UNIVERSITY OF CALIFORNIA, SAN DIEGO

Development and Application of All-Atom Structure-based Models for Studying the Role of Geometry in Biomolecular Folding and Function

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Physics

by

Jeffrey Kenneth Noel

Committee in charge:

Professor José Onuchic, Chair
Professor Patricia Jennings
Professor Herbert Levine
Professor Andrew McCammon
Professor Peter Wolynes

2012

The dissertation of Jeffrey Kenneth Noel is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2012

Every passing hour brings the Solar System forty-three thousand miles closer to Globular Cluster M13 in Hercules – and still there are some misfits who insist that there is no such thing as progress.

—Kurt Vonnegut in *The Sirens of Titan*

TABLE OF CONTENTS

Signature Page iii

Epigraph iv

Table of Contents v

List of Figures viii

List of Tables x

Acknowledgements xi

Vita and Publications xiii

Abstract of the Dissertation xv

Chapter 1 Structure-based Models Capture the Geometric Aspects of Biomolecular Dynamics 1
1.1 Introduction 1
1.2 Structure-based Models 2
1.2.1 Foundations in Energy Landscape Theory 2
1.2.2 Structure-based Model as a Baseline 4
1.3 Implementation of Structure-based Models 5
1.3.1 Structure-based Potentials 6
1.3.2 Choosing a Graining: C_α or All-atom 12
1.3.3 Molecular Dynamics with SBM 13
1.4 Outline of the Thesis 17
1.5 Acknowlegements 21

Chapter 2 An All-atom Structure-Based Potential for Proteins: Bridging Minimal Models with All-atom Empirical Forcefields 22
2.1 Introduction 22
2.2 Results 25
2.2.1 Folding Mechanisms Are Robust to Parameter Changes 25
2.2.2 Fully Folded Backbone Allows for Disordered Side Chains 28
2.2.3 Understanding Free Energy Profiles Through Parametric Variation: Free Energy Profiles Can Be Altered Through Parametric Changes 30
2.2.4 All-Atom Structure-based Simulations Capture C_α Folding Mechanism 34

2.2.5 Native Basin Dynamics of AA Structure-Based Model Correlate with the Dynamics of an All-atom Empirical Forcefield With Explicit Solvent 35
2.3 Discussion 36
2.4 Methods 38
2.4.1 Simulation Details 38
2.4.2 All-Atom Empirical Forcefield Simulations 38
2.4.3 Proline to Alanine Mutations 39
2.4.4 Comparison of Contacts 39
2.5 Acknowledgements 39

Chapter 3 The Shadow Map: A General Contact Definition for Capturing the Dynamics of Biomolecular Folding and Function 41
3.1 Introduction 41
3.2 Methods 43
3.2.1 The All-Atom Structure-Based Model 43
3.2.2 Simulation Details 44
3.2.3 Contact Maps 44
3.2.4 Contact Potential 45
3.2.5 Thermodynamics 48
3.3 Results and Discussion 49
3.3.1 Protein Contact Maps 49
3.3.2 Decoupling the Protein Geometry from the Contact Energy Distribution 55
3.3.3 Shadowing Tends to Increase Folding Cooperativity 58
3.3.4 Dynamics of RNA and Macromolecular Assemblies 60
3.4 Conclusions 64
3.5 Acknowledgments 67

Chapter 4 SMOG@ctbp: Simplified Deployment of Structure-based Models in Gromacs 68
4.1 Introduction 68
4.2 Implementing a Structure-based Model in Gromacs 69
4.2.1 Web Server Interaction 69
4.2.2 Molecular Dynamics with Gromacs 72
4.3 Conclusion 75
4.4 Acknowledgements 75

Chapter 5 The Free Energy Landscape of a Trefoil-Knot Protein: Slipknotting upon Native-Like Loop Formation 76
5.1 Introduction 76
5.2 Results and Discussion 80

5.2.1 Thermodynamically Meta-stable State Precedes Knotting . . . 80
5.2.2 Non-specific Knots, Native Knots and Malformed Knots . . . 82
5.2.3 Folding Mechanism of a Knotted Domain . . . 84
5.2.4 Slipknotting is a General Knotting Mechanism . . . 86
5.2.5 Topological Traps on the Folding Landscape . . . 87
5.2.6 Addition of Side Chains Reduces Topological Trapping 89
5.2.7 Dimerization Occurs After Knotting . . . 91
5.3 Conclusions . . . 94
5.4 Future Work: Detailed Simulations with Anton . . . 95
5.5 Methods and Notes . . . 96
5.5.1 All-Atom Model . . . 96
5.5.2 Reaction Coordinates . . . 97
5.5.3 Route Measure . . . 97
5.5.4 Identification of the Knot Along the Protein . . . 97
5.5.5 Note About Knot Chirality in Proteins . . . 98
5.5.6 Comparing Kinetic Folding . . . 98
5.5.7 Movie of a Slipknot Folding Trajectory . . . 100
5.6 Acknowledgments . . . 100

Chapter 6 Mirror Images as Naturally Competing Conformations in Protein Folding . . . 101
6.1 Introduction . . . 101
6.2 Methods . . . 104
6.2.1 Structure Prediction . . . 105
6.2.2 REMD Simulations . . . 106
6.2.3 Structure-based Models . . . 107
6.2.4 Contact Map Definitions . . . 108
6.3 Results . . . 109
6.3.1 Coarse-grained Native-centric Protein Model Populates the Mirrored Basin . . . 109
6.3.2 Mirrored Structures are Energetically Competitive . 110
6.3.3 Atomistic Simulations Show a Mirror Basin that is Thermodynamically Competitive and Kinetically Accessible . . . 112
6.4 Discussion and Conclusions . . . 119
6.5 Acknowledgements . . . 121

Bibliography . . . 122

LIST OF FIGURES

Figure 1.1: Tertiary interactions in the SBM are defined through the native contact map. . . . 7
Figure 1.2: Comparison of Lennard-Jones and Gaussian contact potentials. . . . 10
Figure 1.3: Comparison of AA SBM and explicit solvent simulations of tRNA accommodation in the ribosome. . . . 12
Figure 1.4: All-atom structure-based simulations of folding for the two-state proteins CI2 and SH3 domain. . . . 16

Figure 2.1: Protein geometrical representations: cartoon, coarse-grained by residue, and all heavy atoms. . . . 23
Figure 2.2: Structures of Protein A, SH3, and CI2. . . . 24
Figure 2.3: Reaction coordinates of the AA model. . . . 26
Figure 2.4: Folding mechanism sensitivity to the balance of local to non-local stabilizing energy. . . . 27
Figure 2.5: Difference in AA contact formation and C_α contact formation. . . . 29
Figure 2.6: Free energy barriers in the AA model. . . . 31
Figure 2.7: Comparison of backbone folding between C_α and AA structure-based models. . . . 33
Figure 2.8: Comparison of contact formation in the native state between the AA structure-based potential and an all-atom empirical forcefield. . . . 35

Figure 3.1: The Shadow contact map screening geometry. . . . 42
Figure 3.2: The versatile Gaussian contact potential. . . . 45
Figure 3.3: Measures of cooperativity. . . . 47
Figure 3.4: The removal of contacts through shadowing. . . . 52
Figure 3.5: Shadow automatically includes contacts where ligands, metal clusters and buried waters are not explicitly represented. . . . 53
Figure 3.6: Excluded volume imparts cooperativity. . . . 56
Figure 3.7: Heat capacity is consistent and folding is cooperative as the cutoff parameter C is varied with the Shadow algorithm. . . . 59
Figure 3.8: Folding an RNA hairpin with the all-atom SBM. . . . 63
Figure 3.9: Distribution of native contact energy between proteins and RNA in the ribosome. . . . 64
Figure 3.10: Structure-based models capture near-native-state fluctuations of both small proteins and large macromolecular assemblies. . . . 65

Figure 4.1: Performance of an all-atom structure-based simulation with Gromacs version 4.5 for a ribosome with 142,196 atoms. . . . 69
Figure 4.2: Flowchart explaining the logic of the SMOG@ctbp web server. . . . 71
Figure 4.3: Structure-based model of the folding of the SH3 domain using SMOG@ctbp with default parameters and Shadow contact map. . . . 73

Figure 4.4: Usage statistics for the SMOG@ctbp web server. 74

Figure 5.1: Structure of the smallest knotted protein. 78
Figure 5.2: Folding routes of a knotted protein. 81
Figure 5.3: Non-specific knots during folding. 83
Figure 5.4: Two possible native loops for threading the trefoil knot. 86
Figure 5.5: Energy landscape of a knotted protein and possible topological traps for a 3_1 knot. 88
Figure 5.6: Specific side-chain packing reduces topological frustration. 90
Figure 5.7: Structure of MJ0366 as a homodimer with β_2-β_2' forming the majority of the dimer interface along with α_2-α_4' and α_4-α_2'. 92
Figure 5.8: Folding with two monomers present. 93
Figure 5.9: Two representative kinetic trajectories from the AA model that fall into topological traps at $T = 0.91T_F$. 99

Figure 6.1: Protein symmetry gives rise to multiple consistent structures. 103
Figure 6.2: Native-centric, coarse-grained protein model populates mirrored structures. 109
Figure 6.3: Distribution of structure prediction energies versus rmsd from the PDB structures for Edpa and α_3d. 111
Figure 6.4: Clustering of the REMD by rmsd. 113
Figure 6.5: Kinetic accessibility and thermodynamic stability of the mirror image. 115
Figure 6.6: Hydrophobic core packing differs between the native and mirror in Bdpa. 117

LIST OF TABLES

Table 3.1: Statistics on various contact maps of model globular proteins. . . . 50
Table 3.2: Comparison of cutoff versus shadowing contact maps in RNA systems. 61

Table 5.1: Relative population of folding routes (in %) for the AA and C_α model. 87

ACKNOWLEDGEMENTS

First and foremost I would like to thank my advisor José Onuchic for his scientific and professional guidance. While usually encouraging me find my own way, he provided a necessary bias to my random walk.

I would also like to thank my coauthors: Paul Whitford, Alex Schug, Shachi Gosavi, Angel Garcia, Karissa Sanbonmatsu, Yoko Suzuki, Joanna Sułkowska, Abhinav Verma, and Wolfgang Wenzel. In particular, I would like to thank Paul and Joanna for their invaluable contributions to my development as a scientific writer and communicator. I am grateful for Patricia Jennings' enthusiasm for my work, scientific camaraderie and willing critical eye. I am also greatly indebted to the many stimulating discussions with the members of the Onuchic and Wolynes research group, especially Peter Wolynes who teaches us how to sharpen our arguments. I have also spent many productive hours speculating with Heiko Lammert and Ryan Hayes and unproductive hours chumming with Anat Burger.

I am indebted to Paul for his willingness to collaborate on all-atom structure-based models and his determination in refining the Fortran code running behind the scenes in SMOG. Also to Joanna for sharing the fascinating problem of folding knotted proteins. Also to Yoko for, every six months, sharing the joys of analytical calculation.

My time in San Diego would have been dull without the friendship and constant stimulation from Jonathan, whose wicked wit won't wane, Matt, who is a solid dude, Alex, with whom I shared many miles through the mountains, Paul, who taught me a little paranoia is valuable, Diego, who shared the zen of mate, and last, but certainly not least, Laura, for Nike rockstar dancehall workout. I would like to thank my family for their support and encouragement. Thanks to Tom for always challenging me and to my mother for fostering adept argumentation. Thanks also to my father for the months spent in his lab, which have been so valuable for understanding and relating to experimentalists.

I acknowledge that most of the text and figures in the thesis have appeared in print elsewhere. Chapter 1, in part, appears in a book chapter "The Many Faces of Structure-Based Potentials: From Protein Folding Landscapes to Structural Characterization of Complex Biomolecules," *Computational Modeling of Biological Systems*,

(2012), Noel and Onuchic. The dissertation author was the primary investigator and author of the chapter. Chapter 2, in part, appears in *Proteins: Structure, Function, Bioinformatics*, (2009), Whitford, Noel, Gosavi, Schug, Onuchic. The dissertation author and Paul Whitford are the primary investigators and coauthors of the paper. Chapter 3, in part, appears in *Journal of Physical Chemistry B*, (2012, in press), Noel, Whitford, Onuchic. The dissertation author is the primary investigator and author of the paper. Chapter 4, in part, appears in *Nucleic Acids Research*, (2010), Noel, Whitford, Sanbonmatsu, Onuchic. The dissertation author was the primary investigator and author of the paper. Chapter 5, in part, appears in *PNAS*, (2010), Noel, Sułkowska, Onuchic. The dissertation author was the primary investigator and author of the paper. Chapter 6, in part, appears in *Journal of Physical Chemistry B*, (2012, in press), Noel, Schug, Verma, Wenzel, Garcia, Onuchic. The dissertation author was the primary investigator and the author of the paper.

Financially, I am indebted to the support provided by a San Diego Fellowship, the NIH Molecular Biophysics Training Program, and a research assistantship from the Center for Theoretical Biological Physics.

VITA

2005	Bachelor of Science in Applied Mathematics, Engineering, and Physics, University of Wisconsin, Madison
2007	Master of Science in Physics, University of California, San Diego
2009	Candidate in Philosophy in Physics, University of California, San Diego
2012	Doctor of Philosophy in Physics, University of California, San Diego

PUBLICATIONS

Sułkowska JI, Noel, JK, Onuchic JN, The free energy landscape of knotted proteins. *Proc. Natl. Acad. Sci.* (submitted)

Noel JK, Whitford PC, Onuchic JN, The Shadow Map: A General Contact Definition for Capturing the Dynamics of Biomolecular Folding and Function. *J. Phys. Chem. B*, 2012. (in press)

Noel JK, Schug A, Verma A, Wenzel W, Garcia AE, Onuchic JN, Mirror Images as Naturally Competing Conformations in Protein Folding. *J. Phys. Chem. B*, 2012. (in press)

Noel JK, Onuchic JN, The many faces of structure-based potentials: From protein folding landscapes to structural characterization of complex biomolecules. Chapter 2 in *Computational Modeling of Biological Systems*; Dokholyan, N. Ed., Springer: New York, 2012.

Suzuki Y, Noel JK, Onuchic JN, A semi-analytical description of protein folding that incorporates detailed geometrical information. *J. Chem. Phys.*, 135(245101), 2011.

Noel JK, Sułkowska JI, Onuchic JN, Slipknotting upon native-like loop formation in a trefoil knot protein. *Proc. Natl. Acad. Sci. USA* 107(15403-8), 2010.

Noel JK, Whitford PC, Sanbonmatsu KY, Onuchic JN, SMOG@ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res.*, 38(W657-61), 2010.

Whitford PW, Noel JK, Gosavi S, Schug A, Onuchic JN, An all-atom structure-based potential for proteins: bridging minimal models with empirical forcefields. *Proteins: Struct. Func. Bioinfo.*, 75(430-441), 2009.

Suzuki Y, Noel JK, Onuchic JN, An analytical study of the interplay between geometrical and energetic effects in protein folding. *J. Chem. Phys.*, 128(025101), 2008.

Vano JA, Wildenberg JC, Anderson MB, Noel JK, Sprott JC, Chaos in low-dimensional Lotka-Volterra models of competition. *Nonlinearity*, 19(2391-404), 2006.

Sprott JC, Vano JA, Wildenberg JC, Anderson MB, Noel JK, Coexistence and chaos in complex ecologies. *Phys. Lett. A* 335(207-212), 2005.

ABSTRACT OF THE DISSERTATION

Development and Application of All-Atom Structure-based Models for Studying the Role of Geometry in Biomolecular Folding and Function

by

Jeffrey Kenneth Noel

Doctor of Philosophy in Physics

University of California, San Diego, 2012

Professor José Onuchic, Chair

Protein dynamics takes place on a rugged funnel-like energy landscape that is biased towards the native state. In naturally occurring proteins, this ruggedness caused by non-native interactions is sufficiently smooth (minimally frustrated) that the landscape is dominated by the native interactions. This provides the theoretical foundation for a class of minimalist protein models called *structure-based models* (SBMs). In the first half of the thesis we develop and characterize an all-atom SBM that seeks to bridge the gulf between coarse-grained SBMs and all-atom empirical models. We report on the robustness of folding mechanisms in the all-atom model and show that the global folding mechanisms in a coarse-grained Cα model and the all-atom model largely agree, although differences can be attributed to geometric heterogeneity in the all-atom model.

We then take a careful look at an important aspect of the SBM, the definition of the native contact map, and propose a general algorithm for generating atomically-grained contact maps called "Shadow." We show that this choice of contact map is not only well behaved for protein folding, since it produces consistently cooperative folding behavior in SBMs, but also desirable for exploring the dynamics of macromolecular assemblies since it distributes energy similarly between RNAs and proteins despite their disparate internal packing. All-atom SBMs employing Shadow contact maps provide a general framework for exploring the geometrical features of biomolecules, especially the connections between folding and function. The second half of the thesis explores the intricacies encountered during folding by proteins at two extremes in structural complexity, complicated folds containing knots and simple folds like three-helix bundles. First we map the full free energy landscape of a knotted protein for the first time and show that a native-biased landscape is sufficient to fold complex topologies. We present a folding mechanism generalizable to all known knotted protein topologies: knotting via threading a native-like loop in a pre-ordered intermediate. Lastly, we discuss a simple three-helix bundle structure, whose structural symmetry opens up a "trap-door" to a competing mirror image structure. The simulations suggest that mirror images might not just be a computational annoyance but are competing folds that might switch depending on environmental conditions or functional considerations.

Chapter 1

Structure-based Models Capture the Geometric Aspects of Biomolecular Dynamics

1.1 Introduction

Structural biology techniques, such as nuclear magnetic resonance (NMR), x-ray crystallography, and cryogenic electron microscopy (cryo-EM), have provided extraordinary insights into the structural details of biomolecules. Recent advances in x-ray crystallography and cryo-EM have even allowed for structural characterization of large molecular machines such as the ribosome, proteasome and spliceosome. The structural data is complemented by experimental techniques capable of probing dynamic information, such as Förster resonance energy transfer (FRET) and stopped flow spectrometry. The ability to combine the low resolution dynamical data with the high-resolution structural data provides tremendous insights into the dynamics of biomolecular systems. Computer simulation of these biomolecular systems provides the necessary bridge that combines static structural data with dynamic experiments at atomic resolution.

Since the first molecular dynamics simulations of bovine pancreatic trypsin inhibitor 35 years ago (*1*), molecular simulations have become indispensable tools in biophysics. Molecular dynamics simulations of biomolecules treat the molecule as a

collection of classical particles interacting through a potential energy function called a forcefield (*2*). The molecule's dynamics are propagated through time by numerical integration of Hamilton's equations resulting in a molecular trajectory. This trajectory can be used to gain a kinetic and thermodynamic understanding of the system. Simulations can be performed using empirically parameterized forcefields that include explicit solvent. In principle, their chemistry-based representation should reproduce the structure and dynamics of a biomolecular system without requiring input from experimental structural data. In practice, making contact with experimental observables poses harsh challenges for these forcefields both due to the level of accuracy required and the long time scales needed (*3–5*). Complementing these detailed models are simplified biomolecular models able to access longer time and larger length scales. One class of simplified biomolecular Hamiltonians, which is theoretically underpinned by the energy landscape theory of protein folding (*6–9*), is called *structure-based models* (SBMs). These models impose a *native bias* by explicitly including structural data in the Hamiltonian. The structural data is derived from experimental techniques (generally x-ray crystallography or NMR) that are able to discern a representative structure of a molecule in a deep free energy basin, *e.g.* a protein native state. The native bias dramatically reduces the complexity of the resulting forcefield. These simplifications allow for a clear physical understanding of a system and open up biologically relevant timescales while retaining the essential dynamical features. SBMs have been validated by their application to protein dynamics, such as folding (*10–14*), oligomerization (*15–18*), and functional transitions (*19–24*). This chapter introduces SBMs and describes the implementations of a commonly used coarse-grained SBM and a new all-atom SBM. The development, characterization, and application of this all-atom SBM is the focus of the thesis.

1.2 Structure-based Models

1.2.1 Foundations in Energy Landscape Theory

The inclusion of a native-bias, the hallmark of a SBM, in protein folding models has a rigorous footing in the energy landscape theory of protein folding (*6–8*). Protein folding is a spontaneous, self-organizing process whereby a protein transitions from a

highly-disordered ensemble (unfolded) to a structured ensemble (folded/native state). The relatively short timescale with which the folded state is reached implies that any competing non-native states (traps) are shallow compared with the overall energy bias to folding. If these traps are sufficiently shallow, the non-native interactions can be grouped into an effective diffusion (*25–27*). In addition, the uniqueness of the folded state implies that it corresponds to the global minimum in the free-energy landscape. The *principle of minimal frustration* (*7*) states that evolution has achieved this folding robustness by selecting for sequences where the interactions present in the native structure are mutually supportive, *i.e.* attractive. The interactions are minimally frustrated or, in other words, maximally consistent. This organization leads to the protein folding on a *funneled landscape* where the energy on average decreases as it forms structures similar to the native structure.

Minimal frustration and the funneled energy landscape give the theoretical foundation for SBMs. A structure-based potential dramatically reduces the biomolecular Hamiltonian's complexity by stabilizing interactions that are spatially close in the native configuration. While real protein funnels have residual energetic frustration caused by non-native interactions, the SBMs discussed here are "perfectly funneled" models, since in the forcefield *all* interactions stabilize the native structure. Non-native interactions are short-range and strictly repulsive. In such a framework, any barriers to folding must be free energy barriers arising from the various ways energy and entropy compensate during folding. The ability of perfectly funneled models to reproduce experimental folding trends and mechanisms shows that geometrical effects like chain connectivity have an enormous influence on protein dynamics (*6, 28, 29*). Since the precise energetics are secondary to the geometry of the protein molecule, this idea leads to the commonly held notion that geometry determines the folding mechanism.

Even though SBMs were formulated in the context of protein folding, their applications are widespread. Folding is only a first step in the lives of proteins which go on to perform a myriad of functions in the cell. The funneled energy landscape upon which the protein folds is the same landscape that controls functional protein motions. Multiple functional conformational states captured by experiment can be naturally included by extending the funneled landscape to have multiple basins. Structured RNAs

must also have evolutionary pressure to reduce the level of frustration or they would encounter their own "Levinthal's paradox." [1] Compared to small proteins, the robust dynamics of large molecular complexes such as the ribosome and proteasome must depend even less on the precise atomic energetic details and more on emergent properties controlled by the geometry of their constituents. While all these systems will have residual levels of frustration, the use of SBMs as a baseline is crucial to partition the global properties, those largely dependent on structure, from the details dependent on specific energetics.

1.2.2 Structure-based Model as a Baseline

Simplified models have a long history of elucidating the organizing principles governing complex systems. A key question is how sensitive a model is to its underlying parameters. Determining the correct value for a parameter is often equally important as understanding the sensitivity to perturbations in that parameter. Since molecular geometry has a central influence on the motions leading to molecular function, simplified models based on low free energy structures are a natural starting point. The simplest models look at the normal modes of an energy landscape created by replacing all short range interactions in a native structure by Hookean springs (*31*). These models can capture relevant rigid body motions. SBM provide an important generalization by allowing the possibility for "cracking," (*19, 32–34*) allowing interactions to break and reform, since the springs are replaced by short range potentials. Thus SBM can capture motion on all scales from native basin dynamics to unfolding.

The straightforward formulation of a structure-based potential allows for sensitivity analysis of the forcefield parameters (*35*) and their simplicity makes them ex-

[1] In a standard illustration of the Levinthal paradox, each bond connecting amino acids can have several (e.g., three) possible states, so that a protein composed of 100 amino acids could exist in $3^{100} = 5 \times 10^{47}$ configurations. If the protein is able to sample new configurations at even the enormous rate of 10^{15} per second, or 3×10^{22} per year, it will take 10^{25} years to try them all. Because of this enormous search time, Levinthal concluded that random searches are not an effective way of finding the correct state of a folded protein. Nevertheless, proteins do fold, and in a time scale of seconds or less. This is the paradox. The paradox can be overcome with a biased search, where the native contacts are on average more favorable than non-native contacts (*7*). If temperature becomes too low, below the so called glass temperature T_G (*26*), the non-native interactions compete strongly enough that the search parallels a random search and the time scales diverge. The folding transition temperature T_F for proteins satisfies $T_F/T_G \gg 1$. A nice discussion of the Levinthal paradox can be found here (*30*).

tremely fast to compute. The forcefield is readily extensible allowing the introduction of complicated effects to be explored parametrically. For example, the effects of electrostatics can be explored by perturbative addition of Coulomb interactions (*36–38*), or the effects of solvent probed by the perturbative addition of desolvation barriers (*39*). A crucial question in the protein folding field has been how proteins manage to achieve such smooth energy landscapes, or equivalently, why do all-atom empirical forcefields and structure prediction schemes have difficulty achieving the level of specificity seen in proteins. Using structure-based potentials with all-atom geometries, we can begin to address this question. These models completely partition energetic effects from geometric effects, and through careful investigation, may discern to what extent energetics contribute to the apparent native specificity in protein structure, folding, and function. While processes like the formation of non-native intermediates during folding (*40–42*) and protein misfolding are clearly cases that perfectly funneled SBM will be unable to fully describe, through adding complexity in a piecemeal fashion to a robust baseline model, a more complete understanding of the interplay between geometry and energy in even these complicated systems will result.

1.3 Implementation of Structure-based Models

SBMs have a long history in the protein folding field. The folding dynamics of minimally frustrated sequences were first tested in lattice models. Early work by Gō and coworkers (*43*) led to "Gō models" as a synonymous name for SBMs. Bryngelson *et al.* (*9*) and Socci *et al.* (*44*) investigated a minimally frustrated lattice model with three types of beads. They found that the dynamics could be well described by diffusion along a small number of collective coordinates on an effective free energy surface defined by those coordinates. As the structural correspondence between cubic lattices and actual proteins is low, Nymeyer *et al.* implemented an off-lattice, coarse-grained model of a protein-like structure. They compared the folding dynamics of an energetically frustrated (*45*) versus a completely unfrustrated β-barrel (*11*). They showed that the completely unfrustrated model, effectively a SBM, exhibited the characteristics of a good folder, specifically, having exponential folding kinetics on a funnel shaped land-

scape that is robust to reasonable perturbations. Following these successes, Clementi *et al.* (*13*) introduced the popular "C_α model," which also had a coarse-grained representation of the protein. This model reproduced the transition state ensembles of several small two and three state proteins. The C_α model has since been adopted by several investigators to explore myriad topics in protein folding (see these references for some highlights (*14, 21, 22, 29, 39, 46–49*)). The off-lattice geometry allowed clear representation of protein structures, making comparisons to experimentally determined dynamics possible. In order to completely partition energetic and geometric effects, we introduced an all-atom SBM (*35*). This model is being used to represent proteins (*35*), RNA/DNA (*50*) and ligands in a consistent fashion for both dynamics (*4, 51, 52*) and molecular modeling (*53–55*). Two models, the all-atom model and the C_α model, are prominently featured in this thesis.

Before these two models are described in detail, we review the key components common to any SBM. The defining characteristic is that the parameters are determined from a native structure. The potential V is composed of three contributions,

$$V = \underbrace{V^{\text{Bonded}} + V^{\text{Repulsive}}}_{\text{Maintain geometry}} + \underbrace{V^{\text{Attractive}}}_{\text{Tertiary structure}} . \tag{1.1}$$

V^{Bonded} includes interactions that maintain the covalently bonded structure and torsional angles. This term often ensures correct chirality. $V^{\text{Repulsive}}$ contains repulsive terms that enforce excluded volume and prevent chain crossings. Collectively these two terms maintain the protein's structure and allowed conformational diversity. $V^{\text{Attractive}}$ contains short range, attractive interactions between atoms (or residues if coarse-graining) close in the native state. The definition of which interactions are given attractive potentials is called a *native contact map* (Figure 1.1).

1.3.1 Structure-based Potentials

There are two SBM potentials that are discussed in this thesis. The first is a popular coarse-grained model called the C_α model (*13*). The second is an all-atom SBM that was developed as part of this thesis called the all-atom model (*35*). They are both in wide use and are publicly available on the SMOG web server (*57*) that was also

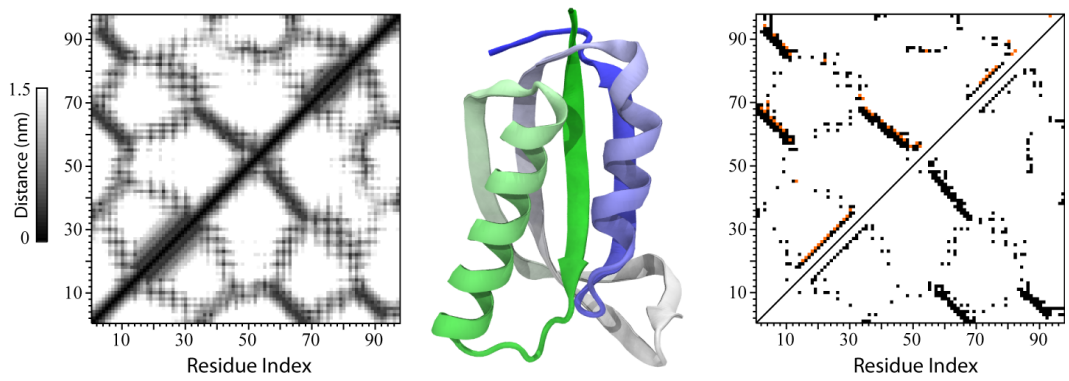

Figure 1.1: Tertiary interactions in the SBM are defined through the native contact map. Structure of the α/β ribosomal protein S6 (PDB code: 1RIS) is shown with the N-terminus (residue 1) colored green. Left panel shows the proximity of the nearest atomic contact for each residue pair up to a maximum of 1.5 nm. Right panel compares two coarse-grained native contact maps. A pair of residues are considered a native contact if they share a native atom-atom contact. Top triangle: 6 Å cutoff. Bottom triangle: a 6 Å cutoff with geometric occlusion using Shadow (*56*). The contacts which are excluded by Shadow are colored orange.

developed as part of this thesis.

C_α Model

The C_α model coarse-grains the protein as single bead of unit mass per residue located at the position of the α-carbon. $\vec{\mathbf{x}}_\mathbf{0}$ denotes the coordinates (usually obtained from the Protein Data Bank http://www.rcsb.org) of the native state and any subscript 0 signifies a value taken from the native state. The potential is given by

$$\begin{aligned} V_{C\alpha}(\vec{\mathbf{x}},\vec{\mathbf{x}}_\mathbf{0}) &= \sum_{\text{bonds}} \varepsilon_r(r-r_0)^2 + \sum_{\text{angles}} \varepsilon_\theta(\theta-\theta_0)^2 + \sum_{\text{backbone}} \varepsilon_D F_D(\phi-\phi_0) \\ &+ \sum_{\text{contacts}} \varepsilon_C C(r_{ij}, r_0^{ij}) + \sum_{\text{non-contacts}} \varepsilon_{NC}\left(\frac{\sigma_{NC}}{r_{ij}}\right)^{12} \end{aligned} \tag{1.2}$$

where the dihedral potential F_D is,

$$F_D(\phi) = [1-\cos(\phi)] + \frac{1}{2}[1-\cos(3\phi)]. \tag{1.3}$$

The coordinates $\vec{\mathbf{x}}$ describe a configuration of the α-carbons, with the bond lengths to nearest neighbors r, three body angles θ, four body dihedrals ϕ, and distance between

atoms i and j given by r_{ij}. C denotes the contact potentials given to the native contacts (see "Contact Potential" below). Protein contacts that are separated by less than 3 residues are neglected. Excluded volume is maintained by a hard wall interaction giving the residues an apparent radius of $\sigma_{\mathrm{NC}} = 4$ Å. The native bias is provided by using the parameters from the native state $\vec{\mathbf{x}}_{\mathbf{0}}$. Setting the energy scale $\varepsilon \equiv k_{\mathrm{B}}$, the coefficients are given the homogeneous values: $\varepsilon_r = 100\varepsilon$, $\varepsilon_\theta = 40\varepsilon$, $\varepsilon_{\mathrm{D}} = \varepsilon_{\mathrm{C}} = \varepsilon_{\mathrm{NC}} = \varepsilon$.

All-atom Model

The all-atom potential is similar to the C_α potential, though representing the all-atom geometry requires some additional terms. In the all-atom model, all heavy (non-hydrogen) atoms are explicitly represented as beads of unit mass, so each interaction is now between atoms as opposed to residues. Bonds, angles, and dihedrals therefore have their traditional chemical meanings. In each residue there two free backbone dihedrals (except Proline) and possibly several side chain dihedrals. Improper dihedrals maintain backbone chirality and, when necessary, side-chain planarity. As in the C_α model, $\vec{\mathbf{x}}_{\mathbf{0}}$ denotes the coordinates of the native state and any subscript 0 signifies a value taken from the native state. The all-atom potential is

$$\begin{aligned} V_{\mathrm{AA}}(\vec{\mathbf{x}}, \vec{\mathbf{x}}_{\mathbf{0}}) = & \sum_{\text{bonds}} \varepsilon_{\mathrm{b}}(r - r_0)^2 + \sum_{\text{angles}} \varepsilon_\theta(\theta - \theta_0)^2 + \sum_{\text{impropers/planar}} \varepsilon_\chi(\chi - \chi_0)^2 \\ & + \sum_{\text{backbone}} \varepsilon_{\mathrm{BB}} F_{\mathrm{D}}(\phi) + \sum_{\text{side chains}} \varepsilon_{\mathrm{SC}} F_{\mathrm{D}}(\phi) \\ & + \sum_{\text{contact map}} \varepsilon_{\mathrm{C}}^{ij} C(r_{ij}, r_0^{ij}) + \sum_{\text{non-contacts}} \varepsilon_{\mathrm{NC}} \left(\frac{r_{\mathrm{NC}}}{r_{ij}} \right)^{12} \end{aligned} \tag{1.4}$$

where F_{D} is given in Equation 1.3. C is the contact potential, an effective short range interaction between atoms i and j that are in contact in the native state (see "Contact Potential" below). The definition of the native contacts is considered in detail in Section 3.2.3. Three criteria define the values of $\varepsilon_{\mathrm{BB}}$, $\varepsilon_{\mathrm{SC}}$, and ε_{C} for a given molecular complex. 1) $\varepsilon_{\mathrm{BB}}$ and $\varepsilon_{\mathrm{SC}}$ are scaled so that $\frac{\varepsilon_{\mathrm{BB}}}{\varepsilon_{\mathrm{SC}}} = R_{\mathrm{BB/SC}}$. 2) The energetic weight of each dihedral and contact is also scaled, such that the ratio of total contact energy to total dihedral energy $\frac{\sum \varepsilon_C}{\sum \varepsilon_{\mathrm{BB}} + \sum \varepsilon_{\mathrm{SC}}} = R_{C/D}$, is satisfied. 3) The total stabilizing energy is set, such that $\sum \varepsilon_C + \sum \varepsilon_{\mathrm{BB}} + \sum \varepsilon_{\mathrm{SC}} = \varepsilon N_{\mathrm{atoms}}$, where ε is the reduced energy unit. This allows a

consistent comparison across parameter sets and gives a folding temperature $k_B T_F \sim 1$. The reduced energy unit ε is defined by the relations $\varepsilon = k_B T^*$ and $k_B T^* = 1$. Here, $R_{BB/SC} = 2$ for protein and $R_{BB/SC} = 1$ for RNA, and $R_{C/D} = 2$. r_0^{ij} is the native distance separation between atoms i and j. $\varepsilon_b = 100\varepsilon$, $\varepsilon_\theta = 20\varepsilon$, $\varepsilon_\chi = 10\varepsilon$, and $\varepsilon_{NC} = \varepsilon$. When improper dihedrals are maintaining the planarity of rings $\varepsilon_\chi = 40\varepsilon$. r_0, θ_0, χ_0, ϕ_0 and r_0^{ij} are given the values found in the native state and $r_{NC} = 1.7$ Å.

A technical issue is normalizing the dihedral energy around each bond. When assigning dihedral strengths, we first group dihedral angles that share the middle two atoms. For example, in a protein backbone, one can define up to four dihedral angles that possess the same $C-C_\alpha$ covalent bond as the central bond. Each dihedral in the group is scaled by $1/N_D$, where N_D is the number of dihedral angles in the group. This normalization is separately maintained for proper and improper dihedrals.

Two ratios determine the distribution of dihedral and contact energies, contact to dihedral ratio $R_{C/D}$ and backbone to side chain ratio $R_{BB/SC}$. In proteins $R_{BB/SC} = \varepsilon_{BB}/\varepsilon_{SC} = 2$ (*35*) and in RNA/DNA $R_{BB/SC} = \varepsilon_{BB}/\varepsilon_{SC} = 1$ (*50*). The contacts and dihedrals are scaled relative to their total contributions, $R_{C/D} = \frac{\sum \varepsilon_C}{\sum \varepsilon_{BB} + \sum \varepsilon_{SC}} = 2$. Lastly, the total contact and dihedral energy is set equal to the number of atoms $N_{atoms} = \sum \varepsilon_C + \sum \varepsilon_{BB} + \sum \varepsilon_{SC}$. This choice gives folding temperatures near 1 in reduced units ensuring a consistent parameterization.

In the all-atom model every term is based on the native structure except the non-contact excluded volume term. In the C_α model all the residues have a homogeneous shape, but in the all-atom model each residue has its unique geometry explicitly represented. This gives the all-atom model structure independent sequence information that adds heterogeneity to the model. This heterogeneity adds geometric frustration to the model, since interactions can only be satisfied if the side chains are correctly oriented (*51*). A question of current interest is whether this sequence dependent information adds constraints to the folding dynamics, allowing the native bias to be relaxed (*35, 58, 59*).

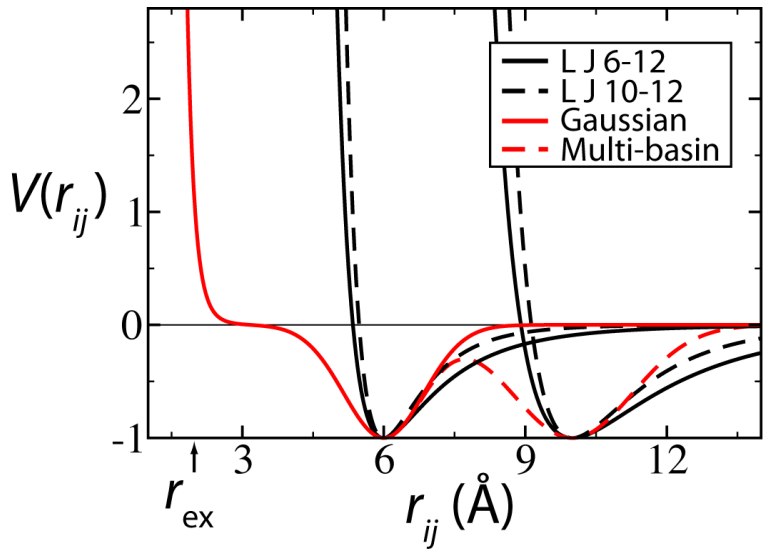

Figure 1.2: Comparison of Lennard-Jones and Gaussian contact potentials. Black curves show LJ contact potentials with minima at 6 Å and 10 Å. The Gaussian contact potential shown in green has an excluded volume $r_{\rm ex}$ that can be set independently of the location of the minimum. The dotted green line shows how the Gaussian contact would change as another minimum at 10 Å is added.

Contact Potential

All of the pair interactions defined in the native contact map interact through a short range, attractive potential, denoted in the SBM potential by $C(r_{ij}, r_0^{ij})$ (Figure 1.2). The contact potential has a minimum at r_0^{ij}, the distance between the pair in the native state. Traditionally, a contact is defined through a Lennard-Jones (LJ) type potential, since the LJ shape is readily available in molecular dynamics packages. In the C_α model a "10-12" LJ potential is used for contacts with the minimum set at the separation between the C_α pair in the native state r_0^{ij},

$$C_{\rm CA}(r_{ij}, r_0^{ij}) = 5\left(\frac{r_0^{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{r_0^{ij}}{r_{ij}}\right)^{10}, \tag{1.5}$$

and in the all-atom model a "6-12" LJ potential with the minimum set at the separation between a contacting atomic pair in the native state,

$$C_{\rm AA}(r_{ij}, r_0^{ij}) = \left(\frac{r_0^{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{r_0^{ij}}{r_{ij}}\right)^{6}. \tag{1.6}$$

Different LJ potentials are used because the native contact distances r_0^{ij} can be much longer in the C_α model. The contacts are coarse-grained to be between the C_α atoms, which can be as distant as 14 Å. The r^{-6} is much broader than the r^{-10} and can lead to unphysical structures in unfolded states as native pairs interact at long distances.

The LJ potentials are well tested and work for many systems, but they have limitations for certain applications because the LJ potential has an excluded volume that moves with the minimum. The effective size of two atoms in contact grows with r_0^{ij}. This additional excluded volume has little effect on the entropy of unfolded conformations since mostly non-contacts are interacting, but has a large effect on the entropy of the folded ensemble where most contacts are formed. In cases where the user wants to control the excluded volume term (*51, 60*), an attractive Gaussian well coupled with a a fixed LJ repulsion,

$$C_{\mathrm{G}}(r_{ij}, r_0^{ij}) = \left(1 + \left(\frac{r_{\mathrm{ex}}}{r_{ij}}\right)^{12}\right)\left(1 + G(r_{ij}, r_0^{ij})\right) - 1 \tag{1.7}$$

where

$$G(r_{ij}, r_0^{ij}) = -\exp\left[-(r_{ij} - r_0^{ij})^2/(2\sigma_{ij}^2)\right]. \tag{1.8}$$

This functional form ensures that the depth of the minimum is -1 (scaled by ε_{C} in Equation 1.4), and r_{ex} sets the excluded volume. r_{ex} has the same function as r_{NC} in Eq. 1.4. If $r_{\mathrm{ex}} = r_{\mathrm{NC}}$, all atomic interactions have an equal excluded volume. For consistency with the LJ potentials, the width of the Gaussian well σ_{ij} models the variable width of the LJ potential. $C_{\mathrm{AA}}(1.2r_0^{ij}, r_0^{ij}) \sim -1/2$ so σ_{ij} is defined such that $G(1.2r_0^{ij}, r_0^{ij}) = -1/2$ giving $\sigma_{ij}^2 = (r_0^{ij})^2/(50\ln 2)$. If r_{ex} is significantly smaller than r_0^{ij} Eq. 3.2 reduces to a more transparent form,

$$C_{\mathrm{G}}(r_{ij}, r_0^{ij}) \to \left(\frac{r_{\mathrm{ex}}}{r_{ij}}\right)^{12} + G(r_{ij}, r_0^{ij}) \quad \text{for } r_{\mathrm{ex}}, \sigma_{ij} \ll r_0^{ij}. \tag{1.9}$$

The flexibility of the Gaussian approach also allows for multiple basin contact potentials for energy landscapes with multiple minima (e.g. Section 6.2.3). Using multiple LJ potentials with different locations of the minima is not a viable option because the longest LJ potential would occlude the others with its excluded volume term. A multi-basin Gaussian potential C_{MB} for minima taken from two structures $\vec{\mathbf{x}}_\alpha$ and $\vec{\mathbf{x}}_\beta$ is given

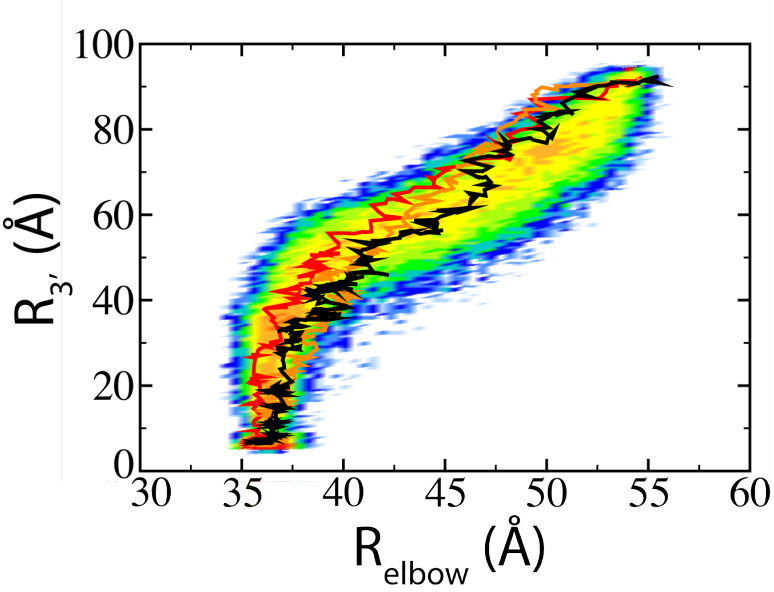

Figure 1.3: Comparison of AA SBM and explicit solvent simulations of tRNA accommodation in the ribosome. Trajectories of three 4 nanosecond explicit solvent targeted molecular dynamics (TMD) overlay the probability distribution of 704 microsecond structure-based TMD runs. With such a short sampling time, the explicit solvent TMD is dominated by steric interactions between the ribosome and the tRNA. The SBM naturally captures the sterics and is consistent with the detailed model. $R_{3'}$ and R_{elbow} monitor the position of the tRNA along the accommodation pathway. Data detailed in (*4*).

by (*60*),

$$C_{\mathrm{MB}}(r_{ij}, r_{\alpha}^{ij}, r_{\beta}^{ij}) = \left(1 + \left(\frac{r_{\mathrm{ex}}}{r_{ij}}\right)^{12}\right)\left(1 + G(r_{ij}, r_{\alpha}^{ij})\right)\left(1 + G(r_{ij}, r_{\beta}^{ij})\right) - 1 \qquad (1.10)$$

Analogous to Eq. 1.7, this construction fixes the depth of both minima at -1. It should be noted that the folding temperature (defined in Section 1.3.3) is typically 0.2 - 0.3 reduced units higher for the Gaussian potential as compared to LJ because the extra excluded volume in the LJ potential destabilizes the native state.

1.3.2 Choosing a Graining: C_α or All-atom

The C_α and all-atom model are both able to describe the properties of the molecular scaffold's geometry. When comparing the two models, C_α and all-atom, the main

advantage of C_α is its speed. Because the all-atom model has roughly eight times more atoms and has slower diffusion due to side chain interactions, the C_α model runs significantly faster than all-atom. This speed is important for studying processes with large barriers, like folding and oligomerization. All-atom can narrow the speed gap with parallelization, but not close it completely. Nonetheless, all-atom has been used to fold small single domain proteins (*35*) and even proteins with complex topologies (*51*). Many processes without large activation barriers, *e.g.* native basin dynamics, have energy landscapes that are easily sampled, and thus the performance hit of all-atom is of no consequence.

The explicit representation of atomic coordinates is advantageous, even for simplified models like SBM. A clear benefit is acting as a bridge between minimalist models and empirical forcefields. Any conformations realized during a simulation of an all-atom SBM can be compared with, and used as input for, empirical forcefields with an explicit solvent. Since the sterics are correct, any process that is dominated by large-scale structural fluctuations should be well represented by an all-atom SBM (*4, 52*). Fig. 1.3 shows targeted molecular dynamics simulations of the tRNA accommodation process in the ribosome, a massive ribonucleoprotein molecular machine ($\sim$ 2.4 MDa). The trajectories from explicit solvent simulations overlay all-atom SBM trajectories. On a smaller scale, the all-atom geometry opens the door to studying side chain degrees of freedom during folding and binding simulations. Constricted conformations like polypeptide slipknots, found in coarse-grained models, are shown to be sterically possible with the all-atom geometry (*51*). Lastly, the all-atom geometry allows a clear way to add perturbative non-native chemical effects like hydrogen bonding (*58*) and partial charges.

1.3.3 Molecular Dynamics with SBM

Molecular dynamics uses Newtonian mechanics to evolve the motions of atoms in time. The interactions defined in the SBM potential define the various forces on the atoms since force is given by the negative gradient of the potential energy. The system is evolved through time in discrete steps. The NVT canonical ensemble is maintained using a thermostat. Thermostats including a drag term, such as stochastic dynamics or Langevin dynamics are used for implicit solvent systems like SBMs. Velocity rescaling

thermostats can introduce heating artifacts when not coupled to an explicit solvent (*61*). Langevin dynamics has been used to model the viscosity of a solvent (*33,62*). The output of a molecular dynamics simulation is a trajectory, a time ordered series of snapshots of the atomic coordinates. The trajectory can be analyzed as a function of time to uncover kinetic properties or, by application of the ergodic theorem, as an ensemble to compute thermodynamic properties.

A molecular dynamics trajectory contains the coordinates of all the atoms in the system, a massive amount of information. Therefore the trajectory is reduced down to one or a few reaction coordinates that monitor the progress of the dynamics under investigation. For protein folding, a useful reaction coordinate would differentiate between the unfolded ensemble, folding intermediates and the folded ensemble. A reaction coordinate for studying a conformational transition would differentiate the various conformers. A natural reaction coordinate for SBMs is Q, the fraction of native contacts formed (*63*). A common definition for a formed contact between the native pair ij is whether it satisfies $r_{ij} < \gamma r_0^{ij}$. Here, $\gamma = 1.2$. In protein folding, low Q would correspond to the unfolded ensemble, medium Q would contain the transition state ensemble and any intermediates, and high Q the folded ensemble. To investigate a conformational transition between two structures A and B, monitoring switching between high Q_{A} and high Q_{B} would indicate transitions. Other possible reaction coordinates are root mean square deviation from a reference structure or radius of gyration. An exciting possibility is to monitor the position of an explicitly represented FRET probe in order to compare with experimental data (*4*).

After the choice of reaction coordinate is made, the value of the coordinate during the trajectory (or several concatenated trajectories) can be histogrammed to obtain a potential of mean force (PMF) along the reaction coordinate. If the chosen coordinate adequately separates two basins, it can be used to identify the transition state at the peak on the free energy landscape. Q has been shown to be a suitable coordinate for protein transitions and thus the peaks in $F(Q)$ can be identified as transition state ensembles (TSE) (*63*) (see Figure 1.4). Great care must be exercised when making quantitative predictions of thermodynamic and kinetic quantities from simplified models. The kinetics of the system are not simply determined by the free energy landscape, but are highly

dependent on diffusion rates. Diffusion rates vary for different molecular systems and must be calibrated separately. For discussion of diffusion in SBM see (*4, 64, 65*). Secondly, the precise values of free energy barriers and thermal stability are a fine balance and depend on the details of the SBM potential. This said, given a constant parameterization, kinetic and thermodynamic quantities tend to scale in a consistent fashion. Fast folding proteins will consistently have smaller free energy barriers than slow folding proteins (*29, 35*). Some quantities are robust to perturbations, in particular the TSE and other so called geometrical features of the energy landscape (*35, 60*).

Protein Folding Example

The most established application of SBMs is to the study of protein folding. Determining the TSE, the shape and size of free energy barriers, and the existence of folding intermediates are all topics of interest. Figure 1.4 shows the result of all-atom SBM folding simulations for two of the most throughly studied proteins, chymotrypsin inhibitor-2 (CI2) and the SH3 domain. These two proteins are two-state folders, meaning the protein only populates two basins spanned by a cooperative transition.

Figure 1.4a,d shows representative traces of Q versus time during constant temperature molecular dynamics near folding temperature T_{F}. T_{F} is the temperature such that the folding and unfolding basins are equally populated. Simulations are performed at T_{F} because it maximizes the sampling rate of the folding transition. T_{F} is determined by running simulations at high and low temperatures, and iteratively converging on a temperature where both folding and unfolding is observed. Q is defined as the fraction of native residue pairs with at least one atom-atom contact within 1.2 times its native separation. Alternative definition of Q, such as the fraction of atom-atom contacts formed, may shift the locations of basins in the resulting free energy landscape, but will preserve the heights of any barriers.

Q traces from long molecular dynamics trajectories at various temperatures can be combined using Weighted Histogram Analysis (WHAM) (*66*), to obtain an optimal density of states. The density of states can then be used to extrapolate $F(Q)$ at any temperature (Figure 1.4b,e). Always, care must be taken to ensure that the trajectories reflect equilibrium. One easy method is to chop all trajectories in half and verify that

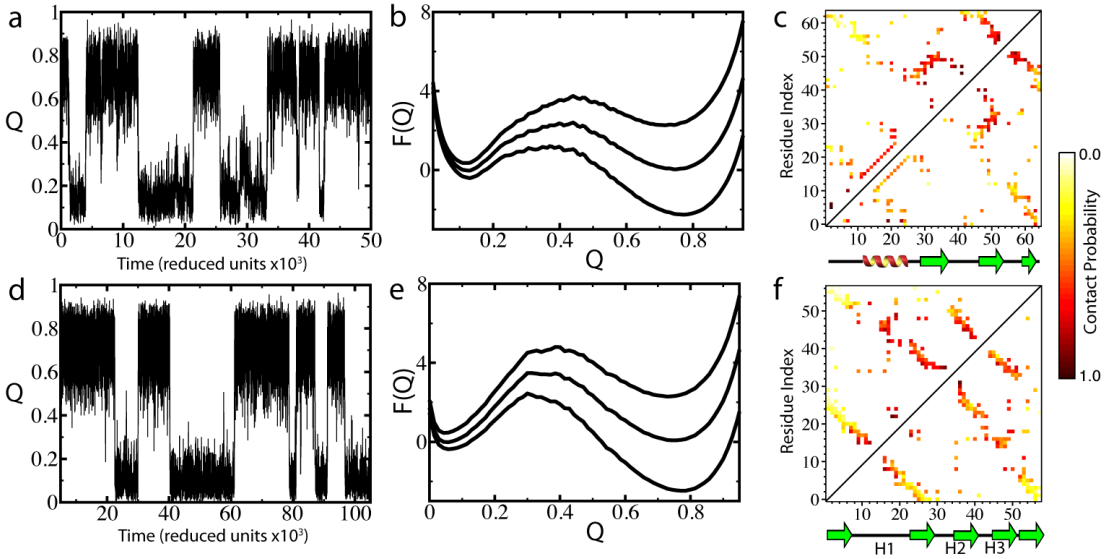

Figure 1.4: All-atom structure-based simulations of folding for the two-state proteins CI2 (**top**) and SH3 domain (**bottom**). PDB codes: 1FMK, 1YPA. **a,d**: The reaction coordinate Q plotted as a function of time for a typical simulation near T_{F}. Both proteins exhibit transitions between a folded ensemble at $Q \sim 0.8$ and an unfolded ensemble at $Q \sim 0.1$. **b,e**: Free energy $F(Q)$ for temperatures $0.98T_{\mathrm{F}}$, T_{F}, and $1.02T_{\mathrm{F}}$ calculated by weighted histogram analysis of long constant temperature MD trajectories. A set of "long" trajectories typically contain thirty folded to unfolded transitions. **c,f**: Transition state ensemble (TSE) for the two proteins. Contact formation probabilities are calculated by an unweighted average of all configurations $0.40 < Q < 0.45$. The upper triangle shows results from the C_α model and the lower triangle shows the AA model. Secondary structure is denoted below the contact maps as are the positions of the three hairpin turns in SH3. CI2 has a diffuse TSE that resembles the native state. The contact probability is roughly predicted by sequence separation. SH3 has a more polarized TSE with contacts from the first ten residues largely absent. The simulations were prepared using SMOG v1.0.6 (*57*) with default parameters.

$F(Q)$ and the TSE are the same for both halves. The TSE is the ensemble of structures that compose the bottleneck to folding. CI2 and SH3 each have a single TSE that connects the unfolded state to the folded state defined by the structure populating the top of $F(Q)$. Figure 1.4c,f shows the average contact maps of the structures with $0.4 < Q < 0.45$. The contact formation probabilities can be connected to Φ-value analysis, an experimental technique that estimates the contribution of a particular residue's contacts to the TSE (*67*). In simulation, Φ_i is given by

$$\Phi_i = \frac{P_i^{\mathrm{TSE}} - P_i^{\mathrm{U}}}{P_i^{\mathrm{F}} - P_i^{\mathrm{U}}} \tag{1.11}$$

where P_i is the probability that residue i forms its contacts and U/F refers to the unfolded/folded ensembles (*68*). Φ_i near 1 means that residue i is very native-like in the TSE and a Φ_i near 0 means that residue i in still unfolded in the TSE.

Since the TSE is a simple average over structures, it can hold hidden complexity. For some proteins with structural symmetry, the TSE is composed of multiple routes through the TSE (*14, 69*). Consider SH3; its TSE could be composed of two routes, a major route where hairpin 2 and hairpin 3 form first and a minor route where hairpin 1 and hairpin 2 form first (Figure 1.4f). Multiple routes can be identified by clustering the contact maps of TSE structures using the number of shared contacts as a similarity measure (*69*). These routes represent entropically viable routes through the TSE. Thus, two real proteins that fold to the same structure may follow seemingly very different paths due to minor energetic differences.

1.4 Outline of the Thesis

The majority of my work has focused on the development and application of all-atom SBMs, and this work is the focus of the thesis. This chapter has already introduced the all-atom SBM Hamiltonian that we developed. Chapters 2, 3, and 4 describe the development and characterization of all-atom SBMs. The last two chapters discuss the application of SBMs for exploring how protein folding is affected by the geometry (and topology) of the protein native state. Chapters 5 and 6 consider proteins at opposite ends of structural complexity. Chapter 5 discusses a complicated structure by exploring

the folding of a knotted protein, while Chapter 6 discusses a simple three-helix bundle structure, whose simplicity opens up a "trap-door" to a competing mirror image structure.

Chapter 2: An All-atom Structure-Based Potential for Proteins: Bridging Minimal Models with All-atom Empirical Forcefields

Coarse-grained SBMs utilize the funneled energy landscape theory of protein folding to provide an understanding of both long time and long length scale dynamics. All-atom empirical forcefields with explicit solvent can elucidate our understanding of short time dynamics with high energetic and structural resolution. Thus, SBMs with atomic details included can be used to bridge our understanding between these two approaches. The robustness of folding mechanisms in one such all-atom model is reported. Results are shown for three two-state globular proteins the B domain of Protein A, the SH3 domain of C-Src Kinase and Chymotrypsin Inhibitor 2. The interplay between side chain packing and backbone folding is explored. We also compare this model to a C_α SBM and an all-atom empirical forcefield. Key findings include 1) backbone collapse is accompanied by partial side chain packing in a cooperative transition and residual side chain packing occurs gradually with decreasing temperature 2) folding mechanisms are robust to variations of the energetic parameters 3) protein folding free energy barriers can be manipulated through parametric modifications 4) the global folding mechanisms in a C_α model and the all-atom model agree, although differences can be attributed to energetic heterogeneity in the all-atom model 5) proline residues have significant effects on folding mechanisms, independent of isomerization effects. Since this SBM has atomic resolution, this work lays the foundation for future studies to probe the contributions of specific energetic factors on protein folding and function.

Chapter 3: The Shadow Map: A General Contact Definition for Capturing the Dynamics of Biomolecular Folding and Function

An important aspect of our SBM Hamiltonian, which has not been explored previously, is the definition of native interactions. The set of native interactions is called a *contact map* and is a ubiquitous tool in the analysis of internal biomolecular inter-

action networks. This chapter presents a general algorithm for generating atomically-grained contact maps called "Shadow." The Shadow algorithm initially considers all atoms within a cutoff distance and then, controlled by a screening parameter, discards the occluded contacts. We show that this choice of contact map is not only well behaved for protein folding, since it produces consistently cooperative folding behavior in SBMs, but also desirable for exploring the dynamics of macromolecular assemblies since it distributes energy similarly between RNAs and proteins despite their disparate internal packing. All-atom SBMs employing Shadow contact maps provide a general framework for exploring the geometrical features of biomolecules, especially the connections between folding and function.

Chapter 4: SMOG@ctbp: Simplified deployment of structure-based models in Gromacs

Molecular dynamics simulations have benefited from years of research on computer algorithms best able to balance speed and efficiency. Molecular dynamics suites like Gromacs, NAMD, and Desmond, package all the necessary algorithms to run stable molecular dynamics and the ability to scale the calculations to thousands of processors. These packages have made homegrown molecular dynamics codes built to run SBM obsolete. SMOG, Structure-based MOdels in Gromacs, is a publicly available web server located at http://smog.ucsd.edu. Any PDB structure consisting of standard amino acids, RNA, DNA and common ligands, can be uploaded to SMOG, which outputs the necessary coordinate, topology and parameter files to run SBM in Gromacs. This provides the flexibility necessary to implement efficient and highly scalable SBM. SMOG in conjunction with Gromacs version 4.5 scales easily to 128 processors when simulating a ribosome, $\sim 150,000$ atoms. Protein folding simulations of much smaller systems scale to ~ 100 atoms per core on a single motherboard.

Chapter 5: The Full Folding Landscape of a Trefoil-Knot Protein: Slipknotting upon Native-Like Loop Formation

Protein knots and slipknots, mostly regarded as intriguing oddities, are gradually being recognized as significant structural motifs. Recent experimental results show that

knotting, starting from a fully extended polypeptide, can be achieved without chaperones. Understanding the nucleation process of folding knots is thus a natural challenge for both experimental and theoretical investigation. In this study, we employ energy landscape theory and molecular dynamics to elucidate the entire folding mechanism. The full free energy landscape of a knotted protein is mapped for the first time using the all-atom structure-based protein model. Results show that, due to the topological constraint, the protein folds through a three state mechanism that contains (i) a precise nucleation site which creates a correctly twisted native loop (first barrier) and (ii) a rate-limiting free energy barrier that is traversed by two parallel knot forming routes. The main route corresponds to a slipknot conformation, a collapsed configuration where the C-terminal helix adopts a hairpin-like configuration while threading, and the minor route to an entropically limited plug motion, where the extended terminus is threaded as through a needle. Knot formation is a late transition state process and results show that random (non-specific) knots are a very rare and unstable set of configurations both at and below folding temperature. Our study shows that a native-biased landscape is sufficient to fold complex topologies and presents a folding mechanism generalizable to all known knotted protein topologies: knotting via threading a native-like loop in a pre-ordered intermediate.

Chapter 6: Mirror Images as Naturally Competing Conformations in Protein Folding

During folding, a protein typically adopts a singular native state. Evolution has selected the protein's sequence to be consistent with the native state geometry, as this configuration must be both thermodynamically stable and kinetically accessible to prevent misfolding and loss of function. For simple protein geometries, such as coiled-coil helical bundles, symmetry introduces a competing, globally-different, near mirror image with identical secondary structure and similar native contact interactions. Experimental techniques like circular dichroism, which rely on probing secondary structure content, cannot readily distinguish these folds. Here, we want to clarify whether the native fold and mirror image are energetically competitive by investigating the free energy landscape of the three proteins *Staphyloccocus aureus* B/E domain of protein A,

and the designed ultra fast-folder α_3d. To prevent a bias from a specific computational approach, the present study employs the structure prediction forcefield PFF01, explicit solvent replica exchange molecular dynamics (REMD) with Amber94 forcefield, and structure-based simulations. We observe that the native fold and its mirror image have similar enthalpic stability and are thermodynamically competitive. REMD predicts the native basin is only 1-2 k_BT more stable than the mirror basin. There is evidence that the mirror fold has faster folding kinetics and could function as a kinetic trap. All together, our simulations suggest that mirror images might not just be a computational annoyance but are competing folds which might switch depending on environmental conditions or functional considerations. Helix swapping may be as common as domain swapping.

1.5 Acknowlegements

Chapter 1, in part, appears in a book chapter "The Many Faces of Structure-Based Potentials: From Protein Folding Landscapes to Structural Characterization of Complex Biomolecules," *Computational Modeling of Biological Systems*, (2012), Noel and Onuchic. The dissertation author was the primary investigator and author of the chapter. Regarding the book chapter, JKN would like to thank Joanna Sułkowska for many helpful discussions and Paul Whitford and Ryan Hayes for a careful reading of the chapter. The work was supported by the Center for Theoretical Biological Physics sponsored by the National Science Foundation (NSF) (Grant PHY-0822283) and NSF Grant NSF-MCB-1051438.

Chapter 2

An All-atom Structure-Based Potential for Proteins: Bridging Minimal Models with All-atom Empirical Forcefields

2.1 Introduction

In recent years the energy landscape theory of protein folding (*6–9, 43*) has been validated through its application to protein folding (*10–14*), oligomerization (*15–18*), functional transitions (*19–24*) and structure prediction (*70, 71*). The theory states that proteins are minimally frustrated, that their energy landscape is funnel shaped and that the folded state of the protein is at the bottom of the funnel. Because of the shape of the landscape there is a strong energetic bias towards the folded state of the protein with relatively infrequent trapping caused by non-native interactions. The resulting heterogeneity observed during folding is due to the geometric constraints of the native structure. Thus, models of proteins that have only the native structure encoded have had great success in determining folding mechanisms. Until recently, most models tended to be coarse-grained, which are very useful in understanding global folding dynamics. In commonly used structure-based (Gō) potentials (*13*), each residue is represented by a bead centered at the location of the C_α atom (Figure 2.1b) and only native interactions are stabilizing.

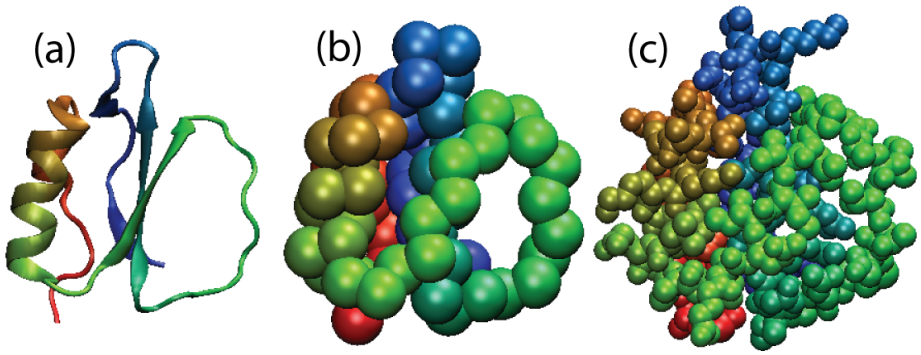

Figure 2.1: CI2 (Protein Data Base Entry 1YPA (*72*)) shown in (a) cartoon representation, (b) C_α representation and (c) all-atom (AA) representation. Structures are colored Red (N-terminus) to Blue (C-terminus). The size of the atoms in the C_α and AA representations correspond to the excluded volume radii used in the C_α (*13*) and AA models studied in this paper. Structures visualized using VMD (*73*).

On the other end of the spectrum of structural and energetic details are the computationally intensive all-atom empirical forcefields (*2, 74–78*). These forcefields include an atomistic representation of a protein either with an implicit or an explicit solvent. In these potentials, the parameters which determine the interaction between atoms, such as partial charges and van der Waals radii, are fit to experimental measurements and quantum mechanical calculations. With accurate calibration, a single parameter set may be applied to any protein and with sufficient computing resources, the dynamics of a protein can be calculated on a computer. The physics-based representation of atom-atom interactions automatically includes electrostatic interactions as well as any non-native interactions that may be present. In principle, these models render knowledge of a native structure unnecessary. A major limitation of these potentials is that they are often too expensive to fold all but small proteins (*79–87*). The timescales that can currently be calculated vary from hundreds of nanoseconds to tens of microseconds, depending on the size of the protein. A notable exception are the millisecond time scales recently achieved using special hardware (*5*). Biological timescales are usually several orders of magnitude larger and these dynamics cannot be accessed using all-atom empirical forcefields. In addition, sensitivity analysis of the dynamics to the parameters is not possible with these all-atom empirical forcefields.

In all-atom empirical forcefields an observed specificity of (i.e. preference for)

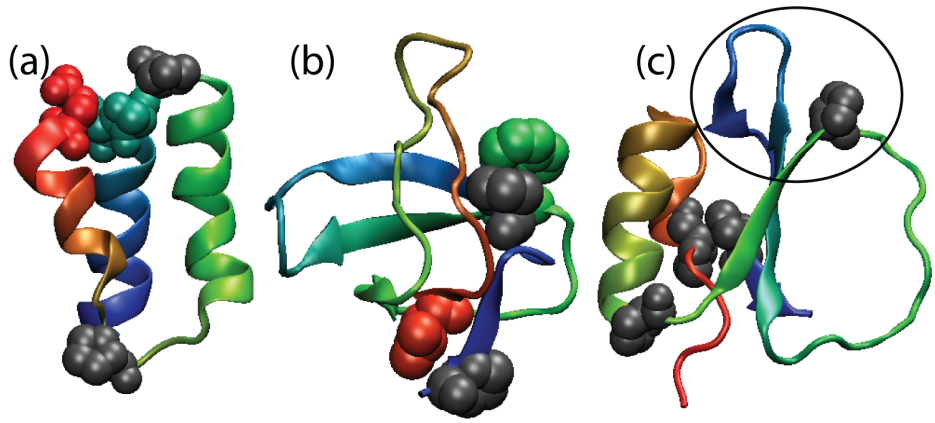

Figure 2.2: Structures of (a) Protein A, (b) SH3 and (c) CI2 (PDB entries 1BDD (*91*), 1FMK (*92*) and 1YPA (*72*)) colored Red (C-terminus) to Blue (N-terminus). These three proteins represent differing structural content and topological complexity. Protein A is a three-helix bundle, SH3 is composed of multiple β strands and in CI2 an alpha helix flanks a β sheet. Proline residues are shown as grey spheres. In Protein A, Gln1 and Ser31 are shown as colored spheres. In SH3, Val4 and Trp35 are shown as spheres. The mini-core of CI2 is circled.

native interactions is seen as a consequence of many energetic contributions. Due to the complex formulation of these potentials, it is impossible to partition geometric effects from energetic ones. There is a similar restriction in coarse-grained models due to their simplicity. Partitioning these effects is often impossible since geometry is included implicitly through energetic interactions. By studying all-atom models with structure-based potentials (*88–90*), since atomic geometry is explicitly included, we can ask to what extent energetics contribute to the apparent native specificity in protein structure, folding and function. In contrast to enzyme catalysis where specific atomic interactions directly control the chemical reactions, in most cases the energetic specificity required in protein folding is less stringent.

Providing a complete picture of specificity in protein folding and function will require the study of many proteins and many parametric variations. In this manuscript, we lay the foundation for this line of investigation through systematic characterization of a completely specific (only, and all, native interactions are stabilizing) AA structure-based model. We study the effect of varying the parameters of the model on folding barriers, mechanisms, contact formation and side chain dynamics. The test proteins, B

domain of Protein A, SH3 domain of C-Src Kinase and Chymotrypsin Inhibitor 2 (CI2) (Figure 2.2) have been experimentally (*93–95*) and computationally (*13, 96–98*) well characterized. Additionally, they possess two-state folding dynamics and represent different secondary and tertiary structures. The present model is energetically unfrustrated, with an explicit representation of all non-hydrogen atoms and homogeneous interaction strengths. We find that folding in the model is robust to parameter changes and dynamics agrees well with both the C_α model and an all-atom empirical forcefield with explicit solvent. Further, side chain ordering can be probed explicitly and the effect of prolines can be calculated. This study and model will serve as a basis for future AA models which incorporate non-specific contributions of energetic frustration, electrostatics and hydration.

2.2 Results

2.2.1 Folding Mechanisms Are Robust to Parameter Changes

We employ a model where the potential energy function is defined by the native state and all heavy (non-hydrogen) atoms are explicitly represented. Any two atoms that are close in the native structure are said to form a native contact. We describe the folding process by using the fraction of native residue pairs in contact Q_{AA} (*see methods*). Figure 2.3a shows Q_{AA}, Q_{CA} (fraction of C_α contacts, *see methods*) and radius of gyration R_g as functions of time for an AA simulation of CI2, near folding temperature. Since Q_{AA} captures the same collapse events as R_g and Q_{CA} (Figure 2.3b), Q_{AA} is a useful measure of backbone folding in addition to side chain packing.

It is crucial to understand the parameter dependence of a model before it can be used to make reliable predictions of folding mechanisms. The robustness of the folding mechanism is probed here by characterizing Protein A, SH3 and CI2 for variants of the AA structure-based energy function. Due to the debate about the balance between secondary and tertiary interactions, we vary the ratio of non-local contact energy to dihedral angles $R_{C/D}$ and the relative strength of backbone dihedral angles to side chain dihedral angles $R_{BB/SC}$ (see *methods*).

To characterize the folding mechanism for different parameter sets we computed

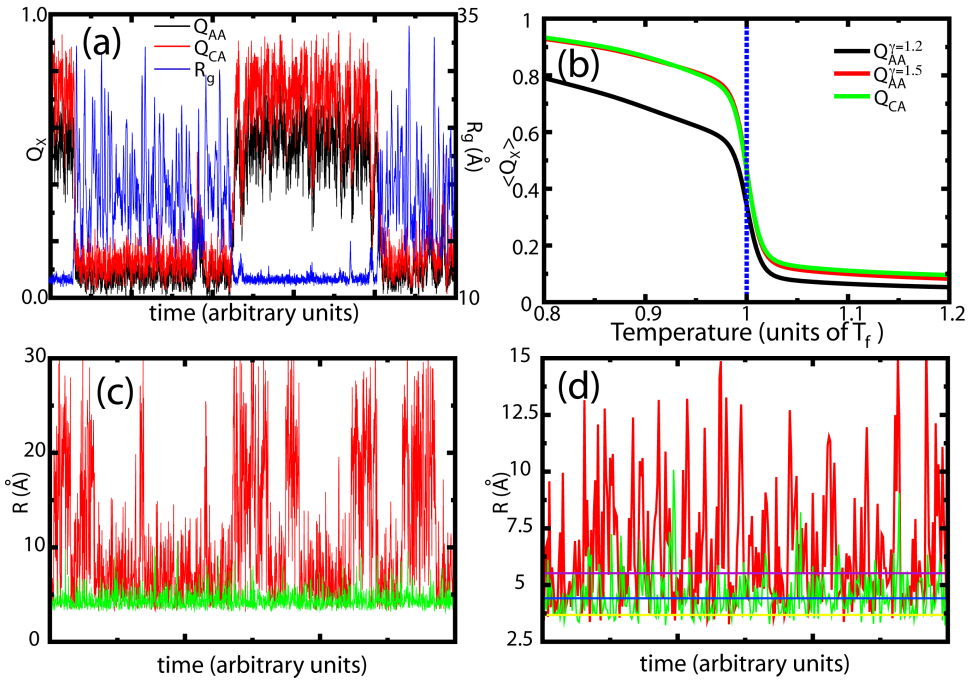

Figure 2.3: (a) Fraction of C_α contacts $Q_{\mathrm{CA}}(t)$, AA contacts $Q_{\mathrm{AA}}(t)$ and Radius of Gyration $R_g(t)$ as functions of time for a representative trajectory of CI2 with the AA model. (b) Average structure formation for several reaction coordinates. A contact between residues is formed when a single atom-atom contact between them is formed. An atom-atom contact is considered formed when the pair is at a distance $r < \gamma\sigma$ where σ is the native pair distance. The fraction of native residue contacts formed Q^X_{AA} is shown for $\gamma = 1.2$ (black) and $\gamma = 1.5$ (red). A C_α contact is formed when the C_α atoms are within 1.2 times their native distance (green). All three coordinates capture the same folding events. (c) Atom-atom distance for a contact in the active loop of CI2 versus time at T_f (red) and $T < T_f$ (green). Large changes in distance (> 20 Å) coincide with folding transitions. Side chain rearrangements in the folded state ($R < 10$ Å) occur on much faster time scales than folding of the entire protein. (d) Same as Figure (c) with time scale decreased by a factor of 100. Horizontal lines correspond to σ (yellow), 1.2σ (blue) and 1.5σ (purple). As temperature is decreased, distance fluctuations and average distances decrease.

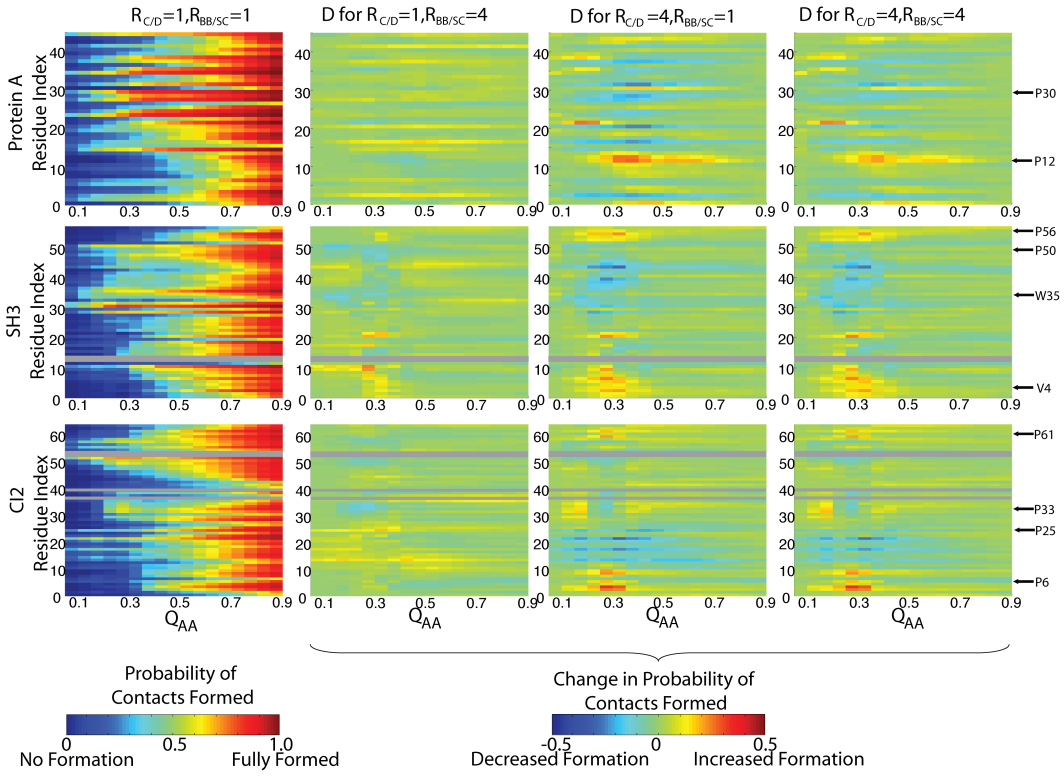

Figure 2.4: The left column shows the probability of contacts being formed for each residue $P(Q_i, Q_{\mathrm{AA}})$ as a function of Q_{AA} for $R_{C/D} = 1.0$ and $R_{BB/SC} = 1.0$. The three right columns show $P(Q_i, Q_{\mathrm{AA}})$ for different Hamiltonians relative to $R_{C/D} = 1.0$ and $R_{BB/SC} = 1.0$. Blue indicates a decrease in formation, relative to $R_{C/D} = 1.0$ and $R_{BB/SC} = 1.0$, and red an increase. Proline containing regions are often sensitive to contact energy. In Protein A, both P12 and P30 fold earlier with increased contact strength. In SH3, the increase in formation of Val4 may be attributed to interactions with Pro56, though Pro50 and Trp35 do not exhibit increased formation. In CI2, both Pro6 and Pro61 exhibit increased formation with increased contact strength. Residues that lack native contacts are shown in grey.

the probability of contacts formed as a function of the folding process $P(Q_i, Q_{\mathrm{AA}})$. $P(Q_i, Q_{\mathrm{AA}})$ is the probability that the set of contacts involving residue i, Q_i, are formed as a function of Q_{AA}. $P(Q_i, Q_{\mathrm{AA}})$ was calculated for the three proteins for 16 different parameter sets (all combinations of $R_{C/D} = 1.0, 2.0, 3.0, 4.0$ and $R_{BB/SC} = 1.0, 2.0, 3.0, 4.0$). Figure 2.4 shows the folding mechanisms for four parameter sets. The difference in folding mechanism between parameter sets i and j can be quantified by the root mean squared deviation in $P(Q_i, Q_{\mathrm{AA}})$ over all Q_{AA} and Q_i,
$P_{rms} = \sqrt{\langle (P_i(Q_i, Q_{\mathrm{AA}}) - P_j(Q_i, Q_{\mathrm{AA}}))^2 \rangle}$. The largest values of P_{rms} for Protein A, SH3 and CI2 were 0.057, 0.097 0.077. SH3 is a complicated fold, Protein A a simple fold and CI2 an intermediate fold (*99*). Thus, it is not surprising that energetic modifications have the largest effects on Protein A and the smallest effects on SH3.

Figure 2.4 shows proline containing regions are less stable to parametric modifications. Regions with prolines, and regions interacting with prolines, form structure earlier (at lower Q) with increased contact strength. This is because contact strength is increased at the expense of dihedral strength. Prolines possess a covalent $C_\delta - N$ bond, which limits the mobility of the ϕ dihedral. Removing energy from the dihedrals does not increase flexibility in prolines. However, adding energy to contacts increases structure formation around prolines. For this reason, increasing $R_{C/D}$ stabilizes and promotes earlier formation of proline containing regions.

2.2.2 Fully Folded Backbone Allows for Disordered Side Chains

While Q_{AA} and Q_{CA} capture the same cooperative folding events, at folding temperature, Q_{CA} is higher than Q_{AA} for the folded ensemble. This suggests that while the backbone structure is native ($Q_{\mathrm{CA}} \approx 0.8$), many of the native residue interactions form as temperature is decreased (Figure 2.3c and d). To account for this structurally and quantitatively, we calculated the difference between the probability of C_α contacts being formed $P(Q^i_{C_\alpha}, Q_{C_\alpha})$ and AA contacts being formed $P(Q^i_{\mathrm{AA}}, Q_{C_\alpha})$ (Figure 2.5). A value of 0 indicates that, on average, the C_α atoms of a residue pair are near their native distance when the side chains are in contact. Positive values are seen when extended side chains are interacting, resulting in the C_α atoms being far from their native distance.

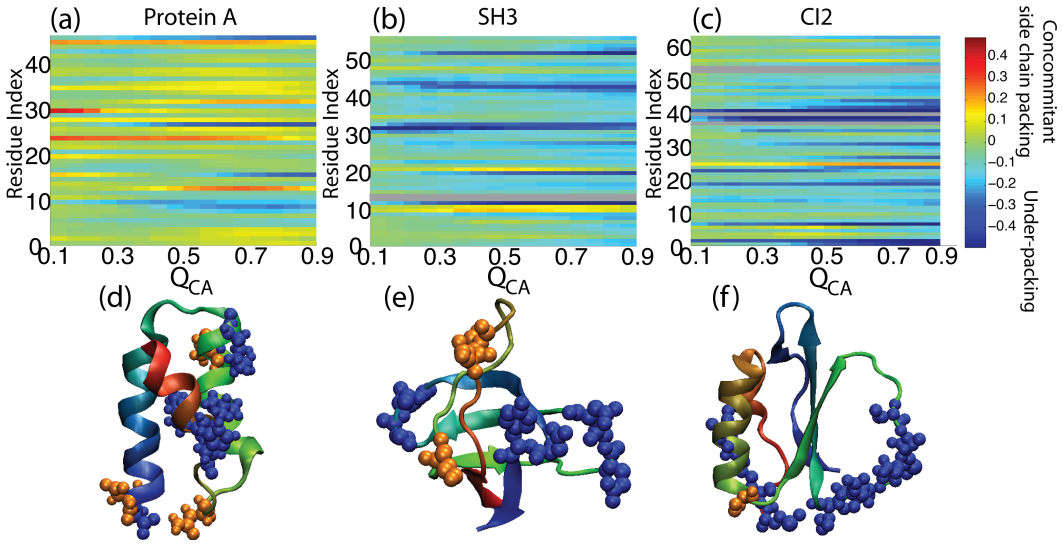

Figure 2.5: Difference in AA contact formation and C_α contact formation $P(Q^i_{\mathrm{AA}}, Q_{C_\alpha}) - P(Q^i_{C_\alpha}, Q_{C_\alpha})$ for (a) Protein A , (b) SH3 and (c) CI2. Positive values (red) indicate that residues are interacting without the C_α atoms being near. Negative values (blue) indicate the residues are "underpacked": the C_α atoms are near each other without the side chains interacting. Residues that lack native contacts are shown in grey. (d-f) Underpacked (blue spheres) and well packed (orange spheres) residues are shown on the native structures. In Protein A, to order the backbone of a helix the side chains must be packed around it. Beta sheets are stabilized by non-local interactions. Thus, a small number of contacts can maintain the tertiary structure of SH3 without the side chains in the turn regions interacting, hence the underpacking. In CI2, the active site loop is significantly underpacked.

Negative values indicate backbone folding precedes side chain ordering[1]. Side chains in Protein A appear to be well-packed, in that there is concomitant side chain and backbone folding. In SH3, the turns have negative values, and are thus underpacked. In CI2, underpacking is primarily found in the active site loop and the C-terminal tail. These results reveal a signature of complicated folds (*98, 99*): a small subset of native contacts is sufficient to constrain the backbone to its native orientation, resulting in significantly underpacked regions in the native state. This occurs in complicated folds because an individual contact can impose a high level of order on the system. In order to form contacts that are distant in sequence a large number of residues must also order. In Protein A, many contacts are local and only constrain single helical turns. In SH3 and CI2, fewer contacts are required to constrain the entire backbone (including the turns and loops).

Figures 2.3c and 2.3d show the dynamics of a typical underpacked contact. As T is lowered below T_f the underpacked contact's average distance and distance fluctuations smoothly decrease. This results in a gradual increase in Q without a noticeable free energy barrier (See Figure 2.6e). We hope that these subtle dynamics will be experimentally probed and tested in the future.

2.2.3 Understanding Free Energy Profiles Through Parametric Variation: Free Energy Profiles Can Be Altered Through Parametric Changes

While the folding mechanisms are stable, the free energy barriers associated with folding and the locations of the folded basins vary systematically with parameters. Figure 2.6 shows free energy profiles for SH3, CI2 and Protein A for several values of $R_{C/D}$ with $R_{BB/SC} = 2.0$. There are four distinct, interrelated, trends shared by all three proteins. First, there are two folding processes: backbone collapse and side chain packing. Second, the free energy minimum for the folded state moves to lower Q with increasing $R_{C/D}$. Third, the free energy barrier decreases with increasing $R_{C/D}$. Finally, increasing $R_{BB/SC}$ has similar effects as increasing $R_{C/D}$ (not shown).

[1] Q_{AA} is a generous definition of side chain packing, since a side chain is "packed" when one or more atom-atom contacts are formed. Thus, "underpacked" residues clearly have very little native structure.

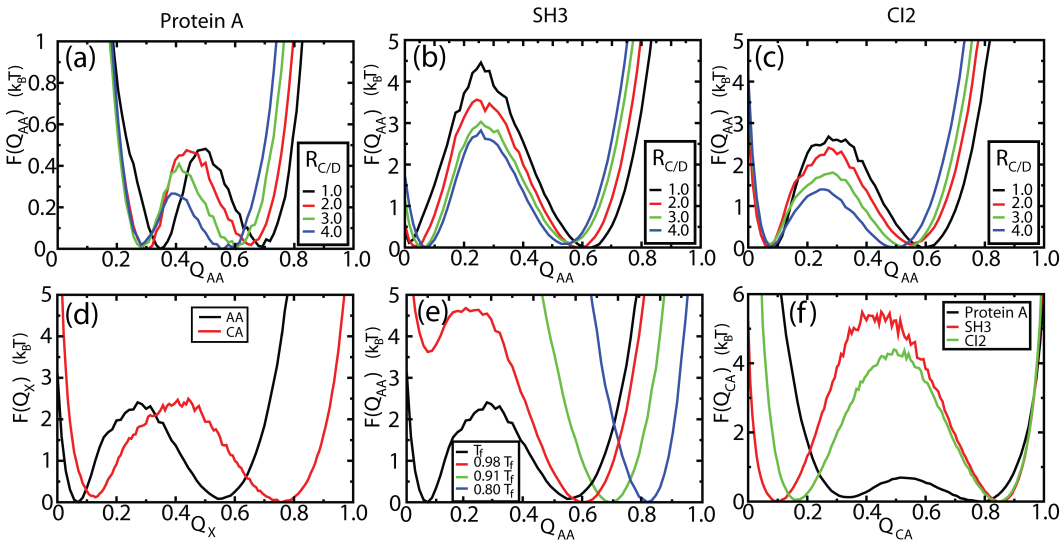

Figure 2.6: Free energy barriers in the AA model for (a) Protein A , (b) SH3 and (c) CI2. Profiles in (a-c) are for $R_{BB/SC} = 2.0$ with $R_{C/D} = 1.0$ (black), $R_{C/D} = 2.0$ (red), $R_{C/D} = 3.0$ (green) and $R_{C/D} = 4.0$ (blue). In SH3 and CI2, barrier height decreases and the folded basins move to lower Q with increasing $R_{C/D}$ and increasing $R_{BB/SC}$. (d) $F(Q_{\mathrm{CA}})$ and $F(Q_{\mathrm{AA}})$ for a typical parameter set demonstrate that the folded basins in (a-c) correspond to collapsed states. (e) Two distinct folding processes observed in our model: backbone collapse and side chain packing. (f) Free energy barriers obtained from C_α structure-based simulations for Protein A, SH3 and CI2. Barrier heights in the C_α simulations are greater than in AA simulations. Both models predict the largest barriers for SH3 and smallest for Protein A.

The free energy basins for the folded states are located at $Q_{\mathrm{CA}} \approx 0.8$ and $Q_{\mathrm{AA}} \approx 0.5$ (Figure 2.6d), indicating that the backbone orders while many native atom-atom interactions remain extended. Thus, the entropy loss during the cooperative folding transition is likely dominated by backbone ordering. Side chain packing occurs both concomitantly with, and after, backbone ordering.

There are likely two major factors that lead to the observed trends. First, increasing $R_{C/D}$ increases contact strength. As seen in other simplified models (*100*), when each contact is stronger, a smaller number of contacts is required (lower Q) to provide an equal amount of stabilizing energy. The second contributing factor is the change in side chain entropy. While entropy loss in the backbone dominates the collapse transition, the gradual side chain packing can also lead to shifting basins. Increasing $R_{BB/SC}$ or $R_{C/D}$ reduces the strength of side chain dihedrals, resulting in more mobile unfolded side chains. Therefore, there is an increased entropy loss per side chain upon folding ΔS_{sc} when $R_{C/D}$ or $R_{BB/SC}$ is increased. Since side chains can pack independently of the collapse transition, when ΔS_{sc} increases, a fraction of the side chain interactions extend, while leaving the overall fold intact. Since the folded basin shifts to lower Q, the overall structure required to form a stable fold is reduced. A reduced barrier height naturally results when the folded basin is less ordered.

Free energy barriers, in conjunction with diffusion constants, provide a direct connection to experimental folding rates (*9, 29, 101*). We find that the relative barrier heights calculated using our AA model are similar to those from a C_α model (Figure 2.6f). The relative barrier heights calculated from this model are known to correlate well with experimental rates (*29*). We note in passing, that the absolute free energy barriers in the AA model can be parametrically changed by up to a factor of two for a given protein and that the relative barrier heights between proteins remain constant. Thus, while the magnitude of the rates will be determined by the diffusion constant, the correlation between experimental folding rates and theoretical barriers is independent of the choice of parameters.

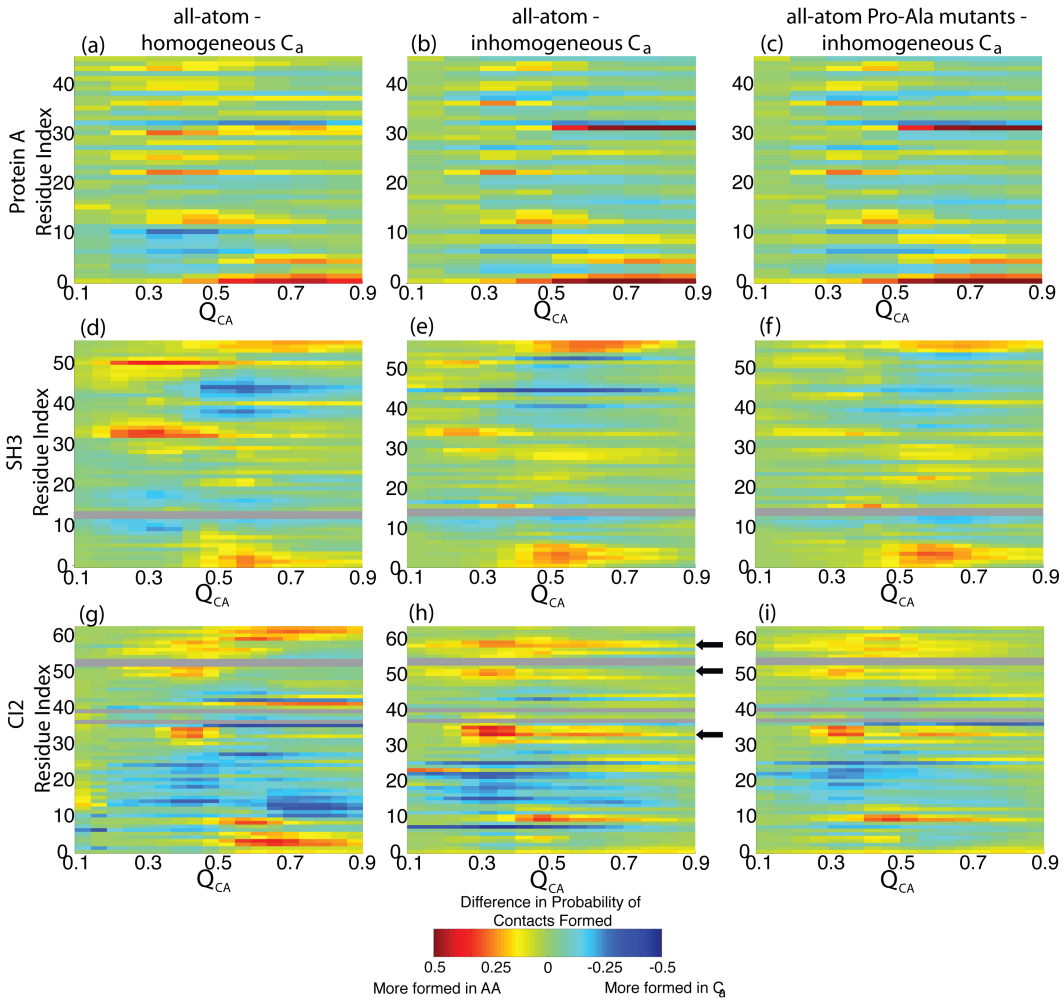

Figure 2.7: Comparison of backbone folding between C_α and AA structure-based models. The probability of contacts being formed in a C_α model, minus the probability of C_α contacts being formed in an AA model, is shown for (a-c) Protein A, (d-f) SH3 and (g-i) CI2. (a, d, g) Comparison of AA simulation to a C_α model with homogenous contact strength. (b, e, h) Comparison between AA results to an energetically inhomogeneous C_α model. Regions of increased formation in the AA representation correspond largely to proline containing regions, or regions that interact with proline, such as the minicore in CI2 (black arrows indicate mini-core residues), the tails of SH3 and turn 2 of Protein A. Increased formation in the tails of CI2 can largely be accounted for by the large number of contacts between GLU4 and ARG62. (c, f, i) The inhomogeneous C_α model compared to the AA model with all prolines mutated to alanines. Mutating proline to alanine improved agreement between models. Residues that lack native contacts are shown in grey.

2.2.4 All-Atom Structure-based Simulations Capture C_α Folding Mechanism

Next we compare the backbone folding mechanisms of our AA model and a commonly used C_α model (*13*). The C_α representation has been successful at capturing experimentally determined protein folding mechanisms (*13, 15*). The first column in Figure 2.7 shows the differences in folding mechanisms between the AA model and an energetically homogeneous C_α model. Every contact and dihedral in the homogeneous C_α model has the same interaction strength. Since the AA model distributes contact energy inhomogeneously between residue pairs, it is not surprising that the mechanisms differ.

To remove differences arising from energetic homogeneity in the C_α model, we modified it such that each contact is weighted by the number of contacts between each residue pair in the AA model (Figure 2.7, second column). For Protein A this modification improves agreement. The remaining difference is in a single turn-to-tail contact (Gln1 with Ser31, Figure 2.2a) that rarely forms in C_α simulations. In SH3, agreement improves around residues Asp34 and Asn52, while differences persist in Gln45 and the tails. The overall effect is increased formation around Gln45 at the expense of the tails. In CI2, there is significant agreement in the tails, though the mini-core still forms earlier (in the AA model), at the expense of the helix. For all three proteins, several regions of disagreement possess proline residues, whose $C_\delta - N$ bond is not included in the C_α model.

To eliminate effects specific to proline, we repeated the AA simulations with all prolines mutated to alanines. The third column of Figure 2.7 shows the Pro-Ala mutants compared to the inhomogeneous C_α model. Improved agreement is observed in Pro-Ala mutants of SH3 and CI2. In both proteins Pro-Ala mutations delay folding of proline regions, in agreement with proline effects on model stability. In SH3 the tails still form slightly earlier in the AA model, at the expense of residues 35-55. In CI2, the balance between minicore and helix formation is clearly improved, highlighting the importance of prolines in the folding process. Pro-Ala mutations have almost no effect on the folding mechanism of P12 and P30 in Protein A and P25 in CI2. This is likely because these prolines are located in turn regions. In our model, turns are highly

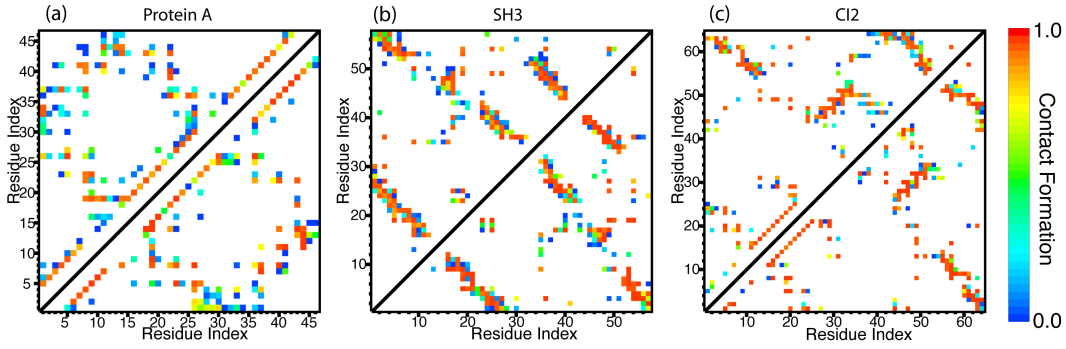

Figure 2.8: Probability of contacts being formed $P(i,j)$ at $T \approx 0.8T_f$ for the AA structure-based potential (top left) and an all-atom empirical forcefield (bottom right) for (a) Protein A, (b) SH3 and (c) CI2. Dark red indicates that residue i (x axis) and residue j (Y axis) are always in contact under native conditions. Dark blue indicates the contact is formed rarely (less than 10% of the time). White indicates $P(i,j) < 0.025$. In all three proteins, contacts are more broadly distributed (higher number of low probability contacts) in the structure-based simulations than in all-atom empirical forcefield simulations (fewer contacts, but with higher probabilities). There are approximately four times as many contacts with $P(i,j) < 0.01$ for the structure-based simulations than are seen in all-atom empirical simulations, indicating more mobile dynamics.

constrained by short range contacts, and the reduced dihedral constraint (imposed by a proline) acts as a small perturbation. The remaining differences between the Pro-Ala AA mutants and the inhomogeneous C_α model demonstrate, to no surprise, that the inclusion of side chains alters the relative entropy of residues.

2.2.5 Native Basin Dynamics of AA Structure-Based Model Correlate with the Dynamics of an All-atom Empirical Forcefield With Explicit Solvent

Two common measures of native state dynamics are native contact formation and root mean squared deviations in structure *rmsd*. Figure 2.8 shows the average contact formation in the native ensemble for the structure-based model and an all-atom empirical forcefield with and explicit solvent. While the average contacts are not identical, no major differences in contact formation are observed. The overlaps between the AA maps and the all-atom empirical forcefield maps of Protein A, SH3 and CI2 are 0.85,

0.97 and 0.84. An overlap of 1 indicates identical maps, and 0 indicates the two maps have no contacts in common.

In a uniquely defined native state, the probability of each contact being formed is 1. Since we sample the native ensemble at finite temperatures, atom mobility leads to additional contacts being formed. In the structure-based model, these additional interactions are strictly repulsive. In an all-atom empirical forcefield these interactions can be attractive, yet they are observed more frequently in the structure-based model[2]. These contacts are likely due to increased mobility in the structure-based simulations. In all-atom empirical forcefields, hydration shells can result in less mobile side chains, and hence a narrower distribution of contacts.

The increased mobility is quantified by the structural *rmsd*. The magnitude of fluctuations in all-atom empirical simulations is much lower than in structure-based simulations (not shown). For the all-atom empirical forcefield at 300 K, the average *rmsd* for Protein A, SH3 and CI2 are 1.53, 1.00 and 0.97 Å. The *rmsd* of the C_α atoms are 1.23, 0.66 and 0.74. The same values are obtained in structure-based simulations at around $T = 0.55T_f$. In real temperature units, $0.55T_f$ corresponds to temperatures significantly less than 300 K. A likely cause for the increased structural fluctuations is hydration effects of explicit solvent molecules in the all-atom empirical forcefield. To compare the distribution of *rmsd* fluctuations between models, correlation coefficients (r) were computed for the *rmsd* by atom in the all-atom empirical forcefield and the structure-based potential. For all parameter sets of the structure-based potential, the $r \approx 0.7$ for CI2 and SH3 and $r \approx 0.8$ for Protein A[3].

2.3 Discussion

In this manuscript, we describe a systematic analysis of an AA structure-based model which bridges the gap between coarse-grained models and all-atom empirical forcefields. We show that in our C_α and AA structure-based models the global folding mechanisms agree and the main differences are largely due to energetic heterogeneity and the explicit representation of prolines in the AA model. Also, the native basin dy-

[2] In Figure 2.8 only interactions present more than 2.5% of the time are shown.

[3] Comparison of *rmsd* of the C_α atoms yields similar values of r.

namics are similar in the AA structure-based model and an all-atom empirical forcefield with explicit solvent. In agreement with previous studies, the folding mechanisms in complicated folds are stable to parametric variation. On the other hand, the free energy barriers associated with folding vary systematically with parameters. Since free energy barriers are not a robust feature of this model, understanding the interplay between barrier heights and diffusion will be important before attempting to predict folding rates (*9, 44, 102*).

Using this model we characterized two folding processes: one associated with backbone collapse and the other with side chain packing. We observed that backbone collapse is accompanied by partial side chain packing in a cooperative transition and residual side chain packing occurs as temperature is reduced below the global folding temperature. One explanation for the partial separation of backbone folding and side chain ordering may be that mobility in specific residues is necessary for the functional properties of proteins. Proteins are selected for their function. Orthogonal networks of residues responsible for stability and function have been proposed (*103, 104*). The observation in our model that some residues are not necessary to maintain the backbone structure is consistent with this proposal. In CI2, the backbone of the active site loop is in the native orientation, yet the side chains are not packed. In SH3, several turns are also disordered. Since binding sites are often found in loops, flexible loops may be more easily adapted to new sequences and functions.

Gradual side chain packing can also allow for proteins to functionally respond to cellular stress by affecting side chain orientations, without denaturing the entire protein. This is consistent with the prediction that localized unfolding, or cracking, is important for biological function of kinases and motor proteins (*19, 19, 22, 32, 33, 105, 106*).

The current model explicitly includes the effects of topological contributions to protein folding, and the role of energetic contributions may now be elucidated. Our results are a significant step forward in understanding protein dynamics from the C_α to the all-atom level. In the coming years, it will be interesting to probe the effects of electrostatics, non-native interactions, water and explicit mutations in this model.

2.4 Methods

The all-atom and C_α models are detailed in Chapter 1 (Equation 1.4). As a reaction coordinate we use Q_{AA} and Q_{CA}. Q_{AA} is the fraction of natively interacting residues that are in contact. Two residues are considered in contact if any native atom-atom interactions between the residues are within 1.2 times the native distance r_0^{ij}. At $1.2r_0^{ij}$ the potential energy of a native pair is approximately half of the minimum. Similarly, Q_{CA} is the fraction of natively interacting residue pairs whose C_α atoms are within 1.2 times their native distance.

2.4.1 Simulation Details

All-atom structure-based simulations were performed using the Gromacs software package (*74*). No modifications to the source code were necessary. Reduced units were used. The timestep τ was 0.0005. The Berendsen algorithm (*107*) was used[4] with the coupling constant of 1. For all folding results in this paper, several constant temperature runs were performed, with temperatures that corresponded to the protein being always folded to always unfolded. The Weighted Histogram Analysis Method (*108,109*) was used to combine data from multiple temperatures into single free energy profiles.

2.4.2 All-Atom Empirical Forcefield Simulations

All-atom empirical forcefield simulations were performed using Gromacs (*74*), with the OPLS-AA forcefield (*110*) with TIP3P water molecules (*111*). Each protein was simulated for 10 ns at T=300K and a pressure of 1 atm. A timestep of 2 fs was used in conjunction with the LINCS algorithm for constraining covalent bonds with hydrogen. Protein A, SH3 and CI2 were simulated with 2810, 3617 and 4644 water molecules in cubic boxes of initial dimensions 45.15 Å, 48.98 Åand 53.07 Å. Temperature was

[4]When using the Berendsen thermostat, numerical instabilities can arise when the bath-molecule coupling timescale is shorter than the timescale for internal energy diffusion. In our experience, these problems tend to surface when you simulate weakly interacting domains with implicit solvation. Since the present study investigates folding of single domain proteins under weak temperature coupling, these features are not likely a source of significant errors. Nonetheless, future work will employ stochastic or Langevin temperature coupling.

maintained using the Berendsen algorithm (*107*). 1 ns was allowed for equilibration. For the remaining 9 ns, structures were saved at 1 ps intervals.

2.4.3 Proline to Alanine Mutations

To investigate the role of proline residues in the AA model, proline to alanine mutants were constructed. This was achieved by removing the C_γ and C_δ atoms of each proline. Native contacts formed with the C_γ and C_δ of a proline were included as contacts with the C_β of the corresponding alanine. This ensured the energetics of the system were unperturbed, and only topology was modified.

2.4.4 Comparison of Contacts

In the all-atom empirical forcefield simulations contacts were determined for each saved structure using CSU (*112*). The average number of contacts $\langle Q \rangle$ was calculated for each protein. The probability of individual contacts being formed was averaged over all structures with $Q = \langle Q \rangle$. With the all-atom empirical potential $\langle Q \rangle$ was 80, 135 and 146 for Protein A, SH3 and CI2. This analysis was repeated for folded simulations with our AA structure-based simulations. For the structure-based simulations $\langle Q \rangle$ was 80, 138 and 144. To compare contact maps, the dot product of the two maps was taken.

2.5 Acknowledgements

Chapter 2, in part, appears in *Proteins: Structure, Function, Bioinformatics*, (2009), Whitford, Noel, Gosavi, Schug, Onuchic. The dissertation author and Paul Whitford are the primary investigators and coauthors of the paper. We would like to thank Angel Garcia and Peter G. Wolynes for useful discussions regarding all-atom modeling. PCW and JKN were supported in part by the National Institutes of Health Molecular Biophysics Training Program at University of California at San Diego Grant T32 GM08326. This work was supported in part by Grants PHY-0216576 and 0225630 from the National Science Foundation (NSF)-sponsored Center for Theoretical Biological Physics, NSF Grant 0543906, the LANL LDRD program and NIH Grant R01-

GM072686.

Chapter 3

The Shadow Map: A General Contact Definition for Capturing the Dynamics of Biomolecular Folding and Function

3.1 Introduction

Chapter 2 characterized an all-atom SBM (Equation 1.4), which explicitly represented the atomic geometry of a biomolecule (*35*). This SBM is a baseline model that can be used to fully discern the role of biomolecular geometry. While our initial study showed the robustness of the all-atom SBM Hamiltonian to changes in many of the energetic parameters (*35*), an important aspect, which has not been explored, is the definition of native interactions. Each native interaction, or "native contact," is formed by an atom-atom pair (or residue-residue pair in a C_α representation) interaction that is proximate in the native state. The set of native contacts is called a *contact map* and is a ubiquitous tool in the analysis of internal biomolecular interaction networks (*31, 99, 113*).

The definitions of contact maps in the literature are nearly as diverse as their applications. The simplest algorithms define contacts between atom (or residue) pairs that are within a cutoff radius of each other (*114*). More complicated algorithms additionally consider, for example, solvent accessibility (*115, 116*) or atomic chemistry (*117*). For protein folding studies, contacts have often been defined through the atomic geometry,

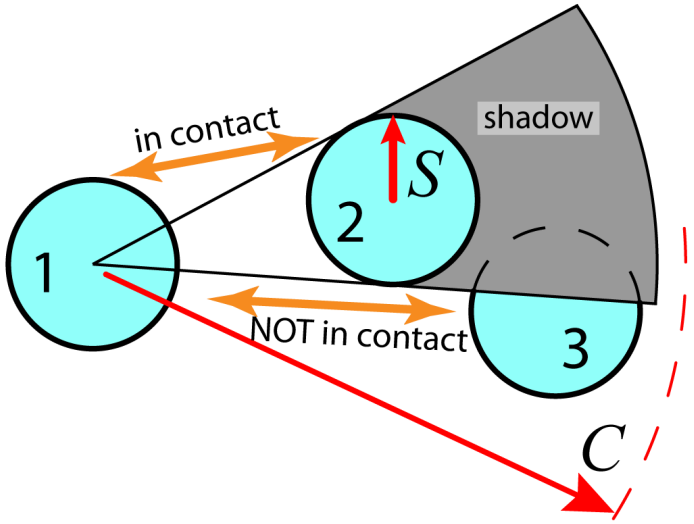

Figure 3.1: The Shadow contact map screening geometry. Only atoms within the cutoff distance *C* are considered. Atoms 1 and 2 are in contact because they are within *C* and have no intervening atom. To check if atom 1 and atom 3 are in contact, one checks if atom 2 shadows atom 1 from atom 3. The three atoms are viewed in the plane and all atoms are given the same shadowing radius *S*. Since a light shining from the center of atom 1 causes a shadow to be cast on atom 3, atoms 1 and 3 are not in contact.

by choosing residue pairs that have heavy atoms within a cutoff distance (4.0 Å to 6.5 Å) (*46, 98, 118, 119*) or atom pairs that shield each other's solvent accessibility (*13, 47*). In a SBM, the native contact map is an integral part of the Hamiltonian, since it defines the distribution of stabilizing energy in the biomolecule. Therefore, as SBMs are being explored in multiple levels of detail and are being applied to increasingly diverse and heterogenous systems, a consistent method for choosing contact maps is desirable.

In this study, we propose a general definition for generating atomically-grained contact maps called "Shadow" (Figure 3.1). It is motivated by the need to satisfy two mutually incompatible features of a simple heavy atom cutoff contact map: to include relevant contacts at distances of at least 6 Å without introducing nonphysical next-nearest neighbor contacts. Long cutoffs enable the map to capture atomic contacts across structural waters or heavy metals that are not explicitly represented. At long cutoff distances though, contacts will be introduced between atom pairs that we do not wish to

model, specifically, those that have an intervening atom. The Shadow algorithm initially considers all atoms within a cutoff distance C and then, controlled by a screening parameter S, discards the occluded contacts. We compare two classes of contact maps: 1) maps based on a simple cutoff distance C and $S = 0$, and 2) maps with $S > 0$. They are compared dynamically by measuring the folding thermodynamics of well studied two-state proteins, the thermodynamics of an RNA hairpin, and the native basin fluctuations of the ribosome. We find that the Shadow contact map gives a consistent definition of atomically-grained native interactions from small proteins up to macromolecular assemblies. Two-state proteins and RNA hairpins show reliably cooperative folding transitions. Also, Shadow contact maps distribute energy similarly between RNAs and proteins despite their disparate internal packing. All-atom structure-based models employing Shadow contact maps are a general framework for exploring the geometrical features of biomolecules, especially the connections between folding and function.

3.2 Methods

3.2.1 The All-Atom Structure-Based Model

The all-atom model is detailed in Chapter 1 (Equation 1.4). Especially important to this chapter is that the total stabilizing energy is set to a constant, $\sum \varepsilon_C + \sum \varepsilon_{\mathrm{BB}} + \sum \varepsilon_{\mathrm{SC}} = \varepsilon N_{\mathrm{atoms}}$, where ε is the reduced energy unit. This means that as the contact map is varied, even though the number of contacts may vary, the net energy contribution from the contacts is constant at $\frac{2}{3}\varepsilon N_{\mathrm{atoms}}$. The energy per contact though will vary. This allows for careful comparison between the different native contact maps. Note that previous implementations of SBMs of RNA (*50*) reduced the strength of contacts between stacked bases by a factor of 3 when using cutoff maps (also see Section 3.3.4). These maps are marked with an asterisk, like M_4^{0*}.

3.2.2 Simulation Details

AA structure-based simulations were performed using the Gromacs software package (*74*). Protein simulations were typically performed on 4 cores and the ribosome simulations were performed on 128, or 256, cores each. The Gromacs source code was modified to include the Gaussian interaction (available at http://smog.ucsd.edu); no further modifications were necessary. The Gromacs topology files were generated with the smog@ctbp webserver (*57*). Reduced units were used. The time step τ was 0.0005. Temperature was controlled through stochastic dynamics with a coupling constant of 2. For all systems simulated in this paper, several constant temperature trajectories were obtained. In the case of folding, temperatures varied from the protein being always folded to always unfolded, and trajectories contained many folding transitions (> 20). The Weighted Histogram Analysis Method (*108, 109*) was used to combine data from multiple temperatures into single free energy profiles. Each ribosome simulation was performed for 2×10^7 time steps, with the second half used for data analysis. Fluctuations in proteins are calculated from 2×10^7 time steps of data. Convergence of native-state fluctuations was reached by 10^7 time steps, since doubling the data gives no discernible difference in the averages.

3.2.3 Contact Maps

Atoms that are spatially near in the native state are considered *contacts* and together the set of all contacts composes a *native contact map*. A contact map encodes which atom pairs ij are given attractive interactions in the SBM potential. In the context of a SBM, the native contact map sets the distribution of renormalized stabilizing enthalpy in the native state.

Here, we propose an algorithm for determining atomic contacts, called Shadow. It uses a heavy-atom cutoff distance together with geometric occlusion. There are two parameters in the algorithm, the cutoff distance C and the screening radius S (Figure 3.1). The algorithm can be metaphorically described: if a light source were located at the center of atom i, and all other atoms were opaque, then all atoms within the cutoff C that have no shadow cast upon them would be considered contacts. To keep the bonded

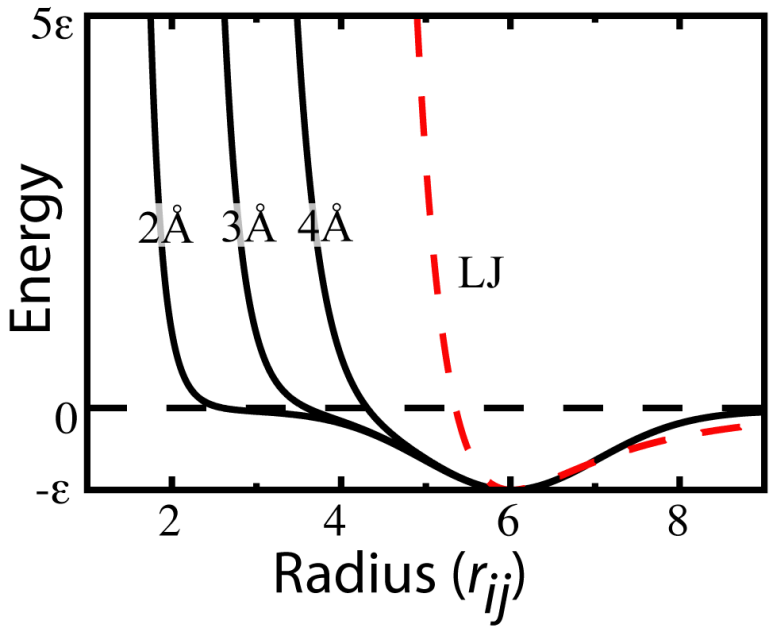

Figure 3.2: The versatile Gaussian contact potential. The repulsive part is constrained to shift with the position of the minimum in the Lennard-Jones potential, which introduces extraneous excluded volume for each native contact. In contrast, the excluded volume can be independently set relative to the contact minimum with the Gaussian contact potential. This decouples the energetics of the contact map from the protein geometry.

atoms from overlapping, S is maintained ≤ 0.5 Å when screening a bonded neighbor. To put shadowing in the context of other approaches, we compare it to the commonly used simple heavy-atom cutoff distance ($S = 0$). We denote a contact map with cutoff distance C and screening radius S as M_C^S. C and S are given in units of Å. Shadow maps were generated with the smog@ctbp webserver (http://smog.ucsd.edu) (*57*). The default map, termed "the Shadow map," refers to M_6^1. Related geometric occlusion methods have been employed by Wu *et al.* (*120*) and by Veloso *et al.* (*117*).

3.2.4 Contact Potential

All of the pair interactions defined in the native contact map interact through an attractive potential, denoted in the SBM potential by $C(r_{ij}, r_0^{ij})$ (Equation 1.4). The contact potential has a minimum at the distance between the pair in the native state r_0^{ij}.

Traditionally, a contact is defined through a Lennard-Jones (LJ) type potential,

$$C_{\mathrm{LJ}}(r_{ij}, r_0^{ij}) = \left(\frac{r_0^{ij}}{r_{ij}}\right)^{12} - 2\left(\frac{r_0^{ij}}{r_{ij}}\right)^{6}. \tag{3.1}$$

The LJ potentials are well tested and perform well for many systems, but they introduce an excluded volume that scales with the contact distance (Figure 3.2). Since the effective volume of two atoms in contact grows with r_0^{ij}, this can lead to complications for certain applications. The variable repulsion introduces heterogeneity into coarse-grained beads, which allows the model to capture effective excluded volume effects. However, it is less clear that this feature is appropriate for all-atom SBM, since the excluded volume should already be explicitly captured by the all-atom geometry.

In this and all following chapters, we employ a contact potential that allows independent control of the excluded volume. By decoupling the protein geometry from the energetics, the contact map definition is independent of the excluded volume. Without this feature, the consequences of varying the contact map will be obscured by the entropic effects of the varying excluded volume. Contact interactions are represented by an attractive Gaussian well coupled to a fixed LJ repulsion,

$$C_{\mathrm{G}}(r_{ij}, r_0^{ij}) = \left(1 + \left(\frac{r_{\mathrm{ex}}}{r_{ij}}\right)^{12}\right)\left(1 + G(r_{ij}, r_0^{ij})\right) - 1 \tag{3.2}$$

where

$$G(r_{ij}, r_0^{ij}) = -\exp\left[-(r_{ij} - r_0^{ij})^2/(2\sigma_{ij}^2)\right]. \tag{3.3}$$

This functional form ensures that the depth of the minimum is -1 (scaled by ε_{C} in Equation 1.4), and r_{ex} sets the excluded volume. r_{ex} has the same function as r_{NC} in Equation 1.4. If $r_{\mathrm{ex}} = r_{\mathrm{NC}}$, all atomic interactions have an equal excluded volume. For consistency with the LJ potentials, the width of the Gaussian well σ_{ij} models the variable width of the LJ potential. $C_{\mathrm{LJ}}(1.2r_0^{ij}, r_0^{ij}) \sim -1/2$ so σ_{ij} is defined such that $G(1.2r_0^{ij}, r_0^{ij}) = -1/2$ giving $\sigma_{ij}^2 = (r_0^{ij})^2/(50\ln 2)$. If r_{ex} is significantly smaller than r_0^{ij} Eq. 3.2 reduces to a more transparent form,

$$C_{\mathrm{G}}(r_{ij}, r_0^{ij}) \to \left(\frac{r_{\mathrm{ex}}}{r_{ij}}\right)^{12} + G(r_{ij}, r_0^{ij}) \quad \text{for } r_{\mathrm{ex}}, \sigma_{ij} \ll r_0^{ij}. \tag{3.4}$$

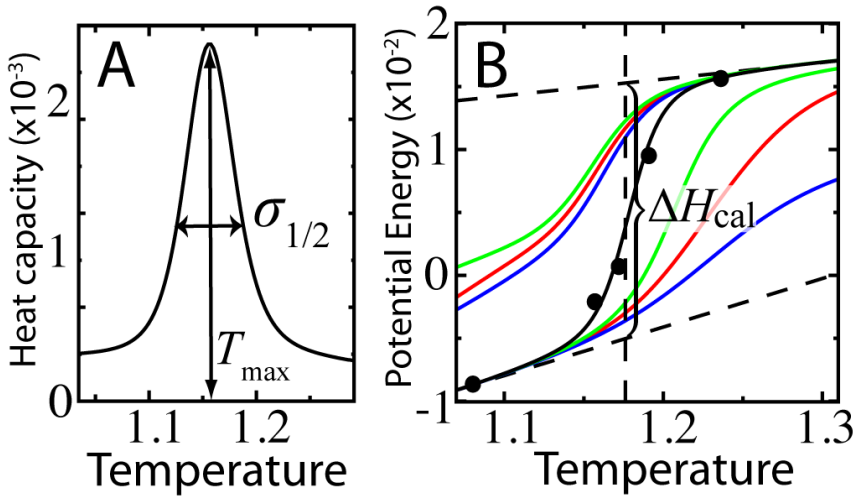

Figure 3.3: Measures of cooperativity. (**A**) $\kappa_1 = \sigma_{1/2}/T_{\max}$. $\sigma_{1/2}$ is the width of the heat capacity at half the maximum. (**B**) κ_2 is a measure of the enthalpy change associated the transition relative to the total enthalpy change ΔH_{cal}. The behavior of the enthalpy in the folded and unfolded states is modeled linearly (horizontal dotted lines). The vertical dotted line marks T_{F}. Weighted histogram analysis gives the continuous lines. Black dots show $\langle H \rangle$ during constant temperature molecular dynamics. The blue, red and green lines show the folded and unfolded state enthalpy for different rmsd cutoffs, d_{c} of 5, 6, and 7 Å, respectively.

3.2.5 Thermodynamics

Folding experiments on small globular proteins have long shown evidence of thermodynamic and kinetic cooperativity, which indicates a phenomenon similar to a first order phase transition between native and denatured states (*95, 121*). To quantify the thermodynamics and cooperativity of the SBM, the heat capacity was calculated. Two different dimensionless measures of cooperativity are considered: the width of the peak in the heat capacity κ_1 and the van't Hoff criterion κ_2. Both are applicable for describing the cooperativity of two-state transitions (*47, 122, 123*).

$$\kappa_1 = \frac{\sigma_{1/2}}{T_{\max}}, \tag{3.5}$$

where $\sigma_{1/2}$ is the full width at half maximum of the heat capacity $C_V \equiv \frac{\partial \langle H \rangle}{\partial T}$ and $T_{\max}$ is the temperature corresponding to the peak in C_V (Figure 3.3). κ_1 is interpreted as a measure of the temperature range over which the transition occurs, where smaller κ_1 indicates a higher degree of cooperativity.

The van't Hoff criterion κ_2 is a measure of cooperativity that is based on the enthalpy distribution during the transition. A cooperative transition has a well defined energy separation between unfolded U and folded F ensembles. With $K_{eq} = [F]/[U]$ as the equilibrium constant of the folding reaction, the van't Hoff criterion is defined at the midpoint of the transition, given by $K_{eq} = \frac{1}{2}$.

$$\begin{aligned} \kappa_2 &= \frac{-k_B T^2}{\Delta H_{cal}} \frac{\partial \ln K_{eq}(T)}{\partial T}\Big|_{K_{eq}=\frac{1}{2}} \\ &= \frac{\langle H \rangle_U - \langle H \rangle_F}{\Delta H_{cal}}\Big|_{K_{eq}=\frac{1}{2}} \end{aligned} \tag{3.6}$$

where ΔH_{cal} is the calorimetric enthalpy change of the transition and $\langle H \rangle_X$ is the average enthalpy of ensemble X. ΔH_{cal} is the integral of C_V over the transition region and is determined by extrapolating the unfolded state enthalpy and the folded state enthalpy to $T_{1/2}$, the temperature where $K_{eq} = \frac{1}{2}$ (Figure 3.3). These extrapolations, known as *baselines*, approximate the temperature dependence of the enthalpy in the absence of the protein transition (*123*). The baselines, H_U and H_F, isolate the heat change of the transition, $\Delta H_{cal} = H_U(T_{1/2}) - H_F(T_{1/2})$. Determination of $T_{1/2}$ requires a definition of the unfolded and folded ensembles. In this investigation, a cutoff in root mean square

deviation from the native state (rmsd) d_c is used to partition configurations (*63, 88*). The "proper" d_c may be determined simply by nuumerically maximizing κ_2, *i.e.* $\partial\kappa_2/\partial d_c = 0$. Note that simplifying the calculation by fixing $\langle H\rangle_F = H_F$ will overestimate κ_2 since $\langle H\rangle_F > H_F$.

3.3 Results and Discussion

Since SBM are applied to diverse biomolecular systems, the present study encompasses a broad range of biomolecular systems, in particular, globular proteins, RNA, and the ribosome. First, we discuss the effects of geometric occlusion on the number and distribution of native contacts in globular proteins and in RNA secondary structure. Then we analyze the sensitivity of both folding thermodynamics and native state fluctuations to the choice of native contacts in model protein and RNA systems. Lastly, we examine the sensitivity of fluctuations to the contact map in a large molecular assembly: the ribosome.

3.3.1 Protein Contact Maps

Protein native structures, as determined by structural biology techniques, are compact and densely-packed structures stabilized by both short- and long-range interactions (*125, 126*). The all-atom SBM encodes the stability imparted by these interactions with short-range attractive potentials between pairs of atoms. These interactions drive the protein towards the low free energy native configuration. The short-range atomic interactions in proteins are on the Å length scale. The closest pairs are the hydrogen bonding interactions between the carboxyl O and amino N found throughout α-helices and β-sheets. The N-O are commonly separated by 2.6-3.0 Å. In the hydrophobic core, carbon pairs are separated by 3.5-4.5 Å. This longer distance is a consequence of the larger Van der Waals radius of carbon compared to oxygen and nitrogen. Salt-bridges exist in protein cores with separations up to 5.5 Å (*127*). Indirect pair interactions mediated through water molecules, either surface or buried, can vary between 5-7 Å (*126*) and are a source of enthalpic stabilization (*124*).

An algorithm to generate protein contact maps that includes all of the above

Table 3.1: Statistics on various contact maps of model globular proteins.

		M_4^0	$M_4^{0.7}$	M_4^1	M_5^1	M_6^1	M_5^0	M_6^0
Protein	Residues	Total contacts						
UBQ[a]	76	387	322	262	625	874	1504	3510
CI2[b]	64	280	229	188	466	597	1117	2566
SH3[c]	57	325	266	185	409	532	1397	3155
BDPA[d]	46	179	153	115	222	329	575	1345
	Contacts[e] per res.	Contacts per atom						
CI2	9.3	0.56	0.45	0.37	0.92	1.2	2.2	5.1
SH3	9.3	0.71	0.58	0.41	0.89	1.2	3.1	6.9
BDPA	7.1	0.49	0.41	0.31	0.60	0.90	1.6	3.6
NHGP[f]	9.1	0.70	0.60	0.43	0.89	1.2	2.6	5.7
		Dispersion in contacts between residues[g]						
NHGP		2.50	2.49	2.05	1.83	1.63	2.06	1.67
		C_V criterion, κ_1						
CI2	3-state	0.038	0.031	0.024	0.032	0.063	0.12	
SH3	0.027	0.025	0.021	0.025	0.028	0.033	0.047	
BDD	0.14	0.087	0.050	0.046	0.046	0.072	0.10	
		van 't Hoff criterion, κ_2						
CI2	3-state	0.66	0.73	0.93	0.84	0.70	0.72	
SH3	0.82	0.91	0.93	0.96	0.94	0.90	0.89	
BDD	0.77	0.71	0.76	0.79	0.84	0.70	0.73	

[a] PDB code 1UBQ
[b] PDB code 1YPA
[c] PDB code 1FMK, residues 84-140
[d] PDB code 1BDD
[e] Contacts per residue with map M_6^1.
[f] Average over a set of 33 non-homologous globular proteins (*124*).
[g] Atom-atom contacts per residue-residue contact, dispersion is normalized by average

mentioned short-range interactions must accommodate pair separations up to at least 6 Å, or more. While a simple cutoff distance criterion will capture all of the essential interactions, there will also be many occluded pair interactions that we are not seeking to model (Figure 3.1). The occluded interactions represent 3-body interactions, and their effects should be considered higher-order corrections. These higher-order, occluded contacts can be identified, and discarded, by using the shadowing geometric criteria described in Section 3.2.3. The parameter choices of $C = 6$ Å and $S = 1$ Å, or M_6^1, define the contact map henceforth called "the Shadow map."

Removal of Contacts Through Geometric Occlusion

The abundance of occluded contacts is checked by constructing various native contact maps M_C^S, where S and C are measured in Å and are described in Figure 3.1. S is the screening strength, which sets the radius of each shadowing atom, and C is the cutoff radius that sets the maximum separation allowed between contacts. Results are summarized for 4 proteins in Table 3.1. To quantify average values and statistical variability in the contact map calculations, we use a standard library of 33 non-homologous globular proteins (NHGP) often used in structure prediction (*124*). Figure 3.4A shows the number of contacts per atom $N(M_C^S)$ as a function of the cutoff radius for different shadowing sizes, averaged over NHGP. M_6^0 has nearly 6 contacts per atom, but for M_6^1 it drops to 1.2 contacts per atom. Thus, geometric occlusion removes 80% of the contacts if shadowing atoms are given a radius of 1 Å. Interestingly, shadowing has a significant effect, even at cutoff distances as small as 4 Å, $N(M_4^0) = 0.70$ while $N(M_4^1) = 0.43$.

While shadowing removes contacts that are occluded by intervening atoms, longer distance contacts that are separated by buried (implicit) solvent are maintained. Figure 3.5A,B show the contact networks in regions of disrupted secondary structure, where buried waters satisfy the left over backbone hydrogen bonds. The black dotted lines highlight contacts that are separated by more than 4.5 Å but that are included in the Shadow map. In both cases shown, there are waters sufficiently localized to be detected in x-ray crystallography, which are depicted by yellow spheres. The water molecules sit in voids in the protein interior, and provide stabilization to a configuration that would otherwise be enthalpically costly. Although the solvent is not explicitly modeled in

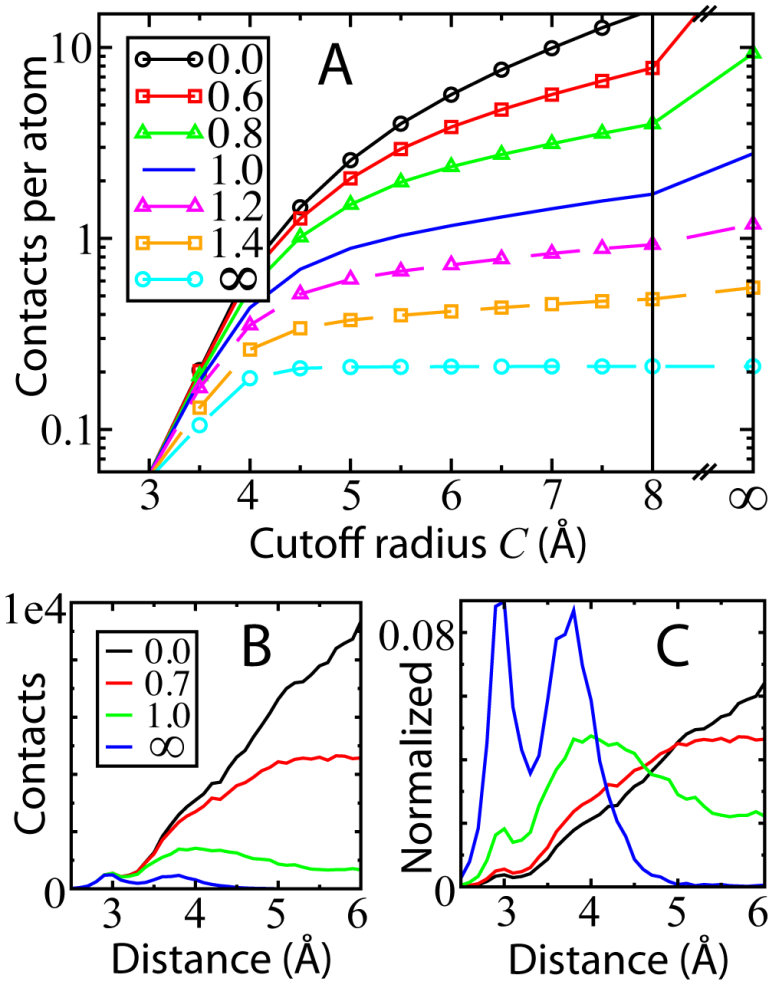

Figure 3.4: The removal of contacts through shadowing. (**A**) Average number of contacts per atom N_C^S as a function of cutoff radius C. Each curve is for a different shadowing size S. Each N_C^S represents an average over all the atoms of the proteins in NHGP. $C,S = \infty$ means C,S greater than the diameter of the protein. (**B**) Contact distance histogram for different S. Curves represent a sum over all the proteins in NHGP. (**C**) Contact distance histogram normalized by total contacts, $N_C^S \times N_{\text{atoms}}$. This corresponds to the distribution of contact energy as a function of contact distance. $S = 0$ is peaked at the cutoff limit whereas $S = 1$ peaks at an intermediate distance of 4 Å.

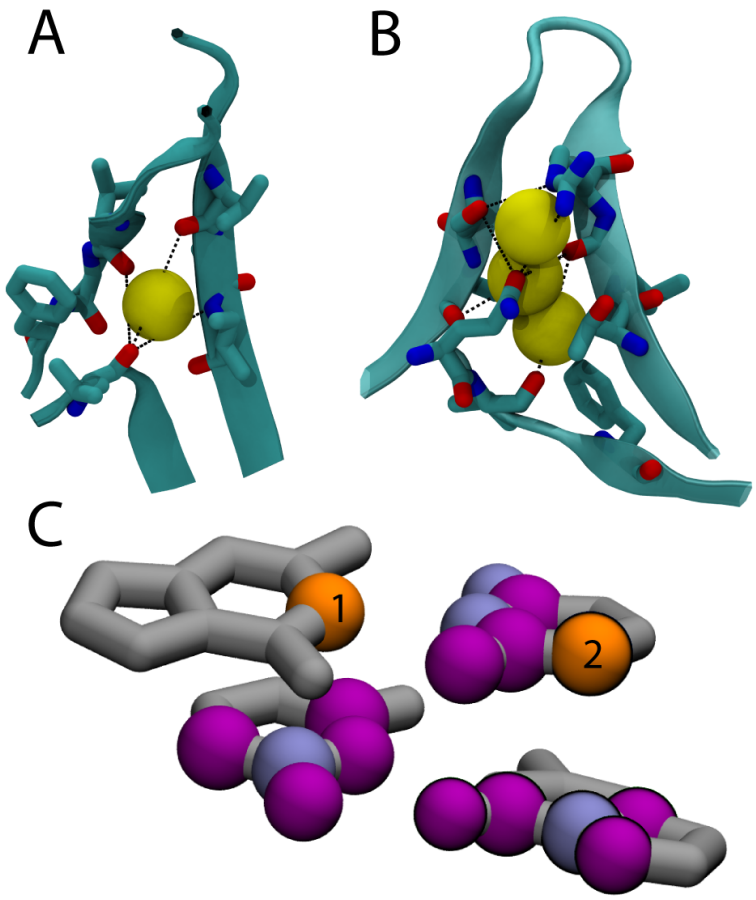

Figure 3.5: Shadow automatically includes contacts where ligands, metal clusters and buried waters are not explicitly represented. (**A**) Buried water in flavodoxin (PDB code: 2FCR) coordinated by VAL87:O, VAL121:O, LEU143:O, and ILE89:N. Black dotted lines show three contacts longer than 4.5 Å involving LEU143:O that are included in M_6^1. (**B**) Buried waters in interleukin 1-β (PDB code: 1L1B) where three β-sheets come together. Several contacts included in M_6^1 and longer than 4.5 Å are shown. (**C**) Shadowed contacts in an RNA helix. Native contacts within 4 Å (M_4^0) of the two numbered atoms (orange) are shown in blue and purple. The contacting atoms that are shadowed (*i.e.* not in M_4^1) are shown purple. Atom 1 has both stacking interactions and base pairing interactions while atom 2 has only stacking interactions

SBM, choosing contacts through shadowing automatically fills these open pockets with compensating contacts because there are no occluding atoms in the void left by the solvent.

There are global differences between the distributions of stabilizing contact energy between cutoff and shadowing maps, *i.e.* $S = 0$ and $S = 1$. The most obvious difference is the significant reduction in the total number of contacts when $S = 1$. This reduction in contacts is strongest for the longest distance contacts, since they are more likely to be occluded. This alters the contact radial distribution function (Figure 3.4B,C). The distribution becomes more heavily weighted towards short-range contacts. Peaks at 3 Å and 4 Å become visible in M_6^1 and are more pronounced for M_∞^∞ (only nearest neighbors). For all contact maps, the 3 Å peak is due to the hydrogen bonding interactions along the secondary structure and the 4 Å peak results from hydrophobic interactions. A more subtle difference is that shadowing tends to smooth the distribution of stabilizing energy between residues. There is a reduction in residue-residue contact energy variance for $S > 0$ (Table 3.1). Residue-residue contact energy is defined as the sum of atom-atom contacts shared between two residues. These differing contact energy distributions will be seen to alter the thermodynamics of protein folding (discussed in Section 3.3.3).

A quantity that shows no systematic variation with contact map is the relative contact order (CO). Averaged over the proteins in NHGP, $\langle \Delta \mathrm{CO} \rangle = \langle \mathrm{CO}_i(M_6^1) - \mathrm{CO}_i(M_6^0) \rangle \approx 0$, and there is little variation from protein to protein since $\langle |\Delta \mathrm{CO}|/\mathrm{CO} \rangle = 0.04$. The constant CO shows that the ratio of long range to short range (in sequence) contacts is constant.

Parameter Reduction: $C \to \infty$ and $S \to \infty$

By increasing $C \to \infty$, a cutoff-invariant definition of contacts is obtained. This corresponds to including as contacts any unshadowed atoms regardless of distance. As mentioned above, any protein interior contacts so generated are likely enthalpically important since an absence of mediating atoms is entropically unlikely. $N(M_\infty^1)$ increases by 1.7 over M_6^1 to 2.9, but for a slightly larger shadow size, $N(M_\infty^{1.2})$ only increases by 0.3 contacts per atom over $N(M_6^{1.2})$. The amount of free space rapidly decreases for

$S > 1$. These additional contacts generated with long cutoffs are dominated by atoms near the protein surface interacting through multiple waters. In order to separate out the desired interior contacts, we would need to introduce a burial parameter, and this is left to future studies.

A parameter-less contact map results from $C \to \infty$ and $S \to \infty$. Since an atom k can only shadow a contact between atom pair ij if $r_{ik}, r_{jk} < r_{ij}$, M_C^∞ only includes the nearest neighbor pairs, $N(M_\infty^\infty) \sim 0.2$. While M_∞^∞ can be used to find nearest neighbor pairs, nearly all interactions longer than 4.5 Å are excluded (Figure 3.4C) and it does not result in cooperative folding (data not shown).

3.3.2 Decoupling the Protein Geometry from the Contact Energy Distribution

If the contact potentials introduce additional excluded volume between native atomic pairs (Figure 3.2), different contact maps will have different amounts of excluded volume. To probe the effects of introducing additional excluded volume in the native contacts, the thermodynamics of CI2 was calculated (Figure 3.6A). Heat capacity (C_V) was compared for M_6^0 with varied native repulsive distances and a constant repulsive distance (r_{NC}) between non-native beads of 1.7 Å. For example, the black curve labeled "4Å" includes a Lennard-Jones-type repulsion (σ_{NC}) at 4 Å between all native pairs of 4 Å or larger, and at the native position for those closer than 4 Å. The C_V becomes sharper (more cooperative) with increasing native repulsion. Also, since the folded basin is being destabilized relative to the unfolded basin, the folding temperature T_F (*i.e.* the temperature at the peak in C_V) decreases. This excluded volume effect makes the Hamiltonian with Lennard-Jones contact potentials ("LJ") markedly more cooperative and less stable than the equivalent Hamiltonian with Gaussian contact potentials ("1.7Å").

The tendency of native excluded volume to alter cooperativity and stability has opposite thermodynamic behavior between Lennard-Jones and Gaussian potentials with M_4^0, M_5^0, and M_6^0 (Figure 3.6B). The Lennard-Jones potentials decrease protein stability since increasing the contact map cutoff C introduces more native contacts, and thus, more native excluded volume. The increased excluded volume decreases the entropy

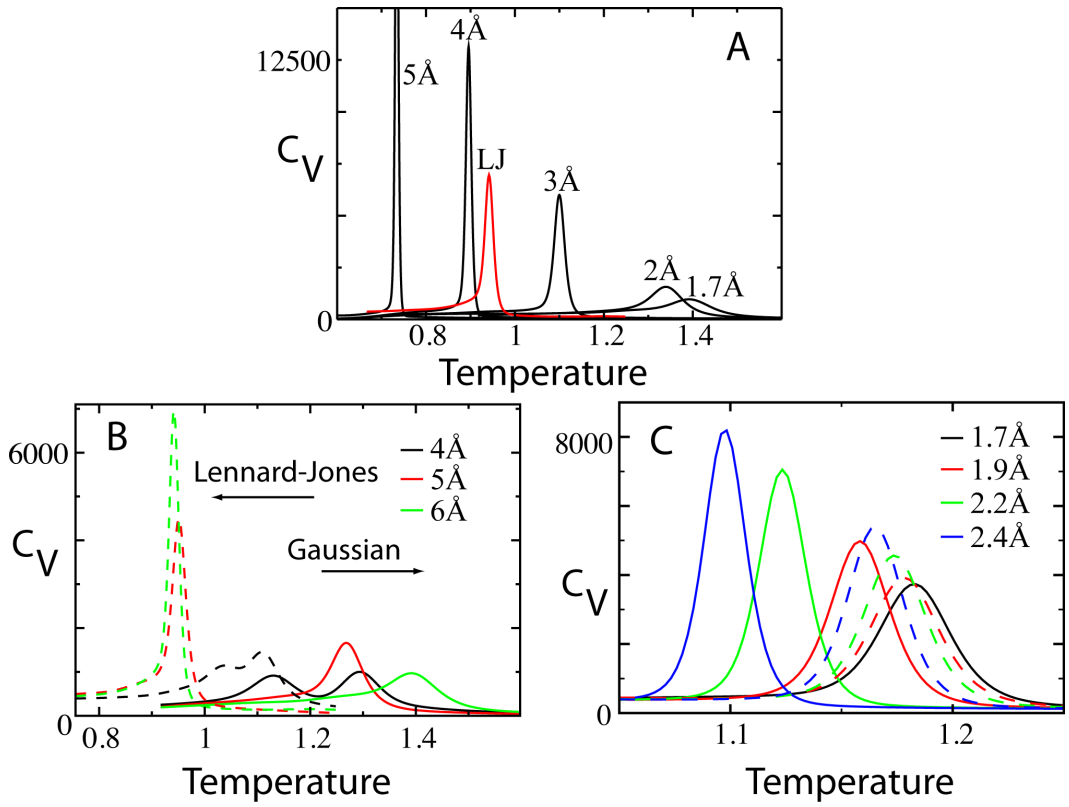

Figure 3.6: Excluded volume imparts cooperativity. (**A**) Heat capacity of the folding transition for M_6^0 and atoms of diameter 1.7 Å. As the excluded volume of only the native contacts is increased, both the folding temperature T_F and cooperativity are dramatically affected. The red curve "LJ" uses Lennard-Jones contact potentials. Though it would seem LJ should be less stable than the 4Å and 5Å curves, the tighter width of the Gaussian potential causes an addition destabilization beyond that of the additional native excluded volume in LJ. (**B**) After removing the extraneous excluded volume, increasing cutoff distance has the opposite effect on T_F. Increased cutoff results in increased cooperativity with Lennard-Jones, whereas the Gaussians show a maximum at intermediate cutoff. (**C**) The black curve denotes the Shadow map with standard parameters and Gaussian contact potentials. Dotted curves correspond to excluded volume altered only between non-native pairs. Solid curves correspond to excluded volume altered between all pairs. $\kappa_1 = 0.032$ for the standard Shadow map and $\kappa_1 = 0.018$ when the atomic repulsive size is increased to 2.4 Å. All data is from CI2.

of the native basin relatively more than the unfolded basin, and therefore decreases the stability of the native state. In contrast, the Gaussian potentials isolate the effects of changing the contact energy distribution by maintaining a constant native excluded volume of 1.7 Å between all atoms. The Gaussian potentials show an opposite behavior, protein stability is increased as C is increased. Now the dominant effect is the increased entropy of the native state as more contacts are introduced. This stabilizing effect will be further discussed in the next section.

Independent of the contact map and contact potential, the repulsive size of the atoms also affects the folding cooperativity and stability. The Gaussian potential allows us to also isolate the effects of changing the atomic repulsion between either only the non-native atomic pairs or all atomic pairs (Figure 3.6C). The Shadow map (M_6^1) is used, non-native excluded volume is controlled by r_{NC} (Equation 1.4), and native excluded volume is controlled by r_{ex} (Equation 3.2). Increasing the size of all the atoms has a similar effect as only increasing the repulsion between native pairs (Figure 3.6A), where κ_1 increases and stability decreases. Since the native state is denser and has more atomic collisions than the unfolded configurations, the entropy of the native basin is relatively smaller when the atoms are larger. Somewhat surprising is that increasing the repulsive size of only the non-native interactions follows the same trend as well. While one might surmise that a larger excluded volume of non-native interactions lowers the entropy of the unfolded basin more than the folded basin, the destabilizing effect shows that in fact non-native interactions are more frequently encountered in the *folded* basin of the all-atom model. This is opposite to the effect seen in a closely related coarse-grained C_α-model (*60*). While the C_α atoms in the backbone are similarly constrained to their native positions in both the coarse-grained and all-atom models (*35*), the all-atom model introduces close-packed side chains that encounter many non-native atomic collisions. In addition to the close atomic distances, there are less native restraints on each atom since M_6^1 gives 1.2 contacts per atom versus 2.6 contacts per residue. We note that the ability to encounter non-native collisions is enhanced by the smooth energy landscape. Previous work showed that an all-atom SBM makes comparatively more non-native contacts in the folded basin than an explicit solvent transferable potential like OPLS (*35*).

3.3.3 Shadowing Tends to Increase Folding Cooperativity

In this section, we explore the effects on folding cooperativity of changing the largest energetic component of the SBM, the native contact map. The native contact map defines the distribution of tertiary stabilizing energy. The effects of changing this distribution are isolated by using a Gaussian contact potential that maintains a constant excluded volume across contact maps (Figure 3.2). The model proteins are three small, fast-folding globular proteins: B-domain of protein A (BDPA), chymotrypsin inhibitor 2 (CI2), and the sh3 domain of c-src kinase (SH3). These three proteins, which we studied previously (*35*), are well studied both experimentally (*93–95*) and theoretically (*13, 97, 98*) and represent simple to complicated folds, respectively (*99*). Differential scanning microcalorimetry has shown that small globular proteins like BDPA, CI2, and SH3 fold cooperatively in a two state manner with singly peaked heat capacity at the folding transition and $\kappa_1 < 0.05$ and $\kappa_2 > 0.95$ (*95, 122, 128*).

We find that using a contact map generated with geometric occlusions consistently increases folding cooperativity relative to a map generated with a cutoff distance. Figure 3.7 shows the heat capacity calculated for two sets of contact maps and three proteins. The first set of maps used a direct cutoff (M_0^4, M_0^5, and M_0^6), while the second set have $S = 1$ (M_1^4, M_1^5, and M_1^6). In every case, the map with $S = 1$ has a smaller κ_1 than the corresponding cutoff map (Table 3.1). In addition to consistently higher folding cooperativity, the thermal stabilities for $S = 1$ vary little in the same protein ($<$5%) and between proteins ($<$10%). The Shadow map (M_6^1) dependably gives folding temperatures near 1.2 for globular proteins. Proteins (PDB codes) not in Figure 3.7 that have been folded with the default all-atom SBM are 3MLG, 1RIS, 2A3D, and 2EFV, and have folding temperatures of 1.12, 1.21, 1.18, and 1.15, respectively.

The thermodynamics of the cutoff contact maps shows some interesting features. First, as C increases the protein becomes more thermally stable seen by the movement of T_F. This is because of two effects: 1) as C increases the contacts are on average wider and 2) the stabilizing energy is more diffuse. Both of these effects increase the entropy of the native state and hence increase stability. The cutoff map contact distance distribution is skewed towards C, and therefore, the average native distance between contacts $\langle r_0^{ij} \rangle$ increases with C (Figure 3.4). A larger native distance produces a wider

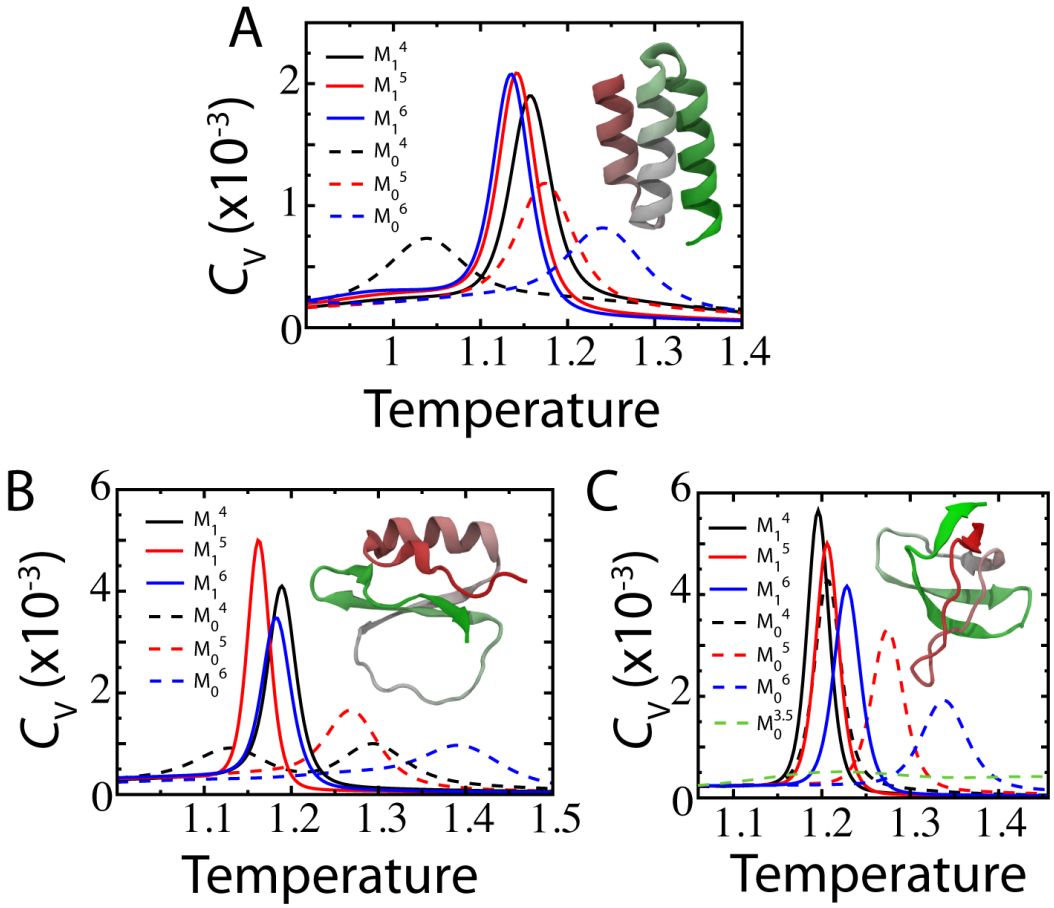

Figure 3.7: Heat capacity is consistent and folding is cooperative as the cutoff parameter C is varied with the Shadow algorithm. Three proteins are shown, (**A**) BDPA, (**B**) CI2, (**C**) SH3. Their native structures are shown as insets. The solid lines are shadowing maps with $S = 1$ and the dotted lines are cutoff maps ($S = 0$).

contact potential since $(r_0^{ij}) \propto \sigma_{ij}$ (Equation 3.3). The energy distribution becomes more diffuse because at higher C there are more total contacts. The total energy available for the contacts is held fixed, so each contact has a smaller share of stabilizing energy. Interestingly, κ_2 does not follow the trend of κ_1 as $C \to 6$ Å, instead staying constant or even increasing. This implies that the increase in κ_1 is not from the introduction of intermediate states, but rather the slow conversion of well defined unfolded and folded ensembles. Second, there is a minimum cutoff distance, below which the protein no longer makes a cooperative transition. Remarkably, at $C = 4$ Å CI2 becomes a 3-state folder, the heat capacity shows a thermodynamic intermediate (*120*) (Figure 3.7B). At $C = 3.5$ Å SH3 resembles a downhill folder (Figure 3.7C). Last, since cooperativity vanishes at both low and high C, there is a peak in cooperativity at an intermediate range of 4 Å $< C <$ 5 Å. The thermal stability T_F of the most cooperative cutoff maps is near the stability of the Shadow map. This property, that the contact maps with similar stabilities have similar cooperativities, was seen to hold among the many variations of M_C^S tested for this paper. It implies that there is an optimal temperature to have a cooperative transition. Perhaps, the Shadow map consistently achieves this stability and thus is cooperative.

3.3.4 Dynamics of RNA and Macromolecular Assemblies

There are many new and exciting areas ripe for exploring through the lens of energy landscape theory, the foundation upon which structure-based models are built. These theoretical tools are already being applied to the study of RNA folding (*50, 129, 130*) and the dynamics of molecular machines composed of either protein, such as kinesin (*33*), or RNA-protein complexes like the ribosome (*4, 131*). Of particular interest is the connection between folding and function, and the role of order-disorder transitions in the control of these molecular machines (*132*). In this section we look beyond protein folding, and show that Shadow contact maps provide a consistent treatment for heterogeneous systems, and thus, a solid framework for addressing the geometrical features of molecular machines.

Table 3.2: Comparison of cutoff versus shadowing contact maps in RNA systems. The RNA helix is helix P2 from the SAM-I riboswitch (*50*) minus the 4 turn residues.

RNA helix		
	M_4^{0*}	M_6^1
Watson-Crick contacts (WC)	137	61
Base-stacking contacts (BS)	232	35
Total contacts (All)	480	190
WC/BS	0.59	1.74
BS/All	0.48	0.18
Ribosome		
	M_4^{0*}	M_6^1
$E_{\text{rna}-\text{rna}}$ contacts	77529	57355
$E_{\text{pro}-\text{rna}}$ contacts	8045	14771
$E_{\text{pro}-\text{pro}}$ contacts	15053	28510
$E_{\text{rna}}^{\text{D}}$ (per RNA atom)[a]	0.37	0.37
$E_{\text{pro}}^{\text{D}}$ (per protein atom)	0.26	0.26
$E_{\text{rna}}^{\text{C+D}}$ (per RNA atom)[b]	1.18	1.01
$E_{\text{pro}}^{\text{C+D}}$ (per protein atom)	0.64	0.97
$\sigma_{E^{\text{C}}}/\overline{E^{\text{C}}}$ in RNA atoms[c]	0.27	0.26
$\sigma_{E^{\text{C}}}/\overline{E^{\text{C}}}$ in protein atoms	0.63	0.49

[a] Dihedral energy in RNA (protein) divided by the number of RNA (protein) atoms

[b] Total contact and dihedral energy in RNA (protein) divided by the number of RNA (protein) atoms.

[c] E^{C} represents the contact energy per atom by residue.

RNA Contact Maps

RNA has three main types of contacts, Watson-Crick (WC) base-pairing, base-stacking (BS) interactions, and tertiary backbone contacts (Figure 3.5C). WC pairs are the hydrogen bonding interactions between complementary RNA bases (*i.e.* A·U and G·C). BS interactions refer to π-π stacking: attractive, non-covalent interactions between the aromatic rings of stacked bases that are adjacent in sequence. Maintaining proper energetic balance between these interactions will be important to the performance of RNA models.

Short-range cutoff contact maps have been shown to overweight the BS interactions relative to WC pairs and tertiary contacts. To maintain a proper balance between secondary and tertiary structure in the study of the folding of the mRNA SAM-I riboswitch with a SBM (*50*), BS interactions were scaled by a factor of $\frac{1}{3}$ when using a 4 Å contact map M_4^0. Here, we denote the cutoff contact map including scaled BS interactions as M_4^{0*}. The over-stabilization of BS interactions in M_4^0 arises from the geometry of closely packed rings. As seen in Figure 3.5C, atoms 1 and 2 are each within 4 Å of five atoms in the adjacent stacked base. This is the case for every atom in the ring, and for every stacked ring in the riboswitch. Interestingly, if geometric occlusion is considered, due to the close packing, the over-counting is avoided. Introducing shadowing with $S = 1$ Å, atoms 1 and 2 each have only a single stacking interaction.

Shadowing naturally gives rise to the approximate $\frac{1}{3}$ scaling in stacking interactions. Table 3.2 compares M_4^0 to M_6^1 for an RNA helix. Base-stacking interactions relative to WC pairs are decreased by a factor of $0.59/1.74 = 0.34$. Relative to all contacts, the BS contacts are decreased by a factor of $0.18/0.48 = 0.37$, which is in surprising agreement with the previous conjecture of $\frac{1}{3}$ (*50*). Thus, the energy distribution between M_4^{0*} and M_6^1 are similar in RNA, but vary by a factor of 2.5 in the number of total contacts. The heat capacity of the isolated 16 residue P2 helix of the SAM-I riboswitch (*50*) was calculated for the two contact maps (Figure 3.8). M_6^1 is more cooperative, while M_4^{0*} is more stable. These trends are in line with those observed in Section 3.3.3 for proteins. M_6^0 in protein is the analog of M_4^{0*} in RNA, it introduced an excess of contacts that increased stability, while the shadow map M_6^1 was less stable but more cooperative. So, while the Shadow map gives a reasonable distribution of energy within RNA, to be

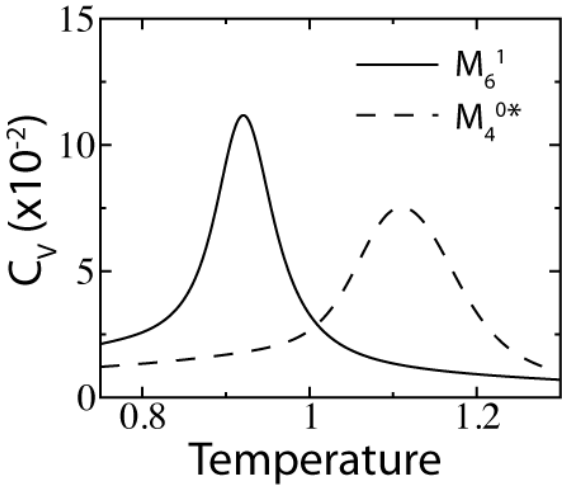

Figure 3.8: Folding an RNA hairpin with the all-atom SBM (16 residue helix P2 of the SAM-I riboswitch, PDB code: 2GIS). The heat capacity C_V is compared between two native contact maps. M_6^1 refers to the normal Shadow contact map and M_4^{0*} refers to a cutoff map where the energy of all base-stacking interactions are scaled by $\frac{1}{3}$. κ_1 is 0.10 for M_6^1 and 0.15 for M_4^{0*}.

applied to RNA-protein assemblies, the shared energy distribution with proteins must also be balanced.

Shadowing in Heterogeneous Assemblies

Tables 3.1 and 3.2 indicate that the atomic packing is very different between RNA and protein. In the RNA hairpin, M_4^{0*} has 150% more contacts than M_6^1, whereas in proteins, M_6^1 has $\sim 70\%$ more contacts than M_4^0. The regularity of base-stacking dominates the short-range contacts in RNA. Proteins, in contrast, have no regular residue packing since the amino acid side chains have a diversity shapes. This difference in packing causes short-range cutoff maps to skew the distribution of stabilizing energy in favor of RNA.

The contact energy per atom by residue E^{C} in the ribosome is shown in Figure 3.9. Even with the BS contacts scaled by $\frac{1}{3}$ in M_4^{0*}, the contact energy in RNA is double that in protein, $\overline{E^{\mathrm{C}}_{\mathrm{rna}}}/\overline{E^{\mathrm{C}}_{\mathrm{pro}}} = 2.1$, where $\overline{E^{\mathrm{C}}}$ is E^{C}_X averaged over all residues of type X. The Shadow map gives a much closer division, $\overline{E^{\mathrm{C}}_{\mathrm{rna}}}/\overline{E^{\mathrm{C}}_{\mathrm{pro}}} = 0.86$. Since RNA has a higher density of dihedrals than protein, if the dihedral and contact en-

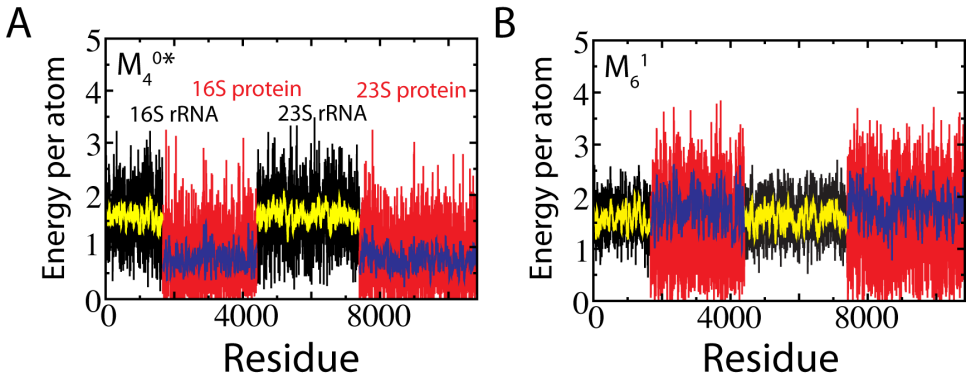

Figure 3.9: Distribution of native contact energy between proteins and RNA in the ribosome. Energy per atom of residue i is calculated as $\varepsilon_C N_C^i / N_A^i$ where N_C^i is the number of atomic contact pairs involving atoms of residue i and N_A^i is the number of atoms in residue i. The previously used (*4, 50*) cutoff contact map M_4^{0*} (**A**) is compared to the Shadow contact map M_6^1 (**B**). Black regions correspond to the 16S and 23S RNA, and the red regions correspond to the proteins associated with the 16S and 23S.

ergy are summed and compared, the Shadow map gives an equal distribution of energy, $\overline{E_{\text{rna}}^{\text{C+D}}} / \overline{E_{\text{pro}}^{\text{C+D}}} = 1.0$. This feature is desirable when simulating heterogeneous molecular assemblies. Fluctuations in the ribosome for two different contact maps, M_4^{0*} and M_6^1, are compared to the fluctuations predicted from the experimental B-factors (Figure 3.10A). For the 23S Ribosomal RNA the correlation between experiment and the SBM with the Shadow map is 0.78. On a smaller scale, to highlight the variability between contact maps, fluctuations are shown for three proteins at $\sim 0.75 T_F$ (Figure 3.10B). While the correlation is high between the between the two maps, deviations can be seen. Future work will have to explore how robust these fluctuations are since deviations in fluctuations between related proteins have been predicted to have functional consequences (*52*).

3.4 Conclusions

We have proposed a general algorithm for generating atomically-grained contact maps called "Shadow" (Figure 3.1). This algorithm enables sufficient contact cutoff distances to capture atomic contacts across structural waters or heavy metals that are not

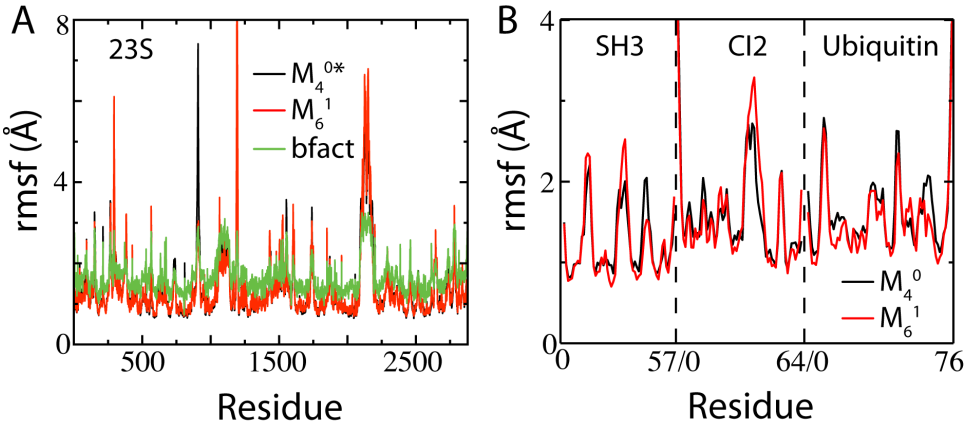

Figure 3.10: Structure-based models capture near-native-state fluctuations of both small proteins and large macromolecular assemblies. (**A**) Comparison of simulated root mean squared fluctuations (rmsf) of each residue in the 23S Ribosomal RNA (PDB codes: 3F1E, 3F1F) between the scaled cutoff map M_4^{0*} and the Shadow map M_6^1. Overlaid are the rmsf by residue, calculated using the experimental B-factors (using the isotropic approximation, rmsf $\equiv \sqrt{\langle r_i^2 \rangle} \sim \frac{3}{8\pi^2} B_i$, where r_i is the displacement of atom i and B_i is the experimental B-factor of atom i (*133*)). A residue rmsf is computed as the arithmetic average of its constituent atoms' rmsf. Overall rmsf agreed with the B-factors for M_4^{0*} at $T = 0.46$, and M_6^1 at $T = 0.71$. The discrepancy in stability is likely due to M_6^1 including double the Mg^{2+}-RNA contacts, which are modeled as harmonic restraints instead of Gaussian contact potentials. Pearson correlation r between M_6^1 and B-factors is 0.78. (**B**) Simulated rmsf of the C_α atoms for three globular proteins at $T = 0.9$. The overall agreement is very close ($r > 0.9$) between two different contact maps, M_4^0 and M_6^1.

explicitly represented, without introducing contacts between atom pairs that one does not wish to model, specifically, those that have an intervening atom. The Shadow algorithm initially considers all atoms within a cutoff distance $C = 6$ Å and then, controlled by a screening parameter $S = 1$ Å, discards the occluded contacts. We showed that this choice of contact map is not only well behaved for protein folding, since it produces consistently cooperative folding behavior, but also desirable in exploring the dynamics of macromolecular assemblies since it distributes energy similarly between RNAs and proteins despite their disparate internal packing.

The study of the connection between the contact distribution and folding cooperativity highlighted that many components of the SBM Hamiltonian affect cooperativity, especially the geometric components. We showed how the Lennard-Jones contact interaction mixes the geometric and energetic parts of the Hamiltonian by changing the excluded volume of native interactions. By decoupling the geometric and energetic parts with the Gaussian contact potential, it was clear that the increased cooperativity obtained through additional Lennard-Jones native contacts was caused by the extra excluded volume. Further, the decoupling showed that the innate cooperativity of the Shadow map was purely an effect of the contact energy distribution. In the case of CI2, the energetic effect of changing contact maps from M_6^0 to M_6^1 decreased κ_1 from 0.12 to 0.032, while the geometric effect of increasing the diameter of the atoms from 1.7 Å to a more realistic 2.4 Å brought κ_1 even further down to 0.018 (experimental range was $\kappa_1 < 0.05$). Other studies have shown that, for example, excluded volume (*134, 135*), backbone stiffness (*35, 100*), contact potential width (*e.g.* σ_{ij} in Eq. 3.3) (*60, 136, 137*) and many-body effects (*47, 135, 138*) affect the cooperativity of protein folding models.

Structure-based models will continue to be an important tool in the characterization of molecular machines and macromolecular assemblies. They are baseline models that can be used to fully discern the role of biomolecular geometry. Going forward, all-atom structure-based models employing Shadow contact maps provide a general framework for exploring the geometrical features of biomolecules, especially the connections between folding and function.

3.5 Acknowledgments

Chapter 3, in part, appears in *Journal of Physical Chemistry B*, (2012, in press), Noel, Whitford, Onuchic. The dissertation author is the primary investigator and author of the paper. JKN wishes to thank Shachi Gosavi for helpful discussion and enthusiasm. This work was supported by the Center for Theoretical Biological Physics sponsored by the National Science Foundation (NSF) (Grant PHY-0822283) and NSF Grant NSF-MCB-1051438 to JNO. This research was also supported in part by the NSF through TeraGrid resources provided by TACC under grant number TGMCB110021. JKN was supported in part by an NIH Molecular Biophysics Training Grant at UCSD (Grant T32 GM08326). PCW was funded by a LANL Directors Postdoctoral Fellowship.

Chapter 4

SMOG@ctbp: Simplified Deployment of Structure-based Models in Gromacs

4.1 Introduction

Molecular dynamics simulations have benefited from years of research on computer algorithms constructed with one goal in mind: speed. Molecular dynamics suites like Gromacs (*74*), NAMD (*76*) and Desmond (*139*), package all the necessary algorithms to run stable molecular dynamics and the ability to scale the calculations to thousands of processors. These packages have made homegrown molecular dynamics codes built to run structure-based models (SBM) obsolete. To reap the benefits of the many features these software suites offer, we have ported SBMs to Gromacs, Structure-based MOdels in Gromacs (SMOG). The SMOG@ctbp web server (http://smog.ucsd.edu) is available to facilitate the creation and use of SBM to investigate the dynamics of proteins, RNA, and DNA. Both the C_α (*13*) and the all-atom (*35, 50*) models discussed in this thesis are available. These SBMs represent baseline models upon which additional complexity can be added by the user.

The purpose of the web server is twofold. First, the webtool simplifies the process of implementing a well-characterized structure-based model on a state-of-the-art, open source, molecular dynamics package, Gromacs (*74*). Second, the tutorial-like format helps speed the learning curve of those unfamiliar with molecular dynamics. A

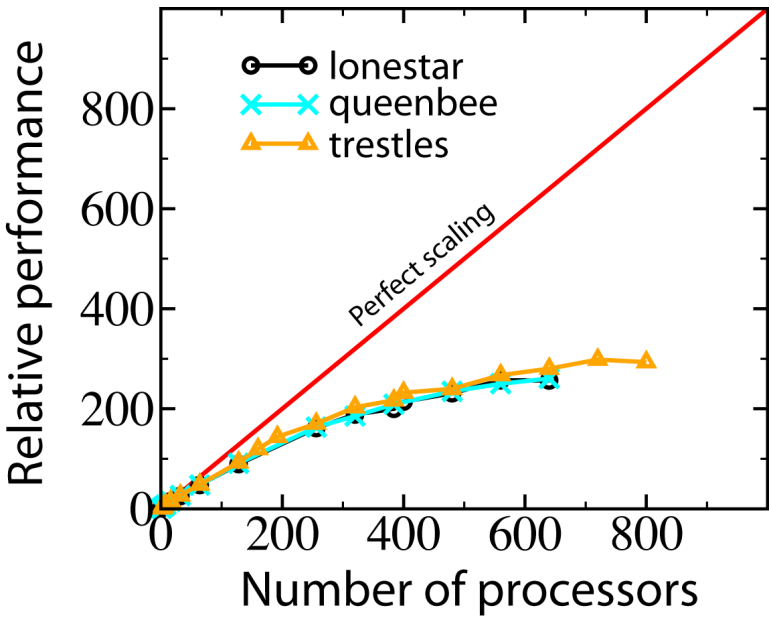

Figure 4.1: Performance of an all-atom structure-based simulation with Gromacs version 4.5 for a ribosome with 142,196 atoms. The system scales up to 200 processors before significant performance loss. Due to large amounts of empty space inside the ribosome, this represents a lower bound on potential scalability. Three different supercomputers are compared with similar results.

webtool user is able to upload any multi-chain biomolecular system consisting of standard RNA, DNA, amino-acids, and a small library of ligands in PDB format and receive as output all files necessary to implement the model in Gromacs. Gromacs has the flexibility necessary to implement an efficient and highly scalable SBM (Figure 4.1). In this chapter we briefly describe the web server interaction and features. Further explanation, tutorials, and FAQs can be found on the web server itself.

4.2 Implementing a Structure-based Model in Gromacs

4.2.1 Web Server Interaction

The main purpose of the web server is to create the input files necessary to simulate a biomolecular system with a SBM in Gromacs (Figure 4.2). A PDB structure that is uploaded from the user's computer is the only required input. While most PDB struc-

tures can be directly downloaded from the PDB database and used with the webtool, users should verify that the PDB file conforms to the guidelines described below and in the supplementary information. A valid PDB file has a `TER` statement (left justified) in between each chain and an `END` statement (left justified) at the end. The following residues are supported by the webtool:

- Protein residues: All standard 20 amino acids (3 letter codes used).
- RNA residues: CYT or C, GUA or G, URA or U and ADE or A.
- DNA residues: DG, DC, DA, DT.
- Ligands: SAM (S-Adenosylmethionine), GNP (Gpp(NH)p), ATP, ADP, AMP, FUA (Fusidic Acid), GTP, GDP
- Ions: BMG (Bound MaGnesium ions), ZN

Upon request, additional ligands may be supported. Ions are given excluded volume and they interact through harmonic potentials (e.g. $k(x-x_0)^2$, where $k = 1.0\ \varepsilon/\text{Å}^2$). For calculating which interactions are included with ions, the same contact rules are used as for ligands.

The webpage where the PDB file is uploaded is entitled "Prepare a Simulation" and is where all user input is obtained. Beyond uploading a PDB file, the web server interface allows the user to customize some basic parameters of the SBM Hamiltonian:

1. The level of graining. It can be varied between all-atom and C_α.
2. The contact map. The user can upload a native contact map or generate a map by choosing either the cut-off or Shadow algorithm. The contact map algorithms are based on the all-atom geometry, thus PDB files that lack some heavy atoms must be manually inspected to ensure proper performance. Contact maps are compared in Chapter 3.
3. The distribution of stabilizing energy. It can be varied between contacts, backbone dihedrals and side chain dihedrals. This is explored in Chapter 2 (*35*).
4. The mass and size of atoms.

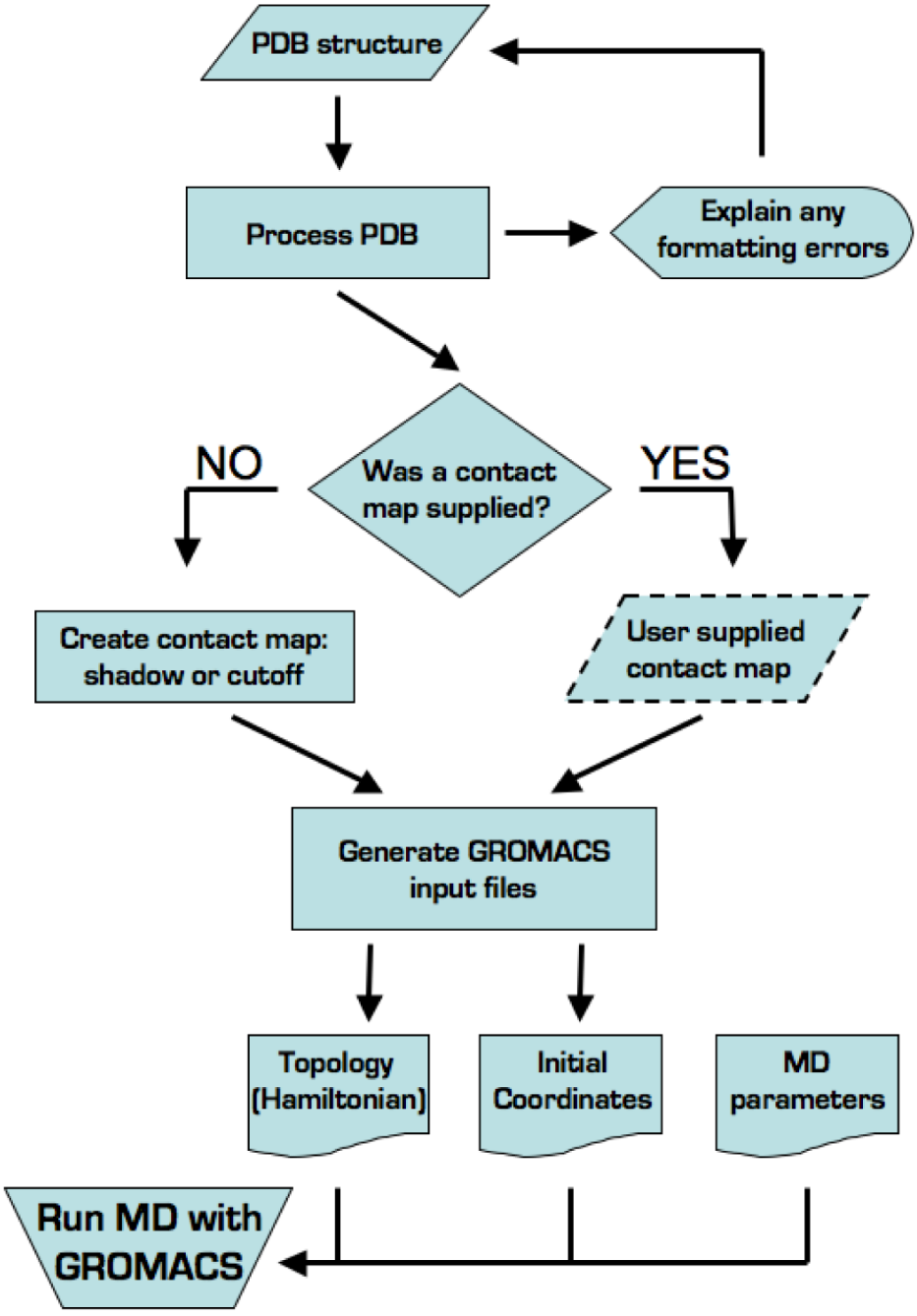

Figure 4.2: Flowchart explaining the logic of the SMOG@ctbp web server.

5. The buffer space. The space between the system and the simulation box is an important parameter. Improved performance and effective parallelization in Gromacs depends on periodic boundary conditions being employed. When using the "dynamic load balancing" features of Gromacs, excessive volumes of empty space can lead to poor scalability. Though, if the simulation box size is too small the system can interact with its image. While the default 10Å buffer is sufficient for many simulations, for folding, the box size should be nearly the linear length of the molecule.

After uploading a PDB file, inspecting the above parameters, and pressing the "Submit" button, the web server will either return a link to the completed output or return an error message describing any formatting inconsistencies. The completed output is a tarball containing:

1. Gromacs coordinate file: the initial structure corresponding to the provided PDB structure; shifted such that the box starts at the origin (`.gro`).

2. Gromacs topology file: describes all the atomic interactions in the SBM Hamiltonian (`.top`).

3. Gromacs index file: convenient for manipulating structures with multiple chains (`.ndx`).

4. Native contact map: if Shadow is selected (`.contact`).

5. Web server output: contains any non-fatal warnings and messages (`.output`).

4.2.2 Molecular Dynamics with Gromacs

In order to run molecular dynamics the user must have access to a compiled Gromacs 4 distribution. The Gromacs source code can be found at http://www.gromacs.org. The topology file and coordinate file, along with a molecular dynamics parameter settings file (`.mdp`) are sufficient to run the SBM in Gromacs. A suggested `.mdp` is available on the web server. Example output for an SH3 domain is shown in Figure 4.3. See the web server or the supplementary information for a brief tutorial highlighting the relevant Gromacs syntax and things to consider.

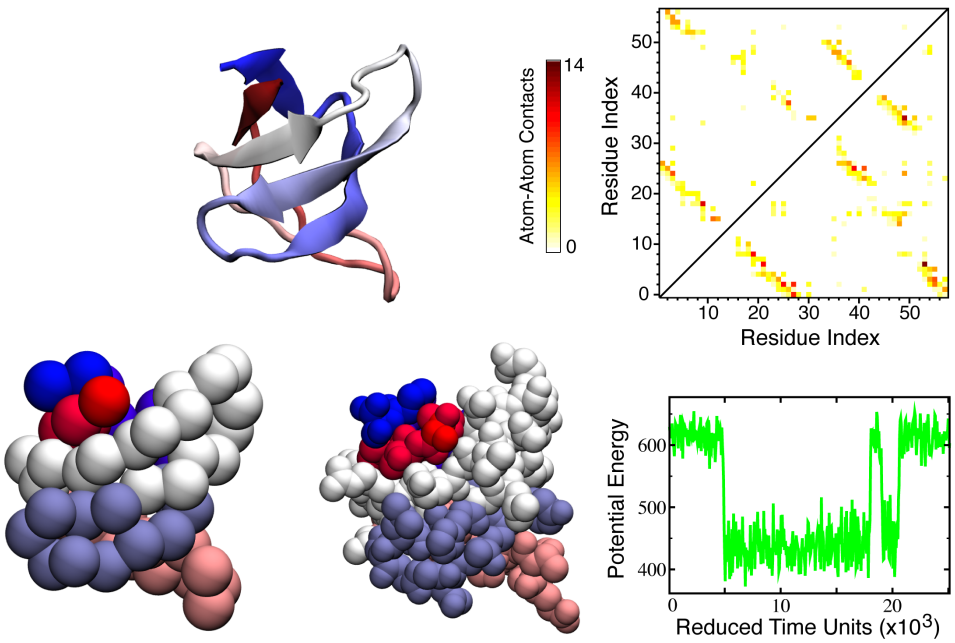

Figure 4.3: Structure-based model of the folding of the SH3 domain. PDB code: 1FMK. **Top Left**: Cartoon representation of SH3 domain. **Bottom Left**: C_α model geometry. **Bottom Middle**: All-atom model geometry. **Top Right**: Contact map for SH3. Upper triangle shows 4 Å cut-off and lower triangle shows Shadow. Coloring is by number of atom-atom pairs per residue-residue contact. **Bottom Right**: Folding of 57-residue SH3 domain at constant reduced temperature $\tilde{T} = 1.0$ with the all-atom model (Gromacs temperature 120 K). Residues 84-140 taken directly from 1FMK.pdb and submitted at SMOG@ctbp with default parameters and Shadow contact map. MD parameters file taken from the web server example.

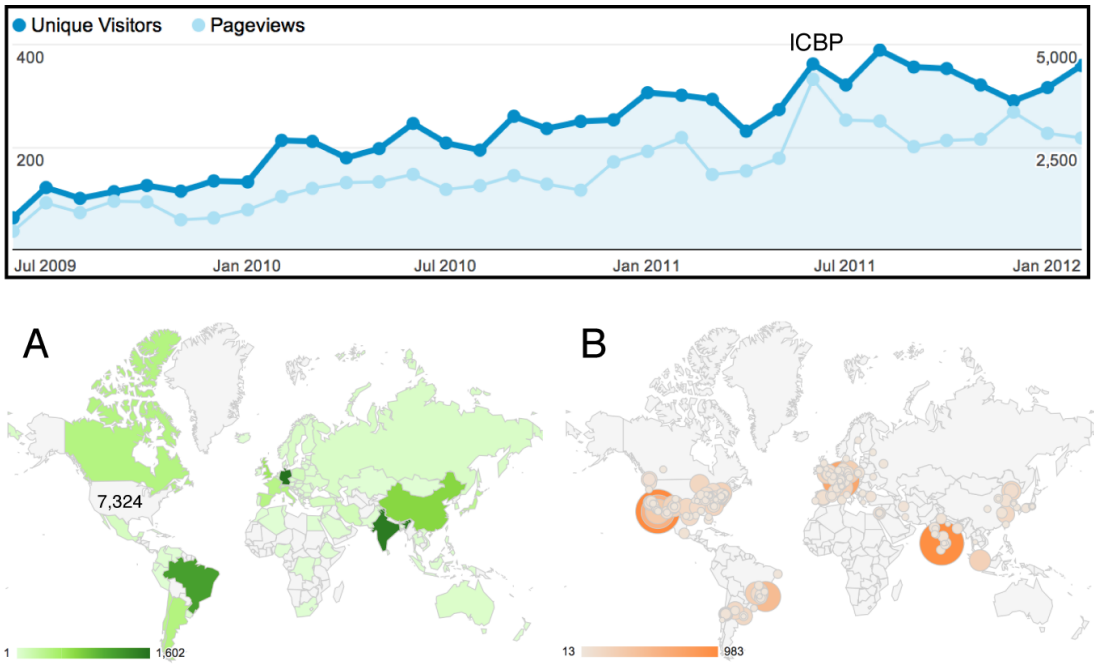

Figure 4.4: Usage statistics for the SMOG@ctbp web server. The top panel shows total page views and unique visitors, each per month, since the web server went online in July 2009. 6162 topologies have been downloaded in that time. (**A**) Total visits by country. The United States, with 7324 visits, is left out of the color scale to allow contrast. (**B**) Total visits by city. The fifteen highest users are: (1) La Jolla, CA, 1292, (2) Bangalore, India, 956, (3) Karlsruhe, Germany, 695, (4) San Diego, CA, 570, (5) Rio de Janeiro, Brazil, 466, (6) State College, PA, 402, (7) Berlin, Germany, 308, (8) Changchun, China, 284, (9) Durham, NC, 273, (10) Kent, OH, 273, (11) Troy, NY, 269, (12) Sao Jose do Rio Preto, Brazil, 261, (13) Singapore, 234, (14) Los Alamos, NM, 230, (15) Madrid, Spain, 219.

4.3 Conclusion

In this chapter we described SMOG@ctbp, a web server that creates the necessary files to simulate a SBM in Gromacs from a provided PDB structure. The web server is receiving wide use from the simulation community (Figure 4.4). The all-atom SBM represents a baseline model that the user is welcome to augment and explore with system dependent details, *e.g.* electrostatics or non-native interactions. The possible applications of SBM go beyond equilibrium and kinetic molecular dynamics. A SBM is a starting point for any study where the overall geometry of the biomolecules is maintained, *e.g.* fitting crystallographic structures into cryoelectron microscopy maps (*140*) and predicting protein-DNA complexes (*55*). Hopefully SMOG@ctbp will enable users to conceive of more new and exciting applications of SBM.

4.4 Acknowledgements

Chapter 4, in part, appears in *Nucleic Acids Research*, (2010), Noel, Whitford, Sanbonmatsu, Onuchic. The dissertation author was the primary investigator and author of the paper. We would like to thank the New Mexico Computing Applications Center for computing time on the Encanto Supercomputer. The work was supported by the Center for Theoretical Biological Physics, sponsored by the National Science Foundation [PHY-0822283, NSF-MCB-0543906], the LANL LDRD program, National Institutes of Health [R01-GM072686], and National Institutes of Health Molecular Biophysics Training Program at University of California at San Diego [T32GM08326 to J.K.N.].

Chapter 5

The Free Energy Landscape of a Trefoil-Knot Protein: Slipknotting upon Native-Like Loop Formation

5.1 Introduction

Protein structures have been observed with several complex folding motifs including knots and slipknots. These include non-trivial topologies containing 3_1, 4_1, 5_2 and 6_1 knots (*141–145*). While the mechanism by which these proteins manage to reliably fold from a disordered linear polypeptide into complicated topologies is still largely a mystery, energy landscape theory is starting to provide us the key to resolve this challenge. In a minimally frustrated, funnel-like energy landscape, one expects that native contacts are on average favorable and dominate over non-favorable non-native ones (*6, 8*). Topological constraints imposed by the existence of a native knot radically alters the funneled landscape. Many routes are barred from reaching the native state due to the obstacle imposed by the knot. Forming a knot requires intricate crossings of the polypeptide, any one made incorrectly leads to an unknotted protein or a wrong chirality. Therefore at first sight the problem of folding knots appears perplexing, but there is no reason to doubt that clues will be found in the native structure itself. Here it is shown how an all-atom structure-based model, which is dominated by native attrac-

tive interactions, is sufficient to uncover the energy landscape and folding routes of the smallest knotted protein.

Experiments have shown that a knotted protein can fold from pre-knotted denatured states (*146*). These experiments monitored the kinetic refolding of homodimeric α/β-knot methyltransferase, YibK, which contains a 3_1 knot, and showed that mutations in the native knotted region slowed the early stages of refolding of the denatured, but still knotted, protein (*146, 147*). More importantly, a recent experiment monitored the folding of YibK immediately post-translation, where the initial protein state is guaranteed to be unknotted. This experiment showed formation of the native knotted state in an *in vitro* environment devoid of chaperones (*148*). Therefore it is now clear that (in at least some cases) the information necessary to fold a knot is wholly contained in the amino acid sequence. Theoretical investigations by Sulkowska *et al.* showed that the native state of YibK is kinetically accessible with a native-biased coarse-grained model through a knotting mechanism where the protein has significant native structure when the knot is created (*49*). In this scenario one of the termini threads a native-like loop through a *slipknot* intermediate, a collapsed configuration where the terminal polypeptide makes native contacts and adopts a hairpin-like configuration while threading (for detailed description of slipknot topology see also (*149*)). An alternative knotting mechanism is a *plug* motion, an extended configuration analogous to threading a needle. This scenario is seen during coarse-grained kinetic simulations of YibK by Wallin *et al.*, but required introduction of attractive non-native interactions around the knotted region (*150*). Here we adopt a different approach from these two previous studies. By applying an all-atom model to a smaller protein, the thermodynamics of folding knots can be studied, rather than only the kinetics.

Beyond tying knots, these proteins must also be able to reliably avoid topological traps, kinetic traps on the landscape whose solution would require chain interpenetration or "chain crossing." Chain crossing is not allowed, thus the connectivity imposes a topological constraint. A simple solution is to evolve a sufficiently sequential or polarized folding route that orders the topological crossings correctly. Typical small and intermediate size proteins fold by a collection of multiple converging pathways towards the native state, a folding funnel. Deviations from this ideal picture can arise from constraints

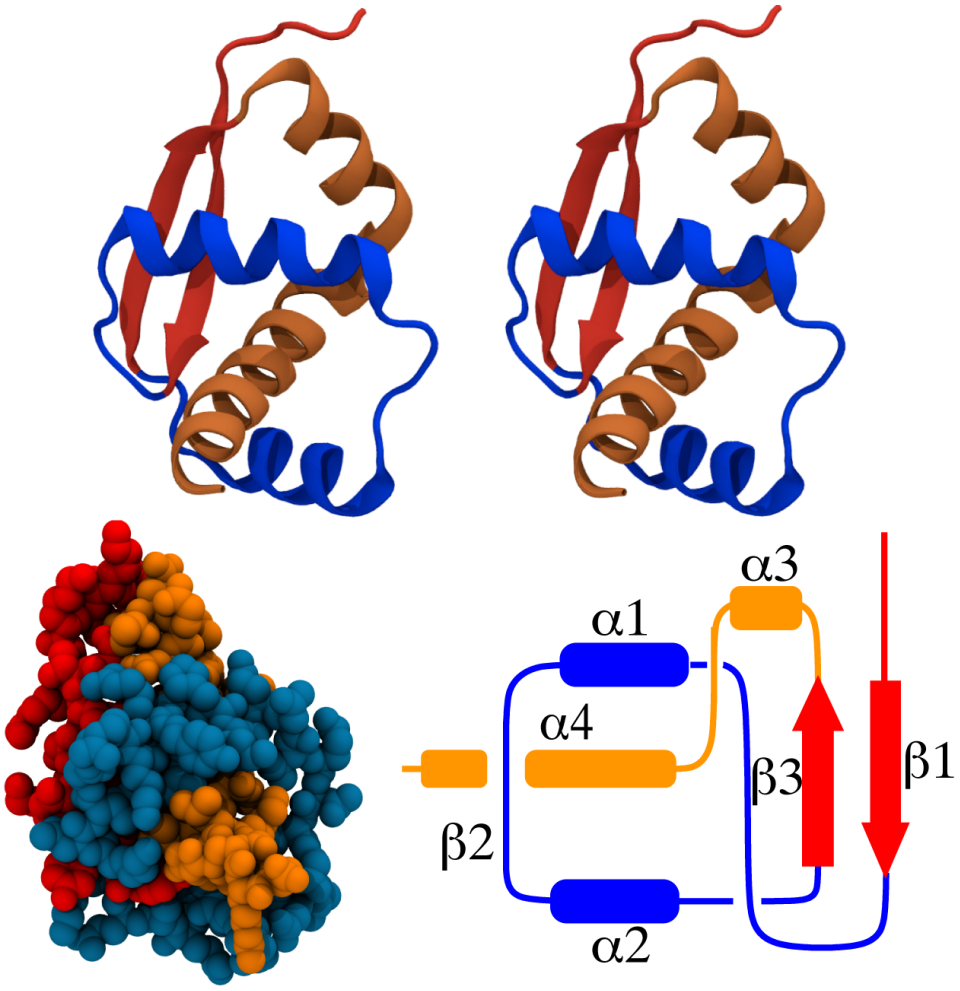

Figure 5.1: Structure of the smallest knotted protein. **Top**: Stereo projection of a cartoon view of the crystal structure (PDB code: 2efv). The coloring corresponds to the schematic view. **Bottom Left**: Van der Waals sphere representation showing the geometry of the all-atom protein model. **Bottom Right**: Schematic view showing the crossings leading to the 3_1 topology. β_2 forms a β-sheet with its image in the homodimer. Only the monomer is shown for simplicity.

imposed by geometric problems related to steric collision or by entropic effects related to chain connectivity (*13, 14, 35, 98, 151*). Such constraints cause the folding mechanism to be dominated by just a subset of the possible folding pathways. The constraints on the folding funnel become severe in the case of knotted proteins, where topological issues become important. The covalently connected nature of the polypeptide backbone makes some of the folding pathways for a knotted protein sterically impossible. Even under this more restricted situation, it has been shown that, just as with traditional proteins, a funnel-like landscape dominated by native interactions manages to fold topologically complex proteins (*49*). The inherent geometric constraints in the structure are able to guide the necessary pre-ordering of chain crossings. This earlier work, however, was qualitative. The kinetic folding simulations had a low success rate of reaching the native knot (<5%), instead becoming trapped in misfolded states in most trajectories. Here we employ an all-atom model to gain a more quantitative understanding of the topological effects. While various kinds of geometrical bottlenecks have been popularly referred to in the protein field as topological constraints, in this paper, we limit the use of the term "topological" to the stricter mathematical meaning related to crossings of the polypeptide chain (*152*).

A natural choice for this study is the smallest knotted protein, MJ0366, from Methanocaldococcus jannaschii (*145, 153*). The protein crystallized as a homodimer and its trefoil knot structure is shown in Figure 5.1. The existence of a knot in a small globular protein gives a unique opportunity to investigate the full process of folding. An all-atom structure-based protein model is employed (*35*) to provide the first in depth characterization of the folding mechanism and free energy landscape of a knotted protein. Despite that MJ0366 is a small globular protein and might therefore be expected to show simple two-state folding behavior, the thermodynamic and kinetic data show a three state system: unfolded, native-like loop formed, native-knotted structure. The correct knotting follows after native loop formation, and never as a random knotting event. This polarized folding pathway is robust to even high concentrations of monomer. The slipknot and plug knotting routes are both populated at folding temperature (Figure 5.2). The comparison between a coarse-grained model and an all-atom model bridges previous work with current results. The comparison illuminates the topological constraints

imposed by the knot and shows how the protein avoids topological trapping.

5.2 Results and Discussion

5.2.1 Thermodynamically Meta-stable State Precedes Knotting

The knotted α/β protein MJ0366 has 82 residues and creates a 3_1 (trefoil) knot (Figure 5.1). We characterize the knot position by monitoring its *depth* (*49*), distance along the sequence K_N, K_C to the knot, respectively from the N-terminal and C-terminal. In the case of a slipknot we additionally monitor depth of a slipknot loop (*149*), which is located between K_N and K_C. In the crystal structure the knot begins at Asn15 and ends at Ala70, hence, $K_N^0 = 15$ and $K_C^0 = 12$ where the superscript denotes the native value. The knot covers $82 - K_N^0 - K_C^0 = 55$ residues, where N is the total number of residues. Helices α_1 and α_2 and their linkers create the loop through which the C-terminal threads. The loop is twisted and its ends are glued by the β-strands. In its native state the knotted domain forms a dense hydrophobic core, largely composed of α_3-α_4 packing with β_1.

Unbiased constant temperature molecular dynamics simulations were performed to obtain the free energy landscape for the monomer structure at folding temperature T_F, the temperature where the unfolded and folded ensembles have equal free energy minima. Each simulation visits both the folded/knotted state and the unfolded/unknotted state. In total, 100 folding/unfolding transitions are included. The progress of folding was monitored with the reaction coordinate Q, the fraction of native residue contacts formed. An advantage of the theoretical approach is the ability to monitor a new coordinate, the precise position of the knot, using the KMT algorithm (*154*). The correctness of the structures was checked by both the number of native contacts and the position K_N, K_C of the knot along the sequence. Figure 5.2 shows the free energy as a function of Q. Three states are clearly seen. Upon folding, the protein must first overcome a barrier to form the β-sheet which defines the loop. Second, after the loop twists correctly, the protein overcomes a larger barrier by threading the C-terminus through the loop. Overall the folding barrier is $5k_BT$. The specific heat shows a single peak at T_F.

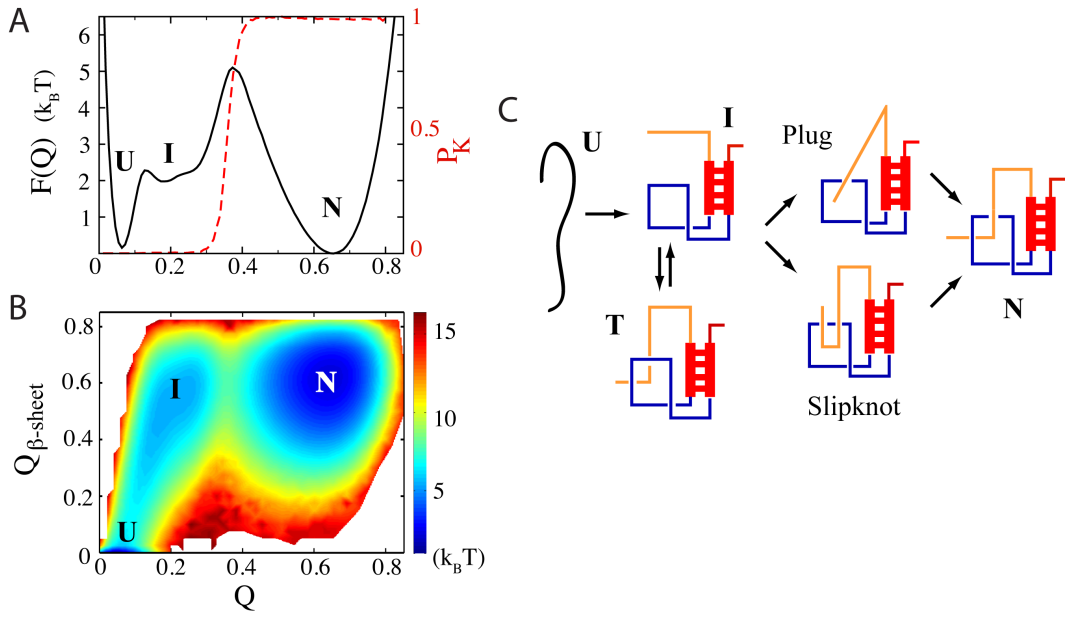

Figure 5.2: Folding routes of a knotted protein. (**A**) Free energy $F(Q)$ is plotted as a function of the global reaction coordinate Q at T_{F}. Three states are clearly seen and labeled **U**, **I** and **N**. The minimum at low Q is the unfolded ensemble, the broad minimum at $Q \approx 0.2$ corresponds to the formation of the β-sheet and correct twisting, and the broad minimum at high Q is the native ensemble. The red dotted line shows the probability P_{K} of finding a knot as a function of Q. The high barrier at $Q = 0.4$ is associated with the entropic cost of forming the native knot. (**B**) Free energy as a function of two coordinates, Q and Q_β, the number of native contacts formed in the β-sheet. The β-sheet is formed prior to the transition state which defines the loop for the C-terminus to thread. (**C**) Folding routes in a topologically frustrated system. A loop with the correct chirality must exist, **I**, which is then threaded by the C-terminus. If the loop twists incorrectly or the C-terminus forms native contacts out of order, the protein may be trapped, unable to proceed to the native state without chain crossing. **T** is an example of such a conformation seen during folding. From state **I** the protein follows two routes to the native ensemble, either plug or slipknot. Cartoon representation of the energy landscape is presented in Figure 5.5.

5.2.2 Non-specific Knots, Native Knots and Malformed Knots

In the case of strings and homopolymers (*155, 156*), there are no preferred locations for nucleation of knots and knots are equally likely to be found anywhere along the sequence. The existence of native structure differentiates a protein from these traditional model systems where knotting is considered. On a funneled landscape, a protein progressively forms native structure, which implies that a protein is more likely to nucleate a knot in a location containing a loop in the native structure. A knot formed by threading a loop consisting of native structure is called a *native* knot, whereas a knot threading a non-native loop is called a *non-specific* (random) knot.

A reasonable criterion to distinguish between these two cases is to define a native knot as when (1) at least one knot crossing differs from the native value by less than ten residues, for example, $K_C^0 - 10 < K_C < K_C^0 + 10$ for the C-terminal crossing, and (2) projection of the knot in the plane gives the native chirality (see Section 5.5.5 for subtleties of the 4_1 knot). This definition includes the following examples of non-specific knotting (Figure 5.3): (*i*) shallow knots, which could easily appear in a long protein with a deep native knot, (*ii*) knots tighter or deeper than native, (*iii*) knots located on the opposite side of the sequence relative to the native position, (*iv*) an incorrect knot (e.g. of the wrong chirality), which would have to untie prior to correct folding. In the unfolded ensemble of MJ0366 we find a non-specific knot less than 0.1% of the time and never find that a non-specific knot nucleates folding. Most of the non-specific knotted configurations were of types *i* and *iii*, however one *iv* case was also found. This is consistent with theoretical evidence that folding nucleation by non-specific knots is entropically unlikely in proteins (*49, 150*). This process should have a barrier with a large entropic contribution since there is little energetic stabilization until the native environment forms around the knot.

Kinetic traps on the folding landscape, whose solution would be a chain crossing, are called *topological traps*. Two types of topological traps can be defined, (a) a non-specific knot of type *ii*, *iii* and *iv* or (b) a malformed topology with some correct crossings but at least one incorrect crossing; these malformed topologies would include cases where the knot is missing. In type *a*, non-specific knots of type *ii* and *iii* must jump along the sequence to find the correct native position, but this process might be

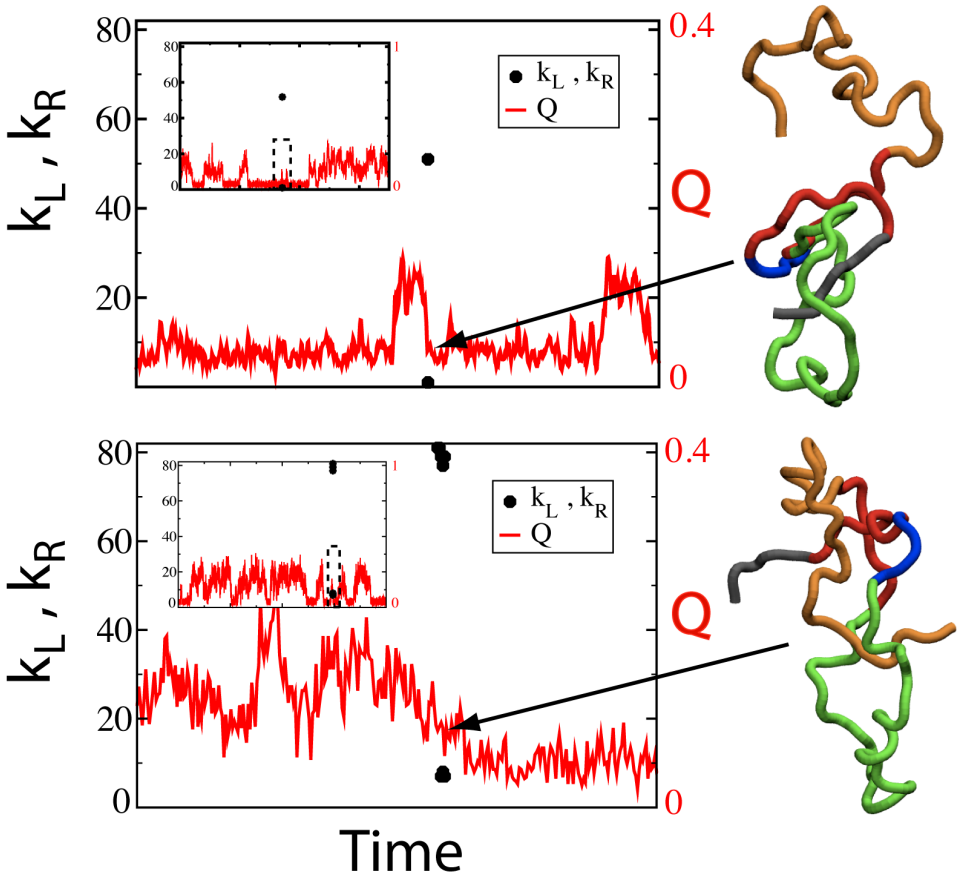

Figure 5.3: Non-specific knots during folding. The red line shows $Q(t)$ and the black dots show the locations of the left k_{L} and right k_{R} ends of a knot. $k_{\mathrm{L}} = K_{\mathrm{N}}$ and $k_{\mathrm{R}} = N - K_{\mathrm{C}}$.The knot is only made transiently in both cases. Both knots are shallow knots, and their structures are shown to the right of the figures. Both of the knots appear in the unfolded ensemble, $Q < 0.1$. The inset shows a longer view of the folding trajectory. **Top**: A non-specific knot (type *iv*) with the wrong chirality. The N-terminal (grey) is threading the C-loop (green). This knot must untie to reach the native state. **Bottom**: Another example of a non-specific knot with the wrong chirality. In this case, the C-terminal (orange) is threading the C-loop (green) in the opposite direction. Rare examples of non-specific knots with correct chirality can also be found (not shown) but are never seen to nucleate folding.

prohibitively slow (*149, 157, 158*). Non-specific knots of type *iv* must backtrack completely. *Backtracking* is the process of breaking a subset of native contacts in order to fall further down the folding funnel (*14*). Traps of type *b* make subtle crossing mistakes and these errors can persist to structures with high Q (Figure 5.5). Since chain crossing is forbidden, large backtracking excursions are required to correct the crossings. Traps of type *a* were transient at T_F and were not observed during kinetic folding. Traps of type *b* are also transient at T_F, but become prevalent at lower temperatures. The specific trapped structures and how side chain packing affects their population is discussed in detail in later sections.

5.2.3 Folding Mechanism of a Knotted Domain

To fold a protein with a 3_1 knot there are three distinct folding routes (*49*): A native-like knot where either the N-terminal protein forms a native-like loop for the C-terminus to thread, or the C-terminal protein forms a native-like loop for the N-terminus to thread, or a non-specific knot with little native structure which then coalesces to the native knot. Figure 5.2c diagrams the folding mechanism at T_F. The protein is never seen to form a knot outside of state **I** that continues to the native state. Non-specific knots of type *ii* are never seen and those of type *i* are exceedingly rare. Types *iii* and *iv* are trapped configurations that have to unfold before proceeding to the native state. Since types *i* and *ii* are not observed to fold native knots, the route to the native state must be through native loop formation. There are two possible loops to form, loops to be threaded by either the C-terminus (C-loop) or N-terminus (N-loop). The C-loop is defined by contacts between Asn15 and Tyr49, a loop length $\Delta L = 34$ residues, and is anchored by the β-sheet. The N-loop can be defined by either Ala42 and Val79, $\Delta L = 37$, or Leu28 and Leu74, $\Delta L = 46$. The N-loop is not anchored by any secondary structure, but is stabilized by the packing of the helices. Since the loop lengths are approximately the same, the choice of which loop is formed first and subsequently threaded is likely determined by energetics. The C-loop is stabilized by $\sim$90 contacts whereas either N-loop is stabilized by only $\sim$60 contacts. Figure 5.2b clearly shows that the intra-β sheet contacts required to form the native C-loop structure occur before the transition state. C-loop formation leads to an unstable intermediate with a free energy barrier of $2k_BT$.

Simply forming the β-sheet is not enough to define the native C-loop, the loop must be twisted correctly. It is possible to twist the β-sheet 360 degrees and arrive at a nearly native configuration that differs only in the crossing near Asn15 and Tyr49 (Figure 5.5c). This minor structural difference gives a topologically incorrect and potentially trapped structure lacking the knot. Forming the twist takes the protein from the metastable intermediate at $Q \sim 0.2$ to the plateau at $Q \sim 0.25$. It is essentially a barrierless process but it must occur before the transition state.

After the C-loop is formed and correctly twisted, the C-terminus must overcome both an entropic barrier and topological barrier to reach the knotted fold. The entropic barrier arises as the C-terminus trades its conformational freedom for the formation of the hydrophobic core, and the topological barrier arises from the excluded volume of the loop and the need to thread the C-terminus through it. This topological barrier manifests as an increased entropic barrier, since the number of routes to the native state are limited by the constraint. Forming the hydrophobic core could be a driving force towards forming the knot since it consists mostly of contacts between β_1 and the threaded α_4. The two possible mechanisms for threading are either a plug or a slipknot intermediate (see Figure 5.2c). The plugging mechanism appears when the C-terminus is the first part of the protein chain to thread the C-loop. Native contacts are not formed until the C-terminus reaches its native position. This mechanism happens through random fluctuations of the C-terminus impinging on the C-loop. The slipknotting mechanism occurs when part of the protein chain (near the C-terminus) threads the C-loop but doubles back so the protein chain stays unknotted, a hairpin-like configuration. The slipknot is stabilized by forming native hydrophobic core contacts between α_3-α_4 and β_1, between Phe10 and Ile63 for example. As the slipknotted intermediate is stabilized by native contacts, the C-terminus has time to thread the loop. The C-loop's ability to accommodate this bulky configuration is facilitated by the flexible five residue chain β_2 and the melting of the C-loop helices. At T_F the protein folds by the plugging mechanism 55% and by the slipknot mechanism 45%. The coexistence of these two pathways was also seen in the folding of YibK (*150*). The equilibrium between these two mechanisms, though, is highly dependent on the length of the threaded C-terminus, and is discussed in the next section.

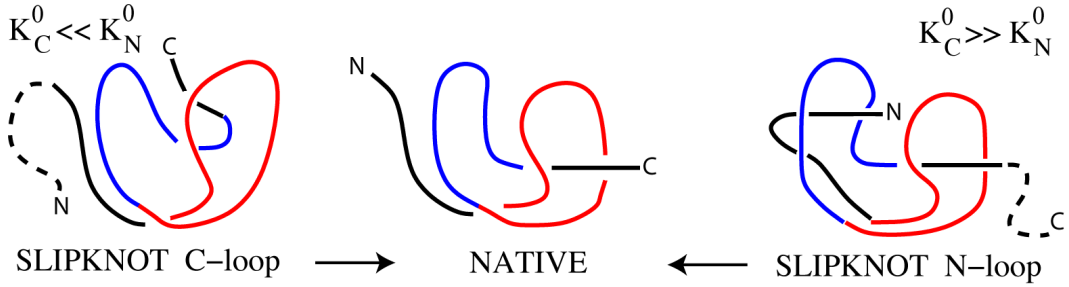

Figure 5.4: Two possible native loops for threading the trefoil knot. K_N^0 and K_C^0 denote the length of the native threaded N-terminal and C-terminal, respectively, and dotted lines show their extension. The dominant route depends on the relative values of K_N^0 and K_C^0. **Center**: Native structure with the two loops highlighted in red and blue. **Left**: Folding via C-loop. **Right**: Folding via N-loop.

5.2.4 Slipknotting is a General Knotting Mechanism

To investigate how the knotting mechanism is affected by the depth of the knot, folding of the MJ0366 structure with an extended C-terminal helix was studied and is summarized in Table 1. Since the actual sequence of MJ0366 used for crystallization has five additional residues at the C-terminus, it is known that MJ0366 is able to knot with a longer C-terminal tail. The simulated C-terminal helix was extended using the five additional residues indicated in the crystal data, increasing K_C^0 to 17. The only native contacts added were local helix contacts in the extended region, no additional contacts with the rest of the protein were added. When kinetic folding is performed at $0.96T_F$, the extended structure folds via the slipknot route 99% of the time. The plugging mechanism is dependent on putting the C-terminal chain into a precise configuration to slide across the loop, which becomes less likely as the entropy of the C-terminal chain increases with additional residues. At the same time, the additional entropy of the extended C-terminus stabilizes the native contacts that support the slipknot intermediate.

A comparison between a coarse-grained (C_α) model which has a single bead-per-residue with the all-atom (AA) model allows more direct connection with previous simulations on knotted proteins (*49, 145, 150*). Two of these studies (*49, 145*) suggested, using a completely funneled C_α folding model, that slipknotting is the likely folding route for a more deeply knotted protein. The AA simulations of the extended C-terminal

Table 5.1: Relative population of folding routes (in %) for the AA and C_α model. Results for the PDB structure and the five residue extended structure are shown. Subscripts N and C denote slipknotting or plugging knotting route via the N- and C-loop, respectively. For example, slip$_C$ denotes slipknotting by threading the C-loop.

	AA model		C_α model	
Route	PDB	PDB+5	PDB	PDB+5
slip $_C$	68	99.3	38	16
slip $_N$	2	0.35	1.5	72
plug $_C$	28	0.35	57.5	3
plug $_N$	-	-	-	1.5
non-specific	-	-	1.5	1.5

tail MJ0366 corroborate this claim. For completeness, kinetic folding simulations of a C_α model at $0.96T_F$ of both the PDB structure and an extended C-terminal structure were performed. The observed folding routes are shown in Table 1. The folding of the unextended structure is reminiscent of the unextended AA model where the protein folds via either slipknotting or plugging through the C-loop. Interestingly, a novel route is seen in the extended structure, slipknotting via the N-loop (Figure 5.4). As the C-terminus is extended, the difficulty of threading it through the C-loop increases relative to the N-terminal threading the N-loop, opening up a new kinetically accessible route. This route is not seen in the AA model, likely because the addition of side-chains makes the N-terminus too bulky to fit through the much tighter N-loop. Although, upon extending the C-terminus beyond ten additional residues, this route may become accessible in the AA model. In summary, the deeper knots follow the same slipknot mechanism, though the protein may switch the threading terminus depending on the depth of the knot at the two termini. This is possible in the 3_1 topology due to the approximate symmetry between the N-loop and C-loop.

5.2.5 Topological Traps on the Folding Landscape

As explained previously, malformed knots with subtle crossing mistakes are topological traps. These traps are regions of configuration space that can be close to native-like states in energy, but are topologically distant since they can only be directly

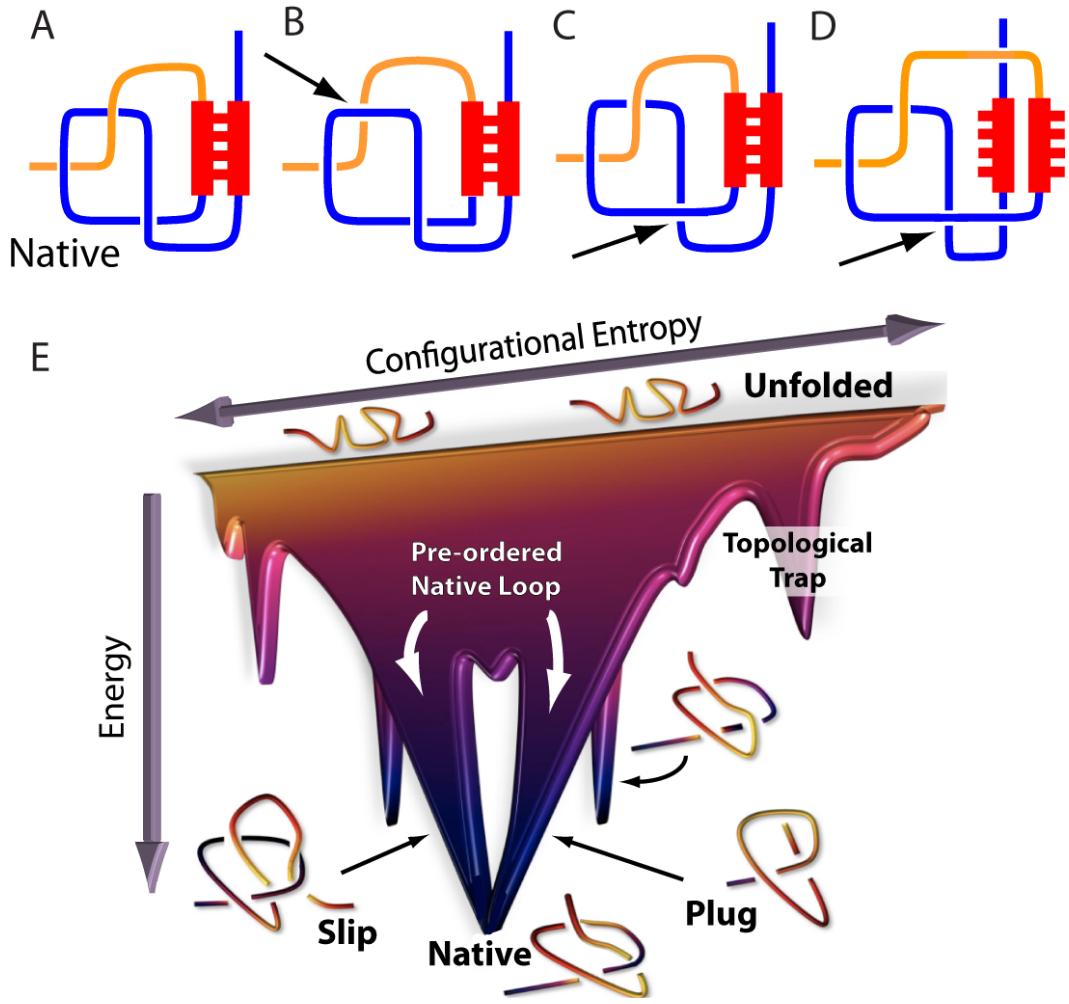

Figure 5.5: Energy landscape of a knotted protein and possible topological traps for a 3_1 knot. (**A**) The native fold. (**B-D**) Examples of topologically trapped structures. Backtracking must occur for these configurations to reach the native knot. The arrows denote the incorrect crossings. Configuration (**B**) is the most prevalent all-atom trap while (**C/D**) are most prevalent in C_α. Configuration (**D**) is not observed in all-atom. Configuration (**C**) is the transition state for the less-accessible slipknot via N-loop route (see Figure 5.4). (**E**) Funneled landscape. The protein is biased to a pre-ordered intermediate that has the relevant polypeptide crossings correct and contains a loop in the native position. α_4 threads the loop via parallel pathways, the plug pathway narrower as it is entropically disfavored. The empty space emphasizes the bifurcation in the landscape. Due to threading, not all routes are allowed. Topologically trapped states exist on the landscape and can differ by only a single crossing. Thus, these traps can be very deep in energy. Since chain crossing is not allowed, they are disconnected from the native state and must backtrack to fold correctly. The higher energy traps correspond to panel (**C**) and the symmetric traps close to the native structure correspond to panel (**B**). In the structure-based model these traps are transient at T_F, but can become kinetic traps at lower temperature.

connected to the native state through chain crossing. Trapped configurations are forced to backtrack in order to reach the native fold.

The folding landscape at T_F is smooth, because the thermodynamic data showed no evidence of long lived trapped states. Since physiological temperatures are below T_F, new features in the landscape can arise from the altered competition between energy and entropy. Collapsed states become more favorable, which impacts topologically frustrated systems that must be able to easily backtrack from the frustrating conformations for efficient folding. As the temperature is lowered, the time for escape from these conformations, the backtracking time, is greater. Investigating folding below T_F can ascertain which topological traps might become important at lower temperatures.

Kinetic folding simulations of MJ0366 were performed, starting from a random unfolded conformation and quenched to temperatures $0.96T_F$, $0.91T_F$, $0.86T_F$. They are summarized in Figure 5.6c,d. As the temperature is decreased the folding time also decreases as the competition between energy and entropy favors an increasingly compact ensemble. If the temperature is decreased far enough, the mean first passage time to reach the native ensemble τ_{mfpt} begins to increase as the protein spends more time in topological traps. At $0.96T_F$ the time spent in traps is negligible compared to τ_{mfpt}, while at $0.91T_F$ 3% of the trajectories visit a trap and at $0.86T_F$ 14% of the trajectories visit a trap. The average time spent in the topological traps increases since the barriers to backtracking are increasing. The topology of the most common trap is shown in Figure 5.5b. The C-terminus makes native hydrophobic core contacts without threading the loop. A second topological trap is seen at $0.86T_F$, the β-sheet forms with the incorrect chirality for the loop. The C-terminus can thread the incorrectly twisted loop and make most of its native contacts, even though the overall topology is trivial. This configuration is shown in Figure 5.5c. These traps, along with others, exist at T_F. They simply have much shorter lifetimes.

5.2.6 Addition of Side Chains Reduces Topological Trapping

A closer look at the comparison between the C_α model and the AA model highlights the role of the geometry in discriminating folding routes. The results show the C_α model is more prone to topologically trapped structures than AA. To quantify the

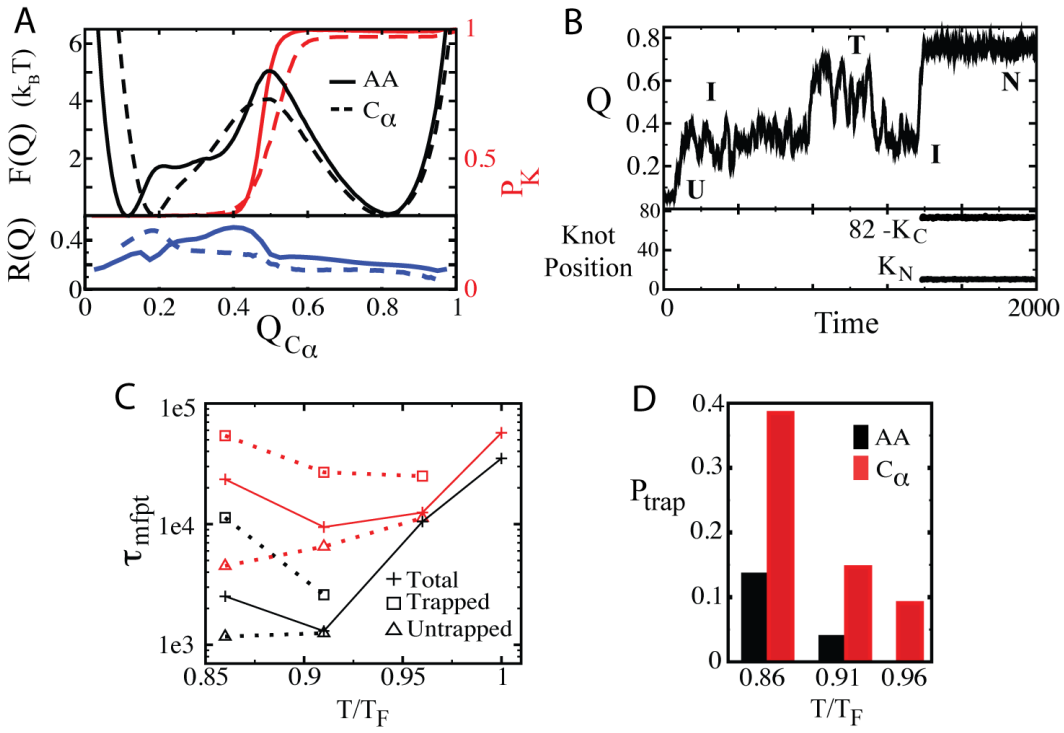

Figure 5.6: Specific side-chain packing reduces topological frustration. (**A**) Free energy as a function of global parameter Q_{CA} are compared between the all-atom(AA) model (dashed) and a one-bead-per-residue(C_α) model (solid). The C_α model has a much less defined shoulder around $Q_{\mathrm{CA}} = 0.25$ and has an unfolded basin corresponding to a formed β-sheet. The probability P_{K} of a knot is shown in red. The knot is formed more gradually along Q_{CA} in the C_α model. Also note that the folded basin is more likely to be unknotted in the C_α model. (bottom) A comparison of the route measure $R(Q)$. The AA model peaks at the transition state where the knot is formed while the C_α model peaks near the unfolded state. The protein configurations leading to the transition state in the AA model are fewer, leading to fewer topological traps. (**B**) An example kinetic folding trajectory with the AA model at $T = 0.91T_F$. The protein spends time in the looped intermediate state before falling into a topological trap **T**. The protein must backtrack before reaching the native state. K_{N} and K_{C} denote the N and C-terminal depths of the knot, respectively. The knot only forms upon folding to the native state. (**C**) Mean first passage times τ_{mfpt} are shown for four different temperatures $0.86T_{\mathrm{F}}$, $0.91T_{\mathrm{F}}$, $0.96T_{\mathrm{F}}$ and T_{F} for the AA model(black) and C_α model(red). τ_{mfpt} is split between the trajectories that spend a significant time in a topological trap(squares) and those that do not(triangles). The overall τ_{mfpt} is shown(crosses). τ_{mfpt} decreases as temperature is lowered from T_{F} and reaches a minimum near $0.91T_{\mathrm{F}}$. Due to trapping the overall τ_{mfpt} is greater at $0.86T_{\mathrm{F}}$ than $0.91T_{\mathrm{F}}$. (**D**) The percentage of trajectories that become trapped P_{trap} for the AA model(black) and C_α model(red) at different temperatures.

ability of the AA model to avoid topological traps, kinetic folding of the C_α model and the AA model are compared. Figure 5.6d shows that the C_α trajectories fall into traps more often than the AA. All of the trapped C_α structures, but only $\approx$20% of the trapped AA structures, were of the types in Figure 5.5c,d. Of the few trapped AA structures, most are of the type shown in Figure 5.5b. The addition of side chains serves to break the symmetry in the C_α geometry, for example in the β-sheet (Figure 5.5c,d).

A clear difference between the two models is captured in Figure 5.6a by comparing route measure $R(Q)$ (*159*) along the folding pathway. $R(Q)$ quantifies the amount of available configuration space that is actually accessed during folding (see SI Appendix). A larger route measure signifies a smaller number of routes traversed during folding. Knot formation is the stage of folding where avoiding incorrect crossings is critical. The C_α model is seen to have a more diverse set of routes leading to the transition state. Also the smaller slope of the knot probability versus Q shows knot formation is less cooperative in the C_α model. The increased persistence length coupled to more precise atomic packing in the AA model imposes an energetic penalty on routes containing improper chain crossings and therefore reduces topological trapping. Due to the importance of correct packing, knotted proteins may be particularly sensitive to mutations in the crossing regions.

5.2.7 Dimerization Occurs After Knotting

Studying the process of dimerization is important as it could have an effect on the folding of the knotted structures. The question is whether the topology forces native-like monomers to fold before dimerization or whether the dimerization step could be coupled to the folding process as in so called obligatory dimers (*68*). YibK, a 3_1 knotted protein, has been shown experimentally to first fold to a native-like monomeric state before a slow dimerization step (*147*). MJ0366 has a similar homodimeric interface as YibK, both interfaces include the C-terminal helix directly involved in the knotted structure (Figure 5.7). MJ0366 has a shallower knot than YibK and a higher proportion of dimeric contacts than YibK, both of which could cause the dimerization to more greatly impact the folding of MJ0366.

The dimerization process was investigated by performing folding simulations

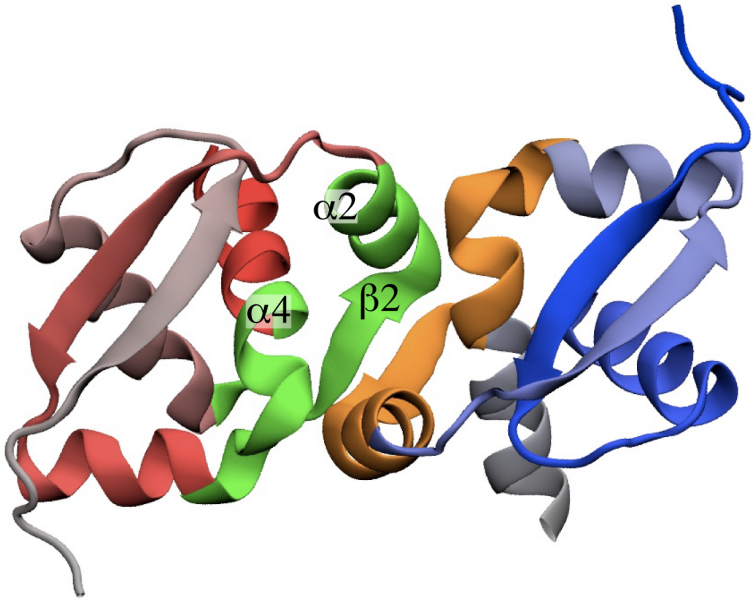

Figure 5.7: Structure of MJ0366 as a homodimer with β_2-β_2' forming the majority of the dimer interface along with α_2-α_4' and α_4-α_2'. The interface contacts (between green and orange) stabilize the monomer from unfolding since unknotting requires unthreading the C-terminal. The structure of the dimer also suggests folding the knot happens before dimerization. Formation of the dimeric interface will cause the C-loop to be tightened through the formation of β_2-β_2'. This impedes threading of the C-loop by the C-terminus. The formation of β_2-β_2' also puts side chains directly in the preferred route of slipknotting. Lastly, the dimeric interface will be less stable with non-native monomers since over half the interface is dependent on contacts from α_4/α_4', the last of the native structure to form.

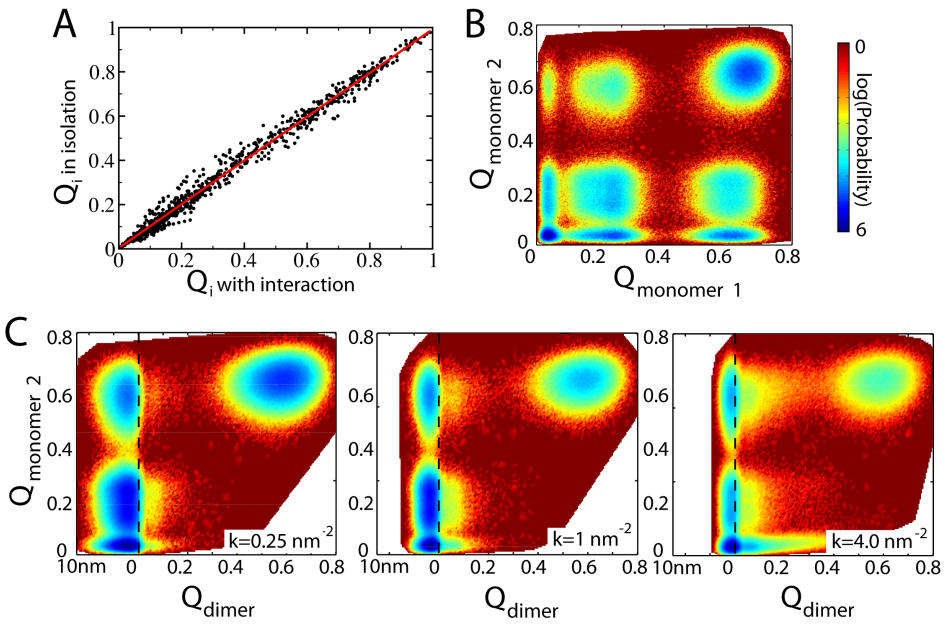

Figure 5.8: Folding with two monomers present. (**A**) Comparison between folding with and without a dimeric partner. Plotted is the correlation between the formation of all 799 atom-atom contacts Q_i in the transition state of a monomer folding in isolation, $k = 0$, or strongly interacting with another monomer $k = 4/\text{nm}^2$. The transition state is taken as $0.39 < Q_{\text{monomer}} < 0.41$. The correlation coefficient is greater than 0.995, which implies the monomer transition state is nearly unaffected by the presence of a strongly crowding dimeric partner. (**B**) Q of the monomers plotted against each other. (**C**) Folding of a monomer is shown versus the state of the dimer, for spring constant $k = (0.25, 1.0, 4.0)/\text{nm}^2$ at $T = 0.98T_F$. Q_{dimer} denotes the number to native inter-monomer contacts formed while Q_{monomer2} is the native intra-monomer contacts formed in one of the monomers. If no inter-monomer contacts are formed ($Q_{\text{dimer}} = 0$) then the distance between the closest inter-monomer native contact is recorded and plotted to the left of $Q_{\text{dimer}} = 0$ and the dotted black line. Notice that the transition state for the monomer is to the left of $Q_{\text{dimer}} = 0$. The monomer preferentially folds without making dimer contacts.

with two monomers present, starting with both unfolded. Contacts between the monomers in the crystal structure were included with the same strength as intra-monomer contacts. Results show that knot formation is unaffected by the presence of another monomer (Figure 5.8). The transition state ensemble is nearly identical whether folding in isolation or in the presence of another dimer. A correlation coefficient of 0.995 is seen between the transition state of an isolated monomer and two monomers held in close proximity by a harmonic spring constant $k = 4\varepsilon/\text{nm}^2$. Also, contacts between monomers are rarely formed in the transition states of the monomers, $Q_{\text{dimer}} < 0.05$ over a broad range of effective monomer concentrations. This remarkable result emphasizes the robustness of the proposed monomeric folding mechanism.

5.3 Conclusions

This study maps the full thermodynamic energy landscape of a knotted protein for the first time. We find that the folding is a thermodynamically three state system: unfolded, loop formation, native knotted structure. Below T_{F}, kinetic folding also follows the same three state mechanism along with increased prevalence of topological traps. An earlier C_α model for folding was shown to overestimate the importance of trapping. At folding temperature two parallel knotting mechanisms are observed, slipknot and plug. At lower temperatures and with an extended C-terminal tail, the mechanism switches exclusively to slipknotting, as the entropically limited plug pathway is suppressed. Further support for the slipknot pathway comes from the observation of slipknots in native protein structures (*144*). This folding route is consistent with previous work on YibK (*49*), so it will be instructive to apply the all-atom model to YibK and other larger knotted proteins.

Our results suggest some general features of folding knots in proteins. More deeply knotted proteins should tend towards creating knots through slipknotting. The viability of the slipknot route assumes that there is some pre-ordered native structure in the knotted domain to provide both scaffolding to anchor the slipknot intermediate and a native loop for the terminus to thread. The inherent geometric constraints in the knotted domain must be sufficient to ensure the correct ordering of crossings. It has been shown

that more complicated "closed" knots, 4_1, 5_2, 6_1, can be unknotted by switching a single crossing (*143, 145*). This is equivalent to the "open" knots being able to be tied through a single loop crossing event. These more complicated protein knots can therefore fit naturally into the mechanism of a pre-ordered domain coupled to a final native loop threading that creates the non-trivial topology. This would extend the folding pathway for the smallest knotted protein to all knotted proteins.

This scenario of pre-ordered native structure preceding knot formation is in direct contrast to folding through a random, non-specific knot which then coalesces into the native knot. The results from the extended C_α model show an underlying plasticity in the folding landscape as the protein can switch the threading terminus if the native geometry is perturbed, obviating a kinetically accessible non-specific knotting route. There are infrequent instances of non-specific knots forming in the unfolded ensemble in our simulations, but these events do not nucleate folding. Instead, the non-specific knots always backtrack. This result is somewhat surprising since the early formation of a knot would seem to surpasses the topological barrier. It shows there are still significant barriers to jumping the position of a random knot to the native position. This is in contrast to the behavior of knots in flexible random polymers.

These observations raise the important question of the dynamics of knots in denatured proteins. Suppose a knotted protein is rapidly denatured in an experiment before the knot can untie. Are the dynamics of the knot on the denatured polypeptide "polymer-like," where a knot is able to become either tightened or slide along the sequence, or are the dynamics "protein-like," where there exist large barriers (*62, 149*) to changing the knot's position? The answer to this question is crucial for interpreting experimental refolding data of knotted proteins and is currently under investigation.

5.4 Future Work: Detailed Simulations with Anton

Sophie Jackson's recent work (*148*) showing that YibK can fold a trefoil knot starting from a nascent (and therefore unknotted) polypeptide without chaperonins highlights the importance and relevance of continuing to investigate the folding process of isolated knotted proteins. She also showed that chaperonins increase the folding rate

of YibK, which implies that there are kinetic traps and intermediates along the folding landscape that can be rectified by chaperones. This chapter suggested that folding knots involves the protein terminal threading a native-like loop formed in a pre-ordered intermediate. The structure-based protein model neglects residual energetic roughness that may become important in exotic protein conformations such as threading polypeptide loops. To investigate the energetic roughness of threading we have performed detailed atomic simulations of these threading events in MJ0366, starting from conformations suggested by structure-based folding trajectories. The simulations were performed on the Anton supercomputer using the AMBER99SB forcefield, totaling 80 microseconds. Completed threading events, both plugging and slipknotting, starting from pre-ordered intermediates were observed with durations of 1 to 4 microseconds. The bulkier slipknotting conformation depended on large loop fluctuations to advance. Due to the lack of backtracking observed, the pre-ordered intermediates represent a significant local minimum on the energy landscape and show that from these intermediates knotting is a downhill process.

5.5 Methods and Notes

5.5.1 All-Atom Model

The all-atom and C_α models were presented in Chapter 1. In the AA model the Gaussian type contact potential (*60*) was used. The native contact map was generated by the Shadow algorithm. The C_α contact map is constructed from the all-atom contact map by including all residue pairs which have at least one atom-atom contact between them. The extended C-terminal residues without crystal coordinates were Glu, Gly, Glu, Arg, Ala. These residues were reconstructed as extending the C-terminal helix using the CHARMM package (*77*) in NAMD 2.6 (*76*). They had no tertiary native contacts. Thermodynamics data was obtained from constant temperature molecular dynamics and histograms from multiple temperatures were combined using the Weighted Histogram Analysis Method (*66*). All structures were visualized using VMD (*73*).

5.5.2 Reaction Coordinates

Q is defined as the fraction of native residues in contact (*35*). A residue contact is formed if any of their native atomic contacts are formed. A contact between atoms i and j is formed if $r_{ij} < 1.2r_{ij}^0$, where $r_{ij}^0$is the pair distance in the native state. Q_{CA} is a coarse-grained version of Q that defines a residue contact as formed if the C_α positions of residues i and j satisfy $r_{ij} < 1.2r_{ij}^0$. Q_β comprises the contacts between residues 7-12 and 49-54.

5.5.3 Route Measure

$R(Q)$ is normalized between 0 and 1 and is defined by

$$R(Q) = \sum_{i=1}^{M} \frac{(\langle Q_i \rangle_Q - Q)^2}{MQ(1-Q)}, \tag{5.1}$$

where M is the number of native contacts and $\langle Q_i \rangle_Q$ is the average formation of the ith contact in all configurations with a particular global Q. $R(Q)$ quantifies the "diversity" of structures seen at each value of Q: $R(Q) = 0$ being maximum diversity and $R(Q) = 1$ being a single route (*159*). At $R(Q) = 0$, all $\langle Q_i \rangle_Q = Q$, meaning all possible configurations of native contacts are sampled equally. At $R(Q) = 1$ only a subset of MQ contacts are formed with $\langle Q_i \rangle_Q = 1$, meaning only one configuration of native contacts is sampled.

5.5.4 Identification of the Knot Along the Protein

Knots observed in proteins are "open" knots, so they differ from the mathematical definition of (closed) knots. Nonetheless, when both termini of the protein are located far enough from its entangled core, they usually can be unambiguously joined by an additional interval which transforms them to a closed loop. If such a loop is not homeomorphic to a circle then the native protein is regarded as representing a nontrivial knot. Under this procedure MJ0366 contains a 3_1 trefoil knot. The native location of the knot $K_{\mathrm{N}}/K_{\mathrm{C}}$ along the sequence, *i.e.* the minimal segment of amino acids that can be identified as a knot, was determined by KMT algorithm (*154*). The positions of $K_{\mathrm{N}}/K_{\mathrm{C}}$

during folding were determined in the same way as the native knot, applying the procedure at each simulation snapshot as described before (*62*). The slipknot conformation was detected as described before in (*149*).

5.5.5 Note About Knot Chirality in Proteins

Topological traps were seen containing knots with the wrong chirality . Mathematically, a chiral knot is a knot that is not equivalent to its mirror image, *i.e.* it cannot be continuously deformed from the image to the mirror image. For a trefoil knot there are two different topologies, the trefoil and its mirror image. A 4_1 knot, which exists in proteins, is not chiral in the mathematical sense. Its mirror image can be deformed into the original knot. This mathematical statement assumes that the polymer (string) itself has no chirality associated with it. In proteins, even with trivial topology, a mirror image cannot be superimposed on the original meaning that a 4_1 knot with all its crossings reversed is in fact a different knot from the protein's perspective. A protein with a 4_1 knotted topology could still fall into a topological trap created by wrong chirality, the knot with all the crossings reversed.

5.5.6 Comparing Kinetic Folding

Rescaling Time Units

The C_α and AA models are both defined in reduced units. One must fill in actual units to be able to compare time scales. The time units are related to the length, mass and energy units,

$$[\text{time}] = \sqrt{\frac{[\text{length}]^2[\text{mass}]}{[\text{energy}]}}. \tag{5.2}$$

Taking the binding energy of the protein to be E_o, the energy unit in AA is $E_o/689$, since the total energy is set to the number of atoms, 689. In the C_α model, each contact and dihedral is given a value of 1, so the energy unit is $E_o/(79+220)$ since there are 79 dihedrals and 220 contacts in the C_α model. The length scale in both models is a nanometer, l_o. The mass scale is different since in C_α a residue-bead is given a mass of 1 whereas in AA an atomic-bead is given a mass of 1. C_α mass unit is therefore 689/82=8.4 times

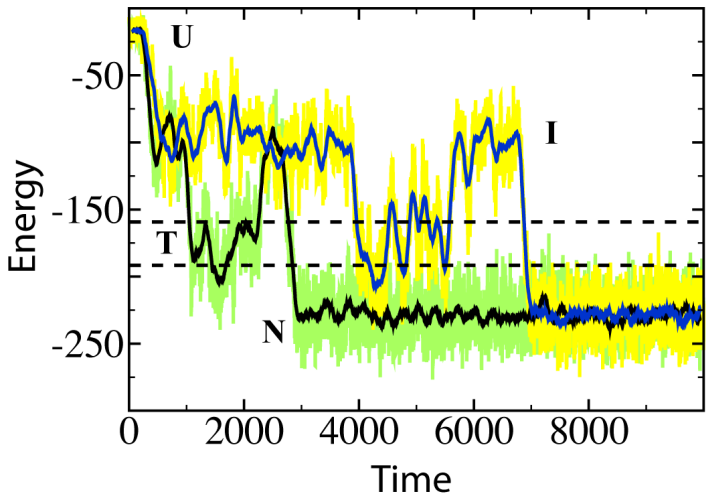

Figure 5.9: Time spent in traps. Two representative kinetic trajectories from the AA model that fall into topological traps at $T = 0.91T_{\mathrm{F}}$. Note four states are clearly seen, **U**, **I**, **T**, **N**. The smoother line is a running average of the previous 100 energy samples, which are sampled every 2.5 time units. Time trapped in state **T** is approximated by counting 2.5 time units times the number of averaged samples lying between the energy values denoted by the dotted lines. The energy of the dotted lines are determined such that all points lying between are distinct from states **I** and **N**.

the AA mass unit m_o since there are 689 atoms and 82 residues. Comparing the time units,

$$\frac{[\text{time}]_{\mathrm{C}_\alpha}}{[\text{time}]_{\mathrm{AA}}} = \frac{\sqrt{\frac{(l_o^2)(8.4m_o)}{E_o/299}}}{\sqrt{\frac{(l_o^2)(m_o)}{E_o/689}}} \approx 2. \tag{5.3}$$

When comparing the mean first passage times τ_{mfpt}, the C_α times are scaled up by a factor of 2 relative to the AA times. This analysis does not take into account differences in diffusion, which would likely increase the relative C_α time further since diffusion in the coarse-grained model should be faster.

Quantifying the Time Spent in Trapped States

This section describes the data comprising Figure 5.6c,d. The overall τ_{mfpt} was determined by averaging the time to knotting for $\geq$200 trajectories started from random

configurations from the unfolded ensemble. The time spent in the trap τ_{trap} is computed by counting the total number of time averaged protein configurations (denoted by energy) that lie within a certain range times 2.5 time units per point (Figure 5.9). If $\tau_{\text{trap}} > \tau_{\text{mfpt}}$ the trajectory is considered "trapped." Figure 5.9 explains the procedure for AA at $T = 0.91T_{\text{F}}$, but the same procedure with slightly altered bounds on the trapped energies is used for other temperatures and for C_α.

5.5.7 Movie of a Slipknot Folding Trajectory

An excerpt from a molecular dynamics trajectory that shows an example of the slipknotting pathway is provided as an animated GIF. This movie was taken from an all-atom folding run at $0.86T_F$ and created using VMD and Tachyon. Helix α_3 first forms its native contacts and then helix α_4 begins to make its hydrophobic core contacts. These interactions stabilize the C-terminus in a bent and strained hairpin-like configuration near the C-loop. After threading, the disordered C-loop orders around the properly threaded C-terminus, which completes folding. This is an excerpt of an all-atom trajectory, but only the positions of the α-carbons are shown for simplicity. The movie can viewed online at http://guara.ucsd.edu/knot/slipMovie.gif.

5.6 Acknowledgments

Chapter 5, in part, appears in *PNAS*, (2010), Noel, Sułkowska, Onuchic. The dissertation author was the primary investigator and author of the paper. This work was supported by the Center for Theoretical Biological Physics sponsored by the NSF (Grant PHY-0822283) with additional support from NSF-MCB-0543906. JKN acknowledges support from NIH Molecular Biophysics Training Program, Grant number: T32GM08326.

Chapter 6

Mirror Images as Naturally Competing Conformations in Protein Folding

6.1 Introduction

Much of protein folding theory works under the assumption that the energy landscape is biased towards a single attractive basin: the native state. Anfinsen's thermodynamic hypothesis states that a protein's native conformation lies in the global minimum of its free-energy landscape (*160*). Evolution achieves this robustness by selecting for sequences in which the interactions present in the native state are mutually supportive and cooperatively lead to the functional structure (*6*). This gives rise to protein sequences that are minimally frustrated, meaning sequences are only consistent with a single native structure. The resulting energy landscape is smooth and funnel-shaped. Any competing or "trapping" protein configurations have an energetic depth much smaller than the overall bias to the native structure. Under this framework, which has been called *energy landscape theory* (*6–8*), protein dynamics is dominated by the geometry of the protein's native configuration. In this scenario, we pose the question: what are the consequences of native configurations with a high degree of symmetry?

Prior studies have investigated the detailed folding mechanisms for several proteins with native structures that have specific domains arranged in a symmetric pattern (*161*). For example, protein L consists of an α-helix packed against two, symmetri-

cally arranged, β-hairpins. Protein L, though, folds asymmetrically, through a transition state ensemble (TSE) consisting of an ordered N-terminal β-hairpin and largely unstructured C-terminal β-hairpin (*162*). A homologous protein, protein G, instead folds by ordering the C-terminal β-hairpin in the TSE (*163*). Simulations have suggested that the detailed side-chain packing determines one folding route over the other (*88, 164*). Another fold family consisting of symmetric domains are the β-trefoils (*165*). A series of theoretical and experimental studies of the Interleukin-1 family of β-trefoils has shown that the the folding degeneracy is broken by functional regions of the protein, which slightly alter the structure among each member (*14, 104, 166*). The overall lesson from these studies is that, while the proteins are able to fold *via* multiple routes, they tend to choose one of the routes allowed by symmetry to dominate the TSE. The choice is based on residual frustration that can arise, for example, from the geometry of the side-chain packing or energetic heterogeneity. These differences are subtle, and not robust, as a couple kcal/mol is enough to bias the folding down a particular route (*6, 167*).

Let us now consider protein structures where symmetry allows, in addition to multiple folding routes, also multiple structures within a nominally singly-funnelled energy landscape. For example, for a three-helix bundle like the B-domain of protein A (Bdpa), the near mirror image [1] of the native state is compatible with the native contact network (Figure 6.1). The three helices pack around a hydrophobic core and arrange (left-) right-handed for the (native) mirror fold. Non-specific interactions, such as hydrophobic interactions, leave little to distinguish between the native and mirror helical packings. This missing specificity towards one particular conformation has been seen in structure prediction, where energy functions are unable, or only weakly able, to discriminate between the native fold and competing mirrored folds/decoys with low energies (*169–173*). For all-helical coiled-coil proteins the composing helices form a hydrophobic core with, and orient against, each other to form the native fold, but a simple reorientation in their mirror image forms a comparable hydrophobic interface.

The ability of proteins to explore multiple structures allowed by symmetry is seen in domain-swapped homodimers (*174*). Domain swapping occurs when structural

[1] Technically, the chirality of naturally occurring L-amino acids precludes folding of a "mathematically" exact mirror image. For the sake of simplicity, we call the mirror arrangement of helix axes a mirror image.

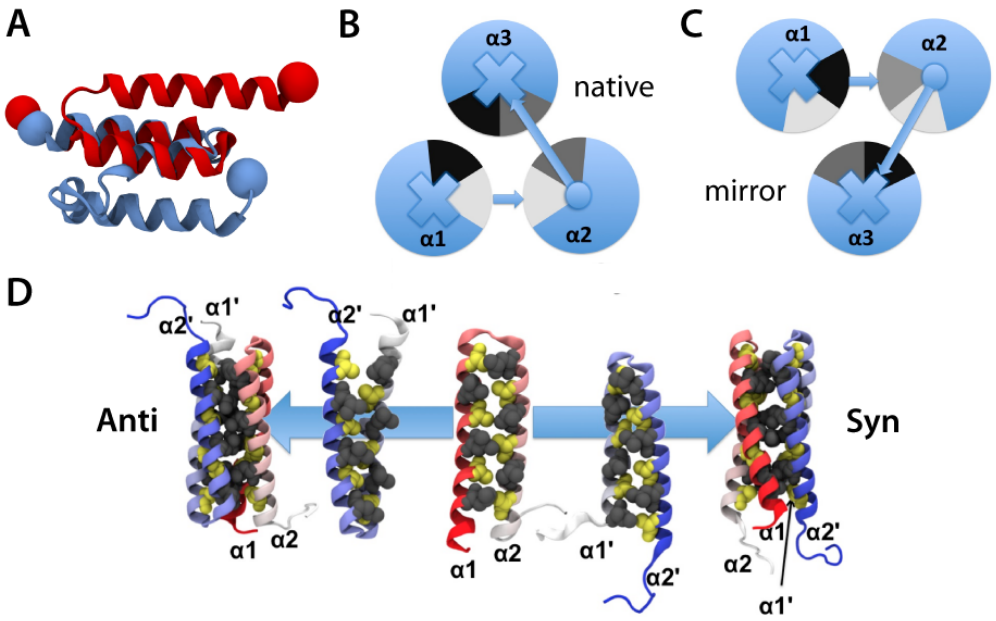

Figure 6.1: Protein symmetry gives rise to multiple consistent structures. **A**: Mirror images of Bdpa, a three-helix bundle. The C-terminal helices are aligned and the native N-terminal helix packs on the top, while the mirror N-terminal helix packs on the bottom. (**B** and **C**) show the packing of amphipathic helices, the hydrophobic core residues shown in gray. Compared to the native fold (**B**), a slight rotation of helices 1 and 2, and reorientation of the third helix, facilitate a hydrophobic core composed of the same residues in the mirror fold (**C**). The packing of interacting residues, however is different. **D**: The Rop-homodimer is composed by two all-helical monomers (chain A red to white, chain B white to blue). If one mutates core residues at the hydrophobic interface into optimally packed Ala (yellow) and Leu (grey), there is competition between two possible arrangements of the monomers (displayed mutant A2L2-6) (*21*). One possibility is the functional WT anti packing in which six small Ala side chains perfectly fit into six larger Leu side chains. For anti, helix α1 is packed against α1' and α2 against α2'. Symmetry, however, enables the competition of a dysfunctional syn packing, in which Ala and Leu form a similarly perfect packing. Here helix α1 is packed against α2' and α2 against α1'. Both states have been measured to compete in single molecule FRET experiments (*168*).

elements crucial for stabilizing the monomeric fold are replaced by the same structural elements from its dimeric partner. As expected in a funneled energy landscape, simulations have shown that the signals for domain swapping are encoded by the monomeric fold (*18*), *i.e.* they arise solely from symmetry. The low specificity towards a singular fold is, perhaps, best exemplified by the Rop-homodimer and its various mutants. While the WT (un)folds slowly ($k_{\mathrm{F}} = 0.013\mathrm{s}^{-1}$), specific mutants have an optimized Ala/Leu core-packing and speed up folding by up to 4 orders of magnitude (*175*). At the same time, these mutations symmetrize the interface (*16*), reduce the specificity to the native fold, and open what has been called a "trapdoor" (*21*), an energetically competitive structure with a different symmetrically related global fold (Figure 6.1). Further, this "trapdoor-fold" is stabilized by small concentrations of denaturant which explains the unusual kinetic behavior (*168*).

The present study investigates the free-energy landscape of three-helix bundle proteins that may have naturally occurring trapdoors built in by symmetry. Their simple coiled-coil structures allow the mirror image structures to be energetically competitive. The B/E-domains of staphylococcal protein A (Bdpa and Edpa) are two of five homologous IgG-binding domains (*176*). They act as pathogenicity factors for the bacterium *Staphylococcus aureus*, by binding tightly to the Fc region of IgG or Fab region of IgM, and each consist of three helices packed against one another. α_3d is a *de novo* superstable designed three-helix bundle which expresses very quick kinetics with folding times around 4 μs (*177*).

6.2 Methods

This study utilizes several simulation methods, not only to guard against forcefield bias, but also in an attempt to gain a complete perspective on the energetic, kinetic, and thermodynamic accessibilities of mirror-symmetrical protein structures. The structure prediction forcefield PFF01/02 shows that mirror configurations of Bdpa, Edpa, and α_3d are enthalpically competitive with the native configurations. This competition between mirrored helical bundles is explored in detail for Bdpa with replica exchange molecular dynamics and structure-based simulations. REMD simulations reveal

two thermodynamically-competitive folded basins, an enthalpically favored native-like basin and an entropically favored mirror-like basin. The structure-based simulations are built from representative REMD structures and corroborate the REMD predictions for the kinetic accessibility of the two basins.

6.2.1 Structure Prediction

Forcefield: PFF01/02

The all-atom (with the exception of apolar CH_N groups) free-energy forcefield PFF01/02 (*83, 173, 178*) models the internal free-energy of the protein along with an averaged implicit solvent interaction. Contributions from backbone entropy are not considered. It has found wide application in structure prediction (*179–181*). The energies can be used to reconstruct folding kinetics. Its functional form is

$$V(\vec{r}) = \sum_{ij} V_{ij} \left[\left(\frac{R_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{ij} \frac{q_i q_j}{\varepsilon_{g(i)g(j)} r_{ij}} + \sum_i \sigma_i A_i + \sum_{\text{hbonds}} V_{\text{hb}}$$

where r_{ij} denotes the distance between atoms i and j and $g(i)$ the type of amino acid containing the atom i. The Lennard-Jones parameters (V_{ij} for potential depth and R_{ij} for equilibrium distance) depend on the type of the atom pair and were adjusted to satisfy constraints from a set of 138 proteins out of the PDB database. The electrostatic interactions contain group-specific dielectric constants $\varepsilon_{g(i)g(j)}$ and partial charges q_i were derived in a potential of mean-force approach. The implicit solvent interaction constants are obtained by a minimal solvent accessible approach and parameterized by free energies per unit area σ_i to reproduce the solvation enthalpies of the Gly-X-Gly family of peptides. A_i corresponds to the area of atom i that is in contact with a fictitious solvent. Hydrogen bonding is modeled as dipole-dipole interactions in the electrostatic term and an additional short-range term for backbone hydrogen bonding (CO to NH). It depends on the OH distance, the angle between N, H and O atoms along the bond and the angle between the CO and NH axis.

Global Optimization

A variety of approaches have been tested to find the global minimum of PFF01/02 (*83, 84*). For the present study we used an evolutionary algorithm for the proteins α_3d and Edpa (*85*). Starting from random initial conformations, a population of structures is fully minimized by repeated rounds of section and minimization. Each round selects a subset of structures that balances energetic favorability with structural diversity. This approach can easily be implemented on heterogeneous computational resources and is very efficient.

6.2.2 REMD Simulations

We use replica exchange molecular dynamics (REMD) to study the folding-unfolding equilibrium of the *Staphylococcus aureus* B domain of protein A (Bdpa) using an all-atom and explicit solvent, forcefield-based model. Bdpa consists of amino acids 1-57 (TADNKFNKEQ QNAFYEILHL PNLNEEQRNG FIQSLKDDPS QSANLLAEAK KLNDAQA) and was modeled with charged C- and N-termini. His 19 was charged. All Glu, Asp, Lys and Arg were modeled as charged. The initial configuration of the system was modeled as an extended PPII structure, without any bias towards the native state. The protein was modeled by the all-atom Amber ff94 forcefield and the solvent was modeled by 6583 TIP3P water molecules, as in our previous calculations on this protein (*82*). An extended, PPII conformation was generated with the Amber leap program. To remove any bias from this initial configuration, the system was heated to 800 K and simulated in vacuum for 100 ps. The resulting collapsed and unfolded configuration was immersed in a cubic box of water molecules and equilibrated at 300 K and 1 atm for 1 ns. The equilibrated protein-water system was a cube 5.97 nm on a side and has a density of 0.9699 g/liter. Copies of this system were then simulated for 1 ns at various T ranging from 275-600 K at constant volume. These configurations were used as initial configurations in the REMD calculation. The REMD simulations were extended for 450 ns per replica. The total simulation time was 28.8 μs. REMD simulations were done over a wide range of temperatures, 287 K to 643 K, chosen to get a uniform exchange rate among replicas sampling neighboring temperatures (*182*).

Exchange attempts were possible at every integration step with a 5% percent probability. The exchange rate was chosen to be close to 20%. On average, exchanges were attempted every 1.4 ps. The average time between successful exchanges for all replicas was 8 ps. The integration time step was 2 fs. Temperatures were maintained using the Nose-Hoover thermostat, with 2 ps coupling time. Hydrogen containing bonds were constrained by SHAKE and SETTLE. We used the particle mesh Ewald summation with a 0.1 nm grid size. Lennard-Jones interactions were truncated at 1.0 nm. Pair interaction lists were updated every 25 integration steps. Figure S1 shows the convergence of the REMD simulation. Thermodynamic averages were calculated over the last 250 ns/replica.

In our experience, Amber ff99SB, which has been shown to describe better dynamics (*183*) and thermodynamics (*184*), does not fold Bdpa nor does it maintain the folded state in REMD simulations (40 μs total) started from the folded state.

6.2.3 Structure-based Models

We use the standard coarse-grained, native-centric structure-based model (SBM), described in Chapter 1. We use the potential described in (*60*), where the usual Lennard-Jones contact potentials are replaced by Gaussian contact potentials. The width of the Gaussians scale linearly with the native contact distance and results in an average width of $\overline{\sigma} \sim 0.7$ Å. The all-atom SBM is also described in Chapter 1. The excluded volume parameters are $\varepsilon_{\mathrm{NC}} = 1$ and $\sigma_{\mathrm{NC}} = 2.1$ Å, giving more realistic atomic sizes. The native contact maps are constructed using Shadow (*57*). The SBM were constructed using the SMOG webserver (*57*). All SBM were sampled with MD using Gromacs v4.5 (*74*) using under-damped stochastic dynamics.

In order to compare the kinetics of the symmetrical structures, we constructed a SBM with "dual-basins" that has equal energetic minima at two structures. This type of SBM has also been used to look at, for example, conformational transitions (*19*) and homodimeric folding (*21, 185*). The two "native" structures, used as input for the dual-basins, are taken as representative members from REMD clusters 1 and 3 from REMD. The structures with the highest number of atomic contacts as determined by Shadow are chosen and minimized in Amber94. The native-like structure thus obtained,

called S_{N}, has 444 atomic contacts, whereas the mirror-like structure S_{M} has 373 atomic contacts. They are shown in Figure 6.1A. In the dual-basin SBM potential all contacts are included. If a contact exists in both structures, but at different distances, it is included using a double-basin Gaussian contact potential, which consists of two Gaussian wells of equal depth with minima at the respective native distances (see Equation 1.10 and (*60, 186*)). To make the basins equally energetically stable, the contact potentials in S_{M} are scaled by $444/373 \approx 1.2$. The torsional angles have a form such that they are minimized at both S_{N} and S_{M}. Improper dihedrals, angles and bond lengths, nearly identical between S_{N} and S_{M}, are taken from S_{N}.

6.2.4 Contact Map Definitions

In order to quantify a structure's similarity to S_{N} and S_{M}, we need to construct two "native" contact maps C_{N} and C_{M}. A contact between two residues is considered "native" to its basin X if, in a subset of structures < 3 Å away from S_X, a contact as defined by the Shadow algorithm (*56, 57*) exists with a probability greater than 0.5. The Shadow algorithm defines residues in contact if they have any directly interacting atoms, *i.e.* two atoms within 6 Å and not occluded by other atoms. A tertiary contact is any contact between residues separated by more than 5 amino acids in sequence. Applying this definition of the contact map, the native basin contact map C_{N} has 112 total contacts and 65 tertiary contacts, and the mirror basin contact map C_{M} has 101 total contacts and 51 tertiary contacts. The reaction coordinate $Q_X = \sum_{ij} \theta(1.2 r_{ij}^X / r_{ij})$ measures the similarity to S_X, where the sum goes over tertiary residue pairs ij in C_X, r_{ij} is the distance between C_α atoms ij, r_{ij}^X is the distance between C_α atoms ij in S_X and θ is the unit step function. In order to separate the native basin from the mirror basin we use the combined coordinate $Q_{\mathrm{N}} - Q_{\mathrm{M}}$. Q_X^{AA} is a finely-grained reaction coordinate that sums over all tertiary atomic pairs from an all-atom contact map C_X^{AA} constructed by running Shadow on S_X. Q_X^{AA} is used with the all-atom SBM. $C_{\mathrm{N}}^{\mathrm{AA}}$ has 444 atomic contacts (254 tertiary) and $C_{\mathrm{M}}^{\mathrm{AA}}$ has 373 atomic contacts (194 tertiary).

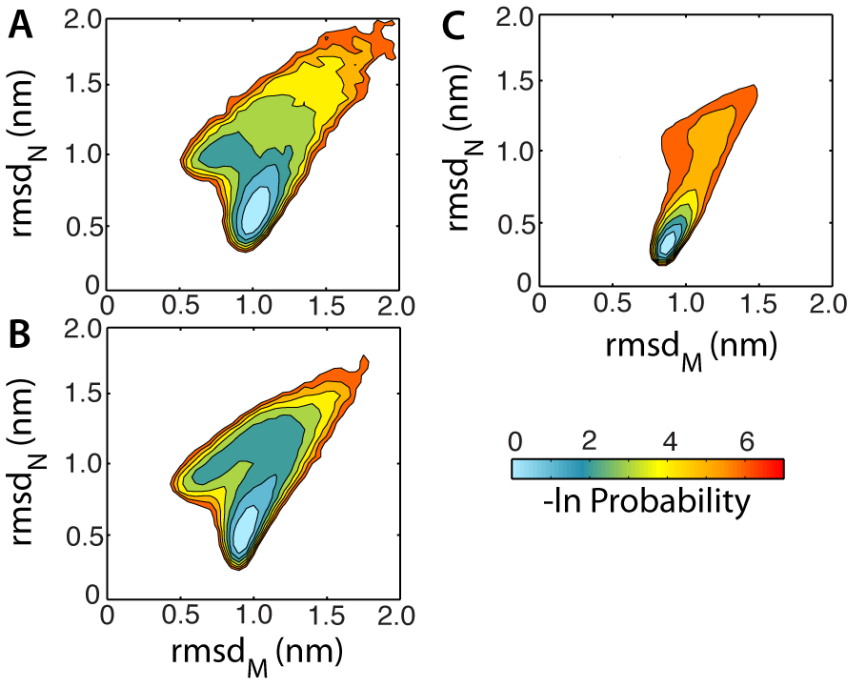

Figure 6.2: Native-centric, coarse-grained protein model populates mirrored structures. Rmsd to the mirror structure (rmsd_M) versus rmsd the native structure (rmsd_N) for Bdpa (**A**), α_3D (**B**), and CI2 (**C**). Histograms are shown at $T = 0.93T_\text{F}$. In (**A**) and (**B**), the mirror image shows up as an accessible conformation for both three-helix bundles, destabilized by $\sim 2\text{k}_\text{B}T$ relative to the native structure. CI2 has an α/β-fold with no symmetrical analog and therefore only populates a single basin.

6.3 Results

6.3.1 Coarse-grained Native-centric Protein Model Populates the Mirrored Basin

The near degeneracy of the native and mirror structures is seen through simulating a coarse-grained SBM (C_α-model) (*13, 60*) of the three-helix bundles. The model includes short range interactions between native contacts and torsional angles, and all interactions have their minima at the native structure. Though the contact interactions on their own are unable to discern between the native structure and its mathematical mirror image because they only depend on scalar distances, the torsional angles are vector quantities, and therefore bias the local chirality to that of the native structure. Since all the helices in these proteins are left-handed, this results in a large energetic penalty for

the right handed helices that would arise in a mathematical mirror image. In the simulations the helices are maintained with left-handed chirality, but the overall packing of the helices switches between the native packing and a mirror image packing. It is this change in helical packing that defines our use of the term *mirror* (Figure 6.1).

Figure 6.2 shows the relative populations of the native and the mirror packings. The native packing is monitored by the root mean squared distance (rmsd) of the C_α atoms from the native structure rmsd_N. The mirror packing is monitored by rmsd_M, the rmsd from a mirror packing of left-handed helices S_M taken from REMD (see Section 6.2.3) The three-helix bundles, Bdpa and α_3d, both show a significant population of the mirror conformation. The mirror is destabilized relative to the native by $\sim 2\text{k}_\text{B}T_\text{F}$. For comparison, CI2, a well-studied α/β two-state protein, is shown in Figure 6.2C. CI2 only populates the native state.

The native and mirror helical packings are not completely degenerate because the helices maintain their left-handed chirality. The structures populating the mirror basin are balancing the energetic cost of straining the native torsional angles to the energetic benefit of forming native tertiary contacts whose distances are not optimized to the mirror helical packing. The ability of the protein to find such structures at a low enough energy is a testament to the plasticity of simple protein structures. While it can be said that the stability of the mirror is overemphasized because the changes in side-chain packing are not described in the coarse-grained model, the energetic heterogeneity is underrepresented by the native-centric energy function. The tertiary contact network is largely hydrophobic, constructed through the packing of amphipathic helices. The nonspecific nature of hydrophobic interactions should allow the mirror packing to utilize additional tertiary contacts beyond the native set. These non-native contacts are not included in the coarse-grained model. The hydrophobic side-chain packing is explored in the next section through a structure prediction forcefield with no knowledge of the native structure.

6.3.2 Mirrored Structures are Energetically Competitive

To assess the stability of native and mirror configurations, we performed structure prediction simulations of Edpa in the all-atom, implicit water forcefield PFF01/02

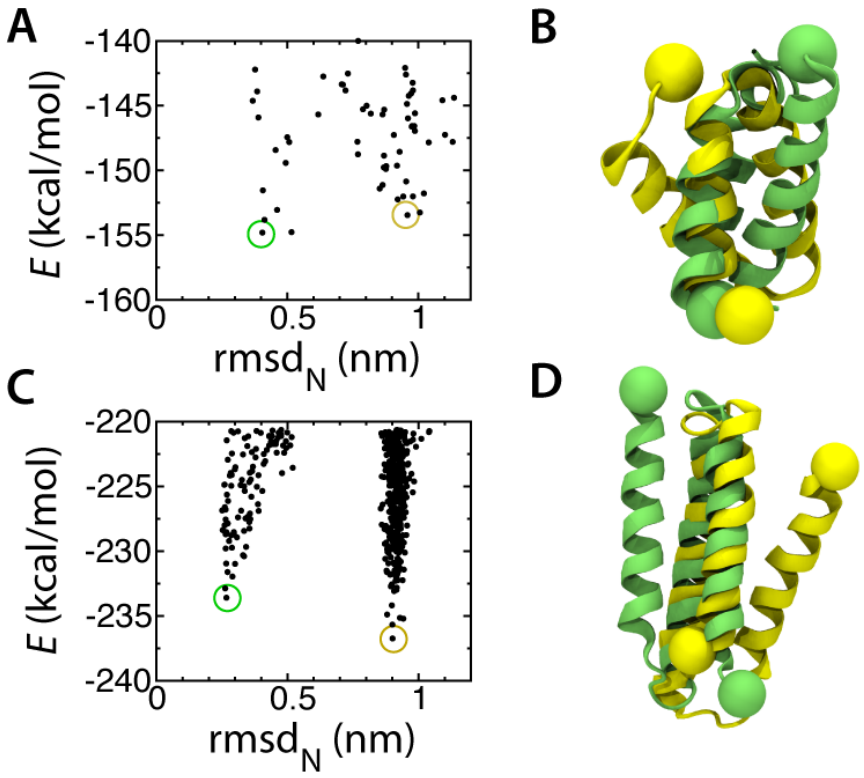

Figure 6.3: Distribution of structure prediction energies versus rmsd from the PDB structures (PDB codes: (**A**) 1edk and (**B**) 2a3d). Edpa (**A** and **B**) and α_3d (**C** and **D**). The lowest energy native-like structure is shown in green and the lowest energy mirror-like structure is shown in yellow. The N-terminal helices are aligned. The energies of the structures shown in (**B**) and (**D**) are indicated by green and yellow circles in (**A**) and (**C**). Two populations with low energy are seen for both proteins. Edpa has more structural diversity than α_3d, probably due to its smaller size.

(*178*) using an evolutionary algorithm (*85*) to minimize the energy and sample the low energy landscape of the protein. Over 5000 protein conformations were visited during the simulations and the energy and rmsd_N (from the PDB structure) of the final population of conformations are shown in Figure 6.3. Edpa shows two broad funnels at around 4 Å and 10 Å rmsd_N. Although the native-like conformation in PFF01/02 is lower in energy than the competing misfolded conformation ($\sim$ 1 kcal/mol), such a small difference is within the resolution attainable by empirical forcefields. The misfolded funnel ($\text{rmsd}_\text{N} \sim 10$ Å) consists of mirrored configurations of Edpa with the N-terminal helix flipping to the other side of the structure formed by middle and C-terminal helix. The native-like and mirrored conformations are shown in Figure 6.3. In these simulations we find no specific rearrangement of the side-chain packing of helices 2 and 3 to accommodate the change in orientation of helix 1 with respect to the native state.

Similarly, we performed simulations for the α_3d protein in PFF01/02. Here we also observe a double funnel in the energy landscape as shown by the distribution of all conformations visited during the simulation (Figure 6.3C). The native-like funnel found its minimum at $\sim$ 2.5 Å while the mirror-like funnel had a minimum at $\sim$ 9 Å. The conformation at the minimum of the mirrored funnel had the same arrangement of helix 1 and 2 as native-like, while the C-terminal helix is on the other side with comparable side-chain packing. Note that the chirality of helical packing in α_3d is opposite of Bdpa/Edpa.

The consistent observation of mirror images in three-helix bundles in PFF01/02 (Bdpa data can be found in (*173*) with the same result) and by others (*169–172, 187, 188*) leads us to speculate that the mirror configurations may affect the folding and function of these proteins. In the next section we examine the thermodynamic competitiveness of the mirror configuration on the folding landscape.

6.3.3 Atomistic Simulations Show a Mirror Basin that is Thermodynamically Competitive and Kinetically Accessible

In order to predict the occupation of the mirror structure, for example in an *in vitro* protein folding experiment, we need to know more than the relative energetic stabilities of the two competing structures. The mirror structure needs to both be ther-

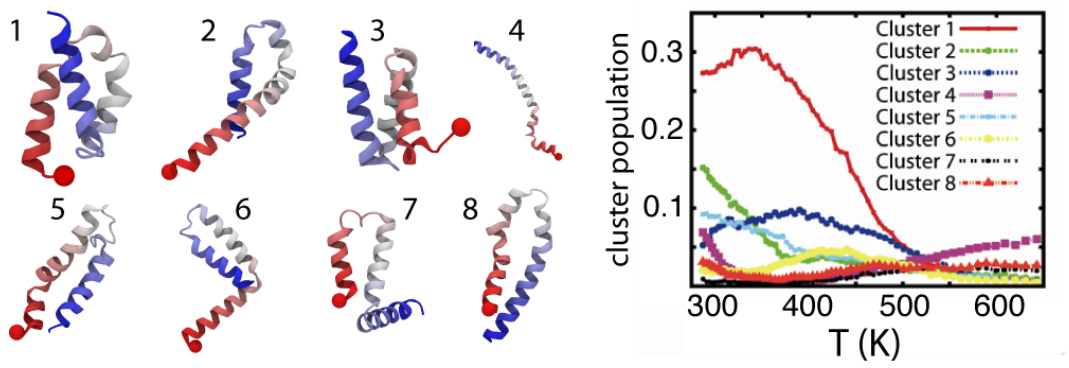

Figure 6.4: **Left**: Central structures from the largest eight structural clusters in the REMD Bdpa simulations at 287K. Cluster 1 is closest to the NMR structure (PDB code: 1bdd (*91*)) and cluster 3 is the mirror image. **Right**: Cluster population as a function of temperature. The native-like cluster is the most populated until 500K.

modynamically competitive, and kinetically accessible, in order for the mirror packing to be visited. To investigate the thermodynamic behavior of symmetrical proteins, we performed replica exchange molecular dynamics (REMD) of Bdpa in an explicit solvent environment. Both helical packings are observed, at a ratio native/mirror of ~ 3. We now present results from both the REMD, and various all-atom structure-based simulations, to show the kinetic accessibility of the native and mirror helical packings.

Replica Exchange Molecular Dynamics of Bdpa

A total of 28.8 microseconds of replica exchange molecular dynamics (REMD) was performed, 450ns for each of 64 replicas of Bdpa. The simulation representation of Bdpa used the same sequence as in the experimental studies of Sato *et al.* (*93*). Blind cluster analysis of the resulting structures shows that, at $T = 287$K, the most populated cluster (cluster 1) corresponds to the native structure and the fifth most populated cluster (cluster 3) corresponds to a mirror helical packing (Figure 6.4). The best agreement between the deposited NMR structure for the protein (*91*) and the simulation is 1.2 Å rmsd of the backbone. From 350K to 525K, the native and mirror clusters become respectively the first and second most populated clusters. The relative thermodynamic stabilization of the mirror structure with temperature is reminiscent of the Rop dimer (Figure 6.1) and suggests that the mirror structure is more entropically favorable.

The REMD simulations provide details of the equilibrium populations of the

symmetric folds, but it does not provide a direct description of the kinetics of the process. Analysis of individual replica trajectories, even though it does not correspond to a single temperature, can give hints of the kinetic properties. The REMD was initialized from random collapsed, but unfolded, conformations. Monitoring the number of replicas that are sampling the native or mirror structural clusters as a function of time, shows that the mirror cluster is initially populated roughly twice as fast as the native cluster (Figure 6.5D).

Dual-Funneled Energy Landscape Description of Bdpa

A funneled energy landscape has a single structure (or rather a small ensemble) that is both the enthalpy minimum and (below folding temperature) the free energy minimum of the energy landscape. "Enthalpy" refers to the renormalized enthalpy obtained by averaging over solvent contributions. Here, for three-helix bundles, our evidence suggests an energy landscape with *two* structures near both the enthalpy minimum (as seen by PFF01/02 in Section 6.3.2), and free energy minimum (as seen by REMD in the previous section), *i.e.* the native and the mirror helical packings. Therefore, a first order approximation to the funneled energy landscape is a SBM that includes "dual-basins," where both the mirror and native structures are made explicit energetic minima.

A coarse-grained dual-basin SBM has been previously applied to a Rop dimer mutant containing homogeneous hydrophobic core packing (*21, 60*). Here, where the side-chain packing is likely of great importance, we use an all-atom SBM that explicitly represents all heavy atoms in the protein. The two "native" structures are representative members of the native and mirror clusters, called S_{N} and S_{M}, taken from the REMD. The dual-basin SBM is set up such that the two structures are equally energetically stable. Thus, the SBM is only testing the relative entropy and kinetic accessibility of the two structures. Constant temperature MD simulations were performed and the results are presented in Figure 6.5. Comparing Figures 6.5A,B shows that, while the dual-basin captures the essence of the REMD landscape, even neglecting the spurious high rmsd structures of the REMD, the SBM misses some of the structural heterogeneity seen in REMD.

Figure 6.5C presents the thermodynamics of three SBMs: two single-basin SBMs

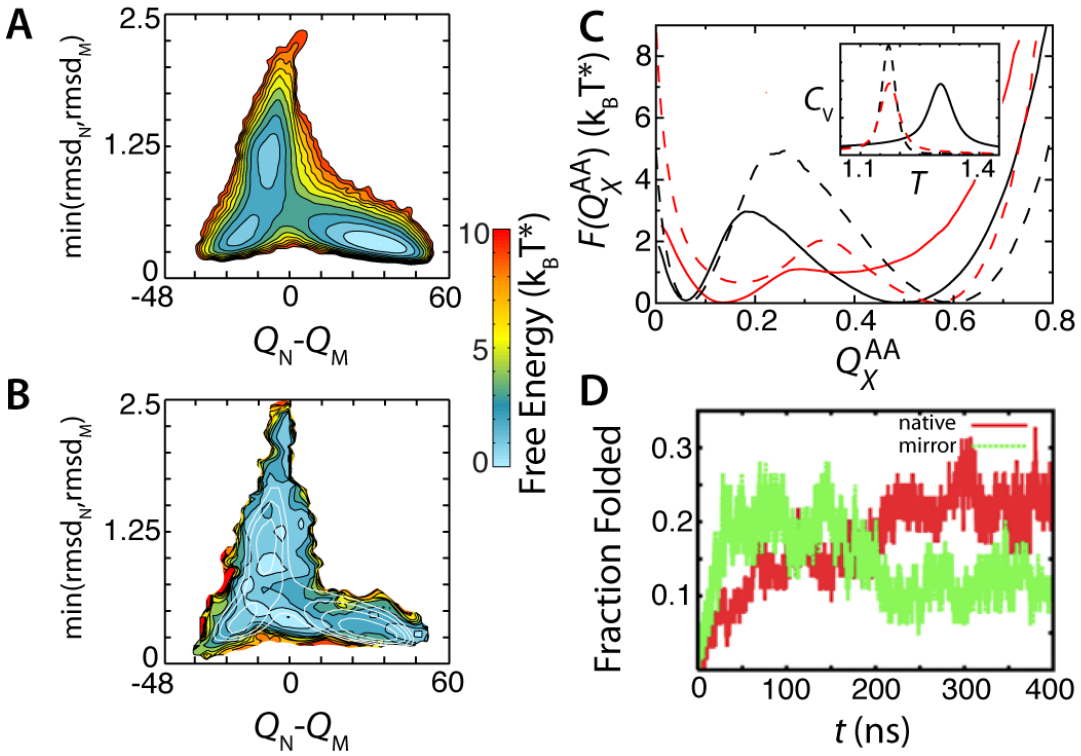

Figure 6.5: Kinetic accessibility and thermodynamic stability of the mirror image. Free energy contours are shown for two models: MD simulations of a dual-basin all-atom structure-based model (SBM) at $\overline{T} = 1.29$ (**A**) and replica exchange MD of Amber94 at $\overline{T} = 0.6$ or $T = 300$K (**B**). The reduced temperature $\overline{T} = T/\mathrm{T}^*$ where $\mathrm{k_B T}^* = 1$. The minimum rmsd to either the native or mirror structure, $\min(\mathrm{rmsd_N}, \mathrm{rmsd_M})$, is plotted versus the difference in native contacts from the mirror contacts, $Q_\mathrm{N} - Q_\mathrm{M}$. Therefore, the abscissa partitions structures between mirror-like and native-like and the ordinate partitions helical-bundle-like structures from extended ones. The SBM basins ($P > 10^{-3}$) are overlaid on the Amber94 as the white lines in (**B**). (**C**) compares barrier heights between three simulations: two control single-basin (SB) SBM (dotted lines) at $\overline{T} = 1.17$ and the dual-basin (DB) SBM (solid lines) at $\overline{T} = 1.29$. Free energy barriers computed with similarity to the native structure Q_N^AA are shown in black, and to the mirror structure Q_M^AA are shown in red. In both the SB and DB simulations, the mirror structure has a lower barrier compared to the native. The mirror structure is more stable than the native structure in SB but it is less stable in DB. Both barriers in the DB simulations are lower than in the SB simulations. The inset in (**C**) shows the specific heat for the three calculations. (**D**) shows corroborating evidence from replica exchange MD that the mirror structure is more kinetically accessible than the native structure. Fraction of structures across all 64 replicas belonging to the folded clusters are plotted as a function of simulation time. The native (cluster 1) is shown in red and the mirror (cluster 3) is shown in green. If $t = 400$ ns corresponds to equilibrium, the occupation of the native relative to the mirror would be ~ 2 in the ensemble of all replicas. The free energies in (**A**), (**B**), and (**C**) are computed using WHAM (*66*).

to each S_N and S_M, and a dual-basin SBM. The two single-basin models indeed have nearly equal thermal stability as seen by the specific heat, and the dual-basin model is considerably more stable. This is expected, because the dual-basin SBM can form contacts from both S_N and S_M and the dual-basin torsional angles are more forgiving. This also decreases the cooperativity of the dual-basin SBM (increases width of C_V) because of the additional structural heterogeneity. The free energy is plotted versus the number of native atomic contacts formed Q_N^{AA} and Q_M^{AA} defined by S_N and S_M, respectively. The heights of the free energy barriers correspond to the kinetics of the transitions (*63*). The single-basin SBMs predict much slower kinetics for the native S_N compared to the mirror S_M. The barriers to both the native basin and the mirror basin decrease in the dual-basin SBM, but the barrier to the native is still 2 k_BT^* larger than to the mirror. The relative thermodynamic stabilities between the native and mirror, favors the mirror structure for the single-basin SBM, but switches to favoring the native structure in the dual-basin SBM by 1 k_BT^*. The dual-basin SBM predicts the following quantities for Bdpa: with the diffusion taken as constant, the folding rate to the mirror is $\exp(-\Delta F^{\ddagger}/k_BT) = \exp(1.8/\overline{T}) = 4.0$ times faster, and the native structure is more stable by a factor of $\exp(\Delta F^{\dagger}/k_BT^*) = \exp(1) = 2.7$, where $F^{\ddagger}$ and $F^{\dagger}$ are the free energies at the barrier and the folded basin, respectively, and Δ implies a subtraction of native from mirror.

The entropic favorability of the mirror basin is suggested by the REMD, since the population of the mirror cluster increases with temperature. This is corroborated by the SBM in two ways. First, in the single-basin SBMs, where S_N and S_M are given equal energetic stability, the mirror is slightly more stable than the mirror (Figure 6.5C). Second, as temperature is increased in the dual-basin SBM, the mirror basin becomes stabilized relative to the native basin. At $\overline{T} = 1.33$, the free energies of the native and mirror become equal, $F_N^{\dagger}(Q_N^{AA} = 0.5) = F_M^{\dagger}(Q_M^{AA} = 0.4)$ (data not shown).

Side-Chain Packing Differences May Lead to Native Preference in Bdpa

The native state of Bdpa consists of three amphipathic helices that turn their hydrophobic faces inward to form a common hydrophobic core. Packing the helices in the mirror configuration disrupts the hydrophobic core, which can be restored through

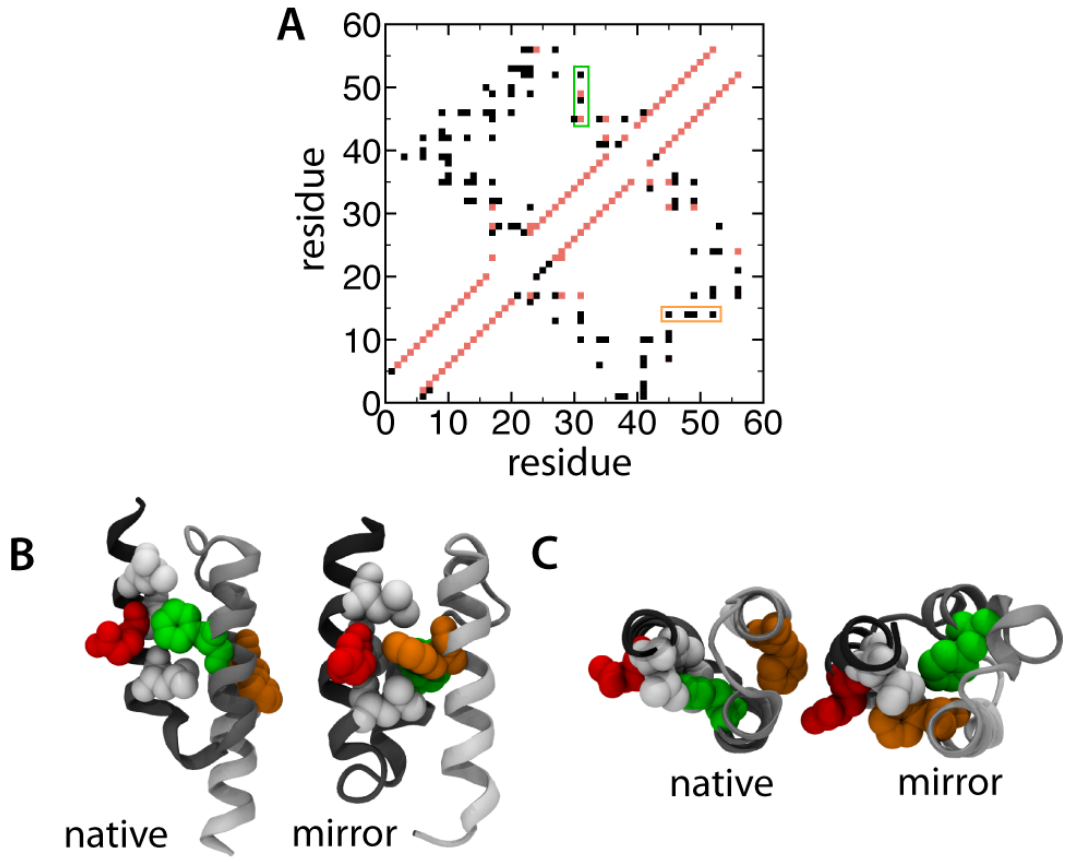

Figure 6.6: Hydrophobic core packing differs between the native and mirror in Bdpa. **A**: Native contact map for the native (black) and mirror (red). The contacts shared between the two structures are colored blue. While the two configurations share nearly all $(i,i+4)$ α-helical contacts, they share only 8 tertiary contacts. **B**: Detailed view of the tertiary packing of Phe14 (orange) and Phe31 (green) with hydrophobic residues Leu45, Ala49, and Leu52 (white) and charged Glu48 (red). **C**: Rotated view shows Phe31 inserted into the heart of the core of the mirror. The rectangles in (**A**) highlight the packing of Phe14 and Phe31.

a reorientation of the side-chain packing (Figure 6.1). This reorientation completely reorganizes the identities of the side chains that pack together. Analysis of REMD shows that the native and mirror configurations have few hydrophobic core contacts in common, and that the native configuration is more tightly packed than the mirror configuration.

Figure 6.6A compares the contact maps of the structures in the native basin to the structures in the mirror basin. (See Section 6.2.4 for a description of the contact maps.) This analysis showed that the native and mirror basins had only 8 tertiary residue contacts in common. While the two basins share nearly all secondary structure α-helical contacts, turn 1 (N-terminal) diverges between the two structures. This difference in turn 1 is a consequence of helix 1 (N-terminal) packing against helices 2 and 3 with different registers[2] between the two basins. Notice that in Figure 6.1A, the N-terminus of the native is one full α-helical turn farther from the C-terminus compared to the mirror. The register shift combined with the helical rearrangement makes the hydrophobic core packing between the two basins very different.

The contact map calculation shows that the native basin, and its representative structure S_{N}, is more tightly packed than its mirror counterparts. The native basin contact map C_{N} has 65 tertiary residue contacts and S_{N}'s contact map $C_{\mathrm{N}}^{\mathrm{AA}}$ has 254 tertiary atomic contacts (444 total), while C_{M} has 51 tertiary residue contacts and $C_{\mathrm{M}}^{\mathrm{AA}}$ has 194 tertiary atomic contacts (373 total). This difference in contact number is a contributing factor in stabilizing the native basin relative to the mirror basin when going from a single-basin SBM to a dual-basin SBM (Figure 6.5C). In order to equalize the energetics between the native and mirror structures, each mirror contact must be strengthened by a factor of $444/373 = 1.2$ (Section 6.2.3). This means that even though the native basin and mirror basin overlap structurally in equal amounts, *i.e.* structures with $Q_{\mathrm{N}}^{\mathrm{AA}} = 0.5$ average 30 $C_{\mathrm{M}}^{\mathrm{AA}}$ tertiary contacts and structures with $Q_{\mathrm{M}}^{\mathrm{AA}} = 0.4$ average 31 $C_{\mathrm{N}}^{\mathrm{AA}}$ tertiary contacts, since a mirror contact formed in the native basin is more favorable than a native contact made in the mirror basin, the overlap of the two folded basins energetically favors the native basin. $\langle Q_{\mathrm{M}}^{\mathrm{AA}}(Q_{\mathrm{N}}^{\mathrm{AA}} = 0.5)\rangle = 0.15$, while $\langle Q_{\mathrm{N}}^{\mathrm{AA}}(Q_{\mathrm{M}}^{\mathrm{AA}} = 0.4)\rangle = 0.12$.

[2] Amphipathic helices can be broken in heptad repeats with the first and fourth residues hydrophobic. These two residues are colored greyscale in Figure 6.1. If two helices have the same interacting heptad repeats as in the native structure, then the hydrophobic packing between them is considered "in register."

The poorly packed mirror conformation may be caused by the difficulty in packing Phe31. In S_N, Phe31 is tightly sandwiched between Leu45 and Leu52 (Figure 6.6B,C), reminiscent of interactions between coiled-coils (*91*). As the hydrophobic core reorganizes to compensate for the mirror helical packing, Phe14 is similarly sandwiched between Leu45 and Leu52. Phe31 though cannot mimic the native packing of Phe14 in the mirror configuration, so it is buried into the middle of the hydrophobic core, disrupting the packing of the core along with distorting both turn 1 and the beginning of helix 2. The benefit of sandwiching Phe14 between Leu45 and Leu52 may be the reason why helix 1 undergoes a register shift in S_M.

6.4 Discussion and Conclusions

We have presented a thorough study of the occurrence of mirror images in simple, symmetrical proteins. A coarse-grained, single-funnel protein model highlights the near structural degeneracy between the native and mirror helical arrangements. It shows that the right-handed mirror helical arrangement is easily accessible, with roughly 5% occupation. The existence of suitable hydrophobic core packing in the mirror conformations is demonstrated by all-atom structure prediction with PFF01/02. It produces competitive folded enthalpies between native and mirror configurations of Bdpa (*173*), Edpa and α_3d. Equilibrium folding simulations of Bdpa, using REMD in Amber94, result in the native cluster being only three times more occupied than the mirror cluster. Relaxation to equilibrium in the REMD suggests that the kinetics of folding to the mirror are faster, and thus that the mirror configuration can function as a kinetic trap during folding. Folding simulations of single- and dual-basin all-atom SBMs constructed from the REMD structures corroborate the REMD thermodynamic and kinetic findings.

These results are not only self consistent, they also agree with previous studies on protein folding and structure prediction using empirical forcefields that also suggested the presence of mirror image folds (*169–173, 187, 188*). Favrin *et al.* (*172*) studied a reduced model protein based on Bdpa using an empirical model focused on hydrogen bonding and hydrophobicity. The authors noted that it was difficult to differentiate between the native fold and its mirror image in their pariwise-additive potential. Scheraga

and coworkers (*170*) studied the folding of Bdpa and apo calbindin D9K in a thermodynamic framework using the UNRES forcefield (*189*). They were successful in locating the native state, but they also encountered mirror images for both proteins, and noted that the mirrors were difficult to discriminate based only on their energies (within a few kcal/mol). More recently, Scheraga and coworkers (*188*) showed using comprehensive sampling of Bdpa with the UNRES forcefield, that not only do folding trajectories at low temperature often visit the mirror configuration as kinetic trap, but also that at folding temperature the native and mirror configurations are equally populated. While the occurrence of mirror images has often been regarded as a deficiency of empirical energy functions, their consistent observation among diverse forcefields, across multiple proteins, and from coarse-grained to all-atom representations, is convincing evidence that mirrored protein conformations are truly competitive.

Solution NMR has been performed on three homologous domains of protein A, Bdpa (*91*), Edpa (*190*), and the Z-domain (Zdpa) (*191*), which has two mutations relative to Bdpa, Ala1 $\rightarrow$ Val and Gly29 $\rightarrow$ Ala. There are also x-ray crystallography structures of Bdpa in complex with human IgG Fc fragment (Bdpa-Fc) (*192*) and Ddpa in complex with human IgM (Ddpa-Fab) (*176*). Bdpa-Fc lacks helix 3 and has no coordinates at all for Ala49-Lys59, while all three solution NMR structures and Ddpa-Fab show a tightly packed three-helix bundle with some N-terminal fraying of helix 1. HD-exchange experiments, though, do show protection of the helix 3 hydrogen bonds in Bdpa-Fc, which suggests that the lack of helix 3 in the crystal structure may be a crystal artifact. The solution NMR structures and Ddpa-Fab are largely consistent, but all show different orientations of helical packings. In Bdpa, helix 1 is tilted 30° with respect to helices 2 and 3, while in Zdpa and Ddpa-Fab, helix 1 is only tilted 15° and in Bdpa-Fc helix 1 and helix 2 are nearly parallel. Though none of the experimental structures display the mirror image, the native structural heterogeneity can be taken as a sign of frustration in the domain. Likely, non-specific hydrophobic interactions give rise to many adequate hydrophobic core packings. Protein A is a virulence factor, one of its goals is to bind strongly to immunoglobulins. Achieving strong binding may require the domain to have considerable flexibility (*167, 193*). Helix swapping may be part of the functional dynamics of the system and these proteins may adopt one conformation

or the other depending on the proteins to which they bind.

Regardless of any functional advantages of mirror images, energy landscape considerations predict that the two symmetric helical arrangements of three-helix bundles should only marginally differ in stability. The principle of minimal frustration (*6, 7*) explains that evolution works to ensure a sequence is consistent with its structure. Geometry therefore becomes the prime determinant of the folding landscape. When symmetry leads to degeneracy between protein structures, the choice between symmetrical helical arrangements occurs at a finer energy scale. Our results show this difference may be as small as a few k_BT. As conditions change, or new interaction partners are introduced, the energy landscape is altered and the protein may fall through a trapdoor (*21*) to its symmetric structural neighbor. In the case of Bdpa REMD, temperature stabilizes the mirror relative to native, similar to the Rop-dimer. Therefore, like the Rop-dimer, the Bdpa mirror may also be stabilized by denaturant (*168*). Another intriguing possibility is that a Phe31 $\rightarrow$ Ala mutant will open the trapdoor. If these mirror conformations indeed exist as meta-stable excitations from the native basin, sensitive NMR experiments should be able to capture their signatures (*194*). More experimental studies are needed in order to quantify the energy landscapes of symmetric protein structures.

6.5 Acknowledgements

Chapter 6, in part, appears in *Journal of Physical Chemistry B*, (2012, in press), Noel, Schug, Verma, Wenzel, Garcia, Onuchic. The dissertation author was the primary investigator and the author of the paper. The work was supported by the Center for Theoretical Biological Physics sponsored by the National Science Foundation (NSF) (Grant PHY-0822283) and NSF Grant NSF-MCB-1051438 to JNO, and NSF-MCB-1050966 to AEG. AS and AV acknowledge the Impuls- und Vernetzungsfond of the Helmholtz Association of German Research Centers. WW is supported by BW Stiftung grant HPC-5.

Bibliography

1. McCammon, J. A, Gelin, B. R, & Karplus, M. (1977) Dynamics of folded proteins. *Nature* **267**, 585–590.

2. Adcock, S. A & McCammon, J. A. (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* **106**, 1589–1615.

3. Shaw, D. E, Maragakis, P, Lindorff-Larsen, K, Piana, S, Dror, R. O, Eastwood, M. P, Bank, J. A, Jumper, J. M, Salmon, J. K, Shan, Y, & Wriggers, W. (2010) Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346.

4. Whitford, P. C, Geggier, P, Altman, R. B, Blanchard, S. C, Onuchic, J. N, & Sanbonmatsu, K. Y. (2010) Accommodation of aminoacyl-trna into the ribosome involves reversible excursions along multiple pathways. *rna* **16**, 1196–1204.

5. Lindorff-Larsen, K, Piana, S, Dror, R. O, & Shaw, D. E. (2011) How fast-folding proteins fold. *Science* **334**, 517–520.

6. Onuchic, J. N & Wolynes, P. G. (2004) Theory of protein folding. *Curr. Opin. Struct. Biol.* **14**, 70–75.

7. Bryngelson, J & Wolynes, P. (1987) Spin glasses and the statistical mechanics of protein folding. *Proc. Nat. Acad. Sci. USA* **84**, 7524.

8. Leopold, P. E, Montal, M, & Onuchic, J. N. (1992) Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Nat. Acad. Sci. USA* **89**, 8721–8725.

9. Bryngelson, J. D, Onuchic, J. N, Socci, N. D, & Wolynes, P. G. (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167–195.

10. Shoemaker, B. A, Wang, J, & Wolynes, P. G. (1997) Structural correlations in protein folding funnels. *Proc. Nat. Acad. Sci. USA* **94**, 777–782.

11. Nymeyer, H, García, A. E, & Onuchic, J. N. (1998) Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Nat. Acad. Sci. USA* **95**, 5921–5928.

12. Clementi, C, Jennings, P. A, & Onuchic, J. N. (2001) Prediction of folding mechanism for circular-permuted proteins. *J. Mol. Biol.* **311**, 879–890.

13. Clementi, C, Nymeyer, H, & Onuchic, J. N. (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.* **298**, 937–953.

14. Gosavi, S, Chavez, L. L, Jennings, P. A, & Onuchic, J. N. (2006) Topological frustration and the folding of interleukin-1 beta. *J. Mol. Biol.* **357**, 986–996.

15. Levy, Y & Onuchic, J. N. (2006) Mechanisms of protein assembly: lessons from minimalist models. *Acc Chem Res* **39**, 135–142.

16. Levy, Y, Cho, S. S, Shen, T, Onuchic, J. N, & Wolynes, P. G. (2005) Symmetry and frustration in protein energy landscapes: a near degeneracy resolves the rop dimer-folding mystery. *Proc. Nat. Acad. Sci. USA* **102**, 2373–2378.

17. Levy, Y, Cho, S. S, Onuchic, J. N, & Wolynes, P. G. (2005) A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. *J. Mol. Biol.* **346**, 1121–1145.

18. Yang, S, Cho, S. S, Levy, Y, Cheung, M. S, Levine, H, Wolynes, P. G, & Onuchic, J. N. (2004) Domain swapping is a consequence of minimal frustration. *Proc. Nat. Acad. Sci. USA* **101**, 13786–13791.

19. Whitford, P. C, Miyashita, O, Levy, Y, & Onuchic, J. N. (2007) Conformational transitions of adenylate kinase: switching by cracking. *J. Mol. Biol.* **366**, 1661–1671.

20. Whitford, P. C, Gosavi, S, & Onuchic, J. N. (2008) Conformational transitions in adenylate kinase. allosteric communication reduces misligation. *J. Biol. Chem.* **283**, 2042–2048.

21. Schug, A, Whitford, P. C, Levy, Y, & Onuchic, J. N. (2007) Mutations as trapdoors to two competing native conformations of the rop-dimer. *Proc. Nat. Acad. Sci. USA* **104**, 17674–17679.

22. Okazaki, K, Koga, N, Takada, S, Onuchic, J. N, & Wolynes, P. G. (2006) Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Nat. Acad. Sci. USA* **103**, 11844–11849.

23. Best, R. B, Chen, Y.-G, & Hummer, G. (2005) Slow protein conformational dynamics from multiple experimental structures: the helix/sheet transition of arc repressor. *Structure* **13**, 1755–1763.

24. Zuckerman, D. M. (2004) Simulation of an ensemble of conformational transitions in a united-residue model of calmodulin. *J. Phys. Chem. B* **108**, 5127–5137.

25. Zwanzig, R. (1988) Diffusion in a rough potential. *Proc. Nat. Acad. Sci. USA* **85**, 2029–2030.

26. Bryngelson, J & Wolynes, P. (1989) Intermediates and barrier crossing in a random energy model (with applications to protein folding). *J. Phys. Chem.* **93**, 6902–6915.

27. Clementi, C & Plotkin, S. S. (2004) The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Sci.* **13**, 1750–1766.

28. Baker, D. (2000) A surprising simplicity to protein folding. *Nature* **405**, 39–42.

29. Chavez, L. L, Onuchic, J. N, & Clementi, C. (2004) Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.* **126**, 8426–8432.

30. Zwanzig, R, Szabo, A, & Bagchi, B. (1992) Levinthal's paradox. *Proc. Nat. Acad. Sci. USA* **89**, 20–22.

31. Tirion, M. (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.* **77**, 1905–1908.

32. Miyashita, O, Onuchic, J. N, & Wolynes, P. G. (2003) Nonlinear elasticity, proteinquakes, and the energy landscapes of functional transitions in proteins. *Proc. Nat. Acad. Sci. USA* **100**, 12570–12575.

33. Hyeon, C & Onuchic, J. N. (2007) Mechanical control of the directional stepping dynamics of the kinesin motor. *Proc. Nat. Acad. Sci. USA* **104**, 17382–17387.

34. Hyeon, C, Jennings, P. A, Adams, J. A, & Onuchic, J. N. (2009) Ligand-induced global transitions in the catalytic domain of protein kinase a. *Proc. Nat. Acad. Sci. USA* **106**, 3023–3028.

35. Whitford, P. C, Noel, J. K, Gosavi, S, Schug, A, Sanbonmatsu, K. Y, & Onuchic, J. N. (2009) An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins* **75**, 430–441.

36. Levy, Y, Onuchic, J. N, & Wolynes, P. G. (2007) Fly-casting in protein-dna binding: frustration between protein folding and electrostatics facilitates target recognition. *J. Am. Chem. Soc.* **129**, 738–739.

37. Cho, S. S, Weinkam, P, & Wolynes, P. G. (2008) Origins of barriers and barrierless folding in bbl. *Proc. Nat. Acad. Sci. USA* **105**, 118–123.

38. Azia, A & Levy, Y. (2009) Nonnative electrostatic interactions can modulate protein folding: Molecular dynamics with a grain of salt. *J. Mol. Biol.* **393**, 527–542.

39. Cheung, M. S, García, A. E, & Onuchic, J. N. (2002) Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Nat. Acad. Sci. USA* **99**, 685–690.

40. Scott, K. A, Batey, S, Hooton, K. A, & Clarke, J. (2004) The folding of spectrin domains i: wild-type domains have the same stability but very different kinetic properties. *J. Mol. Biol.* **344**, 195–205.

41. Ferguson, N, Schartau, P. J, Sharpe, T. D, Sato, S, & Fersht, A. R. (2004) One-state downhill versus conventional protein folding. *J. Mol. Biol.* **344**, 295–301.

42. Sutto, L, Latzer, J, Hegler, J. A, Ferreiro, D. U, & Wolynes, P. G. (2007) Consequences of localized frustration for the folding mechanism of the im7 protein - supplement. *Proc. Nat. Acad. Sci. USA* **104**, 19825–19830.

43. Taketomi, H, Ueda, Y, & Gō, N. (1975) Studies on protein folding, unfolding and fluctuations by computer simulation. i. the effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Prot. Res.* **7**, 445–459.

44. Socci, N, Onuchic, J. N, & Wolynes, P. G. (1996) Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860–5868.

45. Veitshans, T, Klimov, D, & Thirumalai, D. (1997) Protein folding kinetics: timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Folding and Design* **2**, 1–22.

46. Koga, N & Takada, S. (2001) Roles of native topology and chain-length scaling in protein folding: a simulation study with a go-like model. *J. Mol. Biol.* **313**, 171–180.

47. Kaya, H & Chan, H. S. (2003) Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? *J. Mol. Biol.* **326**, 911–931.

48. Andrews, B. T, Gosavi, S, Finke, J. M, Onuchic, J. N, & Jennings, P. A. (2008) The dual-basin landscape in gfp folding. *Proc. Nat. Acad. Sci. USA* **105**, 12283–12288.

49. Sułkowska, J. I, Sułkowski, P, & Onuchic, J. (2009) Dodging the crisis of folding proteins with knots. *Proc. Nat. Acad. Sci. USA* **106**, 3119–3124.

50. Whitford, P. C, Schug, A, Saunders, J, Hennelly, S. P, Onuchic, J. N, & Sanbonmatsu, K. Y. (2009) Nonlocal helix formation is key to understanding s-adenosylmethionine-1 riboswitch function. *Biophys. J.* **96**, L7–9.

51. Noel, J. K, Sulkowska, J. I, & Onuchic, J. N. (2010) Slipknotting upon native-like loop formation in a trefoil knot protein. *Proc. Nat. Acad. Sci. USA* **107**, 15403–15408.

52. Nechushtai, R, Lammert, H, Michaeli, D, Eisenberg-Domovich, Y, Zuris, J. A, Luca, M. A, Capraro, D. T, Fish, A, Shimshon, O, Roy, M, Schug, A, Whitford, P. C, Livnah, O, Onuchic, J. N, & Jennings, P. A. (2011) Allostery in the ferredoxin protein motif does not involve a conformational switch. *Proc. Nat. Acad. Sci. USA* **108**, 2240–2245.

53. Jamros, M. A, Oliveira, L. C, Whitford, P. C, Onuchic, J. N, Adams, J. A, Blumenthal, D. K, & Jennings, P. A. (2010) Proteins at work: a combined small angle x-ray scattering and theoretical determination of the multiple structures involved on the protein kinase functional landscape. *J. Biol. Chem.* **285**, 36121–36128.

54. Ratje, A. H, Loerke, J, Mikolajka, A, Brünner, M, Hildebrand, P. W, Starosta, A. L, Dönhöfer, A, Connell, S. R, Fucini, P, Mielke, T, Whitford, P. C, Onuchic, J. N, Yu, Y, Sanbonmatsu, K. Y, Hartmann, R. K, Penczek, P. A, Wilson, D. N, & Spahn, C. M. T. (2010) Head swivel on the ribosome facilitates translocation by means of intra-subunit trna hybrid sites. *Nature* **468**, 713–716.

55. Schug, A, Weigt, M, Onuchic, J. N, Hwa, T, & Szurmant, H. (2009) High-resolution protein complexes from integrating genomic information with molecular simulation. *Proc. Nat. Acad. Sci. USA* **106**, 22124–22129.

56. Noel, J. K, Whitford, P. C, & Onuchic, J. N. (2012) The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function. *J. Phys. Chem. B* p. 1.

57. Noel, J. K, Whitford, P. C, Sanbonmatsu, K. Y, & Onuchic, J. N. (2010) Smog@ctbp: simplified deployment of structure-based models in gromacs. *Nucleic Acids Res.* **38**, W657–61.

58. de Araujo, A. F. P & Onuchic, J. N. (2009) A sequence-compatible amount of native burial information is sufficient for determining the structure of small globular proteins. *Proc. Nat. Acad. Sci. USA* **106**, 19001–19004.

59. Lammert, H, Wolynes, P. G, & Onuchic, J. N. (2012) The role of atomic level steric effects and attractive forces in protein folding. *Proteins* **80**, 362–373.

60. Lammert, H, Schug, A, & Onuchic, J. N. (2009) Robustness and generalization of structure-based models for protein folding and function. *Proteins* **77**, 881–891.

61. Mor, A, Ziv, G, & Levy, Y. (2008) Simulations of proteins with inhomogeneous degrees of freedom: The effect of thermostats. *J. Comput. Chem.* **29**, 1992–1998.

62. Sułkowska, J, Sułkowski, P, Szymczak, P, & Cieplak, M. (2008) Tightening of knots in proteins. *Phys. Rev. Lett.* **100**, 058106.

63. Cho, S, Levy, Y, & Wolynes, P. G. (2006) P versus q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Nat. Acad. Sci. USA* **103**, 586–591.

64. Kouza, M, Li, M. S, O'brien, E. P, Hu, C.-K, & Thirumalai, D. (2006) Effect of finite size on cooperativity and rates of protein folding. *J. Phys. Chem. A* **110**, 671–676.

65. Oliveira, R. J, Whitford, P. C, Chahine, J, Wang, J, Onuchic, J. N, & Leite, V. B. P. (2010) The origin of nonmonotonic complex behavior and the effects of nonnative interactions on the diffusive properties of protein folding. *Biophys. J.* **99**, 600–608.

66. Kumar, S, Rosenberg, J, Bouzida, D, & Swendsen. (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *J. Comput. Chem.* **13**, 1011.

67. Fersht, A. R. (1995) Characterizing transition-states in protein-folding - an essential step in the puzzle. *Curr. Opin. Struct. Biol.* **5**, 79–84.

68. Levy, Y, Wolynes, P. G, & Onuchic, J. N. (2004) Protein topology determines binding mechanism. *Proc. Nat. Acad. Sci. USA* **101**, 511–516.

69. Baxter, E. L, Jennings, P. A, & Onuchic, J. N. (2011) Interdomain communication revealed in the diabetes drug target mitoneet. *Proc. Nat. Acad. Sci. USA* **108**, 5266–5271.

70. Eastwood, M, Hardin, C, Luthey-Schulten, Z, & Wolynes, P. (2001) Evaluating protein structure-prediction schemes using energy landscape theory. *IBM J. Res. Dev.* **45**, 475–497.

71. Rohl, C. A, Strauss, C. E. M, Misura, K. M. S, & Baker, D. (2004) Protein structure prediction using rosetta. *Methods Enzymol.* **383**, 66–93.

72. Harpaz, Y, Elmasry, N, Fersht, A. R, & Henrick, K. (1994) Direct observation of better hydration at the n terminus of an α-helix with glycine rather than alanine as the n-cap residue. *Proc. Nat. Acad. Sci. USA* **91**, 311–315.

73. Humphrey, W, Dalke, A, & Schulten, K. (1996) Vmd: visual molecular dynamics. *J Mol Graph* **14**, 33–8, 27–8.

74. Hess, B, Kutzner, C, van der Spoel, D, & Lindahl, E. (2008) Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4**, 435–447.

75. Ponder, J. W & Case, D. A. (2003) Force fields for protein simulations. *Adv. Protein Chem.* **66**, 27–85.

76. Phillips, J. C, Braun, R, Wang, W, Gumbart, J, Tajkhorshid, E, Villa, E, Chipot, C, Skeel, R. D, Kalé, L, & Schulten, K. (2005) Scalable molecular dynamics with namd. *J. Comput. Chem.* **26**, 1781–1802.

77. Brooks, B. R, Bruccoleri, R. E, Olafson, B. D, States, D. J, Swaminathan, S, & Karplus, M. (1983) Charmm - a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.

78. Eisenberg, D & McLachlan, A. D. (1986) Solvation energy in protein folding and binding. *Nature* **319**, 199–203.

79. Zhou, R. (2003) Trp-cage: folding free energy landscape in explicit water. *Proc. Nat. Acad. Sci. USA* **100**, 13280–13285.

80. Paschek, D, Nymeyer, H, & García, A. E. (2007) Replica exchange simulation of reversible folding/unfolding of the trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water. *J. Struct. Biol.* **157**, 524–533.

81. Duan, Y & Kollman, P. A. (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**, 740–744.

82. García, A. E & Onuchic, J. N. (2003) Folding a protein in a computer: an atomic description of the folding/unfolding of protein a. *Proc. Nat. Acad. Sci. USA* **100**, 13898–13903.

83. Schug, A, Herges, T, & Wenzel, W. (2003) Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Lett.* **91**, 158102.

84. Schug, A, Herges, T, Verma, A, Lee, K. H, & Wenzel, W. (2005) Comparison of stochastic optimization methods for all-atom folding of the trp-cage protein. *ChemPhysChem* **6**, 2640–2646.

85. Schug, A & Wenzel, W. (2006) An evolutionary strategy for all-atom folding of the 60-amino-acid bacterial ribosomal protein l20. *Biophys. J.* **90**, 4273–4280.

86. Jayachandran, G, Vishal, V, & Pande, V. S. (2006) Using massively parallel simulation and markovian models to study protein folding: examining the dynamics of the villin headpiece. *J. Chem. Phys.* **124**, 164902.

87. Freddolino, P. L, Liu, F, Gruebele, M, & Schulten, K. (2008) Ten-microsecond molecular dynamics simulation of a fast-folding ww domain. *Biophys. J.* **94**, L75–7.

88. Clementi, C, García, A. E, & Onuchic, J. N. (2003) Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: all-atom representation study of protein l. *J. Mol. Biol.* **326**, 933–954.

89. Shimada, J, Kussell, E. L, & Shakhnovich, E. I. (2001) The folding thermodynamics and kinetics of crambin using an all-atom monte carlo simulation. *J. Mol. Biol.* **308**, 79–95.

90. Linhananta, A & Zhou, Y. (2002) The role of sidechain packing and native contact interactions in folding: Discontinuous molecular dynamics folding simulations of an all-atom gō model of fragment b of staphylococcal protein a. *J. Chem. Phys.* **117**, 8983–8995.

91. Gouda, H, Torigoe, H, Saito, A, Sato, M, Arata, Y, & Shimada, I. (1992) Three-dimensional solution structure of the b domain of staphylococcal protein a: comparisons of the solution and crystal structures. *Biochemistry* **31**, 9665–9672.

92. Xu, W, Harrison, S. C, & Eck, M. J. (1997) Three-dimensional structure of the tyrosine kinase c-src. *Nature* **385**, 595–602.

93. Sato, S, Religa, T. L, Daggett, V, & Fersht, A. R. (2004) Testing protein-folding simulations by experiment: B domain of protein a. *Proc. Nat. Acad. Sci. USA* **101**, 6952–6956.

94. Viguera, A. R, Martínez, J. C, Filimonov, V. V, Mateo, P. L, & Serrano, L. (1994) Thermodynamic and kinetic analysis of the sh3 domain of spectrin shows a two-state folding transition. *Biochemistry* **33**, 2142–2150.

95. Jackson, S. E & Fersht, A. R. (1991) Folding of chymotrypsin inhibitor 2. 1. evidence for a two-state transition. *Biochemistry* **30**, 10428–10435.

96. Shea, J.-E, Onuchic, J. N, & Brooks, C. L. (2002) Probing the folding free energy landscape of the src-sh3 protein domain. *Proc. Nat. Acad. Sci. USA* **99**, 16064–16068.

97. Hoang, T & Cieplak, M. (2000) Sequencing of folding events in go-type proteins. *J. Chem. Phys.* **113**, 8319–8328.

98. Shea, J, Onuchic, J, & III, C. B. (1999) Exploring the origins of topological frustration: Design of a minimally frustrated model of fragment b of protein a. *Proc. Nat. Acad. Sci. USA* **96**, 12512–12517.

99. Plaxco, K. W, Simons, K. T, & Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994.

100. Prieto, L & Rey, A. (2007) Influence of the chain stiffness on the thermodynamics of a gō-type model for protein folding. *J. Chem. Phys.* **126**, 165103.

101. Kramers, H. (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **7**, 284–304.

102. Chahine, J, Oliveira, R. J, Leite, V. B. P, & Wang, J. (2007) Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding. *Proc. Nat. Acad. Sci. USA* **104**, 14646–14651.

103. Socolich, M, Lockless, S. W, Russ, W. P, Lee, H, Gardner, K. H, & Ranganathan, R. (2005) Evolutionary information for specifying a protein fold. *Nature* **437**, 512–518.

104. Gosavi, S, Whitford, P. C, Jennings, P. A, & Onuchic, J. N. (2008) Extracting function from a β-trefoil folding motif. *Proc. Nat. Acad. Sci. USA* **105**, 10384–10389.

105. Miyashita, O, Wolynes, P. G, & Onuchic, J. N. (2005) Simple energy landscape model for the kinetics of functional transitions in proteins. *J. Phys. Chem. B* **109**, 1959–1969.

106. Hyeon, C & Onuchic, J. N. (2007) Internal strain regulates the nucleotide binding site of the kinesin leading head. *Proc. Nat. Acad. Sci. USA* **104**, 2175–2180.

107. Berendsen, H. J. C, Postma, J, van Gunsteren, W. F, DiNola, A, & Haak, J. R. (1984) Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690.

108. Ferrenberg, A & Swendsen, R. (1988) New monte carlo technique for studying phase transitions. *Phys. Rev. Lett.* **61**, 2635–2638.

109. Ferrenberg, A & Swendsen, R. (1989) Optimized monte carlo data analysis. *Phys. Rev. Lett.* **63**, 1195–1198.

110. Jorgensen, W. L & Tirado-Rives, J. (1988) The opls [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* **110**, 1657–1666.

111. Jorgensen, W, Chandrasekhar, J, Madura, J. D, Impey, R. W, & Klein, M. L. (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935.

112. Sobolev, V, Wade, R. C, Vriend, G, & Edelman, M. (1996) Molecular docking using surface complementarity. *Proteins* **25**, 120–129.

113. Miyazawa, S & Jernigan, R. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules* **18**, 534–552.

114. Silveira, C. H. D, Pires, D. E. V, Minardi, R. C, Ribeiro, C, Veloso, C. J. M, Lopes, J. C. D, Meira, W, Neshich, G, Ramos, C. H. I, Habesch, R, & Santoro, M. M. (2009) Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins* **74**, 727–743.

115. Sobolev, V, Sorokine, A, Prilusky, J, Abola, E. E, & Edelman, M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics* **15**, 327–332.

116. Sułkowska, J. I & Cieplak, M. (2008) Selection of optimal variants of gō-like models of proteins through studies of stretching. *Biophys. J.* **95**, 3174–3191.

117. Veloso, C. J. M, Silveira, C. H, Melo, R. C, Ribeiro, C, Lopes, J. C. D, Santoro, M. M, & Meira, W. (2007) On the characterization of energy networks of proteins. *Genet. Mol. Res.* **6**, 799–820.

118. Shen, T, Zong, C, Portman, J. J, & Wolynes, P. G. (2008) Variationally determined free energy profiles for structural models of proteins: characteristic temperatures for folding and trapping. *J. Phys. Chem. B* **112**, 6074–6082.

119. Zhang, Z & Chan, H. S. (2010) Competition between native topology and non-native interactions in simple and complex folding kinetics of natural and designed proteins. *Proc. Nat. Acad. Sci. USA* **107**, 2920–2925.

120. Wu, L, Zhang, J, Qin, M, Liu, F, & Wang, W. (2008) Folding of proteins with an all-atom go-model. *J. Chem. Phys.* **128**, 235103.

121. Privalov, P. L & Khechinashvili, N. N. (1974) A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J. Mol. Biol.* **86**, 665–684.

122. Privalov, P. L & Potekhin, S. A. (1986) Scanning microcalorimetry in studying temperature-induced changes in proteins. *Methods Enzymol.* **131**, 4–51.

123. Kaya, H & Chan, H. S. (2000) polymer principles of protein calorimetric two-state cooperativity. *Proteins* **40**, 637–661.

124. Papoian, G. A, Ulander, J, Eastwood, M. P, Luthey-Schulten, Z, & Wolynes, P. G. (2004) Water in protein structure prediction. *Proc. Nat. Acad. Sci. USA* **101**, 3352–3357.

125. Faure, G, Bornot, A, & de Brevern, A. G. (2008) Protein contacts, inter-residue interactions and side-chain modelling. *Biochimie* **90**, 626–639.

126. Williams, M. A, Goodfellow, J. M, & Thornton, J. M. (1994) Buried waters and internal cavities in monomeric proteins. *Protein Sci.* **3**, 1224–1235.

127. Rashin, A. A & Honig, B. (1984) On the environment of ionizable groups in globular proteins. *J. Mol. Biol.* **173**, 515–521.

128. Schafer, H, van Gunsteren, W. F, & Mark, A. E. (1999) Estimating relative free energies from a single ensemble: Hydration free energies. *J. Comput. Chem.* **20**, 1604–1617.

129. Sorin, E. J, Nakatani, B. J, Rhee, Y. M, Jayachandran, G, Vishal, V, & Pande, V. S. (2004) Does native state topology determine the rna folding mechanism? *J. Mol. Biol.* **337**, 789–797.

130. Hyeon, C, Dima, R. I, & Thirumalai, D. (2006) Pathways and kinetic barriers in mechanical unfolding and refolding of rna and proteins. *Structure* **14**, 1633–1645.

131. Whitford, P. C, Ahmed, A, Yu, Y, Hennelly, S. P, Tama, F, Spahn, C. M. T, Onuchic, J. N, & Sanbonmatsu, K. Y. (2011) Excited states of ribosome translocation revealed through integrative molecular modeling. *Proc. Nat. Acad. Sci. USA* **108**, 18943–18948.

132. Whitford, P. C, Sanbonmatsu, K. Y, & Onuchic, J. N. (2012) Energy landscapes of biomolecular folding and function. *Reports on Progress in Physics (in press)*.

133. Garcia, A, Krumhansl, J, & Frauenfelder, H. (1997) Variations on a theme by debye and waller: From simple crystals to proteins. *Proteins* **29**, 153–160.

134. Qi, X & Portman, J. J. (2007) Excluded volume, local structural cooperativity, and the polymer physics of protein folding rates. *Proc. Nat. Acad. Sci. USA* **104**, 10841–10846.

135. Suzuki, Y, Noel, J. K, & Onuchic, J. N. (2011) A semi-analytical description of protein folding that incorporates detailed geometrical information. *J. Chem. Phys.* **134**, 245101.

136. Prieto, L, de Sancho, D, & Rey, A. (2005) Thermodynamics of gō-type models for protein folding. *J. Chem. Phys.* **123**, 154903.

137. Suzuki, Y, Noel, J. K, & Onuchic, J. N. (2008) An analytical study of the interplay between geometrical and energetic effects in protein folding. *J. Chem. Phys.* **128**, 025101.

138. Eastwood, M & Wolynes, P. (2001) Role of explicitly cooperative interactions in protein folding funnels: A simulation study. *J. Chem. Phys.* **114**, 4702.

139. Bowers, K, Chow, E, Xu, H, Dror, R, Eastwood, M, Gregersen, B, Klepeis, J, Kolossvary, I, Moraes, M, Sacerdoti, F, Salmon, J, Shan, Y, & Shaw, D. (2006) Scalable algorithms for molecular dynamics simulations on commodity clusters. *Proc. ACM/IEEE* p. 43.

140. Orzechowski, M & Tama, F. (2008) Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys. J.* **95**, 5692–5705.

141. Mansfield, M. L. (1994) Are there knots in proteins? *Nat. Struct. Mol. Biol.* **1**, 213–214.

142. Virnau, P, Mirny, L. A, & Kardar, M. (2006) Intricate knots in proteins: Function and evolution. *PLOS Comput. Biol.* **2**, 1074–1079.

143. Taylor, W. R. (2007) Protein knots and fold complexity: some new twists. *Comput. Biol. Chem.* **31**, 151–162.

144. King, N. P, Yeates, E. O, & Yeates, T. O. (2007) Identification of rare slipknots in proteins and their implications for stability and folding. *J. Mol. Biol.* **373**, 153–166.

145. Bölinger, D, Sułkowska, J. I, Hsu, H.-P, Mirny, L. A, Kardar, M, Onuchic, J. N, & Virnau, P. (2010) A stevedore's protein knot. *PLOS Comput. Biol.* **6**, e1000731.

146. Mallam, A. L, Rogers, J. M, & Jackson, S. E. (2010) Experimental detection of knotted conformations in denatured proteins. *Proc. Nat. Acad. Sci. USA* **107**, 8189–8194.

147. Mallam, A. L, Morris, E. R, & Jackson, S. E. (2008) Exploring knotting mechanisms in protein folding. *Proc. Nat. Acad. Sci. USA* **105**, 18740–18745.

148. Mallam, A. L & Jackson, S. E. (2011) Knot formation in newly translated proteins is spontaneous and accelerated by chaperonins. *Nat. Chem. Biol.* **8**, 147–153.

149. Sulkowska, J. I, Sulkowski, P, & Onuchic, J. N. (2009) Jamming proteins with slipknots and their free energy landscape. *Phys. Rev. Lett.* **103**, 268103.

150. Wallin, S, Zeldovich, K. B, & Shakhnovich, E. I. (2007) The folding mechanics of a knotted protein. *J. Mol. Biol.* **368**, 884–893.

151. Thirumalai, D & Klimov, D. K. (1999) Deciphering the timescales and mechanisms of protein folding using minimal off-lattice models. *Curr. Opin. Struct. Biol.* **9**, 197–207.

152. Norcross, T. S & Yeates, T. O. (2006) A framework for describing topological frustration in models of protein folding. *J. Mol. Biol.* **362**, 605–621.

153. Bult, C. J, White, O, Olsen, G. J, Zhou, L, Fleischmann, R. D, Sutton, G. G, Blake, J. A, FitzGerald, L. M, Clayton, R. A, Gocayne, J. D, Kerlavage, A. R, Dougherty, B. A, Tomb, J.-F, Adams, M. D, Reich, C. I, Overbeek, R, Kirkness, E. F, Weinstock, K. G, Merrick, J. M, Glodek, A, Scott, J. L, Geoghagen, N. S. M, Weidman, J. F, Fuhrmann, J. L, Nguyen, D, Utterback, T. R, Kelley, J. M, Peterson, J. D, Sadow, P. W, Hanna, M. C, Cotton, M. D, Roberts, K. M, Hurst, M. A, Kaine, B. P, Borodovsky, M, Klenk, H.-P, Fraser, C. M, Smith, H. O, Woese, C. R, & Venter, J. C. (1996) Complete genome sequence of the methanogenic archaeon, methanococcus jannaschii. *Science* **273**, 1058–1073.

154. Koniaris, K & Muthukumar, M. (1991) Knottedness in ring polymers. *Phys. Rev. Lett.* **66**, 2211–2214.

155. Virnau, P, Kantor, Y, & Kardar, M. (2005) Knots in globule and coil phases of a model polyethylene. *J. Am. Chem. Soc.* **127**, 15102–15106.

156. Raymer, D. M & Smith, D. E. (2007) Spontaneous knotting of an agitated string. *Proc. Nat. Acad. Sci. USA* **104**, 16432–16437.

157. Huang, L & Makarov, D. E. (2008) Translocation of a knotted polypeptide through a pore. *J. Chem. Phys.* **129**, 121107.

158. Bornschlögl, T, Anstrom, D. M, Mey, E, Dzubiella, J, Rief, M, & Forest, K. T. (2009) Tightening the knot in phytochrome by single-molecule atomic force microscopy. *Biophys. J.* **96**, 1508–1514.

159. Plotkin, S. S & Onuchic, J. N. (2002) Understanding protein folding with energy landscape theory. part i: Basic concepts. *Q. Rev. Biophys.* **35**, 111–167.

160. Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. *Science* **181**, 223–230.

161. Wolynes, P. G. (1996) Symmetry and the energy landscapes of biomolecules. *Proc. Nat. Acad. Sci. USA* **93**, 14249–14255.

162. Kim, D, Fisher, C, & Baker, D. (2000) A breakdown of symmetry in the folding transition state of protein l1. *J. Mol. Biol.* **298**, 971–984.

163. McCallister, E. L, Alm, E, & Baker, D. (2000) Critical role of β-hairpin formation in protein g folding. *Nat. Struct. Mol. Biol.* **7**, 669–673.

164. Zhou, Y & Linhananta, A. (2002) Role of hydrophilic and hydrophobic contacts in folding of the second β-hairpin fragment of protein g: Molecular dynamics simulation studies of an all-atom model. *Proteins* **47**, 154–162.

165. Broom, A, Doxey, A. C, Lobsanov, Y. D, Berthin, L. G, Rose, D. R, Howell, P. L, McConkey, B. J, & Meiering, E. M. (2012) Modular evolution and the origins of symmetry: reconstruction of a three-fold symmetric globular protein. *Structure* **20**, 161–171.

166. Capraro, D. T, Roy, M, Onuchic, J. N, Gosavi, S, & Jennings, P. A. (2012) β-bulge triggers route-switching on the functional landscape of interleukin-1 beta. *Proc. Nat. Acad. Sci. USA* **109**, 1490–1493.

167. Wolynes, P. G. (2004) Latest folding game results: protein a barely frustrates computationalists. *Proc. Nat. Acad. Sci. USA* **101**, 6837–6838.

168. Gambin, Y, Schug, A, Lemke, E. A, Lavinder, J. J, Ferreon, A. C. M, Magliery, T. J, Onuchic, J. N, & Deniz, A. A. (2009) Direct single-molecule observation of a protein living in two opposed native structures. *Proc. Nat. Acad. Sci. USA* **106**, 10153–10158.

169. Kolinski, A & Skolnick, J. (1994) Monte carlo simulations of protein folding. ii. application to protein a, rop, and crambin. *Proteins* **18**, 353–366.

170. Lee, J, Liwo, A, & Scheraga, H. A. (1999) Energy-based de novo protein folding by conformational space annealing and an off-lattice united-residue force field: application to the 10-55 fragment of staphylococcal protein a and to apo calbindin d9k. *Proc. Nat. Acad. Sci. USA* **96**, 2025–2030.

171. Irbäck, A, Sjunnesson, F, & Wallin, S. (2000) Three-helix-bundle protein in a ramachandran model. *Proc. Nat. Acad. Sci. USA* **97**, 13614–13618.

172. Favrin, G, Irbäck, A, & Wallin, S. (2002) Folding of a small helical protein using hydrogen bonds and hydrophobicity forces. *Proteins* **47**, 99–105.

173. Herges, T & Wenzel, W. (2004) An all-atom force field for tertiary structure prediction of helical proteins. *Biophys. J.* **87**, 3100–3109.

174. Bennett, M. J, Choe, S, & Eisenberg, D. (1994) Domain swapping: entangling alliances between proteins. *Proc. Nat. Acad. Sci. USA* **91**, 3127–3131.

175. Munson, M, Anderson, K. S, & Regan, L. (1997) Speeding up protein folding: mutations that increase the rate at which rop folds and unfolds by over four orders of magnitude. *Folding and Design* **2**, 77–87.

176. Graille, M, Stura, E. A, Corper, A. L, Sutton, B. J, Taussig, M. J, Charbonnier, J. B, & Silverman, G. J. (2000) Crystal structure of a staphylococcus aureus protein a domain complexed with the fab fragment of a human igm antibody: structural basis for recognition of b-cell receptors and superantigen activity. *Proc. Nat. Acad. Sci. USA* **97**, 5399–5404.

177. Zhu, Y, Alonso, D. O. V, Maki, K, Huang, C.-Y, Lahr, S. J, Daggett, V, Roder, H, DeGrado, W. F, & Gai, F. (2003) Ultrafast folding of alpha3d: a de novo designed three-helix bundle protein. *Proc. Nat. Acad. Sci. USA* **100**, 15486–15491.

178. Verma, A & Wenzel, W. (2009) A free-energy approach for all-atom protein simulation. *Biophys. J.* **96**, 3483–3494.

179. Schug, A & Wenzel, W. (2004) Predictive in silico all-atom folding of a four-helix protein with a free-energy model. *J. Am. Chem. Soc.* **126**, 16736–16737.

180. Schug, A, Herges, T, & Wenzel, W. (2004) All-atom folding of the three-helix hiv accessory protein with an adaptive parallel tempering method. *Proteins* **57**, 792–798.

181. Strunk, T, Hamacher, K, Hoffgaard, F, Engelhardt, H, Zillig, M. D, Faist, K, Wenzel, W, & Pfeifer, F. (2011) Structural model of the gas vesicle protein gvpa and analysis of gvpa mutants in vivo. *Mol. Microbiol.* **81**, 56–68.

182. Garcia, A, Herce, H, & Paschek, D. (2006) Simulations of temperature and pressure unfolding of peptides and proteins with replica exchange molecular dynamics. *Annu. Rep. Comp. Chem.* **2**, 83–95.

183. Showalter, S. A & Bruschweiler, R. (2007) Validation of molecular dynamics simulations of biomolecules using nmr spin relaxation as benchmarks: Application to the amber99sb force field. *J. Chem. Theory Comput.* **3**, 961–975.

184. Day, R, Paschek, D, & Garcia, A. E. (2010) Microsecond simulations of the folding/unfolding thermodynamics of the trp-cage miniprotein. *Proteins* **78**, 1889–1899.

185. Baxter, E. L, Jennings, P. A, & Onuchic, J. N. (2012) Strand swapping regulates the iron-sulfur cluster in the diabetes drug target mitoneet. *Proc. Nat. Acad. Sci. USA* **109**, 1955–1960.

186. Noel, J. K & Onuchic, J. N. (2012). pp. 31–54.

187. St-Pierre, J, Mousseau, N, & Derreumaux, P. (2008) The complex folding pathways of protein a suggest a multiple-funnelled energy landscape. *J. Chem. Phys.* **128**, 045101.

188. Maisuradze, G. G, Liwo, A, Ołdziej, S, & Scheraga, H. A. (2010) Evidence, from simulations, of a single state with residual native structure at the thermal denaturation midpoint of a small globular protein. *J. Am. Chem. Soc.* **132**, 9444–9452.

189. Rojas, A. V, Liwo, A, & Scheraga, H. A. (2007) Molecular dynamics with the united-residue force field: ab initio folding simulations of multichain proteins. *J. Phys. Chem. B* **111**, 293–309.

190. Starovasnik, M. A, Skelton, N. J, O'Connell, M. P, Kelley, R. F, Reilly, D, & Fairbrother, W. J. (1996) Solution structure of the e-domain of staphylococcal protein a. *Biochemistry* **35**, 15558–15569.

191. Tashiro, M, Tejero, R, Zimmerman, D. E, Celda, B, Nilsson, B, & Montelione, G. T. (1997) High-resolution solution nmr structure of the z domain of staphylococcal protein a. *J. Mol. Biol.* **272**, 573–590.

192. Deisenhofer, J. (1981) Crystallographic refinement and atomic models of a human fc fragment and its complex with fragment b of protein a from staphylococcus aureus at 2.9- and 2.8-a resolution. *Biochemistry* **20**, 2361–2370.

193. Wahlberg, E, Lendel, C, Helgstrand, M, Allard, P, Dincbas-Renqvist, V, Hedqvist, A, Berglund, H, Nygren, P.-A, & Härd, T. (2003) An affibody in complex with a target protein: structure and coupled folding. *Proc. Nat. Acad. Sci. USA* **100**, 3185–3190.

194. Bouvignies, G, Vallurupalli, P, Hansen, D. F, Correia, B. E, Lange, O, Bah, A, Vernon, R. M, Dahlquist, F. W, Baker, D, & Kay, L. E. (2011) Solution structure of a minor and transiently formed state of a t4 lysozyme mutant. *Nature* **477**, 111–114.