribution function (RDF) for the pair of interacting particles. We therefore begin by briefly describing the RDF, and the algorithm used to compute it.

C.2 Radial Distribution Functions

Consider a system that contains two types of particles: type A and type B. The RDF, \( g_{AB}(r) \), is the average number density of type B particles at a distance \( r \) from a type A particle, \( \langle n_{AB}(r) \rangle \), normalized by the number density of type B particles in the system, \( n_B \). Thus, the equation for the RDF is written:

\[
\frac{g_{AB}(r)}{n_B} = \frac{\langle n_{AB}(r) \rangle}{n_B}.
\]  

(C.1)

The number density, \( n_B \), is the total number of type B particles, \( N_B \), divided by the volume of the system, \( V (n_B = N_B/V) \). The RDF describes the variation in particle density as a function of the particle separation distance relative to the expected number density if the particles were randomly distributed throughout the system. The location of the first minimum of the RDF, \( r_{\text{min}} \), delineates the extent of the first shell of nearest neighbours. The average number of nearest neighbours of type A can be obtained by evaluating the integral:

\[
n_A \int_0^{r_{\text{min}}} 4\pi r^2 g_{AB}(r) dr,
\]  

(C.2)

where \( n_A \) is the number density of type A particles in the system. Thus, it is sensible to choose the separation distance at the first minimum of the RDF as the cutoff distance for an interaction. The RDF is also commonly referred to as the pair distribution function.
In practice, the RDF computed from a simulation trajectory is a histogram with the particle separation distance divided into bins of finite size. Each bin corresponds to a spherical shell surrounding the reference particle of radius \( r_i \) and thickness \( \Delta r \). The number of particles in the spherical shell, \( \text{count}(r_i) \), is counted over the frames in the simulation. In order to normalize the number of particles counted in each bin, we divide the count in each bin by the number of frames in the simulation, \( n_{\text{frames}} \), the number of pairs of particles, \( N_{\text{pairs}} \), the volume of the simulation system, \( V \), and the volume of the spherical shell, \( V_i \), as follows:

\[
g(r_i) = \frac{\text{count}(r_i)}{(n_{\text{frames}}) \left( \frac{N_{\text{pairs}}}{V} \right) (V_i)},
\]

where \( V_i = \left( \frac{4\pi}{3} \Delta r \right) \left( (i + 1)^3 - i^3 \right) \).\(^8\) RDFs in this appendix are computed using equation C.3 and the algorithm outlined by Frenkel and Smit.\(^6\) The STDR simulation configurations obtained at 300 K for the peptide \((GVPGV)\) are used to compute all RDFs and distributions in this appendix (only the equilibrated ensemble of configurations is used; refer to Chapter 6 for a description of this ensemble).

### C.3 Hydrogen bonds

#### C.3.1 Peptide-Peptide Hydrogen Bonds

Hydrogen bonds are energetically-favourable interactions between the partial positive charge of a hydrogen atom and the partial negative charge of an electronegative heteroatom. The “donor” in a hydrogen bond is a highly electronegative atom (e.g. O, N, or F) covalently-bonded to a hydrogen atom, and the “acceptor” is another highly electronegative atom. \textit{Ab initio} calculations have shown that the net stabilization of hydrogen bonds results from the interplay of several energetic contributions that are quantum mechanical in nature. However, as a “first-order” approximation, hydrogen bonds may be thought of as an electrostatic interaction.\(^1\)
In peptide-peptide hydrogen bonds, the nitrogen atoms and oxygen atoms of the polypeptidic backbone act as “donors” and “acceptors”, respectively.\textsuperscript{12,13} Hydrogen bonds can form between the amide groups of different residues. An example of a configuration with a peptide-peptide hydrogen bond involving the amide groups of two valine residues is shown in Figure C.1 (A). The geometry of a peptide-peptide hydrogen bond is illustrated in Figure C.1 (B) with the relevant distance and angle variables.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{peptide_hbonds.png}
\caption{The geometry of peptide-peptide hydrogen bonds}
\end{figure}

Figure C.1 The geometry of peptide-peptide hydrogen bonds
(A) A configuration of the peptide (GVPGV)\textsubscript{7} with an intramolecular hydrogen bond between the amide groups of two valine residues. For clarity, we only show the residues of the peptide in the vicinity of the hydrogen bond. (B) The geometry of a peptide-peptide hydrogen bond. In order to describe the geometry, three variables are used: (1) the distance between the carbonyl oxygen and amide nitrogen, $r_{ON}$, (2) the distance between the carbonyl oxygen and amide hydrogen, $r_{OH}$, and (3) the hydrogen-donor-acceptor (H-N-O) angle, $\theta$.

Peptide-peptide hydrogen bonds impart stability to elements of secondary structure, which include $\alpha$–helices, $\beta$-sheets, and hydrogen-bonded turns.\textsuperscript{12} For this reason, it is essential to accurately count hydrogen bonds as a first step in characterizing the structural properties of disordered proteins. In order to count hydrogen bonds, a specific definition of a hydrogen bond must be adopted. Definitions of hydrogen bonds are inherently empirical in nature, and involve the use of suitable cutoffs selected for the particular application.\textsuperscript{14} It is important to note that in “counting” hydrogen bonds, we are describing the interaction in terms of two states: “intact” and “broken”. Hydrogen bonds may be more realistically described as a continuum of states, and the “intact” and “broken” states may each consist of a distribution of states. Nevertheless, the two-state description of hydrogen bonds has proven to be useful.\textsuperscript{15} One of the major goals of the present study is to quantitatively characterize the relative balance of chain-chain and chain-water
interactions in disordered states. Thus, we need to count both peptide-peptide and peptide-water hydrogen bonds; we proceed by employing a definition of a hydrogen bond that relies on geometric criteria following the approach of Jorgensen and Swenson.\textsuperscript{3,4}

The first necessary condition in our definition of a peptide-peptide hydrogen bond is that the distance between the carbonyl oxygen atom and amide nitrogen atom, $r_{ON}$, is less than 3.3 Å. This cutoff in separation distance, $r_{ON,cutoff}$, corresponds to the first minimum in the pair distribution function of oxygen and nitrogen atoms, $g(r_{ON})$ (Figure C.2).

![Figure C.2 RDF for carbonyl oxygen atoms and amide nitrogen atoms](image)

The first minimum in the RDF for amide nitrogen atoms and carbonyl oxygen atoms is located at 3.3 Å. Accordingly, this distance is chosen as the cutoff distance, $r_{ON,cutoff}$. Oxygen-nitrogen pairs within the same residue and nearest neighbour residues are excluded.

The second criterion in our definition of a hydrogen bond requires that the distance between the carbonyl oxygen atom and the amide hydrogen atom, $r_{OH}$, is less than 2.7 Å. This distance cutoff, $r_{OH,cutoff}$, corresponds to the first minimum in the pair distribution function of carbonyl oxygen atoms and amide hydrogen atoms, $g(r_{OH})$ (Figure C.3). It is also consistent with the distribution of hydrogen-oxygen distances observed in protein structures in the PDB.\textsuperscript{16}

To summarize, our definition of a hydrogen bond relies on the combination of two distance criteria: $r_{ON} < 3.3$ Å and $r_{OH} < 2.7$ Å.
The first minimum in the RDF for carbonyl oxygen atoms and amide hydrogen atoms is located at 2.7 Å. Oxygen-hydrogen pairs within the same residue and nearest neighbour residues are excluded.

Hydrogen bonding criteria in other simulation studies, including our own earlier work,\textsuperscript{17,18} commonly incorporate an angle criterion and an energetic criterion in addition to distance criteria. For example, in the hydrogen-bonding analysis presented in Chapters 4 and 5, we required that the hydrogen-donor-acceptor (H-N-O) angle, $\theta$, is less than 60°, and that electrostatic energy is less than -0.5 kcal/mol.\textsuperscript{17} These criteria were chosen to be consistent with the maximum allowed angle and energy in the Dictionary of Secondary Structure in Proteins (DSSP) algorithm.\textsuperscript{14} We now explain our reasoning in revisiting our definition of a hydrogen bond, and using only the combination of the above two distance criteria to count hydrogen bonds. This revised definition is used in the analysis of Chapters 6, 7, and 8.

First, we revisit the necessity of an angle criterion. Hydrogen bonds have preferred orientations because they involve the interaction of a positively-polarized hydrogen atom with the lone pair electrons of the acceptor atom. Accordingly, directionality of peptide-peptide hydrogen bonds has been observed in globular protein structures from the PDB.\textsuperscript{12} We examine the values of $\theta$ observed if we use only distance criteria to count hydrogen bonds. In Figure C.4, the probability distribution for $\theta$ is shown for all configurations, as well as the subset of
configurations that have a hydrogen bond (that is, configurations that satisfy both $r_{OH} < 2.7 \text{ Å}$ and $r_{ON} < 3.3 \text{ Å}$). We find that counting hydrogen bonds on the basis of the two distance criteria results in values of $\theta$ that are less than $60^\circ$ in more than $99.5\%$ of configurations. Since the angles observed using only distance criteria are typical for hydrogen bonds, there is no need to include an angle cutoff. Furthermore, no requirement for a specific angle, $\theta$, is used by Jorgensen et al. in their definition of a hydrogen bond.\textsuperscript{2,4} We note that the angle cutoff of $60^\circ$ used in Chapters 4 and 5 was not necessary. There is a practical advantage to using fewer criteria in counting hydrogen bonds: it significantly reduces the total computational time required for analysis, which is important when the analysis of hundreds of millions of configurations is required (as in the present study).

![Figure C.4](image)

**Figure C.4** Probability distribution for the hydrogen-nitrogen-oxygen angle, $\theta$

In this graph, two probability distributions of the hydrogen-nitrogen-oxygen angle, $\theta$, are shown. The first probability distribution (in black, with value axis on the left) is computed using all of the configurations in the simulation. The second probability distribution (in red, with value axis on the right) is computed using only the configurations in which a hydrogen bond is present. More than $99.5\%$ of configurations with hydrogen bonds have values of $\theta$ less than $60^\circ$.

C.3.2 Analysis of the Energy of Peptide-Peptide Hydrogen Bonds

In the analysis of Chapters 4 and 5, the electrostatic energy was required to be less than $-0.5 \text{ kcal/mol}$ in order for a hydrogen bond to be counted.\textsuperscript{17} This energetic cutoff is the same as that used in the DSSP algorithm.\textsuperscript{14} DSSP counts hydrogen bonds exclusively on the basis of electrostatic energy, which allows distortions from the ideal hydrogen bond distance and angles (this criterion was motivated by the fact that DSSP was originally developed to analyse protein
structures obtained using x-ray crystallography. Electrostatic energy is computed using Coulomb’s law for each pair of partial charges; in the case of a peptide-peptide hydrogen bond, four charge pairs (O-N, O-H, C-N, and C-H) are involved. The equation for electrostatic energy, $E_{el}$, is:

$$E_{el} = \frac{1}{4\pi \varepsilon_o} \left( \frac{q_O q_N}{r_{ON}} + \frac{q_O q_H}{r_{OH}} + \frac{q_C q_N}{r_{CN}} + \frac{q_C q_H}{r_{CH}} \right);$$  \hspace{1cm} (C.4)

the subscripts $O$, $N$, $C$, and $H$ indicate atom type, $q$ is charge, $r$ is distance, and $\varepsilon_o$ is the permittivity of free space.\textsuperscript{14} When computing the electrostatic energy in the analysis of Chapters 4 and 5, we followed the same procedure as that of the DSSP algorithm, including using the same partial charges as DSSP ($q_C = 0.42e$, $q_O = -0.42e$, $q_N = -0.2e$, and $q_H = 0.2e$, where $e$ is the charge of an electron). However, it is more sensible to use the partial charges from the OPLS-AA/L force field,\textsuperscript{5} since these charges are used in the simulation ($q_C = 0.5e$, $q_O = -0.5e$, $q_N = -0.5e$, and $q_H = 0.3e$). For all configurations, the electrostatic energy computed using the OPLS-AA/L charges is less favourable than the electrostatic energy obtained using the DSSP charges. Because of this difference, we decided to investigate the energetics of hydrogen bonding using the OPLS-AA/L force field in more detail.

Hydrogen-bonding interactions in the OPLS force fields are represented implicitly through the combination of the electrostatic and Lennard-Jones terms in the non-bonded energy between atoms $i$ and $j$:\textsuperscript{2,5}

$$E_{\text{non-bonded}}(i, j) = \frac{q_i q_j e^2}{r_{ij}} + 4\varepsilon_{ij} \left( \frac{\sigma_{ij}^{12}}{r_{ij}^{12}} + \frac{\sigma_{ij}^{6}}{r_{ij}^{6}} \right);$$  \hspace{1cm} (C.5)

The first term in the non-bonded energy represents the electrostatic energy, $E_{el}$; $q_i$ and $q_j$ are the charges on atoms $i$ and $j$, $e$ is the charge of an electron, and $r_{ij}$ is the distance between atoms $i$ and $j$. The parameters in the Lennard-Jones term, $E_{LJ}$, describe the depth of the energy minimum, $\varepsilon_{ij}$, and the particle separation at minimum energy, $2^{1/6}\sigma_{ij}$, respectively. We compute
the non-bonded energy by considering interactions between all pairs of atoms in the two peptide groups involved (see Figure C.5 for a diagram showing the atoms from each residue included in this calculation). This approach is similar to that used by Buck and Karplus in their analysis of the energetics of hydrogen bonding for the CHARMM force field with the TIP3P water model. They found it was necessary to include entire peptide groups (that is, the amide C, O, N, and H atoms of both residues involved in the hydrogen bond), not only the donor and acceptor groups. In addition to the C, O, N, and H atoms of each amide group, we also include the $\alpha$-carbon and $\alpha$-hydrogen of both residues to ensure that the two interacting groups are neutral.

![Figure C.5 Atoms used in the calculation of the non-bonded energy of a hydrogen bond](image)

In computing the non-bonded energy, $E_{\text{non-bonded}}$, using equation C.5, two neutral groups are considered. The first group consists of the atoms surrounding the C=O of residue $i$ ($C_i$, $O_i$, $C_\alpha_i$, $H_\alpha_i$, $N_{i+1}$, and $H_{i+1}$), and the second group consists of the atoms surrounding the N-H of residue $j$ ($N_j$, $H_j$, $C_\alpha_j$, $H_\alpha_j$, $C_{j-1}$, and $O_{j-1}$). The index of each atom indicates the number of the residue to which it belongs.

The probability distributions of the non-bonded energy, $E_{el}$ and $E_{LJ}$ as a function of the oxygen-hydrogen separation distance are shown in Figure C.6 (note that these distributions are normalized by the volume of the spherical shell at each radius, as described in section C.2). We find that the average non-bonded energy of a peptide-peptide hydrogen bond is -5.58 kcal/mol. This value is consistent with both simulations and experiment, which place the expected energy of a peptide-peptide hydrogen bond between -5 and -6 kcal/mol. Since the average $E_{el}$ is -5.43 kcal/mol, the primary energetic stabilization comes from the electrostatic energy. There is also a small stabilization due to the Lennard-Jones term ($E_{LJ}$ is -0.15 kcal/mol on average). The non-
Figure C.6 2D energy distributions for peptide-peptide hydrogen bonds
(A) $E_{\text{non-bonded}}$ vs. $r_{\text{OH}}$ for all configurations. (B) $E_{\text{non-bonded}}$ vs. $r_{\text{OH}}$ for configurations with a hydrogen bond. (C) $E_d$ vs. $r_{\text{OH}}$ for all configurations. (D) $E_d$ vs. $r_{\text{OH}}$ for configurations with a hydrogen bond. (E) $E_{ij}$ vs. $r_{\text{OH}}$ for all configurations. (F) $E_{ij}$ vs. $r_{\text{OH}}$ for configurations with a hydrogen bond. Colour represents the probability of a given $(r_{\text{ON}}, E_{\text{non-bonded}})$ pair (the darkest red colour in each plot corresponds to the most probable pair in the distribution; white regions are the least populated).
bonded energy is negative for 99.98% of configurations with a hydrogen bond (counted using the two distance criteria; see Figure C.6 B, D, and F). For this reason, we do not include an energetic criterion in our definition of a peptide-peptide hydrogen bond. We perform a similar analysis of peptide-water hydrogen bonds in the next section.

C.3.3 Peptide-Water Hydrogen Bonds

In addition to peptide-peptide hydrogen bonds, the amide groups of the polypeptidic backbone can also form hydrogen bonds to water molecules. There are two possible scenarios: (1) the nitrogen of the amide group acts as a donor and the oxygen of a water molecule acts as an acceptor (Figure C.7 A), and (2) the carbonyl oxygen acts as an acceptor and the water oxygen acts as a donor (Figure C.7 B).

![Figure C.7 Geometry of peptide-water hydrogen bonds](image)

(A) Three variables are used to describe the geometry of a hydrogen bond with an amide nitrogen of the peptidic backbone acting as a donor, and the oxygen of a water molecule acting as an acceptor: (1) $r_{NO\text{wat}}$, the distance between the amide nitrogen and water oxygen, (2) $r_{HN\text{wat}}$, the distance between the amide hydrogen and water oxygen, and (3) $\theta$, the hydrogen-donor-acceptor angle. (B) Similarly, three variables are used to describe the geometry of a hydrogen bond with a water oxygen acting as a donor, and a carbonyl oxygen of the peptide backbone acting as an acceptor: (1) $r_{OO\text{wat}}$, the distance between the carbonyl oxygen and water oxygen, (2) $r_{OH\text{wat}}$, the distance between the carbonyl oxygen and water hydrogen, and (3) $\theta$, the hydrogen-donor-acceptor angle.

Analogous to the geometric criteria defined in section C.3.1 for peptide-peptide hydrogen bonds, we define a hydrogen bond between an amide group of the backbone and a water molecule if both of the following criteria are satisfied: $r_{\text{donor-acceptor}} < 3.3$ Å and $d_{H-\text{acceptor}} <$
2.7 Å. These criteria are based on the corresponding RDFs for NH-water distances (Figure C.8 A and C.8 B), and carbonyl oxygen-water distances (Figure C.8 C and C.8 D). These criteria are identical to those used for peptide-peptide hydrogen bonds (see Table C.1).

![RDFs used to determine distance criteria for peptide-water hydrogen bonds](image)

**Figure C.8 RDFs used to determine distance criteria for peptide-water hydrogen bonds**

(A) RDF for amide nitrogen atoms and water oxygen atoms. (B) RDF for amide hydrogen atoms and water oxygen atoms. (C) RDF for amide oxygen and water oxygen atoms. (D) RDF for amide oxygen and water hydrogen atoms. The distance cutoffs are indicated. In each case, we use \( r_{\text{donor-acceptor, cutoff}} = 3.3 \) Å and \( r_{\text{hydrogen-acceptor, cutoff}} = 2.7 \) Å (the same criteria as for peptide-peptide hydrogen bonds).

We can compute the non-bonded energy for peptide-water hydrogen bonds as we did for peptide-peptide hydrogen bonds in section C.3.2. In this case, we use the OPLS-AA/L\(^5\) charges and Lennard-Jones parameters for the backbone atoms, and the TIP3P\(^{22}\) charges and Lennard-Jones parameters for water in equation C.5. The atoms included for each group in computing the non-bonded energy are shown in Figure C.9. The distributions of non-bonded energy for N-H --- O\(_{\text{wat}}\)-H\(_{\text{wat}}\) hydrogen bonds and C=O --- H\(_{\text{wat}}\)-O\(_{\text{wat}}\) hydrogen bonds are shown in Figures
C.10 and C.11, respectively. We find that the non-bonded energies of peptide-water hydrogen bonds are on average less favourable than the average non-bonded energy of peptide-peptide hydrogen bonds (-5.58 kcal/mol). In particular, the average non-bonded energies are -4.10 kcal/mol and -4.32 kcal/mol for N-H --- water and C=O --- water hydrogen bonds, respectively. Since the non-bonded energies are favourable for nearly all hydrogen bonds counted using the combination of distance criteria (the non-bonded energies are negative in 97.9 % and 98.3 % of all N-H --- water and C=O --- water hydrogen bonds, respectively), we do not include an energy criterion in our definition of a hydrogen bond. In the following section, we perform a similar energetic analysis for water-water hydrogen bonds. We then compare the energies of each type of hydrogen bond.

![Figure C.9](image)

**Figure C.9** Atoms used to compute the non-bonded energy of peptide-water hydrogen bonds

In computing the non-bonded energy, $E_{\text{non-bonded}}$, using equation C.5, two neutral groups are considered. (A) For an N-H---water hydrogen bond, all 3 atoms of the water molecule are included, as well as the atoms surrounding the N-H of residue i (N$_i$, H$_i$, Ca$_i$, Ha$_i$, C$_{i-1}$, and O$_{i-1}$). (B) For a C=O---water hydrogen bond, all 3 atoms of the water molecule are included, as well as the atoms surrounding the C=O of residue i (C$_i$, O$_i$, Ca$_i$, Ha$_i$, N$_{i+1}$, and H$_{i+1}$).
Figure C.10 2D energy distributions for N-H --- water hydrogen bonds

(A) $E_{\text{non-bonded}}$ vs. $r_{\text{HOOwat}}$ for all configurations.  
(B) $E_{\text{non-bonded}}$ vs. $r_{\text{HOOwat}}$ for configurations with a hydrogen bond.  
(C) $E_{\text{el}}$ vs. $r_{\text{HOOwat}}$ for all configurations.  
(D) $E_{\text{el}}$ vs. $r_{\text{HOOwat}}$ for configurations with a hydrogen bond.  
(E) $E_{\text{LJ}}$ vs. $r_{\text{HOOwat}}$ for all configurations.  
(F) $E_{\text{LJ}}$ vs. $r_{\text{HOOwat}}$ for configurations with a hydrogen bond.  
Colour represents the probability of a given ($r_{\text{HOOwat}}$, $E_{\text{non-bonded}}$) pair (the darkest red colour in each plot corresponds to the most probable pair in the distribution; white regions are the least populated).
Figure C.11  2D energy distributions for C=O --- water hydrogen bonds

(A) $E_{\text{non-bonded}}$ vs. $r_{\text{OHwat}}$ for all configurations.  (B) $E_{\text{non-bonded}}$ vs. $r_{\text{OHwat}}$ for configurations with a hydrogen bond.  (C) $E_d$ vs. $r_{\text{OHwat}}$ for all configurations.  (D) $E_d$ vs. $r_{\text{OHwat}}$ for configurations with a hydrogen bond.  (E) $E_{\text{LJ}}$ vs. $r_{\text{OHwat}}$ for all configurations.  (F) $E_{\text{LJ}}$ vs. $r_{\text{OHwat}}$ for configurations with a hydrogen bond.  Colour represents the probability of a given ($r_{\text{OHwat}}$, $E_{\text{non-bonded}}$) pair (the darkest red colour in each plot corresponds to the most probable pair in the distribution; white regions are the least populated).
C.3.4 Water-Water Hydrogen Bonds

Two water molecules form a hydrogen bond if the distance between donor and acceptor is less than 3.5 Å (both donor and acceptor are oxygen atoms in this case) and the distance between hydrogen and acceptor is less than 2.4 Å. These cutoffs are chosen to correspond to the minima in the corresponding RDFs (Figure C.12 A and B). We also perform the same analysis of non-bonded energy for water-water hydrogen bonds as we did for peptide-peptide and peptide-water hydrogen bonds using equation C.5 and the TIP3P parameters for water (Figure C.13). The average non-bonded energy of a water-water hydrogen bond is -3.88 kcal/mol, which is greater than the non-bonded energy of both peptide-peptide hydrogen bonds and peptide-water hydrogen bonds. A summary of the average energies of each type of hydrogen bond is provided in Table C.2.

Figure C.12 RDFs for water oxygen atoms and water hydrogen atoms
(A) The RDF of water oxygen pairs. (B) The RDF of water oxygen and water hydrogen pairs (hydrogen and oxygen pairs within the same water molecule are excluded).
Figure C.13  2D energy distributions for water --- water hydrogen bonds

(A) $E_{\text{non-bonded}}$ vs. $r_{\text{Owat-Hwat}}$ for all configurations.  (B) $E_{\text{non-bonded}}$ vs. $r_{\text{Owat-Hwat}}$ for configurations with a hydrogen bond.  (C) $E_{el}$ vs. $r_{\text{Owat-Hwat}}$ for all configurations.  (D) $E_{el}$ vs. $r_{\text{Owat-Hwat}}$ for configurations with a hydrogen bond.  (E) $E_{LJ}$ vs. $r_{\text{Owat-Hwat}}$ for all configurations.  (F) $E_{LJ}$ vs. $r_{\text{Owat-Hwat}}$ for configurations with a hydrogen bond.  Colour represents the probability of a given ($r_{\text{Owat-Hwat}}$, $E_{\text{non-bonded}}$) pair (the darkest red colour in each plot corresponds to the most probable pair in the distribution; white regions are the least populated).
Figure C.14  Probability distributions of $E_{\text{non-bonded}}$ for peptide-peptide, peptide-water and water-water hydrogen bonds
Shown here are the probability distributions of the non-bonded energy (in kcal/mol) for the four possible types of hydrogen bonds (peptide --- peptide, dark green; water --- water, blue; peptide N-H --- water, red; peptide C=O --- water, purple).

Table C.2  Average energies of hydrogen bonds

<table>
<thead>
<tr>
<th>Hydrogen Bond Type</th>
<th>Average Non-bonded Energy</th>
<th>Average Electrostatic Energy</th>
<th>Average Lennard-Jones Energy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peptide-Peptide</td>
<td>-5.58 kcal/mol</td>
<td>-5.43 kcal/mol</td>
<td>-0.15 kcal/mol</td>
</tr>
<tr>
<td>Peptide-Water (N-H)</td>
<td>-4.10 kcal/mol</td>
<td>-4.71 kcal/mol</td>
<td>0.62 kcal/mol</td>
</tr>
<tr>
<td>Peptide-Water (C=O)</td>
<td>-4.32 kcal/mol</td>
<td>-4.93 kcal/mol</td>
<td>0.61 kcal/mol</td>
</tr>
<tr>
<td>Water-Water</td>
<td>-3.88 kcal/mol</td>
<td>-5.18 kcal/mol</td>
<td>1.32 kcal/mol</td>
</tr>
</tbody>
</table>

It is useful to consider the energy of hydrogen bonds in terms of the “hydrogen bond inventory”.\textsuperscript{23,24} We can write this inventory as follows:

$$\text{N-H --- water} + \text{water --- O=C} \rightarrow \text{N-H --- O=C + water---water}$$  \hspace{1cm} (C.6)

Because there are two hydrogen bonds on both sides of this equation, to a first approximation, one might expect this “reaction” to involve no net change in energy.\textsuperscript{23,24} The energies for the peptide-water hydrogen bonds (left side of equation C.6) add up to -8.42 kcal/mol, while the
energy for a peptide-peptide hydrogen bond and the energy for a water-water hydrogen bond add up to -9.46 kcal/mol. There is a net difference in energy of -1.04 kcal/mol in favour of forming a peptide-peptide hydrogen bond and water-water hydrogen bond (that is, favouring the right side of equation C.6). The probability distribution for the total non-bonded energy of each type of hydrogen bond is shown in Figure C.14.

We are interested in this difference in energy because of the results of a project related to this thesis, but which has not been described here because it falls outside the scope of the present study. Briefly, we applied the same methodology described in Chapter 6 to perform massive sampling of the unfolded state of the N-terminal SH3 domain of the protein drk. The results of our STDR simulation, in combination with experimental measurements of the hydrodynamic radius from NMR,\textsuperscript{25} suggest that the unfolded state ensemble obtained using simulation is significantly more compact than the experimentally-observed unfolded state ensemble. In particular, the radius of gyration of the unfolded state obtained using STDR is only 1.1 times that of the folded state, compared to the experimentally-determined ratio of hydrodynamic radii of 1.3. Thus, the configurations in the ensemble generated using the STDR algorithm (with the OPLS-AA/L and TIP3P force fields) are on average significantly more compact than the experimentally-observed unfolded state ensemble. We hypothesize that this significant discrepancy between simulation and experiment may be related to the difference in the non-bonded energy of hydrogen bonds. In particular, the fact that the right side of equation C.6 is more energetically-favourable than the left side favours peptide-peptide self-interactions over peptide-water interactions. If peptide-peptide interactions are overly favourable, this may lead to an unrealistically compact unfolded state ensemble.

More generally, this leads us to ask the question: are the molecular mechanics force fields used here able to represent peptide-water interactions accurately? We note that the OPLS set of force fields were originally parameterized to reproduce experimental thermodynamic properties of liquids, such as the heat of vaporization.\textsuperscript{3} In particular, the non-bonded parameters for the amide group of the polypeptidic backbone are derived from the parameters of N-methylacetamide (NMA). These parameters were based on MC simulations of liquid NMA,
and showed good agreement with experimental heat of vaporization and density (with errors of less than 2 %). However, given that the parameters for NMA were not developed using simulations with water molecules or developed to agree with hydration of amides, there is no reason to expect that these parameters will result in an accurate representation of the interaction between the polypeptidic backbone and water. The hydration of amides was investigated by Jorgensen and Swenson, and they found a favourable non-bonded energy when $E_{\text{non-bonded}}$ was computed for all pairs of atoms in the NMA molecule and a TIP4P water molecule. A deeper investigation of these issues is beyond the scope of this thesis, but will be pursued as a future direction. In particular, we will investigate the energetics of hydrogen bonding in various force field combinations to determine if this observation is a widespread issue.

C.4 Other Chain-Chain Interactions

C.4.1 Contacts Between Residues Based on $\alpha$-carbons (C$\alpha$-contacts)

Contacts between two residues are often defined based on the proximity of their $\alpha$-carbon atoms. Counting contacts in this way does not describe a particular type of interaction (e.g. hydrogen bonding or nonpolar interactions); rather, it counts all types of interactions that bring the residues close in space. Based on the pair distribution function for $\alpha$-carbon atoms (Figure C.15), we may define a cutoff for these interactions of 6.5 Å. This cutoff is located approximately between what appear to be two overlapping peaks, and it is in agreement with the value used in another study of a disordered peptide, KLVFFAE, which is a fragment of Aβ. However, we note that it is quite difficult to define a meaningful cutoff that is applicable to all residue types because of the differing sizes of side chains. For example, there is a shoulder at low $\alpha$-carbon separation distance in the RDF that is due to glycine-glycine contacts. Because the choice of cutoff is unclear in this case, we note that it is difficult to attribute a specific meaning to the C$\alpha$-contacts. For this reason, we do not make use of contacts defined using only the $\alpha$-carbons in this study. It is important to define contacts in a way that accounts for side chains with varying number of atoms, which is the objective of the contacts defined in the next two sections.
The RDF for $\alpha$-carbon atoms of the polypeptidic backbone has a complicated shape. As a cutoff, we choose 6.5 Å, which appears to separate two overlapping peaks. $\alpha$-carbon atoms of nearest neighbour residues are excluded.

**C.4.2 Contacts Between Residues Based on Contacts Between Heavy Atoms (HA-contacts)**

Another common method of defining a contact between two residues utilizes the pairwise distances between their heavy atoms. If any two of their heavy atoms are closer than a specified cutoff distance, then the two residues are said to be in contact with each other. We refer to these contacts as HA-contacts throughout the thesis. Defining contacts between residues in this way is more informative than relying on the $\alpha$-carbons alone because it accounts for the varying sizes of side chains. Based on the RDF for heavy atoms (Figure C.16), we define a cutoff for these interactions of 4.5 Å. We note that the choice of cutoff in this case is not a minimum in the RDF, but appears to correspond to the separation of two overlapping peaks. Using a cutoff of 4.5 Å to define these contacts is in agreement with studies of both disordered peptides and globular proteins, and is sufficiently short to avoid counting solvent separated pairs as contacts. HA-contacts simultaneously count the two types of chain-chain interactions: backbone-backbone and side chain-side chain contacts. Thus, they provide an overall count of chain self-interactions.
The pair correlation function for heavy atoms has a complex shape because it results from several different types of interactions (e.g. both hydrogen-bonded backbone-backbone interactions and nonpolar side chain-side chain interactions). As a cutoff, we choose 4.5 Å, which appears to separate two underlying peaks, and is sufficiently short to exclude the possibility of counting solvent-separated pairs as contacts. Heavy atoms within the same residue and in nearest neighbour residues are excluded.

C.4.3 Contacts Between Residues Based on Contacts Between Nonpolar Carbons (NP-contacts)

Because of the hydrophobic effect, nonpolar groups tend to associate in water. As part of our analysis of chain-chain interactions, we quantify the extent of nonpolar interactions, which we refer to as NP-contacts. An NP-contact between two residues is present if any pair of nonpolar carbons are closer than the cutoff distance, $r_{CC,cutoff}$. A nonpolar carbon is defined here to be any carbon atom covalently-bonded to either two or three hydrogen atoms. We choose a cutoff of 5.4 Å, which corresponds to a local minimum in the pair correlation function of nonpolar carbons (Figure C.17). This cutoff is sufficiently short to avoid miscounting solvent-separated nonpolar carbons as contacts, and is consistent with the cutoff for nonpolar contacts used in another study of an elastin-like peptide.
Figure C.17 RDF for nonpolar carbon atoms

The RDF for nonpolar carbons has two peaks. The first peak corresponds to a nonpolar contact, while the second peak corresponds to a solvent-separated contact. As a cutoff for nonpolar interactions, we choose 5.4 Å, which is the local minimum between these peaks. Nonpolar carbon atoms within the same residue and in nearest neighbour residues are excluded.

C.5 Other Chain-Water Interactions

C.5.1 Hydration Shell Defined Using Heavy Atoms (HA hydration shell)

In addition to counting hydrogen bonds between the polypeptidic backbone and water molecules (section C.3.3), we also consider non-specific interactions with water. The first method we use involves counting the number of water molecules in the hydration shell of the peptide. We define this hydration shell based on the distance between water oxygen atoms and heavy atoms of the peptide; it is therefore called the HA-hydration shell. In particular, a water molecule is considered to be in the hydration shell if its oxygen atom is within a cutoff \( r_{\text{HA-Owat, cutoff}} \) of any heavy atom of the peptide. This cutoff is chosen based on the pair correlation function of peptide heavy atoms and water oxygen atoms (Figure C.18 A). In this case, we do not use the first minimum of the RDF to define the hydration shell because this would result in only counting hydrogen-bonded interactions. We are interested here in counting non-specific interactions in addition to hydrogen-bonded interactions. Thus, we use a cutoff of 5.5 Å. This choice of cutoff to define the hydration shell is similar to that used in another simulation study of an elastin-like peptide: Li and Daggett used a cutoff of 4.5 Å for the distance between water
oxygen atoms and any protein atom.\textsuperscript{30} This is essentially the same as our cutoff because we do not consider peptide hydrogen – water oxygen distances. By only computing distances between peptide heavy atoms and water oxygen atoms (and not considering the peptide hydrogen atoms), the speed of the hydration analysis is dramatically increased.

C.5.2 Hydration Shell Defined Using Nonpolar Carbon Atoms (NP hydration shell)

We define a second type of hydration shell: the nonpolar hydration shell (NP hydration shell). A water molecule is considered to be in the NP hydration shell if its oxygen atom is within a cutoff ($r_{C-O\text{wat, cutoff}}$) of any nonpolar carbon atom of the peptide. For this hydration shell, we use the same cutoff (5.5 Å) as we used for the HA-hydration shell (see Figure C.18 B).

![Figure C.18](image)

**Figure C.18** RDFs used to define the hydration shell and nonpolar hydration shell
(A) The RDF heavy atom and water oxygen pairs. (B) The RDF for nonpolar carbon and water oxygen pairs. In both cases, we use a cutoff of 5.5 Å.

C.6 Radius of Gyration

The radius of gyration is defined in two ways:\textsuperscript{31}

1. As the average distance of particles from the center of mass:

$$R_g^2 = \frac{1}{N} \sum_{i=1}^{N} M_i \left| \vec{r}_i - \vec{r}_{COM} \right|^2.$$  \hspace{1cm} (C.7)
where $\vec{r}_i$ is the position of particle $i$, $N$ is the number of particles, $\vec{r}_{COM}$ is the position of the center of mass of the $N$ particles ($\vec{r}_{COM} = \frac{1}{N} \sum_{i=1}^{N} M_i \vec{r}_i$), and $M_i$ is the mass of particle $i$.

(2) As the average pairwise distance between particles:

$$R_g^2 = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} |\vec{r}_i - \vec{r}_j|^2.$$  \hspace{1cm} (C.8)

We use the first definition of the radius of gyration for monomers (considering all heavy atoms of the backbone). The second definition is used in the analysis of the elastin-like aggregate in Chapter 8; this is done for convenience. Because of periodic boundary conditions, the individual monomers are in different image cells in the trajectory; to avoid the additional overhead of computing the location of the center of mass, the analysis is simplified by using equation C.7 with the minimum image distance between pairs of atoms. Because the sum in equation C.7 is not mass-weighted, this measure of $R_g$ involves the approximation that all heavy atoms have the same mass. That is, C.7 and C.8 are only equivalent if the masses of all atoms are the same. This approximation changes the value of $R_g$ by less than one percent, and it is sensible if we consider that each carbon and nitrogen atom is covalently bonded to at least one hydrogen atom. We note that C.7 and C.8 are equivalent if only $\alpha$-carbons are used in the calculation of $R_g$. 
References


