

**The Dynamics and Folding Pathways  
of Naturally Occurring and Engineered Proteins  
at Atomic Resolution**

Michelle E. McCully

A dissertation  
submitted in partial fulfillment of the  
requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Valerie Daggett, Chair

Wendy E. Thomas

Gabriele Varani

Program Authorized to Offer Degree:

Bioengineering



University of Washington

## **Abstract**

The Dynamics and Folding Pathways of Naturally Occurring and Engineered Proteins  
at Atomic Resolution

Michelle E. McCully

Chair of the Supervisory Committee:

Professor Valerie Daggett

Department of Bioengineering

The protein folding problem, the aim to understand how a protein's amino acid sequence alone is sufficient to dictate its folded structure in a given environment, has confounded scientists for decades. Comprehensive biophysical and structural analysis of individual proteins can provide insight to the forces driving protein folding in general. The synthesis of years of such experimental and computational research on the Engrailed Homeodomain (EnHD) has produced a detailed description of the states populated along its folding pathway. Here, further examination of the folding pathway using all-atom, explicit solvent molecular dynamics simulations provided an atomic-level description of the interactions responsible for transitioning between these states. Simulations of EnHD near its melting temperature showed that the folding and unfolding pathway are the same and support the use of high temperature unfolding simulations to study the folding pathway in reverse.

Multi-molecule unfolding simulations of EnHD gave insight to the effects of intermolecular interactions on folding as well as the types of interactions that drive protein aggregation at high temperature. Further work on an engineered, thermostable variant of EnHD showed that heightened dynamics allowed it to remain stable at high temperature by tolerating the increased thermal fluctuations, whereas EnHD's more geometrically restrictive packing interactions were perturbed at high temperature, causing it to unfold. To investigate how sequence dictates the folded topology of a protein, a unique system consisting of a pair of proteins engineered to have 88% sequence identity but different folds was studied. In this system, seven differing residues alone hold the key to folding to an all- $\alpha$  or  $\alpha/\beta$  structure, and we identified specific contacts in the denatured state stemming from these residues that committed the proteins to their respective folded structures. Further work on this pair of proteins investigated the denatured state of point mutants that knocked out these putative topology-directing interactions. Finally, a review of research on engineered proteins explored the nonnatural folding pathways created by scientists in the absence of natural selection and the contribution of dynamics to successful designs.

## Table of Contents

List of Figures .....	vi
List of Tables .....	ix
Chapter 1: Introduction .....	1
1.1 Challenges and Benefits of Understanding Protein Folding .....	1
1.2 Molecular Dynamics – A High-Resolution Structural Tool .....	3
1.2.1 Simulation Protocol .....	5
1.2.2 Analysis Techniques .....	6
1.3 Studies on the Engrailed Homeodomain .....	9
1.4 Overview of Chapters .....	11
1.5 Insights to Protein Folding and Dynamics .....	14
Chapter 2: Microscopic Reversibility of Protein Folding in Molecular Dynamics	
Simulations of the Engrailed Homeodomain .....	16
2.1 Summary .....	16
2.2 Introduction .....	16
2.3 Methods .....	19
2.3.1 Molecular Dynamics Simulations .....	19
2.3.2 C $\alpha$ RMSD Matrix and 3D MDS .....	20
2.3.3 HIII-Core Distance Calculation .....	21
2.3.4 Average Structures .....	21
2.3.5 HIII-Core Contacts .....	21
2.3.6 Calculation of S-values .....	21
2.3.7 Protein – DNA Interactions .....	22

2.4	Results .....	24
2.4.1	Overview of Simulations .....	24
2.4.2	Native' State at Elevated Temperature .....	27
2.4.3	Properties of the Transition State Ensembles .....	30
2.4.4	Protein – DNA Interactions .....	32
2.4.5	Unfolding and Refolding Pathways .....	33
2.5	Discussion .....	34
Chapter 3: Refolding the Engrailed Homeodomain – Structural Basis for the		
	Accumulation of a Folding Intermediate.....	39
3.1	Summary .....	39
3.2	Introduction .....	39
3.3	Methods .....	42
3.3.1	Simulation Protocol.....	42
3.3.2	Analyses .....	43
3.3.3	Property Space and Reaction Coordinate .....	44
3.3.4	Transition State Selection.....	46
3.4	Results .....	46
3.4.1	Native Simulations .....	47
3.4.2	Property Space Reaction Coordinate.....	47
3.4.3	Quench Simulations: Successful Refolding .....	48
3.4.4	Quench Simulations: Factors Preventing Refolding .....	51
3.5	Discussion .....	54
3.5.1	Property Space Descripton of the Native State and Unfolding Pathway of EnHD.....	54
3.5.2	The Refolded Native State .....	56

3.5.3	Comparison of One Successful Folding Trajectory with 45 Unsuccessful Trials .....	57
3.5.4	Misfolding Traps in the Intermediate Slow Refolding.....	57
3.5.5	Microscopic Reversibility .....	59
Chapter 4: Unfolding in a Test Tube – Multimolecule Atomistic Molecular Dynamics		
	Simulations of the Engrailed Homeodomain .....	60
4.1	Summary .....	60
4.2	Introduction .....	60
4.3	Methods.....	63
4.3.1	Molecular Dynamics Simulation Parameters .....	63
4.3.2	Molecular Dynamics Simulation Analysis.....	64
4.4	Results and Discussion.....	66
4.4.1	Nature of the Intermolecular Interactions .....	67
4.4.2	Native State Behavior in Single-Molecule Versus Test-Tube Simulations.....	70
4.4.3	Effect of Intermolecular Interactions on the Unfolding Pathway of EnHD.....	71
4.5	Conclusions .....	79
Chapter 5: Promiscuous contacts and heightened dynamics increase thermostability in an engineered variant of the Engrailed Homeodomain .....		
	an engineered variant of the Engrailed Homeodomain .....	81
5.1	Summary .....	81
5.2	Introduction .....	81
5.3	Methods.....	84
5.3.1	Simulation Protocol.....	84
5.3.2	Calculation of NMR Comparables .....	85
5.3.3	Simulation Analysis .....	86

5.4	Results .....	87
5.4.1	Validation of Simulations: Comparison with NMR Observables .....	87
5.4.2	Dynamics of EnHD vs. UVF .....	91
5.5	Discussion .....	95
5.6	Conclusions .....	99
Chapter 6: The Denatured State Dictates the Topology of Two Proteins with Almost Identical Sequence but Different Native Structure and Function.....		
		101
6.1	Summary .....	101
6.2	Introduction .....	101
6.3	Methods.....	104
6.3.1	Molecular Dynamics Simulations .....	104
6.3.2	Simulation Analysis .....	104
6.3.3	Transition State Assignment .....	105
6.4	Results .....	106
6.4.1	Equilibrium Unfolding of G <sub>A</sub> 88 and G <sub>B</sub> 88.....	106
6.4.2	Folding and Unfolding Kinetics .....	107
6.4.3	Molecular Dynamics Simulations .....	111
6.5	Discussion .....	116
6.5.1	The Role of Long-Range Interactions in Denatured States.....	116
6.5.2	Do Engineered Proteins Display Cooperative Folding?.....	118
6.5.3	Comparison with Studies on Protein Families .....	119
Chapter 7: Identification and Characterization of Specific Interactions in the Denatured State of G <sub>A</sub> 88 and G <sub>B</sub> 88.....		
		121
7.1	Summary .....	121
7.2	Introduction .....	121
7.3	Methods.....	124



7.3.1	Preparation of Constructs .....	124
7.3.2	Simulation Protocol.....	124
7.3.3	Simulation Analysis .....	125
7.4	Results .....	125
7.4.1	Analysis of the G <sub>A</sub> 88 and G <sub>B</sub> 88 Mutants .....	126
7.4.2	Analysis of all- $\alpha$ and $\alpha/\beta$ Protein Pairs.....	129
7.5	Discussion .....	134
7.6	Conclusions .....	135
Chapter 8: Folding and Dynamics of Engineered Proteins.....		137
8.1	Introduction .....	137
8.2	Proof-of-Principle Protein Designs .....	138
8.2.1	FSD-1, a Heterogeneous Native State and Complicated Folding Pathway .....	139
8.2.2	$\alpha_3D$ , a Dynamic Core Leads to Fast Folding and Thermal Stability.....	143
8.2.3	3-Helix Bundle Thermostabilized Proteins .....	145
8.2.4	Top7, a Novel Fold Topology .....	147
8.2.5	Additional Rosetta Designs .....	151
8.3	Proteins Designed for Function.....	153
8.3.1	Ligands .....	154
8.3.2	Enzymes .....	158
8.4	Conclusions and Outlook .....	163
Bibliography .....		165

## List of Figures

Figure 1.1: Idealized free energy plot of a protein folding pathway .....	2
Figure 1.2: Sample system preparation and simulation .....	5
Figure 1.3: States along the unfolding pathway.....	10
Figure 2.1: General properties for each simulation .....	23
Figure 2.2: Structures and order of population of the 4 states in the 323K/2 simulation.....	25
Figure 2.3: N and N' structures and all-vs.-all core C $\alpha$ RMSD matrix for 2 323K simulations .....	26
Figure 2.4: HIII-core residue pairs fraction of time in contact for N, N', and new TSEs in each simulation.....	28
Figure 2.5: Core C $\alpha$ RMSD between average structures representative of N, N', and TS.....	29
Figure 2.6: S-values for the 13 new TSEs .....	30
Figure 2.7: Comparison of 323K/2 43 ns unfolding and refolding TSEs and pathway.....	31
Figure 2.8: All-vs.-all core C $\alpha$ RMSD matrix of 2 transient N $\rightarrow$ N' $\rightarrow$ N transitions from 323K/3.....	32
Figure 2.9: Structures from 323K/2 fit to DNA.....	33
Figure 3.1: Reaction coordinate for high temperature denaturation simulation .....	48
Figure 3.2: Reaction coordinate for the successful quench simulation .....	49
Figure 3.3: Selected properties from the successful quench simulation.....	50
Figure 3.4: Structures from the successful quench simulation showing the order of HI-HII contacts .....	50
Figure 3.5: Structures from the Successful Quench Simulation.....	52
Figure 3.6: Core C $\alpha$ RMSD for unfolding and refolding .....	53
Figure 3.7: Structures from unsuccessful quench simulations showing nonproductive salt bridges .....	54
Figure 3.8: Contact lifetimes for different contact types .....	55

Figure 4.1: Structures from test-tube simulations.....	62
Figure 4.2: Principal component analysis of 39-dimensional property space .....	66
Figure 4.3: Types of intermolecular vs. intramolecular interactions .....	67
Figure 4.4: Hydrophobic interactions at high temperature .....	69
Figure 4.5: Properties of 498 K unfolding simulations over time .....	71
Figure 4.6: Property space distributions .....	72
Figure 4.7: Final structures plotted in principal component space .....	73
Figure 4.8: Single-molecule vs. test-tube transition state properties .....	76
Figure 4.9: Single-molecule vs. test-tube HI-HII contact loss.....	78
Figure 5.1: Sequence and structure of EnHD and UVF.....	83
Figure 5.2: $^3J_{\text{HNH}\alpha}$ -coupling constants and $S^2$ amide order parameters for EnHD.....	88
Figure 5.3: Core C $\alpha$ RMSD and RMSF at 25 and 100 °C.....	90
Figure 5.4: Native, nonnative, and unique contacts for EnHD and UVF .....	91
Figure 5.5: Main chain and side chain contacts at 25 and 100 °C.....	93
Figure 5.6: Types of side chain contacts at 25 and 100 °C.....	94
Figure 5.7: Stereo image of backbone mobility in EnHD and UVF.....	95
Figure 5.8: Final structures from the 100 °C simulations.....	96
Figure 6.1: G <sub>A</sub> 88 and G <sub>B</sub> 88 sequences and structures .....	102
Figure 6.2: Chemical denaturation of G <sub>A</sub> 88 and G <sub>B</sub> 88 monitored by CD.....	106
Figure 6.3: Folding and unfolding rate constants for G <sub>A</sub> 88 and G <sub>B</sub> 88 .....	107
Figure 6.4: Chevron plots of G <sub>A</sub> 88 and G <sub>B</sub> 88 at varying pH.....	108
Figure 6.5: Rate constants and <i>m</i> -values for G <sub>B</sub> 88 folding and unfolding in salt.....	109
Figure 6.6: Structures of the denatured state at neutral pH.....	110
Figure 6.7: Structures of the transition state at neutral and low pH .....	112
Figure 6.8: Folding pathway and contact maps at neutral pH .....	114
Figure 7.1: Sequences of 16%, 88%, and 98% identical G <sub>A</sub> /G <sub>B</sub> pairs.....	122
Figure 7.2: Interactions in the denatured state of G <sub>A</sub> 88 and G <sub>B</sub> 88 that dictate the folded topology.....	123
Figure 7.3: Putative folding mechanisms for G <sub>A</sub> 88 and G <sub>B</sub> 88.....	126
Figure 7.4: The $\beta$ 1/ $\beta$ 2 Thr1 – Glu19 and $\beta$ 3/ $\beta$ 4 Tyr45 – Asp47 – Lys50 networks from G <sub>B</sub> 88 in G <sub>A</sub> 88/G <sub>B</sub> 88 mutants.....	127

Figure 7.5: The $\alpha 3$ Asp47 – Thr1 – Glu48 network from $G_A88$ in $G_A88/G_B88$ mutants.....	127
Figure 7.6: Percentage $\alpha$ -helix in the denatured state.....	129
Figure 7.7: The $\beta 1/\beta 2$ Thr1 – Glu19 and $\beta 3/\beta 4$ Tyr45 – Asp47 – Lys50 networks from $G_B88$ in swapped sequence/topology pairs.....	130
Figure 7.8: The $\alpha 3$ Asp47 – Thr1 – Glu48 network from $G_A88$ in swapped sequence/topology pairs.....	130
Figure 7.9: Contacts maps and structures from the denatured state.....	132
Figure 8.1: Structures of the proof-of-principle designs.....	140
Figure 8.2: Proteins redesigned by Rosetta.....	151
Figure 8.3: Reaction mechanisms for the retro-aldol and Kemp elimination enzymes.....	159

## List of Tables

Table 3.1: Simulation properties for native simulations .....	43
Table 3.2: Simulation properties for quench simulations .....	44
Table 3.3: Property space weights and values of the reference sets .....	45
Table 4.1: Average properties of the 298 K native simulations .....	70
Table 4.2: Test-tube transition states .....	74
Table 4.3: Single-molecule transition states .....	75
Table 5.1: EnHD comparison between simulation and experiment .....	87
Table 5.2: UVF comparison between simulation and experiment.....	89
Table 6.1: Properties of the native and denatured state ensembles from MD simulations at neutral and low pH .....	111
Table 6.2: C $\alpha$ RMSD of the transition state from neutral and low pH simulations at 498 K.....	115

## **Acknowledgements**

I would like to thank Dr. Valerie Daggett, Dr. Wendy Thomas, Dr. Gabriele Varani, and Dr. Shaoyi Jiang for serving on my supervisory committee. I am grateful to Dr. Darwin Alonso, Dr. David Beck, Dr. Noah Benson, Denny Bromley, Jonathan Cheng, Dr. Gene Hopping, Dr. Amanda Jonsson, Peter Law, Dr. Eric Merkley, Steve Rysavy, Dr. Dustin Schaeffer, Dr. Tom Schmidlin, Dr. Alex Scouras, Dr. Andrew Simms, Dr. Summer Thyme, Dr. Rudesh Toofanny, and Dr. Clare-Louise Towse for their support and collaboration over the years. Additionally, I would like to acknowledge the Biomolecular Structure and Design Program and Department of Bioengineering for administrative support, the National Defense Science and Engineering Graduate Fellowship for financial support, and the National Energy Research Scientific Computing Center for computational resources.

## **Dedication**

This work is dedicated to

dacb, rschaeff, and ajonsson  
for teaching me everything I know

and to Jon  
for his endless patience and support





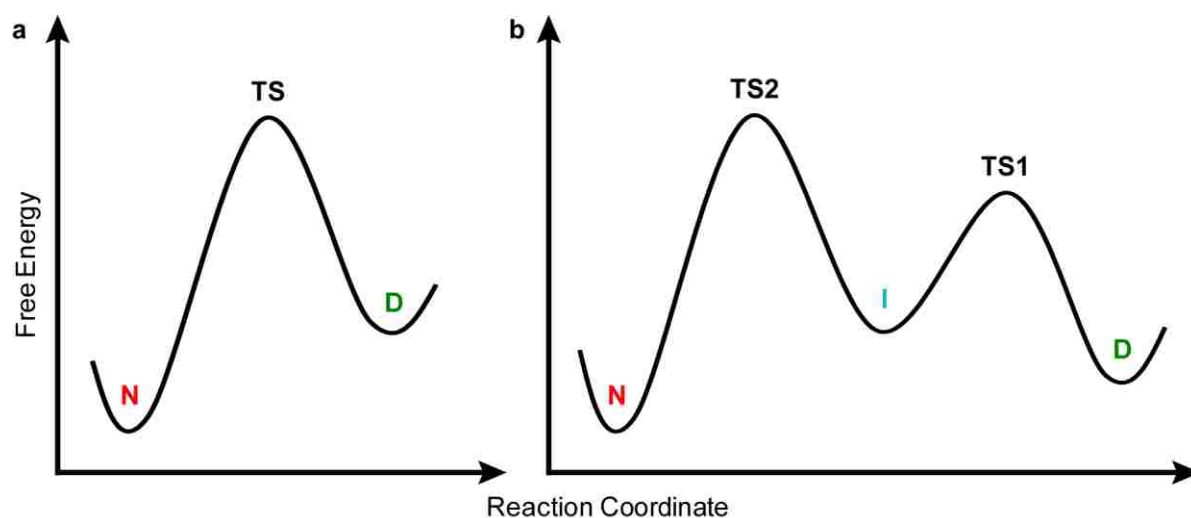
## Chapter 1: Introduction

### 1.1 Challenges and Benefits of Understanding Protein Folding

Developing an understanding of both the general mechanisms of protein folding and the specific pathways that individual proteins take to reach their native, functional structures is a primary goal of structural biology (Berendsen 1998). In the late 1960s, the idea that proteins must fold via a pathway as opposed to a random sampling of states was introduced. Levinthal (1968, 1969) proposed that if a small protein randomly sampled each rotatable bond in its structure, at even a few discrete states during the process of folding, it would take an impossibly long amount of time to reach the native, folded state. It was known at that time that folding occurred within seconds. Therefore proteins could not possibly sample all of conformational space as they folded, in fact they could only sample a very small portion of it. Thus, Levinthal concluded that proteins must fold via discrete, energetically favorable pathways.

Anfinsen (1973) showed that the folding pathway is dependent only upon the amino acid sequence of the protein with a series of protein denaturing and renaturing experiments on ribonuclease. He showed that the activity of the enzyme could be destroyed by adding urea and reducing its disulfide bonds and then restored by returning it to native conditions. Thus, he concluded that the function, and therefore tertiary structure, of the protein in a given environment was encoded purely by its primary, amino acid sequence. Furthermore, he inferred that protein folding was a thermodynamic process where the folded state was a free energy minimum under native, physiological conditions. Proteins at equilibrium traverse the energy landscape between the native and denatured state. The simplest folding model is two-state and can be shown as two energy wells representing the native and denatured states separated by a high-energy transition state (TS; Figure 1.1a). There are often semi-stable states along the folding pathway, called intermediates, separated by smaller energy barriers than the native state (Figure 1.1b).

The idea that protein folding takes place via an energetically favorable pathway, or ensemble of pathways, was followed by proposals of two models of folding pathways: framework (or diffusion-collision) and nucleation-condensation (Fersht 2008). Both models describe protein folding as a process that is primarily driven by the burial of hydrophobic surface area in the core of the protein. In the framework model, folding begins with the formation of structured elements, and then a combination and rearrangement of those elements forms the final, folded structure. Nucleation-condensation stipulates that the protein initially collapses into a molten globule stabilized by native tertiary interactions followed by simultaneous secondary and tertiary structure formation.



**Figure 1.1: Idealized free energy plot of a protein folding pathway**

Examples of a (a) two-state and (b) three-state folding pathway are shown with the native (“N,” red), transition (“TS,” black), intermediate (“I,” cyan), and denatured (“D,” green) states noted. Folding proceeds right-to-left.

Over the past decades, scientists have been using a myriad of experimental and computational techniques to probe the folding pathways of proteins (Fersht 2008, Bartlett and Radford 2009). Kinetic experiments, often based on changes in intrinsic fluorescence or polarization of light due to secondary structure of a protein between the native and denatured states, have probed the folding pathway and detected the presence of intermediate states along the pathway (Serrano *et al.* 2012). Protein engineering techniques can be used to probe specific residues to determine whether they are more or less native-like in the transition state (Fersht *et al.* 1992). Occasionally mutants are found to be stabilized in an intermediate state, which allows investigation of another region of the folding pathway and in some cases, solution of the structure.

There are many benefits of understanding how proteins fold. Once we understand folding pathways, we will be better equipped to understand how and why proteins misfold. There are many diseases that are caused by misfolded proteins, including bovine spongiform encephalopathy (mad cow disease, or in humans, Creutzfeldt-Jakob disease), Alzheimer's, Parkinson's, and amyotrophic lateral sclerosis (Chiti and Dobson 2009, Eisenberg and Jucker 2012). A better understanding of misfolding will lead to diagnostics and treatments of these disorders.

Understanding how proteins fold will also help predict the folded structure of a protein based on its sequence. While protein sequences are straightforward and relatively inexpensive to obtain, protein structures require much more care, time, and money to solve. The two most popular techniques used to solve protein structures are X-ray crystallography and nuclear magnetic resonance, but both have their drawbacks (Wagner *et al.* 1992). Crystallography requires growing crystals, getting good diffraction patterns, and solving the phasing problem; and nuclear magnetic resonance (NMR) spectroscopy is most useful for small, soluble proteins stable at high concentration. The ability to accurately predict protein structures computationally will remove these obstacles.

Given a protein structure, there are many more ways to study and modify its function than from the sequence alone. For example, the structure of an unknown protein may hint at its physiological function even if it has low sequence similarity to known proteins. Specific arrangements of amino acids to form catalytic motifs are not always apparent from a protein sequence. Knowing the structures of two proteins that interact, the interaction could be predicted and then inhibited or enhanced depending on the situation. Structure-based drug design has been very successful in recent years (Congreve *et al.* 2005), but of course, the structure of the protein drug target is required. Based on a structure, scientists can select a region on a protein to bind a drug in order, for example, to stabilize a critically destabilized protein (Boeckler *et al.* 2008) or inhibit a binding event (Varghese 1999).

## **1.2 Molecular Dynamics – A High-Resolution Structural Tool**

The molecular interactions driving protein folding occur at the atomic level and on picosecond timescales. Experimental techniques currently do not allow concurrent spatial

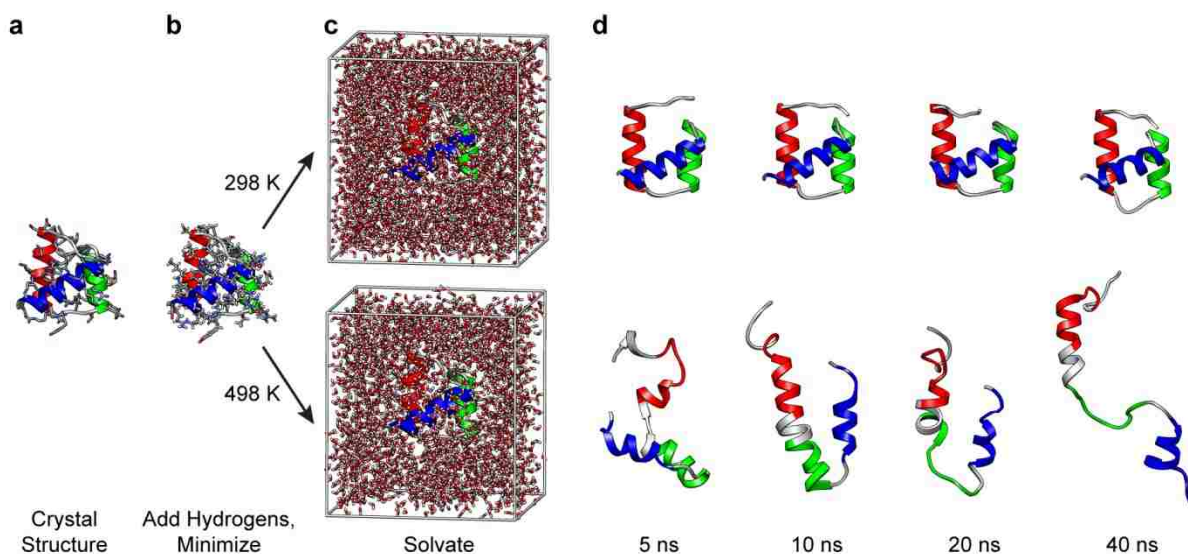
and temporal observations at these resolutions, so we must depend on computational modeling. All-atom, explicit solvent molecular dynamics (MD) simulations offer such spatial-temporal resolution, and present computer technology allows for simulation timescales on the order of nanoseconds to microseconds. MD simulations have been shown to predict and reproduce experimental observables in simulations of the protein native state for a wide variety of proteins (Beck *et al.* 2008b), in the transition and denatured state from thermal unfolding simulations (Li and Daggett 1996, Mayor *et al.* 2003b), and in the dynamical changes of a protein due to a naturally occurring, pathogenic mutation (Rutherford and Daggett 2009).

All-atom, explicit solvent MD simulations model each atom in the simulation system, of both the solute and solvent, and the forces between them. Each atom is represented as a charged sphere and covalent bonds as springs. The geometry of the bonded atoms is maintained through springs on the angles and a rotational, cosine-based term on the torsional angles. The interactions between nonbonded atom pairs are modeled with van der Waals and Coulombic terms. These forces are calculated for each atom in three-dimensional space at each time step, and Newton's second law of motion ( $\text{force} = \text{mass} \times \text{acceleration}$ ) is applied to get the accelerations for the next time point. From the accelerations, the velocities and then new spatial coordinates are calculated. Repetition of this process results in the trajectory of each atom in the system over the entire simulation time.

The potential energy, which is the sum of the bond, angle, torsion, van der Waals, and Coulombic terms, is summed with the kinetic energy, which is calculated directly from the velocities of the atoms, to attain the total energy of the system. The energy of the system must be kept constant since the NVE microcanonical ensemble requires a constant number of particles, simulation volume, and total energy. By scaling the velocities in the system when the energy reaches a maximum threshold, this total energy is kept constant over the course of the simulation. The temperature of the system is calculated directly from the atomic velocities, and scaling the velocities similarly maintains a desired temperature.

This work has employed the *in lucem* molecular mechanics (*ilmm*) simulation package, which was developed in-house for calculating and subsequently analyzing MD trajectories (Beck *et al.* 2000-2012). The force constants for the springs, van der Waals radii,

partial charges, and other system constants were taken from the Levitt *et al.* (1995) force field and flexible three-center (F3C) water model (Levitt *et al.* 1997). Our standard simulation protocol was generally followed (Beck *et al.* 2004); the specific protocols used for each system studied here are detailed in the respective methods sections.



**Figure 1.2: Sample system preparation and simulation**

(a) The starting structure is usually selected from the PDB. Shown here is the Engrailed Homeodomain (PDB id: 1enh) colored by helix (HI: 10-22 red, HII: 28-38 green, HIII: 42-55 blue). (b) The PDB structure has any missing atoms added and is then minimized. (c) The structure is solvated in a preequilibrated box of water at a density dependent on the target simulation temperature. (d) The simulation is carried out (side chains and water molecules are not displayed here for clarity). Room temperature, or 298 K, is typically used for native simulations, and temperatures up to 498 K are employed to observe the unfolding pathway.

### 1.2.1 Simulation Protocol

A starting structure is typically procured from the Protein Data Bank (PDB; Figure 1.2a); structures solved by X-ray crystallography or NMR are most commonly used, though homology models or structures from simulations may also be utilized. Any missing atoms (often hydrogens) are built in, and the structure is briefly minimized (Figure 1.2b). Next, the protein is put in a box of preequilibrated F3C water molecules with a density dependent on the desired simulation temperature (Figure 1.2c; Kell 1967, Haar *et al.* 1984). The protein and water are then alternately minimized and heated for short periods of time to prepare the system. At this point, the production simulation begins as the system is heated to the desired temperature (Figure 1.2d). The production run continues until it reaches the desired simulation time, with structures saved to disk every 1 ps.

### 1.2.2 Analysis Techniques

After the MD simulation is complete, the atomic coordinates may be post-processed in a variety of ways in order to elucidate the details of the dynamics or un/folding pathway of the simulated protein. Several of the most commonly used analysis methods in this work are introduced here. Where it is applicable, experimental measurements that can be compared with the calculated values from the simulation analyses are discussed as well.

Native protein simulations begin from the experimentally determined NMR or crystal structure and are performed at temperatures below the  $T_m$  of the protein. Such simulations may be validated by comparing interatomic distances with distance restraints derived from nuclear Overhauser effect (NOE) crosspeaks measured by NMR experiments. NOE equivalent distances are measured from our simulation as the  $\langle r^{-6} \rangle$  weighted distance between the closest protons that were identified as contributing to the NOE. The NOE is considered satisfied if this distance is below 5.0 Å or the restraint distance, whichever is longer.

The root-mean-square deviation (RMSD) from the starting structure and root-mean-square fluctuation (RMSF) about the mean structure both give an indication of how much movement there is in the protein. Traditionally these measurements are taken over only  $C\alpha$  atoms after alignment.  $C\alpha$  RMSD indicates movement away from the starting structure, and higher values indicate more motion. If a simulation has reached equilibrium, the  $C\alpha$  RMSD will level off. However, for a protein under destabilizing conditions, the value will continually increase over time.  $C\alpha$  RMSF is a measure of how much variation there is in the position of each  $C\alpha$  atom. It is usually viewed as a time-averaged value on a per-residue basis so that regions with higher flexibility, and thus a higher  $C\alpha$  RMSF, such as tails and loops, are apparent.

Tertiary structure may also be investigated through patterns in contacts. A contact is defined as two carbon atoms that are  $\leq 5.4$  Å apart, or any other pair of non-hydrogen atoms that are  $\leq 4.6$  Å apart. Hydrogen bonds may be specifically investigated, and the distance cutoff for donor hydrogen – acceptor atom pair is  $\leq 2.6$  Å where the donor – hydrogen – acceptor angle is within 45° of linearity. These contacts are often classified as to whether they happen within or between main chain (N,  $C\alpha$ , C, O, H) or side chain atoms. It is also helpful to classify contacts as to whether they are present in the starting structure (native) or

gained during the course of the simulation (nonnative) in order to assess how much the protein has deviated from its starting structure. Tracking specific contacts over time gives insight to interactions of interest, and residue-vs.-residue contact maps may be used to simply visualize contacts by region of the protein.

Secondary structure in MD simulations may be identified in two ways. First, main chain hydrogen bonding patterns identify helices and sheets, and this technique is called the dictionary of protein secondary structure (DSSP; Kabsch and Sander 1983). For example, a series of backbone hydrogen bonds between residues that are four positions apart in sequence designate an  $\alpha$ -helix. Second, patterns in backbone  $\phi/\psi$  angles also indicate secondary structure, with repeating  $260^\circ < \phi < 330^\circ$  and  $280^\circ < \psi < 355^\circ$  identifying regions of  $\alpha$ -helix, for example. Tracking which residues adopt different secondary structures vs. random coil across the simulation helps point to the stability of a specific region. Secondary structure may be confirmed experimentally using far-UV circular dichroism (CD) spectroscopy, which identifies secondary structure content in the protein by detecting the polarization fingerprints of  $\alpha$ -helix,  $\beta$ -sheet, and random coil (Greenfield and Fasman 1969).

The solvent accessible surface area (SASA) of a protein measures how compact the structure is; as a protein unfolds, the SASA, especially of its hydrophobic residues, increases. The SASA is calculated using a probe radius of 1.4 Å, which is the radius of a water molecule. The probe is rolled across the surface of the protein, and the area that it can reach is summed to give the SASA. SASA may be broken down into hydrophobic, polar, main chain, or side chain depending on what atoms make up the surface. The SASA of aromatic residues, such as tryptophan, may be compared with experimental fluorescence data, as these residues shift wavelengths depending on whether they are buried or solvent exposed. In addition,  $m$ -values, which are a measure of how dependent the folding kinetics are on the denaturant concentration, are proportional to the change in SASA upon unfolding (Myers *et al.* 1995).

In cases where there is an especially large number of structures to consider, it is helpful to combine data from many properties into a single value. This is accomplished by creating a multi-dimensional property space, where each property represents one dimension. Once this space has been assembled and normalized for a collection of structures, the

Euclidean distance may be taken between two structures to represent how similar they are to each other. A reference state is assembled from a single or collection of native simulations, and the average distance between a structure and all structures in the reference state is the mean distance to reference (MDTR). Plotting a histogram of the MDTR for a simulation gives a reaction coordinate (Toofanny *et al.* 2010). For a simulation of a protein at equilibrium in one state, there is a single peak on the reaction coordinate. For an unfolding simulation, there is a larger distribution along the reaction coordinate, and peaks represent semi-stabilized states, such as intermediates. Alternatively, a principal component (PC) analysis may be performed on the space, and a histogram of the projection onto the first, second, or more PCs may be plotted (Toofanny *et al.* 2010). Using more dimensions allows better discrimination between states, which is especially useful when looking for intermediates on the un/folding pathway. The disadvantage is that more than three dimensions are difficult to visualize. Once again, peaks in the histogram represent stable states, or energy wells. Property space is a powerful tool for considering a wide variety of properties in a simple visualization.

Major transition states in the folding pathway can be probed experimentally, and thus identifying them in our simulations is useful for gaining structural information about transition states as well as for validating simulations against experimental observables. Transition states in our unfolding simulations are selected using a three-dimensional multidimensional scaling (MDS) of the all-against-all C $\alpha$  RMSD matrix (Li and Daggett 1994). For a 498 K unfolding simulation, the matrix is typically calculated for structures in a 1-, 2-, or 5-ns timeframe. In the 3D MDS, the native cluster is identified, and the TS is the final structure in the cluster, with the previous 5 ps additionally comprising the transition state ensemble (TSE). In the case of refolding, the TSE is made up of the native cluster reentry as well as the next 5 ps (McCully *et al.* 2008). S-values may be calculated for each residue of the protein in the TSE as the product of S $_2^\circ$  and S $_3^\circ$  (Daggett *et al.* 1996). S $_2^\circ$  is the extent of native secondary structure for a given residue, and S $_3^\circ$  is the ratio of the number of contacts the residue makes in the TSE to the number of contacts it makes in the simulation starting structure. S-values typically range from 0 to 1 with higher values indicating more native-like structure. S-values are directly comparable to experimentally determined  $\Phi$ -values.  $\Phi$ -values are calculated by mutating a residue and measuring the ratio of  $\Delta\Delta G_{\ddagger-D}$  to

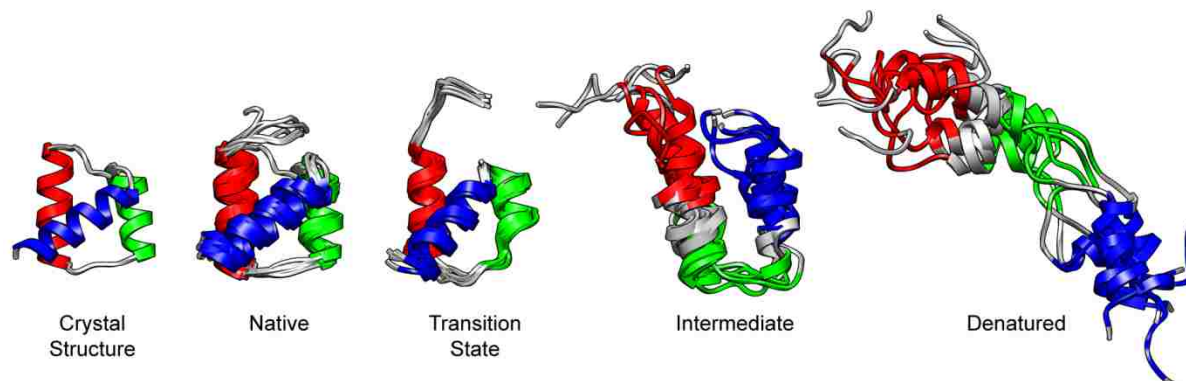


$\Delta\Delta G_{N-D}$  due to the mutation (Fersht *et al.* 1992), where  $\Delta G_{N-D}$  is the free energy of folding and  $\Delta G_{\ddagger-D}$  is the activation energy. Similar to S-values,  $\Phi$ -values generally range from 0 to 1, where a value of 1 indicates the TS is equally perturbed by the mutation as the native state, signifying that the residue is as structured in the TS as when the protein is folded. Likewise, a  $\Phi$ -value of 0 indicates that the residue is as unstructured in the TS as in the denatured state. By validating the TSE via comparison of S-values with available  $\Phi$ -values, we can have full confidence in the transition state structures provided by unfolding or refolding simulations.

### 1.3 Studies on the Engrailed Homeodomain

The Engrailed Homeodomain (EnHD) is a 61-residue, three-helix bundle protein found in *Drosophila melanogaster* (Figure 1.2a). Natively it is a transcription factor that controls the *engrailed* gene locus, which is responsible for proper development of thoracic segments in the fruit flies during embryogenesis (Kornberg 1981; Liu *et al.* 1990). The folding pathway of EnHD has been investigated over the years using a variety of experimental (Mayor *et al.* 2000, Gianni *et al.* 2003, Mayor *et al.* 2003a, Mayor *et al.* 2003b, Religa *et al.* 2005, Religa *et al.* 2007, Huang *et al.* 2008) and computational (Mayor *et al.* 2000, Mayor *et al.* 2003b, Gianni *et al.* 2003, DeMarco *et al.* 2004, Zhang *et al.* 2005, Hubner *et al.* 2006a, Beck and Daggett 2007, Huang *et al.* 2008, Li *et al.* 2008) techniques. It is a popular model system for folding studies because of its ultra-fast folding and unfolding kinetics. From this work, we have a thorough description of the different states populated along the folding pathway.

The native state (N) of EnHD is barely stable with  $\Delta G_{D-N} = 2.5$  kcal/mol and  $T_m = 325$  K (52 °C; Mayor *et al.* 2003a, Mayor *et al.* 2003b). Fluorescence and near-UV CD ellipticity indicate the presence of a folded aromatic core (Mayor *et al.* 2000, Mayor *et al.* 2003a). Analysis of MD simulations reports a C $\alpha$  RMSD of  $\sim 2-4$  Å to the crystal structure and  $\sim 70\%$   $\alpha$ -helix (Figure 1.3; DeMarco *et al.* 2004).



**Figure 1.3: States along the unfolding pathway**

Structures are shown left-to-right from a single high temperature unfolding simulation (498 K, run 2): crystal structure (PDB id: 1enh), native state (0-60 x 10 ns), transition state (255-260 x 1 ps), intermediate state (12-20 x 2 ns), and denatured state (35-60 x 5 ns).

The structure of the transition state ensemble of EnHD was first predicted by MD simulation (Figure 1.3; Mayor *et al.* 2000, Mayor *et al.* 2003b) and later investigated by  $\Phi$ -value analysis, which confirmed the TS predictions (Gianni *et al.* 2003). The TS structures show HIII pulling away from the hydrophobic core with a  $C\alpha$  RMSD of  $\sim 3$ -5 Å, and a framework folding mechanism was proposed (Mayor *et al.* 2000, Gianni *et al.* 2003, DeMarco *et al.* 2004).

A U-shaped intermediate state (I) with high native and nonnative helical content and  $C\alpha$  RMSD  $\sim 9$ -12 Å was predicted by MD simulation (Figure 1.3; Mayor *et al.* 2000, Mayor *et al.* 2003b, DeMarco *et al.* 2004), examined experimentally (Mayor *et al.* 2003a), and later determined by NMR for an engineered, intermediate-stabilized mutant (Religa *et al.* 2005). It was shown that the mutation of Leu16 to Ala caused EnHD to be stabilized in I under native conditions. The denatured state (D) of EnHD lacks an aromatic core as measured by near-UV CD ellipticity and Trp48 fluorescence (Mayor *et al.* 2003a). Thermal denaturation experiments (Mayor *et al.* 2003a) and high temperature MD simulations (Figure 1.3; Mayor *et al.* 2000, DeMarco *et al.* 2004) yield highly extended conformations with little to no stable helical structure and a  $C\alpha$  RMSD  $> 14$  Å.

Formation of I from the unfolded state is very fast, with  $t_{1/2} \approx 1.5$   $\mu$ s at 298 K (25 °C), whereas folding from I to N is an order of magnitude slower ( $t_{1/2} \approx 15$   $\mu$ s; Mayor *et al.* 2003b). The folding and unfolding kinetics of EnHD are ultra-fast, in fact they were the fastest directly measured rate constants to date when they were initially determined (Mayor

*et al.* 2000). Because EnHD folds on such a fast timescale, it is a perfect candidate for MD un/folding simulations, which are currently computationally limited to timescales on the order of ns –  $\mu$ s.

Thus, EnHD is a well-understood system with a plethora of experimental observations to use for comparison with simulation data. This body of knowledge exemplifies the power of combining experimental and computation data to get a richer understanding of protein folding. While each of the individual states (N, TS, I, D) is very well characterized, the pieces that are left to be detailed are the transitions between these states, which is the folding pathway itself. Additionally, a structural explanation of why the intermediate forms in the first place has never been offered.

## 1.4 Overview of Chapters

This work sets out to investigate the individual interactions that dictate the dynamics and folding pathway of naturally occurring and engineered proteins at atomic-level detail. Engineered proteins provide a unique opportunity to investigate what types of dynamics and folding pathways occur when the protein has not been subject to the process of evolution. EnHD is studied to represent native proteins because of the large body of existing data describing the states along its folding pathway and its amenability to molecular dynamics folding studies. With a detailed description of the dynamics and folding pathway of EnHD, it is possible to compare them with data from engineered proteins. We investigate the structural bases for the thermostability of one engineered protein and the divergent folding pathways of a pair of engineered proteins with high sequence identity. MD simulations are the perfect tool to address these questions because of the detailed spatial-temporal resolution that they provide.

Chapter 2 opens with a discussion of the principle of microscopic reversibility as it applies to protein folding and unfolding. In structural terms, this principle states that for a protein folding reaction, the conformations populated during folding are the same as those populated during unfolding, just in the opposite order. It has been shown previously that this principle holds true for folding of the chymotrypsin inhibitor 2 (Day and Daggett 2007), and it is shown here to be true for EnHD as well. Structural similarity along the folding and

unfolding pathways as well as the order of contact gain and loss showed that the folding pathway was the reverse of the unfolding pathway. The consequence of this finding is that it is possible to learn about the folding pathway by studying the unfolding pathway. Thus, high-temperature unfolding MD simulations, which are significantly less computationally intensive than folding simulations, are a valid way to investigate protein folding pathways.

In Chapter 3, a series of “quench” simulations are discussed. These simulations began with a denatured structure from a thermal unfolding simulation of EnHD and were run under folding-permissive conditions (e.g. lower temperature). The one simulation in which EnHD successfully refolded to the native state was our first direct look at the folding pathway of this protein at atomic-level detail. Refolding simulations showed that the Leu16 – Leu34 contact between HI and HII must be formed in order for folding to proceed from the intermediate state to the native state, which explains why the Leu16Ala mutant is stabilized in the intermediate state. They also pointed to the formation of nonnative salt bridges as the reason that folding from the intermediate to the native state is slower than folding from the unfolded state to the intermediate.

We next investigated the unfolding pathway of EnHD at high concentration in Chapter 4. We created a single simulation system with 32 copies of the protein and a concentration of 18 mM, and we simulated it under native and thermally denaturing conditions. This system allowed us to investigate the types of contacts (hydrogen bonds, hydrophobic, or nonspecific contacts) that occur within one molecule of EnHD as compared with the types of contacts that occur between molecules. Our simulations showed extensive aggregation of EnHD, especially at high temperature, and analysis of the intermolecular contacts showed that aggregation was driven by the burial of hydrophobic residues and that aggregation could bury more hydrophobic surface area than folding. We also compared the native state and unfolding pathway at this higher concentration with our traditional single-molecule simulations. Despite the high concentration of protein in our system, the native state and unfolding pathway were both largely unaffected by the presence of neighboring molecules.

The engineered protein, UVF, discussed in Chapter 5, is a thermostable version of EnHD designed by the Mayo group (Gillespie *et al.* 2003, Shah *et al.* 2007). The protein was

designed to contain only hydrophobic residues at buried locations and only polar and charged residues at surface locations. The melting temperature was successfully raised from 325 K for EnHD to >373 K for UVF (52 °C and >99 °C, respectively). This engineered protein provides an excellent opportunity to investigate the molecular-level details of thermostability in comparison with its naturally occurring counterpart. Our simulations of EnHD and UVF showed that UVF was more dynamic than EnHD at room temperature and made more unique contacts. When the temperature was raised to 373 K, UVF maintained its increased dynamics without unfolding whereas EnHD unfolded. Analysis of the number and types of contacts that occurred within and between the buried and surface regions of the two proteins provided a structural explanation for why UVF was more stable than EnHD at high temperature. These conclusions also provide insight for future protein designs of thermostable proteins and suggest there is a balance that must be maintained between stability and rigidity.

In Chapter 6, the folding pathways of pair of proteins that were engineered to have 88% sequence identity (49 of 56 residues) but different tertiary structures is discussed.  $G_A88$  is a three-helix bundle with a long N-terminal tail whereas  $G_B88$  is a single  $\alpha$ -helix across two  $\beta$ -hairpins that form a sheet (Alexander *et al.* 2007, He *et al.* 2008). In combination with experimental work done in the Gianni Group, our simulations point to specific contacts that form in the denatured state that commit the protein to the all- $\alpha$  or  $\alpha/\beta$  fold. These interactions also explain experimental data that show  $G_B88$  to have higher native helical content in the denatured state than  $G_A88$ .

The study of  $G_A88$  and  $G_B88$  continues in Chapter 7 with the creation of mutants to knockout and add interactions in the denatured state that were found in the previous chapter to commit the proteins to their native topologies. We were successful in knocking out the identified interactions through charge-neutralizing mutations and creating several of the desired interactions in our unfolding simulations, and experimental work to test these predictions is ongoing in the Gianni Lab. A further iteration of this pair of proteins was considered where a single residue determined whether the protein folded to the all- $\alpha$  or  $\alpha/\beta$  fold, as were the two original template proteins for the engineering process. Analysis of unfolding simulations of these four proteins in addition to  $G_A88$  and  $G_B88$  showed that the

two hairpins of the  $\alpha/\beta$  fold formed independently of each other and that the all- $\alpha$  fold had more long-range interactions in the denatured state.

Finally, the work concludes in Chapter 8 with a review discussing the folding and dynamics of engineered proteins. The review focuses primarily on proteins that were designed computationally and whose dynamics and folding pathway have been studied by MD simulation. It discusses the kinetics of the simplest proof-of-principle designs that successfully repacked the hydrophobic core of existing proteins. It then moves on to explore the stability and folding pathway of proteins that were not designed based on a preexisting structure, so the backbone scaffold had to be designed in addition to the side chain sequence and packing. Finally, it discusses how native dynamics affected proteins that were designed for function, either to bind a ligand or catalyze a chemical reaction.

## 1.5 Insights to Protein Folding and Dynamics

This work has described the folding pathway and dynamics of EnHD and several engineered proteins in more detail than ever before, and the conclusions found here can be extended to better understand folding in other proteins as well. Thermal unfolding simulations have been shown here to reproduce the folding pathway in reverse, justifying such studies in other proteins. The Leu16Ala mutation in EnHD was found here to stall folding in the intermediate state due to the importance of the Leu16 – Leu34 contact to the folding pathway. If similarly important contacts are identified in other proteins, they may also be interrupted to create intermediate-stabilized mutants for further study. Residues involved in the nonnative salt bridges that slowed folding in EnHD could potentially be mutated to create an even faster folding variant of the protein (and in fact have; Gianni *et al.* 2003). However, nonnative salt bridges and hydrogen bonds were also shown to be useful in  $G_A88$  and  $G_B88$  for committing the proteins to their all- $\alpha$  or  $\alpha/\beta$  topology, respectively. This finding emphasizes the importance of nonnative salt bridges in both slowing and directing the folding pathway. Scientists designing proteins should consider the folding traps they might be creating when utilizing charged residues and balance them with stabilizing, on-pathway salt bridges in their designs.

Work on the dynamics of designed proteins has shown the importance of considering entropy in the design process. UVF was designed to be thermostable by maximizing enthalpic interactions in EnHD, its native template. UVF was shown here to be more dynamic than its naturally occurring counterpart, and this gain in entropy helped it maintain its folded structure at high temperature. The idea of purposefully increasing the entropy when designing thermostable proteins could be incorporated into future design strategies. However, this technique may be less useful for design targets that must maintain the geometry of a specific binding site. A collection of studies on engineered proteins was reviewed here and further emphasized the importance of dynamics to successful designs. Native dynamics, entropy, and interactions in the denatured state are all difficult to target directly given current computational design methodology, but they would be beneficial to incorporate into the design process in the future.

## **Chapter 2: Microscopic Reversibility of Protein Folding in Molecular Dynamics Simulations of the Engrailed Homeodomain**

### **2.1 Summary**

The principle of microscopic reversibility states that at equilibrium the number of molecules entering a state by a given path must equal those exiting the state via the same path under identical conditions, or in structural terms, that the conformations along the two pathways are the same. There has been some indirect evidence indicating that protein folding is such a process, but there have been few conclusive findings. In this study, we performed molecular dynamics simulations of an ultra-fast unfolding and folding protein at its melting temperature in order to observe, on an atom-by-atom basis, the pathways the protein followed as it unfolded and folded within a continuous trajectory. In a total of 0.67  $\mu\text{s}$  of simulation in water, we found 6 transient denaturing events near the melting temperature (323 K and 330 K) and an additional refolding event following a previously identified unfolding event at high-temperature (373 K). In each case unfolding and refolding transition state ensembles were identified, and they agreed well with experiment based on comparison of S- and  $\Phi$ -values. Based on several structural properties, these 13 transition state ensembles agreed very well with each other and with 4 previously identified transition states from high-temperature denaturing simulations. Thus, not only were the unfolding and refolding transition states part of the same ensemble, but in five of the seven cases, the pathway the protein took as it unfolded was nearly identical to the subsequent refolding pathway. These events provide compelling evidence that protein folding is a microscopically reversible process. In the other two cases, the folding and unfolding transition states were remarkably similar to each other, but the paths deviated.

### **2.2 Introduction**

In 1925, Richard C. Tolman coined the term “microscopic reversibility” in reference to chemical reactions:



In recent years increasing use has been made of a new postulate which perhaps cannot yet be stated in its final form, but which requires in a general way in the case of a system in thermodynamic equilibrium not only that the total number of molecules leaving a given state in unit time shall on the average equal the number arriving in that state in unit time, but also that the number leaving by any particular path shall on the average be equal to the number arriving by the reverse of that particular path, thus excluding any cyclical maintenance of the equilibrium state. The writer has ventured to name this postulate *the principle of microscopic reversibility* (Tolman 1925).

This description was recast into structural terms in 1967 by Frank H. Westheimer (1968) and was adopted by the IUPAC in 1999:

In the case of  $S_N2$  reactions at tetrahedral centers implying a formation of the trigonal bipyramid transition state (or intermediate) structure, the original formulation of the principle was extended in the following way: if a molecule or reactant enters a trigonal bipyramid at an apical position, this (or another) molecule or reactant must likewise leave the trigonal bipyramid from an apical position (Minkin 1999).

The hypothesis that protein folding may, like chemical reactions, be microscopically reversible has since been offered. If this hypothesis is true, one would expect to observe identical transition states for folding and unfolding, and major events on the folding pathway would occur in reverse order in the unfolding pathway. Supporting evidence has been presented by Jackson *et al.* who showed through  $\Phi$ -value analysis on chymotrypsin inhibitor 2 (CI2) that folding and unfolding transition states are the same, suggesting that the pathways are also the same (Jackson *et al.* 1993). Additionally, molecular dynamics (MD) generated unfolding transition states (TS) of CI2 are in quantitative agreement with experimental data collected for both unfolding and refolding (Li and Daggett 1994; Daggett *et al.* 1996). Furthermore, the unfolding and direct refolding pathways of CI2 were shown to be the same in a single continuous MD trajectory (Day and Daggett 2007).

The latter MD study was done at the melting temperature ( $T_m$ ) of the protein, at which point the folding and unfolding rates are equal and  $\Delta G = 0$ . For these reasons, exchange between the folded and unfolded states is dependent on the energy barrier,  $\Delta G^\ddagger \approx 2.3$  kcal/mol (Itzhaki *et al.* 1995), and unfolding and refolding may occur in a single trajectory on time scales tractable by MD. The protein, CI2, passed through three different states, native (N), nearly native (N'), and denatured (D), then returned to N' over a time period of about

60 ns. When moving from N' to D and back, unfolding and refolding transition states were identified, and they were the same. The C $\alpha$  root-mean-square deviation (RMSD) to the native structure and internal contacts were analyzed to differentiate between the three different states. Day and Daggett (2007) defined N', an alternate, stable state for CI2 at elevated temperature. The protein passed through N' before it moved through its TS to D. N' was characterized by many near-native interactions but elongated contact distances. In particular, Trp5, a fluorescence unfolding probe, was buried in both N and N', but not in D, as would be expected if both N and N' were native but D was not. D had a disrupted hydrophobic core and loss of secondary structure. This CI2 study showed direct unfolding and refolding in a single continuous trajectory by the same structural pathways for the first time. Consequently, this behavior needs to be demonstrated in another system to ensure it is reproducible, which we describe here.

The engrailed homeodomain (EnHD) of *Drosophila melanogaster* is a 61-residue three-helix bundle. It is ultra-fast folding ( $k_F = 37,500 \text{ s}^{-1}$  at 298 K and  $51,000 \text{ s}^{-1}$  around 315 K) and unfolding ( $k_U = 1,100 \text{ s}^{-1}$  at 298 K and  $205,000 \text{ s}^{-1}$  at 336 K), and its folding and unfolding pathways have been extensively characterized through combined experimental and MD studies (Mayor *et al.* 2000; Gianni *et al.* 2003; Mayor *et al.* 2003a; Mayor *et al.* 2003b). Folding for EnHD follows the framework model involving the docking of HI (residues 10-22), HII (28-38), and HIII (42-55; DeMarco *et al.* 2004). These properties make EnHD especially well-suited for MD folding studies.

In this study, we performed MD simulations of EnHD near its  $T_m$ ,  $52 \text{ }^\circ\text{C} = 325 \text{ K}$  (Mayor *et al.* 2003), to compare unfolding and refolding under identical conditions. We analyzed 5 simulations, 3 at 323 K and 2 at 330 K. We compared them to 4 previously described thermal denaturation simulations, 2 each at 373 K and 498 K, and one native simulation at 298 K. The first of the 373 K simulations was found to contain a region of particular interest that had not been previously reported, so that simulation was also analyzed in detail. We identified and characterized 3 different states populated during unfolding and refolding: N, N', and D. We also found 6 transient denaturing events in which EnHD partially unfolded and refolded in 3 of the 5  $T_m$  simulations; from this 12 unfolding and refolding transition state ensembles (TSE) were identified. EnHD unfolded in the 373K/1

simulation, and a TSE in agreement with experiment was reported previously (Mayor *et al.* 2000a; Gianni *et al.* 2003; DeMarco *et al.* 2004). Further investigation of this simulation showed that EnHD later refolded, so we also describe this high temperature refolding TSE. These 13 TSEs agree well with the 4 previously identified unfolding TSEs from the 4 high-temperature simulations. Besides defining TSEs, we analyzed the entire pathway EnHD followed as it unfolded and refolded. Five of the 7 refolding pathways were nearly identical to the unfolding pathways that preceded them. These 5 examples are further evidence that the ensembles of folding and unfolding pathways are one and the same, and that protein folding is a microscopically reversible process. However, in the other 2 cases, EnHD passed through remarkably similar unfolding and refolding transition states, but only a portion of the actual refolding pathway was similar to the unfolding pathway.

## 2.3 Methods

### 2.3.1 Molecular Dynamics Simulations

A total of 9 MD simulations are addressed in this paper at the following temperatures with simulation times in parentheses: 298 K (100 ns), 323 K (100 ns, 50 ns, 42 ns), 330K (100 ns, 100 ns), 373 K (24 ns, 75 ns), 498 K (20 ns, 60 ns), for a total of 0.67  $\mu$ s. All 4 of the 373 K and 498 K simulations have been described previously (Mayor *et al.* 2000a; Gianni *et al.* 2003; DeMarco *et al.* 2004), as have the first 2 323 K and the 298 K simulations (Beck and Daggett 2007).

Both of the 330 K simulations were performed using our in-house molecular dynamics package, *in lucem* molecular mechanics (*ilmm*; Beck *et al.* 2000-2012) with the Levitt *et al.* (1995) force field using previously described protocols (Beck and Daggett 2004). The crystal structure (PDB ID: 1enh) was minimized for 1000 steps and solvated in a box of F3C water molecules (Levitt *et al.* 1997) such that there was at least 12 Å between the protein and the edge of the periodic box. The density was set to 0.985 g/mol in agreement with the experimentally-determined liquid-vapor coexistence curve for this temperature (Kell 1967). 1000 steps of steepest descent minimization were performed on the water alone followed by 1 ps of dynamics. Next, the water and the protein were independently minimized for an additional 500 steps. Production simulations were performed for 100 ns

allowing all atoms to move with structures written out every 1 ps. Long-range interactions were truncated after 8 Å using a force-shifted nonbonded cutoff. Our force-shifted cutoff method at this distance is the most effective treatment of long-range interactions based on computational savings, energy conservation, and ability to reproduce experimental results (Beck *et al.* 2005). The 323K/3 simulation followed the same protocol, except there was only 8 Å of padding between the protein and the edge of the box, the protein was minimized for 200 steps before adding water, and the simulation was run for 42 ns.

### 2.3.2 C $\alpha$ RMSD Matrix and 3D MDS

All-vs.-all C $\alpha$  RMSD matrices were calculated to identify clusters of structures with similar conformations. Granularities were chosen to give 1000-5000 time points over the period of interest. The C $\alpha$  RMSD between each structure and every other structure was computed, resulting in a matrix with 1000<sup>2</sup>-5000<sup>2</sup> data points. Low C $\alpha$  RMSD boxes on the diagonal represent a period of time during which the protein stayed in a particular conformation. When these boxes lie off the diagonal, they indicate conformations of similar structure visited discontinuously in time. As described previously (DeMarco *et al.* 2004), the “core” (residues 8-53) was usually used to calculate C $\alpha$  RMSD, rather than the whole protein (residues 3-56). Since the fluctuations of the terminal residues are not indicative of the overall motion of the protein and introduce noise, the 5 residues at the N-terminus and 3 at the C-terminus were not included where specified.

Using the program R (Team 2004), multidimensional scaling (MDS) was performed to project the matrix down to 3 dimensions. This scaling results in a 3D plot in which each point represents a structure, and the distance between any two points is proportional to the C $\alpha$  RMSD between the respective structures. The points are connected in order of time for the period of interest. As with the matrix, a series of points close together indicates that the structures are similar. Using this plot, TSEs were chosen as the structures representing the last 5 ps leaving the extended native cluster for an unfolding event (Li and Daggett 1994) or the structures representing the first 5 ps upon returning to the native cluster for a refolding event.

### 2.3.3 HIII-Core Distance Calculation

The distance between the closest backbone atoms in the C-terminus of HIII and the HI-HII scaffold in the crystal structure was chosen to represent the movement of HIII. The atoms chosen were the C $\alpha$  of Phe20 and the backbone carbonyl C of Lys52, and the distance was measured at 10-ps granularity. Since the 373K/1 simulation was so short, a granularity of 1 ps was used to give consistent sampling.

### 2.3.4 Average Structures

Average structures were calculated using 100 ps granularity for the long N and N' time-spans. For TS structures, all 6 structures in the TSE were included in the average. The C $\alpha$  RMSD of the core residues (8-53) between the average structures was then calculated.

### 2.3.5 HIII-Core Contacts

A contact for a pair of residues was defined based on whether any one of the heavy atoms in the first residue was below a set cutoff of any heavy atom in the second residue. This cutoff was defined as 5.4 Å for carbon-carbon distances and 4.6 Å for all other atom pairs. For Figure 2.4, the calculation was taken over the time period of interest with 1-ps granularity, and the percentage of structures in which the two specified residues were in contact was reported for the average measurements. For the whole-simulation graphs (Figure 2.1c), each of the contacts is listed as a separate horizontal line, and if the contact was present at the given time point along the x-axis, a cross (+) was plotted. A granularity of 10 ps was used.

To choose which contacts to report, we identified residue pairs in which one member of the pair was in HIII and the other was not. Of these, the only pairs that were selected were those that were in contact at least 25% of the time in the native (298 K) simulation.

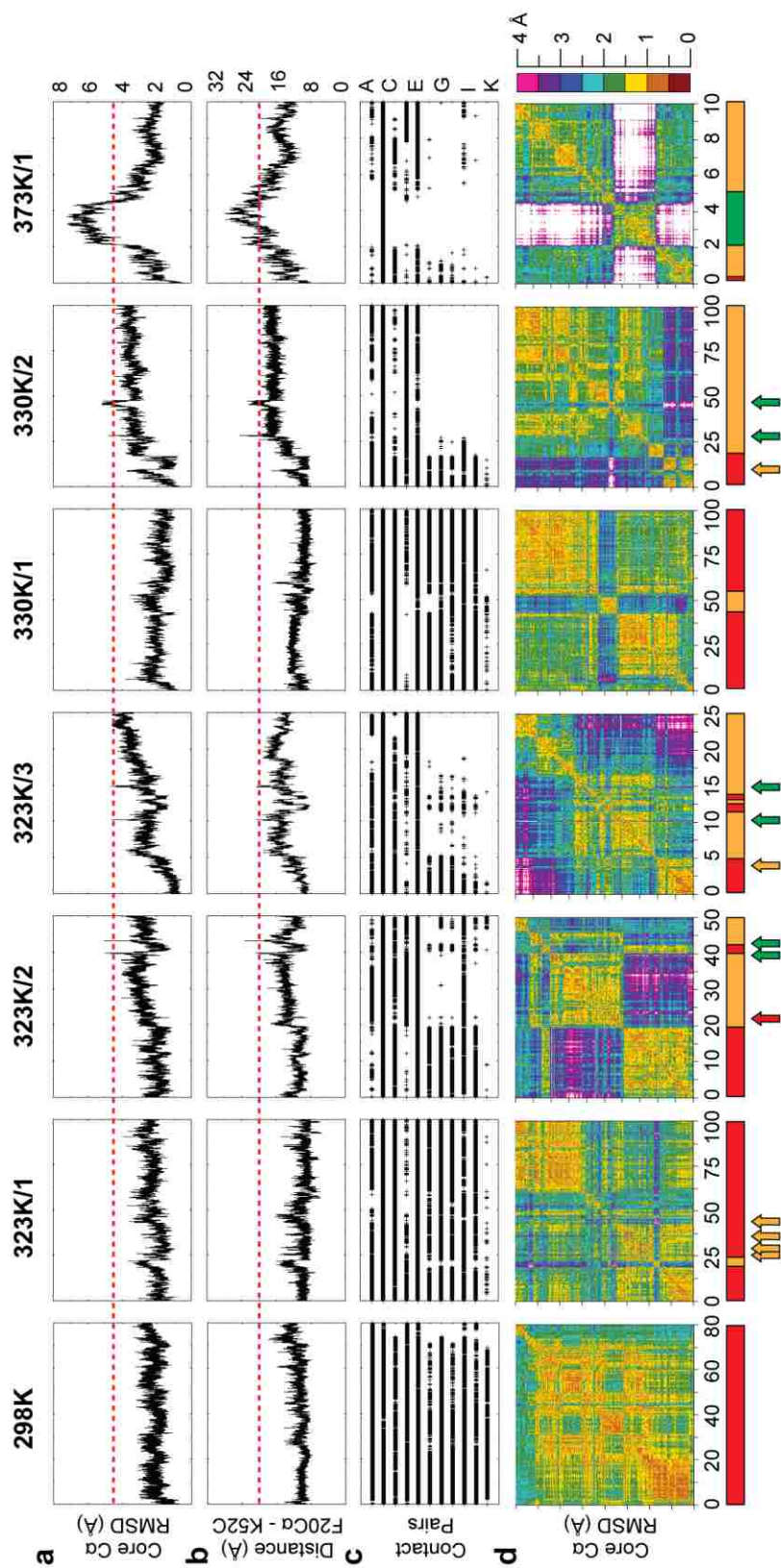
### 2.3.6 Calculation of S-values

The S-value is a semi-quantitative structure index that provides an overall measure of the secondary and tertiary structure for each residue of the protein (Daggett *et al.* 1996). S-values agree well with experimental  $\Phi$ -values for a variety of proteins (Daggett *et al.* 1996; Daggett *et al.* 1998; Li and Daggett 1998; Fulton *et al.* 1999; Gianni *et al.* 2003; Day and Daggett 2005). The S-value is the product of the extent of native secondary structure (S<sub>2°</sub>)

and native and nonnative tertiary contacts ( $S_{3^\circ}$ ) present in a given structure relative to the number of contacts in the crystal structure. A value of 1 for  $S$  corresponds to native-like extent of structure in the TS, while a value of 0 suggests the residue is unstructured. As previously described (Gianni *et al.* 2003),  $S_{3^\circ}$  was used in place of  $S$  for residues Phe8, Leu26, and Leu40. For these three residues, side-chain interactions were maintained despite disorder in the main chain. Consequently, the product of  $S_{2^\circ}$  and  $S_{3^\circ}$  did not accurately represent the degree of structure retention.

### 2.3.7 Protein – DNA Interactions

In order to generate a semi-quantitative measure of whether MD-generated EnHD conformers bind DNA, we measured distances between the DNA sugar-phosphate backbone and the HI-HII helical hairpin using the crystal structure of EnHD bound to DNA (PDB ID: 3hdd). The crystal structure contains two nearly identical EnHD structures (core  $C\alpha$  RMSD = 0.27 Å), so we selected the one bound to the ideal TAATTA sequence for all measurements (Fraenkel *et al.* 1998). EnHD binds DNA primarily through residues in HIII (major groove) and the N-terminus (minor groove) (Kissinger *et al.* 1990; Fraenkel *et al.* 1998). Since the N-terminus becomes structured only upon binding DNA (Fraenkel *et al.* 1998), its conformation during our simulations should have no bearing on whether free EnHD is structured enough to bind DNA. Using Profit (Martin 1992-2001), the MD structures were first aligned to the DNA-bound structure based on a least-squares fit of the  $C\alpha$  atoms in HIII, the DNA-binding helix. We selected a pairs of residues for the distance measurements, representative of one of two hydrogen bonds in the DNA-bound crystal structure that did not involve HIII or the N-terminus of EnHD. The atoms chosen for the measurement were the  $C\alpha$  of Tyr25 from EnHD and backbone P of Thymine28 from the DNA. For measurements over time, 10 ps granularity was used. A period of time beginning 1 ns after the TS and continuing for 1 ns was selected to represent D for all 4 high-temperature unfolding simulations. Since there was not a full nanosecond of denatured time for most of the 4 lower-temperature simulations, the most denatured structure based on 3D MDS of the  $C\alpha$  RMSD matrix was used.



**Figure 2.1: General properties for each simulation**

(a)  $C\alpha$  RMSD of the core (residues 8-53) calculated over time for each simulation relative to the 0 ns structure. Values above  $\sim 4.5$  Å are indicative of movement to D, as indicated by the dashed red line. (b) Distance between the  $C\alpha$  of Phe20 and the backbone C of Lys52 over time.  $N'$  is characterized by values of  $\sim 15$  Å, while distances of greater than 20 Å (dashed red line) appear when the protein moves to D. (c) Contacts made between residue pairs. Alternate pairs are labeled on the right from top to bottom: (A) Ile45 – Leu38, (B) Ile45 – Leu40, (C) Trp48 – Leu16, (D) Phe49 – Leu16, (E) Phe49 – Phe20, (F) Phe49 – Arg24, (G) Phe49 – Leu26, (H) Lys52 – Phe20, (I) Arg53 – Arg24, (J) Arg53 – Arg24, (K) Arg53 – Leu26. Contacts 49-24, 49-26, 52-20, and 53-24 are characteristically present in N but not  $N'$ , while additional contacts are lost during D. (d) An all-vs.-all core  $C\alpha$  RMSD matrix. Low- core  $C\alpha$  RMSD squares on the diagonal represent a period of time with similar structures, and when they are off the diagonal, they indicate that the structures from the two corresponding time periods are similar. Below each matrix is a timeline depicting the different states the protein takes in each simulation: N (red),  $N'$  (orange), and D (green). Arrows represent transiently occupied states.

## 2.4 Results

A total of 10 MD simulations were performed at 5 different temperatures (298K, 323K, 330K, 373K, and 498K). We describe the major conformational states of EnHD in each of the 10 simulations: N, N', TS, and D. When EnHD was in N, HIII was docked against the HI-HII scaffold. N' was characterized by a slight movement of HIII towards the N-terminus without losing many contacts or solvating the hydrophobic core. When HIII moved out and away from the HI-HII scaffold, exposing the hydrophobic core, EnHD was deemed to be in D. The protein was not necessarily unfolded in D, but it was not native nor, by definition, biologically active. Whenever the protein moved from N' to D or from D back to N', a TS was identified. These four states will be discussed further in the context of each simulation.

### 2.4.1 Overview of Simulations

EnHD remained folded in the native 298 K simulation with a core C $\alpha$  RMSD of  $2.1 \pm 0.3$  Å (average  $\pm$  1 standard deviation) and an HIII-core distance (Phe20 C $\alpha$  – Lys52 carbonyl C, see Methods) of  $10.6 \pm 1.4$  Å for the first 80 ns of the 100 ns simulation (Figure 2.1a,b). During the final 20 ns of the simulation, the 9 C-terminal residues formed a  $\pi$ -helix, but HIII remained docked to the HI-HII scaffold. For this reason, the final 20 ns are not considered here.

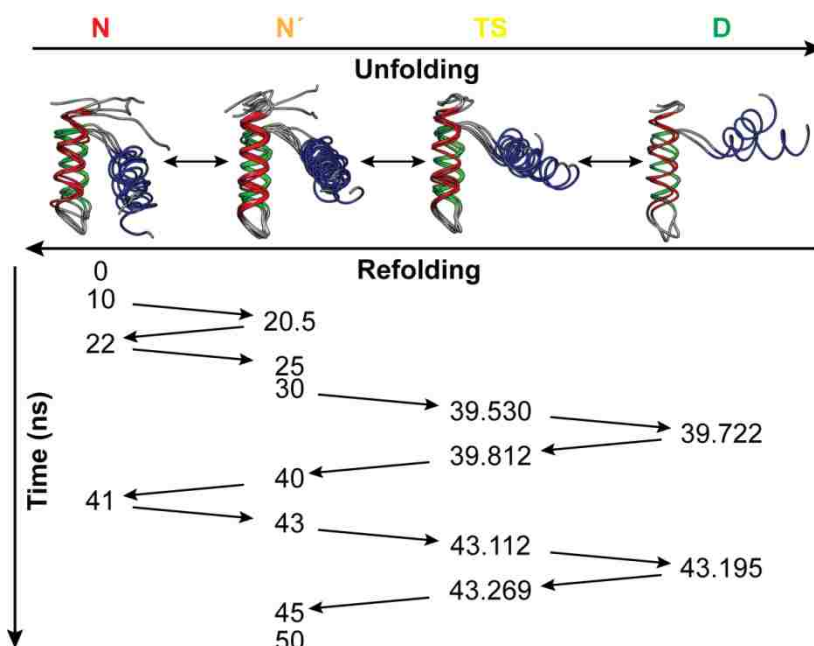
In the 323K/1 100 ns simulation, EnHD stayed mostly in N but briefly moved to N' from 18-23 ns. It transiently moved to N' about 4 more times over the next 30 ns then stabilized in N for the remainder of the simulation (Figure 2.1). The protein did not populate D during this simulation, so there were no TSs identified.

EnHD was in N for the first  $\sim$ 20 ns of the 323K/2 simulation. At 19 ns, HIII moved  $\sim$ 10 Å towards the N-terminus, entering N' (Figure 2.3a). It remained in N' until 39 ns when there was a large jump in core C $\alpha$  RMSD and HIII-core distance, reflecting the undocking of HIII and entrance into D. HIII moved far enough away from the core (20 Å) for it to lose 10 of its 11 native core contacts (Figure 2.1c) and for the hydrophobic core to be solvated. This altered position was only transient, however, and HIII moved back to its position in N' a short 0.28 ns later. Over the next 1 ns, HIII moved back to its N position where it stayed for  $\sim$ 3 ns. HIII then returned to its N' position transiently before the protein once again entered



D at 43 ns. This transition was marked by another jump in core C $\alpha$  RMSD, HIII-core distance, and loss of contacts (Figure 2.1). After 0.16 ns, the protein returned to N', where it remained for the duration of the 50 ns simulation. The structures of EnHD in all four of its states are shown in Figure 2.2, and these structures are representative of those seen in the other 6 simulations.

The first 5 ns of the 323K/3 simulation were spent mostly in N, after which EnHD stabilized in N' for another 5 ns. There was a transient movement to D at 10 ns, indicated by a spike in core C $\alpha$  RMSD and HIII-core distance. There was a transition back to N for 2 ns with a transient jump to N' during that time. Then at 13 ns, there was a more stable transition to N', interrupted only by a transient jump to D at 14 ns. EnHD remained in N' until the end of the simulation. Only the first 25 ns of the 42 ns simulation are considered here in order to focus on the transitions of interest.

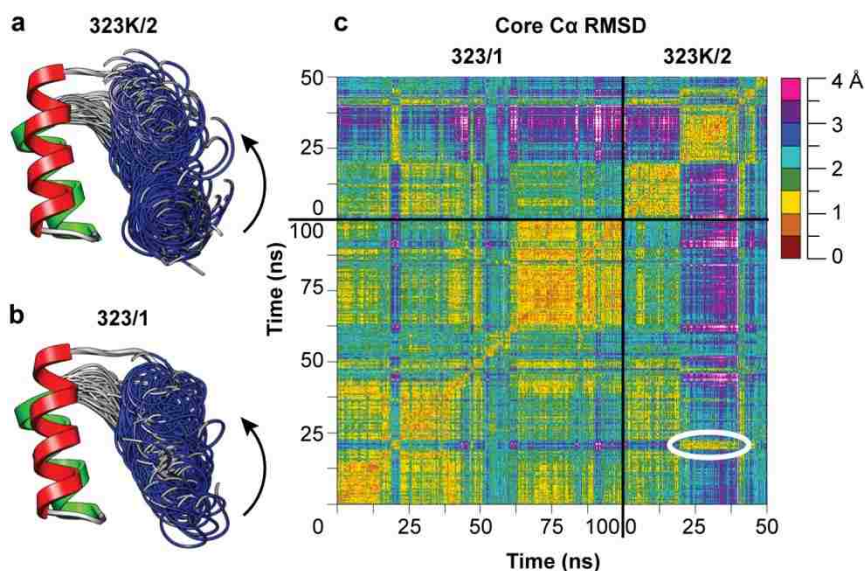


**Figure 2.2: Structures and order of population of the 4 states in the 323K/2 simulation**

Under each state is a set of representative structures from the course of the simulation overlaid and fit on HI-HII C $\alpha$  atoms. HI (residues 8-20) is colored in red, HII (26-36) in green, and HIII (42-55) in blue. The time point in ns of each structure is listed below, and arrows connect them in the order they occurred. Structures within each state are similar, while each state is distinct. In N and N', HIII forms a  $\sim 15^\circ$  angle with the HI-HII scaffold. When EnHD reaches the TS, the angle is  $\sim 30^\circ$ , and it becomes even wider in D.

EnHD was stable in N for the majority of the 330K/1 simulation, with the exception of 43-53 ns (Figure 2.1). During this time, HIII unwound slightly at the N-terminus causing

it to move towards the N-terminus by a register, much like the movement previously described for N'. Because the movement was due to unwinding rather than a simple loop movement, N' in this simulation had somewhat different properties (core C $\alpha$  RMSD HIII-core and contact pattern) than N' in the 323 K simulations (Figure 2.1). The 330K/2 simulation began with 8 ns of N, moved to N' for 1 ns, returned to N for 8 ns, then shifted to N' again. At 27 ns, the protein moved from N' to D for 0.33 ns, then returned to N' at 28 ns. Again, at 44 ns, EnHD moved from N' to D and stayed there until 47 ns when it returned to N'. It remained in N' for the duration of the 100 ns simulation (Figure 2.1). Each of the three non-transient N' states here agree very well with N' in the 323 K simulations as shown in Figure 2.2.



**Figure 2.3: N and N' structures and all-vs.-all core C $\alpha$  RMSD matrix for 2 323K simulations**

(a) N and N' from the first 38 ns of the 323K/2 simulation taken every 0.5 ns for HIII. (b) N and N' from the first 100 ns of the 323K/1 simulation with structures taken every 1 ns for HIII. HI-HII is from the 0 ns structure. HIII moves towards the N-terminus but not out, so the hydrophobic core is not solvated. (c) All-vs.-all core C $\alpha$  RMSD matrix for the 323K/1 and 323K/2 simulations. 323K/1 and 323K/2 are in N' for the longest time from 19-20 ns and 20-39 ns, respectively. The color of the circled, low core C $\alpha$  RMSD box off the diagonal indicates that these two N' states are as similar to each other as they are to themselves. Smaller boxes can be seen off the diagonal for the points in the 323K/1 simulation when EnHD moves to N' transiently indicating these transient N' states are the same as the longer two.

EnHD moved from N to N' within the first 0.5 ns of the 373K/1 simulation, then proceeded on to D. An unfolding TS was previously identified at 1.720 ns (Mayor et al. 2000; Gianni et al. 2003; DeMarco et al. 2004), and we identified a new refolding TS shortly

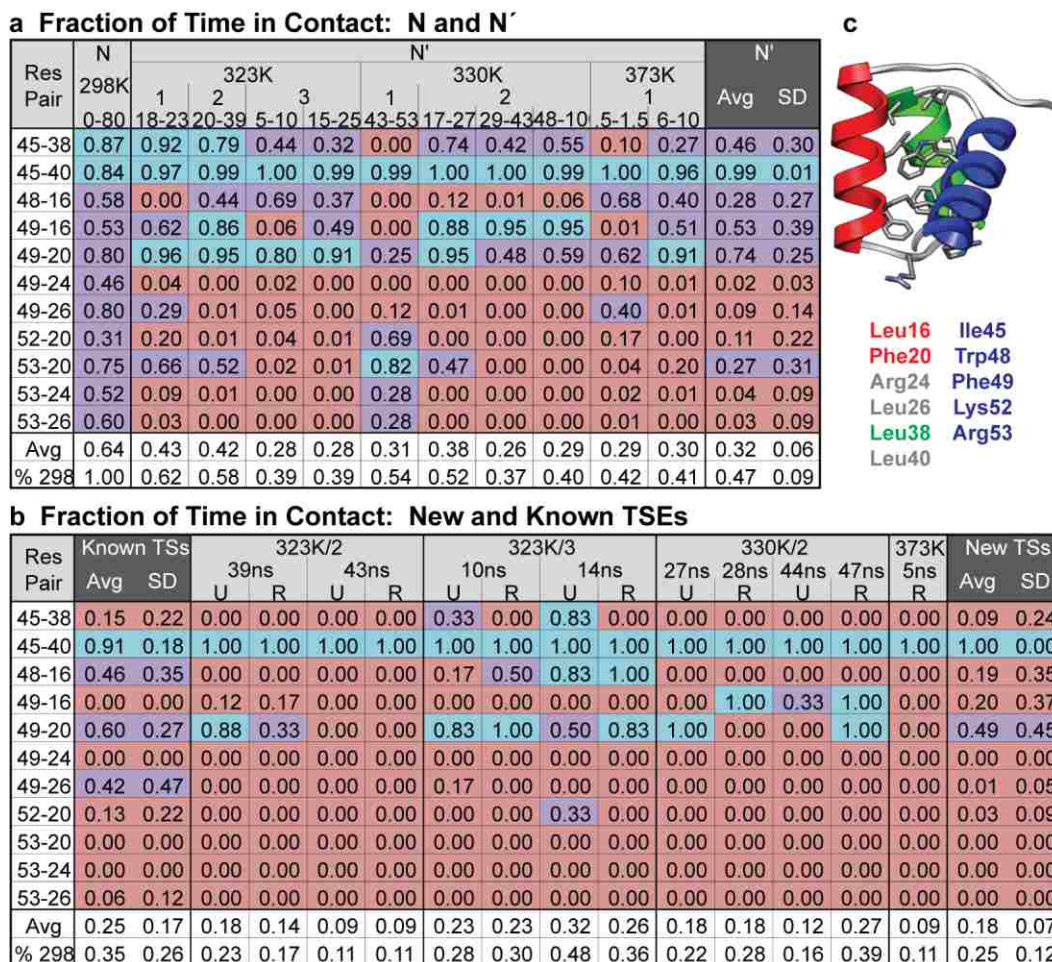
after 5 ns. After EnHD returned to N', it remained there for the remainder of the 24 ns simulation. We will focus only on the first 10 ns of the simulation here. After refolding, N' did not agree very well with N' in the other simulations near the T<sub>m</sub>, due to the fluctuating  $\pi$ -helical structure adopted by the C-terminus of HIII around 7 ns (data not shown). The temperature of this simulation was nearly 40 K above the T<sub>m</sub> of EnHD, so refolding is expected to occur only rarely and incompletely. We were lucky to observe such an event so that we can compare folding and unfolding within a continuous trajectory at high temperature.

## 2.4.2 Native' State at Elevated Temperature

Before HIII moved far away enough from the HI-HII scaffold for EnHD to become “denatured,” it occupied two distinct positions. The conformation of EnHD with HIII positioned most similarly to the crystal structure is N, and we define N' as the state in which HIII slides ~10 Å towards the N-terminus (Figure 2.3a,b).

In all three of the 323 K simulations and the second 330 K simulation, the same N' structures were observed based on core C $\alpha$  RMSD and HIII-core contact similarity (Figure 2.3c, 2.4a, 2.5a). Based on the core C $\alpha$  RMSD matrix of the first two 323 K simulations (Figure 2.3c), the conformations from 18-23 ns and the subsequent transient deviations from N in 323K/1 were the same as those from 20-40 ns and 42-50 ns (with the exception of the unfolding events) in 323K/2. The core C $\alpha$  RMSD between average structures of N' in the 323K/1-3 and 330K/2 simulations also showed this similarity (Figure 2.5a). The 0.5-1.5 ns N' in 373K/1 was in good agreement with N' in 323K/1-3 and 330K/2 based on core C $\alpha$  RMSD, but the 6-10 ns N' was less similar due to disruption of HIII.

The fraction of time the HIII-core contacts were made was remarkably similar for all 10 non-transient N' conformations, with the exception of 330K/1 (Figure 2.4a). The average standard deviation for contact time over all 11 residue pairs was 19%, and if the 330K/1 simulation was excluded, it dropped to 15% (Figure 2.4). Contact vs. time plots are useful to probe which pairs are in contact over the course of a simulation (Figure 2.1c). Based on this, contacts Phe49 – Arg24 (F), Phe49 – Leu26 (G), Lys52 – Phe20 (H), and Arg53 – Arg24 (K) were lost when EnHD moved from N to N' and were mostly regained if it moved back to N from N'.



**Figure 2.4: HIII-core residue pairs fraction of time in contact for N, N', and new TSEs in each simulation**

(a)-(b) are colored: < 25% pink, 25-75% purple, >75% blue. (a) Contacts for N in the 298K simulation and N' in the 323, 330, and 373K simulations. The temperature of the simulation, simulation number, and time span in ns of the N' state is given. (b) Contacts for new and previously identified TSEs. The simulation temperature and number is given (373K is 373K/1) as well as the ns during which the unfolding (U) or refolding (R) TSE occurred. The average fraction of time in contact was reported for each simulation as was the fraction of time in contact relative to the native (298K) simulation. Additionally, the average and standard deviation of the fraction of time in contact was calculated for each contact pair across all compared time spans. The average fraction of contacts is quite different for each of the states, N, N', and TS, with a low standard deviation, and those contacts that are lost are lost fairly consistently. (c) Each of the selected contact residues is shown on the EnHD structure colored by atom.

**a Core C $\alpha$  RMSD for N and N' Average Structures**

		N	323K						330K				373K	
		min	1	2	3	1	2	1	2	1	2	1	2	
		298K 0-80	18-23	20-39	5-10	15-25	43-53	17-27	28-43	48-100	5-15	6-10		
N	min	0.00	1.82	2.22	2.72	2.41	3.10	2.20	2.68	3.10	3.22	2.05	2.22	
	298K 0-80ns	0.00	1.69	2.48	2.23	2.82	2.26	2.41	2.62	2.70	1.91	3.02		
N'	323K/1 18-23ns		0.00	0.96	1.22	1.60	1.77	1.01	1.57	1.58	1.53	2.82		
	323K/2 20-39ns			0.00	1.57	1.27	2.26	0.60	1.39	1.28	2.02	2.90		
	323K/3 5-10ns				0.00	1.68	1.74	1.66	1.72	1.86	1.09	2.94		
	323K/3 15-25ns					0.00	2.54	1.36	1.35	1.26	2.20	3.25		
	330K/1 43-53ns						0.00	2.38	2.61	2.67	1.86	2.80		
	330K/2 17-27ns							0.00	1.30	1.28	1.98	2.80		
	330K/2 28-43ns								0.00	0.40	2.12	3.10		
	330K/2 48-100ns									0.00	2.28	3.19		
	373K/1 0.5-1.5ns										0.00	2.79		
	373K/1 6-10ns											0.00		

**b Core C $\alpha$  RMSD for New and Known TSE Average Structures**

		N	New TS										Known TS						
		min	323K					330K					373K	498K					
			39 U	39 R	43 U	43 R	10 U	10 R	14 U	14 R	27 U	28 R	44 U	47 R	5 R	1 U	1 U	1 U	2 U
New TS	min	0.00	3.46	3.35	3.17	3.50	2.58	2.60	3.02	2.86	3.30	3.62	4.10	3.01	3.39	2.53	2.68	2.86	2.93
	323K/2 39ns Unfold		0.00	0.65	2.27	2.53	2.70	2.83	2.74	1.77	2.55	1.99	2.20	2.35	3.16	2.00	3.10	4.14	3.95
	323K/2 39ns Refold			0.00	2.09	2.32	2.49	2.63	2.45	1.41	2.42	1.86	2.21	2.24	3.00	1.89	3.02	3.84	3.63
	323K/2 43ns Unfold				0.00	0.65	1.31	1.35	1.74	1.84	1.11	1.74	2.14	2.01	1.91	1.70	2.22	2.72	2.90
	323K/2 43ns Refold					0.00	1.39	1.36	1.74	1.99	1.35	2.01	2.30	2.34	2.23	1.87	2.54	2.72	3.01
	323K/3 10ns Unfold						0.00	0.74	1.30	1.84	1.67	2.36	2.72	2.08	2.24	1.55	2.36	2.35	2.72
	323K/3 10ns Refold							0.00	1.56	1.92	1.81	2.51	2.79	2.25	2.41	1.46	2.14	2.09	2.63
	323K/3 14ns Unfold								0.00	1.57	2.18	2.41	2.71	2.23	2.70	1.92	2.50	2.65	2.97
	323K/3 14ns Refold									0.00	2.34	2.00	2.56	2.29	2.85	1.39	2.52	3.12	3.07
	330K/2 27ns Unfold										0.00	1.48	2.17	1.90	1.67	2.16	2.66	2.93	3.14
	330K/2 28ns Refold											0.00	1.75	1.96	2.03	2.37	3.00	3.66	3.60
	330K/2 44ns Unfold												0.00	1.77	2.65	2.60	3.14	3.93	3.95
	330K/2 47ns Refold													0.00	2.15	2.28	2.63	3.21	3.29
	373K/1 5ns Refold														0.00	2.71	2.94	3.08	3.06
	Known TS	373K/1 1ns Unfold														0.00	2.24	2.84	2.92
373K/2 1ns Unfold																0.00	2.45	2.80	
498K/1 0ns Unfold																	0.00	1.84	
498K/2 0ns Unfold																			0.00

**Figure 2.5: Core C $\alpha$  RMSD between average structures representative of N, N', and TS**

C $\alpha$  RMSDs reported are for the core residues of the average structure over the time periods indicated and are colored to visualize trends 0-1 Å (red), 1-2 Å (orange), 2-3 Å (yellow), 3-4 Å (green), 4-5 Å (blue). (a) Core C $\alpha$  RMSDs for average N' structures. The simulation temperature, number, and time span of N' is given. Excluding the later 373K/1 N' structure, all of the N' structures are within 3 Å core C $\alpha$  RMSD of each other, and with the exception of 330K/1, all of the low-temperature N' structures are within 2 Å. (b) Core C $\alpha$  RMSD for the 13 new average TSE structures and 4 previously identified. The simulation temperature, number, and time during which the unfolding (U) or refolding (R) TSE occurred is given. The average TSE structures are most similar at the same temperatures, and all are about the same core C $\alpha$  RMSD from N (~2.5-3.5 Å).

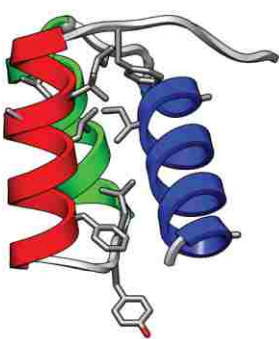
### 2.4.3 Properties of the Transition State Ensembles

A total of 13 new unfolding and refolding TSEs were identified in 4 simulations at 323 K, 330 K, and 373 K. The TSEs all had low core C $\alpha$  RMSD to each other ( $2.1 \pm 1.0$  Å), particularly the  $T_m$  TSEs ( $1.7 \pm 0.9$  Å; Figure 2.5b). The  $T_m$  TSEs were less similar to the high-temperature TSEs, with lower core C $\alpha$  RMSDs to the 373 K TSEs than the 498 K TSEs, but most were 2-4 Å core C $\alpha$  RMSD between any two high temperature and  $T_m$  TSE average structures. The core C $\alpha$  RMSD to the native state was  $3.1 \pm 0.4$  Å over all 17 TSEs, and the lowest core C $\alpha$  RMSDs were observed between the exit and reentry TSEs at 39 and 43 ns in the 323K/2 simulation ( $0.65$  Å in both cases).

The HIII-core contacts agreed very well between the new and previously identified TSEs (Figure 2.4b). The Ile45 – Leu40 contact was consistently sustained in all of the TSEs, likely due to the residues' positions in the HII-HIII loop and at the N-terminal end of HIII, respectively. Where there were dissimilarities in contacts made between the known and new TSEs, it was usually the case that there were less contacts made in the new, lower temperature TSEs.

Correlation Coefficients (R): S vs. $\Phi$ For Unfolding and Refolding TS Ensembles			
<b>323K/2</b>			
<u>39ns Unfold</u>	<u>39ns Refold</u>	<u>43ns Unfold</u>	<u>43ns Refold</u>
0.79	0.76	0.86	0.8
<b>323K/3</b>			
<u>10ns Unfold</u>	<u>10ns Refold</u>	<u>14ns Unfold</u>	<u>14ns Refold</u>
0.74	0.86	0.78	0.71
<b>330K/2</b>			
<u>27ns Unfold</u>	<u>28ns Refold</u>	<u>44ns Unfold</u>	<u>47ns Refold</u>
0.78	0.79	0.10	0.03
<b>373K/1</b>			
	<u>5ns Refold</u>		
	0.60		

- Phe8Ala\*
- Leu13Ala
- Ala14Gly
- Leu16Val
- Phe20Ala
- Tyr25Gly
- Ala25Gly
- Leu26Ala\*
- Leu38Ala
- Leu38Val
- Gly39Ala
- Leu40Ala\*
- Ala43Gly
- Ile45Val
- Ala54Gly

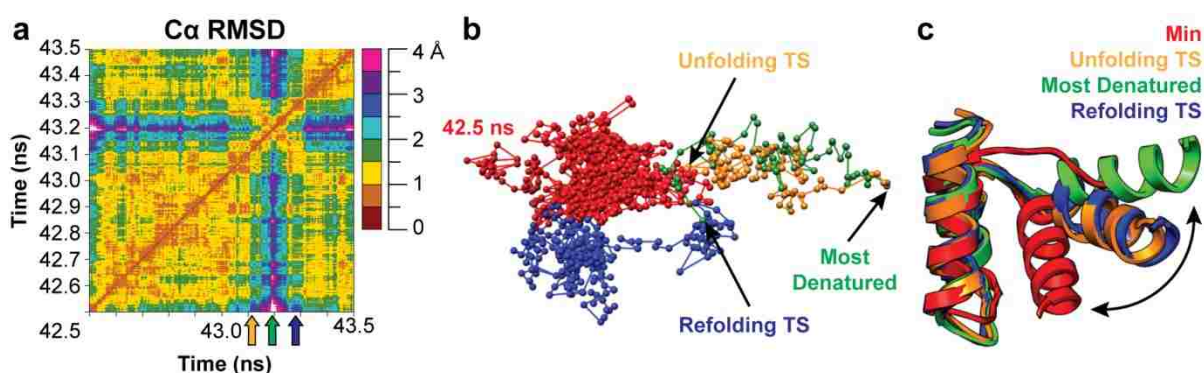


**Figure 2.6: S-values for the 13 new TSEs**

The correlation between calculated S-values and experimentally determined  $\Phi$ -values is quite good for the first 10 TSEs. The S- and  $\Phi$ -values do not agree as well for the 330K/2 44-47ns or 373K/1 TSEs, which is due to loss of secondary structure or altered packing. For the 3 residues noted by an asterisk (\*), only  $S_3$  was reported (see Methods). The 13 residues for which  $\Phi$ -values have been determined are shown in sticks.

There was a pattern of gain and loss of HIII-core contacts preceding and following the TSEs across the simulations. There were 6 residues involved in contacts characteristically lost in N': Phe20, Arg24, Leu26, Phe49, Lys52, and Arg53; and aside from

the 4 pairs that lost contact in N', these 6 residues were also involved in 3 more contacts: Phe49 – Phe20, Arg53 – Phe20, and Arg53 – Leu26. Of these 3 pairs, the 53-26 contact was lost early in all 6 simulations at elevated temperature (Figure 2.1c). In the case of the 12 new TSEs in the 323 K and 330 K simulations, the 49-20 contact was lost within 1 ns before or immediately after the exit TS, and it was reformed within 0.1 ns following the reentry TSs. The 53-20 contact was also lost during in the same time (with one exception: 323K/3 10 ns), and it was regained within 1 ns of all 4 reentry TSs in the 323 K simulations, but it never reformed after the first reentry TS in the 330K/1 simulation. In the 373K/1 simulation the 49-20 pair was lost 0.5 ns after the exit TS and regained 0.1 ns after the reentry TS, and the 53-20 pair was lost 1.5 ns before the exit TS and regained 1 ns after reentry.

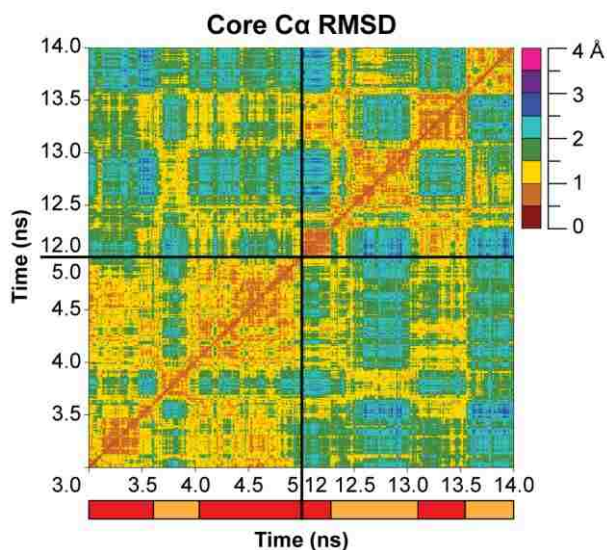


**Figure 2.7: Comparison of 323K/2 43 ns unfolding and refolding TSEs and pathway**

(a) An all-vs.-all Cα RMSD matrix showing the unfolding TS (43.112 ns, orange arrow), most denatured state (43.195 ns, green arrow), and the refolding TS (43.269 ns, blue arrow). The visible “X” is evidence that the conformations the protein takes as it leaves N' to go to D are the same as those it takes when it returns to D, but in the reverse order. (b) The 3D MDS projection of the matrix from (a) in which each of the points represents a structure, and the distance between any two points is proportional to the Cα RMSD between the respective structures. The colors denote different periods in time: 42.5 ns to the unfolding TS (red), unfolding TS to the most denatured conformation (orange), most denatured to the refolding TS (green), and refolding to 43.5 ns (blue). That the paths that the protein followed as it moved from N' to D and back are overlaid in the 3D projection indicates the conformations the protein took were very similar and in reverse order. (c) Structures of the unfolding TS (orange), refolding TS (blue), most denatured conformation (green), and the starting minimized structure (red). The structures were fit based on the Cα atoms of HI-HII. The 2 TS structures are nearly identical, while they are distinct from both N and D.

S-values, which quantify local structure in TSEs, were calculated for the 13 TSEs and compared with experimentally determined  $\Phi$ -values. The S- and  $\Phi$ -values agree very well for 10 of the 13 new TSEs, with linear correlation coefficients ranging from  $R = 0.71$ - $0.86$  (Figure 2.6). The two 330K/2 44-47 ns TSEs did not agree well with experiment (S vs.  $\Phi$ ,  $R = 0.10$  and  $0.03$ ). Loss of secondary structure in the termini of HIII explains some of the disagreement. For example, Ala43 from the N-terminal end of HIII completely lost its

secondary structure, giving it an S-value of 0 rather than its  $\Phi$ -value of 1.05. The 373K/1 5 ns reentry TS had a slightly lower correlation coefficient ( $R = 0.60$ ), with the largest disagreements seen for residues Leu26, Leu38, and Gly39 which are located at the ends of HIII. The secondary structure was as expected for these residues (turn, helix, helix; respectively), so the discrepancy is due to altered packing.



**Figure 2.8: All-vs.-all core C $\alpha$  RMSD matrix of 2 transient N $\rightarrow$ N' $\rightarrow$ N transitions from 323K/3**

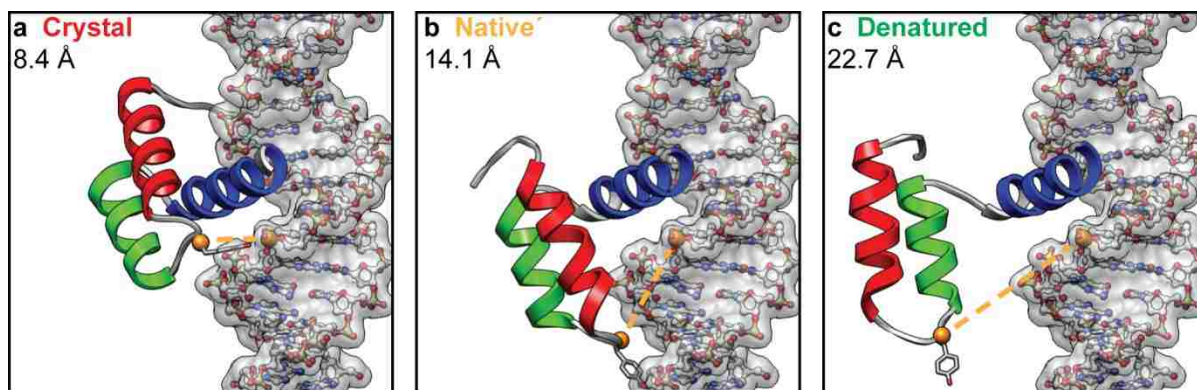
Much like for the N' $\rightarrow$ D $\rightarrow$ N' transition in the 323K/2 43 ns TS, an “X” is visible on the core C $\alpha$  RMSD matrix for both N $\rightarrow$ N' $\rightarrow$ N transitions. In this case, there is a third “X” that is apparent off the diagonal at the intersection of the times at which the other two “X”s occur. This suggests that not only are the N $\leftrightarrow$ N' paths very similar, but also the N $\rightarrow$ N' $\rightarrow$ N path from the first transition is remarkably similar to that from the second.

#### 2.4.4 Protein – DNA Interactions

When EnHD can no longer bind DNA, it loses its biological activity and is therefore denatured, although it may not be unfolded. Consequently, it is informative to determine whether the structures we identified as N, N', and D can bind DNA. In an effort to quantify EnHD's DNA binding ability, we fit EnHD to DNA based on the C $\alpha$  atoms of the major binding helix (HIII) and assessed to what extent the other binding interactions could be formed. Tyr25 is not in HIII or the unstructured N-terminus, and it makes a hydrogen bond to the DNA backbone (Kissinger *et al.* 1990; Fraenkel *et al.* 1998). Since we forced HIII to bind, this residue was selected for distance calculations. Across the simulations, the distances for Tyr25 – Thymine28 show a general trend of being longer for D than N, but the distance is also longer in N' than N. Structures from the 323K/2 simulation and the Tyr25 –



Thymine28 distance are given in Figure 2.9 and are representative of what we saw in the other simulations.



**Figure 2.9: Structures from 323K/2 fit to DNA**

EnHD colored with Tyr25 in sticks. EnHD's Tyr25 C $\alpha$  and the DNA's Thymine28 phosphate are shown in orange spheres, and the distance between these two atoms is given and marked with an orange dashed line. (a) The crystal structure of EnHD bound to DNA (PDB ID: 3hdd) (b) N' at 30 ns in the 323K/2 simulation. (c) The most denatured state (D) from the 43 ns N'→D→N' transition, 43.195 ns. The structures show that EnHD moved slightly away from the DNA in N' compared to N, and much farther in D.

#### 2.4.5 Unfolding and Refolding Pathways

There were 7 unfolding and refolding events identified in 4 simulations during the following times: 39 and 43 ns in 323K/2, 10 and 14 ns in 323K/3, 27-28 and 44-47 ns in 330K/2, and 1-5 ns in 373K/1. Comparison of the structures indicates that the protein moved through the same conformations when it refolded as when it unfolded for a given unfolding/refolding event. This path retracing can be seen as an “X” on the core C $\alpha$  RMSD matrix and as overlaid paths in the 3D projection of the matrix (Figure 2.7). As the protein moved from its most denatured point back to N', a line perpendicular to the diagonal of the matrix is apparent. This line represents a series of conformations that is very similar to the series of conformations the protein passed through just previously, but in reverse order. In the 3D projection, this same phenomenon is seen as overlapping points along the path from N' to the most denatured point, then back to N'. This evidence was present to different extents for each of the 7 unfolding/refolding events, but it is the most striking for 323K/2 39 and 43 ns, 323K/3 10 ns, 330K/2 27-28 ns, and 373K 1-5 ns.

While N' is not in the denatured ensemble, it is distinct from N, and thus there must exist a low-energy pathway to move between the two states. As with the unfolding and refolding pathways, there was an “X” on the core C $\alpha$  RMSD matrix when EnHD transiently

moved from N' to N and back. In the 323K/3 simulation, there were 2 transient N→N'→N movements. There was an “X” visible on the core C $\alpha$  RMSD matrix around both N→N'→N transitions, and there was a third “X” off the diagonal, around the intersection of the times corresponding to both of the individual N→N'→N transitions (Figure 2.8). This third “X” indicates that not only were the N→N' and N'→N pathways very similar for a single transition, but also that the two N'→N→N' transitions were almost equally as similar.

## 2.5 Discussion

Three different conformational states were populated by EnHD in our 7 simulations: N, N', and D. The protein's state was determined based on a combination of measurements: core C $\alpha$  RMSD, HIII-core distance, and HIII-core contact pattern (Figure 2.1). Measuring the core C $\alpha$  RMSD to the minimized crystal structure was the foremost method in determining EnHD's state. The core C $\alpha$  RMSD generally fluctuated between 1-3 Å when the protein was in N and 2-4 Å for N' (Figure 2.1a, 2.5). Values over 4.5 Å usually indicated a departure from N' to D. The Phe20 C $\alpha$  – Lys52 C distance, representative of the distance between HIII and the HI-HII scaffold, was also a good indicator of state. When the protein was in N, the distance fluctuated around  $10 \pm 2$  Å, while an increase to  $14 \pm 2$  Å indicated movement to N' (Figure 2.1b). Distances over 20 Å occurred when EnHD moved to D. Movement between states was more clearly discerned based on the HIII-core distance than on core C $\alpha$  RMSD. There was nearly always a clean jump in distance between the different states, which suggests that N and N' are distinct states despite both being “native.” The pattern of contacts between HIII and the rest of the protein also helped discriminate different states. The contact pairs selected for analysis were deliberately chosen to be good representatives of N. Jumps in core C $\alpha$  RMSD to 2-4 Å and HIII-core distance to  $\sim 14$  Å, which were characteristic of N', coincided with loss of 4 contacts pairs: Phe49 – Arg24, Phe49 – Leu26, Lys52 – Phe20, and Arg53 – Arg24 (Figure 2.1c).

To estimate the likelihood that the 3 different EnHD conformations binds DNA, we fit structures from our MD simulations onto the DNA-bound crystal structure and took a distance measurement that might discern between native (N or N') and D. Representative structures from the 323K/2 simulation and distances are given in Figure 2.9. Even though

both the N' and D structures were distinct from the crystal structure, N' conformations could more easily move back to N and be in a position to bind the DNA (and thus be biologically active) than those in D. For N'→N movement to occur, HI and HII would have to slide along the DNA and HIII so that they could dock against HIII as in N. This movement was what we saw for all N→N' transitions in our simulations without DNA. For a D structure to move to N, it would have to expel the water from its solvated hydrophobic core before HI and HII could dock back onto HIII and the DNA. Overall, D appears to be too distorted to function properly, and N' falls somewhere between N and D. While N' may be able to recover and clamp down on the DNA, it is also possible that it is too distorted and therefore inactive.

Having defined three different states for EnHD in 7 independent simulations, it is interesting to consider the variations within a state and how EnHD passes between them. N' was observed in all 6 of the elevated temperature simulations, but it was more similar in 5 of them (323K/1-3, 330K/2, and 373K/1) than it was in 330K/1 based on core C $\alpha$  RMSD and HIII-core contacts (Figure 2.4a, 2.5a). Also, the first period of N' in the 373K/1 simulation matched the first 4 simulations, while the second period was ambiguous. In all cases, HIII moved a register towards the N-terminus, away from the HI-HII turn, but N' in 330K/1 and the end of 373K/1 was somewhat different from N' in the other 4 simulations and in the beginning of 373K/1, despite having the same overall topology. This difference suggests that there may be multiple, subtly different N' states.

In the 4 simulations where there was N'↔D movement, a total of 13 new TSEs were identified. The TSEs within one temperature were most similar, and the best agreement was between the unfolding and refolding TSEs for a single transient unfolding/refolding event (Figure 2.5b). In these cases, only a small portion of D was sampled so it was likely that the protein would find a refolding path very similar to its unfolding path from the ensemble of paths available.

Based on the 11 HIII-core contact pairs selected for analysis, there was good agreement among all 17 TSEs (13 new and 4 previously identified; Figure 2.1c, 2.4b). In almost all of the cases where there was disagreement between the new and previously identified TSEs, it was the case that there were more contacts present in the high-temperature TSEs, which suggests they were more native-like. It is expected that high-temperature TSEs

will more closely resemble N than TSEs at the protein's  $T_m$  due to Hammond effects. This phenomenon causes the structure of the TSE to become more native-like upon destabilization, in this case by heat (Matthews and Fersht 1995; Daggett *et al.* 1998; Day *et al.* 2002).

Not only were the 13 new TSEs consistent with those previously identified, but 10 of them were in good agreement with experiment based on comparing calculated S-values to experimental  $\Phi$ -values ( $R = 0.71-0.86$ ; Figure 2.6). The correlation was somewhat lower ( $R = 0.60$ ) for the 373K/1 refolding TSE and significantly lower ( $R = 0.10$  and  $0.03$ ) for the 330K/2 44-47 ns TSEs. The 330K/2 44-47 ns unfolding/refolding event followed a cyclical path as it folded and unfolded, yet its unfolding and refolding TSEs were very similar to each other (average structure core  $C\alpha$  RMSD = 1.77 Å; Figure 2.5b). The lack of agreement in this case may be illustrating discrepancies between single-molecule behavior versus bulk measurements. That is, our aberrant TSE pair may be extreme members of the much broader ensemble probed experimentally.

The 13 new TSEs generally agreed well with each other, the 4 previously identified TSEs from high-temperature unfolding simulations, and experiment. The unfolding and refolding TSEs were equally similar, which is evidence that all 17 TSEs come from the same global ensemble of transition states for EnHD folding. Further, our data suggest that the ensemble of paths for unfolding and refolding is also the same, which supports our long-standing contention that protein folding is a microscopically reversible process.

The symmetrical order of contact pair loss and gain upon unfolding and refolding is further evidence that protein folding is microscopically reversible. The 6 residues involved in the 4 contact pairs that were characteristically lost in N' made 3 additional contacts (Figure 2.1c, 2.4). One pair was lost early in the simulations, and the other two, Phe49 – Phe20 and Arg53 – Phe20, were usually lost just before the unfolding TS and regained right after the refolding TS in the 7 unfolding/refolding events. These 6 residues make up the half of the HIII-core contacts closest to the C-terminus (Figure 2.4c). Arg24, Leu26, and Lys52 lost all of their HIII-core contacts in N', but Phe20, Phe49, and Arg53 maintained 2 of the original 7 contact pairs. It was not until right after these 2 contact pairs were lost that EnHD reached its TS and became denatured, and they reformed right after reentering N' from D in most cases.

Thus, loss and gain of these 6 hydrophobic core contacts are critical steps on the  $N \leftrightarrow N'$  and  $N' \leftrightarrow D$  pathways, based on our 323 and 330 K simulations. This is evidence for microscopic reversibility in protein folding because there is a consistent order of loss of contacts in unfolding that is repeated in reverse order upon refolding. However, the pattern of loss by these 6 contact pairs was not consistently repeated in the 4 high-temperature unfolding simulations previously run at 373 and 498 K (data not shown).

The all-vs.-all core  $C\alpha$  RMSD matrix and its 3D projection are arguably the best ways to observe the similarity of MD structures over time. Indeed, there was a visible “X” on the core  $C\alpha$  RMSD matrix for 5 of the 7 unfolding/refolding events. Similarly, the structures from the unfolding TS to the most denatured point back to the refolding TS overlay on the 3D projection of the core  $C\alpha$  RMSD matrix for these 5 unfolding/refolding events (Figure 2.7). Both the “X” and the overlaid paths indicate that the conformations EnHD moved through from the unfolding TS to the most denatured point had low core  $C\alpha$  RMSD to the conformations EnHD took as it moved back to the refolding TS, but in reverse order.

“X”s were also visible for transient movements from N to N'. In 323K/3, not only was the  $N \rightarrow N'$  path similar to the  $N' \rightarrow N$  path, but both  $N \rightarrow N' \rightarrow N$  movements followed highly similar paths. There is an “X” on the diagonal of the all-vs.-all core  $C\alpha$  RMSD matrix for each  $N \rightarrow N' \rightarrow N$  movement, and there is also an off-diagonal “X” at the intersection of the times corresponding to each of the  $N \rightarrow N' \rightarrow N$  events. While these four  $N \leftrightarrow N'$  pathways are not complete folding or unfolding pathways, they are transitions between discrete states along the full folding pathway.

EnHD never moved from N to D without first passing through N' each of the 7 times it unfolded. In fact, in the 323K/2 and 323K/3 simulations, it moved back to N between the two unfolding/refolding events, passing through N' on the way. Based on our simulations, the  $N \leftrightarrow N'$  paths were part of the same ensemble as were the  $N' \leftrightarrow D$  paths, and N' is a necessary step between N and D. Together, these findings are evidence that the entire  $N \rightarrow N' \rightarrow D$  and  $D \rightarrow N' \rightarrow N$  pathways are mirror images of the same process, and thus protein folding is a microscopically reversible process.

We identified and characterized four distinct states of EnHD: N, N', TS, and D which were consistent across simulations at 298, 323, 330, and 373 K. Core C $\alpha$  RMSD, HIII-core contacts, HIII-core distance, and predicted DNA binding ability were used to discriminate between the states and place them on the folding pathway. We identified 7 transient denaturing events in 6 simulations and identified 13 new unfolding and refolding TSEs. The 13 new TSEs agreed well with 4 previously identified TSEs based on core C $\alpha$  RMSD and HIII-core contacts as well as with experimental data based on  $\Phi$ - and S-values. In 5 of the 7 transient denaturing events, the unfolding pathway was nearly identical to the refolding pathway. We also found two N $\leftrightarrow$ N' transitions that followed the same pathway in the folding and unfolding directions for both transitions. These phenomena are evidence that the ensemble of folding and unfolding pathways is one and the same and that protein folding can be a microscopically reversible process.

## Chapter 3: Refolding the Engrailed Homeodomain – Structural Basis for the Accumulation of a Folding Intermediate

### 3.1 Summary

The ultrafast folding pathway of the Engrailed Homeodomain has been exceptionally well-characterized by experiment and simulation. Helices II and III of the 3-helix bundle protein form the native helix-turn-helix motif as an on-pathway intermediate in a few  $\mu\text{s}$ . The slow step is then the proper docking of the helices in  $\sim 15 \mu\text{s}$ . But, there is still the unexplained puzzle of why helix docking is relatively slow, which is part of a more general problem of why rearrangements of intermediates can be slow. To address that problem, we performed 46 all-atom molecular dynamics refolding simulations in explicit water totaling  $15 \mu\text{s}$  of simulation time. The simulations started from an intermediate state structure that was generated in an unfolding simulation at 498 K and was then “quenched” to folding-permissive temperatures. The protein refolded successfully in only one of the 46 simulations, and in that case, the refolding pathway mirrored the unfolding pathway at high temperature. In the 45 simulations in which the protein did not fully fold, nonnative salt bridges trapped the protein, which explains why the protein folds relatively slowly from the intermediate state.

### 3.2 Introduction

The homeodomain superfamily has been especially interesting in the development of theories and the application of methods to protein folding. The engrailed homeodomain (EnHD) is a 3-helical bundle protein (helices HI, HII, and HIII), and its native state is barely stable with  $\Delta G_{\text{D-N}} = 2.5 \text{ kcal/mol}$  (Mayor *et al.* 2003a; Mayor *et al.* 2003b). It folds via a proven on-pathway folding intermediate (Mayor *et al.* 2003a; Mayor *et al.* 2003b; Religa *et al.* 2005) that is formed at  $\sim 3 \times 10^5 \text{ s}^{-1}$  and rearranges at  $\sim 5 \times 10^4 \text{ s}^{-1}$  at  $42 \text{ }^\circ\text{C}$  ( $t_{1/2} = 15 \mu\text{s}$  at  $25 \text{ }^\circ\text{C}$ ). Those rate constants were, at the time, the fastest observed for a protein (Mayor *et al.* 2000; Mayor *et al.* 2003b), which made it a prime target for molecular dynamics (MD)

simulation (Mayor *et al.* 2000; Mayor *et al.* 2003b; DeMarco *et al.* 2004; Gianni *et al.* 2003). Further, the folding pathway can be blocked at the rearrangement step by protein engineering so that the intermediate is stable under “physiological conditions” and its structure was solved by NMR (Religa *et al.* 2005). It contains the HII-turn-HIII motif in the native structure, but the helices aren’t docked. The motif is, in fact, an independently folding domain (Religa *et al.* 2007).

MD simulation predicted the complete description of the folding pathway, in reverse by simulating unfolding, and was later benchmarked and validated by experiment simulation (Mayor *et al.* 2000; Mayor *et al.* 2003b; DeMarco *et al.* 2004; Gianni *et al.* 2003). For example, the high temperature simulation used here as a starting point for the current study is in excellent agreement with experiment: the MD-predicted transition state (TS) is in quantitative agreement with experiment, and the MD predictions (Mayor *et al.* 2000) were published three years before the experimental work (Mayor *et al.* 2003b; Gianni *et al.* 2003). The MD-predicted intermediate is also in agreement with experiment (Mayor *et al.* 2000; DeMarco *et al.* 2004) and the MD-generated structure was confirmed through direct NMR experiments (Religa *et al.* 2005) five years after the prediction. Coarse-grain refolding models and atomic-level Monte Carlo methods have also reproduced structures on EnHD’s folding pathway (Zhang *et al.* 2005; Hubner *et al.* 2006a; Li *et al.* 2008). Thus, this is a well-characterized system for protein folding studies.

Nevertheless, there remains an outstanding problem: why, when a folded helix-turn-motif is formed in a few  $\mu$ s, does simple docking of the helices take 15  $\mu$ s? This is part of a general problem that the major structural parts of a protein can be formed rapidly but the final formation of native structure is slow (Englander *et al.* 2007). To address this problem for EnHD, we have conducted MD simulations of “refolding,” starting from the folding intermediate, which was generated at high temperature and then quenched to folding permissive conditions by lowering the temperature. Here, we report 46 independent MD quench simulations in explicit water at 310, 314, and 319 K (37, 41, and 46 °C) totaling nearly 15  $\mu$ s of simulation time completed over ~8 years of computer time. The starting structure was from the experimentally verified, 498 K unfolding simulation discussed above (Mayor *et al.* 2000; Mayor *et al.* 2003b; DeMarco *et al.* 2004; Figure 3.1). Experimentally,



the intermediate state is the denatured state under folding conditions; that is, the intermediate is the starting point for folding (Mayor *et al.* 2000; Mayor *et al.* 2003a; Mayor *et al.* 2003b; Religa *et al.* 2005). Consequently, we chose a snapshot from the thermal unfolding simulations corresponding to the native end of the intermediate ensemble as the starting structure for multiple, independent quench simulations.

The use of a large number of parallel simulations is a convenient method for studying first-order reactions since such processes are stochastic and a small number  $N$  of the total processes  $N_T$  will have gone to completion in a short period of time ( $\delta t$ ):  $N/N_T = \delta t / t_{1/2}$  (Fersht 2002). For this reason we performed 46 simulations with an average simulation time of 326 ns, which yields  $\delta t/t_{1/2} = 326 \text{ ns} / 15,000 \text{ ns} = 0.0217$ , given the experimental  $t_{1/2}$  for  $I \rightarrow N$ . So, from this simple analysis we would expect only one of the simulations to refold. The problem is that short simulations might become trapped in on- or off-pathway events by formation of transiently stable structures.

Determining whether a protein has refolded in simulation is another important, though less appreciated, challenge. Many individual properties or pairs of properties have been used in the literature to determine whether a protein has refolded, including radius of gyration,  $C\alpha$  root mean squared deviation (RMSD), native contacts, and solvent accessible surface area (SASA). These properties, when considered individually, are insufficient to prove the protein has reached the native state. For example, proteins may achieve a native-like  $C\alpha$  RMSD and radius of gyration, yet make few native contacts. However, many properties in combination can provide satisfying proof that a protein has refolded.

Here, we found that 45 of the 46 simulations explored off-pathway events, and one folded successfully. Our previous unfolding simulations capture the reverse of productive folding pathways, as described above, and the productive refolding simulation presented here mirrors the previously simulated unfolding pathway. In addition the single simulated refolding process observed here is consistent with our earlier study where microscopic reversibility was observed directly for EnHD at its  $T_m$  (McCully *et al.* 2008). Furthermore, the unproductive refolding simulations have now captured off-pathway events that explain why the docking reaction is slowed down.

### 3.3 Methods

#### 3.3.1 Simulation Protocol

The starting structure for all of the quench simulations came from a thermal denaturation simulation at 498 K, which was previously verified against experimental data (Mayor *et al.* 2000; Mayor *et al.* 2003b; DeMarco *et al.* 2004; Gianni *et al.* 2003). The 5 ns structure was strategically chosen from the native end of the intermediate state ensemble, as it was poised for refolding, and therefore did not require computational resources to simulate a portion of the folding pathway that was not directly of interest. We wanted to begin from the same state as experimental folding studies, and the intermediate is the denatured state under folding conditions. The native simulations began from the crystal structure of EnHD (PDB ID: 1enh; Clarke *et al.* 1994). Our in-house MD software, *in lucem* molecular mechanics (*ilmm*; Beck *et al.* 2002-2012), with the all-atom Levitt *et al.* (1995) force field, was used for all simulations. The starting structure was minimized for 150 steps for the MD-derived starting structure and 1000 for the crystal structure. The minimized structure was solvated in a box of F3C water molecules (Levitt *et al.* 1997) with 8-10 Å of padding between the protein and edge of the periodic box, and the water density was set based on the simulation temperature according to the experimentally determined liquid-vapor coexistence curve; 298 K: 0.997 g/ml, 310: 0.993, 314: 0.992, 319: 0.990 (Kell 1967). Standard protocols were used to complete preparation of the system and for the production run (Beck and Daggett 2004). The NVE microcanonical ensemble was employed with 2 fs timesteps and structures written out every 1 ps. Nonbonded interactions were truncated at 8 Å with a force-shifted cutoff (Beck *et al.* 2005), and the nonbonded list was updated every two steps.

The fastest experimental folding rate constant for EnHD was measured at  $5.1 \times 10^4 \text{ s}^{-1}$  around 315 K (42 °C; Mayor *et al.* 2000). Accordingly, we chose to run our quench simulations at 310, 314, and 319 K. Forty-six simulations were run at these three temperatures for varying lengths of time, resulting in a total simulation time of 14.996 μs. The average simulation time was 326 ns, and the longest was 793 ns (Table 3.1, 3.2). We employed various computing clusters including Intel, AMD, and Power5 architectures. The total wall-clock time for the quench simulations was nearly 8 years.

**Table 3.1: Simulation properties for native simulations**

Temperature (K)	Run	Total Time (ns)	Reference Set*	% NOEs Satisfied †
298	1	80	all	83.0
	2	70	all	82.6
	3	50	all	88.8
	4	50	all	87.5
	5	20	all	85.6
	6	20	all	90.7
	7	20	all	88.1
310	1	100	all	82.0
	2	50	all	87.8
	3	100	0-94 ns	87.2
	4	100	0-4.6, 5.3-5.9, 6.2-100 ns	82.0
	5	20	all	88.5
	6	20	all	88.1
314	1	78	all	87.9
	2	79	all	85.9
	3	79	all	87.8
319	1	335	0-37 ns	83.5
	2	334	0-61 ns	87.5
	3	166	0-21 ns	85.8
	4	166	0-64 ns	88.1
	5	166	0-137 ns	87.2
Summary	21	2103	1249 ns	86.4

\* Time during which EnHD occupied the native state. Only the time spans in this column were used when calculating the reference set for the reaction coordinate.

† A total of 654 reported NOEs were used for comparison (Religa 2008). NOEs were calculated over the reference set in column 4. An NOE was considered satisfied if the  $r^{-6}$  weighted distance between protons was  $\leq 5.5$  Å.

### 3.3.2 Analyses

The  $C\alpha$  RMSD was monitored over time for the core residues (8-53) since the N- and C-termini have large fluctuations that are not representative of the general structure and dynamics of the protein. Solvent accessible surface area of the core residues and Trp48, the fluorescence probe of folding, was calculated using our in-house implementation of the Lee and Richards (1971) algorithm, and secondary structure was calculated with our in-house implementation of the DSSP algorithm (Kabsch and Sander 1983).

Eleven contacts between HIII and the HI-HII scaffold were identified previously as key indicators of “foldedness” (McCully *et al.* 2008), and an additional 5 contacts between

HI and HII were also selected. Residues were considered in contact if they contained atoms that met at least one of the following criteria: (1) carbon – carbon distance  $< 5.4 \text{ \AA}$ , (2) hydrogen bond acceptor – hydrogen distance  $< 2.6 \text{ \AA}$  and donor – hydrogen – acceptor angle within  $45^\circ$  of linearity, or (3) heavy atom – heavy atom distance  $< 4.6 \text{ \AA}$  for atoms that do not satisfy (1) or (2). The distance between the COM of each residue was also calculated for all 16 pairs. Contacts were monitored over all residues in the protein and categorized based on whether they were present in the crystal structure (native, otherwise nonnative) and whether the contact pair consisted of main chain atoms, side chain atoms, or both. Contact lifetimes were also calculated at 1-ps granularity. Carbon atoms were classified as nonpolar, and all other atoms were considered polar. Two residues were considered in contact if the distance between a non-hydrogen atom from each residue fell below a cutoff of  $5.4 \text{ \AA}$  for carbon – carbon pairs and  $4.6 \text{ \AA}$  for all other pairs.

**Table 3.2: Simulation properties for quench simulations**

Temperature (K)	Number of Simulations	Total Time ( $\mu\text{s}$ )
310	7	4.682
314	9	4.743
319	30	5.495
Total	46	14.920

Experimental NOE values were obtained from the Biological Magnetic Resonance Data Bank, entry 15536 (Religa 2008). An NOE was considered satisfied in our simulations if the  $r^{-6}$  weighted distance between closest protons during the simulation was  $\leq 5.5 \text{ \AA}$ , which was the maximum cutoff employed by Religa in building his NMR structure (Religa 2008).

### 3.3.3 Property Space and Reaction Coordinate

A total of 35 physical properties of the protein were selected to create a multidimensional property space and are listed with their average values and standard deviations in Table 3.3. The properties were calculated at 10-ps granularity and were normalized over the simulations being compared. The principal components of this space were calculated for the 498 K denaturing simulation and all of the 298 K native simulations, and each property's contribution to the first principal component is reported in Table 3.3.

**Table 3.3: Property space weights and values of the reference sets**

Property	PC 1				
	Weight*	298 K	310 K	314 K	319 K
Core C $\alpha$ RMSD <sup>†</sup> (Å)	0.99	2.19 ± 0.58	2.37 ± 0.54	2.31 ± 0.69	2.20 ± 0.60
Fraction $\alpha$ -Helix <sup>‡</sup>	0.90	0.72 ± 0.05	0.66 ± 0.07	0.67 ± 0.03	0.68 ± 0.04
COM Distance <sup>§</sup> Arg30 - Glu19 (Å)	0.95	6.99 ± 0.79	6.70 ± 0.61	6.50 ± 0.60	6.80 ± 0.86
COM Distance Leu4 - Arg15 (Å)	0.95	7.49 ± 0.91	7.80 ± 1.47	7.03 ± 0.88	7.27 ± 1.09
COM Distance Leu4 - Leu16 (Å)	0.96	8.37 ± 0.75	8.25 ± 0.94	7.52 ± 0.75	8.01 ± 0.87
COM Distance Glu7 - Arg15 (Å)	0.94	8.16 ± 0.68	8.49 ± 1.24	8.02 ± 0.64	8.47 ± 1.49
COM Distance Leu38 - Gln12 (Å)	0.95	7.17 ± 1.19	7.43 ± 1.27	6.87 ± 1.02	7.14 ± 1.25
COM Distance Ile45 - Leu38 (Å)	0.72	9.03 ± 1.10	9.54 ± 1.13	9.65 ± 1.29	9.54 ± 1.17
COM Distance Ile45 - Leu40 (Å)	0.67	6.74 ± 0.82	6.56 ± 0.62	6.61 ± 0.66	6.64 ± 0.70
COM Distance Trp8 - Leu16 (Å)	0.94	9.84 ± 1.72	10.08 ± 2.07	9.37 ± 2.03	9.15 ± 1.56
COM Distance Phe49 - Leu16 (Å)	0.93	9.41 ± 1.41	9.08 ± 1.62	9.75 ± 2.07	8.96 ± 1.60
COM Distance Phe49 - Phe20 (Å)	0.91	8.32 ± 1.59	7.50 ± 1.23	7.63 ± 1.02	7.38 ± 0.98
COM Distance Phe49 - Arg24 (Å)	0.90	9.70 ± 2.00	10.98 ± 3.01	11.39 ± 2.99	11.25 ± 2.64
COM Distance Phe49 - Leu26 (Å)	0.92	7.96 ± 1.69	8.92 ± 2.16	9.18 ± 2.35	8.87 ± 2.43
COM Distance Lys52 - Phe20 (Å)	0.90	9.55 ± 2.03	10.37 ± 2.65	10.65 ± 2.44	10.22 ± 2.69
COM Distance Arg53 - Phe20 (Å)	0.92	9.14 ± 1.94	8.78 ± 2.08	9.34 ± 2.14	9.19 ± 2.05
COM Distance Arg53 - Arg24 (Å)	0.92	8.33 ± 2.83	10.02 ± 4.11	10.95 ± 3.72	10.31 ± 3.85
COM Distance Arg53 - Leu26 (Å)	0.91	9.78 ± 2.59	12.09 ± 2.64	12.51 ± 2.42	11.92 ± 3.14
Core Main Chain SASA <sup>¶</sup> (Å <sup>2</sup> )	0.93	459 ± 33	462 ± 42	455 ± 44	465 ± 39
Core Main Chain Polar SASA (Å <sup>2</sup> )	0.93	276 ± 27	288 ± 34	282 ± 37	288 ± 35
Core Side Chain SASA (Å <sup>2</sup> )	0.91	3282 ± 184	3255 ± 171	3217 ± 166	3233 ± 183
Core Side Chain Non-Polar SASA (Å <sup>2</sup> )	0.92	1729 ± 152	1710 ± 145	1647 ± 144	1665 ± 131
Core Non-Polar SASA (Å <sup>2</sup> )	0.91	1912 ± 158	1884 ± 151	1821 ± 147	1842 ± 135
Core Polar SASA (Å <sup>2</sup> )	0.88	1829 ± 84	1833 ± 89	1852 ± 102	1856 ± 125
Core Total SASA (Å <sup>2</sup> )	0.95	3741 ± 191	3718 ± 187	3672 ± 176	3697 ± 200
Trp48 SASA (Å <sup>2</sup> )	0.60	95 ± 32	98 ± 36	77 ± 38	76 ± 38
Native MC-MC Contacts <sup>  </sup>	0.96	122.4 ± 2.1	120.9 ± 2.9	121.0 ± 2.5	121.4 ± 2.7
Native MC-SC Chain Contacts	0.96	117.1 ± 5.8	111.7 ± 7.9	112.6 ± 7.5	113.2 ± 8.4
Native SC-SC Contacts	0.93	63.1 ± 6.4	59.4 ± 6.6	60.3 ± 7.6	60.4 ± 7.0
Total Native Contacts	0.97	152.7 ± 7.0	149.7 ± 7.4	151.0 ± 8.4	151.5 ± 7.7
Nonnative MC-MC Contacts	0.85	4.1 ± 2.4	7.4 ± 3.9	6.7 ± 3.4	6.3 ± 3.0
Nonnative MC-SC Chain Contacts	0.87	4.9 ± 3.0	10.8 ± 5.6	11.1 ± 6.4	9.6 ± 5.0
Nonnative SC-SC Contacts	0.77	6.8 ± 3.3	12.6 ± 6.2	12.5 ± 5.7	12.2 ± 5.8
Total Nonnative Contacts	0.83	12.0 ± 4.6	20.2 ± 8.4	19.8 ± 8.3	18.6 ± 7.1
Total Contacts	0.90	164.8 ± 6.6	169.9 ± 7.3	170.7 ± 6.2	170.1 ± 6.3

\* Weights are reported for the first principal component of the property space of a previously validated 498 K unfolding trajectory and the 298 K reference set.

<sup>†</sup> RMSD was calculated over the C $\alpha$  atoms of residues 8-53. All properties are given as the average  $\pm$  1 s.d.

<sup>‡</sup> DSSP (Kabsch and Sander 1983) was used to determine what fraction of the 54 residues was in  $\alpha$ -helix.

<sup>§</sup> Center of mass was calculated over all non-hydrogen atoms for both residues listed, and the distance between the two centers of mass is reported.

<sup>¶</sup> SASA was calculated as specified over residues 8-53 or just Trp48 using the Lee and Richards (1971) algorithm.

<sup>||</sup> Two residues were considered to be in contact if the distance between at least one heavy atom from each was less than 5.4 Å for carbon/carbon pairs and 4.6 Å for all other pairs.

The distance in property space between any two time points was calculated as the average Euclidean distance between the 35-dimensional points. The average distance in property space between one point and a set of points was calculated as the sum of the distance between the point of interest and each point in the set divided by the number of points in the set. For a detailed explanation, see Beck and Daggett (2007), in particular

Equations 1 and 2, as well as the original paper introducing property space analysis and PCA of MD trajectories by Kazmirski *et al.* (1999).

We created a one-dimensional reaction coordinate by calculating the mean distance in property space from each time point in the simulation of interest to a reference set as described above and in more detail by Toofanny *et al.* (2010). The reference set for a given temperature was made up of all time points from the native simulations at that temperature during which EnHD was in the native state.

### 3.3.4 Transition State Selection

The mean distance to the 298 K reference set in property space was used to select TS ensembles in the native simulations. The TS was defined as the final time point that fell below a cutoff defined based on the simulation's distribution on the reaction coordinate. This method is a variation on the method described by Toofanny *et al.* (2010).

Three-dimensional projection using multidimensional scaling (MDS) of the all-against-all C $\alpha$  RMSD matrix for the quench simulation was performed using the statistical package, R (Team 1925). TS ensembles were selected as the native cluster exit and preceding 5 ps for unfolding and as the native cluster entry and subsequent 5 ps for refolding, as described previously (Li and Daggett 1994; Li and Daggett 1996; McCully *et al.* 2008).

The S-value, a residue-based measure of structure that is comparable to experimental  $\Phi$ -values, is a product of the extent of native secondary structure ( $S_{2^\circ}$ ) and native and nonnative tertiary contacts ( $S_{3^\circ}$ ) in a given residue in the TS ensemble relative to the crystal structure (Daggett *et al.* 1996).  $S_{3^\circ}$  only was reported for residues Phe8, Leu26, and Leu40, as described previously (Gianni *et al.* 2003).

## 3.4 Results

We first discuss simulations of EnHD at 298-319 K and the stability of the native state. Next, we validate a 35-dimensional property space and use it to identify states in a high-temperature unfolding simulation. Finally, we discuss our refolding, or “quench,” simulations, which were compared with the native simulations using our property space. One of the 46 quench simulations was found to successfully refold, and that simulation is

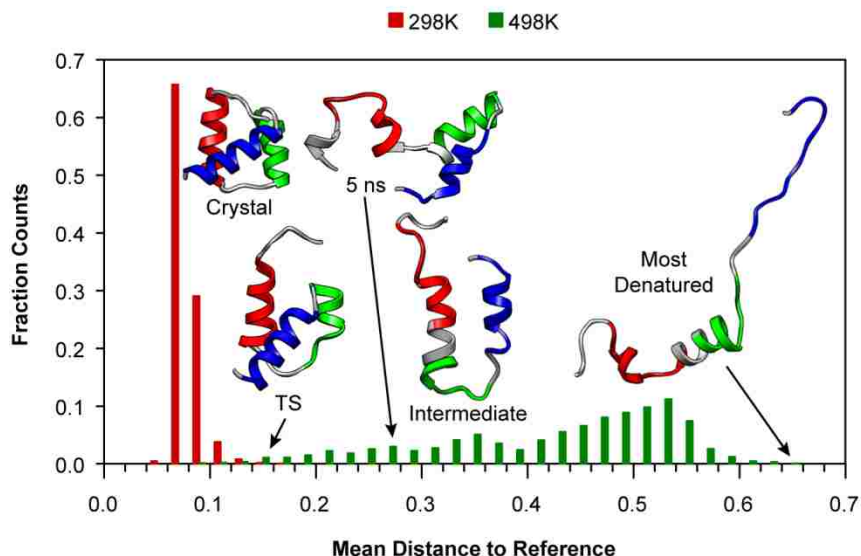
described in detail. Additionally, we identify the interactions inhibiting the other 45 quench simulations from refolding to the native state.

### 3.4.1 Native Simulations

We constructed a reference set for native EnHD from the native simulations at 298, 310, 314, and 319 K (Table 3.1, 3.2). Note, however, that many of these “native” simulations are intentionally at elevated temperature where both the folding and unfolding rates are high. Overall, EnHD was stable in the native state with a core (residues 8-53)  $C\alpha$  root-mean-square deviation (RMSD) of  $2.27 \pm 0.60$  Å (Table 3.3) and on average  $86 \pm 2\%$  of the Nuclear Overhauser effect crosspeaks (NOEs) were satisfied (Table 3.1). The residues that consistently had the most violations were Phe20 (HI) and Leu26 (HI-HII loop), which both pack into the hydrophobic core. When EnHD is in the nearly native (N') state (McCully *et al.* 2008), HIII translates toward the N-terminus, which breaks about 75% of the native contacts made by Phe20 and Leu26. Since N' involves reorientation of HIII along HI and HII without exposing the hydrophobic core and it becomes more prevalent as the temperature rises (298 K, 6% population; 310, 28%; 314, 21%; 319, 33%), and since the NOEs were measured at 278 K, we expect such NOE violations.

### 3.4.2 Property Space Reaction Coordinate

A multidimensional-embedded one-dimensional reaction coordinate based on physical properties of the protein was calculated for the high temperature unfolding simulation using the 298 K native simulations as reference (Figure 3.1, and see Methods for details). The reaction coordinate for the high-temperature unfolding showed that the transition state (TS) selected previously (Mayor *et al.* 2000; Mayor *et al.* 2003b; DeMarco *et al.* 2004; Gianni *et al.* 2003) fell at 0.16 along the reaction coordinate, just outside the reference distribution (Figure 3.1), as would be expected for this method. The intermediate and denatured populations formed broad, connected peaks in the 498 K distribution. The intermediate spanned a mean distance to reference of  $\sim 0.24$ - $0.40$ , and the starting structure for the quench simulations was at 0.27. The starting structure contains the HII-HIII helix-turn-helix motif (Figure 3.1), and it is 10.5 Å  $C\alpha$  RMSD from the crystal structure or 8.0 Å over core residues.



**Figure 3.1: Reaction coordinate for high temperature denaturation simulation**

The mean Euclidean distance in our 35-dimensional property space was calculated to the reference set for each structure in the 298 native set and 498 K unfolding simulation, and a histogram of those distances is shown here. The reference set was each structure in the 298 K native simulation, so more native-like structures have a lower mean distance to reference. The crystal structure is shown to represent the 298 K reference set, and the following structures are shown with their mean distance to reference in 35-dimensional property space: TS (0.16); 5 ns structure, which is the starting structure for the quench simulations (0.27); folding intermediate ensemble (~0.24-0.40); and most denatured structure (0.66). EnHD is colored by helix: HI residues 10-22, HII 28-38, and HIII 42-55.

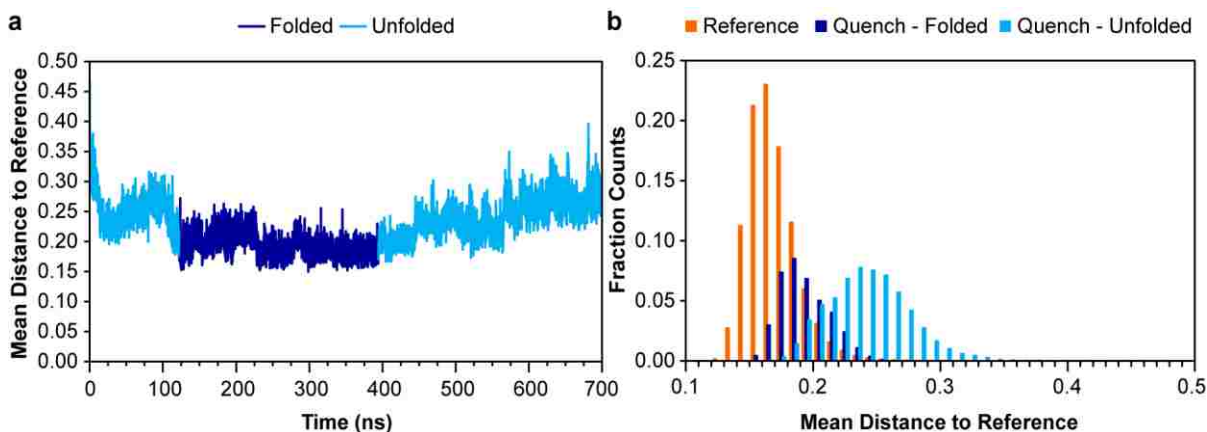
Principal component analysis (PCA) was run on this dataset to determine which properties contributed most to the reaction coordinate, and the resulting weights are listed in Table 3.3. Nearly all of the properties had weights above 0.90, and the highest weights were observed for core C $\alpha$  RMSD (0.99) and total native contacts (0.97). Center of mass distances between the five residue pairs between HI and HII all had very high weights, as did the various types of native contacts. The lowest weights were observed for the Trp48 solvent accessible surface area (SASA) and COM distances of residues in or near the HII-HIII turn.

### 3.4.3 Quench Simulations: Successful Refolding

Of the 46 quench simulations performed, one of them refolded based on property space analysis. This simulation was run at 319 K for 698 ns. A reaction coordinate was calculated using the native portions of the 319 K native state simulations as the reference set (Figure 3.2, with the native state at  $0.169 \pm 0.019$  along this 319 K reaction coordinate). Boundaries for the transition state ensembles were chosen at 122.210 ns (+ 5 ps) for refolding and at 394.596 ns (- 5 ps) for the subsequent unfolding. The refolding and unfolding TS



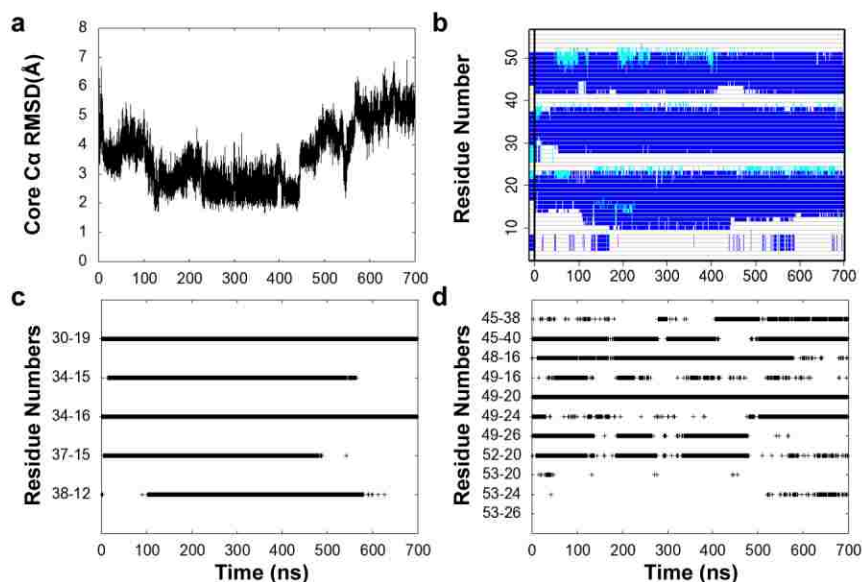
ensembles were more native-like than expected (0.194 and 0.170 mean distance to reference for refolding and unfolding, respectively, Figure 3.2a) which probably resulted from the broad native state population at 319 K.



**Figure 3.2: Reaction coordinate for the successful quench simulation**

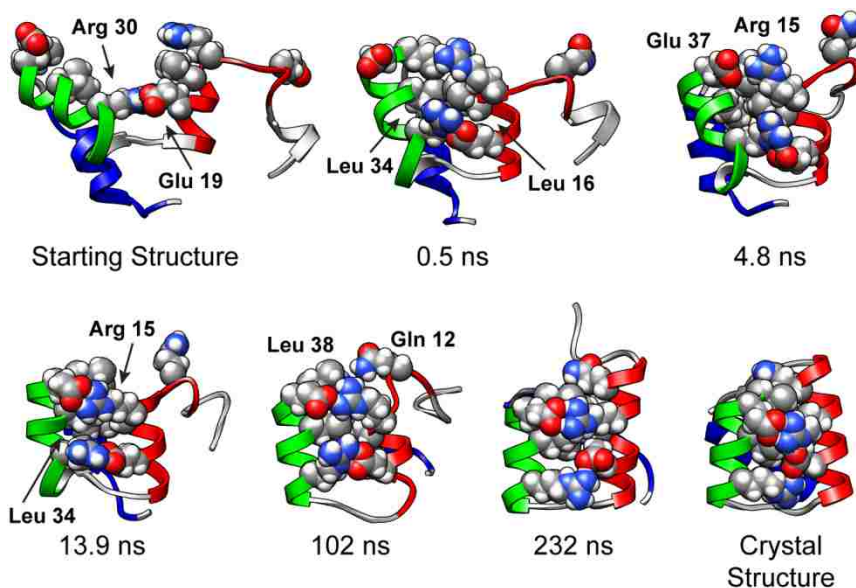
(a) Mean distance to the 319 K reference set in 35-dimensional property space is plotted over time for the quench simulation. The folded portion spans 122.210 – 394.596 ns with a mean distance to reference =  $0.19 \pm 0.02$ , and the denatured portion has a mean distance to reference of  $0.24 \pm 0.03$ . (b) The distribution of the mean distance to reference for the 319 K reference set (see Table 3.1) is shown ( $0.17 \pm 0.02$ ) along with the folded portion of the quench simulation and the denatured portion of the quench simulation. The folded portion of the quench simulation falls within the 319 K native state distribution, which indicates that the folded portion of the quench simulation is as similar to the native state as the native state is to itself.

The semiquantitative structure index (S-value) is a residue-based measure of secondary and tertiary structure, and it can be compared with experimental  $\Phi$ -values (Daggett *et al.* 1996). Both typically range from 0 to 1, and higher values reflect increased local structure in the TS ensemble. The correlation between our S-values and experimental  $\Phi$ -values was 0.79 for the refolding TS ensemble and 0.54 for the subsequent unfolding TS ensemble. The correlation between S and  $\Phi$  was low for the unfolding TS ensemble (394 ns) due to a different motion of HIII relative to HI and HII than is usually observed for EnHD. The unfolding TS ensemble was characterized by HIII pulling away from the core by first rotating such that it became parallel to HI and HII, whereas the N-terminus of HIII pulled away from HI and HII with the HII-HIII loop acting as a hinge for the refolding TS identified here and previously observed TSs (Mayor *et al.* 2000; Mayor *et al.* 2003b; DeMarco *et al.* 2004; Gianni *et al.* 2003). As a result, the following residues in the unfolding TS ensemble had more contacts with HII and thus a higher S than expected: Phe8 (N-term), Leu13 (HI), Leu16 (HI), Phe20 (HI), and Leu26 (HI-HII loop).



**Figure 3.3: Selected properties from the successful quench simulation**

(a)  $C\alpha$  RMSD of the core residues (8-53) to the minimized crystal structure. (b) DSSP showing the secondary structure for each residue over time:  $\alpha$ -helix (blue),  $\pi$ -helix (cyan), and 3/10-helix (magenta). (c) HI-HII contacts. (d) HIII-core contacts. For (c) and (d), a cross (+) was plotted at each time point when the two residues were in contact.



**Figure 3.4: Structures from the successful quench simulation showing the order of HI-HII contacts**

Arg30 – Glu19 was present in the starting structure. Leu34 – Leu16 formed at 0.5 ns, then Glu37 – Arg15 at 4.8 ns, and Leu34 – Arg15 at 13.9 ns. Finally, Leu38 – Gln12 formed at 102 ns. The “best” structure from the quench simulation as well as the crystal structure are also shown for comparison.

Several properties were plotted over the course of the quench simulation (Figure 3.3) to determine an order of events. Only one of the 5 HI-HII key contacts was present in the

starting structure: Arg30 – Glu19 (Figure 3.3c, 3.4). The next contact formed was Leu34 – Leu16 at 0.5 ns, then Glu37 – Arg15 at 4.8 ns, and Leu34 – Arg15 at 13.9 ns. In the successful quench simulation, these residues came in contact at 102 ns. The N-terminal end of HIII reformed  $\alpha$ -helix at 4 ns, the N-terminal end of HII at 52 ns, and the N-terminal end of HI at 111 ns (Figure 3.3b). In the 39 simulations where EnHD maintained the Arg30 – Glu19 contact, the Leu34 – Leu16 contact always formed first, and it formed very early in the simulation. Glu37 – Arg15 and Leu34 – Arg15 formed next in the 7 simulations where they formed at all. The last HI-HII contact, Leu38 – Gln12, only formed in the presence of the previous 4 contacts in 5 of our simulations, including the successful one.

Formation of these contacts and helices is apparent in the structures of EnHD over the time course of the quenched refolding simulation (Figure 3.4, 3.5). HI and HII snapped together in the first 2 ns of the simulation. Around 100 ns, the N-terminus of HI developed  $\alpha$ -helix and the final HI-HII contact formed. For the next 20 ns, HIII reoriented on the HI-HII scaffold. EnHD passed through the TS at 122 ns with HIII docking in the native orientation. At 232.260 ns, EnHD achieved its lowest mean distance to reference of 0.15 and a core C $\alpha$  RMSD of 2.30 Å. After the unfolding TS at 394 ns, HIII continued reorienting over the HI-HII scaffold, and around 570 ns, HI and HII came apart.

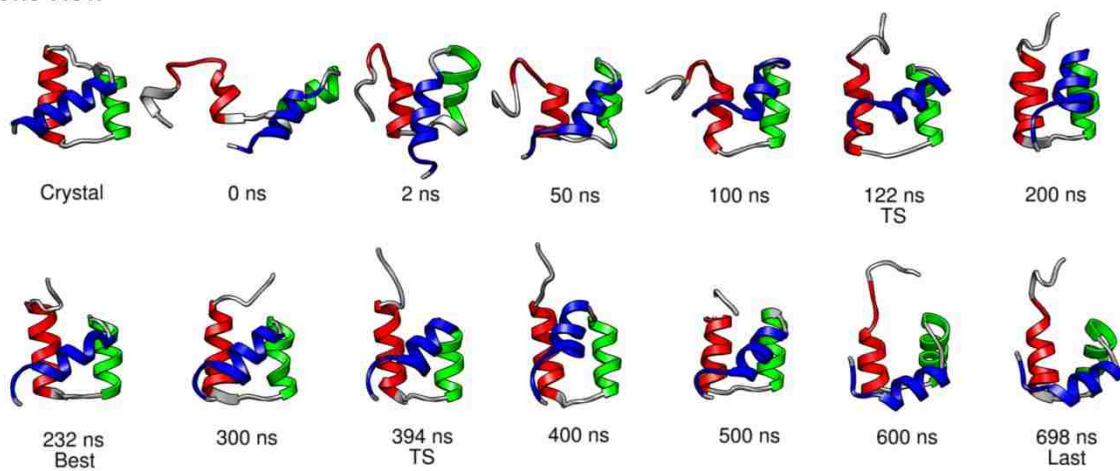
The core C $\alpha$  RMSD of EnHD is plotted for three simulations in Figure 3.6. The structure from 5 ns into the 498 K unfolding simulation was quenched at 319 K. Eventually it attained a core C $\alpha$  RMSD that would be expected for a protein in the native state at 319 K.

#### **3.4.4 Quench Simulations: Factors Preventing Refolding**

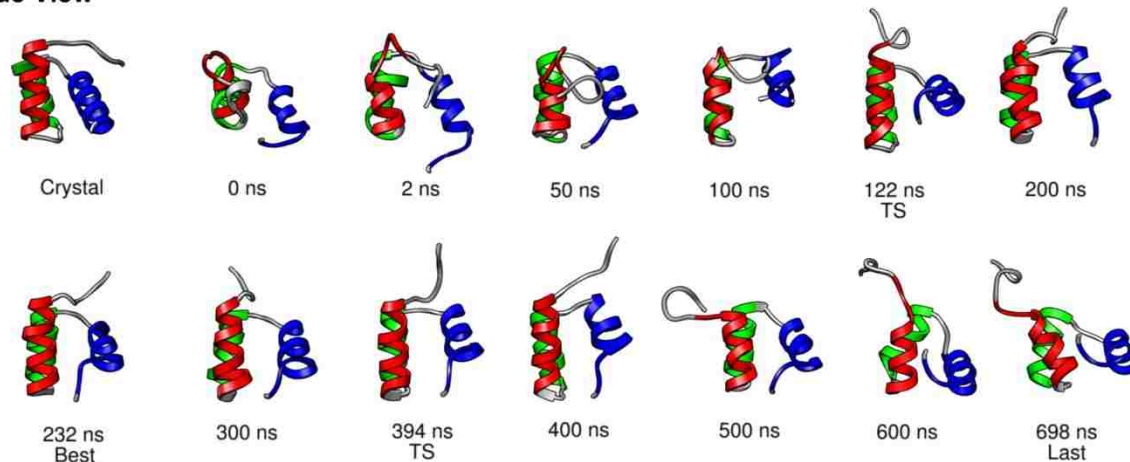
We identified several common motifs that prevented folding from proceeding by analyzing the 45 simulations ( $1.4 \times 10^7$  structures, 14.2  $\mu$ s of sampling) that did not refold. Many nonnative salt bridges formed in the unsuccessful quench simulations that kept residues involved in native contacts from finding each other. For example, Arg15 (in HI) often formed a salt bridge with Glu19 (HI), which prevented Arg30 (HII) and Glu19 (HI) from adopting their native arrangement (Figure 3.7a). This nonnative salt bridge also kept Arg15 (HI) from finding Glu37 (HII) and pulling HI and HII into their anti-parallel, native orientation. Arg29 (HII) often formed salt bridges with Glu37 (HII), again deterring the native Arg 15 (HI) – Glu37 (HI) salt bridge from forming and also kinking the N-terminal

end of HIII (Figure 3.7a). The placement and orientation of the N-terminus (residues 3-10) had a strong influence over whether the N-terminus of HI would form  $\alpha$ -helix. For example, Lys17 (HI) often interacted with the carbonyl groups in the backbone of Phe8 (N-term) and Ser9 (N-term), which along with a number of salt bridges that often formed between either Arg3 (N-term) or Arg5 (N-term) and Glu22 (HI), locked the N-terminus to HI and made it impossible for the N-terminus of HI to adopt native  $\phi/\psi$  angles and backbone hydrogen bonding patterns or the native Leu38 (HII) – Gln12 (HI) contact to form (Figure 3.7b,c).

### Front View

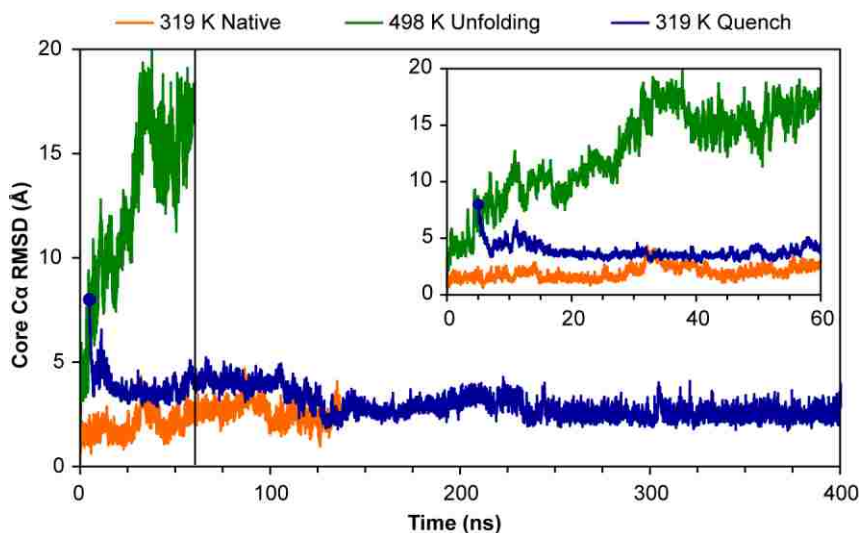


### Side View



### Figure 3.5: Structures from the Successful Quench Simulation

Front and side views of structures from the quench simulation are shown every 100 ns as are structures from significant time points in the simulation. 2 ns: 4 HI-HII contacts formed; 100 ns:  $\alpha$ -helix formed in the N-terminus and the final HI-HII contact formed; 122.210 ns: refolding TS; 232.260 ns: lowest mean distance to 319 K reference set (0.15); 394.596 ns: unfolding TS; 570 ns: HI-HII break apart; 698.600 ns: final structure from the simulation.

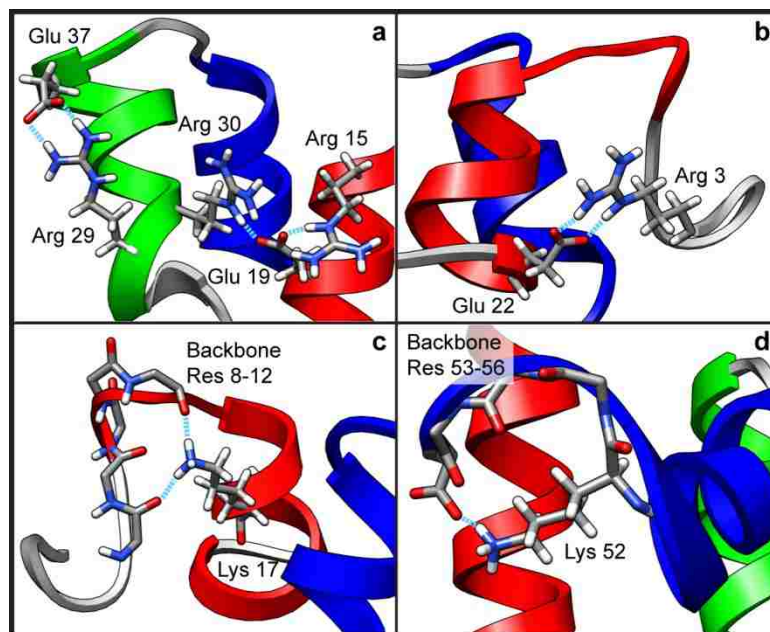


**Figure 3.6: Core C $\alpha$  RMSD for unfolding and refolding**

Core (residues 8-53) C $\alpha$  RMSD to the minimized crystal structure is plotted for a 319 K native (137 ns), 498 K unfolding (60 ns), and 319 K quench (395 ns) simulation. The structure from 5 ns into the 498 K unfolding simulation (8 Å core C $\alpha$  RMSD, emphasized on plot) was quenched to 319 K and after an additional 122 ns (at 127 ns on this plot), refolded to the native state for 272 ns (it unfolded at 399 ns on this plot). The inset shows the detail of the first 60 ns.

Even in the few cases where all 5 of the of the native HI-HII contacts formed, HIII never packed exactly correctly against the HI-HII scaffold. The C-terminus of HIII never fully formed  $\alpha$ -helix, usually due to a salt bridge between Lys52 and the carboxy group of the C-terminal residue, which made it impossible to pack all of the the core residues from the C-terminus into the core (Figure 3.7d). In one case, a salt bridge between Arg28 (HII) and Lys46 (HIII) caused the HII-HIII turn to kink, and it disrupted the  $\alpha$ -helix at the N-terminus of HIII as well as the orientation of HIII relative to HI and HII.

When there were long-lived, long-range, nonnative hydrophobic interactions, they were always accompanied by polar interactions (Figure 3.8). However, these polar interactions, particularly salt bridges, were often present in the absence of nonpolar interactions. In contrast, while there were long-lived nonpolar interactions, they were always accompanied by longer-lived polar interactions, particularly salt bridges.



**Figure 3.7: Structures from unsuccessful quench simulations showing nonproductive salt bridges**

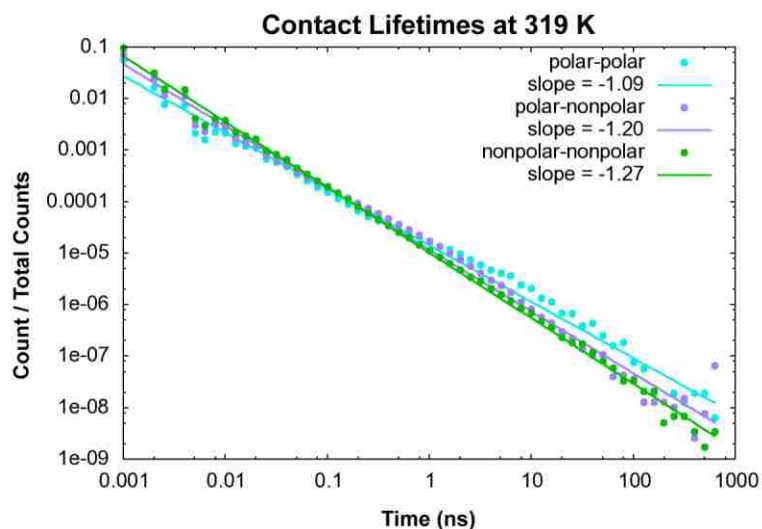
Salt bridges (cyan dashed lines) in several unsuccessful quench simulations that hindered refolding. (a) The Glu37 – Arg29 and Arg30 – Glu19 – Arg15 salt bridges inhibited the native Arg30 – Glu19 contacts from forming and caused HI and HIII to skew relative to each other. (b) The Glu22 – Arg3 salt bridge is an example of several salt bridges that formed between the N-terminus and Glu22 and kept HI from fully forming. (c) Lys17 coordinated the backbone carbonyl groups of Phe8, Ser9, and Gln12, which stabilized the N-terminus in a nonnative orientation. (d) Lys52 formed a salt bridge with the C-terminus, kinking the end of HIII.

## 3.5 Discussion

### 3.5.1 Property Space Description of the Native State and Unfolding Pathway of EnHD

Our 35-dimensional property space analysis allows us to compare the native state of EnHD at different temperatures better than any one of the individual properties alone. The 310, 314, and 319 K native state overlapped in property space with the 298 K native state, but the higher-temperature native states spanned regions that the 298 K native state did not. In other words, the 298 K ensemble was a subset of the 310, 314, and 319 K native states. This is due to the higher prevalence of the N' state at slightly elevated temperature. In N', HIII translates  $\sim 10$  Å towards the N-terminus (McCully *et al.* 2008), which has little effect on most of the reported properties. For example, the COM Distances are slightly, though not significantly, higher for residue pairs that have a member in the C-terminal portion of HIII as is the core C $\alpha$  RMSD. The number of native contacts is slightly, though again not

significantly, lower. The percentage of NOEs satisfied for the native state at 310, 314, and 319 K ( $86 \pm 2\%$ ) was the same as for the native state at 298 K ( $85 \pm 3\%$ ). So, despite the fact that the native state in the elevated temperature simulations was broader than at 298 K, its agreement with the native state as measured by experiment was comparable. Because the temperature was slightly elevated, EnHD was more likely to overcome the energy barriers that confined it to a smaller portion of the native energy well at lower temperatures.



**Figure 3.8: Contact lifetimes for different contact types**

Contact lifetimes for polar-polar (cyan), polar-nonpolar (purple), and nonpolar-nonpolar (green) contacts for one 700-ns simulation at 319 K. A line was fitted to the data on a log-log plot, and the resulting slopes are indicated in the legend. More negative slopes indicate that shorter lifetimes dominate over longer, thus polar-polar contacts are longer-lived than nonpolar-nonpolar contacts at this temperature.

From our 35-dimensional property space, we calculated a multidimensional embedded 1D reaction coordinate. The reaction coordinate for the high-temperature unfolding shows that the TS determined previously (Mayor *et al.* 2000; Gianni *et al.* 2003; Mayor *et al.* 2003b; DeMarco *et al.* 2004) falls just outside the reference distribution (Figure 3.1), as would be expected for this method. The intermediate and denatured populations formed separate but connected peaks in the 498 K distribution. Instead of the generic 15 properties used for high-throughput analysis of data in the lab's Dynameomics database (Toofanny *et al.* 2010), many of the properties employed here were specific to EnHD. Better discrimination between the native, transition, intermediate, and denatured states was possible using the 35 properties listed in Table 3.3 than their standard 15 property subset. This

underscores the importance of using multiple, protein-specific properties to determine the state of a protein.

### 3.5.2 The Refolded Native State

A drop in mean distance to reference to  $0.193 \pm 0.019$  was observed upon refolding (Figure 3.2a) as well as a drop in core C $\alpha$  RMSD to  $2.69 \pm 0.39$  Å (Figure 3.3a). For comparison, the reference native state at 319 K had a mean distance to reference of  $0.169 \pm 0.019$  (Figure 3.2b) and a core C $\alpha$  RMSD of  $2.20 \pm 0.60$  (Table 3.3). The refolded native state kept all five native HI-HII contacts and had intermittent HIII-core contacts (Figure 3.3c,d), as did the reference native state (data not shown). However, HIII was more variable in the refolded native state, both in fraction of  $\alpha$ -helix and position relative to HI and HII (Figure 3.5). HIII spent more time in alternate orientations than the two seen primarily in the reference native state. The first was characterized by HIII laying diagonally across HI and HII as in the crystal structure (Figure 3.5) and the second by HIII moving  $\sim 10$  Å towards the N-terminus, although other arrangements were observed as well (McCully *et al.* 2008). The mobility of HIII is apparent in Figure 3.3d, which shows the residues between HIII and the core going in and out of contact.

The refolded native state had a slightly lower fraction  $\alpha$ -helix ( $0.61 \pm 0.04$ ; Figure 3.3b) than the reference native state ( $0.68 \pm 0.04$ , Table 3.3), which was due to fraying at the C-terminus of HIII. This fraying was also apparent in the NOE satisfaction calculated from simulation. The refolded native state had 77% of the NOEs satisfied, while the reference native state had 86% (Table 3.1). The most severe violations were in the C-terminus since the last turn of HIII never formed due to a stable salt bridge between Lys52 and the C-terminal carboxyl group (Figure 3.7d), but this is consistent with experimental data. The C-terminus has higher B-factors in the crystal structure ( $27 \pm 15$  Å<sup>2</sup>) than the core of the protein ( $19 \pm 11$  Å<sup>2</sup>; Clarke *et al.* 1994), and NMR experiments show higher J-couplings and lower backbone order parameters in the C-terminus, which is attributed to helix fraying (Religa 2008). Otherwise, the violations were due to Leu26, which did not have all of its NOEs satisfied in the native simulations either. Additionally, due to the orientation of the the N-terminus, Phe8, which accounted for many of the violations, interacted with other residues in the N-terminus rather than residues in HI and HIII as in the native simulations.



### 3.5.3 Comparison of One Successful Folding Trajectory with 45 Unsuccessful Trials

We performed 46 independent “refolding” simulations beginning from the EnHD intermediate state, which is the starting point for experimental folding under “physiological conditions” (Mayor *et al.* 2003a; Mayor *et al.* 2003b). We generated the intermediate by unfolding the protein at 498 K and quenching it at experimentally determined refolding temperatures. Besides providing an atomic-level description of aspects of the folding pathway, our quench simulations give insight into energy traps the protein might encounter as it navigates the energy landscape between the denatured and native states. For the most part, EnHD got stuck in nonproductive conformations due to the formation of nonnative salt bridges, though these interactions were stabilized by nonpolar interactions in some cases. EnHD contains 9 Arg, 4 Lys, and 6 Glu residues, for a total of 19 of 54 residues, or 35%. While native salt bridges stabilized the native state, and their formation contributed to key steps along the folding pathway (e.g. Arg30 – Glu19 initiating HI and HII zipping together), EnHD can form many more nonnative salt bridges that can stabilize nonnative conformations (Figure 3.7). Salt bridges were the strongest and longest-lived noncovalent interaction in our simulations (Figure 3.8). So, once a favorable nonnative interaction formed, it often did not break on the timescale of our shorter simulations. In the longer unproductive quench simulations, nonnative salt bridges gave way to native interactions, and we expect that if the simulations were extended, more of them would eventually refold to the native state.

### 3.5.4 Misfolding Traps in the Intermediate Slow Refolding

The protein correctly refolded in just one of the simulations, but in all 46 of our quench simulations, the zipping together of HI and HII was a common initial event. The order of contact formation between HI and HII observed in the productive refolding simulation was representative of other quench simulations: Arg30 – Glu19 was present in the starting structure, Leu34 – Leu16 formed early on, Leu34 – Arg15 and Gly37 – Arg15 were formed next, and Leu38 – Gln12 was last to form (Figure 3.3c, 3.4). All 5 contacts were formed at the same time in only 5 of the quench simulations. The Leu34 – Leu16 contact was second to form in all 40 quench simulations when the Arg30 – Glu19 contact remained intact. Additionally, both Leu16 and Leu34 are highly conserved among all 84 homeodomains in *D. melanogaster* (Noyes *et al.* 2008). The consistent order of contact

formation and the evolutionary conservation of these two Leu residues suggest that the Leu34 – Leu16 contact is critical for folding. If true and EnHD were unable to form this contact, folding would be halted in the intermediate state. Indeed, when Leu 16 is mutated to the smaller Ala, EnHD preferentially populates the folding intermediate (Mayor *et al.* 2003a; Religa *et al.* 2005).

The trapped intermediate states were characterized by HIII never being completely correctly packed onto the HI-HII scaffold. The C-terminus of HIII never fully formed  $\alpha$ -helix, usually because of a salt bridge between its terminal carboxy group and Lys52. Indeed, the Fersht Lab has shown that removing this interaction via mutation of Lys52 to Ala leads to faster folding:  $6.4 \times 10^4 \text{ s}^{-1}$  versus  $3.8 \times 10^4 \text{ s}^{-1}$  for wild type (Gianni *et al.* 2003). These data are consistent with the implication from the simulations that there is an interaction involving Lys52 competing with productive folding. For comparison,  $\alpha_3\text{D}$ , a designed 3-helical bundle protein folds faster than EnHD and does not populate an intermediate (Zhu *et al.* 2003). For  $\alpha_3\text{D}$ , folding rates of  $3.1 \times 10^5 \text{ s}^{-1}$  were measured at 49 °C ( $t_{1/2} \approx 4.8 \mu\text{s}$  at 25 °C) at a pH of 2.2. Because folding took place at low pH, all Asp and Glu residues were protonated, and nonnative salt bridges could not cause bottlenecks in the folding pathway.

In another study of 3-helical bundle protein folding, the R16 and R17 domains of  $\alpha$ -spectrin were shown to fold  $\sim 3$  orders of magnitude slower than the R15 domain due to the internal friction of the proteins (Wensley *et al.* 2010). Based on  $\Phi$ -value analysis and measurement of the internal friction of the three proteins and several variants, the authors proposed that transient misdocking of the helices slowed folding in the cases of R16 and R17. Similarly, the nonnative interactions seen in our refolding simulations kept the helices from packing correctly (Figure 3.7) and slowed folding.

Accordingly, our MD simulations strongly imply that the intermediate refolds slowly because of diversions, and such transient off-pathway traps prevent the protein from finding productive folding pathways on the timescale of our simulations. The combination of unfolding simulations with many parallel folding simulations is seen to be very powerful (Fersht 2002); unfolding simulations mirror productive, on-pathway folding events measured experimentally (Mayor *et al.* 2000; Mayor *et al.* 2003b; Gianni *et al.* 2003), and refolding simulations detect both on-pathway folding directly (described here and in simulations at the

$T_m$ ; Day and Daggett 2007; McCully *et al.* 2008) as well as off-pathway events in the nonproductive refolding simulations. These quench simulations have resolved the problem of why the folding intermediate folds relatively slowly: the transient off-pathway traps slow the reaction by 1-2 orders of magnitude.

### 3.5.5 Microscopic Reversibility

As protein folding and unfolding have been shown to follow the same pathway for EnHD and CI2 at their melting temperatures (Day and Daggett 2007; McCully *et al.* 2008), it is interesting to consider the order that the HI-HII contacts were gained in the quench simulations compared with the order in which they were lost in the high-temperature unfolding direction, noting that unfolding and folding are occurring under different conditions. The order of loss for the five contact pairs in the high temperature denaturation simulation was Leu34 – Arg15 (0.2 ns), Leu38 – Gln12 (1.2 ns), Glu37 – Arg15 (3.0 ns), Leu34 – Leu16 (4.2 ns), and Arg30 – Glu19 (8.4 ns). The order the contacts were gained in the successful quench simulation was Arg30 – Glu19 (present at start), Leu34 – Leu16 (0.5 ns), Glu37 – Arg15 (4.8 ns), Leu34 – Arg15 (13.9 ns), and Leu38 – Gln12 (102 ns; Figure 3.3c, 3.4). The order of gain and loss is nearly identical with the only exception being swapping of the Leu34 – Arg15 and Leu38 – Gln12 interactions, although they together occurred first in unfolding and last in folding. We note that this comparison is subject to the caveat that we have only one successful refolding trajectory. However, when these 5 key HI-HII contacts formed in the 45 simulations that never fully refolded, they were gained in the same order as the successful quench simulation, even though all 5 contacts only formed in 4 of the unsuccessful simulations.

Formation of the HI-HII scaffold is a critical step in the folding pathway for EnHD and must be completed before HI and HIII can dock in their native orientation. The scaffold forms in a stepwise manner beginning with contacts forming near the HI-HII loop, including the critical Leu34 – Leu16 contact, and continuing with combined HI helix formation and the zipping together of HI and HII.

## Chapter 4: Unfolding in a Test Tube – Multimolecule Atomistic Molecular Dynamics Simulations of the Engrailed Homeodomain

### 4.1 Summary

Molecular dynamics simulations of protein folding or unfolding, unlike most *in vitro* experimental methods, are performed on a single-molecule. The effects of neighboring molecules on the un/folding pathway are largely ignored experimentally and simply not modeled computationally. Here, we present two all-atom, explicit solvent molecular dynamics simulations of 32 copies of the Engrailed Homeodomain (EnHD), an ultra-fast folding and unfolding protein for which the un/folding pathway is well characterized. These multimolecule simulations, in comparison with single-molecule simulations and experimental data, show that intermolecular interactions have little effect on the un/folding pathway. EnHD unfolded by the same mechanism whether it was simulated in water only or also in the presence of other EnHD molecules. It populated the same native state, transition state, and folding intermediate in both simulation systems and was in good agreement with experimental data available for each of the three states. Unfolding was slowed by interactions with neighboring proteins, which were mostly hydrophobic in nature and ultimately caused the proteins to aggregate. Protein-water hydrogen bonds were also replaced with protein-protein hydrogen bonds, additionally contributing to aggregation. Yet, despite the increase in protein-protein interactions, the protein aggregates formed in simulation did not do so at the total exclusion of water. These simulations support the use of single-molecule techniques to study protein unfolding and also provide insight to the types of interactions that occur as proteins aggregate at high temperature at an atomic level.

### 4.2 Introduction

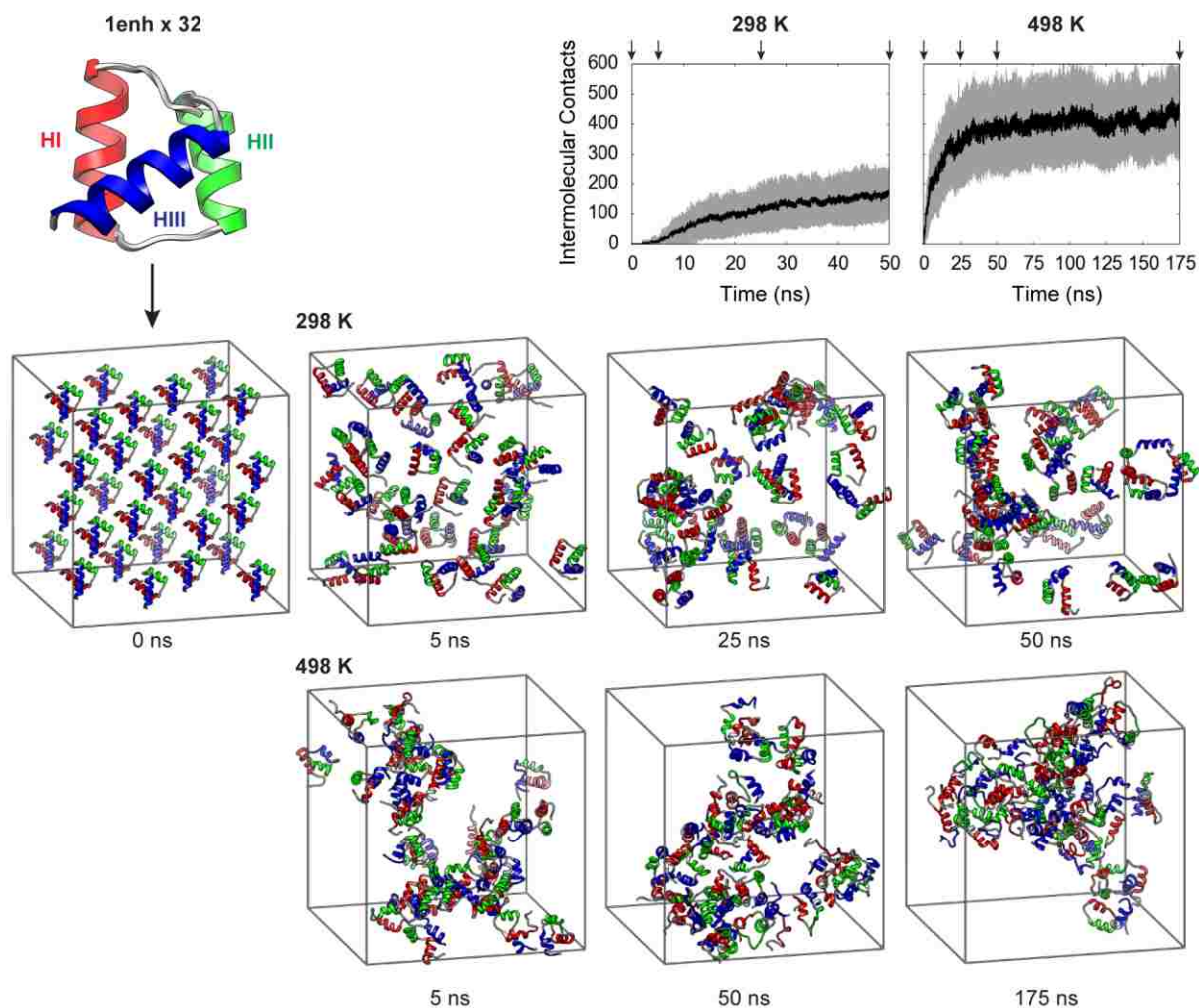
The folding pathway of the Engrailed Homeodomain (EnHD) has been extensively characterized through a combination of experimental and computational techniques. EnHD is a 3-helix bundle protein, with helices I and II packing in parallel and helix III docking

across them (Figure 4.1). It folds and unfolds on ultrafast timescales, which makes its folding pathway a good candidate to be studied by simulation. The structure of the transition state was first predicted by molecular dynamics (MD) simulations (Mayor *et al.* 2000) and later validated by experimental techniques (Gianni *et al.* 2003). A folding intermediate was identified by experiment (Mayor *et al.* 2003a), structurally characterized by MD (Mayor *et al.* 2003b), and the MD-predicted structure was later validated by NMR (Religa *et al.* 2005). In addition, it has been shown that the protein adheres to the principle of microscopic reversibility in simulations at its melting temperature ( $T_m$ ) such that unfolding and refolding occur in a single continuous trajectory (McCully *et al.* 2008). Furthermore, when high temperature unfolding simulations are quenched to folding permissive temperatures, the protein refolds by the reverse of the unfolding (McCully *et al.* 2010). Despite the fact that the experimental techniques used provide ensemble measurements and the MD simulations are single-molecule in nature, the agreement has been very good, allowing for a much richer description of the folding/unfolding process of EnHD than would be possible through either approach alone.

Single-molecule vs. ensemble measurements aside, protein concentration is a variable that differs greatly between simulation, *in vitro* experiments, and protein folding *in vivo*. MD simulations of protein un/folding are effectively carried out at infinite dilution. Structural and kinetic measurements of EnHD have employed protein concentrations on the order of 10  $\mu\text{M}$  – 10 mM. On the other end of the scale, proteins fold *in vivo* in a crowded cellular environment with predicted concentrations of  $\sim 300$  mg/mL (Silverman and Glick 1969). For EnHD, a 7 kDa protein, this cellular concentration equates to  $\sim 40$  mM. The effect of neighboring molecules on the folding pathway has been largely ignored, particularly in computational studies, without regard to whether the low concentration condition is realistic. Given the wealth of information regarding EnHD folding, it is a good system for investigating the effect of concentration on protein behavior.

Here, we present MD simulations of protein unfolding in a multimolecular system, which we refer to as “test-tube” simulations. Our system consisted of 32 copies of EnHD solvated with explicit water, resulting in a concentration of 18 mM. This system was heated to 298 K to probe the dynamics and conformational properties of the native state of EnHD in

the presence of neighboring molecules. In addition, the system was heated to 498 K to investigate the thermal unfolding pathway in a high concentration of protein. Previous studies have shown that the unfolding is an activated process and the pathway is independent of temperature from 373 – 498 K, such that the process is just faster at higher temperature (Mayor *et al.* 2000, Gianni *et al.* 2003, Mayor *et al.* 2003b, DeMarco *et al.* 2004). The test-tube simulations were compared with multiple independent single-molecule simulations (10 simulations at 298 K and 7 at 498 K) to determine the effect of neighboring molecules on the protein and protein-water interactions.



**Figure 4.1: Structures from test-tube simulations.**

32 copies of the crystal structure (1enh) were placed in a simulation system and were heated to 298 and 498 K. Structures are colored by helix (HI: 10-22 red; HII: 28-38 green; HIII: 42-55 blue), and water is not displayed. The average number of intermolecular contacts per molecule and standard deviations are plotted over time. Proteins in the high temperature unfolding simulation formed many more contacts than those in the 298 K simulation.

Both the native dynamics and unfolding pathway were largely unaffected by the presence of neighboring molecules, though unfolding progressed more slowly in the test-tube simulation. The native, transition, and intermediate states populated during unfolding agree equally well with experiment when considering the single-molecule and test-tube simulations. Based on similarity between the test-tube and single-molecule simulations, it is not necessary to consider neighbor effects to accurately reproduce the unfolding pathway of a protein. The resources required to perform a simulation on a multimolecular system do not generally justify the information gleaned here on the unfolding pathway; however, a similar analysis should be performed on additional protein systems to ensure that this conclusion is generally valid.

Molecules in the high-concentration test-tube systems aggregated, with the high temperature simulation showing one main cluster and the low temperature simulation showing many smaller, dynamic clusters. Most of the contacts between proteins were hydrophobic in nature, unlike most of the contacts within the single protein simulations. At high temperature, nonpolar packing interactions (or hydrophobic interactions) that were lost upon unfolding were replaced with nonpolar, hydrophobic interactions between neighboring proteins, which resulted in aggregation. Hydrogen bonds also formed between protein molecules, many at the exclusion of water, further promoting aggregation. Thus, these simulations provide a molecular picture of protein aggregation at elevated temperature.

## 4.3 Methods

### 4.3.1 Molecular Dynamics Simulation Parameters

The molecular dynamics simulations were generated using the package, *in lucem* molecular mechanics (*ilmm*; Beck *et al.* 2000-2012) with the Levitt *et al.* (1995) force field. Several of the single-molecule simulations (all runs at 298 K, and runs 1 and 2 at 498 K) were reported previously (Mayor *et al.* 2000, Gianni *et al.* 2003, Mayor *et al.* 2003b, McCully *et al.* 2010). The starting structure for the molecular dynamics simulations was the crystal structure (PDB ID: 1enh; Clarke *et al.* 1994; Figure 4.1). To create the multimolecule, “test-tube” system, 32 of these structures were arranged in a face centered cubic lattice with sides of length  $\sim 144 \text{ \AA}$  giving concentrations of  $\sim 18 \text{ mM}$ . The system was

solvated with flexible F3C water (Levitt *et al.* 1997), and the water density was set based on the simulation temperature according to the experimentally determined liquid-vapor coexistence curve (298 K: 0.997 g/mL (Kell 1967), 498 K: 0.829 g/mL (Haar *et al.* 1984)). The resulting systems had 85,230 and 71,148 water molecules for a total of 285,994 and 243,748 atoms at 298 and 498 K, respectively. The NVE microcanonical ensemble (constant number of particles, volume, and energy) was used with 2-fs timesteps, and structures were saved every 1 ps for analysis. An 8 Å force-shifted nonbonded cutoff was employed (Beck *et al.* 2005), and the nonbonded list was updated every two steps. For the single-molecule simulations, there were a total of 7 performed at 298 K (2 x 80 ns, 2 x 50 ns, 3 x 20 ns) and 10 at 498 K (1 x 39 ns, 1 x 60 ns, 8 x 50 ns) totaling 819 ns. The test-tube system was simulated for 50 ns at 298 K and 175 ns at 498 K for a total of 225 ns, giving an equivalent of 7.2  $\mu$ s of single-molecule data. In total, there was over 8  $\mu$ s of simulation data in this study.

The NMR structure of the L16A mutant of EnHD (a surrogate for the intermediate state) has been solved (Mayor *et al.* 2003a, Religa *et al.* 2005). Each of the 25 models in the NMR structure (PDB ID: 1ztr) were truncated to residues 3-56, and the same properties were calculated for each model as for the MD-generated structures.

#### 4.3.2 Molecular Dynamics Simulation Analysis

A total of 39 physical properties were monitored for all of the simulations to create an alternate description of the trajectory in property space, which can be very helpful for comparing different trajectories (Kazmirski *et al.* 1999, Beck and Daggett 2007, Toofanny *et al.* 2007, McCully *et al.* 2010; Figure 4.2). C $\alpha$  RMSD to the minimized crystal structure was calculated for the core residues (8-53), which excludes the floppy N- and C-termini. The percentage of residues forming  $\alpha$ -helix was calculated by our in-house implementation of the DSSP algorithm, which bases secondary structure assessments on hydrogen bonding (Kabsch and Sander 1983). Center-of-mass (COM) distances were calculated between 16 residue pairs previously found to be indicative of the folded native state (McCully *et al.* 2008, McCully *et al.* 2010). The number of native residue-residue contacts were counted and classified as occurring between main chain and side chain atoms and whether they were present in the starting structure (native/nonnative). If non-sequential residues contained



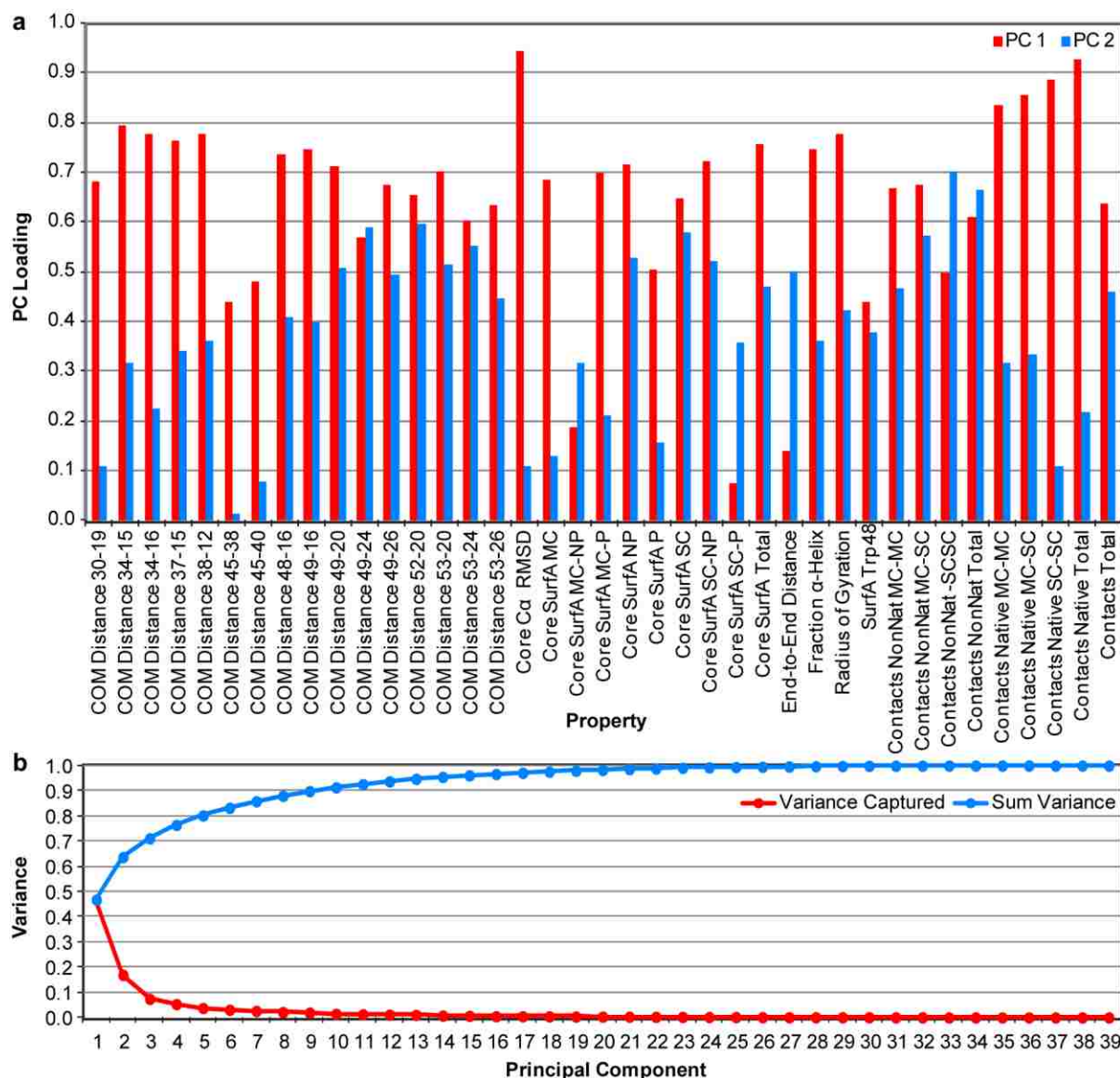
carbon atoms that were  $\leq 5.4 \text{ \AA}$  or any other non-hydrogen atoms that were  $\leq 4.6 \text{ \AA}$  apart, the residues were considered in contact. Solvent accessible surface area was calculated for the core residues using our in-house implementation of the Lee and Richards (1971) algorithm and a probe radius of  $1.4 \text{ \AA}$ . The resulting SASA was classified as main chain or side chain and polar or nonpolar. SASA of Trp48, the fluorescence probe for folding, was included as well, as were the radius of gyration and end-to-end distance. Principal component analysis was carried out on the resulting normalized 39-dimensional property space as described previously (Toofanny *et al.* 2010).

Contacts were also classified by type. Hydrogen bonds were defined as when the donor atom's hydrogen was  $\leq 2.6 \text{ \AA}$  from the acceptor atom, and the angle between the three atoms was within  $45^\circ$  of linearity. Hydrophobic interactions were for aliphatic carbons atoms separated by  $\leq 5.4 \text{ \AA}$ . Any other pair of non-hydrogen atoms that did not meet the aforementioned criteria but were  $\leq 4.6 \text{ \AA}$  apart were classified as an "other," nonspecific contact. Contacts between atoms of neighboring residues were not counted.

The 298 K simulations were compared with experiment via NOE satisfaction. 654 NOEs are available for our construct, residues 3-56 (Religa 2008). An NOE was considered satisfied if the  $\langle r^{-6} \rangle$  weighted distance between the closest protons in the NOE was less than  $5.5 \text{ \AA}$ , which was the maximum cutoff published in the EnHD experimental set.

In both the single-molecule and test-tube 498 K simulations, TS ensembles for the unfolding and any refolding events were determined. Unfolding TSs were considered the point of no return from the native-like cluster of the 3-dimensional multidimensional scaling (3D MDS) of the all-against-all  $C\alpha$  RMSD matrix (Li and Daggett 1994). To create a TS ensemble, the TS is taken as the final point in the native-like cluster and the previous 5 ps. For refolding, the TS is the first point in the native cluster, and the TSE is that point along with the subsequent 5 ps (McCully *et al.* 2008). The structure index, or S-value, was calculated over the ensemble as the product of  $S_{2^\circ}$  and  $S_{3^\circ}$  for each residue (Daggett *et al.* 1996).  $S_{2^\circ}$  is the fraction of native secondary structure, and  $S_{3^\circ}$  is the fraction of native and nonnative contacts in the TSE relative to the number of contacts in the crystal structure. S-values are a semi-quantitative reflection of structure in the TS. Experimental  $\Phi$ -values are based on energetics but are used to infer structural attributes of the TS (Gianni *et al.* 2003).

Both take values between 0 and 1 and can be compared and used for validation of the MD-generated TS structures (Gianni *et al.* 2003, Mayor *et al.* 2003b).



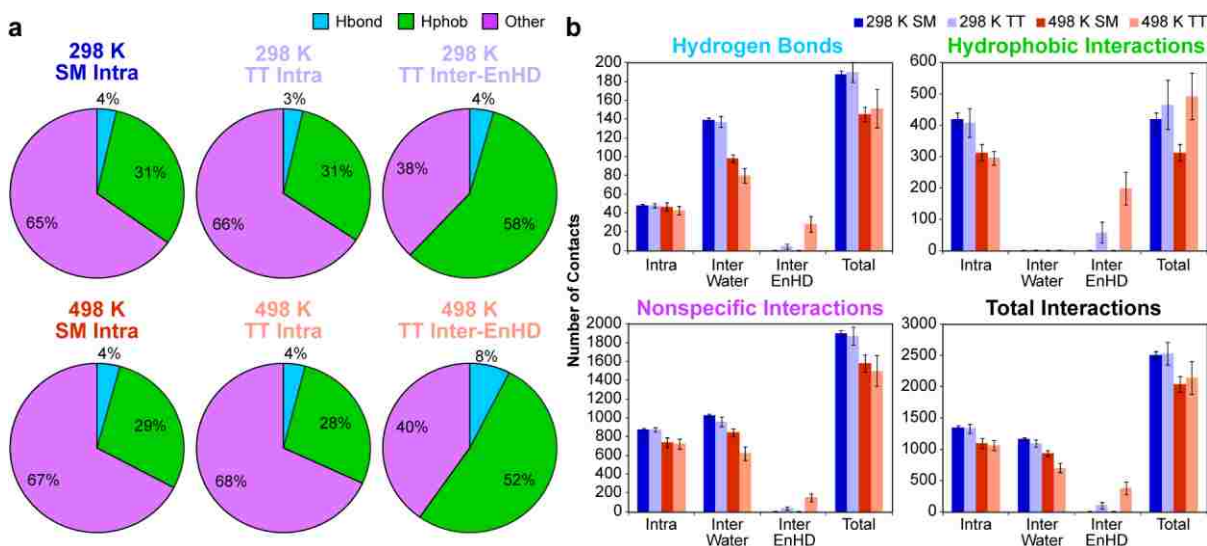
**Figure 4.2: Principal component analysis of 39-dimensional property space**

(a) The 39 properties making up the property space are plotted along with their loadings in the first (red) and second (blue) PC. (b) The variance (red) and total variance (blue) captured is plotted for each PC.

#### 4.4 Results and Discussion

To create the multimolecule, “test-tube” system, 32 copies of EnHD were placed on a lattice and solvated by water giving a concentration of  $\sim 18$  mM (Figure 4.1). The distance between the center of mass of any molecule and that of its closest neighbor in the initial

lattice was a minimum of 51 Å, and the closest atoms were 23 Å apart. After constructing the system, the temperature was brought to either 298 or 498 K. Figure 4.1 shows the evolution of the two test-tube systems over time. At high temperature the molecules quickly began to interact, aggregate, and form a large number of intermolecular contacts. In contrast, at 298 K the proteins moved more freely through solution, transiently interacting with neighboring molecules.



**Figure 4.3: Types of intermolecular vs. intramolecular interactions**

(a) The fraction of contacts, as classified as hydrogen bonds (Hbond, cyan), hydrophobic interactions (Hphob, green), and nonspecific interactions (Other, purple) are plotted for interactions that occur between atoms within a single-molecule (Intra) or between protein molecules (Inter-EnHD). Simulations are additionally grouped by single-molecule (SM, lighter) or test-tube (TT, darker) and temperature (298 K blue, 498 K red). (b) The number of contacts, classified by type of interaction, made within a single EnHD molecule (Intra), between EnHD and water (Inter Water), between EnHD and other protein molecules (Inter EnHD), and summed over the three classes of interacting partners (Total) for each set of simulations/molecules (298 K single-molecule, dark blue; 298 K test-tube, light blue; 498 K single-molecule, dark red; 498 K test-tube, light red).

#### 4.4.1 Nature of the Intermolecular Interactions

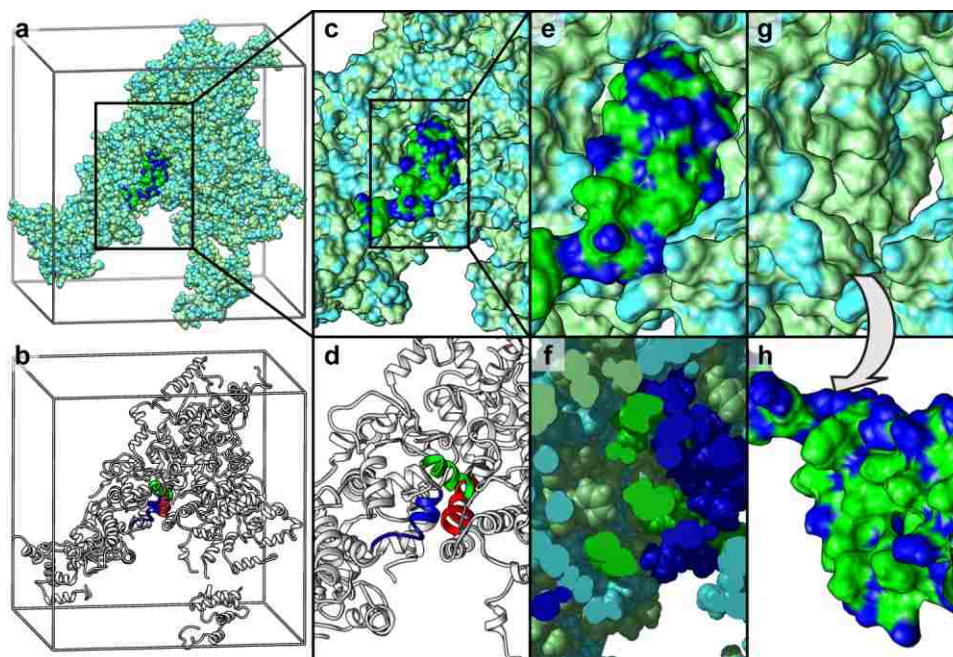
The types of contacts that occurred within a molecule (i.e. contacts with self) differed in proportion and number from the types of contacts that occurred between molecules. Figure 4.3a shows the proportion of hydrogen bonds (Hbond), hydrophobic interactions (Hphob), and nonspecific interactions (Other) that occur as intramolecular and intermolecular contacts. Nonspecific contacts include interactions between hydrophobic and polar groups, as well as polar or charged interactions that do not satisfy the angular geometry of a hydrogen bond. In both the single-molecule (SM) and test-tube (TT) simulations, intramolecular contacts were primarily nonspecific (65-68%), followed by hydrophobic (28-

31%). However, the contacts between protein molecules in both the 298 and 498 K simulations were primarily hydrophobic (58 and 52%) with only 38 and 40% nonspecific interactions.

The total number of interactions present within molecules, between molecules, and with water was the same for single-molecule vs. test-tube simulations at a given temperature (Figure 4.3b). For each of the four systems, contacts were about equally split between intramolecular and intermolecular. Overall, there were fewer contacts at high temperature than at low temperature because the density of water is lower at 498 K. Additionally, there were fewer intramolecular contacts at 498 K because EnHD unfolded. Contacts with water were lost in the two test-tube simulations and replaced with contacts between proteins. Despite the aggregation, EnHD still had twice as many contacts with water than with other protein molecules at 498 K and 10 times as many at 298 K (Figure 4.3b).

Hydrogen bonds were the only type of intramolecular interaction to remain constant between the native and high-temperature simulations (Figure 4.3b). EnHD made ~45 hydrogen bonds with itself, primarily within the helical backbone, in all four systems (single-molecule and test-tube simulations at both low and high temperature). The native hydrogen bonds that were lost during unfolding were replaced with nonnative hydrogen bonds, primarily between side chains. It has been shown previously that the denatured state of EnHD contains many such nonnative interactions (McCully *et al.* 2010). In the high temperature test-tube simulation, EnHD made fewer hydrogen bonds with water than in the single-molecule simulations, replacing them with hydrogen bonds to neighboring protein molecules.

The most dramatic increase in contacts in the high temperature test-tube simulation reflected hydrophobic interactions (Figure 4.3b). Proteins in the single-molecule and test-tube simulations at both temperatures had the same number of intramolecular hydrophobic interactions, but in the test-tube simulations there was an increase in the number of intermolecular protein-protein interactions. EnHD gained hydrophobic contacts with neighboring molecules without a net loss of intramolecular hydrophobic contacts (Figure 4.3b).



**Figure 4.4: Hydrophobic interactions at high temperature**

Molecule 23 at 175 ns in the 498 K test-tube simulation is shown throughout with hydrophobic groups colored green or light green and polar groups colored blue or cyan in the space-filling representation. In the ribbon representations, molecule 23 is shown colored by helix, and the remaining molecules are colored white. (a) and (b) show molecule 23 in the context of the whole system. (c) and (d) focus on molecule 23, and (e) further zooms into a hydrophobic patch. Hydrophobic patches (green and light green) on the surface of the molecules came together to form hydrophobic clusters as the proteins aggregated. (f) shows a slice through the proteins in the same orientation as e, and even more hydrophobic surface area can be seen buried between the molecules. Molecule 23 has been removed in (g) to show the interacting surface of the neighboring molecules. (h) shows only molecule 23 rotated 180° to display the surface it presented to the neighboring molecules.

While many hydrophobic residues made contacts with neighboring molecules, many more were exposed to solvent and did not make favorable intermolecular interactions. It is geometrically impossible for all hydrophobic residues on EnHD to be buried, even in the native state, so solvent exposure of some residues is expected. Hydrophobic clusters also frequently formed within a molecule (intramolecular) and consisted of both native and nonnative interactions. At the end of the high temperature test-tube simulation, molecule 23 had a hydrophobic patch on helix III that was buried in a hydrophobic pocket created by two other copies of EnHD (Figure 4.4). Figure 4.4a,c,e shows progressively closer views of molecule 23 within the context of the 31 other molecules in the simulation colored by hydrophobic (green) and polar (blue) groups. A slice into the binding surface (Figure 4.4f) shows the interactions between hydrophobic groups of molecule 23 and the other proteins that formed the binding pocket. The binding pocket with molecule 23 removed (Figure 4.4g)

shows a large hydrophobic patch (green), which was matched on the binding surface of molecule 23 (Figure 4.4h).

Protein folding is driven by release of water from exposed nonpolar groups and the burial of hydrophobic amino acids in the core of the protein. This is apparent in the decrease in the total number of hydrophobic contacts in the single-molecule native vs. unfolding simulations (Figure 4.3b). In the test-tube simulations, there was an increase in total hydrophobic interactions over the number in the native single-molecule simulations. Just as folding is driven by the need to bury hydrophobic groups, aggregation was also dominated by hydrophobic interactions. Indeed, the hydrophobic contacts that were lost upon unfolding were replaced with intermolecular hydrophobic contacts due to aggregation. In terms of burial of hydrophobic surface area, aggregation is as effective as folding.

**Table 4.1: Average properties of the 298 K native simulations**

Property	Single-Molecule Simulations	Test-Tube Simulation
Core C $\alpha$ RMSD* (Å)	2.22 $\pm$ 0.61	2.58 $\pm$ 1.52
Fraction $\alpha$ -Helix <sup>†</sup>	0.71 $\pm$ 0.05	0.69 $\pm$ 0.03
Fraction Native Contacts <sup>‡</sup>	0.77 $\pm$ 0.04	0.78 $\pm$ 0.05
Fraction NOEs Satisfied <sup>§</sup>	0.87 $\pm$ 0.03	0.85 $\pm$ 0.03

\* RMSD was calculated over the C $\alpha$  atoms of residues 8-53. All properties are given as the average  $\pm$  1 s.d.

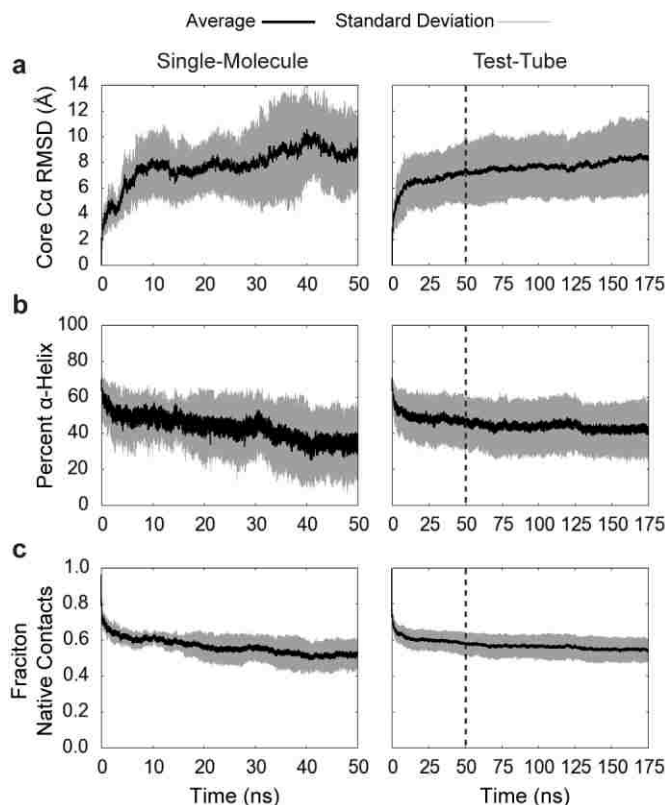
<sup>†</sup> DSSP was used to determine what fraction of the 54 residues was in  $\alpha$ -helix.

<sup>‡</sup> Two residues were considered to be in contact if the distance between at least one non-hydrogen atom from each was less than 5.4 Å for carbon/carbon pairs and 4.6 Å for all other pairs. A contact was considered native if it was present in the minimized crystal structure.

<sup>§</sup> A total of 654 reported NOEs were used for comparison. An NOE was considered satisfied if the average  $r^{-6}$  weighted distance between closest equivalent protons was  $\leq$  5.5 Å.

#### 4.4.2 Native State Behavior in Single-Molecule Versus Test-Tube Simulations

The test-tube simulation at 298 K is in good agreement with previously reported single-molecule simulations (Mayor *et al.* 2000, McCully *et al.* 2010) as shown in Table 4.1 based on core (residues 8-53) C $\alpha$  RMSD, fraction  $\alpha$ -helix, and fraction of native contacts satisfied. Overall, the properties of the molecules in the test-tube simulation indicated that EnHD was slightly, but not significantly, less folded than in the single-molecule simulations. However, in comparison with experiment, the fraction of nuclear Overhauser effect crosspeaks (NOEs) was satisfied as well in the test-tube system as in the single-molecule simulations (85  $\pm$  3% vs. 87  $\pm$  3, respectively).

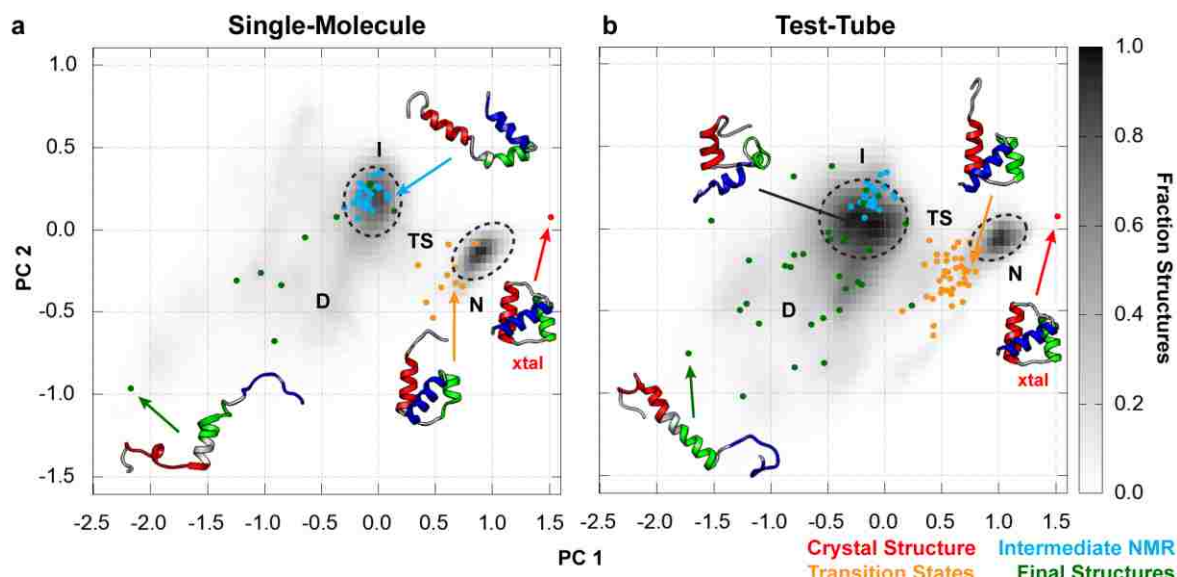


**Figure 4.5: Properties of 498 K unfolding simulations over time**

(a) C $\alpha$  RMSD of the core residues (8-53) to the crystal structure is plotted over time. (b) Percentage of residues in  $\alpha$ -helix, as defined by DSSP, is plotted over time. (c) Fraction of native residue-residue contacts (present in the minimized crystal structure) is plotted over time. On average, EnHD unfolded more and faster in the single-molecule simulations. However, by the ends of the respective simulations, EnHD unfolded about the same amount, on average.

#### 4.4.3 Effect of Intermolecular Interactions on the Unfolding Pathway of EnHD

The average and standard deviation of the core C $\alpha$  RMSD, percentage  $\alpha$ -helix, and fraction of native contacts across the 10 independent single-molecule simulations and all 32 molecules in the 498 K test-tube simulation is plotted over time (Figure 4.5). At 50 ns, on average, EnHD had unfolded less in the test-tube simulation relative to the single-molecule simulations, though not significantly. However, by the end of the test-tube simulation (175 ns), the extent of unfolding was comparable to the shorter single-molecule simulations. The standard deviation of core C $\alpha$  RMSD was larger in the test-tube simulation indicating more variation in unfolding.

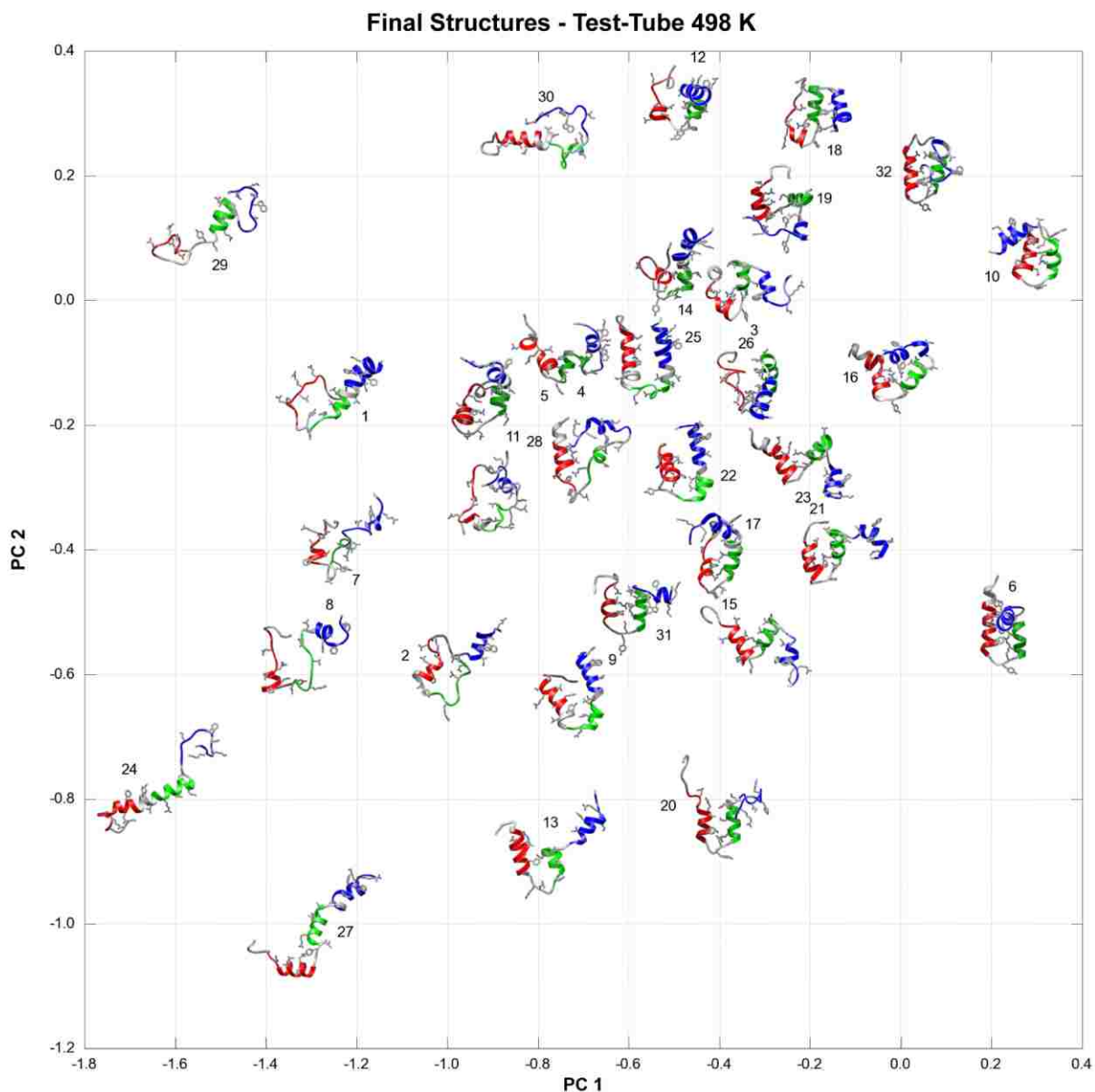


**Figure 4.6: Property space distributions**

The distributions of structures in the first and second principal components of the 39-dimensional property space are plotted for simulations at 298 and 498 K for the (a) single-molecule and (b) test-tube systems. The 298 and 498 K distributions were independently normalized to the bin with the maximum number of structures and plotted from white (0) to black (1). The crystal structure (red), transition state structures (orange), structures from the NMR ensemble of the intermediate (cyan), and final structures from the simulations (green) are additionally plotted in property space. Regions representing the native state (N), transition state (TS), folding intermediate (I), and denatured state (D) are noted. Images of several structures are shown to visualize the space. SM: crystal structure (1.51, 0.08); run 2 transition state, 0.26 ns (0.35, -0.21); L16A NMR model 1 (-0.09, 0.23); run 2, 60 ns (-2.17, 0.96). TT: crystal structure (1.51, 0.08); molecule 24 transition state, 0.435 ns (0.73, -0.26); molecule 24 intermediate, 5.670 ns (-0.10, 0.07); molecule 24 final structure, 175 ns (-1.72, -0.76).

Instead of considering many properties independently, as above, these and other properties can be combined to create a property space that describes molecular structures and folding pathways and rigorously compares different simulations. Structures that are close together when plotted in property space, or a projection thereof, have similar structural properties. General, protein-independent property spaces have been successfully employed to compare folding pathways of different molecules and identify transition states for unfolding (Kazmirski *et al.* 1999, Toofanny *et al.* 2010). A property space containing protein-specific properties has also been used previously to describe the unfolding and refolding pathway of EnHD (Beck and Daggett 2007, McCully *et al.* 2010).





**Figure 4.7: Final structures plotted in principal component space**

The structure of each of the 32 molecules from the 498 K test-tube simulation at 175 ns is plotted at its value in the first two dimensions of PC space. Proteins are colored by helix, and residues whose COM distances contribute to the property space are shown in sticks and colored by atom (Gln12, Arg15, Leu16, Glu19, Phe20, Arg24, Leu26, Leu 34, Glu37, Leu38, Leu40, Arg30, Ile45, Trp48, Phe49, Lys52, Arg53).

Here, 39 properties were included in a protein-specific property space, and a principal component analysis was performed. The properties with the highest weights in the first principal component were core  $C\alpha$  RMSD, native contacts, and COM distances between residues in HI and HII (Figure 4.2a). The first and second components accounted for 64% of the total variance (Figure 4.2b).

**Table 4.2: Test-tube transition states**

<b>Molecule Number</b>	<b>Unfolding/ Refolding</b>	<b>Time (ps)</b>	<b>Core C<math>\alpha</math> RMSD* (Å)</b>	<b># Native Contacts<sup>†</sup></b>	<b># Inter Contacts<sup>‡</sup></b>	<b>S vs. <math>\Phi</math> Correlation<sup>§</sup></b>
1	U	185	3.24	137	0	0.717
2	U	713	2.80	135	25	0.591
3	U	318	2.07	146	0	0.638
4	U	291	2.84	127	0	0.649
5	U	173	2.74	143	0	0.638
6	U	276	2.66	141	0	0.767
6	R	346	2.31	142	0	0.792
6	U	1797	3.16	128	0	0.498
7	U	299	2.81	147	0	0.681
8	U	568	2.46	146	24	0.803
9	U	1373	3.55	130	34	0.788
10	U	527	2.38	151	0	0.537
11	U	369	3.19	136	3	0.622
12	U	364	2.79	140	0	0.709
13	U	178	1.90	142	0	0.697
14	U	109	2.10	146	0	0.758
14	R	229	1.60	151	0	0.722
14	U	400	2.38	143	0	0.730
15	U	121	2.39	146	0	0.681
15	R	303	1.78	144	0	0.727
15	U	361	2.04	149	0	0.818
16	U	96	2.68	145	0	0.761
17	U	1366	3.35	140	13	0.726
18	U	375	2.42	146	0	0.457
19	U	407	2.51	149	10	0.619
20	U	632	3.71	133	0	0.831
21	U	2147	3.92	134	0	0.406
22	U	328	2.35	143	0	0.855
23	U	472	2.79	143	0	0.683
24	U	453	2.17	146	0	0.826
25	U	166	3.46	135	0	0.692
26	U	253	2.48	142	0	0.655
27	U	455	2.21	149	20	0.808
28	U	358	2.44	136	0	0.665
29	U	244	3.40	138	0	0.838
30	U	376	3.08	137	11	0.644
31	U	301	2.22	144	0	0.736
32	U	496	2.99	139	0	0.603
32	R	1375	2.56	144	0	0.638
32	U	1668	3.40	130	0	0.624
Average		532	2.68	141	4	0.691
Std Dev		495	0.55	6	8	0.103

\* C $\alpha$  RMSD of residues 8-53 to the starting structure<sup>†</sup> Number of intramolecular residue pairs in contact that were also in contact in the starting structure<sup>‡</sup> Number of atoms in contact with other molecules of EnHD<sup>§</sup> Correlation between S-values calculated for the transition state and experimentally-determined  $\Phi$ -values

Structures from all of the simulations were plotted in property space by their first and second principal components (Figure 4.6). The most unfolded structures were found in the more negative regions of the plot, and the crystal structure (red dot) is found at (1.51, 0.08). The single-molecule and test-tube simulations occupied much of the same space in this representation, with the majority of structures contributing to two peaks, denoted “N” and “I” (Figure 4.6). In the test-tube simulations, there was a tail off the native peak, indicative of the few structures that unfolded. The transition states (orange dots, “TS”), the major peak in the denatured state (I), and final structures from the 498 K simulations (green dots) are in the same regions in both the single-molecule and test-tube simulations indicating similarity between the unfolding pathways. Images of the 32 final structures from the unfolding test-tube simulation are plotted in the first two principal components in Figure 4.7 to better visualize the space.

**Table 4.3: Single-molecule transition states**

Run Number	Unfolding/ Refolding	Time (ps)	Core C $\alpha$ RMSD* ( $\text{\AA}$ )	# Native Contacts <sup>†</sup>	S vs. $\Phi$ Correlation <sup>‡</sup>
1	U	170	3.33	140	0.739
2	U	260	3.03	131	0.688
3	U	412	3.34	143	0.702
4	U	611	2.32	142	0.659
5	U	580	3.22	143	0.718
6	U	180	2.68	149	0.670
7	U	414	3.23	137	0.706
8	U	444	3.02	138	0.773
9	U	106	1.85	149	0.700
10	U	334	2.63	138	0.774
Average		351	2.87	141	0.713
Std Dev		172	0.49	5	0.039

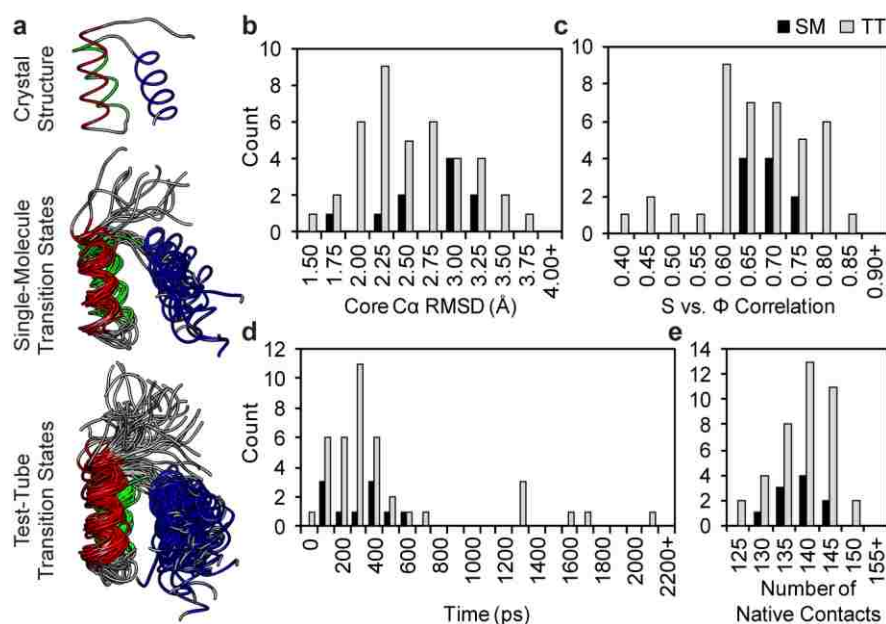
\* C $\alpha$  RMSD of residues 8-53 to the starting structure

<sup>†</sup> Number of intramolecular residue pairs in contact that were also in contact in the starting structure

<sup>‡</sup> Correlation between S-values calculated for the transition state and experimentally-determined  $\Phi$ -values

Regions in property space with high populations represent the more favorable states populated. For example, the native state (N) was represented by a peak around (0.20, 0.20). The most populated nonnative region, around (-0.25, -0.25), represented a stable nonnative state (I) that was also occupied by the NMR-derived ensemble of the folding intermediate. The intermediate was engineered to be highly populated under physiological conditions through mutation of Leu16 to Ala so that structural studies could be performed (Mayor *et al.* 2003a). The L16A mutant is globular, highly helical, and very mobile, and a solution

structure of this L16A mutant was solved by NMR (Religa *et al.* 2005). The 25 models from this NMR structure were plotted in our property space, and they fell in a highly populated region of the space characterized by high  $\alpha$ -helical content and low center-of-mass distances between the 16 key HI-HII and HIII-core residue pairs. Thus, our simulations in combination with the property space analysis correctly identified the intermediate state.



**Figure 4.8: Single-molecule vs. test-tube transition state properties**

(a) The crystal structure (1enh) as well as structures from all of the transition states in the single-molecule and test-tube simulations are colored by helix. (b-e) Histograms of properties of the transition states in the single-molecule (SM, black) and test-tube (TT, gray) simulations: (b) core Ca RMSD, (c) correlations of S- and  $\Phi$ -values, (d) time point at which the transition state occurred, and (e) number of native residue-residue intramolecular contacts made at the transition state. The single-molecule and test-tube transition states have good agreement with each other as well as with experiment.

Transition state (TS) ensembles for the high temperature unfolding simulations were identified for each molecule in the test-tube simulation (Table 4.2) and the 8 new single-molecule simulations (in addition to the 2 previously reported; Mayor *et al.* 2000; Structures from all of the simulations were plotted in property space by their first and second principal components (Figure 4.6). The most unfolded structures were found in the more negative regions of the plot, and the crystal structure (red dot) is found at (1.51, 0.08). The single-molecule and test-tube simulations occupied much of the same space in this representation, with the majority of structures contributing to two peaks, denoted “N” and “I” (Figure 4.6). In the test-tube simulations, there was a tail off the native peak, indicative of the few

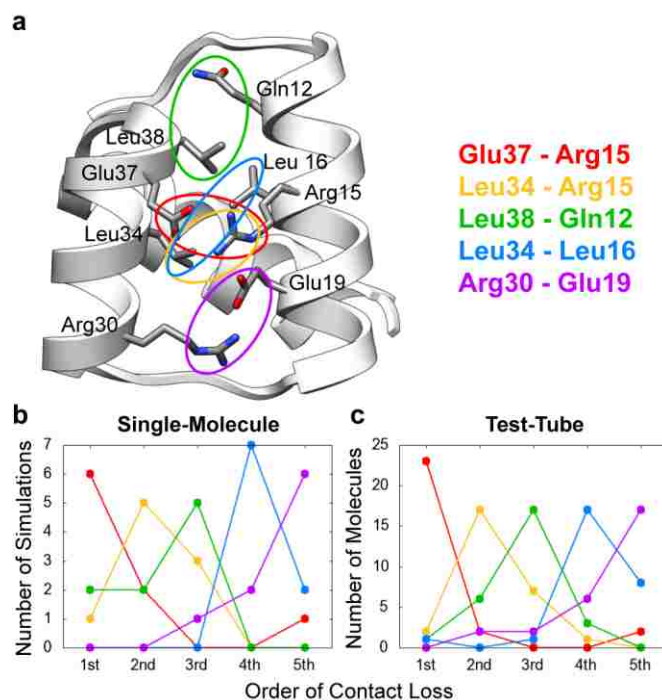
structures that unfolded. The transition states (orange dots, “TS”), the major peak in the denatured state (I), and final structures from the 498 K simulations (green dots) are in the same regions in both the single-molecule and test-tube simulations indicating similarity between the unfolding pathways. Images of the 32 final structures from the unfolding test-tube simulation are plotted in the first two principal components in Figure 4.7 to better visualize the space.

Table 4.3) using multidimensional scaling of the  $C\alpha$  RMSD matrix, as described previously (Li and Daggett 1994) and in the Methods Section. A total of 40 TS ensembles were identified for the 32 molecules in the test-tube simulation, as 4 molecules experienced transient refolding events. Past work has demonstrated that the transition state for EnHD is characterized by HIII pulling away from the HI-HII scaffold and exposing the hydrophobic core (Mayor *et al.* 2000, Gianni *et al.* 2003, Mayor *et al.* 2003b), and this is indeed the case for the transition states described here (Figure 4.8a).

Distributions of the core  $C\alpha$  RMSD, time when the TS occurred, and the number of native contacts present in the TS were very similar for EnHD in the single-molecule vs. test-tube simulations (Figure 4.8b,d,e). There were a few late TSs in the test-tube simulations, which is consistent with the slower kinetics observed for the three properties illustrated in Figure 4.8. A semiquantitative structure index (S-value) was calculated for each of the TS ensembles as the ratio of the fraction of native secondary structure and the fraction of contacts present in the TS relative to the native state. S-values can be compared with experimental  $\Phi$ -values, which have been measured for EnHD (Gianni *et al.* 2003).  $\Phi$ -values indicate the amount of structure in a given residue during the transition state, where a value of 0 suggests denatured-like extent of structure and 1 suggests native-like extent of structure. Correlations between S- and  $\Phi$ -values were good for all 50 transition states, with the majority falling within  $R = 0.60 - 0.85$  (Figure 4.8c).

The order that the five key contacts between HI and HII were gained upon refolding is generally consistent in temperature quenched simulations of EnHD (McCully *et al.* 2010). In refolding, the contacts form in the order: Arg30 – Glu19, Leu34 – Leu16, Glu37 – Arg15 and Leu34 – Arg15, then Leu38 – Gln12. In both the single-molecule and multi-molecule unfolding simulations presented here, the five contacts were usually lost in the same order:

Glu37 – Arg15, Leu34 – Arg15, Leu38 – Gln12, Leu34 – Leu16, then Arg30 – Glu19 (Figure 4.9). The last two contacts lost upon unfolding (Leu34 – Leu16 and Arg30 – Glu19) were the same two that were gained first in refolding, in opposite order. Notably the Arg 30 – Glu 19 contact was present in the starting structure for the refolding simulations, so it was always the first contact to form. The first three contacts lost in unfolding (Glu37 – Arg15, Leu34 – Arg15, Leu38 – Gln12) were gained last in refolding. However, they were gained in the same order they were lost, rather than the opposite. Leu38 – Gln12 was consistently the last and least likely contact to reform in the refolding simulations, but it was usually lost third in these unfolding simulations. Curiously, there was not a single simulation or individual molecule in which the five contacts were lost in the exact reverse order that they were consistently gained in the quenched refolding simulations.



**Figure 4.9: Single-molecule vs. test-tube HI-HII contact loss**

(a) Five contact pairs between HI-HII, circled by color: Glu37 – Arg15 (red), Leu34 – Arg15 (orange), Leu38 – Gln12 (green), Leu34 – Leu 16 (cyan), Arg30 – Glu19 (purple). (b) Number of times each contact pair was lost, 1st, 2nd, 3rd, 4th, and 5th in the 9 single-molecule simulations where all five contact pairs were lost. (c) Number of times each contact pair was lost in the 27 molecules in the test-tube simulations where all five contact pairs were lost.

## 4.5 Conclusions

Here we present two test-tube simulations that probe the interactions between molecules in the native state and during thermal unfolding. EnHD formed clusters to varying extents in the multimolecule, “test-tube” simulations at both low and high temperature. Hydrophobic packing interactions that were lost upon unfolding were replaced with intermolecular protein-protein hydrophobic contacts in the high-temperature test-tube simulation. EnHD gained protein-protein hydrogen bonds in the test-tube simulations while losing such interactions with water. Yet, there were overall many fewer contacts made between protein molecules than within, and most of the intermolecular interactions were with water rather than with other proteins. While the molecules were largely interacting with each other in the test-tube simulations, it was not at the total exclusion of water.

The unfolding pathway was largely unaffected by the presence of neighboring protein molecules, though it was moderately slowed down. The structures from the single-molecule and test-tube simulations occupied the same regions of property space with similar distributions. TS ensembles agreed well with each other, as well as with experiment, based on several individual properties. The 39-dimensional property space correctly identified the folding intermediate as compared with structures from the NMR ensemble. The order of contact loss between HI and HII was consistent between the single-molecule and test-tube simulations. However, it did not precisely agree with the order these same contacts were gained in refolding simulations performed previously.

Despite the fact that MD is typically a single-molecule technique, it consistently reproduces ensemble experimental measurements. Here, we created a small ensemble, though still many orders of magnitude smaller than the number of molecules probed by experimental methods, and we obtained good agreement with experimental NOEs in the native state, correlation between  $S$ - and  $\Phi$ -values for transition states, and overlap in property space with the folding intermediate. The native state behavior and unfolding pathways were remarkably similar in the single-molecule and test-tube simulations despite the high degree of aggregation observed, particularly at high temperature. While neighboring molecules did not perturb the unfolding pathway, they did alter the kinetics, slowing down the process. These test-tube simulations provide insight into the nature of interactions in protein

80

aggregates, showing hydrophobic aggregation through both folded and unfolded segments of the structure.



## Chapter 5: Promiscuous contacts and heightened dynamics increase thermostability in an engineered variant of the Engrailed Homeodomain

### 5.1 Summary

A thermostabilized variant (UVF) of the Engrailed Homeodomain (EnHD) was previously engineered by Mayo and coworkers (Gillespie *et al.* 2003, Shah *et al.* 2007). The melting temperature of the nonnatural, designed protein is 50 °C higher than the natural wild-type protein (> 99 vs. 52 °C), and the two proteins share 22% sequence identity. We have performed extensive (1  $\mu$ s) all-atom, explicit solvent molecular dynamics simulations of the wild type and engineered proteins to investigate their structural and dynamic properties at room temperature and at 100 °C. Our simulations are in good agreement with NMR data available for the two proteins (NOEs, J-coupling constants, and order parameters for EnHD; and NOEs for UVF) and show we reproduce the backbone dynamics and side chain packing in the native state of both proteins. UVF was more dynamic at room temperature than EnHD, with respect to both its backbone and side chain motion. When the temperature was raised, the thermostable protein maintained this mobility while retaining its native conformation. EnHD, on the other hand, was unable to maintain its more rigid native structure at higher temperature and began to unfold. Heightened protein dynamics leading to promiscuous and dynamically interchangeable amino acid contacts made UVF more tolerant to increasing temperature, providing a molecular explanation for heightened thermostability of this protein.

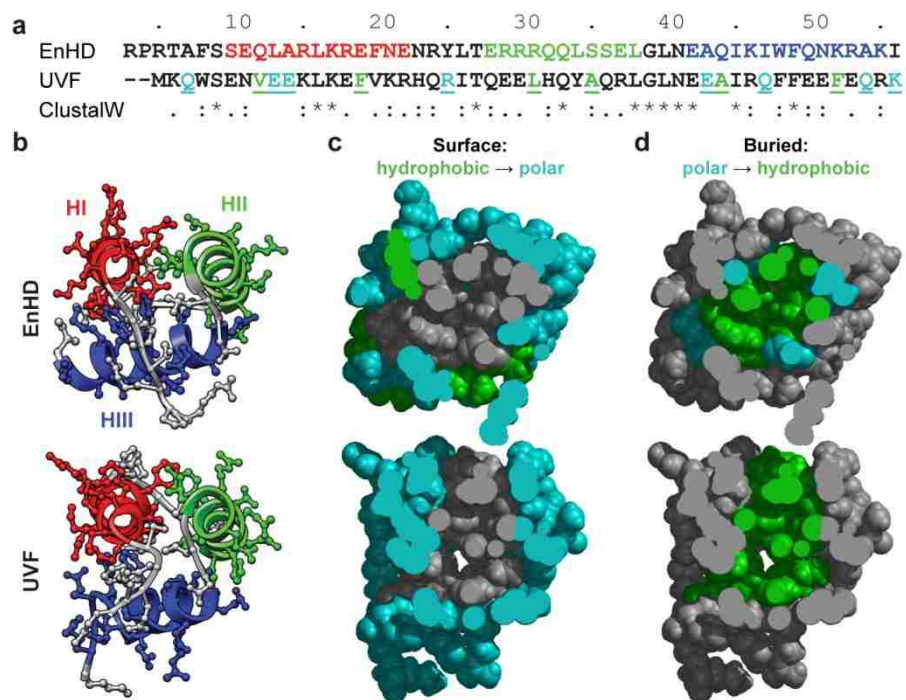
### 5.2 Introduction

The Engrailed Homeodomain (EnHD) has been a popular target for protein folding studies, both experimental and computational, due to its ultrafast folding and unfolding times as well as its low melting temperature and  $\Delta G_{D-N}$  ( $T_m = 52$  °C,  $\Delta G_{D-N} = 1.7$  kcal/mol; Mayor *et al.* 2000, Mayor *et al.* 2003b, Gianni *et al.* 2003, DeMarco *et al.* 2004, McCully *et al.*

2008, McCully *et al.* 2010). In addition, its native state structure and dynamics have been characterized extensively (Mayor *et al.* 2003a, McCully *et al.* 2008, McCully *et al.* 2010, Religa 2008). Over the past decade, this protein has been re-engineered to increase its thermostability by the Mayo group using full-sequence computational design (Marshall and Mayo 2001, Marshall *et al.* 2002, Gillespie *et al.* 2003, Shah *et al.* 2007). They used a limited amino acid library that selected only hydrophobic amino acids (Ala, Val, Leu, Ile, Phe, Tyr, Trp) for buried residues, polar and charged amino acids for the surface, and statistically probable amino acids as helix-capping residues. In 2007, the group created UVF, which is very stable with a  $T_m > 99$  °C and  $\Delta G_{D-N} = 4.2$  kcal/mol (Shah *et al.* 2007) with 22% sequence identity to EnHD (Figure 5.1). Most of the mutations relative to EnHD remove polar residues from the buried regions and hydrophobic residues from the surface. An NMR structure of UVF along with a set of nuclear Overhauser effect crosspeaks (NOEs) was published simultaneously (Shah *et al.* 2007).

This pair of proteins provides an excellent opportunity to investigate the detailed structural and dynamic bases for the thermostability of such engineered proteins and thermostabilized proteins in general. Differences in the sequences hint at reasons for increased thermostability: hydrophobic residues in the core and not on the surface increase the strength of the hydrophobic effect, and electrostatic interactions on the surface and at helix-capping residues further increase stability (Marshall and Mayo 2001, Marshall *et al.* 2002). Mayo and coworkers attributed the increased melting temperature to decreasing the overall charge of the protein from +7 to 0 and to incorporating more favorable electrostatic interactions. These new interactions included additional salt bridges between surface side chains as well as side chain – backbone interactions that stabilized helix-dipoles and capped the helices. Yet, a full description of the protein's dynamics at room temperature and how the thermostabilized protein maintains a folded structure at higher temperatures is lacking. To that end, we have performed all-atom, explicit solvent molecular dynamics (MD) simulations of EnHD and UVF at 25 and 100 °C totaling 1  $\mu$ s of simulation time. In addition, we have compared our simulations with NMR observables where they are available: NOEs, J-coupling constants, and backbone amide  $S^2$  order parameters for EnHD; and NOEs for UVF. Our simulations are in good agreement with the available NMR data for

both proteins, validating the proteins' backbone movement and tertiary interactions in our simulations.



**Figure 5.1: Sequence and structure of EnHD and UVF**

(a) Alignment and ClustalW score for the sequences of both proteins. The EnHD sequence is colored by helix (HI: 10-22 red; HII: 28-38 green; HIII: 42-53 blue). Hydrophobic residues in EnHD that were mutated to polar residues in UVF are colored in cyan on the UVF sequence; polar to hydrophobic mutations are colored green. (b) Experimental structures (EnHD: 1enh crystal structure, top; UVF: 2p6j NMR structure, bottom) are shown colored by helix with non-hydrogen side chain atoms shown as ball and sticks. (c) Surface residues in EnHD (top) and UVF (bottom) are colored as hydrophobic (Ala, Val, Leu, Ile, Phe, Tyr, Trp; green) or polar (cyan). Buried residues (8, 12, 16, 19, 20, 26, 31, 34, 35, 38, 40, 44, 45, 48, 49, 52) are colored gray. (d) Buried residues in EnHD (top) and UVF (bottom) are colored by polar (cyan) and hydrophobic (green). Surface residues are colored gray. UVF was designed to remove all hydrophobic residues from the surface and polar residues form the core. Hydrophobic/polar residues are more segregated between buried and surface residues in UVF, and the hydrophobic core is more loosely packed.

Our native simulations of the two proteins indicate that UVF is more dynamic at room temperature than EnHD. When the temperature was increased to 100 °C, EnHD became more mobile while the engineered protein maintained dynamics on par with what was observed at room temperature. The increased dynamics in EnHD at the higher temperature caused it to lose many of its native core contacts and unfold. In contrast, the thermostabilized protein maintained its more nonspecific and dynamic core interactions with increased temperature, and the structure easily tolerated the increase. In a sense, the more

promiscuous and dynamic contacts in the designed protein better absorbed energy, leading to a higher effective heat capacity and thermostability.

## 5.3 Methods

### 5.3.1 Simulation Protocol

All simulations were performed using our in-house molecular dynamics software, *ilmm* molecular mechanics (*ilmm*; Beck *et al.* 2000-2012), with the Levitt *et al.* (1995) force field. The starting structure for simulations of EnHD was the crystal structure (PDB id: 1enh), which was determined to a resolution of 2.1 Å and includes residues 3-56 (Clarke *et al.* 1994). A set of native simulations were also begun from model 1 of the NMR ensemble of EnHD (PDB id: 2jwv; Religa 2008) to control for any differences due to crystal vs. NMR structures. The simulations of UVF began with the first model of the NMR ensemble (PDB id: 2p6j; Shah *et al.* 2007). The UVF structure were renumbered to align with EnHD's PDB numbering, and they are discussed using residue numbers 5-56 instead of 1-52 to avoid confusion.

Each system was prepared as per our standard protocols (Beck and Daggett 2004) using a water box that extended 10 Å past the edge of the protein on all sides filled with flexible, explicit water (Levitt *et al.* 1997) at a density consistent with the experimentally determined liquid-vapor coexistence curve (25 °C: 0.997 g/mL; 100 °C: 0.958 g/mL; Kell 1967). Energies and forces were calculated using 2 fs timesteps. The NVE (constant number of particles, volume, and total energy) microcanonical ensemble was employed using periodic boundary conditions. An 8 Å force-shifted cutoff was used for nonbonded interactions (Beck *et al.* 2005), and the nonbonded pair list was updated every 2 steps. Structures were saved every 1 ps for analysis, and 5 independent simulations were performed for 50 ns each for both EnHD and UVF at 25 and 100 °C. The EnHD NMR control structure was only simulated at 25 °C (5 50-ns simulations), and those simulations will not be discussed in depth. Additional control simulations at 225 °C verified that the thermostabilized protein did indeed unfold at very high temperature, but they are not discussed here. The simulation time totals 1 μs and resulted in 1,000,000 structures for detailed analysis.

### 5.3.2 Calculation of NMR Comparables

NMR experiments were performed on a 61-residue construct of EnHD (residues -1 to 59) and deposited in the Biological Magnetic Resonance Data Bank (BMRB, accession number 15536 and block id 276091 for PDB 2jwv; Religa 2008). This dataset contains 45  $^3J_{\text{HNH}\alpha}$  coupling constants, 58  $S^2$  N-H order parameters, and 675 NOEs. The coupling constants and NOEs were measured at 5 °C and the order parameters at 25 °C. In addition, 1151 NOEs measured at 20 °C were found for UVF in the BMRB's Filtered Restraints Database (BMRB block ids 449656 and 449657; Doreleijers *et al.*, 2005, Shah *et al.* 2007). Similar values were calculated from the atomic coordinate data in our simulations to compare with these experimental observables.

J-coupling constants were calculated from backbone  $\phi$  angles using the Karplus equation (Karplus 1959, Beck *et al.* 2008).  $^3J_{\text{HNH}\alpha}$  was calculated at each step and averaged over the simulation for each residue to obtain  $\langle ^3J_{\text{HNH}\alpha} \rangle$ . Experimental coupling constant data were available for residues 10-55 of EnHD except Glu37.

The  $S^2$  order parameters were calculated from the MD simulations as the autocorrelation of the backbone N-H bond vector over a sliding, finite time window (Lipari and Szabo 1982, Levitt 1983, Wong and Daggett 1998). We calculated  $S^2$  for EnHD using a sliding window of 10 ns. Before doing the  $S^2$  calculation, all structures were aligned on the backbone atoms (N, C $\alpha$ , C, O) of the core (residues 8-53) to remove rotational and translational motion. Experimental amide  $S^2$  order parameters were available for all residues in our construct, so we compared  $S^2$  as measured from our simulations to experiment for residues 3 to 56, with the exception of Pro 4 (for which there is no N-H bond).

A total of 675 NOEs were deposited for EnHD, 654 of which correspond to residues in our construct, and 1151 NOEs were reported for UVF. We calculated the distance between the closest equivalent protons in the NOE using an  $\langle r^{-6} \rangle$  weighted distance. If this distance was less than 5.5 Å, which was the longest cutoff published for the EnHD experimental set, the NOE was considered satisfied. For UVF, the NOE was considered satisfied if  $r$  was  $\leq 5.5$  Å or the  $r_{\text{far}}$  value of the NOE from the experimentally derived NOE list, whichever was longer (the largest  $r_{\text{far}}$  was 6 Å). To calculate NOEs for the crystal

structure of EnHD, hydrogen atoms were added and minimized for 100 steps using steepest descent minimization.

### 5.3.3 Simulation Analysis

The root-mean-square deviation of the C $\alpha$  atoms (C $\alpha$  RMSD) to the simulation starting structure (minimized crystal structure for EnHD or NMR structure for UVF) was calculated for the core residues (8-53) of all proteins. The calculation was limited to the core residues because the N- and C-termini have large movements that are not representative of the dynamics of the structured region of the proteins. In addition, this truncation allows us to directly compare the RMSD between both proteins, as their number of residues differs. The C $\alpha$  root-mean-square fluctuation (C $\alpha$  RMSF) about the mean structure over time was also calculated for the core residues. The mean structure for each protein was calculated by averaging the coordinates for each core C $\alpha$  atom across simulation time.

Contacts were counted and classified in several different ways. First, contacts were defined as native or nonnative based on whether they occurred in the starting structure. They were also classified based on whether the contacting atoms were in the main chain (N, C $\alpha$ , C, O) or side chain. Atoms were considered in contact for carbons that were  $\leq 5.4$  Å apart or non-carbon atoms that were  $\leq 4.6$  Å apart. Hydrogen atoms were not considered, nor were interactions within a residue or between neighboring residues. Residues were considered in contact if they contained at least one contacting atom pair.

Next, side chain – side chain contacts were further classified as making hydrogen bond, hydrophobic, or “other” interactions. Three atoms were defined as being in a hydrogen bond if the hydrogen and acceptor atoms were  $< 2.6$  Å, the donor-hydrogen-acceptor angle was within  $45^\circ$  of linearity, the charges on the donor and acceptor were  $< -0.3$ , and the charge on the hydrogen was  $> +0.3$ . Two carbons were defined as participating in a hydrophobic interaction if they were  $< 5.4$  Å apart and each carbon was bound to at least one hydrogen atom. If two non-hydrogen atoms that didn't satisfy the previous two contact types were  $< 4.6$  Å apart, they were defined as making a nonspecific, “other” interaction. Finally, the core residues (8-53) were further classified by whether they were buried (8, 12, 16, 19, 20, 26, 31, 34, 35, 38, 40, 44, 45, 48, 49, 52) or on the surface (9-11, 13-15, 17, 18, 21-25, 27-30, 32, 33, 36, 37, 39, 41-43, 46, 47, 50, 51, 53).

## 5.4 Results

### 5.4.1 Validation of Simulations: Comparison with NMR Observables

#### 5.4.1.1 J-Coupling Constants

$^3J_{\text{HNH}\alpha}$  coupling constants averaged over all 5 native simulations of EnHD at 25 °C were reported in Figure 5.2a, and the correlations were reported in Table 5.1. J-coupling constants calculated from the crystal structure and NMR ensemble (25 structures) were also plotted (Figure 5.2a). Helical residues (10-22, 28-38, 42-55) had the best agreement with experiment and were consistently  $< 6.0$  Hz, as expected for helices (Pardi *et al.* 1984). The correlations ranged from 0.56 to 0.70, and the average over all 5 simulations was 0.67. Although some correlations were low, the RMSD was always within 1.5 Hz, which is the error range for such experiments and the conversion of backbone  $\phi$  angles from MD to a coupling constant using the Karplus equation.

**Table 5.1: EnHD comparison between simulation and experiment**

Temp (K)	Run	Sim Length (ns)	$^3J_{\text{HNH}\alpha}$ <sup>†</sup>		$S^2$ (10 ns) <sup>‡</sup>		NOEs <sup>§</sup>	
			Correlation	RMSD (Hz)	Correlation	RMSD	% NOEs Satisfied	Mean Viol Dist (Å)
298	1	50	0.59	1.32	0.86	0.08	83.8	0.90
	2	50	0.70	1.17	0.83	0.10	87.6	1.10
	3	50	0.61	1.28	0.91	0.07	92.0	0.45
	4	50	0.65	1.23	0.80	0.10	91.1	0.82
	5	50	0.56	1.35	0.83	0.09	86.2	1.01
	All	250	0.67	1.21	0.87	0.08	92.0	0.65
1enh (Crystal)	1*		0.83	0.90			96.8	0.43
2jw (NMR)	25*		0.94	0.55			98.8	0.19

\* Number of structures, not simulation length.

<sup>†</sup>  $^3J_{\text{HNH}\alpha}$  coupling constants were calculated based on the backbone  $\phi$  angle and the Karplus Equation, and the correlation and RMSD to the experimental values are reported.

<sup>‡</sup>  $S^2$  order parameters were calculated using a sliding 10-ns window, and the correlation and RMSD to the experimental values are reported.  $S^2$  was not calculated for 1enh and 2jw because there were not enough structures.

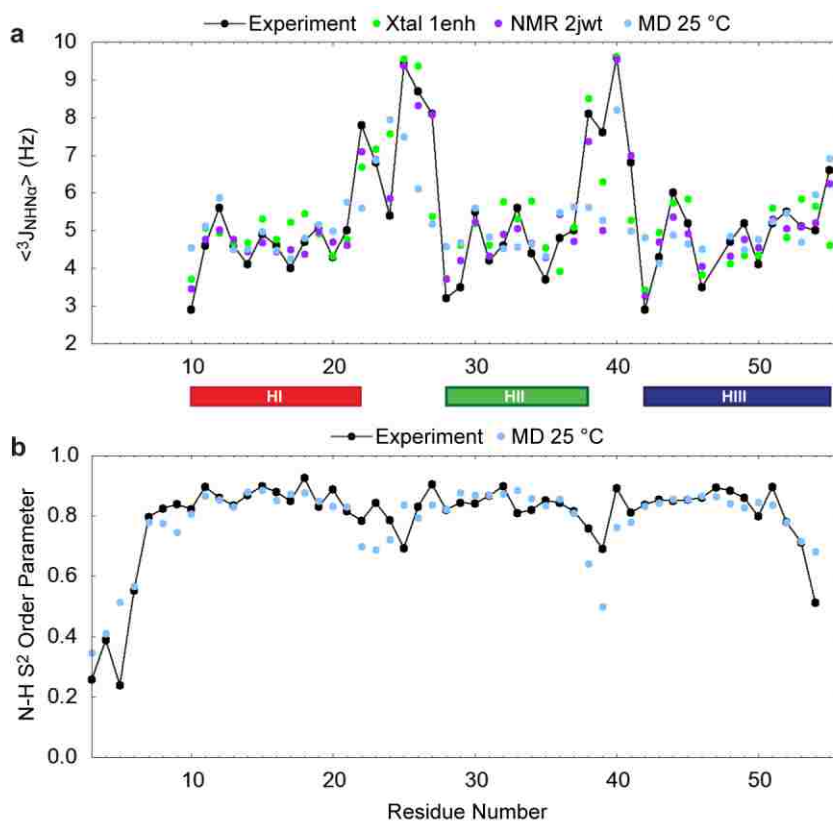
<sup>§</sup> An NOE was considered satisfied if the  $\langle r^{-6} \rangle$  distance between closest equivalent protons was  $\leq 5.5$  Å. For those that were not satisfied, the mean violation distance is reported.

The residues with the worst agreement were Arg24, Tyr25, Leu26, and Thr27 in the HI-HII loop. This was due to Leu26 pointing out past HIII rather than into the core of the protein. Its rotation disrupted the  $\phi$  angles of the surrounding residues. Leu38, Gly39, and Asn41 in the HII-HIII loop also had poor agreement with experiment. In most simulations, Asn41 alternated between  $\phi \approx +70^\circ$  and  $-50^\circ$ . In simulations with the best agreement, Asn41

spent the most time with  $\phi \approx +70^\circ$ , which corresponds to values of  ${}^3J_{\text{HNH}\alpha}$  near the experimental value of 6.8 Hz. However, in several simulations, Asn41 favored  $\phi \approx +140^\circ$  where J is very low ( $\approx 2$ ).

#### 5.4.1.2 Order Parameters

The  $S^2$  order parameter describes the amount of motion of a particular bond vector where values range from 0 (no order) to 1 (completely rigid).  $S^2$  values for the backbone amide bonds are shown in Figure 5.2b as calculated for all 5 native EnHD simulations using a 10-ns window. Correlations with experiment for each simulation and for all simulations at 25 °C are presented in Table 5.1. Experimentally derived  $S^2$  values are available for all residues in our construct, with the exception of Pro4.



**Figure 5.2:**  ${}^3J_{\text{HNH}\alpha}$ -coupling constants and  $S^2$  amide order parameters for EnHD

(a) Coupling constants are plotted as measured experimentally (black) and as calculated for the crystal structure (green), NMR structures (purple), and MD simulations at 25 °C (blue) for residues 10-55, except for Glu37 for which no experimental data were available. Agreement is best for helical areas (residues 10-22, 28-38, 42-55), and correlations over all 45 residues were 0.83, 0.94, and 0.67 for the crystal structure, NMR structures, and MD, respectively. (b) Order parameters are plotted for residues 3-56, with the exception Pro 4. Helical regions have very good agreement with experiment. The correlation between experiment and the MD simulations for the 53 residues was 0.87.



Correlations with experiment ranged from 0.80-0.91 for the 5 simulations independently, and the correlation was 0.87 when they were pooled together. The three helices and the HI-HII loop had the best agreement with experiment. The N- and C-termini as well as the HII-HIII loop deviated from the experimental values. Since our construct is missing four residues on the N-terminus (-1 to 2) and three on the C-terminus (57 to 59), it is reasonable to suspect that these missing residues affect the motions of the termini. Residues in the HII-HIII loop, particularly Leu38 and Gly39, were less rigid in our simulations than their experimental values indicated.

#### 5.4.1.3 Nuclear Overhauser Effect Crosspeaks

The percentage of NOEs that were satisfied by our simulations of EnHD and the average violation distances for those that were not were calculated (Table 5.1). When the 5 native simulations were all pooled together, 92% of the 654 reported NOEs for our construct were satisfied. The mean violation distance for the NOEs that were not satisfied was 0.65 Å.

**Table 5.2: UVF comparison between simulation and experiment**

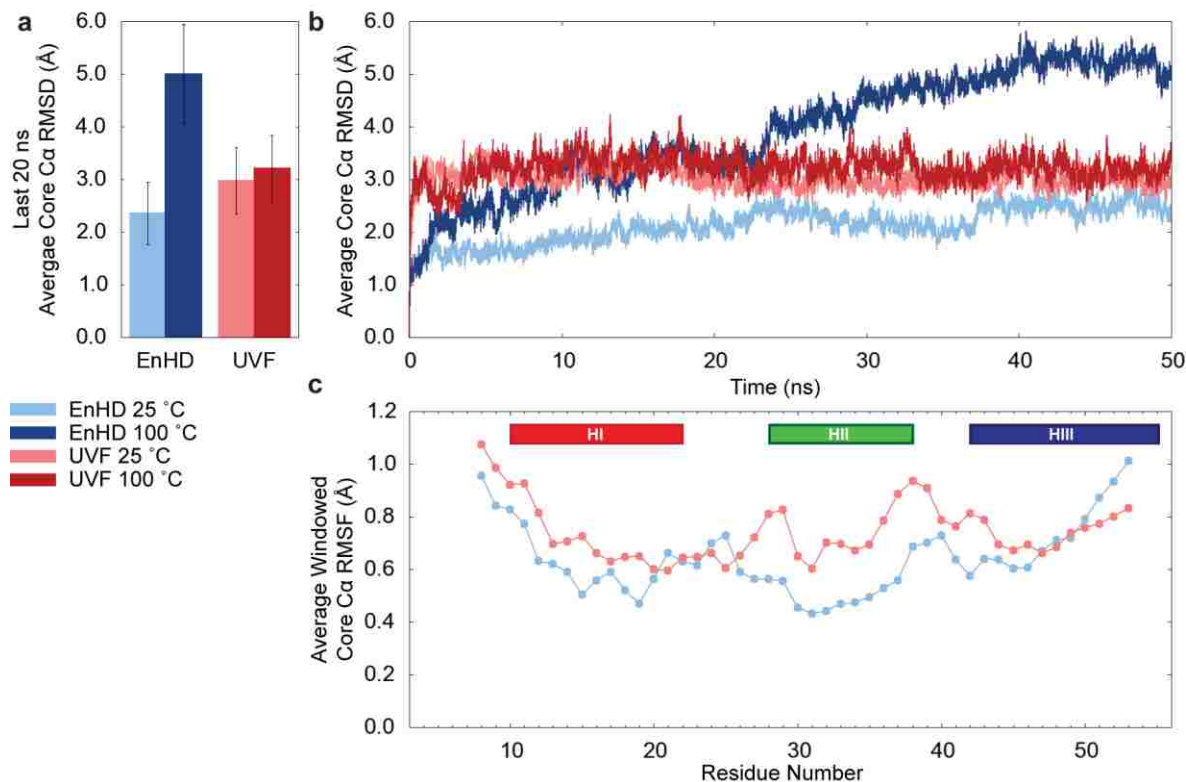
Temp (K)	Run	Sim Length (ns)	NOEs <sup>†</sup>	
			% NOEs Satisfied	Mean Viol Dist (Å)
298	1	50	91.1	1.99
	2	50	88.0	2.10
	3	50	93.3	1.45
	4	50	94.2	1.15
	5	50	92.4	1.84
	All	250	95.0	1.25
373	1	50	92.7	1.48
	2	50	92.9	1.51
	3	50	93.4	1.35
	4	50	92.7	1.81
	5	50	94.6	1.41
	All	250	95.1	1.41
All Simulations		500	95.8	1.31
2p6j (NMR)		25*	97.0	0.47

\* Number of structures, not simulation length.

<sup>†</sup> An NOE was considered satisfied if the  $\langle r^{-6} \rangle$  distance between closest equivalent protons was  $\leq 5.5$  Å or the distance specified in by the restraint, which ever was farther. For those that were not satisfied, the mean violation distance is reported.

There were only three individual NOEs that were violated by  $> 1$  Å in all 5 simulations of EnHD: Glu19 H $\gamma$  – Phe49 H $\zeta$ , Leu26 H $\delta$  – Trp48 H $\epsilon$ , and Arg53 H – Lys55 H $\epsilon$ . The core packing was looser in our simulations than in the crystal structure or NMR ensemble, and Phe49, in particular, flipped around in the hydrophobic core, sometimes

moving farther away from Glu19 than in the crystal or NMR structures. The Leu26 – Trp48 NOE was not satisfied in the crystal structure either due to a flipped orientation of Leu26, which is maintained in the simulations. Similarly, Lys55 adopts a rotamer that tilts the terminal amine toward the backbone of Arg53 in the crystal and NMR structures without making any direct contacts, but it was very dynamic in the simulations and that orientation was not maintained.

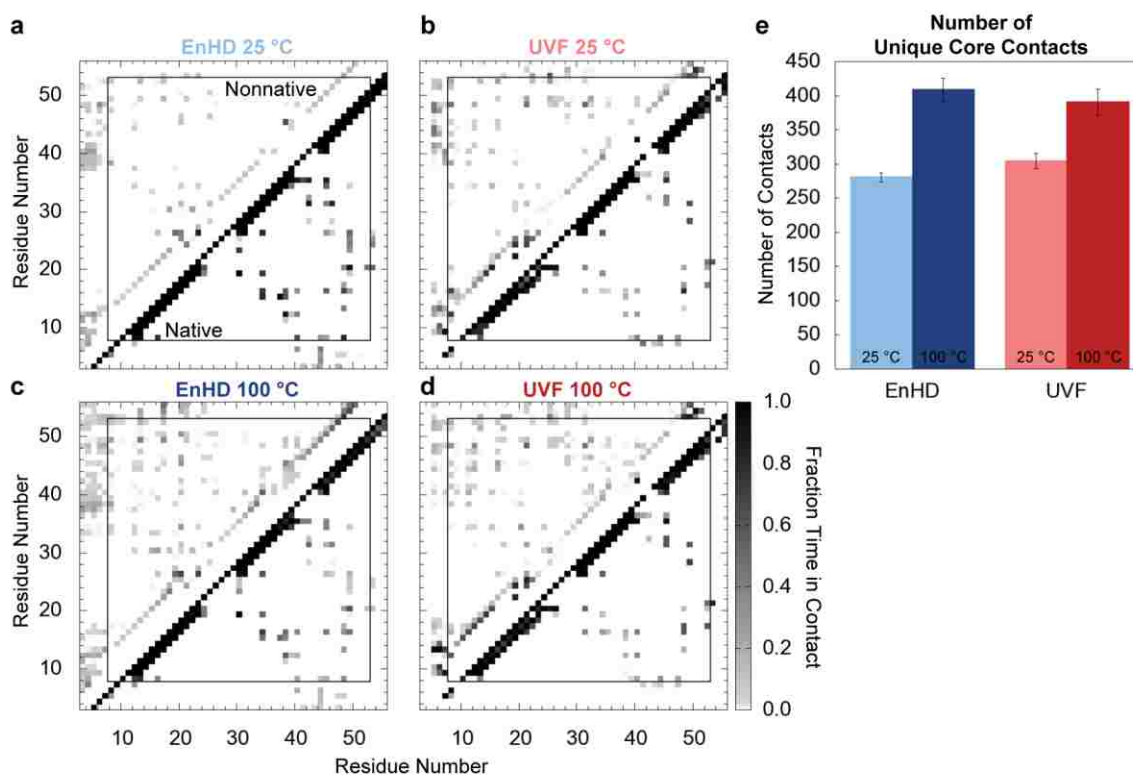


**Figure 5.3: Core Ca RMSD and RMSF at 25 and 100 °C**

(a) The average Ca RMSD of the core residues (8-53) over the last 20 ns of each of the five independent simulations is plotted for EnHD (blue) and UVF (red) at 25 (lighter) and 100 °C (darker) with the standard deviation denoted by error bars. (b) The core Ca RMSD was averaged at each time point over the five independent simulations and is plotted over time for both proteins at both temperatures. The RMSD was higher for UVF at 25 °C, relative to EnHD. When the temperature was raised to 100 °C, the RMSD increased for EnHD as it unfolded. However in the case of the thermostabilized proteins, there was no change in RMSD when the temperature was raised. (c) The core Ca RMSF is plotted for each residue in EnHD and UVF at 25 °C. The RMSF was lower for EnHD than UVF in HI and HIII indicating less fluctuation in the backbone.

NOE satisfaction was also calculated for UVF and is reported in Table 5.2. NOE satisfaction was 95% at both 25 and 100 °C, and all 500 ns of simulation time pooled together had nearly as many NOEs satisfied as the NMR ensemble (96 vs. 97%). Residues Glu14, Glu18, Arg22, and Glu42 had no long-range (between residues with a sequence

separation of  $\geq 5$ ) NOEs satisfied by simulation or in the NMR ensemble. Gly39 had 0 of 8 long-range NOEs satisfied by simulation but all 8 were satisfied in the NMR ensemble. All of these residues are found on the surface of UVF. In our simulations, Glu18 and Arg22 often formed a salt bridge with each other, Glu14 and Glu42 interacted with solvent, and Gly39 was loosely part of HII.



**Figure 5.4: Native, nonnative, and unique contacts for EnHD and UVF**

Fraction time in contact for the 25 °C simulations of (a) EnHD and (b) UVF and the 100 °C simulations of (c) EnHD and (d) UVF. The box indicates the core residues (8-53). Native contacts (present in the starting structure) are plotted below the diagonal and nonnative contacts above. Fraction time in contact ranges from 0.0 (white) to 1.0 (black). (e) The total number of unique core residue-residue contacts for EnHD (blue) and UVF (red) at 25 (lighter) and 100 °C (darker) with error bars indicating the standard deviation across simulations. UVF had more unique residue-residue contacts than EnHD at 25 °C, particularly nonnative ones. When the temperature was raised, EnHD picked up more contacts in nonnative regions of the contact plot while UVF increased the number of contacts without losing its overall tertiary structure.

## 5.4.2 Dynamics of EnHD vs. UVF

### 5.4.2.1 Backbone Motion

At room temperature (25 °C), EnHD had a lower core (residues 8-53) C $\alpha$  RMSD ( $2.3 \pm 0.6$  (s.d.) Å) than its thermostabilized counterpart ( $3.3 \pm 0.7$  Å) over the last 20 ns of the

simulations (Figure 5.3a,b). When considering all 50 ns, the core C $\alpha$  RMSD was  $2.1 \pm 0.6$  vs.  $3.0 \pm 0.6$  Å for EnHD and UVF, respectively. UVF was more mobile at room temperature, with larger rearrangements of the three helices. HI and HII did not move apart, but they did sometimes tilt relative to one another, and HIII slid across helices I and II. In EnHD, there was a loosening of the structure relative to the crystal structure; however, all three helices held their original conformations with HI and HII parallel and HIII docked across them. Notably, this difference was not simply due to crystal vs. NMR starting structures; when simulations starting from the EnHD's NMR structure were performed, the core C $\alpha$  RMSD was  $2.3 \pm 0.6$  Å for 5 50-ns simulations and  $2.5 \pm 0.5$  Å for the final 20 ns of the same simulations.. In the last 20 ns of the simulations at 100 °C, the core C $\alpha$  RMSD of EnHD increased to  $5.0 \pm 0.9$  Å. Not only did the core C $\alpha$  RMSD itself increase, but so did its standard deviation, indicative of more structural heterogeneity at the higher temperature. However, the core C $\alpha$  RMSD of UVF was unchanged with a 75 °C increase in temperature.

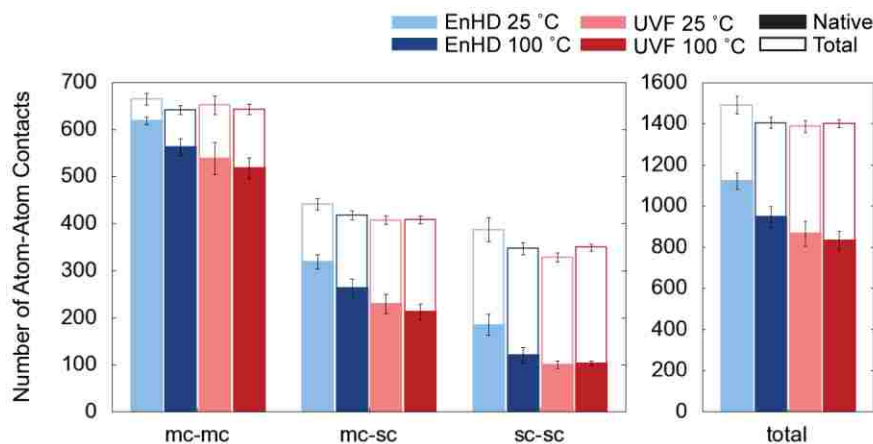
The C $\alpha$  RMSF was calculated for the core residues over all simulations at 25 °C (Figure 5.3c). Again, residues in EnHD had lower fluctuations about the mean structure than those in UVF, indicative of a more rigid structure. As expected, the most rigid regions were the three helices, where the amide atoms are involved in hydrogen bonding.

#### 5.4.2.2 Core Contacts Analysis

Native (present in the simulation starting structure) and nonnative contacts in the 25 and 100 °C simulations of EnHD and UVF are plotted as maps in Figure 5.4a-d, with native contacts below the diagonal and nonnative above. The three helices are apparent along the diagonal with  $i \rightarrow i + 2$ ,  $i + 3$ , and  $i + 4$  contacts present during the entire simulation. The off-diagonal contacts are indicative of the HI-HII interactions and the interactions of HIII with HII and the C-terminus of HI. While there were more nonnative contacts present in UVF at both temperatures, these contacts were present in the same regions as the native contacts. This consistency suggests there was no loss of overall structure, but rather there were local rearrangements in a fluid core.

At 100 °C, EnHD made many nonnative contacts, but they were smeared all over the protein, indicative of nonnative tertiary structure and unfolding. Though EnHD had more contacts on average than UVF (Figure 5.5), it made fewer unique contacts (Figure 5.4e).

This difference suggests that UVF made more promiscuous contacts, yielding a more fluid core. There were more unique contacts at 100 than 25 °C for both proteins, indicative of increased fluidity in the protein core as the temperature rose.



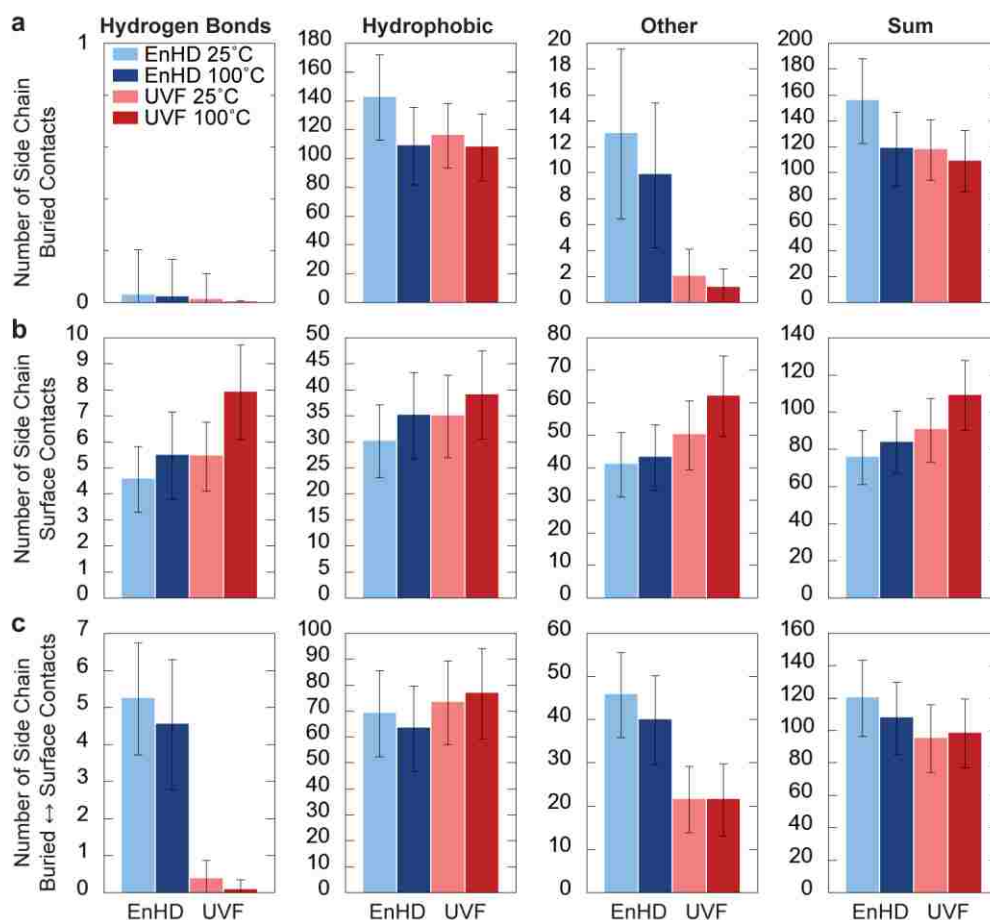
**Figure 5.5: Main chain and side chain contacts at 25 and 100 °C**

Total number of contacts for the core residues (8-53) of EnHD (blue) and UVF (red) at 25 (lighter) and 100 °C (darker) are plotted. The solid portion of the bars indicates native contacts (present in the starting structure) and white portions indicates nonnative contacts. The error bars indicate the standard deviation over time for the native contacts (below) and total (above). Contacts are classified as main chain – main chain, main chain – side chain, side chain – side chain, and total.

Figure 5.5 shows the average number of contacts per frame in simulations at 25 and 100 °C organized by whether they occurred between atoms in the main chain (N, C $\alpha$ , C, O) or side chain and further classifies them as native (solid region) or nonnative (white region). EnHD retained more of its native contacts at 25 °C, while UVF gained more nonnative contacts to replace any lost native contacts, particularly when considering side chain – side chain contacts only. UVF had little change in the number of contacts over all eight classifications between 25 and 100 °C. However, EnHD lost contacts, especially native side chain contacts, at 100 relative to 25 °C while maintaining main chain – main chain contacts.

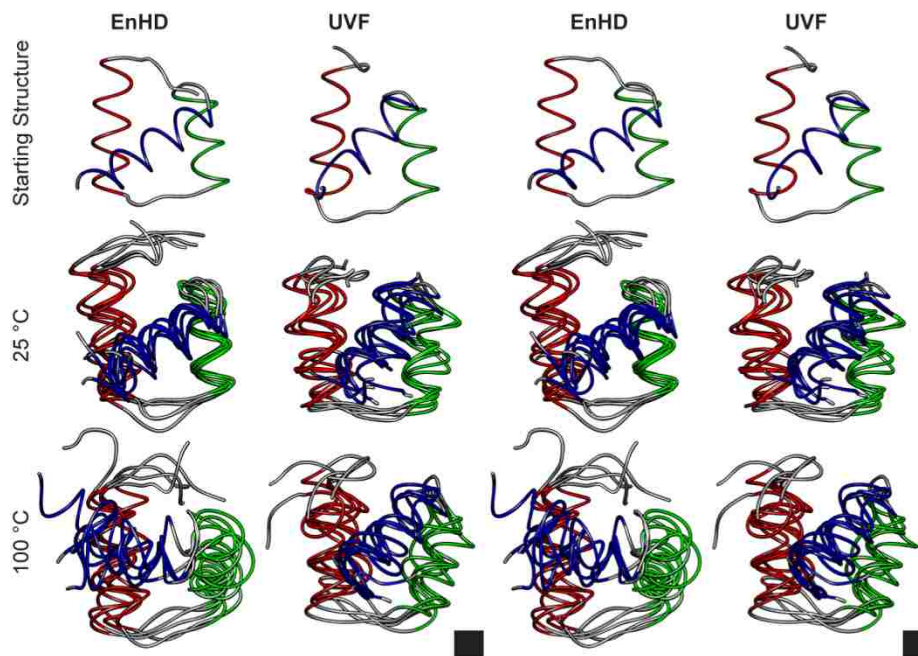
Side chain contacts were further analyzed by type of contact and whether contacts were made between buried residues, between surface residues, or between the buried and surface residues (Figure 5.6). While UVF had about the same number of buried hydrophobic side chain contacts as EnHD, EnHD had significantly more buried “other” contacts (Figure 5.6a). This was due to limiting the core to hydrophobic residues in the design of UVF. EnHD lost buried contacts, especially hydrophobic, at 100 °C relative to 25 °C, consistent with unfolding. EnHD did not have any fewer hydrogen bonds than UVF, which is due to

the fact that the polar residues removed from the core when engineering UVF did not form hydrogen bonds with each other in EnHD. On the surface, there was no significant difference between EnHD and the thermostabilized variant, despite the eight polar to hydrophobic mutations in UVF (Figure 5.6b). However, there was a significant loss in hydrogen bonds and “other” interactions between buried and surface residues in UVF compared to EnHD (Figure 5.6c). Overall, there were more hydrophobic interactions between buried residues and more hydrogen bonds and nonspecific interactions between surface residues for both proteins (Figure 5.6a,b), as expected. There were few differences in contacts between the simulations at 25 vs. 100 °C, especially for UVF.



**Figure 5.6: Types of side chain contacts at 25 and 100 °C**

Total number of contacts for the side chains of the core residues (8-53) of EnHD (blue) and UVF (red) at 25 (lighter) and 100 °C (darker) are plotted. Contacts were further classified by whether they were between residues that were (a) both buried (residues 8, 12, 16, 19, 20, 26, 31, 34, 35, 38, 40, 44, 45, 48, 49, 52), (b) both on the surface (9-11, 13-15, 17, 18, 21-25, 27-30, 32, 33, 36, 37, 39, 41-43, 46, 47, 50, 51, 53), or (c) one buried and one on the surface. Contacts are plotted left-to-right by whether they were hydrogen bonds, hydrophobic interactions, nonspecific interactions, or the total of the previous three groups. Error bars indicate the standard deviation across simulations.



**Figure 5.7: Stereo image of backbone mobility in EnHD and UVF**

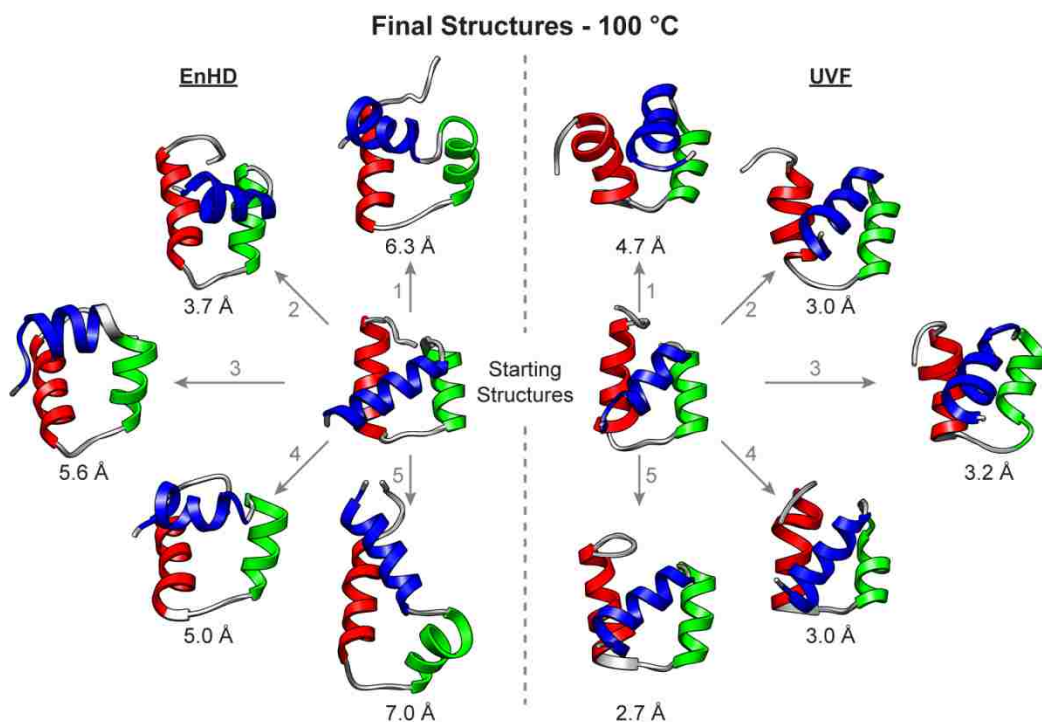
EnHD (left) and UVF (right) are colored by helix (HI red, HII green, HIII blue). The simulation starting structures are shown (top) along with the 50-ns structures from the 5 independent simulations at 25 °C (middle) and 100 °C (bottom) overlaid. UVF was more mobile at both temperatures than EnHD at 25 °C, but EnHD unfolded at 100 °C whereas UVF did not. The images can be seen in cross-eyed stereo by overlapping the two black boxes in the lower right corners.

The increased native backbone motion in UVF relative to EnHD at 25 °C can be seen in stereo images of the final structures from the 5 independent simulations (Figure 5.7). There was little difference between the motion of the backbone in UVF as the temperature increased from 25 to 100 °C. However, EnHD became more heterogeneous at 100 relative to 25 °C as it began to unfold (Figure 5.8). In Figure 5.8, the final (50 ns) structures from the high temperature simulations of EnHD show HIII pulling away from HI and HII, consistent with the first steps in unfolding. UVF, on the other hand, maintained a folded, three-helix bundle conformation with lower final core C $\alpha$  RMSD's to the starting structure.

## 5.5 Discussion

Increased backbone dynamics and promiscuous core contacts in UVF allowed it to maintain its native, folded structure in our simulations at high temperature, showing little difference in behavior with a 75 °C increase in temperature. We validated our native

simulations by comparing available NMR data to corresponding properties from our simulations. The  $^3J_{\text{HNH}\alpha}$  coupling constants and backbone amide  $S^2$  order parameters calculated from 250,000 structures from our 5 independent MD simulations of EnHD agree well with experiment. These measurements demonstrate that our simulations reproduce the backbone structure ( $\phi$  angles) and main chain dynamics (N-H bond motion) that have been observed by NMR (Religa 2008). The three helices were intact in our simulations, as reflected in coupling constants of less than 6.0 Hz (Pardi *et al.* 1984) and high order parameters in the helical regions. The HIII-HIII loop showed some disagreement with the NMR observables, particularly due to residues Leu38 and Gly39, which were more mobile in our simulations than the experimental observables suggest. Our good agreement with NOEs for EnHD and UVF show that both proteins maintain the overall structure and packing of the protein in our simulations.



**Figure 5.8: Final structures from the 100 °C simulations**

Simulation starting structures and final (50 ns) structures from the five independent runs of EnHD (left) and UVF (right) at 100 °C. Below each structures is its core C $\alpha$  RMSD to the starting structure. While EnHD unfolds at high temperature, UVF maintains the heightened dynamics observed at 25 °C.

EnHD was more rigid in our simulations than its engineered counterpart (Figure 5.7). In particular, there was less fluctuation in the backbone as observed by a lower core C $\alpha$



RMSD and C $\alpha$  RMSF in EnHD relative to UVF at 25 °C (Figure 5.3). When the temperature was raised to 100 °C, EnHD unfolded, reaching core C $\alpha$  RMSDs > 5 Å by the end of the simulations, whereas UVF remained as stable as in the simulations at 25 °C. UVF had more nonnative contacts and more diversity in contacts than EnHD (Figure 5.4,5.5). It is important to note that while UVF accumulated many nonnative contacts, these contacts were still in the same regions as the native contacts (HI to HII, and HII to HII and the C-terminus of HI), and the protein's NOEs were well satisfied (Table 5.2). Together, these data indicate that there were no large structural changes in UVF that would indicate it was unfolding. In EnHD at 100 °C, however, nonnative contacts were formed between all regions of the protein, indicative of loss of native tertiary structure and unfolding. The increased backbone flexibility and higher number of nonnative and unique contacts in UVF than EnHD reflect a more fluid, flexible core in the thermostabilized protein.

While the main chain mobility was higher in UVF relative to EnHD, the effect was enhanced further in the side chain motions. This increased motion of the buried side chains in the core of UVF relative to EnHD can be observed qualitatively when looking at structures from the simulations at 25 and 100 °C (Figure 5.7). The types of residues in the core of the different proteins help shed light on why this increased fluidity occurs in UVF (Figure 5.6). In designing UVF, Mayo and coworkers used a restricted library of amino acids that would only allow hydrophobic residues in the core and polar or charged residues on the surface (Marshall and Mayo 2001, Gillespie *et al.* 2003, Shah *et al.* 2007). Six mutations present in UVF increased the hydrophobic content of the core: Gln12Val, Glu19Phe, Arg31Leu, Ser35Ala, Gln44Ala, Trp48Phe, and Lys52Phe. None of these residues in EnHD formed hydrogen bonds with each other; instead they interacted with surface polar and buried hydrophobic residues.

Removing polar-to-hydrophobic contacts is entropically favorable, but why removing buried-to-surface hydrogen bonds results in heightened thermostability is less straightforward. In EnHD, the small movement of a buried polar residue could greatly weaken the contribution of a hydrogen bond with a surface residue in maintaining a folded structure. If the buried polar residue were to exchange hydrogen bonding partners on the surface, this new nonnative interaction might stabilize a conformation of the residue that

disrupts the native packing of the hydrophobic core. In contrast, in UVF, a small shift in a buried residue would simply exchange hydrophobic binding partners with a residue nearby and maintain a favorable energetic contribution. In this way, the thermostabilized protein could withstand the increased motion and entropy at higher temperatures by shifting interactions rather than losing them and their associated favorable enthalpic contributions.

When the temperature was raised to 100 °C, a temperature at which UVF was experimentally observed to be folded (Shah *et al.* 2007), the core C $\alpha$  RMSD did not change at all relative to 25 °C. However, 100 °C is well above the  $T_m$  of EnHD (52 °C; Mayor *et al.* 2000), and there was a marked increase in the average core C $\alpha$  RMSD for EnHD relative to 25 °C (Figure 5.3a). While EnHD did not become fully unfolded at 100 °C in these relatively short simulations, it did lose hydrophobic side chain contacts between the helices (Figure 5.5,5.6a). Consistent with the framework mechanism for folding, which has been observed for this protein previously (Mayor *et al.* 2003b, DeMarco *et al.* 2004), EnHD maintained its helical main chain – main chain contacts while losing hydrophobic and side chain contacts. UVF, on the other hand, had no change in the number or type of side chain contacts when the temperature was raised from 25 to 100 °C (Figure 5.6a).

Other proteins that have been engineered with the same intent of localizing polar residues to the surface and hydrophobic to buried regions have likewise been found to have a fluid core.  $\alpha_3D$ , for example, had similar backbone movement to natural proteins but more dynamic side chains in its core, as seen by N-H and methyl S<sup>2</sup> order parameters (Walsh *et al.* 1999, Walsh *et al.* 2001a). This protein, though it was designed for stability, is extremely fast-folding ( $t_{1/2} \approx 5 \mu s$  at 25 °C), which the authors attributed to a loose transition state ensemble caused by pre-existing hydrophobic clusters, mid-range interactions in the turns, a predisposition to form helices, and an imprecise arrangement of the orientation of buried side chains (Zhu *et al.* 2003). In agreement with these findings for  $\alpha_3D$ , kinetic studies of UVF revealed it folded almost twice as fast as EnHD ( $t_{1/2} \approx 9$  and 15  $\mu s$ , respectively at 25 °C; Gianni *et al.* 2003, Gillespie *et al.* 2003), and unlike EnHD, UVF had a loosely packed transition state.

It has been shown previously that addition of salt bridges, especially on the protein surface, results in increased thermostability (Strickler *et al.* 2006). Chan *et al.* (2011)

showed that this stabilization is due to a lower change in heat capacity ( $\Delta C_p$ ) upon folding, with contributions of  $\sim 0.2 \text{ kcal mol}^{-1} \text{ K}^{-1}$  per salt bridge. In designing UVF, 8 hydrophobic surface residues on EnHD were mutated to polar, and 5 of these polar residues were charged: Leu13Glu, Ala14Glu, Tyr25Arg, Ala43Glu, and Ile56Lys. Glu14 was the only one of these residues to make a salt bridge in the UVF NMR structure, but in our simulations all 5 residues spent most of the time making salt bridges with other polar/charged surface residues. Indeed, UVF had an increase in the number of surface hydrogen bonds (most of which were also salt bridges) over EnHD, especially at 100 °C (Figure 5.6b).

In an earlier study, Bolon and Mayo (2001) surveyed 263 globular proteins to determine the distribution of buried polar residues in proteins and found that  $\sim 1/3$  of buried residues were polar. In addition, they mutated polar residues in the core of *E. coli* thioredoxin to hydrophobic amino acids, and a quintuple mutant was shown to have a heterogeneous, though folded, conformation under native conditions. The authors suggested that buried polar residues more uniquely specify the folded structures due to the directional nature of their interactions and hydrophobic aversion, consistent with the known specific nature of polar interactions versus nonspecific hydrophobic interactions. EnHD, in agreement with the 263 natural proteins surveyed, has 6 polar out of 16 buried residues. However, these six residues do not form hydrogen bonds with each other, but instead they interact with other polar residues on the surface of EnHD. When these six residues were mutated to hydrophobic amino acids in UVF, the resulting structure was likewise found to have a fluid core and be heterogeneous in nature. So it seems that while mutating buried polar amino acids to hydrophobic residues results in thermostabilized proteins, the effect is due to removing specific interactions between the buried and surface residues rather than among the buried polar residues themselves, creating a core stabilized by dynamic, nonspecific side chain packing.

## 5.6 Conclusions

We have compared the dynamics of EnHD and its engineered thermostabilized variant, UVF, at 25 and 100 °C. We validated our simulations against NMR observables including  $^3J_{\text{HNH}\alpha}$  coupling constants, backbone amide  $S^2$  order parameters, and NOEs. Our

simulations suggest that UVF is able to maintain a folded structure at a higher temperature due to increased flexibility in its core interactions. Although UVF was designed to maximize enthalpy upon folding, we have shown that the entropic contribution is essential for maintaining the stability of UVF at high temperature.

## **Chapter 6: The Denatured State Dictates the Topology of Two Proteins with Almost Identical Sequence but Different Native Structure and Function**

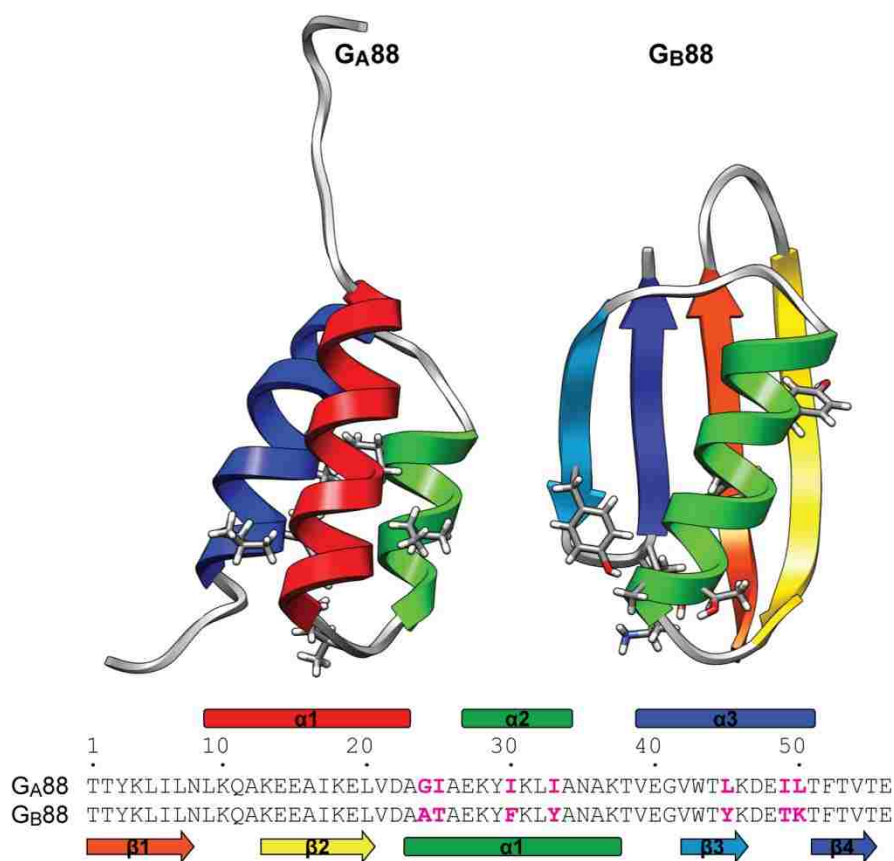
### **6.1 Summary**

The protein folding problem is often studied by comparing the mechanisms of proteins sharing the same structure but different sequence. The recent design of the two proteins  $G_A88$  and  $G_B88$ , which display different structures and functions while sharing 88% sequence identity (49 out of 56 residues), allows the unique opportunity for a complementary approach. At which stage of its folding pathway does a protein commit to a given topology? Which residues are crucial in directing folding mechanisms to a given structure? By using a combination of biophysical and computational techniques we have characterized the folding of both  $G_A88$  and  $G_B88$ . We show that contrary to expectation,  $G_B88$ , which is characterized by a native  $\alpha+\beta$  fold, displays greater extent of native helical structure in the denatured state than  $G_A88$ , which is all- $\alpha$  in its native state. Both experiments and simulations indicate that such residual structure may be tuned by changing pH. Thus, despite the high sequence identity, the folding pathways for these two proteins appear to diverge as early as in the denatured state. Our results suggest a mechanism whereby protein topology is committed very early along the folding pathway, being imprinted in the residual structure of the denatured state.

### **6.2 Introduction**

Understanding the rules that govern the folding of proteins is one of the main unsolved problems in modern science. Current knowledge on the protein folding reaction has been achieved by extensively characterizing the folding mechanisms of simple globular proteins (Fersht 2008), but a comprehension of the folding pathways of larger multi-domain systems is still far from being achieved. Given the diversity of protein structures and amino acid compositions, it is extremely difficult to draw general rules by studying folding kinetics

of individual proteins. In fact, when considering the folding of different proteins, a comparison may be jeopardized by the variability in amino acid sequence and the 3D structure of the native and denatured states.



**Figure 6.1: G<sub>A</sub>88 and G<sub>B</sub>88 sequences and structures**

Structures are shown in ribbons and colored by secondary structure (G<sub>A</sub>88: α1 residues 9-23 red, α2 27-34 green, α3 39-51 blue; G<sub>B</sub>88: β1 1-8 orange, β2 13-20 yellow, α1 23-36 green, β3 42-46 cyan, β4 51-55 blue) with the 7 differing residues displayed in sticks. A sequence alignment and secondary structure are shown below.

A powerful approach to elucidate relationships between sequence information and folding mechanism is to study proteins that differ in sequence but share the same overall fold (Chiti *et al.* 1999, Clarke *et al.* 1999, Martínez and Serrano 1999, Riddle *et al.* 1999, Travaglini-Allocatelli *et al.* 2003, Friel *et al.* 2003, Travaglini-Allocatelli *et al.* 2005, Chi *et al.* 2007, Calosci *et al.* 2008). This strategy assumes that general correlations between amino acid sequences and folding pathways may be extrapolated by comparing folding processes of different members of a given protein family.

Generally, proteins with significant sequence similarity are expected to have a similar fold. In fact, analysis of the protein data bank (PDB) reveals that a sequence similarity of 40% nearly always leads to a similar fold (Wilson *et al.* 2000). This observation provoked Rose and Creamer in 1994 to issue the “Paracelsus Challenge,” whereby the protein folding community was confronted with the task of designing two proteins that were at least 50% identical but possessed different folds (Rose and Creamer 1994). Amazingly, this goal was fully achieved in only three years, when Regan and co-workers designed a sequence that, in spite of being 50% identical to a mostly  $\beta$ -sheet protein, folded into a four-helix bundle (Dalal *et al.* 1997). Since then, others have achieved similarly impressive results (Rose 1997, Davidson 2008).

Recently, ambitious work by Bryan and co-workers led to the design of pairs of proteins with an extraordinarily high degree of sequence identity but different folds and different functions (Alexander *et al.* 2007, He *et al.* 2008). In particular, the sequences of two streptococcal protein G domains were subjected to an iterative design of heteromorphous pairs, leading the authors to produce two protein G variants.  $G_A88$  is which is mostly  $\alpha$ -helical (the 3-helix bundle protein A fold), and  $G_B88$  displays the  $\alpha+\beta$  protein G fold (Figure 6.1). These two proteins share 88% sequence identity (49 out of 56 residues), yet display two different structures and functions that are similar to the respective wild type proteins. In parallel with studies on protein families, this protein engineering achievement offers the unique opportunity for a complementary study on protein folding mechanism that addresses two key questions: (1) At which stage of its folding pathway does a protein commit to a given topology? and (2) Which residues are crucial in directing folding to a given native structure?

Here we present an extensive characterization of the folding mechanisms of  $G_A88$  and  $G_B88$  by experiment and molecular dynamics (MD) simulation. The results obtained under a variety of solvent conditions suggest the presence in the denatured state of  $G_B88$  of pH-sensitive residual structure, as indicated by a pH dependence of its  $m_{D-N}$  value, which is not observed for  $G_A88$ . The MD simulations are consistent with these findings, showing that for  $G_B88$  the nonpolar solvent accessible surface area decreases markedly at low pH. Interestingly, in agreement with earlier observations on a similar heteromorphous protein A/G pair sharing 59% sequence identity (A219 and G311; Scott and Daggett 2007), the extent of

native-like helical structure in the denatured state of G<sub>B</sub>88 ( $\alpha+\beta$  fold) is greater than that of denatured G<sub>A</sub>88 (all- $\alpha$  fold). Both our current and earlier studies of this system suggest that protein topology is committed very early along the folding pathway and “imprinted” in the residual structure of the denatured state. This weak, loosely defined topology is sufficient to dictate the pathway of folding. The significance of these observations from the perspective of previous work on the folding of proteins with the same topology but very different sequences is discussed.

## 6.3 Methods

### 6.3.1 Molecular Dynamics Simulations

We performed 11 all-atom, explicit solvent molecular dynamics (MD) simulations for each protein, G<sub>A</sub>88 and G<sub>B</sub>88: 298 K (30 ns), 498 K neutral and low pH (3 x 50 ns, 2 x 5 ns), for a total of 700 ns (0.7  $\mu$ s) of simulation time. The starting structures were taken from the published NMR ensembles (G<sub>A</sub>88: PDB ID 2jws, model 1; G<sub>B</sub>88: PDB ID 2jwu, model 3 (298 K), model 1 (498 K); He *et al.* 2008). Low pH systems were created by protonating all aspartate and glutamate residues (neither protein contains a histidine). The simulations were all performed using our in-house MD software, *in lucem* molecular mechanics (*ilmm*; Beck *et al.* 2000-2012), with the Levitt *et al.* (1995) all-atom force field and the microcanonical ensemble (NVE, constant number of particles, system volume, and total energy). Nonbonded terms were treated with an 8 Å (at 498 K) or 12 Å (298 K) force-shifted cutoff. The proteins were minimized and solvated with explicit F3C flexible waters (Levitt *et al.* 1997) using our standard protocol (Beck and Daggett 2004). Briefly, the protein was treated *in vacuo* for 1000 steps of steepest descent minimization. Pre-equilibrated F3C water was added within 1.8 Å of the protein to a box extending at least 10-12 Å from the protein on all sides. The water was then minimized for 1000 steps, and then subjected to 500 ps of dynamics with 2-fs timesteps and an additional 500 steps of steepest descent minimization. Finally, the protein was minimized for 500 steps.

### 6.3.2 Simulation Analysis

The denatured state ensemble was defined as the final 30 ns of the three longer (50 ns) 498 K simulations. The C $\alpha$  root-mean-square deviation (RMSD) was calculated over



all 56 residues as well as for just the core residues. The core was defined as the consecutive residues beginning with the N-terminal residue of the first secondary structure element as defined by He *et al.* (2008) to the C-terminus of the final element (G<sub>A</sub>88: residues 9-51; G<sub>B</sub>88: 1-55). The percentage helix and solvent accessible surface area (SASA) were calculated using our in-house implementations of the Dictionary of Protein Secondary Structure (DSSP; Kabsch and Sander 1983) and Lee and Richards (1971) algorithms, respectively. The percentage of  $\alpha$ -helix or  $\beta$ -sheet structure was reported over the total 56 residues as the number of structured residues as defined by DSSP, and nonpolar SASA was reported as the sum of the SASA for all nonpolar residues (Ala, Val, Phe, Pro, Met, Ile, Leu, Trp, Gly). When values were reported relative to the native state, the average value over the 30 ns 298 K simulation was used as the native value.

### 6.3.3 Transition State Assignment

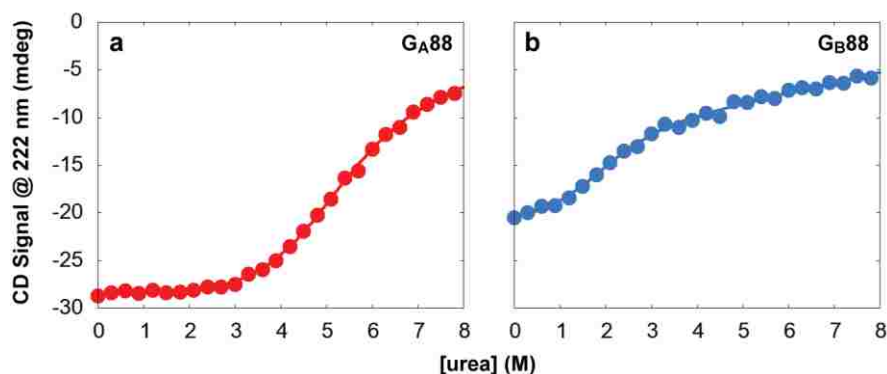
Multidimensional scaling (MDS) of the all-against-all C $\alpha$  RMSD matrix was performed using the R statistical package to assign conformational ensembles. The exit from the native-like cluster, or point of no return, in the 3 dimensional projection of the MDS was defined as the transition state (TS) and the preceding 5 ps as the transition state ensemble, as described previously (Li and Daggett 1994, Li and Daggett 1996). The average C $\alpha$  RMSD to the starting structure was reported for each TS structure as was the average pairwise C $\alpha$  RMSD between all TS structures in the ensemble.

The fraction of time that residues were in contact was calculated for the native ensembles (all 30 ns of the native state), the TS ensembles (5 ps from each 498 K unfolding simulation), and the denatured state (the last 30 ns of the 3 long 498 K simulations). Two residues were considered in contact if they contained carbon atoms that were  $\leq 5.4$  Å apart or any other two non-hydrogen atoms that were  $\leq 4.6$  Å apart. Hydrogen bonds were measured in the denatured state (last 30 ns of the 3 50 ns 498 K simulations) for specific residue pairs using the following criteria: (1) the distance between the donor hydrogen and acceptor atom was  $\leq 2.6$  Å; (2) the donor – hydrogen – acceptor angle was within 45° of linearity and (3) the charges on the donor and acceptor atoms were  $\leq -0.3$  and the charge on the hydrogen was  $>+0.3$ .

## 6.4 Results

### 6.4.1 Equilibrium Unfolding of G<sub>A</sub>88 and G<sub>B</sub>88

To study the folding mechanism of G<sub>A</sub>88 and G<sub>B</sub>88, Morone *et al.* (2011) carried out both equilibrium and kinetic (un)folding experiments. The results of urea-induced equilibrium denaturation of G<sub>A</sub>88 and G<sub>B</sub>88 measured at 10 °C, pH 7.2 in 50 mM sodium phosphate buffer, as monitored by far UV circular dichroism (CD) spectroscopy, are provided in Figure 6.2. The observed transitions follow simple two-state behavior, suggesting the absence of stable equilibrium intermediates for both proteins and indicating that these designed variants are capable of cooperative (un)folding reactions. Furthermore, the reaction was fully reversible under all conditions explored.



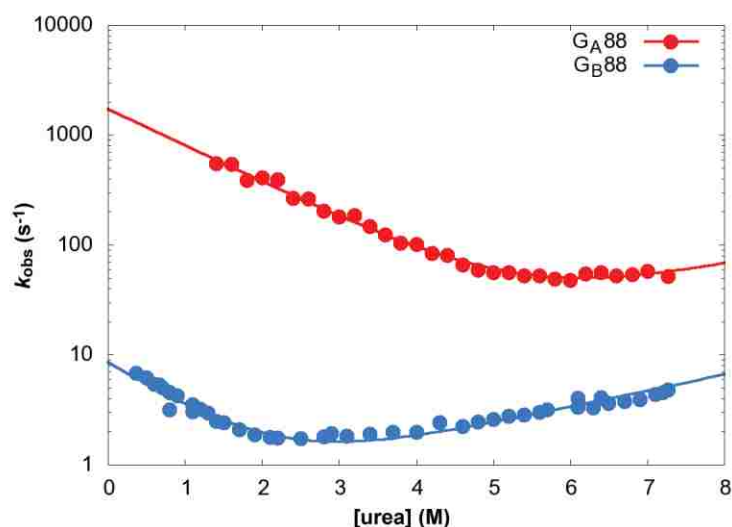
**Figure 6.2: Chemical denaturation of G<sub>A</sub>88 and G<sub>B</sub>88 monitored by CD**

Equilibrium denaturation of (a) G<sub>A</sub>88 and (b) G<sub>B</sub>88 monitored by CD in 50mM sodium phosphate buffer at pH 7.2 and 10 °C. Lines are the best fit to a two-state unfolding mechanism. Data provided by Angela Morrone (Morrone *et al.* 2011).

At physiological pH and in the absence of denaturant, the unfolding free energy of G<sub>A</sub>88 derived from a two-state analysis and a global fit of 60 wavelengths (from 250 to 220 nm) is  $\Delta G_{D-N} = 3.00 \pm 0.18$  kcal/mol with a  $m_{D-N} = 0.62 \pm 0.04$  kcal mol<sup>-1</sup> M<sup>-1</sup>. In the case of G<sub>B</sub>88,  $\Delta G_{D-N} = 2.35 \pm 0.30$  kcal/mol and  $m_{D-N} = 1.10 \pm 0.08$  kcal mol<sup>-1</sup> M<sup>-1</sup>. Considering that the G<sub>B</sub>88 construct contains ~10 structured residues more than G<sub>A</sub>88, both  $m_{D-N}$  values are consistent with those expected for proteins of this size, according to the BPPred database (Geierhaas *et al.* 2007). Hence, the seeming difference in cooperativity, as reflected by the different  $m_{D-N}$  values, can be accounted for by the difference in the number of structured residues between the two proteins.

## 6.4.2 Folding and Unfolding Kinetics

Morrone *et al.* (2011) carried out extensive fluorescence kinetic experiments on both proteins under a variety of different experimental conditions. In particular, the folding and unfolding kinetics were investigated at several pH values, ranging from 10 to 2. In the case of G<sub>A</sub>88, it was not possible to measure reliable folding and unfolding rate constants at 25 °C over a wide range of denaturant concentrations because the rates were too fast for our stopped-flow apparatus. Thus, kinetic folding data for the two proteins were recorded at 10 °C to slow the process for G<sub>A</sub>88 and to obtain values under the same conditions for comparison in the case of G<sub>B</sub>88. In all cases, the folding and unfolding time courses were fitted satisfactorily to a single exponential decay at any final denaturant concentration.

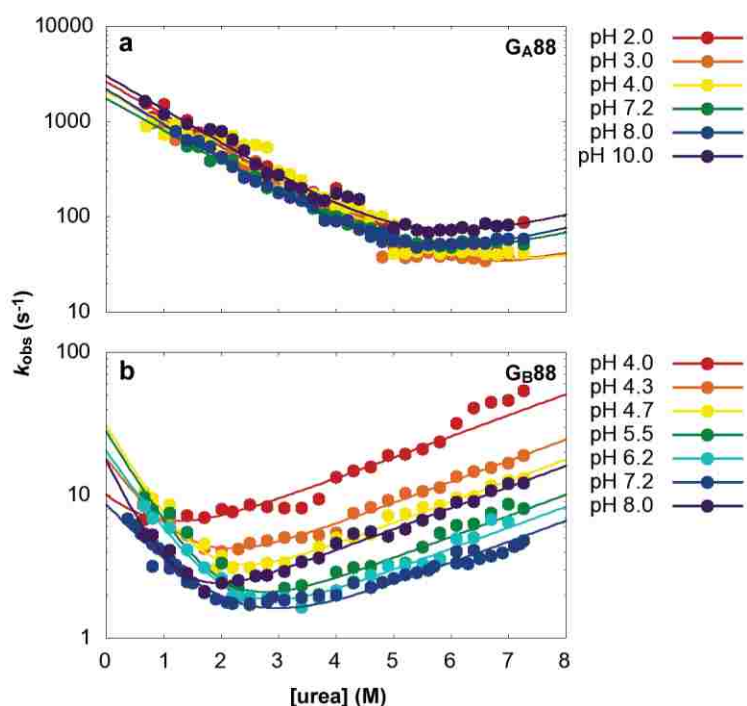


**Figure 6.3: Folding and unfolding rate constants for G<sub>A</sub>88 and G<sub>B</sub>88**

Semilogarithmic plot of the observed rate constant for folding and unfolding of G<sub>A</sub>88 (red) and G<sub>B</sub>88 (blue) versus [urea] at pH 7.2 in 50 mM NaP<sub>i</sub> obtained at 10 °C, monitored by fluorescence emission. Data provided by Angela Morrone (Morrone *et al.* 2011).

Semi-logarithmic plots of the observed folding/unfolding rate constants of G<sub>A</sub>88 and G<sub>B</sub>88 versus denaturant concentration (i.e. chevron plots) at pH 7.2 are presented in Figure 6.3. Both proteins displayed a V-shaped chevron plot, a hallmark of two-state folding (Jackson and Fersht 1991). In the case of G<sub>A</sub>88 there was excellent agreement between the thermodynamic parameters obtained by equilibrium and kinetic data. However, in the case of G<sub>B</sub>88 there was a minor deviation. The  $m_{D-N}$  value was  $1.10 \pm 0.08$  kcal mol<sup>-1</sup> M<sup>-1</sup> from equilibrium experiments and  $0.90 \pm 0.05$  kcal mol<sup>-1</sup> M<sup>-1</sup> from chevron plot analysis. As observed previously for other small single domain proteins (Mayor *et al.* 2003b, Religa *et al.*

2005, White *et al.* 2005), a significant deviation of  $m_{D-N}$  from equilibrium and kinetic data suggests that there is residual structure and/or changes in the exposure of nonpolar residues in the denatured state of the protein. Since the small deviation observed for  $G_{B88}$  is at the limit of experimental detection, we performed additional experiments under various conditions as well as MD simulations to further investigate these options, as described below.

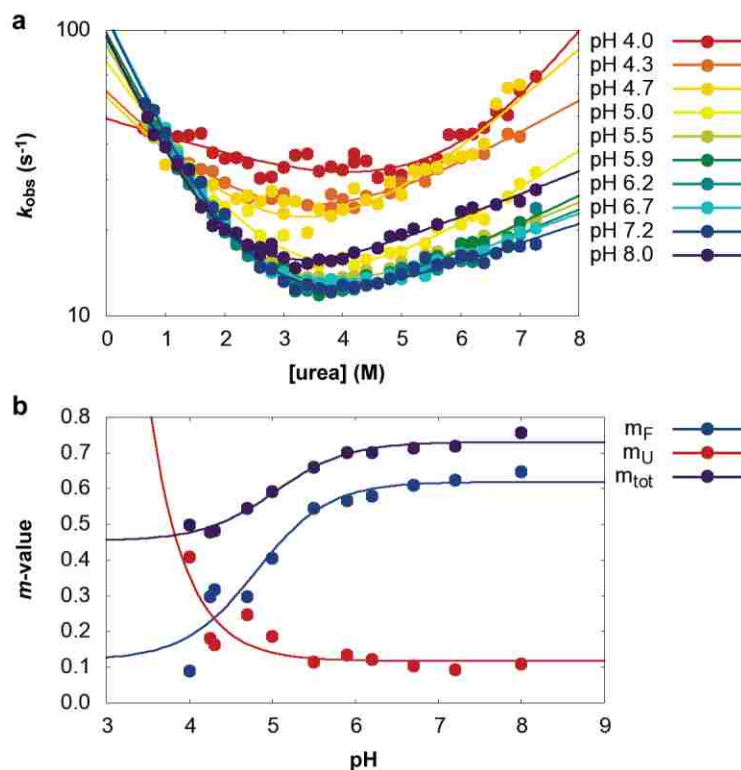


**Figure 6.4: Chevron plots of  $G_{A88}$  and  $G_{B88}$  at varying pH**

Semilogarithmic plots of the observed rate constants for folding and unfolding versus [urea] measured at different pH values and 10 °C. The lines are the best fit to a two-state model. (a)  $G_{A88}$ : pH 2.0 red, 3.0 orange, 4.0 yellow, 7.2 green, 8.0 blue, 10.0 purple. (b)  $G_{B88}$ : pH 4.0 red, 4.3 orange, 4.7 yellow, 5.5 green, 6.2 cyan, 7.2 blue, 8.0 purple. Data provided by Angela Morrone (Morrone *et al.* 2011).

A powerful method to address the global properties of folding transition and denatured states is the analysis of chevron plots recorded under different experimental conditions or on various site-directed variants (Sánchez and Kiefhaber 2003). In fact, since the  $m$ -values (slopes of the unfolding and refolding arms of the chevron plots) reflect the change in accessible surface area upon (un)folding, analysis of their dependence on reaction conditions may be of diagnostic value to identify transition state movements along the reaction coordinate, as well as denatured state collapse or residual structure. In this study, we compared the folding kinetics of  $G_{A88}$  and  $G_{B88}$  at various pH values ranging from 10 to 2. Inspection of Figure 6.4 reveals that, while both the stability and  $m$ -values of  $G_{A88}$  are

insensitive to pH and this protein is fully native even at pH 2.0, G<sub>B</sub>88 is destabilized at pH < 5. However, the low stability of the latter protein and the poor definition of the observed refolding arms prevented a quantitative analysis of the  $m$ -values. Therefore as detailed below, we resorted to investigating the folding of G<sub>B</sub>88 under stabilizing conditions.

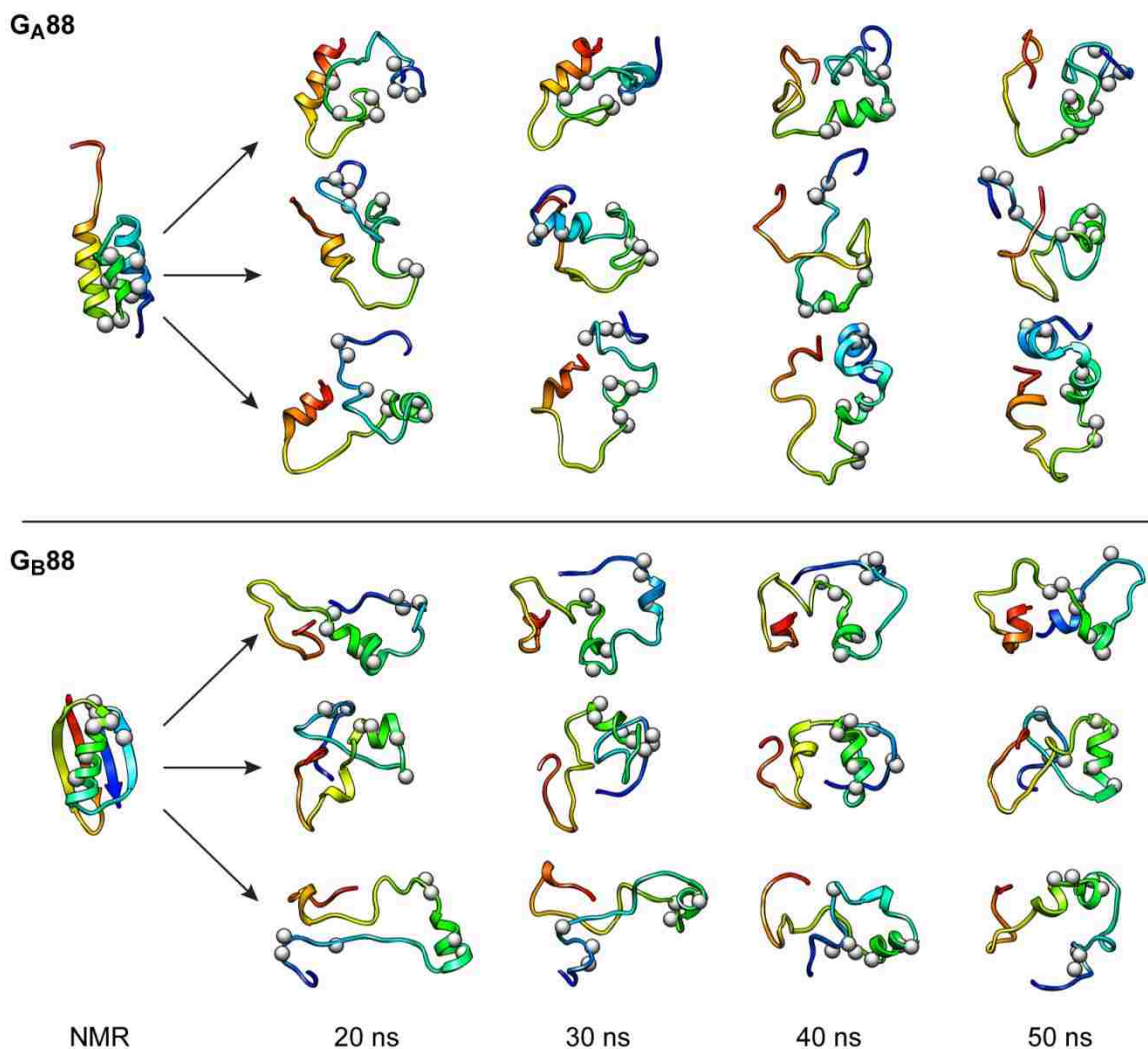


**Figure 6.5: Rate constants and  $m$ -values for G<sub>B</sub>88 folding and unfolding in salt**

(a) Semilogarithmic plot of the observed rate constants for folding and unfolding of G<sub>B</sub>88 versus [urea] measured at different pH values in the presence of 0.4 M sodium sulfate at 25 °C. pH 4.0 red, 4.3 red-orange, 4.7 orange, 5.0 yellow, 5.5 yellow-green, 5.9 green, 6.2 blue-green, 6.7 cyan, 7.2 blue, 8.0 purple. Lines are the best fit to a two-state model. (b) Dependence of  $m_F$  (blue),  $m_U$  (red), and  $m_{D-N}$  (purple) on pH. Lines are the best fit to an equation implying a single protonation site with pK<sub>a</sub> ~5. Data provided by Angela Morrone (Morrone *et al.* 2011).

Certain inorganic salts, such as phosphates and sulfates, favor compact protein conformations because of preferential exclusion of solvent from the protein surface (Timasheff 1993); this makes them potent stabilizers of both the native and the partially folded states. The chevron plots of G<sub>B</sub>88 measured at different pH values and in the presence of 0.4 M sodium sulfate are provided in Figure 6.5a. As expected, the stabilizing salt allows for a better definition of the refolding arms of the chevron plots. Consequently, we carried out a quantitative analysis of folding parameters over a wide range of pH conditions in the presence of salt. Figure 6.5b shows the dependence on pH of calculated  $m_{D-N}$ ,  $m_F$  and  $m_U$

values for  $G_{B88}$ . The data fit to the protonation of a single titratable group with an apparent  $pK_a \sim 5$ . Interestingly, the  $m_{D-N}$  decreases with decreasing pH values, suggesting that the denatured state of this small, single-domain protein becomes more compact at acidic conditions. Importantly, however, even at neutral pH, the observed  $m_{D-N}$  is lower than that calculated from equilibrium experiments (i.e.  $0.93 \pm 0.05 \text{ kcal mol}^{-1} \text{ M}^{-1}$ ), suggesting the presence of residual structure in the denatured state under physiological conditions.



**Figure 6.6: Structures of the denatured state at neutral pH.**

$G_{A88}$  (above) and  $G_{B88}$  (below) are colored red  $\rightarrow$  blue from the N  $\rightarrow$  C. The  $C\alpha$  atoms of differing residues (24, 25, 30, 33, 45, 49, 50) are shown as gray balls. The NMR structures of the native states are shown on the left, and the calculated structures at 20, 30, 40, and 50 ns for each of the 3 long simulations at 498 K (1 top, 2 middle, 3 bottom) are depicted.

### 6.4.3 Molecular Dynamics Simulations

To further investigate the differences between  $G_A88$  and  $G_B88$ , MD simulations were conducted. Five independent thermal unfolding simulations were performed for each protein at 498 K at both neutral and low pH, in addition to a simulation at room temperature (298 K) for each protein as a control. Snapshots from the thermal unfolding at neutral pH are presented in Figure 6.6. The sequence positions where the two proteins display different amino acids are highlighted as balls. In  $G_B88$  the first hairpin ( $\beta1/\beta2$ ) had a tendency to be loosely maintained in the denatured state, whereas the second hairpin ( $\beta3/\beta4$ ) was more extended. In  $G_A88$  the C-terminal region tended to collapse down more than in  $G_B88$ , leading to slightly more interactions involving  $\beta3$  and  $\beta4$  (residues 42-55). Specifically, these residues had  $21.6 \pm 5.5$  internal residue-residue contacts in  $G_A88$  vs.  $18.8 \pm 4.5$  contacts for  $G_B88$ . Finally, the central helix (displayed in green in Figure 6.6) was fairly well preserved in the denatured state of  $G_B88$ .

**Table 6.1: Properties of the native and denatured state ensembles from MD simulations at neutral and low pH**

	pH and Temp (K)	C $\alpha$ RMSD (Å)*	% $\alpha$ -Helix <sup>†</sup>	% Native $\alpha$ -Helix <sup>‡</sup>	NP SASA (Å <sup>2</sup> ) <sup>§</sup>	% Native NP SASA <sup>¶</sup>
$G_A88$	N, 298	$3.4 \pm 0.4$	$67 \pm 3$	100	$1330 \pm 83$	100
	N, 498	$11.4 \pm 1.3$	$15.8 \pm 10.7$	$24 \pm 16$	$2424 \pm 204$	$182 \pm 19$
	L, 498	$11.8 \pm 2.5$	$26.4 \pm 11.8$	$40 \pm 18$	$2309 \pm 224$	$174 \pm 20$
$G_B88$	N, 298	$2.5 \pm 0.4$	$26 \pm 1$	100	$1185 \pm 72$	100
	N, 498	$11.3 \pm 1.5$	$15.5 \pm 7.6$	$59 \pm 29$	$2062 \pm 198$	$174 \pm 20$
	L, 498	$12.7 \pm 1.2$	$27.4 \pm 13.8$	$104 \pm 52$	$1867 \pm 191$	$158 \pm 19$

\* All properties were averaged over the final 30 ns of the 3 long 498 K simulations, and the average  $\pm$  standard deviation is given. The values for the 298 K simulations are defined as 100% native.

<sup>†</sup>  $\alpha$ -Helix was defined using DSSP (Kabsch and Sander 1983).

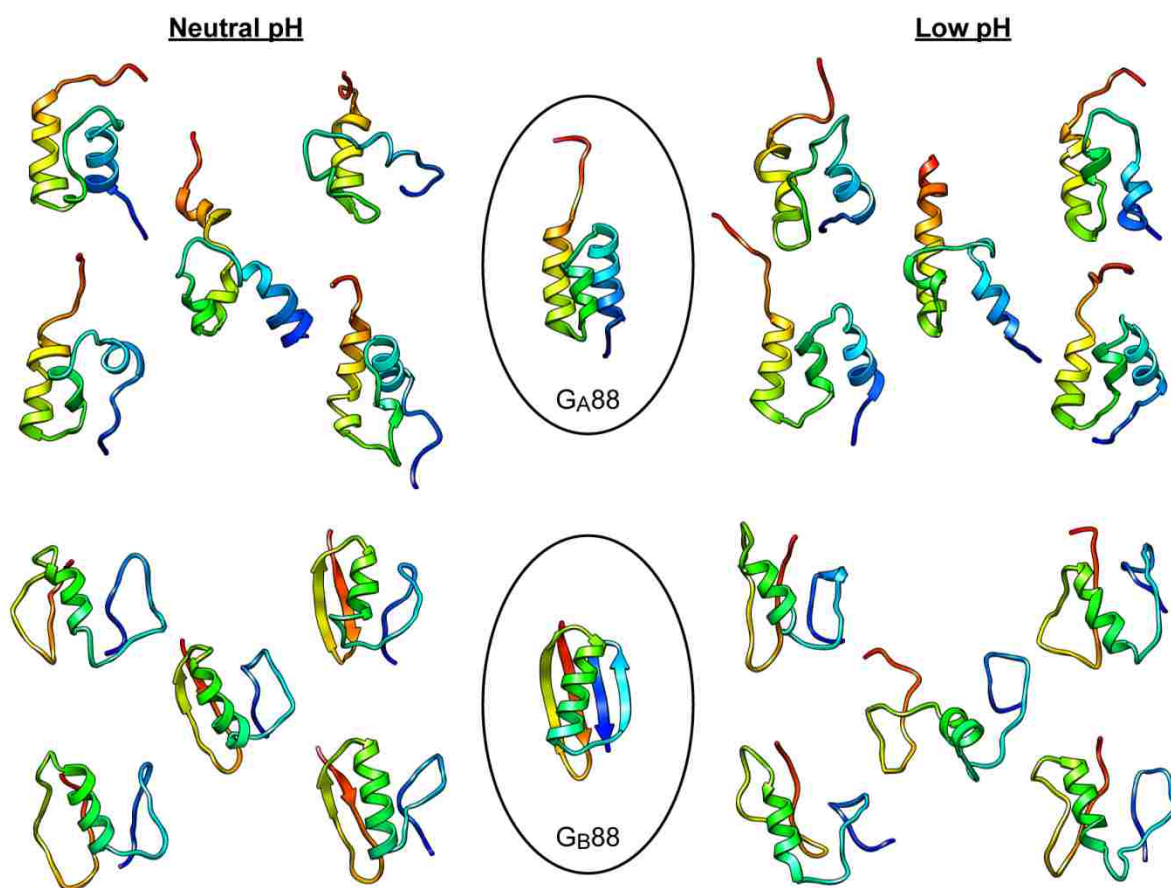
<sup>‡</sup> Percentage native  $\alpha$ -helix was calculated relative to the percent of  $\alpha$ -helix in the 30 ns 298 K native simulation.

<sup>§</sup> Nonpolar SASA was defined as the sum of the SASA for all hydrophobic residues (see Methods).

<sup>¶</sup> Percent native nonpolar SASA was calculated relative to the nonpolar SASA in the 30 ns 298 K native simulation.

A quantitative comparison of the unfolding simulations is provided in Table 6.1, where we report the average properties over the three independent unfolding simulations for each protein. The C $\alpha$  RMSD relative to the starting structure reached over 11 Å in all cases. While there was little change in the C $\alpha$  RMSD upon lowering the pH for  $G_A88$ , in the case of  $G_B88$  the C $\alpha$  RMSD increased at low pH. The overall residual helical content in the denatured state at neutral pH was similar for  $G_A88$  and  $G_B88$ . The helix content of both proteins increased when the pH was lowered. Interestingly, the helical content in the

denatured state of  $G_B88$  at low pH was surprisingly high, being 104% of the native extent of helix content. In contrast  $G_A88$  contained 39% of its native helix content. While we could not directly address the helical content of the denatured states, we experimentally observed the  $m_{D-N}$  value of  $G_B88$  to decrease with decreasing pH (Figure 6.5b). This observation is consistent with the MD simulations suggesting the denatured state of  $G_B88$  to be more structured at acidic than neutral pH. Importantly, such dependence was not observed in  $G_A88$ , whose  $m_{D-N}$  value was insensitive to pH.



**Figure 6.7: Structures of the transition state at neutral and low pH**

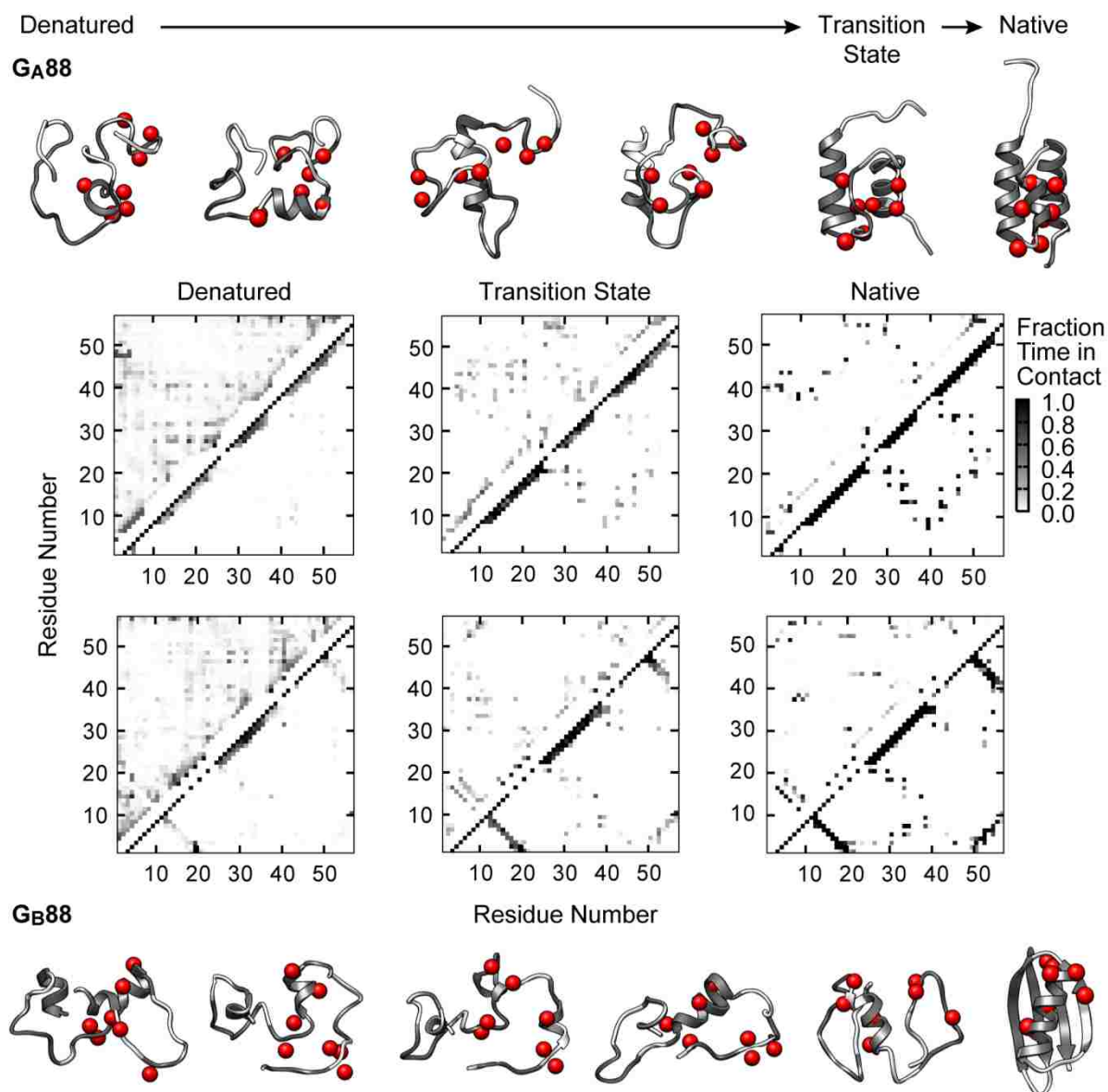
$G_A88$  (above) and  $G_B88$  (below) are colored red  $\rightarrow$  blue from N  $\rightarrow$  C. The NMR structure of the native state is shown in a circle in the middle for each of the two proteins. For each set, runs 1-5 are ordered: top left, top right, middle, bottom left and bottom right. Transition state times for  $G_A88$  at neutral pH are at the following simulation times: 486, 483, 249, 331, and 225 ps for runs 1-5, respectively.  $G_A88$  at low pH: 362, 227, 842, 77, 165 ps.  $G_B88$  at neutral pH: 422, 449, 248, 276, 338 ps.  $G_B88$  at low pH: 159, 102, 164, 305, 147 ps. Transition states were selected as the first cluster exit in the 3D MDS of the all-against-all  $C_\alpha$  RMSD matrix. For more details, see the Methods section.



The relative compaction of G<sub>B</sub>88 compared with G<sub>A</sub>88 can also be seen in Figure 6.6. This effect is reflected in the nonpolar solvent accessible surface areas (SASA) of the two denatured states relative to their control native states (Table 6.1). The nonpolar SASA at neutral pH was approximately 1000 Å<sup>2</sup> lower in the native state (2424 ± 204 vs. 1330 ± 83 and 2062 ± 198 1185 ± 72 Å<sup>2</sup> for G<sub>A</sub>88 and G<sub>B</sub>88, respectively). Relative to the native state, the nonpolar SASA of the denatured state at neutral pH increased by 182% in the case of G<sub>A</sub>88 and by 174% for G<sub>B</sub>88. Although both proteins had some reduction of nonpolar SASAs when the pH was lowered, G<sub>B</sub>88 was more sensitive to acidification, and the increase in nonpolar exposure upon unfolding was reduced relative to neutral pH. This finding parallels the experimental observations, which clearly indicate the denatured state of G<sub>B</sub>88 was more compact at acidic conditions, as reflected by a decreased  $m_{D-N}$  value (Figure 6.5b).

Transition state (TS) ensembles were identified for each of the simulations described above, as well as the two shorter simulations for each protein at both pH conditions (Figure 6.7). The representative TS structures for G<sub>A</sub>88 show that their helical content and overall size were quite similar at both pHs. On the other hand, the G<sub>B</sub>88 TS ensemble was very sensitive to pH, and the structures at neutral pH were much more native-like than those at low pH. The β1/β2 hairpin was more robust than the β3/β4 hairpin in the TS. Furthermore, the two hairpins did not appear to interact directly and instead behaved as different entities physically separated by the central helix.

Interestingly, the heterogeneity of the G<sub>A</sub>88 and G<sub>B</sub>88 TS ensembles at neutral pH were similar, particularly when the effect of the unstructured N-terminus of G<sub>A</sub>88 was accounted for (compare the core C $\alpha$  RMSD “To Self” values in Table 6.2). Both ensembles are approximately 5 Å C $\alpha$  RMSD from their respective starting structures. The pH sensitivity of the G<sub>B</sub>88 TS ensemble was dramatic: the C $\alpha$  RMSD to the starting structure increased by 1.5 Å and the average C $\alpha$  RMSD between structures within the TS ensemble also increased by 1.4 Å. The core C $\alpha$  RMSD was 6.2 Å for the G<sub>B</sub>88 TS ensemble at low pH and it was quite distorted with a C $\alpha$  RMSD of 7.0 Å from its starting structure (Figure 6.7). In contrast the G<sub>A</sub>88 TS remained within ~4 Å of its starting structure and the spread within a TS ensemble was ~4.5 Å (excluding the N-terminus of G<sub>A</sub>88).



**Figure 6.8: Folding pathway and contact maps at neutral pH**

G<sub>A</sub>88 (above) and G<sub>B</sub>88 (below) are colored white with secondary structure elements colored dark gray (G<sub>A</sub>88: residues 9-23, 27-34, 39-51; G<sub>B</sub>88: residues 1-8, 13-20, 23-36, 42-46, 51-55). Structures are from 50, 40, 30, 20 ns, and the transition state (G<sub>A</sub>88: 0.486 ns; G<sub>B</sub>88: 0.422 ns) of run 1 at 498 K. The NMR structures are also shown to the right. C $\beta$  atoms of differing residues (24, 25, 30, 33, 45, 49, 50) are shown as red balls (or Ca for Gly24 in G<sub>A</sub>88). Fraction time in contact for G<sub>A</sub>88 (above) and G<sub>B</sub>88 (below) are plotted with nonnative contacts in the top-left triangle of each panel and native contacts (present in the starting structure) in the bottom-right of each panel. Native contacts are reported for the full 30 ns 298 K simulation (N = 30,000); transition state contacts are reported for all transition state ensembles (N = 30); and denatured state contacts are plotted for the last 30 ns of the 3 long 498 K simulations (N = 90,000). Two residues were considered in contact if they contained carbon atoms that were  $\leq 5.4$  Å apart or any other non-hydrogen atoms were  $\leq 4.6$  Å apart. Contacts were colored from white (never occurred) to black (present 100% of the time).

The overall folding pathway from representative simulations (run 1 in each case) of G<sub>A</sub>88 and G<sub>B</sub>88 are presented in Figure 6.8 as the reverse of the simulated unfolding process.

It appears that the topology, residual structure, and interactions in the denatured state direct whether the protein will fold into the helical  $G_A88$  structure or the mixed  $\alpha/\beta$   $G_B88$  structure. In  $G_A88$  the protein has some dynamic residual structure, while the main chain is fairly fluid with different main chain interactions occurring over time within the collapsed state; the productive interactions with respect to folding are local along the sequence, with folding of kernels of helical structure that then dock together and consolidate in the TS ensemble. In contrast, in the case of  $G_B88$ , the approximate topology of the native state was already apparent in the denatured state, with segregation of the two hairpin regions by the central helix. This helix was more stable in  $G_B88$  than in  $G_A88$  due to improved packing interactions between residues 30 and 33, which in  $G_A88$  are both Ile, while in the case of  $G_B88$  are Phe and Tyr. Moreover, substitution of Gly to Ala at residue 24 and Ile to Thr at residue 25 in  $G_B88$  also increased the helical propensity of the region.

**Table 6.2: C $\alpha$  RMSD of the transition state from neutral and low pH simulations at 498 K**

	pH	C $\alpha$ RMSD (Å)		Core* C $\alpha$ RMSD (Å)	
		To Start <sup>†</sup>	To Self <sup>‡</sup>	To Start	To Self
$G_A88$	N	5.5 ± 0.7	6.7 ± 1.0	4.2 ± 0.5	4.7 ± 0.4
	L	4.9 ± 0.5	5.5 ± 0.9	3.5 ± 0.6	3.7 ± 0.5
$G_B88$	N	5.5 ± 1.1	4.8 ± 0.9	5.4 ± 1.1	4.7 ± 0.9
	L	7.0 ± 1.6	6.2 ± 1.4	6.9 ± 1.5	6.2 ± 1.4

\* The “core” is residues 9-51 for  $G_A88$  and 1-55 for  $G_B88$ . The core value removes the effect of the unstructured N-terminus in  $G_A88$ . The core C $\alpha$  RMSD for  $G_A88$  at both low and neutral pH is 4 Å, and for  $G_B88$  it is 6 Å. The average ± standard deviation is reported.

<sup>†</sup> The C $\alpha$  RMSD to the simulation starting structure was reported for all 5 transition state structures (N = 5).

<sup>‡</sup> The pairwise C $\alpha$  RMSD between all 5 transition state structures was reported (N = 10).

Interestingly, even though the sequence of the N-terminal region is identical in the two proteins, there was a tendency for the region to form a helix in  $G_A88$  and a loose hairpin in  $G_B88$  (Figure 6.7 and 6.8). This difference was, in large part, due to a hydrogen bond between Thr1 and Glu19, which was present in 99.8% of the  $G_B88$  denatured state structures at neutral pH. At low pH, however, the loose  $\beta 1/\beta 2$  hairpin was not present since Glu19 was protonated. Although positions 1 and 19 are identical in  $G_A88$  and  $G_B88$ , in the case of the latter, we did not observe this hydrogen bond in the denatured state at neutral pH. Instead, Glu19 tended to interact with the solvent, and Thr1 either interacted with solvent or formed hydrogen bonds with Asp47 and Glu48 (over 67% and 60% of the denatured state, respectively). The backbone  $\phi/\psi$  angles of residues 47 and 48 when interacting with Thr1 were compatible with forming  $\alpha 3$  in  $G_A88$ . These observations indicate that long-range

interactions play a critical role in the residual structure in the denatured state of these proteins.

The  $\beta 3/\beta 4$  turn was also fractionally present in the denatured state of  $G_B88$ , due to the presence of two side chain hydrogen bonds: Asp47 – Lys50 and Asp47 – Tyr45 (57% and 14% of the time at neutral pH, respectively). Of note, these hydrogen bonds were not present in the low pH denaturing simulations of  $G_B88$  due to the protonation of Asp47, nor were they present in the denatured state of  $G_A88$  where both Lys50 and Tyr45 are mutated to Leu. This interaction appears to stabilize the  $\beta 3/\beta 4$  turn in  $G_B88$ , thus preventing these residues from assuming an  $\alpha$ -helical structure, as they do in  $G_A88$ .

## 6.5 Discussion

Critical insights on many problems in biology have been classically achieved using simplified model systems. While a comprehensive understanding of the folding of large multidomain proteins is still an aspiration, the successful design of two heteromorphic proteins sharing 88% sequence identity, called  $G_A88$  and  $G_B88$  (Alexander *et al.* 2007), provides a unique opportunity to unveil the mechanism whereby a few key residues commit the polypeptide chain to its characteristic and functionally competent native topology. Here we have characterized the folding and unfolding kinetics of  $G_A88$  and  $G_B88$  by experiment and simulation. The key findings show that both engineered proteins appear to fold via a two-state mechanism, and protein topology is committed very early along the folding pathway.

### 6.5.1 The Role of Long-Range Interactions in Denatured States

For the purpose of this study, it is essential to understand the mechanism whereby a few key residues univocally determine native topology, i.e. which structural determinants preclude the sequence of  $G_A88$  from adopting the  $G_B88$  topology and vice-versa? Experimentally, we clearly detected a difference in denatured state properties of the two proteins. In fact, we observed the  $m_{D-N}$  value of  $G_B88$  to decrease with decreasing pH (Figure 6.5b). This observation is consistent with the MD simulations suggesting the denatured state of  $G_B88$  to be more structured at acidic than neutral pH, as mirrored both by its native helical content and by its solvent accessible surface area (Table 6.1). Importantly,

such a dependence was not observed in G<sub>A</sub>88, whose  $m_{D-N}$  value was found experimentally to be insensitive to pH. In summary, although only 7 of the 56 amino acids are different between G<sub>A</sub>88 and G<sub>B</sub>88, only the latter displays a detectable residual structure in its denatured state. Such a structure may be tuned by changing pH, as reflected both by analysis of  $m_{D-N}$  values as a function of pH (apparent pK<sub>a</sub> ~5) and by comparison of the MD simulations at neutral and low pH. Surprisingly, the residues that are different between the two proteins (Figure 6.1) do not include amino acids titrating below neutral pH, suggesting that the observed compaction of the denatured state of G<sub>B</sub>88 originates from nonlocal effects. Indeed, the MD simulations of G<sub>B</sub>88 highlight the presence of side chain hydrogen bonds in the  $\beta$ 3/ $\beta$ 4 hairpin turn, in the denatured state at neutral pH. These hydrogen bonds involve residue Asp47, which is protonated in the low pH simulations and does not form hydrogen bonds with either Tyr45 or Lys50 as it does at neutral pH. Interestingly, residues 45 and 50 are both mutated to Leu in G<sub>A</sub>88 and are part of the  $\alpha$ 3 helix. However, Asp47 along with Glu48 tends to interact with Thr1 in the denatured state of G<sub>A</sub>88. This interaction favored backbone dihedral angles, which were compatible with  $\alpha$ -helix rather than the  $\beta$ 3/ $\beta$ 4 turn, as in G<sub>B</sub>88.

According to AGADIR (Munoz and Serrano 1997) and our MD simulations, the 7 residue difference in sequence between G<sub>A</sub>88 and G<sub>B</sub>88 yields an increased helical propensity for G<sub>B</sub>88. In particular, the four residues forming the loop connecting  $\alpha$ 1 and  $\alpha$ 2 in G<sub>A</sub>88 adopt a helical conformation in the structure of native G<sub>B</sub>88. Furthermore, MD reveals that substitution of Gly to Ala at residue 24 and Ile to Thr at residue 25 increases the helical propensity of the region. Thus, experiments and simulations converge in supporting the hypothesis that a longer, more stable  $\alpha$ -helix in G<sub>B</sub>88 prevents the latter sequence from folding to the G<sub>A</sub>88 structure. Interestingly, it was recently shown that a switch between the G<sub>A</sub> and G<sub>B</sub> structures may be obtained even with a single amino-acid substitution (Alexander *et al.* 2009). Under conditions where the G<sub>A</sub> fold is >90% populated, mutation of Leu45 into Tyr shifts the population to >90% of the G<sub>B</sub> fold. Surprisingly, position 45 is not in the loop where G<sub>A</sub>88 and G<sub>B</sub>88 display a different helical propensity (Figure 6.1), indicating that the greater helical content of the denatured state of G<sub>B</sub>88 is affected by long-range interactions.

Overall, comparison of the folding of  $G_A88$  and  $G_B88$  highlights a conundrum: while only a few residues (or even a single one) are responsible for the selective stabilization of the two alternative topologies, information on the folding mechanism indicates that no single residue appears to act as a unique gatekeeper in the selection of protein topology. Both experiments and simulations on the folding of  $G_A88$  and  $G_B88$  suggest that native topology might be pre-sculpted in the denatured state, where incipient nuclei are present. Stabilization of such nuclei is affected by long-range interactions, and commitment to the native fold occurs by selective stabilization of these incipient nuclei, rather than by actively blocking alternative pathways.

### **6.5.2 Do Engineered Proteins Display Cooperative Folding?**

An intriguing general question is whether folding is under evolutionary pressure. This issue was recently discussed by Baker and co-workers in a study on the folding of a *de novo* designed protein, Top7, characterized by a novel non-natural topology (Watters *et al.* 2007). While most small, naturally occurring proteins fold in a cooperative mode (Tanford 1970), Top7 displays a non-cooperative folding mechanism, suggesting that cooperativity may be a result of natural selection. In this context, it may seem somewhat puzzling that both  $G_A88$  and  $G_B88$  fold in a cooperative manner, displaying single exponential folding and unfolding kinetics as well as V-shaped chevron plots, a hallmark of two-state folding (Jackson and Fersht 1991). However, a possible explanation to reconcile these apparently conflicting observations is that, while the structure of Top7 is non-natural (Kuhlman *et al.* 2003), both  $G_A88$  and  $G_B88$  have been engineered starting from naturally occurring frameworks (Alexander *et al.* 2007).

It was reported recently that malleability of protein folding pathways stems from the existence of multiple nuclei within a given protein structure (Lindberg and Oliveberg 2007). Implicit in this view, a natural topology would contain one or more nucleation motifs, representing the minimal units to encode for the final structure (Hubner *et al.* 2006b). Nucleation of the motif is the basis for cooperative folding and the number of accessible pathways is related to the number of nucleation motifs within a protein (Haglund *et al.* 2008). On the basis of the folding pathways observed for Top7,  $G_A88$ , and  $G_B88$ , it is tempting to speculate that, since naturally occurring topologies contain nucleation motifs, productive

folding of these substructures would result in cooperative transitions. In contrast, non-natural proteins may not contain stable nucleation motifs and, thus, they appear to fold non-cooperatively. We may conclude that evolution does not directly select for cooperative folding, but rather it selects for topologies that can fold in a cooperative manner.

### 6.5.3 Comparison with Studies on Protein Families

The study of homologous proteins has represented a powerful approach to obtain insight into protein folding (Chiti *et al.* 1999, Clarke *et al.* 1999, Martínez and Serrano 1999, Riddle *et al.* 1999, McCallister *et al.* 2000, Friel *et al.* 2003, Travaglini-Allocatelli *et al.* 2003, Travaglini-Allocatelli *et al.* 2005, White *et al.* 2005, Chi *et al.* 2007, Calosci *et al.* 2008, Wensley *et al.* 2009), especially when combined with structural information on intermediate or transition states. In a recent study, Calosci *et al.* (2008) addressed the structural features of the early and late transition states of two homologous three-state proteins, PSD-95 PDZ3 and PTP-BL PDZ2. For different PDZ domains, we observed that the late folding transition states (TS2) are more similar to each other than the early transition states (TS1). This observation would suggest that, while native topology defines the late stages of folding in a unique way, significant freedom in creating structural contacts is observed for the early events. In this perspective, it is of interest to compare the cases of G<sub>A</sub>88/G<sub>B</sub>88, whose folding appears to diverge early in the denatured state, with that of the PDZ family, where a strong native bias is seen only at the late stages of folding. A plausible scenario to reconcile these apparently contrasting results would imply the presence of multiple nucleation motifs at the early stages of PDZ folding. In fact, the apparent structural divergence of TS1 for PDZ2 and PDZ3 most likely arises from selective stabilization of alternative nuclei in the two denatured proteins, which may then appear to explore distinct early folding pathways. When folding proceeds to the native state, the alternative nuclei all consolidate in a native-like conformation and folding pathways appear to converge. Support for this hypothesis comes from circular permutation experiments on PDZ2, whereby alternative nuclei may be selectively stabilized and alter the early events of folding without affecting the late ones (Ivarsson *et al.* 2008, Ivarsson *et al.* 2009). On the other hand, G<sub>A</sub>88 and G<sub>B</sub>88 do not appear to contain alternative nuclei; moreover, since they display two completely different structures, their respective nucleation motifs may be completely independent such that their folding pathways diverge from the very early stages. Future

work based on protein engineering experiments, MD simulations, and  $\Phi$ -value analysis (Fersht *et al.* 1992) will further address structural determinants in the folding of this heteromorphic protein pair in an effort to identify crucial residues for the stabilization of these two alternative topologies and their respective folding motifs, and to extend the experimental and theoretical work to more complex multi-domain systems.



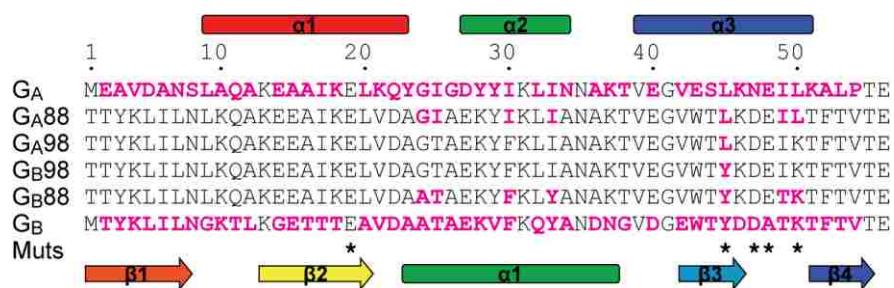
## Chapter 7: Identification and Characterization of Specific Interactions in the Denatured State of G<sub>A</sub>88 and G<sub>B</sub>88

### 7.1 Summary

A pair of proteins was recently engineered to have 88% identity but populate two different folds, either a 3-helix bundle or the mixed  $\alpha/\beta$  protein G fold. Using a combination of folding kinetics and molecular dynamics simulations, we previously predicted interactions in the denatured state that promoted folding to one topology or the other. Here, we created mutants to knock out these interactions or insert them into the opposite sequence to test our predictions. Our high-temperature, unfolding simulations showed that the mutants were successful in creating and destroying the desired interactions and modulating the amount and distribution of  $\alpha$ -helix in the denatured state. Experimental work to test these predictions is ongoing. We also simulated a further iteration of the designed protein pair that has 98% sequence identity between the all- $\alpha$  and  $\alpha/\beta$  protein as well as the naturally-occurring pair of proteins, which was the starting point for the engineering work and has 16% identity. Our results identified interactions in the denatured state common among the proteins that fold to the same topology. The two  $\beta$ -hairpins in the  $\alpha/\beta$  fold formed independently of each other, separated sequentially by a helix present in the denatured state. The all- $\alpha$  topology was characterized by more long-range interactions.

### 7.2 Introduction

G<sub>A</sub>88 and G<sub>B</sub>88 were engineered by Alexander *et al.* (2007) to have different folds despite sharing 88% sequence identity (Figure 7.1). G<sub>A</sub>88 was based on the A domain of protein G and forms a 3-helix bundle, and G<sub>B</sub>88 was based on the B domain of the same protein and forms a mixed  $\alpha/\beta$  fold (Figure 6.1). Alexander *et al.* (2009) continued refining their structures and came up with two sequences differing by only one residue (98% identity) that fold to the two distinct topologies. In G<sub>A</sub>98, when residue 45 is a Leu, the protein takes on the all- $\alpha$  fold, and in G<sub>B</sub>98, Tyr45 promotes the  $\alpha/\beta$  fold.



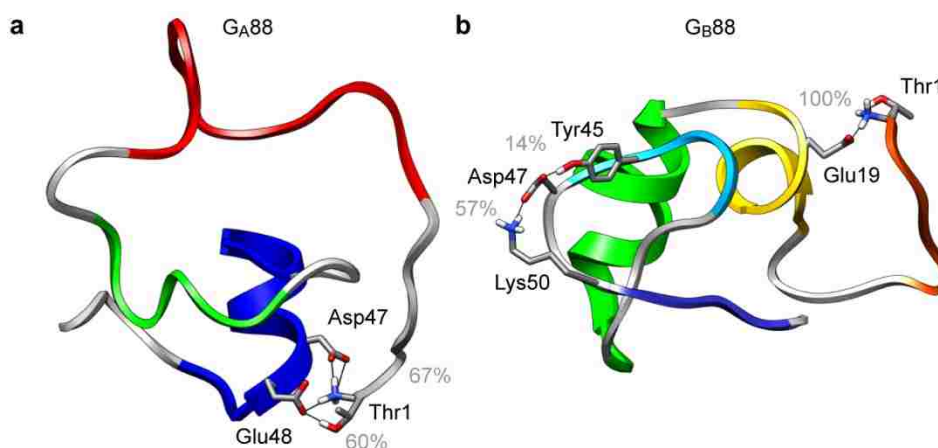
**Figure 7.1: Sequences of 16%, 88%, and 98% identical  $G_A/G_B$  pairs**

Sequences of the 16%, 88% and 98% identical  $G_A/G_B$  pairs are shown with differing residues between the pairs highlighted in magenta. Mutations to  $G_{A88}$  or  $G_{B88}$  performed here are indicated by asterisks (\*) underneath the sequences. Secondary structure assignments from  $G_{A88}$  and  $G_{B88}$  are shown above and below the sequences, respectively ( $G_{A88}$ :  $\alpha 1$  9-23 red,  $\alpha 2$  27-34 green,  $\alpha 3$  39-51 blue;  $G_{B88}$ :  $\beta 1$  1-8 orange,  $\beta 2$  13-20 yellow,  $\alpha 1$  23-36 green,  $\beta 3$  42-46 cyan,  $\beta 4$  51-55 blue).

In our previous work on  $G_{A88}$  and  $G_{B88}$ , interactions in the denatured state were identified that determined the final folded structure (Morrone *et al.* 2011). In particular, there were two clusters of hydrogen bonding residues in  $G_{B88}$  that favored hairpin formation, and there was one in  $G_{A88}$  that promoted formation of the  $\alpha 3$  helix (Figure 7.2). In order to follow up on these predictions of denatured state interactions in  $G_{A88}$  and  $G_{B88}$  that commit the proteins to their respective topologies, we made a series of mutants (Figure 7.1) with the intent of 1) knocking out these hydrogen bonds / salt bridges; 2) adding residues participating in the hydrogen bonds / salt bridges from one fold into the other; or 3) seeing whether the interactions are inherent in the sequence or dependent on the starting structure for our unfolding simulations. We performed one native (298 K) and three unfolding (498 K) all-atom, explicit solvent molecular dynamics simulations of each of these mutant structures. Additionally, we simulated the 98% and 16% identical structures; we threaded the  $G_{B88}$  and  $G_{B98}$  sequences onto the all- $\alpha$  structure and vice-versa and unfolded them; and we simulated  $G_{A88}$  and  $G_{B88}$  starting from an extended structure.

Analysis of the  $G_{A88}/G_{B88}$  mutant simulations showed success in creating and breaking the predicted interactions. The Asp47 – Thr1 – Glu48 network present in the denatured state of  $G_{A88}$  was disrupted by pH (protonation of the Asp and Glu) as well as the Asp48Asn and Glu48Gln mutations. The network was also absent when the  $G_{B88}$  sequence was threaded onto the all- $\alpha$  structure. When the Leu50Lys and Leu50Lys/Leu45Tyr mutations (that are present in  $G_{B88}$ ) were introduced into the  $G_{A88}$  structure, only the double mutant abolished the Asp47 – Thr1 – Glu48 network in favor of the Tyr45 – Asp47 – Lys50

network that is present in the denatured state of  $G_B88$ . In addition, the Glu19Gln and Lys50Met mutations that were meant to knock out interactions in the denatured state of  $G_B88$  promoted, to some extent, the Asp47 – Thr1 – Glu48 network present in that of  $G_A88$ . Experimental work to test these observations is currently ongoing in the Gianni group.



**Figure 7.2: Interactions in the denatured state of  $G_A88$  and  $G_B88$  that dictate the folded topology**

(a) The Thr1 – Asp47 – Glu48 hydrogen bond network promotes helix formation in  $G_A88$  (Run 2, 36.0 ns). (b) The Thr1 – Glu19 and Tyr45 – Asp47 – Lys50 hydrogen bonds promote formation of the  $\beta 1/\beta 2$  and  $\beta 3/\beta 4$  hairpins (Run 2, 31.0 ns). Gray numbers indicate the percentage of time in the denatured state that at least one hydrogen bond was present between the corresponding residue pair.

We also unfolded the 98% and 16% identical pairs of proteins in order to compare the interactions in the denatured state. The  $\beta 1/\beta 2$  and  $\beta 3/\beta 4$  interactions discussed above were present in the denatured state of  $G_B98$  and  $G_B$ , and the Asp47 – Thr1 – Glu48 network was present in one of three independent simulations of  $G_A98$ . In simulating  $G_A88$  and  $G_B88$  at high temperature starting from an extended conformation, we observed the Asp47 – Thr1 – Glu48 network in one run of  $G_A88$  and the  $\beta 3/\beta 4$ -promoting interactions in  $G_B88$ . The  $\beta 1/\beta 2$ -promoting salt bridge was only present minimally in one run of  $G_B88$  and, surprisingly, in one run of  $G_A88$ . The  $G_B$ -based sequences tended to have interactions within the two  $\beta$ -hairpins in the denatured state but not between them, whereas the  $G_A$ -based sequences had more long-range contacts.

## 7.3 Methods

### 7.3.1 Preparation of Constructs

In order to knock out the Asp47 – Thr1 – Glu48 network in G<sub>A</sub>88, the Glu48Gln single-mutant and Glu48Gln/Asp47Asn double mutants were created starting with G<sub>A</sub>88 (PDB id: 2jws) using the Dunbrack backbone-dependent rotamer library in UCSF Chimera (Dunbrack 2002). The  $\beta$ 3/ $\beta$ 4-promoting residues, Tyr45 – Asp47 – Lys50, were added into the G<sub>A</sub>88 structure via the Leu50Lys and Leu50Lys/Leu45Tyr mutations. Additionally, the 7-mutation change to thread the G<sub>B</sub>88 sequence on the all- $\alpha$  structure, the 3 mutations to create the G<sub>A</sub>98 structure (Ile25Thr, Ile30Phe, Leu50Lys), and the 4 mutations to put the G<sub>B</sub>98 sequence on the all- $\alpha$  structure (Ile25Thr, Ile30Phe, Leu50Lys, Leu45Tyr) were made.

Starting with the  $\alpha/\beta$  G<sub>B</sub>88 structure (PDB id: 2jwu), three mutants were created to knock out the  $\beta$ 1/ $\beta$ 2- and  $\beta$ 3/ $\beta$ 4-promoting interactions: Glu19Gln, Lys50Met, and Glu19Gln/Lys50Met. Similarly, the G<sub>A</sub>88 and G<sub>A</sub>98 sequences were threaded onto the  $\alpha/\beta$  structure making 7 and 4 (Ala24Gly, Tyr33Ile, Thr49Ile, Tyr45Leu) mutations, respectively. G<sub>B</sub>98 was created by making 3 mutations: (Ala24Gly, Thr33Ile, Thr49Ile)

G<sub>A</sub> and G<sub>B</sub> (Alexander *et al.* 2007), the 16% identical starting structures for the engineering process, were also simulated. G<sub>A</sub> (or PSD1) has an NMR structure (PDB id: 2fs1; He *et al.* 2006), and G<sub>B</sub> (B1 domain of protein G) has a 2.07 Å resolution crystal structure (PDB id: 1pga; Gallagher *et al.* 1994). Extended structures with the G<sub>A</sub>88 and G<sub>B</sub>88 sequences were created in Ribosome (Srinivasan 1997) and simulated at 498 K only (3 x 50 ns). The 3 long (50 ns) 498 K low-pH WT simulations from our previous study (Morrone *et al.* 2011), where all aspartate and glutamate residues were protonated, were also considered.

### 7.3.2 Simulation Protocol

One native (30 ns) and three unfolding (50 ns) molecular dynamics simulations were run for each of the structures described above using our in-house software package, *in lucem* molecular mechanics (*ilmm*; Beck *et al.* 2000-2012), with the Levitt *et al.* (1995) force field and flexible three-center water model (F3C; Levitt *et al.* 1997). The NVE (constant number of particles, volume, and energy) ensemble was employed, and nonbonded terms were

treated with an 8 Å (at 498 K) or 12 Å (298 K) force-shifted cutoff. Missing hydrogen atoms were added, 500 steps of steepest descent minimization was performed, and any mutated residues were further minimized for 100 steps. Finally, the whole structure was minimized for 1000 steps. Solvation and heating followed our standard protocols (Beck and Daggett 2004): the proteins were placed in a preequilibrated box of F3C water with the proper density as determined by the liquid-vapor coexistence curves [298 K: 0.997 g/mL (Kell 1967), 498 K: 0.829 g/mL (Haar *et al.* 1984)] with the box extending 10-12 Å past the edge of the protein. The water only was minimized for 1000 steps and then subjected to 500 ps of dynamics with 2-fs timesteps. Finally, the water and then the protein were each minimized for an additional 500 steps. At this point the production run began, with structures saved out every 1 ps for analysis. In total, 21 constructs were simulated, giving 3.66  $\mu$ s of simulation time and 3,660,000 structures.

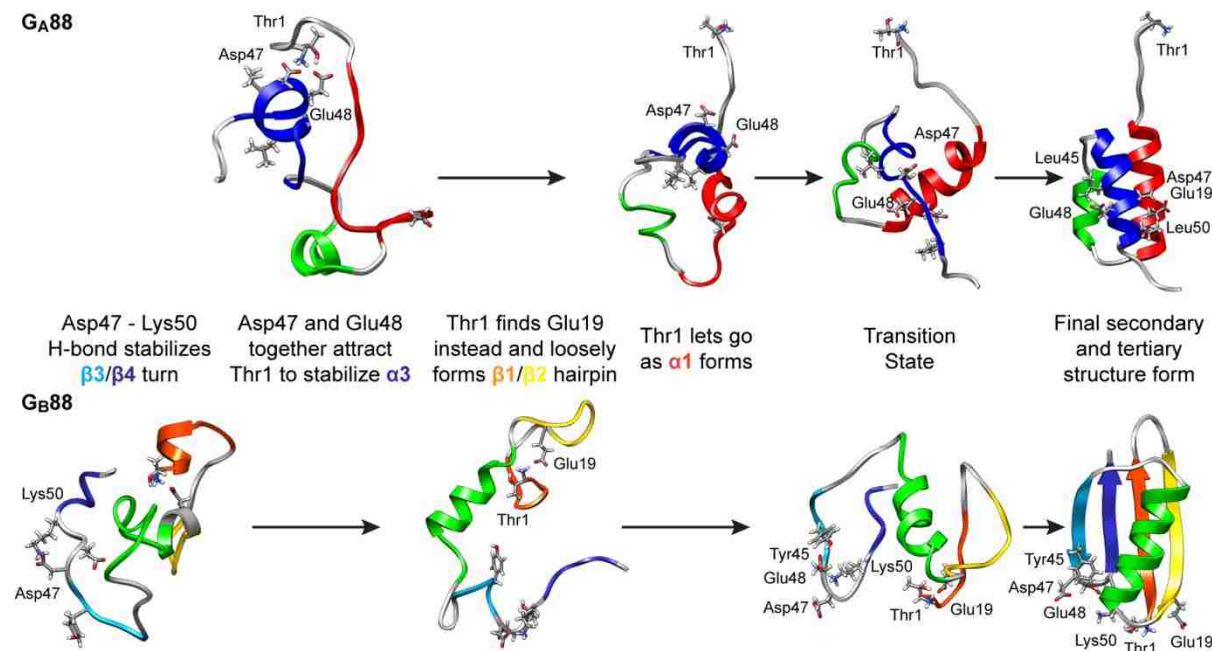
### 7.3.3 Simulation Analysis

Contact analysis was performed to determine what fraction of the simulation time each residue pair of interest made at least one hydrogen bond or salt bridge. A contact was present if the distance between the donor hydrogen and acceptor atom was  $\leq 2.6$  Å; the donor – hydrogen – acceptor angle was between 135 and 180°; and the absolute values of the charges on the donor, acceptor, and hydrogen atoms were  $> 0.3$ . Secondary structure was defined using our in-house implementation of the DSSP algorithm (Kabsch and Sander 1983). The fraction time in contact for the selected contact pairs and percentage time in  $\alpha$ -helix were reported over the final 30 ns of the 3 independent 50-ns 498 K unfolding simulations.

## 7.4 Results

Based on the interactions observed previously (Morrone *et al.* 2011), a mechanism for folding is proposed as shown in Figure 7.3. In  $G_A88$ , the nonnative Asp47 – Thr1 – Glu48 salt bridge network stabilized a backbone conformation that was compatible with  $\alpha$ -helix formation and promoted formation of  $\alpha 3$ . Thr1 eventually broke away, which allowed the rest of the helix structure to form. In  $G_B88$ , Asp47 instead interacted with Lys50 (which is Leu50 in  $G_A88$ ) forming the turn of the  $\beta 3/\beta 4$  hairpin. Because Asp47 was interacting with

Lys50, the C-terminus of Thr1 instead found Glu19. This Thr1 – Glu19 interaction caused loose formation of the  $\beta 1/\beta 2$  hairpin. Final folding for  $G_B88$  consisted of firming up of these two hairpins as the nonnative salt bridges broke and  $\alpha 1$  formed.

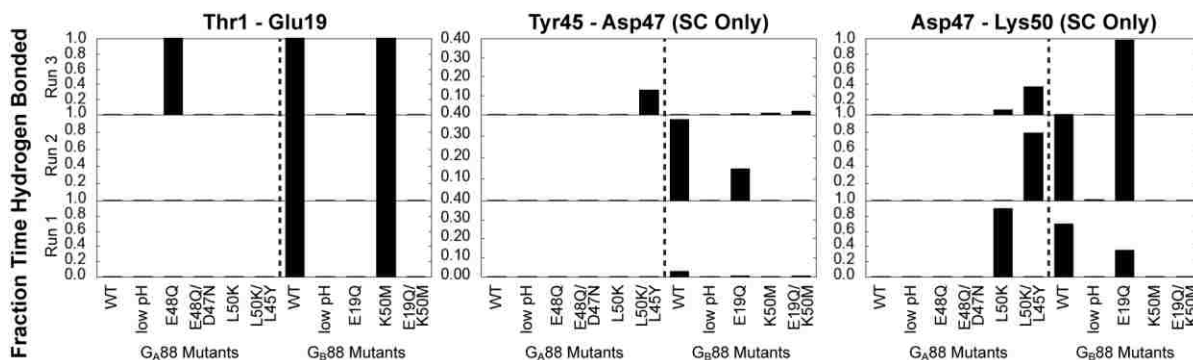


**Figure 7.3: Putative folding mechanisms for  $G_A88$  and  $G_B88$**

Hydrogen bond networks in the denatured state of  $G_A88$  (above) and  $G_B88$  (below) are predicted to determine the final folded structure as described.  $G_A88$  structures are from 498 K run 2: 37.0, 5.0, 0.483 ns;  $G_B88$  498 K run 1: 41.0, 19.0, 0.422 ns; and final structures are the NMR structures.

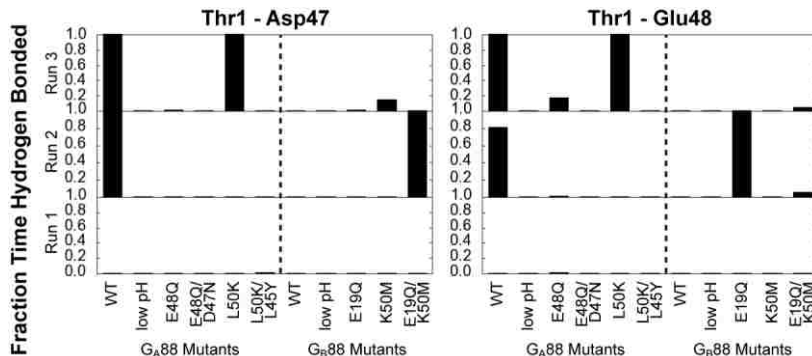
#### 7.4.1 Analysis of the $G_A88$ and $G_B88$ Mutants

Hydrogen bond interactions between Thr1 – Glu19 and Tyr45 – Asp47 – Lys50 in the denatured state of  $G_B88$  promoted formation of the  $\beta 1/\beta 2$  and  $\beta 3/\beta 4$  hairpins, respectively. The Thr1 – Glu19 interaction was not present at low pH (where Glu is protonated) nor in the Asp19Asn or Asp19Asn/Lys50Met mutants of  $G_B88$ , as expected (Figure 7.4). It was present when the Asp47 – Thr1 – Glu48 interaction (which promotes formation of  $\alpha 3$  in  $G_A88$ ) was knocked out in run 3 of the  $G_A88$  Glu48Gln mutant, though curiously it was not present in the double mutant (Asp47Asn/Glu48Gln). Disruption of the Thr1 – Glu19 hydrogen bond in the double mutant may be due to disruption of the Tyr45 – Asp47 – Lys50 network that promoted the  $\alpha/\beta$  fold. The Thr1 – Glu19 hydrogen bond was also present to a minimal extent when the sequence of  $G_B88$  was threaded on the  $G_A88$  structure (as predicted) and to a large extent when the  $G_A88$  sequence was threaded on the  $G_B88$  structure (not expected; Figure 7.7).



**Figure 7.4: The  $\beta 1/\beta 2$  Thr1 – Glu19 and  $\beta 3/\beta 4$  Tyr45 – Asp47 – Lys50 networks from  $G_B88$  in  $G_A88/G_B88$  mutants**

The fraction of time at least one hydrogen bond was present between the Thr1 – Glu19 (left), Try45 – Asp47 (middle), and Asp47 – Lys50 (right) residues pairs is plotted for each of the 3 independent runs at 498 K. In each plot, the six structures on the left are mutations to  $G_A88$ , and the five on the right are mutations to  $G_B88$ . Tyr45 – Asp47 and Asp47 – Lys50 reported side chain – side chain contacts only.



**Figure 7.5: The  $\alpha 3$  Asp47 – Thr1 – Glu48 network from  $G_A88$  in  $G_A88/G_B88$  mutants**

The fraction of time at least one hydrogen bond was present between the Thr1 – Asp47 (left) and Thr1 – Glu48 (right) residue pairs is plotted for each of the 3 independent runs at 498 K. In each plot, the six structures on the left are mutations to  $G_A88$ , and the five on the right are mutations to  $G_B88$ .

The Tyr45 – Asp47 – Lys50 hydrogen bond network, which promoted formation of the  $\beta 3/\beta 4$  hairpin in the denatured state of  $G_B88$ , was knocked out by low pH (due to protonation of Asp47), the Lys50Met mutation, and the Lys50Met/Glu19Gln double mutation, as expected (Figure 7.4). Introducing the Leu45Tyr and Leu50Lys mutations into  $G_A88$  promoted formation of the Tyr45 – Asp47 and Asp47 – Lys50 interactions, respectively, as anticipated. The Glu19Gln mutation to  $G_B88$  (which should disfavor formation of the  $\beta 1/\beta 2$  hairpin in  $G_B88$ ) did not knock out this network, which suggests that the two hairpins form independently.

We introduced the charge-neutralizing Glu48Gln and Asp47Asn mutations into  $G_A88$  in order to break the Asp47 – Thr1 – Glu48 salt bridge network in the denatured state. The

low pH simulations also removed the charges from Asp47 and Glu48. In the Glu48Gln single mutant, in the Glu48Gln/Asp47Asn double mutant, and at low pH, the interactions did not form, as expected (Figure 7.5). These two interactions were present in one simulation of the Leu50Lys and Leu50Lys/Leu45Tyr mutants, where residues involved in the  $\beta 3/\beta 4$ -promoting salt bridge network were introduced. In these two mutants, formation of the  $\beta 3/\beta 4$  turn (Figure 7.4) disfavored the two competing,  $\alpha 3$ -promoting salt bridges (Thr1 – Asp47 and Thr1 – Glu48).

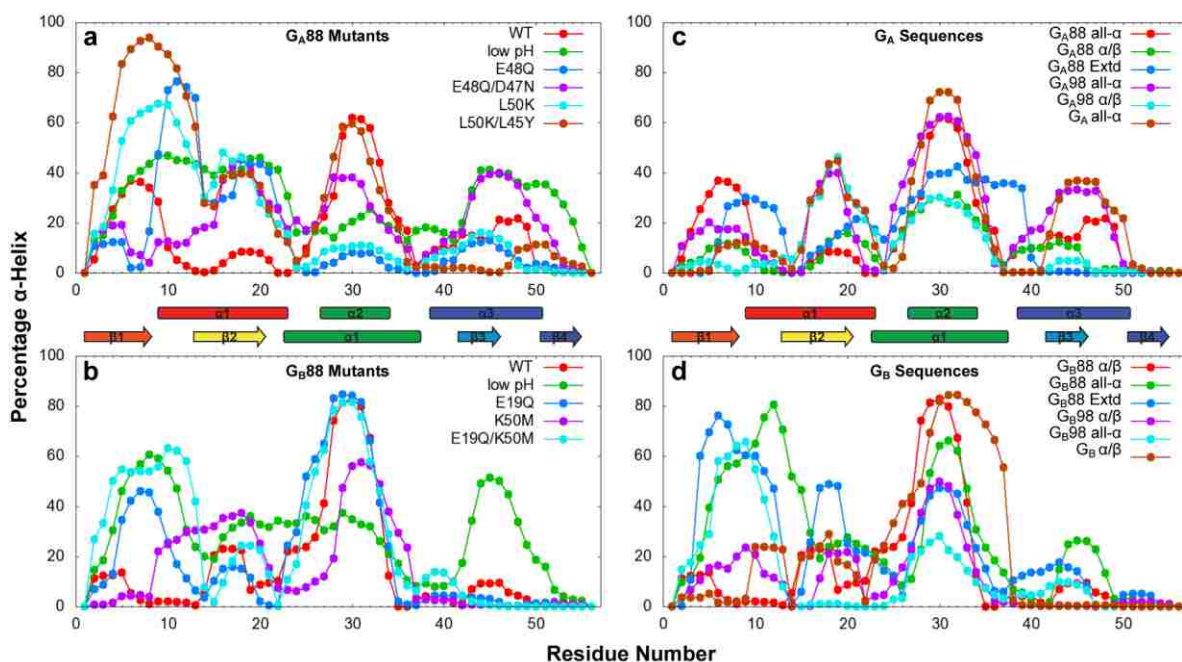
Overall, our charge-neutralizing mutations were successful in knocking out the desired interactions in the denatured state, as anticipated. Additionally, several of the interactions that we attempted to add in the denatured state were successful in not only forming but also inhibiting competing interactions. For example, the Glu19Gln mutation on  $G_B88$  that knocked out the  $\beta 1/\beta 2$  Thr1 – Glu19 salt bridge (Figure 7.4) promoted formation of the Thr1 – Glu48 salt bridge that was present in the  $G_A88$  denatured state in one simulation (Figure 7.5). This switch illustrates the competition for interaction with the N-terminus between Gln19 in  $G_B88$  (promoting formation of  $\beta 1/\beta 2$ ) and Glu48 in  $G_A88$  (promoting formation of  $\alpha 3$ ). The  $\beta 3/\beta 4$ -stabilizing salt bridge between Asp47 and Lys50 was successfully added in run 1 of the Leu50Lys and runs 2 and 3 in the Leu50Lys/Leu45Tyr mutants of  $G_A88$  at the exclusion of the  $\alpha 3$ -promoting Asp47 – Thr1 – Glu48 salt bridges.

Figure 7.6a,b shows the percentage of time each residue adopted  $\alpha$ -helical structure in the denatured state for  $G_A88$ ,  $G_B88$ , and their respective mutants.  $G_A88$  had more helix content in residues 1-10 than  $G_B88$ , where residues 1-8 form  $\beta 1$ . The Glu19Gln mutation in  $G_B88$ , as well as the Glu19Gln/Lys50Met double mutation and low pH simulations, which all destroyed the  $\beta 1/\beta 2$ -promoting Thr1 – Glu19 salt bridge, had more  $\alpha$ -helix in this region (Figure 7.6b), consistent with  $G_A88$ . Likewise, the Glu48Gln/Asp47Asn double mutation to  $G_A88$ , which removed the competing salt bridge interaction with the N-terminus, promoted loss of  $\alpha$ -helix in the same region (Figure 7.6a).

The  $G_A88$  Glu48Gln/Asp47Asn double mutant is representative of what happens towards the end of folding when Thr1 no longer coordinates Glu48 and Asp47 (Figure 7.3). The  $\alpha$ -helix distribution in this mutant was most similar to the native state, with three peaks consistent with  $\alpha 1$ ,  $\alpha 2$ , and  $\alpha 3$  (Figure 7.6a). The Leu50Lys and Leu45Tyr mutations to



$G_A88$  that were intended to promote  $G_B88$ -like structure were successful in making the desired local salt bridges that lead to kinking of the  $\beta3/\beta4$  turn (Figure 7.4). However, they also resulted in extensive  $\alpha$ -helix content in the N-terminus (Figure 7.6) and no formation of the Thr1 – Glu19 salt bridge (Figure 7.5), which is incompatible with the  $\beta1/\beta2$  hairpin. This discontinuity suggests that the two hairpins form independently in  $G_B88$ .



**Figure 7.6: Percentage  $\alpha$ -helix in the denatured state**

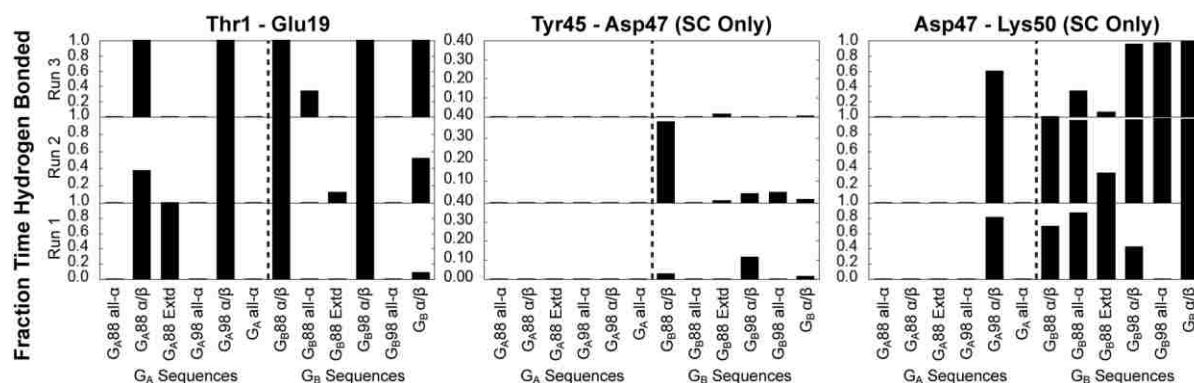
The percentage of time each residue was in  $\alpha$ -helix (as determined by hydrogen bonding pattern; Kabsch and Sander 1983) during the final 30 ns of the 3 498 K simulations is plotted for (a)  $G_A88$  and mutants, (b)  $G_B88$  and mutants, (c) simulations of  $G_A$ -based sequences with the given starting fold (all- $\alpha$ ,  $\alpha/\beta$ , or extended), and (d) simulations of  $G_B$ -based sequences with the given starting fold. Native secondary structure for  $G_A88$  and  $G_B88$  is denoted between the plots.

## 7.4.2 Analysis of all- $\alpha$ and $\alpha/\beta$ Protein Pairs

### 7.4.2.1 $G_A88$ and $G_B88$

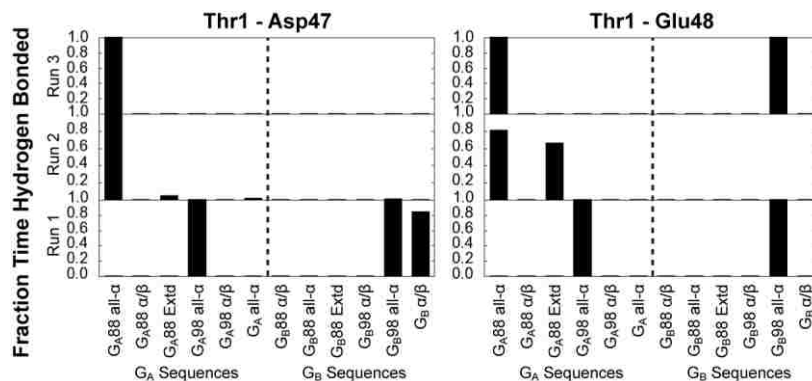
Threading the  $G_A88$  sequence on the  $G_B88$  structure helped determine whether the interactions present in the denatured state (Figure 7.2) were sequence-specific or dependent on the starting structure for the unfolding simulations. With the  $G_A88$  sequence on the  $\alpha/\beta$  structure, the  $\beta3/\beta4$ -compatible Tyr45 – Asp47 – Lys50 network was knocked out (due to the Tyr45Leu and Lys50Leu mutations in  $G_A88$ ) but not the  $\beta1/\beta2$ -promoting Thr1 – Glu19 interaction (Figure 7.7). When the  $G_B88$  sequence was threaded on the all- $\alpha$  structure, the Asp47 – Lys50 salt bridge that stabilized the  $\beta3/\beta4$  turn formed (though not the Tyr45 –

Asp47 hydrogen bond), and the  $\alpha$ 3-promoting Thr1 – Glu19 salt bridge was rarely present. The specific  $\beta$ 3/ $\beta$ 4 interactions were present to a larger extent in the simulation that started from the extended structure of  $G_B88$  than that of  $G_A88$ , but this was not the case for the  $\beta$ 1/ $\beta$ 2 interaction.



**Figure 7.7:** The  $\beta$ 1/ $\beta$ 2 Thr1 – Glu19 and  $\beta$ 3/ $\beta$ 4 Tyr45 – Asp47 – Lys50 networks from  $G_B88$  in swapped sequence/topology pairs

The fraction of time at least one hydrogen bond was present between the Thr1 – Glu19 (left), Try45 – Asp47 (middle), and Asp47 – Lys50 (right) residues pairs is plotted for each of the 3 independent runs at 498 K. In each plot, simulations of  $G_A$ -based sequences with the given starting fold (all- $\alpha$ ,  $\alpha/\beta$ , or extended) are on the left, and simulations of  $G_B$ -based sequences with the given starting fold are on the right.



**Figure 7.8:** The  $\alpha$ 3 Asp47 – Thr1 – Glu48 network from  $G_A88$  in swapped sequence/topology pairs

The fraction of time at least one hydrogen bond was present between the Thr1 – Asp47 (left) and Thr1 – Glu48 (right) residue pairs is plotted for each of the 3 independent runs at 498 K. In each plot, simulations of  $G_A$ -based sequences with the given starting fold (all- $\alpha$ ,  $\alpha/\beta$ , or extended) are on the left, and simulations of  $G_B$ -based sequences with the given starting fold are on the right.

The interaction between the N-terminus and the  $\alpha$ 3 region was not present when the  $G_B88$  structure was threaded onto the all- $\alpha$  fold, despite the fact that these residues are the same in the two sequences (Figure 7.8). It also was not present in the  $G_B88$  extended

structure but did occur transiently in run 2 of the G<sub>A</sub>88 extended structure. However, this  $\alpha$ 3-promoting interaction never formed when the G<sub>A</sub>88 sequence was threaded on the  $\alpha/\beta$  fold.

G<sub>A</sub>88 and G<sub>B</sub>88 had about the same amount of  $\alpha$ -helical content in the denatured state (Figure 7.6c,d). According to AGADIR (Muñoz and Serrano 1997), G<sub>B</sub>88 is predicted to have higher native helix content than G<sub>A</sub>88. While this is not the case in the native state, the denatured state of G<sub>B</sub>88 does have an increase in helical content relative to the native state, particularly in the region predicted by AGADIR, whereas G<sub>A</sub>88 has a decrease (Morrone *et al.* 2011). When the sequence of G<sub>B</sub>88 was put on the all- $\alpha$  structure, its helicity in the denatured state increased, particularly in the N-terminal region, whereas G<sub>A</sub>88 maintained its low helix content when it began unfolding from the  $\alpha/\beta$  topology. The same trend was observed when G<sub>A</sub>88 and G<sub>B</sub>88 were simulated starting from an extended structure.

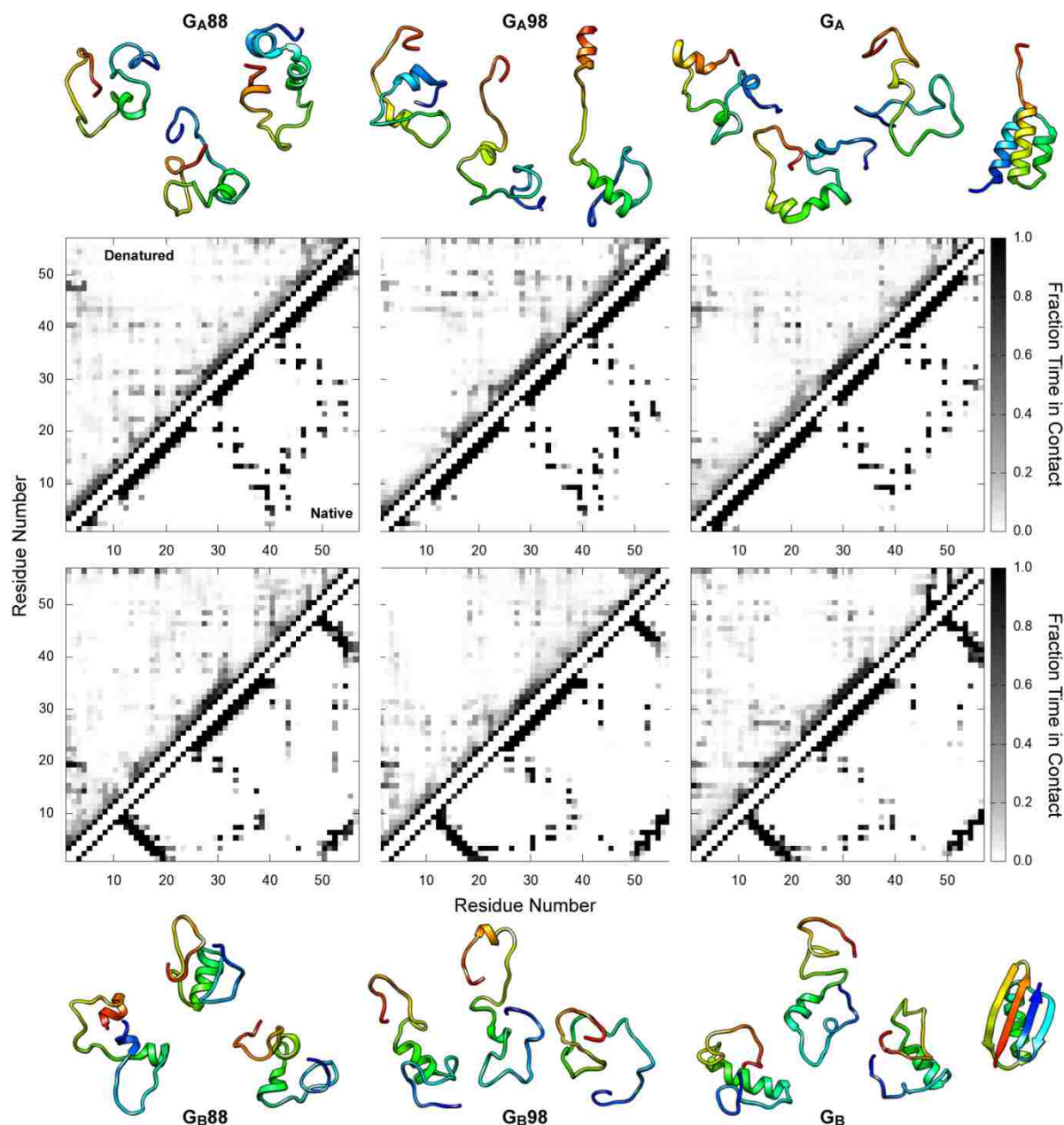
#### 7.4.2.2 G<sub>A</sub>98 and G<sub>B</sub>98

G<sub>A</sub>98 and G<sub>B</sub>98 were stable in the native 298 K simulation, with an average C $\alpha$  RMSD to the simulation starting structure of  $3.2 \pm 0.4$  and  $2.0 \pm 0.2$  Å, respectively. For comparison the C $\alpha$  RMSD of the 88% identical pair was  $3.4 \pm 0.4$  and  $2.5 \pm 0.4$  Å. The all- $\alpha$  structures had a higher C $\alpha$  RMSD due to the unstructured N-terminal tail, and if only the structured residues, 9-51, were considered, the C $\alpha$  RMSD was  $1.3 \pm 0.1$  for both G<sub>A</sub>98 and G<sub>A</sub>88. Contacts were long-lived and in the regions expected based on the all- $\alpha$  and  $\alpha/\beta$  topologies (Figure 7.9).

The denatured state of G<sub>B</sub>98 was similar to that of G<sub>B</sub>88, with the  $\beta$ 1/ $\beta$ 2 region and  $\beta$ 3/ $\beta$ 4 turn maintaining contacts and forming separately from each other (Figure 7.9). Also in agreement with G<sub>B</sub>88,  $\alpha$ -helix was largely maintained in the  $\alpha$ 1 region in G<sub>B</sub>98. G<sub>A</sub>98 was more expanded in the denatured state than G<sub>A</sub>88. There were multiple regions in the denatured state that maintained their native helicity, but they were not consistent between the three simulations. G<sub>A</sub>98 had more separation of the N-terminal region from the remainder of the protein than any of the other proteins. In G<sub>A</sub>98 there were more contacts between the  $\alpha$ 2 and  $\alpha$ 3 regions, whereas G<sub>A</sub>88 had the most contact between  $\alpha$ 1 and  $\alpha$ 2.

Unlike for the 88% identical pair, G<sub>A</sub>98 had more  $\alpha$ -helix content in the denatured state than G<sub>B</sub>98 (Figure 7.6). However, the set of simulations of the 98% identical proteins

with the highest helical content was that which the G<sub>B</sub>98 sequence was threaded on the all- $\alpha$  topology. Residues 1-23 are the same in the 88% and 98% identical sequences, but there was a wide variety of helical content in this region, especially considering the simulations that started from the alternate fold topology.



**Figure 7.9: Contacts maps and structures from the denatured state**

Contacts are plotted for the denatured state (top left; last 30 ns of the 3 50-ns 498 K simulations) and native state (bottom right; all 30 ns of the single 298 K simulation) of the 88% (left), 98% (middle), and 16% (right) identical G<sub>A</sub> (above) and G<sub>B</sub> (below) pairs. Above and below the contact maps are the respective final structures from the three independent 498 K simulations. At the right are the NMR structures of G<sub>A</sub>88 and G<sub>B</sub>88. Structures are colored red  $\rightarrow$  blue N  $\rightarrow$  C and shown in ribbons.

The Thr1 – Glu19 and Tyr45 – Asp47 – Lys50 networks found in G<sub>B</sub>88 were also present in G<sub>B</sub>98 but not G<sub>A</sub>98 (Figure 7.7). However, since residue 50 is also a Lys in G<sub>A</sub>98 (it is a Leu in G<sub>A</sub>88), the Asp47 – Lys50 salt bridge was present when the G<sub>A</sub>98 sequence was threaded onto the  $\alpha/\beta$  structure. The Asp47 – Thr1 – Glu48 network from G<sub>A</sub>88 was present in run 1 of G<sub>A</sub>98 (Figure 7.8), but it was also present in G<sub>B</sub>98 when starting from the all- $\alpha$  structure. Indeed, G<sub>A</sub>98 is less stable than G<sub>A</sub>88 ( $T_m = 37$  vs.  $69$  °C; Alexander *et al.* 2009), and this may be partly due to the competition between the local Asp47 – Lys50 salt bridge that stabilizes the  $\beta3/\beta4$  turn and formation of the native  $\alpha3$  helix.

#### 7.4.2.3 G<sub>A</sub> and G<sub>B</sub>

At 298 K, G<sub>A</sub> and G<sub>B</sub> had C $\alpha$  RMSDs of  $2.3 \pm 0.3$  and  $2.2 \pm 0.4$  Å to the starting structure, respectively. G<sub>A</sub> had a core C $\alpha$  RMSD of  $1.1 \pm 0.1$  Å to the starting structure. The reason the C $\alpha$  RMSD was lower for G<sub>A</sub> than G<sub>A</sub>88 and G<sub>A</sub>98 was because  $\alpha1$  extends from residues 4-24 in G<sub>A</sub> (He *et al.* 2006) as opposed to residues 9-23 in G<sub>A</sub>88. The extra 5 residues of helix at the N-terminus of  $\alpha1$  were maintained in the native simulation; G<sub>A</sub> had  $75 \pm 2\%$   $\alpha$ -helix at 298 K whereas G<sub>A</sub>88 and G<sub>A</sub>98 had only  $67 \pm 3$  and  $70 \pm 3\%$ , respectively.

G<sub>B</sub> had helix in the  $\alpha1$  region in the denatured state like the other G<sub>B</sub>-based sequences, and most of the contacts in the denatured state occurred between  $\beta1$  and  $\beta2$  (Figure 7.9). There were fewer interactions between  $\beta3$  and  $\beta4$  in this protein, and G<sub>B</sub> was the only protein of G<sub>B</sub>, G<sub>B</sub>88, and G<sub>B</sub>98 that had much interaction between the two hairpin regions. G<sub>A</sub>, more like G<sub>A</sub>88 than G<sub>A</sub>98, had more contacts between  $\alpha1$  and  $\alpha2$  than between  $\alpha2$  and  $\alpha3$ .

Glu19 is conserved in all six of the sequences in this study (Figure 7.1). As in the other  $\alpha/\beta$  structures, the Thr1 – Glu19 salt bridge was present in the denatured state of G<sub>B</sub> (Figure 7.7). Likewise, Tyr45, Asp47, and Lys50 were conserved in G<sub>B</sub>, and the Asp47 – Lys50 salt bridge was present, as was the Tyr45 – Asp47 salt bridge, to a minor extent. This network was not present in G<sub>A</sub>, where these three residues are Leu, Asn, and Leu, respectively. The Thr1 – Asp47 salt bridge was only present in run 1 of G<sub>B</sub>, but not in G<sub>A</sub> where residue 47 is Asn (Figure 7.8). Despite having the lowest sequence identity of the three pairs, G<sub>A</sub> and G<sub>B</sub> had the most similar  $\alpha$ -helix distributions in the denatured state (Figure 7.6). G<sub>A</sub> had overall more  $\alpha$ -helix than G<sub>B</sub>, especially in the C-terminal region.

## 7.5 Discussion

The mutants we made here were successful in breaking the hydrogen bonds that we previously identified in the denatured state of G<sub>A</sub>88 and G<sub>B</sub>88 (Figure 7.2; Morrone *et al.* 2011) through charge-neutralizing substitutions. The mutations also changed the distribution of  $\alpha$ -helix in the denatured state, with the most deviation in the N-terminus. Surprisingly, the N-terminus is the region with the least sequence variation between the proteins.

The Glu48Gln mutation in G<sub>A</sub>88 was successful in knocking out the  $\alpha$ 3-promoting Thr1 – Glu48 salt bridge (Figure 7.5), and it promoted formation of the competing,  $\beta$ 1/ $\beta$ 2-promoting Thr1 – Glu19 interaction (Figure 7.4). The Glu48Gln/Asp47Asn double mutant had extensive  $\alpha$ -helix content (Figure 7.6), indicative of what occurs in the final steps of folding when the Asp47 – Thr1 – Glu48 reaction is released in the absence of the Thr1 – Glu19 salt bridge. We predict that these mutations would shift the equilibrium of G<sub>A</sub>88 towards the  $\alpha/\beta$  topology.

The Leu50Lys/Leu45Tyr double mutant in G<sub>A</sub>88 successfully created the  $\beta$ 3/ $\beta$ 4-promoting Tyr45 – Asp47 – Lys50 network (Figure 7.4) while knocking out the native,  $\alpha$ 3-compatible Thr1 – Asp47 – Glu48 network (Figure 7.5). However, it did not also promote formation of  $\beta$ 1 and  $\beta$ 2. This duality supports the hypothesis that the two hairpins form independently. This mutation is predicted also to shift the equilibrium towards the  $\alpha/\beta$  topology and lower the stability.

The Glu19Gln mutation in G<sub>B</sub>88, which knocked out the  $\beta$ 1/ $\beta$ 2-promoting Thr1 – Glu19 salt bridge, resulted in helix formation in the N-terminus (Figure 7.6) and the presence of the competing Thr1 – Glu48 salt bridge (Figure 7.5). The Lys50Met mutation in G<sub>B</sub>88 knocked out the Tyr45 – Asp47 – Lys50 interaction, as anticipated (Figure 7.4). The Glu19Gln/Lys50Met double mutant also resulted in loss of these interactions and formation of G<sub>A</sub>88-like helical structure. In these mutants, especially the double mutant, we predict more structure in the denatured state, similar to what was observed for wild type G<sub>B</sub>88 under acidic conditions.

There was some consistency in contact patterns in the denatured state between the G<sub>A</sub>-based vs. G<sub>B</sub>-based sequences (Figure 7.9). For example, there were more short-range

interactions, especially between the  $\beta 1$  and  $\beta 2$  regions in all three  $G_B$ -based proteins. The two hairpins tended to interact with themselves in the denatured state, separated by the  $\alpha 1$  helix, suggesting that the hairpins form independently. The  $G_A$ -based proteins had more long-range contacts, shown in the top-left corner of the contact maps. In particular, there were consistently interactions between the N-terminal and  $\alpha 3$  regions. In  $G_B$ , unlike in the  $G_B88$  and  $G_B98$ , there was often a strong interaction between the N- and C-termini.

The specific hydrogen bonds we saw when threading one sequence onto the opposite fold were more dependent on the starting conformation than the sequence itself, especially the competition between Glu19 and Asp47/Glu48 for interacting with the N-terminus. Folding to the alternate conformation only occurs  $< 3\%$  of the time for the 88% identical pair (He *et al.* 2006) and  $< 10\%$  of the time in the 98% identical pair (Alexander *et al.* 2009). Notably, we were simulating the minor folding pathway in reverse in these cases, so our simulations showed the interactions that would lead the sequences to the opposite fold. It is therefore not surprising that the interactions that we saw correlated with the final folded structure more than the protein sequence.

It is possible that our simulations, if run for much longer periods of time, would converge on the interactions in the denatured state that were observed for the  $G_A88/G_B88$  pair. We have found that salt bridges in the denatured state last hundreds of nanoseconds (McCully *et al.* 2010), so it may not be computationally reasonable to extend our simulations long enough to see whether the structure-specific interactions break and the anticipated interactions form.

## 7.6 Conclusions

We were successful in knocking out the putative, topology-directing interactions previously identified in the denatured state of the  $G_A88/G_B88$  pair of proteins. In the cases of the Leu50Lys/Leu45Tyr double mutant to  $G_A88$  and the Glu19Gln mutant to  $G_A88$ , the mutated residues that promoted formation of interactions in the denatured state that were compatible with the opposite topology. These hypotheses are currently being tested experimentally in the Gianni group.

The  $G_B$ -based sequences, when unfolded from the native  $\alpha/\beta$  structure, generally had more interactions within the two hairpin regions but few interactions between them. The  $G_A$ -based sequences, when unfolding from the all- $\alpha$  structure, had longer-range interactions, especially between the N-terminus and  $\alpha 3$ . Our simulations of unfolding using the opposite fold as the starting structure maintained interactions more consistent with the starting fold than sequence, suggesting our simulations were too short.



## Chapter 8: Folding and Dynamics of Engineered Proteins

### 8.1 Introduction

Scientists have been rationally engineering proteins for over two decades, and great progress has been made in a relatively short amount of time. The simplest designs include repacking cores of small globular proteins, and the most intricate involve designing and inserting a catalytic site into a protein scaffold. The principles learned from studying the simple designs are important to incorporate into the design process for engineering the more complex proteins. While much attention is given to the design target and its validation by structural techniques, less thought has been directed toward the role of dynamics in these designed proteins or how the folding/unfolding pathway has been affected. Perhaps the most intriguing information to come from the biophysical studies of such proteins are the ways in which the engineers have inadvertently created especially strange folding pathways, intermediates, and native state dynamics. These observations of “unnatural” proteins provide insight into what nature has purposefully avoided or preferred in the evolution of functional, stable proteins, which can then be incorporated into future design strategies.

This review discusses an array of studies that probe the folding and dynamics of designed globular proteins in comparison with their naturally occurring counterparts. These studies use a range of biophysical techniques including both computational and experimental approaches, and the main techniques discussed in this chapter are introduced briefly below. Molecular dynamics (MD) simulations model the structures of proteins using the laws and equations of Newtonian physics. MD provides theoretical atomic-level descriptions of the motions and interactions in proteins under various temperature and solvent conditions. Experimental techniques help describe a protein’s structure, kinetics, and thermodynamics. X-ray crystallography is used to obtain model structures of proteins based on their electron density and can identify flexible regions of the protein. Nuclear magnetic resonance (NMR) studies provide a wide variety of information about interactions within proteins and their backbone and side chain dynamics. Circular dichroism (CD) can detect signatures of

secondary structure elements such as  $\alpha$ -helices,  $\beta$ -sheets, and random coils. Fluorescence studies monitor the solvent exposure of certain amino acids such as tryptophan and other nonnatural residues. When CD and fluorescence studies are combined with thermal or chemical denaturation, folding rate constants ( $k_f$ ) and half lives ( $t_{1/2}$ ), free energies of folding ( $\Delta G_U$ ), melting temperatures ( $T_m$ ), and more can be estimated to describe the kinetic and thermodynamic stability of the proteins. The focus of this review is on proteins that were designed and whose dynamics and folding pathway were studied by computational methods as well as the experimental validation of these findings.

The first section of this review addresses relatively simple proof-of-principle designs that repack side chains onto a given backbone structure. FSD-1 (a zinc finger fold) and the three-helix bundles discussed below are examples of simple hydrophobic core repacking. The design of  $\alpha_3D$  involved a new backbone conformation, and Top7 was specifically designed to have a fold topology not found in nature. This section ends with a more systematic study of several protein designs to compare the amount of variation seen in proteins all designed with the same computational method. The second section describes the dynamics of proteins that were designed for function. A pair of 4-helix bundle proteins designed to bind two divalent cations is discussed as well as a protein that was designed to bind a dipeptide. In addition, two enzyme designs are presented that catalyze a retro-aldol and Kemp elimination reaction, respectively.

## 8.2 Proof-of-Principle Protein Designs

The early protein designs aimed simply to create a stable structure on an existing backbone scaffold by changing the amino acid sequence. As protein folding is driven by burial of hydrophobic residues, protein design has largely focused on sequestering hydrophobic amino acids in the core of the protein and placing polar and charged amino acids on the surface to make favorable interactions with the solvent. Choosing one of the 20 naturally occurring amino acids along with its orientation at every position to interact favorably with its neighbors in the sequence is a huge combinatorial problem even for small proteins. Computational methods aided in simplifying this problem. Rotamers, structures of amino acid side chains with the most probable  $\chi$  angles, were used to vastly decrease the

structural space that each residue could assume. Scoring functions were used to select structures that buried hydrophobic residues away from the surface, made hydrogen bonds, and had no atomic clashes. Monte Carlo methods were used to sample various structures to (hopefully) reach the best energy score, or most stable structure.

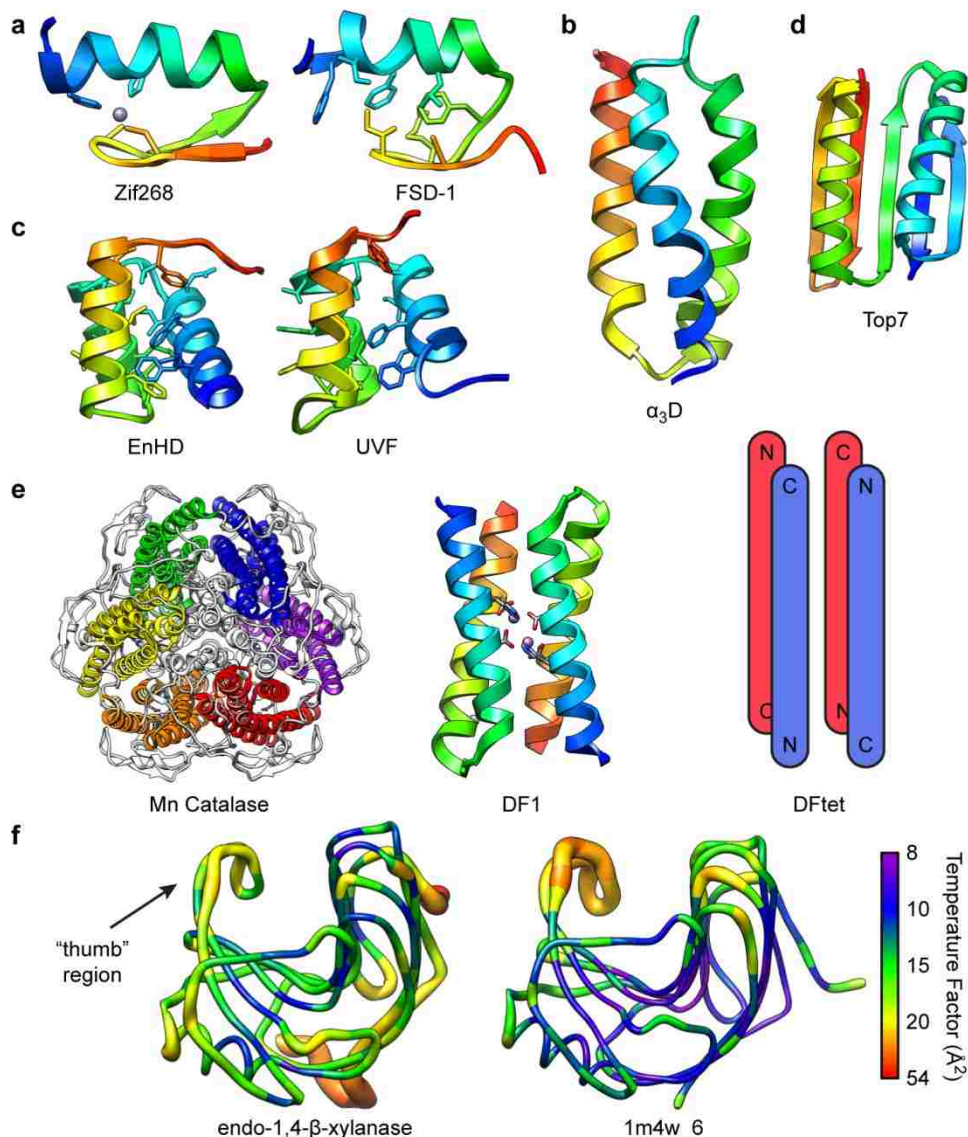
Some of these relatively simple designed proteins were expressed and characterized biophysically and will be discussed in this section. Studying these designs and comparing them with naturally occurring proteins is first and foremost necessary for analyzing and improving the scoring functions and design protocol. Yet perhaps more importantly, analyzing the properties of proteins that have not evolved naturally over millions of years can give new insight to selective pressures in nature.

### **8.2.1 FSD-1, a Heterogeneous Native State and Complicated Folding Pathway**

The first fully automated protein sequence design was presented in 1997 by the Mayo Group (Dihyat and Mayo 1997). This 28-residue protein, named FSD-1, was built based on the  $\beta\beta\alpha$  fold of Zif268 (Pavletich and Pabo 1991), a zinc finger (Figure 8.1a). However, in place of the bound zinc ion in Zif268, a hydrophobic core was added. Their methodology used a dead-end elimination algorithm to select side chain rotamers and score them based on solvation, packing, hydrogen bonding, secondary structure propensity, and van der Waals terms. Further, core positions were limited to hydrophobic amino acids and surface to polar, and boundary position were unrestricted. The resulting structure was soluble, folded, and monomeric (Dahiyat and Mayo 1997). The authors reported weakly cooperative and completely reversible folding with a melting temperature ( $T_m$ ) of 39 °C by circular dichroism (CD) measurements. An NMR structure was solved and had a root-mean squared deviation (RMSD) of 2.0 Å to the designed structure over the backbone atoms of residues 3-26.

The small size of FSD-1 made it a popular target for computational folding studies. Two all-atom, explicit solvent molecular dynamics (MD) studies by Lei *et al.* (2004, 2006) probed the native state and transition state as well as the unfolding and early refolding pathways. FSD-1 has a broad native state with considerable structural deviation at 27 °C, despite remaining compact, as reflected in the radius of gyration ( $R_g$ ) and backbone RMSD measurements of ~9.5 Å and ~2.0 Å, respectively. When the temperature in the simulations was raised to 57 °C, above the previously measured  $T_m$ , the structure became even more

heterogeneous. There was a greater increase in root-mean squared fluctuation (RMSF) about the mean structure for the  $\beta$ -hairpin than the  $\alpha$ -helix at the higher temperature, and half of the tertiary contacts between the hairpin and helix were lost.



**Figure 8.1: Structures of the proof-of-principle designs**

Structures are colored red  $\rightarrow$  blue from the N  $\rightarrow$  C-terminus unless otherwise indicated. (a) Zif268 and FSD-1 with zinc-binding residues or hydrophobic core residues shown in sticks. FSD-1 can fold without a zinc ion due to its redesigned hydrophobic core. (b)  $\alpha_3D$ . (c) EnHD and UVF with core residues shown in sticks. The core of UVF is packed more loosely than in the natural protein. (d) Top7. (e) Manganese catalase with its six 4-helix bundle active sites colored by chain (right). DF1 was designed based on the 4-helix bundles of Mn catalase (middle). A schematic of the antiparallel nature of two pairs of helices, DFtet-A and DFtet-B, which form a heterotetramer in DFtet. (f) Endo-1,4- $\beta$ -xylanase and designed 1m4w\_6, which was unsuccessful in binding the D-Ala-D-Ala dipeptide. Structures are shown in worms and colored based on the B-factors in the respective crystal structures. The “thumb” region was more flexible in the designed protein.

High temperature unfolding simulations gave the first look at the denatured state and unfolding pathway of FSD-1 (Lei *et al.* 2006). The  $R_g$  of the denatured state ranged from native levels ( $\sim 9.5$  Å) to much higher ( $\sim 20$  Å), and the RMSD reached  $>10$  Å. The  $\beta$ -hairpin was completely lost, whereas there was some fluctuating  $\alpha$ -helix present. In the unfolding pathway, the  $\beta$ -hairpin was observed to melt first followed by the  $\alpha$ -helix. In refolding simulations, the early steps of folding were slowed by the formation of a nonnative salt bridge between Glu17 and Arg19 in the  $\alpha$ -helical region as well as some helix content in the  $\beta$ -hairpin region. Ala5 and Lys6 nucleated helix formation in the denatured state, likely due to their high helical propensities, suggesting that limiting unfavorable secondary structure propensity and the potential to form nonnative salt bridges in the denatured state should be taken into account in the design stage.

Two all-atom, explicit solvent replica exchange MD simulations (REMD) of FSD-1 were performed by Li *et al.* (2007) to probe the native, transition, and denatured states, as well as the  $T_m$ . The authors calculated a  $T_m$  of 147 °C based on the fraction of native contacts over the various simulated temperatures, which is considerably higher than the 39 °C measured experimentally by Dihatay *et al.* (1997). Individually, the  $T_m$  was 169 °C for the  $\alpha$ -helix based on the fraction of residues in  $\alpha$ -helix and 136 °C for the  $\beta$ -hairpin as measured by the fraction of native contacts in residues 1-14. The authors also defined a putative folding intermediate which had a well structured  $\alpha$ -helix, but little to no  $\beta$ -hairpin content. Folding was initiated by collapse of the hydrophobic core along with formation of the  $\alpha$ -helix to reach the intermediate. It then continued with formation of the  $\beta$ -hairpin and final hydrophobic packing.

A later REMD study by a different group found slightly lower melting temperatures (Wu and Shea 2010). Based on the  $R_g$  of the hydrophobic core residues (Ala5, Ile7, Phe12, Leu18, Phe21, Ile22, and Phe25) and RMSD of the  $\alpha$ -helix,  $\beta$ -hairpin, and whole protein, these authors calculated a  $T_m$  of 32 °C for the overall tertiary structure, 34 °C for the  $\beta$ -hairpin, 50 °C for the hydrophobic core, and 100 °C for the  $\alpha$ -helix. In their heat capacity calculations, they found broad transitions at 48 and 114 °C, and they attributed the first one to the  $\beta$ -hairpin unfolding while exposing the hydrophobic core and the second to unfolding of the  $\alpha$ -helix. Further folding simulations in the same study depicted a folding pathway

initiated by rapid helix formation then slowly followed by formation of the  $\beta$ -hairpin and overall native tertiary structure.

More work on the kinetics of FSD-1 folding was completed by Sadqui *et al.* (2009), pinning the  $T_m$  at 89 °C by CD and FRET in thermal denaturation experiments. Fluorescence was measured by incorporating two nonnatural aromatic amino acids at the C-terminus (residue 27) and in the  $\beta$ -hairpin (residue 7). This mutant, FSD-1ss, was an ultra-fast folder fitting a double-exponential decay with relaxation half lives of 0.15 and 4.5  $\mu$ s. In agreement with the previous studies, the native state was structurally degenerate, and the NMR solution structure of FSD-1ss was more expanded than the wild type structure.

As it was becoming clear that the folding pathway of FSD-1 was anything but straightforward, further thermodynamic experiments in conjunction with REMD simulations by Feng *et al.* (2009) were presented. Repeated CD and experiments over 4-80 °C gave very similar results to the original experiments (Dahiyat and Mayo 1997), but Feng *et al.*, while agreeing with a  $T_m$  of 41 °C, noted that there was no baseline and the transition was broad across all temperatures. Differential scanning calorimetry (DSC) experiments also showed a broad melting transition around 41 °C. However, in comparing DSC results between FSD-1 and a 50-residue  $\alpha$ -helical peptide, the authors proposed that the unfolding transition observed at 41 °C was due to a helix-to-coil transition rather than global formation of the hydrophobic core. These CD and DSC results agreed with their REMD and other MD/REMD simulations indicating a flexible, heterogeneous native state with a barely stable  $\beta$ -sheet. Additionally, their REMD simulations again pointed to a much higher  $T_m$  of 125 °C.

The folding pathway of FSD-1 has confounded experimentalists and computational scientists over the last decade. Computational folding, unfolding, and REMD studies more or less consistently agree upon a folding pathway beginning with fast formation of the  $\alpha$ -helix into an intermediate followed by formation of the  $\beta$ -hairpin and collapse of the hydrophobic core. Both computation and experiment agree that the native state is highly heterogeneous and that the  $\beta$ -hairpin is only barely stable. Measuring the  $T_m$ , on the other hand, has proved to be rather contentious. Most computational studies pin a global unfolding  $T_m$  well over 100 °C, with the  $\beta$ -hairpin lower than the  $\alpha$ -helix when calculated separately. The  $T_m$  measured experimentally by various methods is  $\sim$ 40 °C with a very broad transition

(or 89 °C for FSD-1ss). The finding that FSD-1 likely folds in two phases may simplify the issue. Experimental techniques may be picking up the late folding phase for formation of the  $\beta$ -hairpin and hydrophobic core, which one REMD study identified as occurring over 32-50 °C, or on a helix-to-coil transition. In addition, experimental measurements have only been taken up to ~100 °C, whereas the  $T_m$  of the early phase of helix formation is predicted to be >100 °C. This stage of folding may have a very short half-life, making it also difficult to detect kinetically ( $t_{1/2} = 150$  ns for FSD-1ss). However, REMD is known to exaggerate melting temperatures, and in a matched study of a natural protein, REMD gave an abnormally high  $T_m$  while conventional MD unfolding simulations were in good agreement with the experimental melting temperature (Beck *et al.* 2007).

It is informative to compare the folding pathway of FSD-1 (or as much of it as we can discern!) to its naturally occurring template protein, Zif268, a zinc finger transcription factor in mice. While FSD-1 was designed to fold without a zinc ion, Zif268 requires one (Miura *et al.* 1998). In the absence of Zn(II), the 27-residue peptide forms mostly  $\beta$ -sheet in solution. When Zn(II) is added, it is first coordinated by Cys3 and Cys6 (both in the native  $\beta$ -sheet), forming the hairpin turn, breaking the nonnative  $\beta$ -sheet present in the apo form, and increasing helical content. His19 and His23 (both in the native  $\alpha$ -helix) bind the zinc ion next, helping form the  $\alpha$ -helix and the final tertiary  $\beta\beta\alpha$  fold. In contrast, the  $\alpha$ -helix is the most stable part of FSD-1 and may be present to some extent in the denatured state. While zinc ion coordination drives formation of the  $\beta$ -hairpin in Zif268, hydrophobic collapse is responsible for the hairpin formation in FSD-1. The zinc ion tightly holds the geometry of the hairpin in place relative to the helix in the native state of Zif268, whereas FSD-1's native state is highly heterogeneous and the  $\beta$ -sheet is barely stable. The fluidity of the hydrophobic core imposes little geometrical restraint on the  $\beta$ -sheet and  $\alpha$ -helix in FSD-1.

### 8.2.2 $\alpha_3$ D, a Dynamic Core Leads to Fast Folding and Thermal Stability

While FSD-1 is a 28-residue protein designed based on the backbone of an existing protein,  $\alpha_3$ D is a 73-residue 3-helix bundle and is not based on a naturally occurring protein (Figure 8.1b; Walsh *et al.* 1999). The DeGrado group used their previously designed “Coil-Ser” coiled coil as the starting point for  $\alpha_3$ C, shortening the helices, adding helix capping boxes, designing interhelix electrostatic interactions, and repacking the hydrophobic core

using a genetic algorithm (Bryson *et al.* 1998). An additional five mutations were added to create  $\alpha_3$ D: Met1 and Gly2 were prepended to the sequence; and surface mutations, Glu9Gln, Ser16Thr, and Ser65Asp were introduced to increase sequence variation between the three helices (Walsh *et al.* 1999). An NMR solution structure of  $\alpha_3$ D was determined, and it had a backbone RMSD of 1.9 Å from the designed model. 14 of the 18 hydrophobic residue side chain  $\chi_1$  angles could be defined as a single rotamer by H $\alpha$ -H $\beta$  coupling constants and an additional two by NOE restraints. The two remaining residues, Trp4 and Tyr45, located near the termini of helices 1 and 2, respectively, adopted multiple rotamers. Further NMR studies probed the internal motions of  $\alpha_3$ D (Walsh *et al.* 2001a). Backbone N-H  $S^2$  order parameters were similar to those observed in naturally occurring proteins, ranging from 0.80 to 0.85 in helical regions and lower in loops. Similarly, backbone C $\alpha$ -H $\alpha$  order parameters were very high ( $\sim$ 0.90) in agreement with natural proteins. However, the  $S_{\text{axis}}^2$  values, averaged by type of side chain methyl group, were about 0.20 lower than the same groups in natural proteins, indicating that hydrophobic side chains in  $\alpha_3$ D had larger amplitude motions. This distribution of order parameters is indicative of a rigid, largely immobile backbone and a highly dynamic hydrophobic core.

Thermodynamic, kinetic, and MD studies gave the first indication that  $\alpha_3$ D was a very thermally stable, ultra-fast folding protein (Walsh *et al.* 2001b, Zhu *et al.* 2003). In water its  $T_m$  was  $>90$  °C, but when the pH was lowered to 2.2, a  $T_m = 73$  °C could be measured. At 25 °C its folding time was  $\sim$ 4.8  $\mu$ s, and it had a maximum folding rate at 49 °C with  $t_{1/2} = 3.2$   $\mu$ s. At low denaturant concentrations, the presence of a putative folding intermediate was detected kinetically. MD simulations suggested the structure of this intermediate had helix 2 docked to a partially denatured helix 1 and helix 3 in solution. MD detected various unfolding pathways, and in agreement with IR data, the denatured state was observed to have fluctuating  $\alpha$ -helix as well as some  $\pi$ - and  $3_{10}$ -helix. The fast folding rates were attributed to the presence of helix and medium-range interactions in the denatured state. The transition state was fairly loose and could potentially be approached by the many folding pathways observed by simulation. Additionally, the core of  $\alpha_3$ D has no geometrically restrictive polar interactions, so the relatively nonspecific hydrophobic packing can occur quickly.



Two mutants of  $\alpha_3D$ , Ala60Leu and Ala60Ile, added additional hydrophobic volume to its core (Walsh *et al.* 2001b). While this sort of addition usually destabilizes proteins by 1.5-5.0 kcal/mol due to disrupted hydrophobic packing (Liu *et al.* 2000), it resulted in increased stability in  $\alpha_3D$ . While  $\alpha_3D$  has  $\Delta G_U = 6.8$  kcal/mol, the Ala60Leu and Ala60Ile mutants had  $\Delta G_U = 7.9$  and 7.6 kcal/mol, respectively. Based on the  $^1H$  and  $^{13}C$  chemical shift dispersion in NMR spectra, both mutant proteins had more dynamic hydrophobic cores than  $\alpha_3D$ . The authors concluded that  $\alpha_3D$  must be more malleable than naturally occurring proteins not only to tolerate the increased volume in its core but actually to be stabilized by the mutations.

Like FSD-1,  $\alpha_3D$  has a highly dynamic hydrophobic core, which seems to lead to its high thermal stability and rapid folding. A fluorescent mutant of FSD-1 was likewise observed to fold at “ultra-fast” speeds ( $t_{1/2} = 4.5 \mu s$ ), and both proteins have putative folding intermediates characterized by near-native amounts of  $\alpha$ -helix. Both proteins were designed largely on the principle of packing the core with hydrophobic residues and putting polar and charged residues on the surface. Yet, without buried, geometrically specific polar interactions, both proteins have a very dynamic hydrophobic core, which acts as an entropy sink, stabilizing the native state.

### 8.2.3 3-Helix Bundle Thermostabilized Proteins

The Mayo Group worked for several years designing and redesigning variants of the Engrailed Homeodomain (EnHD) 3-helix bundle, a 56-residue, ultra-fast folding transcription factor from *D. melanogaster*. Their goal was always to thermostabilize the protein, but their methods evolved over time. Their initial design, NC0, incorporated polar and charged residues on the surface of the protein using their in-house program, ORBIT (Marshall *et al.* 2002). While 11 of EnHD's 29 surface residues are charged, resulting in an overall charge of +7, NC0 has 22 charged surface residues and no overall charge. Despite predictions of more hydrogen bond and salt bridge interactions in NC0, it had  $\Delta G_U = 2.3$  kcal/mol and a  $T_m = 53$  °C compared with  $\sim 1.8$  kcal/mol and 52 °C for EnHD, respectively (Mayor *et al.* 2000, Gianni *et al.* 2003). However, when only the four amino acids with the highest N-capping propensity (Ser, Thr, Asn, Asp) were allowed at the N-capping positions on the three helices, and positively (Lys, Arg, His) and negatively (Asp,

Glu) charged residues were excluded from N- and C-terminal locations, respectively, NC3-Ncap, was much more stable than EnHD or NC0 with  $\Delta G_U = 5.9$  kcal/mol and  $T_m = 88$  °C. Unlike FSD-1 and  $\alpha_3D$ , whose stabilities were largely due to increased entropy, NC3-Ncap was stabilized despite adding low-entropy, geometrically restrictive surface interactions. The decreased entropy in NC3-Ncap was likely compensated for by the increased enthalpic contribution of the engineered surface interactions.

Two additional variants were designed, this time taking into account buried residues as well as surface and helix-capping positions (Shah *et al.* 2007). Buried positions were limited to Ala, Val, Leu, Ile, Phe, Tyr, and Trp whereas surface positions were chosen from Ala, Asp, Asn, Glu, Gln, His, Lys, Ser, Thr, and Arg. Capping residues were chosen as in the previous study (Marshall *et al.* 2002). Monte Carlo methods and in-house scoring functions identified UVF and UMC, with 39 and 40 mutations relative to EnHD respectively, as the best designs. CD and NMR studies on both sequences indicated well-folded,  $\alpha$ -helical proteins with  $\Delta G_U = 2.3$  kcal/mol and  $T_m > 99$  °C. NOEs were collected for UVF, and its structure was determined (Figure 8.1c). It indeed formed a 3-helix bundle with the best defined regions being helices 1 and 2, the N-terminal turns of helix 3, and the helix 2-3 turn in agreement with  $^3J_{\text{HNH}\alpha}$  coupling constant values and  $S_{\text{axis}}^2$  side chain order parameters. Similarly, the C-terminus of helix 3 in EnHD is only loosely folded (Clarke *et al.* 1994, Religa 2008).

Kinetic folding studies of both NC3-Ncap and UVF (also referred to as ENH-FSM1 in the literature) were performed as well (Gillespie *et al.* 2003). Laser T-jump relaxation studies found folding dynamics to fit a double exponential decay and measured  $t_{1/2} = 23.9$   $\mu\text{s}$  in water for NC3-Ncap and 8.8  $\mu\text{s}$  for UVF compared with  $\sim 15$   $\mu\text{s}$  for EnHD (Mayor *et al.* 2003b). The relative solvent accessibility of the transition state was 0.52 and 0.39 for NC3-Ncap and UVF, respectively compared with 0.85 for EnHD, indicating a more expanded structure for the engineered proteins.

MD studies of EnHD and UVF showed that the thermostabilized protein was more dynamic at room temperature than EnHD based on RMSD/F (McCully, Beck, and Daggett, *Accepted*). When the temperature of the simulations was raised to 100 °C, UVF maintained the same level of dynamics, whereas EnHD became more flexible. Examining the number

and types of contacts in these proteins sheds some light as to how UVF can tolerate such heightened dynamics. Removing polar and charged residues from the core of UVF resulted in fewer less-favorable hydrophobic-polar interactions between buried residues and between buried and surface residues. So, while UVF successfully maximized enthalpically favorable interactions, as designed, it also seemed to benefit from higher entropy, as seen by its heightened dynamics relative to EnHD.

NC0, NC3-Ncap, UVF, UMC, and others were all designed with the goal of thermostabilizing EnHD, and three proteins (NC3-Ncap, UVF, and UMC) were successful in reaching a higher  $\Delta G_U$  and  $T_m$ . Despite designing for thermostability, NC3-Ncap and UVF both retained the fast folding properties of the natural protein. Similarly,  $\alpha_3D$  and FSD-1ss, though not designed specifically to be thermostable, were found to be fast folding. On the other hand, while the folding pathways of these three-helix bundles were relatively simple, that of FSD-1 was definitely not straightforward. While all of these proteins were engineered to be stable and employed a similar basic design strategy (burying hydrophobic residues and adding polar interactions on the surface), there was never a goal of creating fast-folding proteins. Once the rules surrounding protein folding are better understood, perhaps folding pathways themselves will be able to be engineered for speed or to have (or not have) specific folding intermediates.

#### **8.2.4 Top7, a Novel Fold Topology**

FSD-1 was built based on the backbone of Zif268, and all of the Mayo group three-helix bundle designs were based on EnHD.  $\alpha_3D$  did not specifically use a naturally occurring protein as a template, but it was based on previous Coil-Ser designs and was ultimately a three-helix bundle. Top7, on the other hand, was intentionally designed to have a unique backbone topology never observed in nature (Kuhlman *et al.* 2003). A 93-residue,  $\beta\beta\alpha\beta\alpha\beta\beta$  globular fold was created with the two  $\alpha$ -helices lying parallel on the  $\beta$ -sheet (Figure 8.1d). While all of the designed proteins discussed here thus far have only optimized side chain residues and orientations on a fixed backbone scaffold, Kuhlman *et al.* used their program, Rosetta, to iteratively optimize the backbone and sequence/rotamers. In the initial sequence, the 22 surface positions were limited to polar amino acids and all others to anything but cysteine. The final sequence had no significant similarity to any naturally occurring proteins,

just as the backbone topology was novel. Based on biophysical characterization, Top7 was highly soluble, monomeric, thermally stable ( $T_m > 99$  °C), a cooperative folder, and exceptionally stable ( $\Delta G_U = 13.2$  kcal/mol). A high-resolution crystal structure was determined, and the backbone RMSD between the crystal structure and designed model was 1.2 Å. Side chain orientations also agreed well between the two structures.

Initial folding studies on Top7 in 2004 indicated it was extremely stable up to ~6 M guanidine and that the folding pathway was very complicated (Scalley-Kim and Baker 2004). At high denaturant concentration folding was two-state, but at lower concentrations it became three-state with  $t_{1/2} = \sim 0.87$  and  $\sim 0.12$  s. A follow-up study in 2007 found that the kinetics were at least four-state (U, I1, I2, F) and identified three folding phases: fast, medium, and slow (Watters *et al.* 2007). The relative amount of surface area buried during a folding phase can be estimated based on the dependence of the rate constant on the denaturant concentration. These measurements suggested that the most surface area was buried during the fast phase whereas the medium and slow phases had little to no surface area burial. CD measurements indicated that secondary structure also forms during the fast phase with 80% of the native signal forming in 2 ms. Therefore, hydrophobic collapse and secondary structure formation likely occur in the first fast phase, and the medium and slow phases are due to the rearrangement of the collapsed structure. The authors suspected the medium phase was due to the final transition into the native state due to the continuity of the rate constants of the medium phase with the single-phase rate constant at high guanidine concentrations. The slow phase is harder to pin down and could be due to the formation of a parallel- or on-pathway intermediate. The authors suspected the former case because a slow obligatory phase preceding the formation of the folded state in the medium phase would not be kinetically observable.

NMR experiments detected a conformation at guanidine concentrations between 4.0 and 6.5 M that had nonnative interactions along with some native interactions (Watters *et al.* 2007). CD and fluorescent signals became native-like at higher denaturant concentrations (they form before this intermediate on the U→F folding pathway), which suggests this intermediate is largely folded but stabilized by nonnative interactions. Point mutations and truncations gave additional insight to the complicated folding pathway of Top7. Comparing

the kinetics of various mutants with the wild type protein suggested that the middle region of Top7 was involved with the formation or stability of one or more of the intermediates. The C-terminal region seemed to be involved in the final stage of folding into the native state, and a fragment of just the final  $\alpha$ -helix and three  $\beta$ -sheets folded with single-phase kinetics. Such kinetics in the C-terminal fragment also suggest that the N-terminal portion of the protein is responsible for the nonnative interactions that cause formation of the intermediates. A different study using single-molecule force spectroscopy along with steered MD simulations found that the dominant step in unfolding was indeed separation of the C-terminal fragment from the N-terminal region (Sharma *et al.* 2007).

For its size, Top7 is extremely stable and has a complicated folding pathway. It's difficult to say whether and to what extent these attributes are due to the design process and/or its unique topology. As has been observed in other designed proteins discussed thus far, the basic design strategy (hydrophobic residues in the core, polar on the surface) is quite good at creating very stable proteins, often more stable than their naturally occurring counterparts. Watters *et al.* (2007) suggested that since Top7 is so stable and has seven secondary structural elements, it is likely that it can form substructures that are also relatively stable leading to populated folding intermediates. However, many of the mutations that were tested destabilized Top7 without simplifying the folding kinetics, so its stability alone does not cause its complicated folding pathway. The design process only took into account the final folded native structure without any negative design to prevent stabilization of nonnative interactions or intermediates. As was noted by Lei *et al.* (2006) in regards to FSD-1 and others, nonnative interactions likely slow folding and should be taken into account in the design stages. Yet, it is easier said than done, and the denatured state is still largely ignored when designing proteins.

The work of Dallüge *et al.* (2007) answers the question of whether the unique topology of Top7 dictates its complicated folding pathway by redesigning the sequence of Top7. The authors selected nearly 50% of the residues by hand, placing bulky hydrophobic amino acids in the core and favorable electrostatic interactions on the surface, and the remaining residues were placed computationally using tetrapeptide fragments. Two of the eight designed sequences, M5 and M7, were selected to express and characterize. Both

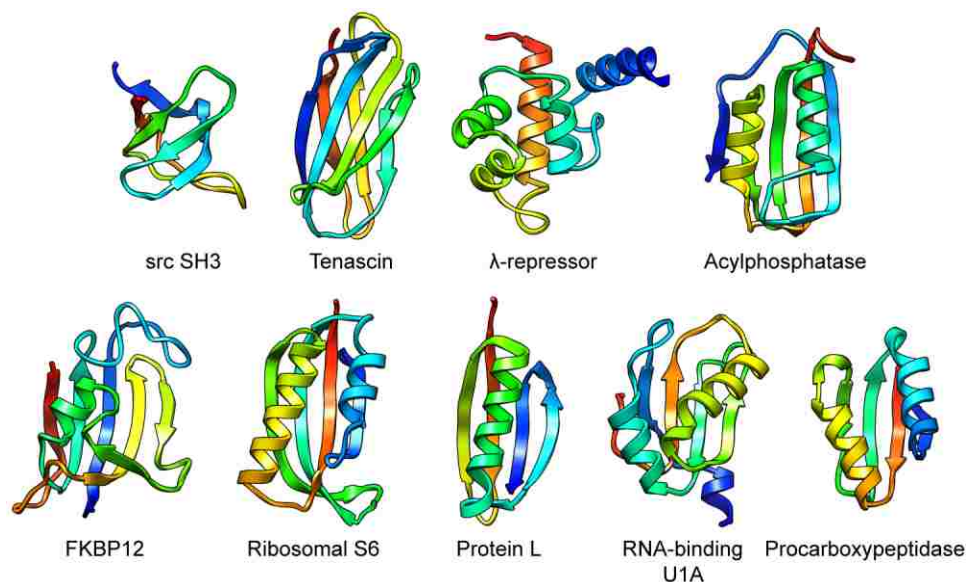
proteins were monomeric in solution, and CD and NMR spectra were indicative of the Top7 secondary structure. The two proteins were stable up to nearly 7 M guanidine by CD, and  $\Delta G_U$  was measured at 9.0 and 19.0 kcal/mol for M5 and M7, respectively ( $\Delta G_U = 13.2$  kcal/mol for Top7; Kuhlman *et al.* 2003). Similar values were extrapolated for thermal denaturation in guanidine, and folding fit a two-state model. Though a structure of either of these proteins was never determined, the CD and NMR data suggest that M5 and M7 fold to the same topology as Top7. However, both redesigned proteins were observed to follow a simple two-state folding pathway, suggesting that Top7's complicated folding pathway is not dictated by its unique topology but rather its sequence.

As Top7 was designed to have a fold topology not observed in nature, it is interesting to speculate as to the effects of natural selection on the folding pathway and structure of natural proteins relative to Top7. The most obvious difference between computational design and natural selection is that computational design, at least as it has been used in the studies discussed here, only designs toward stability (or "energy score") of the final folded state. Focusing on building a well-packed hydrophobic core creates fast-folding proteins in some cases (3-helix bundles), yet yields proteins with intermediates that slow folding in others (Top7 and FSD-1). Proteins in nature likely evolve to avoid kinetic traps that greatly slow folding since partially folded proteins are prone to aggregation and are targeted for degradation in cells. The denatured state is also under evolutionary pressure in nature, as favorable interactions in the denatured state result in a smaller free energy of folding, effectively destabilizing the native state. Unfortunately, intermediate and denatured states are not fully understood in general and are difficult to account for in design algorithms. Evolution has converged on many of the same protein folds multiple times, so it is tempting to think that Top7's unique fold has never been sampled because of its unfavorable folding intermediates. However, the mutational studies done by Dallüge *et al.* (2007) indicate the fold is capable of two-state folding with a different sequence and comparable stability. Top7 is notably more stable than other naturally occurring proteins of similar size, and many of its mutants were also quite stable. As nature typically creates only barely stable proteins, it is possible that such a stable protein as Top7 would be disfavored in nature. The design of Top7 was a remarkable feat in the field of protein engineering and serves to underscore the

challenges of computational design methodologies and the beautiful complexity of protein folding in nature.

### 8.2.5 Additional Rosetta Designs

The design studies discussed thus far have each redesigned a single protein using a different computational method. The biophysical characterizations of each resulting designed protein have provided new insights to the important forces driving protein folding and stability. It is also useful to consider many proteins all designed by the same method to determine how much of the variation in the designed proteins is due to the method itself or to the diversity of proteins in general. Dantas *et al.* (2003) performed such a study, redesigning nine small proteins using Rosetta (src SH3, tenascin,  $\lambda$ -repressor, acylphosphatase, immunophilin FKBP12, ribosomal S6, protein L – two designs, RNA-binding U1A, and procarboxypeptidase; Figure 8.2).



**Figure 8.2: Proteins redesigned by Rosetta**

Nine proteins from the Dantas *et al.* (2003) and Scalley-Kim and Baker (2004) studies are shown here colored red→blue, N→C.

The resulting ten designs were expressed and then characterized using CD, size exclusion chromatography, chemical and thermal denaturation, and one-dimensional NMR. One of the designed proteins was random coil in solution (SH3), and one aggregated (tenascin). The designed ribosomal S6 would not denature and was likely multimeric in solution, as was FKBP12. The remaining six designed proteins appeared to be folded by the

various methods, their CD spectra matched the respective wild type proteins, and chemical denaturation indicated their folding was two-state. Two proteins, designs of  $\lambda$ -repressor and acylphosphatase, were well folded, but thermal melts and NMR spectra indicated they were less rigid than their natural counterparts. Both protein L designs were well folded and stable, though one was less stable than the wild type. Designs of U1A and procarboxypeptidase were both well folded and more stable than the naturally occurring proteins by  $\sim 2$  and  $\sim 7$  kcal/mol, respectively. Five of the proteins (acylphosphatase, both protein L designs, U1A, and procarboxypeptidase), were thermostabilized relative to the respective wild type proteins.

In a later study, the kinetics of several of these proteins were measured and compared to the wild type (Scalley-Kim and Baker 2004). First, a successful SH3 design was presented that was less stable than the wild type, but it folded  $\sim 10$ -fold slower. Designed procarboxypeptidase was found to fold as fast as the wild type, and designed protein L and acylphosphatase folded 10-fold and an amazing  $10^5$ -fold faster, respectively.

Despite using identical protocols, the designed proteins ranged from more to less stable, unfolded to folded to aggregated, and slower to faster folding. What then was the reason for the variation? Three designs were significantly more stable than the wild type protein (ribosomal S6, U1A, and procarboxypeptidase) and two were less stable (protein L and  $\lambda$ -repressor). The three proteins with increased stability had more hydrophobic sequences relative to wild type, and likewise, the two with decreased stability had less hydrophobic sequences. For the two proteins that were folded but multimeric (ribosomal S6 and FKBP12), both had an increased hydrophobic surface area relative to the wild type. And finally, the SH3 redesign was found to be unfolded due to a clash between two bulky hydrophobic residues in the core. This clash was allowed in the design protocol because the atomic radii are scaled down to account for the rigidity imposed by using discrete rotamers. Kinetic analysis suggested that the transition state ensemble was preferentially stabilized in the proteins that folded faster than their wild type counterparts. The authors suggested this stabilization could be accomplished by broadening the transition state ensemble; the increased hydrophobic content of the core increases the potential number of favorable partially folded conformations in the ensemble.



These observations give good suggestions for improvements to the design protocol. It is computationally simple to test for hydrophobic patches on the surface that promote aggregation, and clashes that cannot be eliminated by small deviations from the inserted rotamer could similarly be assessed during the design process. Simply increasing the hydrophobic content of the sequence to improve stability is a bit more complicated because clearly all residues cannot be hydrophobic. The energy contribution of additional hydrophobic interactions must be balanced with the increasingly dynamic and molten cores that result depending on the requirements of the engineered protein. Stabilizing the transition state ensemble in the way discussed above may likewise be accomplished by increasing the hydrophobic content of the core, but directly incorporating the structure of the transition state in the design process is more difficult.

### **8.3 Proteins Designed for Function**

Many “rules” for creating stable proteins were defined from the simplistic proof-of-principle protein designs. First and foremost, hydrophobic residues should be in the core, not on the protein surface where they promote self-aggregation. These residues largely contribute to the stability of the protein overall, but must be balanced against the fact that they lead to more molten structures. Polar and charged residues that make more specific, geometrically restrictive interactions create less molten structures. Penalizing interactions that stabilize the denatured state and folding intermediates with the scoring function would be ideal, but it is difficult to account for them in practice. Designing a broad transition state ensemble might speed folding, but it is again difficult to model.

Proteins are the workhorses of cells and organisms acting as structural elements, catalysts in reactions, signaling molecules, and more. It is only natural that the next step is to design proteins that can carry out these and other novel functions. To do this, protein engineers must not only create stable proteins that follow the rules of protein folding, but also consider an additional set of rules that comes along with binding molecules and carrying out chemical reactions.

### 8.3.1 Ligands

Creating a protein to bind a ligand is far from a straightforward task. The binding site must be compatible with the ligand, both geometrically and chemically. The decrease in enthalpy due to favorable binding interactions must be balanced with the loss of degrees of freedom upon binding. A binding sight may need to be general enough to recognize a family of ligands or so specific that it can discern between enantiomers. Additionally, the ligand needs to get to the binding site. If the ligand is bound in a cleft or pocket, the protein will likely need to move to accommodate the insertion of the molecule without distorting the binding site too much. This section discusses the dynamics of two 4-helix bundles designed to bind metals and another protein designed to bind a small peptide.

#### 8.3.1.1 Metal-Binding 4-Helix Bundles, the Effectiveness of Negative Design

Due Ferro 1 (DF1) was designed by Lombardi *et al.* (2000) to contain the 4-helix active site scaffold found in manganese catalase (Figure 8.1e). DF1 is made up of a homodimer of helix-loop-helix motifs. The active site contains one His and two Glu residues from a Glu-Xxx-Xxx-His motif and a single Glu in both monomers binding each of two Mn(II) ions. Polar amino acids were added to stabilize the six active site residues, and the rest of the core and interface positions were packed with hydrophobic and polar amino acids. More polar residues were added to create favorable interactions with the solvent.

The resulting dimer was expressed and characterized. CD and analytical ultracentrifugation indicated the dimer was formed and was helical. DF1 bound Zn(II), Co(II), and Fe(II), and spectral analysis with cobalt was indicative of the correct active site geometry as observed in naturally occurring proteins. The dimer in complex with Zn(II) was successfully crystallized, and the structure showed the expected geometry. The two monomers had a backbone RMSD of 0.6 Å from each other, and the dimer had an RMSD of 1.6 Å from the designed structure. The dimer folded and dimerized in a single transition with a free energy of dimerization of  $-12.8 \pm 0.6$  kcal/mol and  $K_d = 0.41$  nM.

Unfortunately, DF1 was found to be inactive, as it could not stabilize the oxidized form of the active site (Spiegel *et al.* 2006). The authors proposed that this was because the protein scaffold did not allow for the geometry of the oxidized state in the active site. In order to investigate this hypothesis, Spiegel *et al.* performed MD simulations of the natural

enzyme, Mn-catalase from *L. plantarum*, and DF1. The RMSF of the C $\alpha$  atoms of the active site was the same in DF1 as in the natural protein, but it was higher in the termini of the helices in DF1. The authors attributed this to the more restrictive environment of the 4-helix bundle within the whole Mn-catalase complex rather than in solution as in DF1. The per-atom RMSD of the C $\alpha$  atoms in DF1 was higher than in the 4-helix bundles of Mn-catalase (1.2 vs. 0.8 – 0.9 Å) due to a conformational change in DF1. This change involved a sliding motion of two of the helices away from each other along the central axis of the bundle leading to an increased distance between the two Mn(II) ions. The same motion was also observed in Mn-catalase, but not to the same extent. The lack of activity in DF1 may also be due to one of the active site Glu residues and a water molecule blocking the substrates (O<sub>2</sub> and H<sub>2</sub>O<sub>2</sub>) from entering the active site. The corresponding Glu residue in the natural protein lies above the active site in a hydrophobic patch and is thought to be involved in proton shuttling during the reaction. This inserted water molecule was also responsible for the helix sliding and increased interionic distance in DF1.

This design was improved upon by Summa *et al.* (2002), who used the DF1 dimer to create an A<sub>2</sub>B<sub>2</sub> heterotetramer, again to bind two divalent metals, employing elements of both positive and negative design. This work designed two helices (A and B) to associate with one another in the desired tetrameric conformation such that they would also bind a divalent metal, and the tetramer was called DFtet. The Glu-XXX-XXX-His motif in DF1 and naturally occurring diiron proteins was incorporated in the B helix (DFtet-B), and an additional Glu was in the A helix (DFtet-A). The binding site for the two metals involved coordination by all four Glu and both His residues and the remainder of the core was packed with hydrophobic amino acids. The rest of the residues were placed with the goal of stabilizing the desired anti-parallel heterotetrameric topology while destabilizing the other two possible anti-parallel A<sub>2</sub>B<sub>2</sub> configurations. Placing positively and negatively charged residues in the helix-helix interfaces using a scoring function allowed this discernment. 3:1 combinations, homotetrameric, and all parallel combinations were not specifically designed against, as they would not have the correct ligand-binding geometry. However, the resulting designed structure was not expected to form stable parallel homo or heterotetramers by visual inspection due to unfavorable electrostatic interactions.

The designed structure was helical by CD for a solution of DFtet-A and DFtet-B but random coil for the individual peptides (Summa *et al.* 2002). The stoichiometry was determined to be 1:1 based on the CD spectra of solutions with different molar ratios of the peptides. When Zn(II) was added, the CD signal of the tetramer was unchanged as was that of the DFtet-A peptide. However, DFtet-B did form secondary structure in the presence of the zinc ion, presumably forming homooligomers to bind the metal. Sedimentation equilibrium ultracentrifugation indicated a 1:1 tetrameric mixture of both peptides in DFtet. DFtet-B seemed to be aggregating as a dimer or trimer.

Thermal unfolding studies indicated the melting temperature was 75 °C without metal and 95 °C in the presence of Zn(II) (Summa *et al.* 2002). The stability of the complex was also lower at pH 6.0 relative to 7.4, presumably due to protonation of the active site His residues. Spectroscopic binding studies with Co(II) produced a spectrum similar to those from other proteins with active site geometries like DFtet, and the desired ratio of 2 Co(II) : 1 DFtet was observed. When Fe(II) was added, it was oxidized to Fe(III) more rapidly than the uncatalyzed reaction, but not as fast as in naturally occurring diiron proteins.

Summa *et al.* designed a pair of peptides that associated in the desired binding stoichiometry and configuration and coordinated divalent metals as intended. In particular, they were successful in achieving the correct orientations of the four independent peptides in the tetramer, in large part due to negative design against undesired arrangements. Interestingly, homotetramers were not explicitly designed against, and DFtet-B was found to homooligomerize in solution indicating homooligomers should have been included in negative design. Not only was the correct geometry of four peptides and two metal ions obtained, but DFtet was more stable than most naturally occurring proteins ( $T_m = 95$  °C with ions bound).

In summary, the DeGrado group was successful in transplanting a naturally occurring diiron binding geometry from Mn-catalase into DF1 and DFtet. These peptides were successfully designed to self-associate even in the absence of their coordinating metal ions and coordinate them with the correct geometry when bound. Both proteins were remarkably stable with  $\Delta G_{\text{dimer}} = -12.8$  kcal/mol and nanomolar binding affinity for the dimers of DF1 and  $T_m = 95$  °C for DFtet. DFtet was capable of single-turnover oxidation of Fe(II) to

Fe(III), though DF1 was unable to catalyze the conversion of oxygen and hydrogen peroxide to water. The major helical motions in DF1 were also present to a smaller extent in Mn-catalase, suggesting the rest of the scaffold in the natural protein contributes to maintaining the active site geometry. And finally, negative design was imperative for producing a tetramer with the correct geometry and stoichiometry in designing DFtet.

### 8.3.1.2 Peptide Binding

Morin *et al.* recently attempted the design of a protein to bind and sequester a small peptide (2011). The D-Ala-D-Ala dipeptide, a necessary precursor to the peptidoglycan cell wall in gram-positive bacteria and the target of the powerful antibiotic, Vancomycin, was the target ligand of their studies. The protein, endo-1,4- $\beta$ -xylanase, was chosen as the scaffold due to its thermostability, available high-resolution structure, and geometry of its enzymatic cleft. The dipeptide ligand contained 25 atoms, a considerable jump in complexity from the two metal ions bound by DF1 and DFtet. The ligand was placed manually into the enzymatic cleft of the protein, and the Rosetta program was used to optimize the ligand position and protein sequence at the same time. The best three of the resulting designs were selected to express and characterize, and 1m4w\_6 was selected for structural characterization. Unfortunately, none of the three proteins yielded specific, high-affinity binding. The crystal structure of 1m4w\_6 showed an expansion of the binding pocket characterized by a 1.3 Å outward motion of the protein “thumb” region and an increase in solvent accessible surface area by 2.5x relative to the wild type structure. Many of the interface residues had high crystallographic temperature factors (B-factors) suggesting heightened motion (Figure 8.1f), and the predicted binding contacts were disrupted geometrically. Mutating two of the residues thought to contribute to this “open” conformation back to the wild type aided in reverting the structure back to the desired “closed” conformation. However, the mutant did not display any appreciable binding affinity.

This study underscores the difficulties of designing a functional binding pocket and inserting it into a protein scaffold. These authors may have designed a pocket that could have successfully bound the peptide ligand, but the dynamics of the protein scaffold in the region they inserted the pocket distorted the geometry of the binding residues and rendered the designed protein ineffective. The authors selected the endo-1,4- $\beta$ -xylanase scaffold

because they reasoned its thermostability would help it accommodate the series of potentially destabilizing mutations needed to create a binding pocket. However, in the designed proteins discussed here thus far, many of the thermostabilized designs had increased dynamics relative to their natural counterparts. The thumb region in the wild type protein was the most flexible region of the protein based on B-factors to begin with, and indeed, in the designed proteins the B-factors in this area further increased indicative of distortion of the designed binding site. Even if the binding site geometry allowed for ligand binding, the protein would pay a huge entropic cost for restricting motion in the thumb region relative to the unbound form, which would destabilize the bound conformation. Clearly these entropic costs must be taken into account when selecting a protein scaffold and in the design procedure.

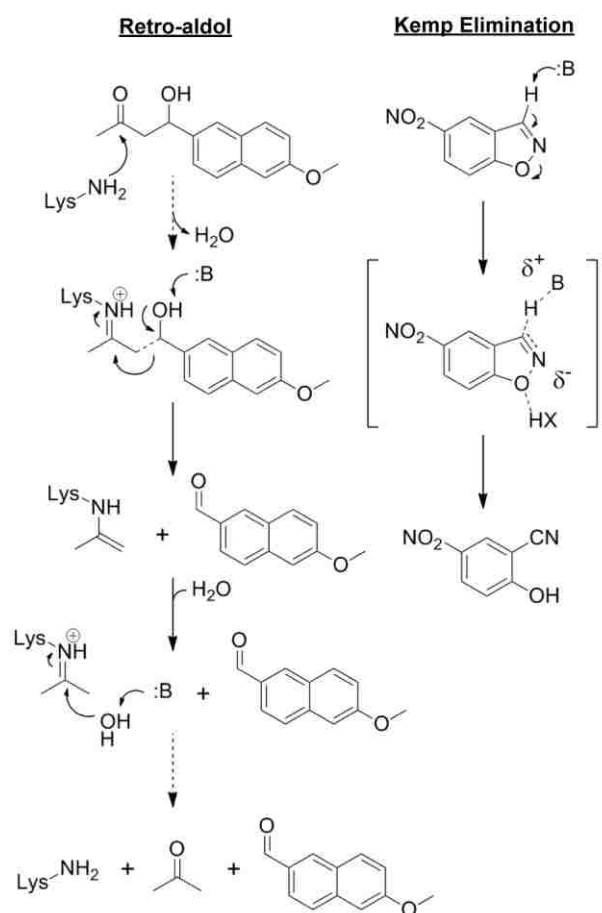
### 8.3.2 Enzymes

Binding ligands is a difficult task of balancing enthalpic and entropic contributions of the ligand and protein binding site. Requiring the ligand to undergo a chemical reaction while it is bound adds an additional element of complexity. As enzymes work by stabilizing the transition state in the reaction they are catalyzing, the “ligand” in designing an enzyme is the transition state structure. In addition to the transition state structure, quantum mechanics calculations are employed to define the precise geometry of the catalytic residues in relation to the transition state. This “theozyme” is then placed in a protein scaffold and other residues are selected to stabilize the conformation (Jiang *et al.* 2008). This section discusses biophysical studies on the dynamics and kinetics of two designed enzymes, a retro-aldolase and Kemp eliminase.

#### 8.3.2.1 Retro-Aldol Enzyme, Accommodating a Two-Step Reaction

The retro-aldol (RA) reaction involves breaking a carbon-carbon bond using a Lys residue as a Schiff base (Figure 8.3). The first step is binding of the substrate aldol molecule to the Lys residue with release of a water molecule. Next, the carbon-carbon bond is broken expelling the first product, an aldehyde. Finally, the second product, a ketone, is released from the bound Lys with addition of water. Four potential theozymes were created and insertion was attempted in 71 protein scaffolds using Rosetta (Jiang *et al.* 2008). A total of 72 designs was tested experimentally, 70 of which were successfully expressed. 32 designs were active with modest rate enhancements of  $\sim 10^2$ - $10^4$  representing two of the four

designed active site motifs. These rates are comparable to reaction rates observed for catalytic antibodies, but not for naturally occurring enzymes. A range of kinetic behaviors were observed from a lag phase, to steady-state, to an initial burst step. The linear kinetics were generally observed in the jelly-roll scaffold, while TIM-barrels tended to have more complex kinetics, possibly due to their more buried binding pockets. Structures of two of the active designs (RA22 and RA61) were solved by X-ray crystallography, and the active sites were nearly identical to the designed models with heavy atom RMSDs of 1.1 and 0.8 Å.



**Figure 8.3: Reaction mechanisms for the retro-aldol and Kemp elimination enzymes**

Mechanisms that were targeted for the *de novo* designed enzymes are shown for the retro-aldol (left) and Kemp elimination (right) reactions.

In a later study, MD simulations of RA22 were performed in complex with the substrate, 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone (Ruscio *et al.* 2009). Both the bound and unbound forms of the substrate were simulated. RA22 was one of the TIM barrel designs observed to have burst-phase kinetics, which Jiang *et al.* attributed to product

inhibition. Ruscio *et al.* were unable to reproduce such a burst-phase and concluded that release of the product was not a limiting factor in catalysis. In the MD simulations, the authors observed two major orientations of the substrate relative to the active site, one of which (O2) was more amenable to the first nucleophilic attack step and another (O1) that suited the second proton abstraction step. In both the bound and unbound MD simulations, the O2 orientation was preferred. The geometry necessary to perform the first step was satisfied in 49% of the frames, and for the second step only 30%. In addition, the geometry of the active site was found to be distorted relative to the optimal designed structure in the majority of the simulations due to protein dynamics. Thus, the authors predict that the limiting step in the reaction is the proton abstraction leading to the carbon-carbon bond cleavage and suggest reoptimizing the enzyme to favor the O2 substrate orientation as well.

Despite successfully designing an ideal active site as recapitulated in the crystal structure of the enzyme, the dynamics of the protein distorted the active site and favored an orientation of the substrate that was only compatible with one of the two major reaction steps. RA22 has a  $1.2 \times 10^3$  rate enhancement over the uncatalyzed reaction, but it is still orders of magnitude away from the rates of naturally occurring enzymes. The MD simulations suggested that the reaction is stalling at the second step and point out areas for improvement in subsequent rounds of design that could further increase the catalytic rate.

### **8.3.2.2 Kemp Elimination Enzyme, Rigid Active Site Geometry Promotes Catalysis**

Röthlisberger *et al.* (2008) designed an enzyme to carry out a reaction for which there is no naturally occurring enzyme, Kemp elimination (KE). In this reaction, a proton is extracted from a carbon by a catalytic base creating a nitrile group (Figure 8.3). Of 59 designs in 17 different protein scaffolds using two catalytic motifs, eight had measurable activity when expressed. Rate accelerations of  $\sim 10^3$ - $10^5$  over the uncatalyzed reaction were observed, slightly better than for the designed retro-aldolases, with multiple turnovers. One of the active designs, KE07, was crystallized and had an active site RMSD of 1.0 Å from the designed model. This design was further modified using *in vitro* directed evolution techniques. Nine evolved variants of KE07 were presented from ten rounds of directed



evolution, the best of which had a rate increase of  $1.2 \times 10^6$  over the uncatalyzed reaction and about ~100-fold over KE07.

Several of the evolved mutants of KE07 were further characterized by Khersonsky *et al.* (2010). The two major types of mutations that were incorporated involved replacing a Lys that quenched the catalytic Gly and tuning the environment of the catalytic base via a network of hydrogen bonds in the active site. The melting temperatures of the various mutants from the ten rounds of evolution were measured. Notably, the original protein scaffold used to create KE07 was an enzyme from *T. maritima*, a thermophilic bacterium. For the original design,  $T_m = 95$  °C, whereas it was down to 72 °C by the seventh round of design. In addition, KE07 had a temperature dependence similar to that of the uncatalyzed reaction. As the protein went through more rounds of evolution, the temperature dependence was nearly lost. So as enzymatic activity increased, stability decreased in this case, though such an inverse dependence need not occur. The authors noted that the majority of mutations made to proteins, particularly point mutations, are destabilizing, and it seems that mutations to create a catalytic site are no different. The active site may have an effectively lower  $T_m$  than the rest of the protein, meaning the ideal catalytic geometry seems to be lost before the overall structure of the protein scaffold. It is likely that as the active site becomes preorganized for binding and transition state stabilization, strain is introduced, decreasing the overall stability of the active site. Indeed, two of the mutations found during directed evolution had unfavorable backbone  $\phi/\psi$  angles in the design's crystal structure.

Quantum mechanics with molecular mechanics (QM/MM) and classical MD studies were done on a series of successful and inactive Kemp elimination designs to attempt to explain the differences in activity of the designs (Kiss *et al.* 2010). In QM/MM simulations, the active site was solvated with explicit water. The QM layer included the substrate and side chains of three catalytic residues involved in the theozyme, and the MM layer included a 10 Å sphere of water around the active site and the rest of the protein. Using this protocol, the activation barrier was estimated for six active and four inactive designs with an  $R = 0.76$  correlation and a slope of 1.5. In addition, the active site geometry was lost in two of the four inactive designs over the course of the simulation (KE66 and KE38). This correlation is

good, but the authors decided it was not worth the computational expense associated with the calculation.

Next, several systems were subjected to 20 ns of classical MD, including the 23 KE designs with their substrate, a catalytic antibody (34E4) for the KE reaction, and a naturally occurring enzyme for a different reaction (cathepsin K; there is no naturally occurring enzyme for the KE reaction; Kiss *et al.* 2010). In the case of KE07 and the KE catalytic antibody, both maintained their active site geometries throughout the simulations with active site side chain RMSDs of  $1.2 \pm 0.2$  and  $1.3 \pm 0.3$  Å, respectively. The strength of the hydrogen bonds within the substrate and the catalytic residues was a good predictor of the activity of the enzyme. While the designed structure had near-ideal hydrogen bonding angles and distances, this geometry was not maintained in the MD simulations of the inactive designs. Active designs had hydrogen bond distances below 3.2 Å and angles above 90°. When these criteria were used to classify the enzymes as active or inactive, there were only two false negatives (predicted inactive but had activity) and one false positive out of the 23 designs. Notably the two false negatives were among the three least active designs in the set. Neither the number of water molecules near the catalytic base oxygen(s) nor the RMSD of the active site backbone or side chains had any predictive value for the activity. While the dynamics of the entire protein were necessary to capture changes in the active site geometry, the activity could be predicted with considerable accuracy simply based on the hydrogen bonds between the substrate and catalytic residues. Cathepsin K, the naturally occurring enzyme, had even more ideal hydrogen bond geometries with distances of 2.1 and 1.8 Å and angles of 160 and 161°, suggesting that if better hydrogen bonding geometries could be realized in the KE designs, the activity could be further improved.

A second round of directed evolution along with MD studies was applied to a different design, KE70 (Khersonsky *et al.* 2011). Eleven new designs resulted from nine rounds of evolution. In the MD simulations, the RMSD of the active site backbone atoms decreased with increasing rounds of design, indicating the active site became more rigid as the directed evolution progressed. The RMSF, on the other hand, initially decreased but leveled off after two rounds of design. Both observations are indicative of increased preorganization of the active site in the more active designs. Once again, the deviation from

ideal hydrogen bond geometry was an excellent predictor of the reaction rate ( $k_{\text{cat}}$ ) and activity ( $k_{\text{cat}}/K_m$ ). In fact, the correlation between the square of the deviation from the ideal distance and  $-\ln(k_{\text{cat}})$  or  $-\ln(k_{\text{cat}}/K_m)$  was  $R = 0.83$  and  $0.88$ , respectively.

While engineered thermostabilized proteins benefited from increased dynamics, it seems that enzymes do not. The enzymes studied here benefit from rigid active sites and idealized geometries of the catalytic residues. Notably, the thermostability of the KE enzymes decreased as the active sites rigidified and activity increased. However, the discussion of dynamics in the enzymes described here has been limited to the active site. These studies have not delved into the overall dynamics of the protein, which forms the scaffold for the active site. It would be interesting to see how the dynamics of the whole protein evolved with increased activity and rigidity of the active site. Do they become more rigid as well? Or do they become more mobile to compensate for the loss of entropy in the active site?

## 8.4 Conclusions and Outlook

*De novo* protein design has come a long way from repacking and thermostabilizing globular proteins, to creating new protein topologies, to engineering functional proteins. Multiple computational algorithms have been developed that can successfully pack amino acids on a given backbone to create a stable structure. Completely repacking the protein core with hydrophobic residues tends to result in a highly dynamic, sometimes molten structure that often has high thermostability and fast-folding kinetics. This outcome is likely due to the geometrically unrestrictive nature of hydrophobic packing interactions as well as the increased force of the hydrophobic effect driving folding and stabilizing the native state. Creating a more rigid protein, on par with naturally occurring proteins, is more challenging and requires designing in polar interactions to restrict movement. Computational algorithms do not tend to place these sorts of interactions without deliberate instruction from the user.

Besides simply designing the final folded structure, the denatured state and entire folding pathway should ideally be taken into account. Favorable interactions in the denatured state can destabilize the native state, and the inadvertent creation of folding intermediates can lead to degradation or aggregation of the protein. It seems that simple two-

state folding pathways generally observed in nature for small globular proteins are not accidental and have been selected for over the course of evolution. Unfortunately, considering multiple states in design, especially since the denatured state and intermediates are difficult to pin down structurally, is both theoretically and computationally difficult. However, there are studies describing successful designs targeting transition and intermediate states based on MD-generated models from thermal unfolding simulations (White *et al.* 2005, Ladurner *et al.* 1998).

Despite plenty of room remaining for improvement remaining, functional proteins have been successfully designed. The 4-helix bundle proteins engineered by the DeGrado group recapitulated the binding geometry of naturally occurring diiron proteins and bind divalent metals (Summa *et al.* 2002, Spiegel *et al.* 2006). The Baker group designed a retroaldolase (Jiang *et al.* 2008) and Kemp eliminase (Röthlisberger *et al.* 2008) with reaction rates comparable to existing catalytic antibodies but lower than natural enzymes. Creating proteins to bind more diverse ligands and designing protein-protein interactions present future challenges for protein engineers. Consideration of the dynamics of the binding site and full protein scaffold will be important in the design process. Lessons learned in designing thermostable proteins will need to be applied to these designed enzymes and binding proteins if they are to be used in industrial processes. And as protein engineers begin to create proteins that are as functional and efficient as those that evolved naturally, the next question becomes, can we do even better than nature?

## Bibliography

- Alexander P.A., He Y., Chen Y., Orban J., and Bryan P.N. (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci USA*. 104:11963-11968.
- Alexander P.A., He Y., Chen Y., Orban J., and Bryan P.N. (2009) A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci USA*. 106:21149-21154.
- Anfinsen C.B. (1973) Principles that Govern the Folding of Protein Chains. *Science*. 181:223-230.
- Bartlett A.I. and Radford S.E. (2009) An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nat Struct Mol Biol*. 16:582-588.
- Beck D.A.C., McCully M.E., Alonso D.O.V., and Daggett V. (2000-2012) *In lucem* molecular mechanics (*ilmm*). University of Washington, Seattle, USA.
- Beck D.A.C. and Daggett, V. (2004) Methods for molecular dynamics simulations of protein folding/unfolding in solution. *Methods*. 34:112-120.
- Beck D.A.C., Armen R.S., and Daggett V. (2005) Cutoff size need not strongly influence molecular dynamics results for solvated polypeptides. *Biochemistry*. 44:609-616.
- Beck D.A.C. and Daggett V. (2007) A One-Dimensional Reaction Coordinate for Identification of Transition States from Explicit Solvent  $P_{\text{fold}}$ -Like Calculations. *Biophys J*. 93:3382-3391.
- Beck D.A.C., White G.W.N., and Daggett V. (2007) Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations. *J Struct Biol*. 157:514-523.
- Beck D.A.C., Alonso D.O.V., Inoyama D., and Daggett V. (2008a) The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. *Proc Natl Acad Sci USA*. 105:12259-12264.
- Beck D.A.C., Jonsson A.L., Schaeffer R.D., Scott K.A., Day R., Toofanny R.D., Alonso D.O.V., and Daggett V. (2008b) Dynameomics: mass annotation of protein dynamics and unfolding in water by high-throughput atomistic molecular dynamics simulations. *Protein Eng Des Sel*. 21:353-368.
- Berendsen H.J.C. (1998) Protein Folding: A Glimpse of the Holy Grail? *Science*. 282:642-643.

- Boeckler F.M., Joerger A.C., Jaggi G., Rutherford T.J., Veprintsev D.B., and Fersht A.R. (2008) Targeted Rescue of a Destabilized Mutant of p53 by an *in silico* Screened Drug. *Proc Natl Acad Sci USA*. 105:10360-10365.
- Bolon D.N. and Mayo S.L. (2001) Polar Residues in the Protein Core of *Escherichia coli* Thioredoxin Are Important for Fold Specificity. *Biochemistry*. 40:10047-10053.
- Bryson J.W., Desjarlais J.R., Handel T.M., and DeGrado W.F. (1998) From coiled coils to small globular proteins: design of a native-like three-helix bundle. *Protein Sci*. 7:1404-1414.
- Calosci N., Chi C.N., Richter B., Camilloni C., Engström Å., Eklund L., Travaglini-Allocatelli C., Gianni S., Vendruscolo M., and Jemth P. (2008) Comparison of successive transition states for folding reveals alternative early folding pathways of two homologous proteins. *Proc Natl Acad Sci USA*. 105:19241-19246.
- Chan C.-H., Yu T.-H., and Wong, K.-B. (2011) Stabilizing Salt-Bridge Enhances Protein Thermostability by Reducing the Heat Capacity Change of Unfolding. *PLoS One*. 6: e21624.
- Chi C.N., Gianni S., Calosci N., Travaglini-Allocatelli C., Engström Å., and Jemth P. (2007) A conserved folding mechanism for PDZ domains. *FEBS Lett*. 581:1109-1113.
- Chiti F., Taddei N., White P.M., Bucciantini M., Magherini F., Stefani M., and Dobson C.M. (1999) Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat Struct Bio*. 6:1005-1009.
- Chiti F. and Dobson C.M. (2009) Amyloid formation by globular proteins under native conditions. *Nat Chem Biol*. 5:15-22.
- Clarke J., Cota E., Fowler S.B., and Hamill S.J. (1999) Folding studies of immunoglobulin-like  $\beta$ -sandwich proteins suggest that they share a common folding pathway. *Structure*. 7:1145-1153.
- Congreve M., Murray C.W., and Blundell T.L. (2005) Keynote review: Structural biology and drug discovery. *Drug Discov Today*. 10: 895-907.
- Daggett V., Li A., Itzhaki L.S., Otzen D.E., and Fersht A.R. (1996) Structure of the transition state for folding of a protein derived from experiment and simulation. *J Mol Biol*. 257:430-440.
- Daggett V., Li A., and Fersht A.R. (1998) Combined molecular dynamics and  $\Phi$ -value analysis of structure-reactivity relationships in the transition state and unfolding pathway of barnase: Structural basis of Hammond and anti-Hammond effects. *J Am Chem Soc*. 120:12740-12754.
- Dahiyat B.I. and Mayo S.L. (1997) *De Novo* Protein Design: Fully Automated Sequence Selection. *Science*. 278:82-87.

- Dalal S., Balasubramanian S., and Regan L. (1997) Protein alchemy: changing  $\beta$ -sheet into  $\alpha$ -helix. *Nat Struct Biol.* 4:548-552.
- Dallüge R., Oschmann J., Birkenmeier O., Lücke C., Lilie H., Rudolph R., and Lange C. (2007) A tetrapeptide fragment-based design method results in highly stable artificial proteins. *Proteins.* 68:839-849.
- Dantas G., Kuhlman B., Callender D., Wong M., and Baker D. (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J Mol Biol.* 332:449-460.
- Davidson A.R. (2008) A folding space odyssey. *Proc Natl Acad Sci USA.* 105:2759-2760.
- Day R., Bennion B.J., Ham S., and Daggett V. (2002) Increasing temperature accelerates protein unfolding without changing the pathway of unfolding. *J Mol Biol.* 322:189-203.
- Day R. and Daggett V. (2005) Sensitivity of the folding/unfolding transition state ensemble of chymotrypsin inhibitor 2 to changes in temperature and solvent. *Protein Sci.* 14:1242-1252.
- Day R. and Daggett V. (2007) Direct observation of microscopic reversibility in single-molecule protein folding. *J Mol Biol.* 366:677-686.
- DeMarco M.L., Alonso D.O.V., and Daggett V. (2004) Diffusing and colliding: the atomic level folding/unfolding pathway of a small helical protein. *J Mol Biol.* 341:1109-1124.
- Doreleijers J.F., Nederveen A.J., Vranken W., Lin J., Bonvin A.M.J.J., Kaptein R., Markley J.L., and Ulrich E.L. (2005) BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures. *J Biomol NMR.* 32:1-12.
- Dunbrack R.L. Jr. (2002) Rotamer libraries in the 21st century. *Curr Opin Struct Biol.* 12:431-440.
- Eisenberg D. and Jucker M. (2012) The Amyloid State of Proteins in Human Diseases. *Cell.* 148:1188-1203.
- Englander S.W., Mayne L., and Krishna M.M. (2007) Protein folding and misfolding: mechanism and principles. *Q Rev Biophys.* 40:287-326.
- Feng J.A., Kao J., and Marshall G.R. (2009) A Second Look at Mini-Protein Stability: Analysis of FSD-1 Using Circular Dichroism, Differential Scanning Calorimetry, and Simulations. *Biophys J.* 97:2803-2810.
- Fersht A.R., Matouschek A., and Serrano L. (1992) The folding of an enzyme: Theory of protein engineering analysis of stability and pathway of protein folding. *J Mol Biol.* 224:771-782.

- Fersht A.R. (2002) On the simulation of protein folding by short time scale molecular dynamics and distributed computing. *Proc Natl Acad Sci USA*. 99:14122-14125.
- Fersht A.R. (2008) From the first protein structures to our current knowledge of protein folding: delights and skepticisms. *Nat Rev Mol Cell Biol*. 9:650-654.
- Fraenkel E., Rould M.A., Chambers K.A., and Pabo C.O. (1998) Engrailed Homeodomain – DNA complex at 2.2 Å resolution: a detailed view of the interface and comparison with other engrailed structures. *J Mol Biol*. 284:351-361.
- Friel C.T., Capaldi A.P., and Radford S.E. (2003) Structural analysis of the rate-limiting transition states in the folding of Im7 and Im9: similarities and differences in the folding of homologous proteins. *J Mol Biol*. 326:293-305.
- Fulton K.F., Main E.R., Daggett V., and Jackson S.E. (1999) Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J Mol Biol*. 291:445-461.
- Gallagher T., Alexander P., Bryan P., and Gilliland G.L. (1994) Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry*. 33:4721-4729.
- Geierhaas C.D., Nickson A.A., Lindorff-Larsen K., Clarke J., and Vendruscolo M. (2007) BPPred: a Web-based computational tool for predicting biophysical parameters of proteins. *Protein Sci*. 16:125-134.
- Gianni S., Guydosh N.R., Khan F., Caldas T.D., Mayor U., White G.W., DeMarco M.L., Daggett V., and Fersht A.R. (2003) Unifying features in protein-folding mechanisms. *Proc Natl Acad Sci USA*. 100:13286-13291.
- Gillespie B., Vu D.M., Shah P.S., Marshall S.A., Dyer R.B., Mayo S.L., and Plaxco K.W. (2003) NMR and temperature-jump measurements of *de novo* designed proteins demonstrate rapid folding in the absence of explicit selection for kinetics. *J Mol Biol*. 330:813–819.
- Greenfield N.J. and Fasman G.D. (1969) Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry*. 8:4108-4116.
- Haar L., Gallagher J.S., and Kell G.S. (1984) *NBS/NRC Steam Tables: Thermodynamic and Transport Properties and Computer Programs for Vapor and Liquid States of Water in SI Units*. Hemisphere, Washington, USA.
- Haglund E., Lindberg M.O., and Oliveberg M. (2008) Changes of protein folding pathways by circular permutation: Overlapping nuclei promote global cooperativity. *J Biol Chem*. 283:27904-27915.
- He Y., Rozak D.A., Sari N., Chen Y., Bryan P., and Orban J. (2006) Structure, Dynamics, and Stability Variation in Bacterial Albumin Binding Modules: Implications for Species Specificity. *Biochemistry*. 45:10102-10109.



- He Y., Chen Y., Alexander P., Bryan P.N., and Orban J. (2008) NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc Natl Acad Sci USA*. 105:14412-14417.
- Huang F., Settanni G., and Fersht A.R. (2008) Fluorescence resonance energy transfer analysis of the folding pathway of Engrailed Homeodomain. *Protein Eng Des Sel*. 21:131-146.
- Hubner I.A., Deeds E.J., and Shakhnovich E.I. (2006a) Understanding ensemble protein folding at atomic detail. *Proc Natl Acad Sci USA*. 103:17747-17752.
- Hubner I.A., Lindberg M., Haglund E., Oliveberg M., and Shakhnovich E.I. (2006b) Common motifs and topological effects in the protein folding transition state. *J Mol Biol*. 359:1075-1085.
- Itzhaki L.S., Otzen D.E., and Fersht A.R. (1995) The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J Mol Biol*. 254:260-288.
- Ivarsson Y., Travaglini-Allocatelli C., Morea V., Brunori M., and Gianni S. (2008) The folding pathway of an engineered circularly permuted PDZ domain. *Protein Eng Des Sel*. 21:155-160.
- Ivarsson Y., Travaglini-Allocatelli C., Brunori M., and Gianni S. (2009) Engineered symmetric connectivity of secondary structure elements highlights malleability of protein folding pathways. *J Am Chem Soc*. 131:11727-11733.
- Jackson S.E. and Fersht A.R. (1991) Folding of chymotrypsin inhibitor 2: Evidence for a two-state transition. *Biochemistry*. 30:10428-10435.
- Jackson S.E., elMasry N., and Fersht A.R. (1993) Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: a critical test of the protein engineering method of analysis. *Biochemistry*. 32:11270-11278.
- Jiang L., Althoff E.A., Clemente F.R., Doyle L., Röthlisberger D., Zanghellini A., Gallaher J.L., Betker J.L., Tanaka F., Barbas C.F., Hilvert D., Houk K.N., Stoddard B.L., and Baker D. (2008) *De Novo* Computational Design of Retro-Aldol Enzymes. *Science*. 319:1387-1391.
- Kabsch W. and Sander C. (1983) Dictionary of Protein Secondary Structure – Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers*. 22:2577-2637.
- Karplus M. (1959) Contact Electron-Spin Coupling of Nuclear Magnetic Moments. *J Chem Phys*. 30:11-15.
- Kazmirski S.L., Li A., and Daggett V. (1999) Analysis methods for comparison of multiple molecular dynamics trajectories: Applications to protein unfolding pathways and denatured ensembles. *J Mol Biol*. 290:283-304.

- Kell G.S. (1967) Precise Representation of Volume Properties of Water at One Atmosphere. *J Chem Eng Data*. 12:66-69.
- Khersonsky O., Röthlisberger D., Dym O., Albeck S., Jackson C.J., Baker D., and Tawfik D.S. (2010) Evolutionary Optimization of Computationally Designed Enzymes: Kemp Eliminases of the KE07 Series. *J Mol Biol*. 396:1025-1042.
- Khersonsky O., Röthlisberger D., Wollacott A.M., Murphy P., Dym O., Albeck S., Kiss G., Houk K.N., Baker D., and Tawfik D.S. (2011) Optimization of the *In-Silico*-Designed Kemp Eliminase KE70 by Computational Design and Directed Evolution. *J Mol Biol*. 407:391-412.
- Kiss G., Röthlisberger D., Baker D., and Houk K.N. (2010) Evaluation and ranking of enzyme designs. *Protein Sci*. 19:1760-1773.
- Kissinger C.R., Liu B.S., Martin-Blanco E., Kornberg T.B., and Pabo C.O. (1990) Crystal structure of an Engrailed Homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. *Cell*. 63:579-590.
- Kornberg T. (1981). Engrailed: A Gene Controlling Compartment and Segment Formation in *Drosophila*. *Proc Natl Acad Sci USA*. 78:1095-1099.
- Kuhlman B., Dantas G., Ireton G.C., Varani G., Stoddard B.L., and Baker D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*. 302:1364-1368.
- Ladurner A.G., Itzhaki L.S., Daggett V., and Fersht A.R. (1998) Synergy between simulation and experiment in describing the energy landscape of protein folding. *Proc Natl Acad Sci USA*. 95:8473 -8478.
- Lee B. and Richards F.M. (1971) Interpretation of Protein Structures – Estimation of Static Accessibility. *J Mol Biol*. 55:379-380.
- Lei H. and Duan Y. (2004) The role of plastic  $\beta$ -hairpin and weak hydrophobic core in the stability and unfolding of a full sequence design protein. *J Chem Phys*. 121:12104-12111.
- Lei H., Dastidar S.G., and Duan Y. (2006) Folding Transition-State and Denatured-State Ensembles of FSD-1 from Folding and Unfolding Simulations. *J Phys Chem B*. 110:22001-22008.
- Levinthal C. (1968) Are there pathways for protein folding? *J Chim Phys*. 65:44-45.
- Levinthal C. (1969) How to Fold Graciously. In *Mossbauer Spectroscopy in Biological Systems*. Eds. Debrunner P., Tsibris J.C.M., and Munck E. University of Illinois Press, Urbana, USA. pp. 22-24.
- Levitt M. (1983) Molecular dynamics of native protein: Analysis and nature of motion. *J Mol Biol*. 168:621-657.

- Levitt M., Hirshberg M., Sharon R., and Daggett V. (1995) Potential Energy Function and Parameters for Simulations of the Molecular Dynamics of Proteins and Nucleic Acids in Solution. *Comp Phys Comm.* 91:215-231.
- Levitt M., Hirshberg M., Sharon R., Laidig K.E., and Daggett V. (1997) Calibration and testing of a water model for simulation of the molecular dynamics of proteins and nucleic acids in solution. *J Physl Chem B.* 101:5051-5061.
- Li A. and Daggett V. (1994) Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2. *Proc Natl Acad Sci USA.* 91:10430-10434.
- Li A. and Daggett V. (1996) Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. *J Mol Biol.* 257:412-429.
- Li A. and Daggett V. (1998) Molecular dynamics simulation of the unfolding of barnase: characterization of the major intermediate. *J Mol Biol.* 275:677-694.
- Li W., Zhang J., and Wang W. (2007) Understanding the folding and stability of a zinc finger-based full sequence design protein with replica exchange molecular dynamics simulations. *Proteins.* 67:338-349.
- Li D.W., Yang H., Han L., and Huo S. (2008) Predicting the folding pathway of Engrailed Homeodomain with a probabilistic roadmap enhanced reaction-path algorithm. *Biophys J.* 94:1622-1629.
- Lindberg M.O. and Oliveberg M. (2007) Malleability of protein folding pathways: a simple reason for complex behaviour. *Curr Opin Struct Biol.* 17:21-29.
- Lipari G. and Szabo A. (1982) Model-Free Approach to the Interpretation of Nuclear Magnetic-Resonance Relaxation in Macromolecules: Theory and Range of Validity. *J Am Chem Soc.* 104:4546-4559.
- Liu B., Kissinger C.R., and Pabo C.O. (1990) Crystallization and preliminary X-ray diffraction studies of the Engrailed Homeodomain and of an Engrailed Homeodomain/DNA complex. *Biochem Biophys Res Commun.* 171:257-259.
- Liu R., Baase W.A., and Matthews B.W. (2000) The introduction of strain and its effects on the structure and stability of T4 lysozyme. *J Mol Biol.* 295:127-145.
- Lombardi A., Summa C.M., Geremia S., Randaccio L., Pavone V., and DeGrado W.F. (2000) Retrostructural analysis of metalloproteins: Application to the design of a minimal model for diiron proteins. *Proc Natl Acad Sci USA.* 97:6298-6305.
- Marshall S.A. and Mayo S.L. (2001) Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol.* 305:619-631.
- Marshall S.A., Morgan C.S., and Mayo S.L. (2002) Electrostatics significantly affect the stability of designed homeodomain variants. *J Mol Biol.* 316:189-199.

- Martin A.C.R. (1992-2001) ProFit: Protein Least Squares Fitting. University College London, London, England.
- Martínez J.C. and Serrano L. (1999) The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat Struct Biol.* 6:1010-1016.
- Matthews J.M. and Fersht A.R. (1995) Exploring the energy surface of protein folding by structure-reactivity relationships and engineered proteins: observation of Hammond behavior for the gross structure of the transition state and anti-Hammond behavior for structural elements for unfolding/folding of barnase. *Biochemistry.* 34:6805-6814.
- Mayor U., Johnson C.M., Daggett V., and Fersht A.R. (2000) Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc Natl Acad Sci USA.* 97:13518-13522.
- Mayor U., Grossmann J.G., Foster N.W., Freund S.M., and Fersht A.R. (2003a) The denatured state of Engrailed Homeodomain under denaturing and native conditions. *J Mol Biol.* 333:977-991.
- Mayor U., Guydosh N.R., Johnson C.M., Grossmann J.G., Sato S., Jas G.S., Freund S.M., Alonso D.O.V., Daggett V., and Fersht A.R. (2003b) The complete folding pathway of a protein from nanoseconds to microseconds. *Nature.* 421:863-867.
- McCallister E.L., Alm E., and Baker D. (2000) Critical role of  $\beta$ -hairpin formation in protein G folding. *Nat Struct Biol.* 7:669-673.
- McCully M.E., Beck D.A.C., and Daggett V. (2008) Microscopic reversibility of protein folding in molecular dynamics simulations of the Engrailed Homeodomain. *Biochemistry.* 47:7079-7089.
- McCully M.E., Beck D.A.C., Fersht A.R., and Daggett V. (2010) Refolding of the Engrailed Homeodomain: Structural basis for the accumulation of a folding intermediate. *Biophys J.* 99:1628-1636.
- Minkin V.I. (1999) Glossary of Terms Used in Theoretical Organic Chemistry. *Pure Appl Chem.* 71:1919-1981.
- Miura T., Satoh T., and Takeuchi H. (1998) Role of metal-ligand coordination in the folding pathway of zinc finger peptides. *Biochim Biophys Acta.* 1384:171-179.
- Morin A., Kaufmann K.W., Fortenberry C., Harp J.M., Mizoue L.S., and Meiler J. (2011) Computational design of an endo-1,4- $\beta$ -xylanase ligand binding site. *Protein Eng Des Sel.* 24:503-516.
- Morrone A., McCully M.E., Bryan P.N., Brunori M., Daggett V., Gianni S., Travaglini-Allocatelli C. (2011) The denatured state dictates the topology of two proteins with almost identical sequence but different native structure and function. *J Biol Chem.* 286:3863-3872.

- Muñoz V. and Serrano L. (1997) Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers*. 41:495-509.
- Myers J.K., Pace C.N., and Scholtz J.M. (1995) Denaturant  $m$  values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. *Protein Sci*. 4:2138-2148.
- Noyes M.B., Christensen R.G., Wakabayashi A., Stormo G.D., Brodsky M.H., and Wolfe S.A. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*. 133:1277-1289.
- Pardi A., Billeter M., and Wuthrich K. (1984) Calibration of the angular dependence of the amide proton-C alpha proton coupling constants,  $^3J_{\text{HNH}\alpha}$ , in a globular protein: Use of  $^3J_{\text{HNH}\alpha}$  for identification of helical secondary structure. *J Mol Biol*. 180:741-751.
- Pavletich N.P. and Pabo C.O. (1991) Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science*. 252:809-817.
- Religa T.L., Markson J.S., Mayor U., Freund S.M., and Fersht A.R. (2005) Solution structure of a protein denatured state and folding intermediate. *Nature*. 437:1053-1056.
- Religa T.L., Johnson C.M., Vu D.M., Brewer S.H., Dyer R.B., and Fersht A.R. (2007) The helix-turn-helix motif as an ultrafast independently folding domain: the pathway of folding of Engrailed Homeodomain. *Proc Natl Acad Sci USA*. 104:9272-9277.
- Religa T.L. (2008) Comparison of multiple crystal structures with NMR data for Engrailed Homeodomain. *J Biol NMR*. 40:189-202.
- Riddle D.S., Grantcharova V.P., Santiago J.V., Alm E., Ruczinski I., and Baker D. (1999) Experiment and theory highlight role of native state topology in SH3 folding. *Nat Struct Biol*. 6:1016-1024.
- Rose G.D. and Creamer T.P. (1994) Protein folding: predicting predicting. *Proteins*. 19:1-3.
- Rose G.D. (1997) Protein folding and the Paracelsus challenge. *Nat Struct Biol*. 4:512-514.
- Röthlisberger D., Khersonsky O., Wollacott A.M., Jiang L., DeChancie J., Betker J., Gallaher J.L., Althoff E.A., Zanghellini A., Dym O., Albeck S., Houk K.N., Tawfik D.S., and Baker D. (2008) Kemp elimination catalysts by computational enzyme design. *Nature*. 453:190-195.
- Ruscio J.Z., Kohn J.E., Ball K.A., and Head-Gordon T. (2009) The Influence of Protein Dynamics on the Success of Computational Enzyme Design. *J Am Chem Soc*. 131:14111-14115.
- Rutherford K. and Daggett V. (2009) A Hotspot of Inactivation: The A22S and V108M Polymorphisms Individually Destabilize the Active Site Structure of Catechol *O*-Methyltransferase. *Biochemistry*. 48:6450-6460.

- Sadqi M., de Alba E., Pérez-Jiménez R., Sanchez-Ruiz J.M., and Muñoz V. (2009) A designed protein as experimental model of primordial folding. *Proc Natl Acad Sci USA*. 106:4127-4132.
- Sánchez I.E. and Kiefhaber T. (2003) Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *J Mol Biol*. 325:367-376.
- Scalley-Kim M. and Baker D. (2004) Characterization of the Folding Energy Landscapes of Computer Generated Proteins Suggests High Folding Free Energy Barriers and Cooperativity may be Consequences of Natural Selection. *J Mol Biol*. 338:573-583.
- Scott K.A. and Daggett V. (2007) Folding mechanisms of proteins with high sequence identity but different folds. *Biochemistry*. 46:1545-1556.
- Serrano A.L., Waagele M.M., and Gai F. (2012) Spectroscopic studies of protein folding: Linear and nonlinear methods. *Protein Sci*. 21:157-170.
- Shah P.S., Hom G.K., Ross S.A., Lassila J.K., Crowhurst K.A., and Mayo S.L. (2007) Full-sequence computational design and solution structure of a thermostable protein variant. *J Mol Biol*. 372:1-6.
- Sharma D., Perisic O., Peng Q., Cao Y., Lam C., Lu H., and Li H. (2007) Single-molecule force spectroscopy reveals a mechanically stable protein fold and the rational tuning of its mechanical stability. *Proc Natl Acad Sci USA*. 104:9278-9283.
- Silverman L. and Glick D. (1969) Measurement of protein concentration by quantitative electron microscopy. *J Cell Biol*. 40:773-778.
- Spiegel K., DeGrado W.F., and Klein M.L. (2006) Structural and dynamical properties of manganese catalase and the synthetic protein DF1 and their implication for reactivity from classical molecular dynamics calculations. *Proteins*. 65:317-330.
- Srinivasan R. (1997) Ribosome v. 1.0. Johns Hopkins University, Baltimore, USA.
- Strickler S.S., Gribenko A.V., Gribenko A.V., Keiffer T.R., Tomlinson J., Reihle T., Loladze V.V., and Makhatadze G.I. (2006) Protein Stability and Surface Electrostatics: A Charged Relationship. *Biochemistry*. 45:2761-2766.
- Summa C.M., Rosenblatt M.M., Hong J.-K., Lear J.D., and DeGrado W.F. (2002) Computational *de novo* Design, and Characterization of an A<sub>2</sub>B<sub>2</sub> Diiron Protein. *J Mol Biol*. 321:923-938.
- Tanford C. (1970) Protein denaturation: Theoretical models for the mechanism of denaturation. *Adv Protein Chem*. 24:1-95.
- Team R.C. (2004) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

- Timasheff S.N. (1993) The control of protein stability and association by weak interactions with water: how do solvents affect these processes? *Annu Rev Biophys Biomol Struct.* 22:67-97.
- Tolman R.C. (1925) The Principle of Microscopic Reversibility. *Proc Natl Acad Sci USA.* 11:436-439.
- Toofanny R.D., Jonsson A.L., and Daggett V. (2010) A comprehensive multidimensional-embedded one-dimensional reaction coordinate for protein unfolding/folding. *Biophys J.* 98:2671-2681.
- Travaglini-Allocatelli C., Gianni S., Morea V., Tramontano A., Soulimane T., and Brunori M. (2003) Exploring the cytochrome *c* folding mechanism: cytochrome *c*<sub>552</sub> from *Thermus thermophilus* folds through an on-pathway intermediate. *J Biol Chem.* 278:41136-41140.
- Travaglini-Allocatelli C., Gianni S., Dubey V.K., Borgia A., Di Matteo A., Bonivento D., Cutruzzolà F., Bren K.L., and Brunori M. (2005) An obligatory intermediate in the folding pathway of cytochrome *c*<sub>552</sub> from *Hydrogenobacter thermophiles*. *J Biol Chem.* 280:25729-25734.
- Varghese J.N. (1999) Development of neuraminidase inhibitors as anti-influenza virus drugs. *Drug Develop Res.* 46:176-196.
- Wagner G., Hyberts S.G., and Havel T.F. (1992) NMR structure determination in solution: a critique and comparison with X-ray crystallography. *Annu Rev Bioph Biom.* 21:167-198.
- Walsh S.T.R., Cheng H., Bryson J.W., Roder H., and DeGrado W.F. (1999) Solution structure and dynamics of a *de novo* designed three-helix bundle protein. *Proc Natl Acad Sci USA.* 96:5486-5491.
- Walsh S.T.R., Lee A.L., DeGrado W.F., and Wand A.J. (2001a) Dynamics of a *De Novo* Designed Three-Helix Bundle Protein Studied by <sup>15</sup>N, <sup>13</sup>C, and <sup>2</sup>H NMR Relaxation Methods. *Biochemistry.* 40:9560-9569.
- Walsh S.T.R., Sukharev V.I., Betz S.F., Vekshin N.L., and DeGrado W.F. (2001b) Hydrophobic Core Malleability of a *de novo* Designed Three-helix Bundle Protein. *J Mol Biol.* 305:361-373.
- Watters A.L., Deka P., Corrent C., Callender D., Varani G., Sosnick T., and Baker D. (2007) The Highly Cooperative Folding of Small Naturally Occurring Proteins Is Likely the Result of Natural Selection. *Cell.* 128:613-624.
- Wensley B.G., Gärtne M., Choo W.X., Batey S., and Clarke J. (2009) Different members of a simple three-helix bundle protein family have very different folding rate constants and fold by different mechanisms. *J Mol Biol.* 390:1074-1085.

- Wensley B.G., Batey S., Bone F.A., Chan Z.M., Tumelty N.R., Steward A., Kwa L.G., Borgia A., and Clarke J. (2010) Experimental evidence for a frustrated energy landscape in a three-helix-bundle protein family. *Nature*. 463:685-688.
- Westheimer F.H. (1968) Pseudo-rotation in the hydrolysis of phosphate esters. *Accounts Chem Res*. 1:70-78.
- White G.W., Gianni S., Grossmann J.G., Jemth P., Fersht A.R., and Daggett V. (2005) Simulation and experiment conspire to reveal cryptic intermediates and a slide from the nucleation-condensation to framework mechanism of folding. *J Mol Biol*. 350:757-775.
- Wilson C.A., Kreychman J., and Gerstein M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*. 297:233-249.
- Wong K.-B. and Daggett V. (1998) Barstar has a highly dynamic hydrophobic core: evidence from molecular dynamics simulations and nuclear magnetic resonance relaxation data. *Biochemistry*. 37:11182-11192.
- Wu C. and Shea J.-E. (2010) On the origins of the weak folding cooperativity of a designed  $\beta\beta\alpha$  ultrafast protein FSD-1. *PLoS Comput Biol*. 6:e1000998.
- Zhang M., Chen C., He Y., and Xiao Y. (2005) Improvement on a simplified model for protein folding simulation. *Phys Rev E*. 72:051919.
- Zhu Y., Alonso D.O.V., Maki K., Huang C.Y., Lahr S.J., Daggett V., Roder H., DeGrado W.F., and Gai F. (2003) Ultrafast folding of  $\alpha_3D$ : a *de novo* designed three-helix bundle protein. *Proc Natl Acad Sci USA*. 100:15486-15491.



## Vita

Michelle McCully was born and grew up in Chicago, IL. She received her Bachelor of Science degree in 2006 from Washington University in St. Louis, where she majored in Biomedical Engineering and minored in Computer Science. She developed an interest in scientific research working in the lab of Dr. David Sept at Washington University, where she investigated the binding of signaling molecules to cytoskeletal proteins. In 2012 she earned her Doctor of Philosophy degree from the University of Washington as part of the Interdisciplinary Program in Biomolecular Structure and Design and the Department of Bioengineering. Her publications include:

Beck D.A.C., **McCully M.E.**, Alonso D.O.V., and Daggett V. (2000-2012) *In lucem* molecular mechanics (*ilmm*). University of Washington, Seattle, USA.

**McCully M.E.**, Beck D.A.C., and Daggett V. (2008) Microscopic reversibility of protein folding in molecular dynamics simulations of the engrailed homeodomain. *Biochemistry*. 47:7079-89.

**McCully M.E.**, Beck D.A.C., Fersht A.R., and Daggett V. (2010) Refolding of the Engrailed Homeodomain: Structural basis for the accumulation of a folding intermediate. *Biophys J*. 99:1628-1636.

Morrone A., **McCully M.E.**, Bryan P.N., Brunori M., Daggett V., Gianni S., and Travaglini-Allocatelli C. (2011) The Denatured State Dictates the Topology of Two Proteins with Almost Identical Sequence but Different Native Structure and Function. *J Biol Chem*. 286:3863-3872.

Wang D., Robertson I.M., Li M.X., **McCully M.E.**, Crane M.L., Luo Z., Tu A.Y., Daggett V., Sykes B.D., and Regnier M. (2012) Structural and functional consequences of the cardiac troponin C L48Q Ca<sup>2+</sup>-sensitizing mutation. *Biochemistry*. 51:4473-4487.

**McCully M.E.** and Daggett V. (2012) Folding and Dynamics of Engineered Proteins. In *Protein Engineering Handbook, vol. III*. Eds. Lutz S. and Bornscheuer U.T. Wiley-VCH, Weinheim. pp. 89-114.

**McCully M.E.**, Beck D.A.C., and Daggett V. Promiscuous contacts and heightened dynamics increase thermostability in an engineered variant of the Engrailed Homeodomain. *Protein Eng Des Sel*. *Accepted*.

**McCully M.E.**, Beck D.A.C., and Daggett V. Unfolding in a Test Tube: Multimolecule Atomistic Molecular Dynamics Simulations of the Engrailed Homeodomain. *Submitted.*