2017-11-14

Exploring the Role of Large Clusters of Branched Aliphatic Residues on the Folding Free Energy Landscape of (βα)8 TIM Barrel Proteins

Kevin T. Halloran
University of Massachusetts Medical School

EXPLORING THE ROLE OF LARGE CLUSTERS OF BRANCHED ALIPHATIC RESIDUES ON THE FOLDING FREE ENERGY LANDSCAPE OF $(\beta\alpha)_8$ TIM BARREL PROTEINS

A Dissertation Presented

By

Kevin Terence Halloran

Submitted to the Faculty of the
University of Massachusetts Graduate School of Biomedical Sciences, Worcester
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

November 14, 2017

EXPLORING THE ROLE OF LARGE CLUSTERS OF BRANCHED ALIPHATIC RESIDUES ON THE FOLDING FREE ENERGY LANDSCAPE OF $(\beta\alpha)_8$ TIM BARREL PROTEINS

A Dissertation Presented

By

Kevin Terence Halloran

This work was undertaken in the Graduate School of Biomedical Science
Program in Biochemistry and Molecular Pharmacology

Under the mentorship of

C. Robert Matthews, Ph.D., Thesis Advisor

Brian Kelch, Ph.D., Member of Committee

Francesca Massi, Ph.D., Member of Committee

Lawrence Stern, Ph.D., Member of Committee

Doug Barrick, Ph.D., External Member of Committee

Anthony Carruthers, Ph.D., Chair of Committee, Dean of the Graduate School of Biomedical Sciences

November 14, 2017

Dedication

To my family, especially my wife Laura, for supporting me throughout my graduate studies

Acknowledgments

I would like to first thank Dr. Matthews for providing me the opportunity to join his lab. His mentorship over the past 7 years have been invaluable to my development as a scientist. Thanks also goes to my thesis advisory committee, Dr. Brian Kelch, Dr. Francesca Massi, Dr. Lawrence Stern, and Dr. Anthony Carruthers for keeping me focused on the goals of my research.

I would also like to thank Dr. Osman Bilsel and Dr. Jill Zitzewitz for their guidance and training in the lab. They have been crucial in the development of my training as an independent scientist. To the past and present members of the Matthews group, Can, Sagar, Divya, Ganga, Vijay, Paul, Brian, Noah, Yvonne, Ere, Rohit, Sujit, Nidhi, and Kevin, it has been a truly enjoyable experience in lab with everyone.

A special thanks to Dr. Srinivas Chakravarthy from Bio-CAT, who, along with Osman and Sagar, spent many a night awake at the beamline working to help develop the continuous flow SAXS. To Dr. Payal Das, Dr. Ruhong Zhou, Yanming Wang, Dr. Karunesh Arora, and Dr. Charles Brooks III. Their expertise in computational biology have brought many new insights to the experimental projects I have worked on in the lab.

Finally, a special thank you to the people of the Biochemistry and Molecular Pharmacology Department, especially the Kelch, Schiffer, and Massi Labs. And to the support staff, Karen, Luca, Maria, and Josh, for making life in the lab run smoothly.

Abstract

$(\beta\alpha)_8$ TIM barrel proteins are one of the most common structural motifs found in biology. They have a complex folding free energy landscape that includes an initial off-pathway intermediate as well as two on-pathway intermediates. The formation of these intermediates is hypothesized to be driven by large clusters of the branched chain amino acids, isoleucine, leucine, and valine (ILV).

All-atom MD simulations and circular dichroism experiments on polar mutants of the hydrophobic clusters of α-Trp synthase, a TIM barrel protein, revealed the importance of dehydrating the clusters on intermediate states. Custom, single-piece microfluidic chips were interfaced with small angle x-ray scattering and time resolved FRET experiments to monitor the role of a large ILV cluster on the microsecond timescale in a second TIM barrel protein, sIGPS. Dimensional analysis of the initial misfolded intermediate showed an ILV cluster was responsible for the initiation of structure in the intermediate. Early structure formation in the ILV cluster was confirmed by coarse grained simulations. Native state hydrogen exchange experiments were used to probe the higher energy species that are in equilibrium with the native state. Results from the NMR experiment complement the kinetic studies as the core of stability found by NMR mapped back to the same region of the ILV cluster that was found to initiate folding.

When taken together, the results show the importance of hydrophobic clusters on the entire free energy surface of TIM barrel proteins.

Contents

List of Figures

List of Tables

List of Abbreviations

BASiC: Branched Aliphatic Side Chain
ILV: Isoleucine, Leucine, Valine
αTS: Alpha subunit of Tryptophan Synthase
sIGPS: Indole-3-glycerol phosphate synthase
BPHC: 2,3-dihydroxy-biphenyl dioxygenase
FRET: Förster resonance energy transfer
SAXS: Small angle x-ray scattering
CD: Circular dichroism
HX: Hydrogen exchange
MD: Molecular dynamics
CF: Continuous flow
MRE: Mean Residue Ellipticity
θ_N: Native MRE Signal
θ_I: Intermediate MRE signal
θ_U: Unfolded MRE signal
TSE: Transition State Ensemble
Rg: Radius of gyration
μs: Microsecond
ms: Millisecond
MEM: Maximum Entropy Modeling
TCSPC: Time correlated single photon counting
q: Scattering angle
Q: Fraction native contacts

Chapter I: Introduction

The proteomes of single cell and multicellular organisms are dynamic and complex. Proteins catalyze chemical reactions, provide structural support to the cell, and receive and transmit extracellular signals to ensure organism survival[1–3]. The majority of all proteins involved with such functions require a competently structured and folded form. This initial folding reaction occurs at the stage of translation where, as the genetic code is translated, the newly synthesized polypeptide chain folds spontaneously, or with the help of chaperones, to achieve its native functional state[4]. How a polypeptide chain spontaneously folds is not well understood and has been a major question in structural biology since the initial crystal structures of proteins were determined. It is important to understand how the polypeptide chain spontaneously folds as the unfolded state will be sampled with some frequency under a dynamic equilibrium. Populations of unfolded or intermediate states can lead to misfolding and aggregation, which have been previously implicated in a number of disease such as ALS, Alzheimer's, phenylketonuria, and cystic fibrosis[5–8]. As more protein-based therapeutics are developed, a better understanding of the relationship between sequence, stability, and function will allow for better engineering of these therapeutics.

The crystal structure of myoglobin was solved to 6 angstroms in 1958 by Kendrew and colleagues[9]. In reporting the findings, the authors point out that

one of the most surprising aspects of the protein's structure was the lack of symmetry. Due to the lack of symmetry, they could not postulate how the protein folded properly to achieve its native state solely based on the native structure.

Christian Anfinsen's work with Ribonuclease A meanwhile focused on the relationship between enzyme activity and tertiary structure[10,11]. Anfinsen and his colleagues showed that purified enzyme lost activity upon the addition of denaturant and reducing agents. Surprisingly, a large fraction of the activity was regained when the denaturant was diluted out and the disulfide bonds formed by oxidation. Because the random formation of all possible disulfide bonds would have resulted in ~1% recovery of activity, the sequence of the protein contains all the information required to direct the formation of the native state. This work formed the foundation for the thermodynamic hypothesis of protein folding - the three-dimensional shape of a protein was determined by the lowest energy state[12]. Cyrus Levinthal noted the process of going from the unfolded state to the native state could not be a random search as if the polypeptide chain searched through all possible confirmations, folding would not take place on a biologically relevant timescale. Therefore, Levinthal proposed that folding was biased through a preferred pathway[13].

Interactions that drive folding and models of protein folding

Even before the first crystal structure was solved, Linus Pauling theorized that hydrogen bonding of the peptide backbone must drive folding[14]. While some contemporary studies suggest that backbone hydrogen bonding plays an

important role in folding[15], it cannot be the sole determinant as the homogenous chemical diversity of backbone hydrogen bonding is, alone, insufficient to describe all three dimensional shapes proteins take within nature. Rather, it is the twenty different side chains of the primary sequence that encodes the secondary and tertiary structure[12,16]. The ability to change the primary sequence and therefore the ionic, hydrophobic, and van der Waal interactions of the side chains allows proteins to adopt many different folds required to complete the various biological activities. Due to the relatively low stability of the folded, native state when compared to the unfolded state, one must consider all these different interactions when studying how proteins fold as subtle changes to hydrogen bond networks and side chain packing can greatly affect the process[17].

Hydrophobic interactions have been known to be important to the folding of proteins for many years due to the energetic penalty of solvation in aqueous solution[18–20]. It was argued that the backbone hydrogen bonds could not drive folding as the it would be energetically neutral to break the intermolecular backbone-water hydrogen bonds to form intramolecular hydrogen bonds. However, because the process of burying hydrophobic side chains is non-specific, that force alone could not wholly account for structural specificity in folding[12]. Thus, several models of protein folding based on structure formation were developed in an attempt to describe the folding process.

There were three initial models to describe the folding process. In the nucleation-condensation model [21] and the framework model [22], secondary

structure within the polypeptide chain is assumed to form rapidly. The framework model then assumes that the tertiary structure of the protein is achieved through simple diffusion with the joining of multiple elements of secondary structure being rate limiting[23]. The nucleation model assumes that after the nucleation site forms, further secondary structure and tertiary structure forms in a hierarchical manner[21]. In the hydrophobic collapse model, entropic gains of quickly, but not necessarily specifically, burying hydrophobic side chains are thought to rapidly collapse the peptide chain to a relatively dense structure. This rapid collapse of the chain limits the conformational search with secondary structure and tertiary structure forming after the collapse[24].

As computational and theoretical studies of protein folding advanced, new three dimensional models, called free energy landscapes, based on entropy and enthalpy were developed[25,26]. Free energy landscapes propose that there are multiple pathways that the polypeptide chain can take to reach the native state and can be represented pictorially by a funnel. In the unfolded state, many states that are in rapid equilibrium with each other are present and as the chain begins to fold, the path of least resistance to the native state is taken. As the protein folds and chain entropy decreases, intermediates can be described as local minima in the funnel. While intermediates may slow the overall rate at which the protein folds, the formation of the intermediates limits the possible number of conformations the chain can take[27]. The existence of multiple pathways for folding creates an issue, however, as experimental folding kinetic data follows

simple exponentials. The issue of multiple pathways not being seen by experiments is thought to reflect the sensitivity of the experiments only to the slowest timescales of each reaction[28].

Sequence, Topology and the BASiC hypothesis

In nature, proteins with varying topologies are found to have vastly different folding rates. It is theorized that the differing rates are due to the complexity of the folded native state in different topologies[29]. Calculating the absolute contact order of the protein, a parameter that calculates the distance in chain length from atoms that are in contact with each other averaged over all contacts and normalized for the total chain length, can give a rough estimate of the expected folding rate[30]. However, even when looking within one protein topology family, different rates are found. These differences are thought to be due to the differences in sequences, as even single point mutations within a protein can cause drastic changes.

Despite the changes in the sequences within a topology family, proteins are often found to densely pack their cores with hydrophobic residues[19]. This provides multiple, energetically favorable interactions, the burial of the hydrophobes and formation of multiple van der Waal interactions. Based upon the side chain partitioning scales [31], the Branched Aliphatic Side Chain (BASiC) hypothesis was developed to describe the relationship between hydrophobic side chain burial and protein stability[32]. Through the burial of isoleucine, leucine, and

valine (ILV) residues, the protein is able to exclude water from its core and in turn lower the local dielectric constant [33]. The drop in dielectric constant strengthens the underlying hydrogen bond network and as a result decreases the volume and increases the packing density[34]. As a result, van der Waal interactions are increased creating link between secondary and tertiary structure formation. The exclusion of water from tightly packed ILV residues therefore enhances the cooperativity of the folding reaction, for both the formation of intermediates as well as the native state[32]. The clusters of ILV residues are able to serve as cores of stability for both the native and intermediate states[35–37].

TIM barrels as a model system

The TIM barrel $(\beta\alpha)_8$ family of proteins is one of the most common folds found in biology[1]. They are often involved in metabolic pathways and perform a variety of chemical reactions. The canonical TIM barrel fold is made up of eight central hydrophobic beta strands that form a barrel and is surrounded by eight amphipathic alpha helices. The beta strands and alpha helices are connected by long βα loops that contain the active site and short, tight αβ loops that are thought to provide stability to the protein[38]. The TIM barrel motif is a useful target for folding studies due the large number of sequences with the same structural motif. This allows for a test of the BASiC hypothesis as evolution as caused the sites of ILV clusters within the TIM barrel family to change in size and location.

The TIM barrel family has a highly conserved folding mechanism that contains an off-pathway intermediate that is followed by two on-pathway

intermediates. The folding model appears to be determined by the topology as proteins with vastly different protein sequences maintain the mechanism[36,37,39,40], including artificial TIM barrels that have been synthesized by other groups[41]. Of interest is the off-pathway intermediate that forms during the burst phase of stopped-flow kinetic experiments. The intermediate is considered off-pathway due to the unfolding-like kinetic phase that is found on the millisecond during protein refolding experiments[39]. The unfolding reaction under refolding conditions indicates that the intermediate is at least partially unfolding or structurally rearranging during the reaction due to improper contacts. While topology determines the overall folding mechanism, the protein sequence determines the location of structure in the intermediate states. In particular, the location of large clusters of branched aliphatic residues determines where the structure forms[36,42].

The α-subunit of tryptophan synthase from E. coli

The α subunit of tryptophan synthase (αTS) is responsible for the synthesis of indole from indole-3-glycerol phosphate[43]. Its folding mechanism has been well characterized using multiple spectroscopic techniques[39,44,45]. The protein has one off-pathway intermediate, I_{BP}, and two on-pathway intermediates, I_2 and I_1. The initial step in folding is the formation of the off-pathway intermediate which has been shown to form on the microsecond timescale[46]. The unfolding of this intermediate controls the formation of the I2 intermediate which then progresses through the I1 intermediate. Three prolines, P28, P217, and P261,

are known to isomerize and slow the interconversion of the intermediates and act as rate limiting step to reach the native state [47].

αTS has three ILV clusters, a large N-terminal cluster containing 31 residues, a cluster internal to the barrel containing 8 residues, and a C-terminal cluster containing 12 residues. Native state hydrogen exchange experiments have shown strong protection in the N-terminal half of the protein, corresponding to the large ILV cluster for the I_{BP} intermediate, suggesting that there is significant secondary structure in that region for the intermediate[45]. This was confirmed by monitoring the sub-millisecond folding reaction by time resolved FRET and SAXS which found near native like distances for the N-terminal region of the protein within 50 microseconds [46]. The rapid formation of N-terminal region has been attributed to the proper formation of the ILV cluster as simple alanine mutations to a subset of the 31 ILV residues in the N-terminal cluster have been shown to eliminate the I_{BP} intermediate[35].

(A) (B)

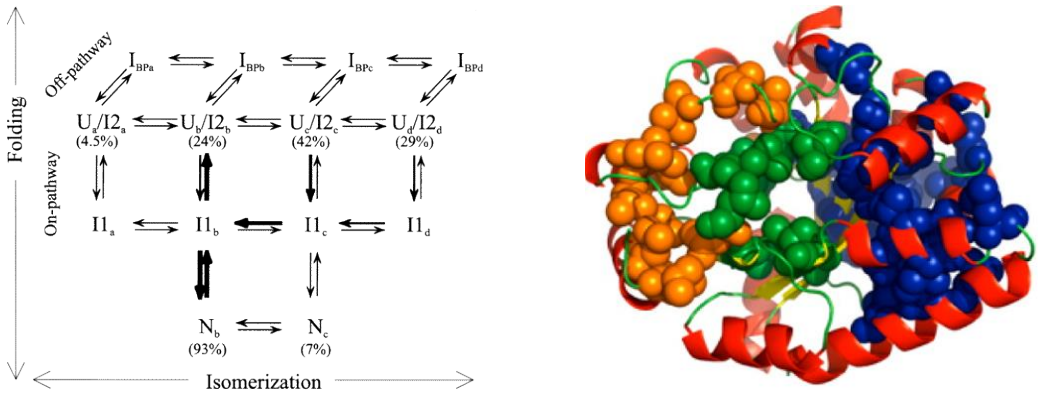


Figure 1. 1 (A) Folding mechanism of αTS including the proline isomerization reactions. The mechanism contains multiple parallel channels due to prolines contained within the sequence. (B) Ribbon diagram highlighting the 3 ILV clusters of αTS located at the N-terminal half (blue spheres), internal to the barrel (green spheres), and the C-terminal cluster (orange spheres).

Figure 1A is adapted from "Folding Mechanism of the α-Subunit of Tryptophan Synthase, an αβ Barrel Protein: Global Analysis Highlights the Interconversion of Multiple Native, Intermediate, and Unfolded Forms through Parallel Channels" copyright Biochemistry, 1999

Indole-3-glycerol phosphate synthase from S. solfataricus

The indole-3-glycerol phosphate synthase (sIGPS) is responsible for the conversion of 1-(2-carboxyphenylamino)-1-deoxyribulose 5 phosphate to 3-glycerol phosphate. For stability reasons, the first twenty-five residues have been cut off for the work in this dissertation[48]. Like αTS, the folding mechanism of sIGPS has previously been studied extensively. It has been confirmed to initially fold via an off-pathway intermediate that forms faster than 5 milliseconds. This intermediate at least partially unfolds before traversing through two on pathway intermediates[36,49,50].

Unlike αTS, sIGPS has a single large ILV cluster that spans roughly from β3 through β6. When monitored by pulse quench hydrogen exchange mass

spectrometry, it was this region that showed protection in 75 milliseconds with the strongest protection in the $(\beta\alpha)_4$ region. The strong protection in this βα pair is thought to be due to the high concentration of branched aliphatic residues, with 11 out of 22 being I, L, or V. The central region of the protein seems to act as a nucleation site for folding as the on-pathway intermediates showed protection patterns that expanded out to cover β_2 though β_7[36].

$$I_{BP} \rightleftharpoons U \rightleftharpoons I_A \rightleftharpoons I_B \rightleftharpoons N$$

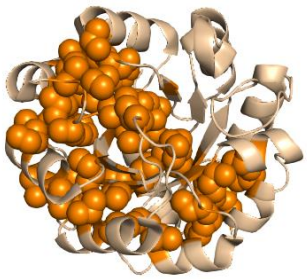


Figure 1. 2 Folding mechanism and ribbon diagram of sIGPS highlighting the large ILV cluster. The protein first folds from the unfolded state (U) to the off-pathway intermediate (I_{BP}). I_{BP} must then at least partially unfold before the on-pathway intermediates I_A and I_B form. The ILV cluster (orange spheres) spans from the β_3 through β_7 region.

Scope

The scope of the dissertation is to better understand the role of clusters of branched aliphatic residues in contributing stability to TIM barrel proteins throughout the entire free energy landscape. The hydrophobic packing of I, L, and V residues are thought not only are important to the native state but also the

formation of the intermediates on the folding pathway. The work in the dissertation focuses on two members of the TIM barrel family, αTS and sIGPS. As they are members of the TIM barrel family, both fold via one off-pathway intermediate and two on-pathway intermediates. Despite similar native topologies, their respective ILV clusters are quite different. The work will look at the contributions of the ILV clusters and how cluster size and location may be affecting the free energy landscape.

In Chapter II, the role of desolvating the hydrophobic clusters is investigated through the introduction of isosteric polar mutations to the ILV clusters of αTS. The equilibrium studies in Chapter II are complemented by simulation work performed by Ruhong Zhou's group at the IBM Watson Center. Both experiments and simulations show the importance of properly desolvating the large N-terminal ILV cluster of αTS for both the formation of the equilibrium intermediate and the native state.

Chapter III examines the earliest folding events of sIGPS and the role the large ILV cluster plays in initiating the folding process. Experimentally, sub-millisecond kinetics are studied using custom microfluidic chips interfaced with time resolved small angle x-ray scattering and time resolved Forester resonance energy transfer experiments. Additional structural insights on the folding intermediates are gained by Gō-simulations performed by Charles Brook's laboratory at the University of Michigan.

In Chapter IV, the role of the ILV clusters to sIGPS are probed by native state hydrogen exchange monitored by nuclear magnetic resonance (NMR). Due to the high kinetic stability of sIGPS, we are able to monitor the native basin of the free energy landscape and gain insights to the cores of stability and the role of ILV clusters.

The final chapter, Chapter V, discusses the role of ILV clusters and how they are important across the entire folding free energy landscape. Preliminary data on methionine based labeling is presented, along with a brief discussion of how the technique might be used in future experiments to address issues in early folding events.

Chapter II: The relationship between desolvating hydrophobic side chains and stability

This chapter has previously been published as:

Interplay between Drying and Stability of a TIM Barrel Protein: A Combined Simulation–Experimental Study
Payel Das, Divya Kapoor, **Kevin T. Halloran**, Ruhong Zhou, and C. Robert Matthews
Journal of the American Chemical Society 2013 135 (5), 1882-1890
DOI: 10.1021/ja310544t

This chapter was a collaborative effort with the lab of Dr. Ruhong Zhou. The simulations were performed by Dr. Payel Das and Dr. Zhou. The experiments were performed by myself and Dr. Divya Kapoor. All authors took part in the analysis of the data.

Introduction

The folding of proteins following synthesis on a ribosome or dilution from a chemically-denatured state involves the formation of numerous van der Waal's interactions, hydrogen bonds and electrostatic interactions that stabilize the compact native conformation. It is widely accepted that a necessary structural consequence of the protein folding reaction is the exclusion of water from the side chains and main chains that become buried in the native state. The thermodynamic consequence of the dehydration reaction reflects the substantial gain in entropy realized by freeing water during folding.

The role of water in protein folding reactions has been examined by both experimental and computational approaches. Mutational analyses, in which nonpolar side chains are replaced with isosteric polar side chain analogs, have shown that water is selectively shed prior to the appearance of the native state to enable the formation of critical cores of stability in early intermediates[51] or transition state ensembles[52,53]. By contrast, time-resolved infrared spectroscopy analysis revealed dehydration of the main-chain amides in the final step of folding from the alkaline-denatured states of both α-helical[54] and β-sheet proteins[55]. A third experimental approach towards examining the role of water in folding monitors the protection of main chain amide hydrogens against exchange in deuterated water in partially-folded states[56,57] and folding intermediates[36,58,59]. When hydrogen exchange (HX) techniques were applied to a pair of $(\beta\alpha)_8$ TIM barrel proteins[36,45], protection against exchange in folding intermediates was found to be selectively associated with clusters of branched aliphatic side chains,

isoleucine, leucine and valine (ILV). The molecular rationale for this behavior was ascribed to the preferential partitioning of side chain analogs of saturated hydrocarbon moieties into the vapor phase, relative to their aromatic, sulfur or polar-containing counterparts that spontaneously dissolve in water[31]. The Branched Aliphatic Side Chain (BASiC) Hypothesis was formulated on the basis of these differential solubilities and proposes that clusters of ILV side chains play crucial roles in stabilizing folding intermediates in TIM barrel proteins by selectively excluding water from their interiors[35,60].

From a computational perspective, nanoscale dewetting transitions[61–63] between hydrophobic surfaces have long been of interest for both physical[62,64–68] and biological systems[62,65,69–73]. Previous molecular dynamics (MD) simulation studies have identified several proteins or peptides in which a dewetting transition was observed prior to the docking of preformed elements of secondary structure. For example, a remarkable dewetting transition was observed within the nanoscale channel between the four melittin α-helices each of whose hydrophobic interface comprises 3 isoleucines, 4 leucines, 1 tryptophan and 2 valines[73]. A subsequent study on a variety of protein complexes (dimers, tetramers and two-domain proteins) found that dewetting required large complementary hydrophobic surfaces with significant contributions from isoleucines, leucines and valines[71]. In contrast, a marked decrease in water density was not detected at the domain interface in the two-domain 2,3-

dihydroxy-biphenyl dioxygenase (BPHC)[69]. The domain interface in BPHC is relatively heterogeneous in nonpolar side chains.

Building on the results of the previous experiments and MD simulations, we adopted a combined experimental and computational approach to test the conjecture that large ILV-rich clusters in TIM barrel proteins are prone to undergo dewetting from their interiors. As a target, we chose the alpha subunit of tryptophan synthase, αTS, a ~28 kDa TIM barrel $(\beta\alpha)_8$ protein that is a component of the $\alpha_2\beta_2$ tetrameric tryptophan synthase complex. Previously, the protein was observed to offer strong and selective protection against HX in an on-pathway intermediate associated with a large N-terminal ILV cluster[57]. A smaller C-terminal ILV cluster does not offer protection against HX and provides an internal control. As a surrogate for the polarity introduced by water, two buried leucines in the N-terminal cluster and a single leucine in the C-terminal cluster were individually replaced with the isosteric and polar asparagine. The effects of these mutations on the water density within the clusters were predicted by MD simulations of artificially-displaced versions of their preformed β-sheet and α-helical components. These predictions were then compared with the effects of the mutations on the experimentally-determined stabilities and structures of the native state and the folding intermediate. The possibility that wetting could also be enhanced by replacing a buried cysteine adjacent to the N-terminal ILV cluster with an asparagine was also studied. The combined results support the conclusion that the drying of the large N-terminal ILV cluster is crucial to the

stability and structure of the native state and of a productive folding intermediate in a TIM barrel protein.

Results

A ribbon diagram of αTS and the location of its three ILV clusters are shown in Figure 2.1a. Cluster 1, containing 31 ILVs, forms the interface between the exterior of the β-barrel and the interior of the α-helical shell in $(\beta\alpha)_{1\text{-}4}$. Cluster 2, containing 12 ILVs, is found at the interface between the β-barrel and the α-helical shell in $(\beta\alpha)_{5\text{-}6}$. Cluster 3 is located in the interior of the cylindrical barrel and is formed from 8 ILVs individually contributed by 7 of the 8 β-strands and α-helix 0 at the N-terminus.

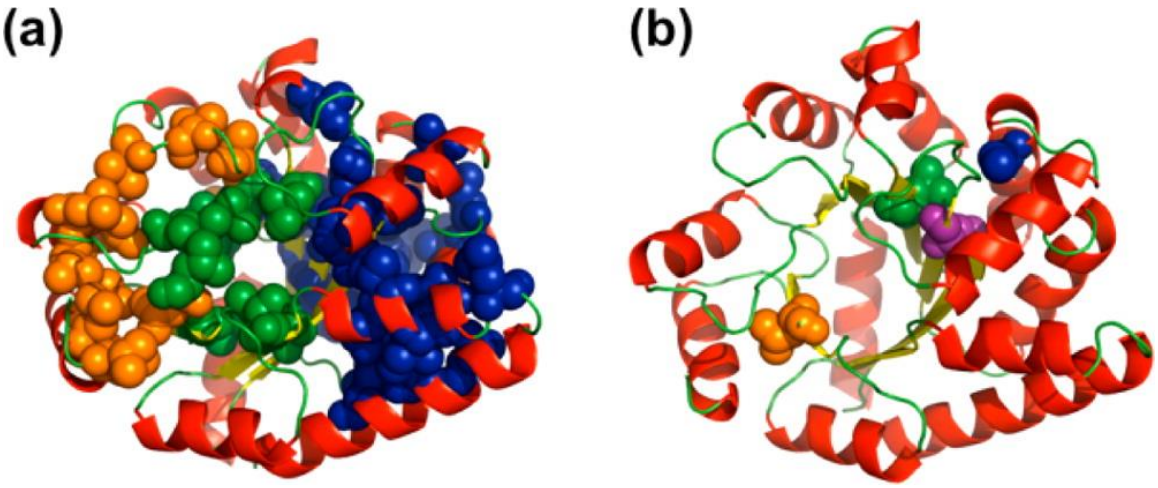


Figure 2.1 Ribbon diagrams of αTS (a) highlighting the three hydrophobic clusters formed by the ILV residues [cluster 1 (blue), cluster 2 (orange) and cluster 3 (green)] (b) showing the location of hydration mutations in the crystal structure [Leu50 in β2 (purple), Cys81 in α2 (blue), Leu99 in β3 (green), and Leu176 in β6 (orange)]. Coordinates of αTS from Salmonella typhimurium were used to generate the figure from a refined version of PDB entry 1BKS

MD Simulations of Hydration in ILV Clusters

Hydration in Cluster 1. A cavity inside Cluster 1, with an estimated volume of ~1300 $Å^3$, was created by pulling the α_1 and α_2 helices away from the β_1, β_2 and β_3 strands by a separation distance d varying from 4 to 6 Å (Figure 2.2a) and filled with water molecules. Previous work have reported nanoscopic dewetting transitions in proteins with cavity volumes of a similar order[70,74]. During the 16 ns simulation time, the water molecules were free to move, but the protein heavy atoms remained fixed. Figure 2.2b shows the water density plots as a function of simulation time for Cluster 1 at separation distances of 4 Å and 6 Å. The cavity undergoes intermittent transitions between wet and dry states at a separation distance of 4 Å, however, no drying transition was observed at d = 6 Å. These drying transitions typically occur in 200-300 ps. To check the convergence of our results and if the system reached equilibrium, we started the simulations from two different initial states with d = 4 Å: one from the ‘wet’ state and a second starting from a ‘dry’ state, in which all of the initial water molecules were removed manually to create a dry cavity (Figure 2.2c)

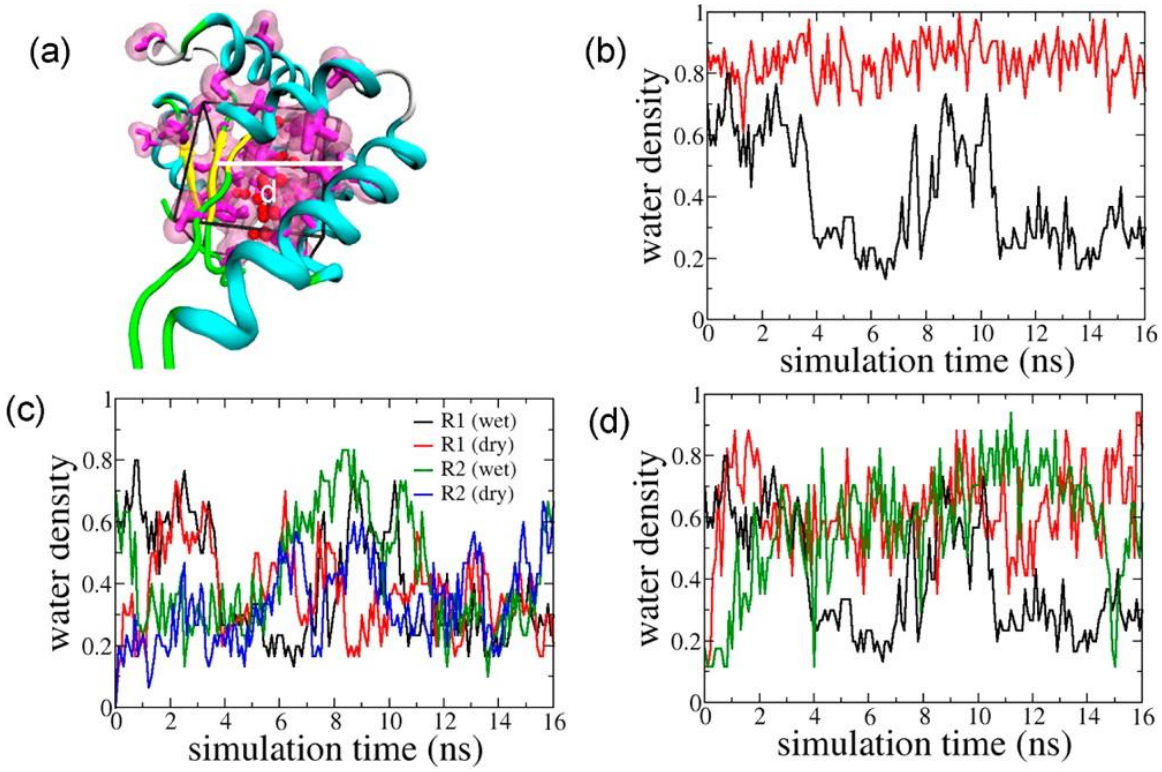


Figure 2.2(a) Ribbon representations of ILV cluster 1, in which the helices (shown in cyan) are manually separated from the β-strands (shown in yellow) by a separation distance d to create a hydrophobic cavity. The cavity was initially filled with water molecules (red spheres). The observation volume of ~1300 $Å^3$ is shown with black lines. Branched aliphatic side chains (heavy atoms only) are shown as sticks and as molecular surfaces. The cluster is oriented in such a way that the bottom of the figure is the N-terminus of the β-strands. (b) Plots of normalized water density as a function of simulation time of ILV cluster 1 for d = 4 Å (black) and 6 Å (red). The normalized water density was obtained by dividing the number of water molecules by the maximum number of water molecules inside the cavity (Nw,max = 30). (c) Plot of water density as a function of simulation time for four different trajectories, two starting from the "wet" initial state and two others starting from the "dry" initial state, with d = 4 Å. (d) Plots of normalized water density as a function of simulation time for the three ILV clusters of αTS (cluster 1 in black, cluster 2 in red, and cluster 3 in green).

Within the first 1-4 ns, the cavity underwent wetting/dewetting transitions in which the normalized water density inside the cavity switched between a maximum of 0.8 (wet) and a minimum of 0.2 (dry) from both initial states. The normalized water density is obtained by dividing number of water molecules with maximum number of water molecules inside the cavity. Snapshots of the cavity in the wet and dry states suggested that the water density was lowest near the center of the cavity, as a vapor bubble was frequently formed in this region and was stable for several nanoseconds. The two termini of the β-strand triplet remained relatively wet, the N-terminus being drier than the C-terminus. The latter results are consistent with the stronger protection against amide hydrogen exchange (HX) with solvent in this region observed in native-state HX experiments[45].

To provide insight into the role of individual ILV side chains to the dehydration observed in Cluster 1, 10 of its constituent members were individually substituted with alanines, and the simulations were performed on these alanine mutants. The residues selected (Figure 2.3a), V23, L25, I37, I41, L48, L50, L85, I95, I97 and L99, have previously been shown to eliminate an early kinetic trap in folding when replaced by alanine and all but I41A and L85A significantly destabilize the on-pathway equilibrium intermediate[35]. The minimal effects of the I41 and L85 variants are thought to reflect their location in helices α_1 and α_2, self-contained elements of secondary structure on the surface of the

protein that can more readily mitigate the effect of mutations on stability than their β-barrel counterparts.

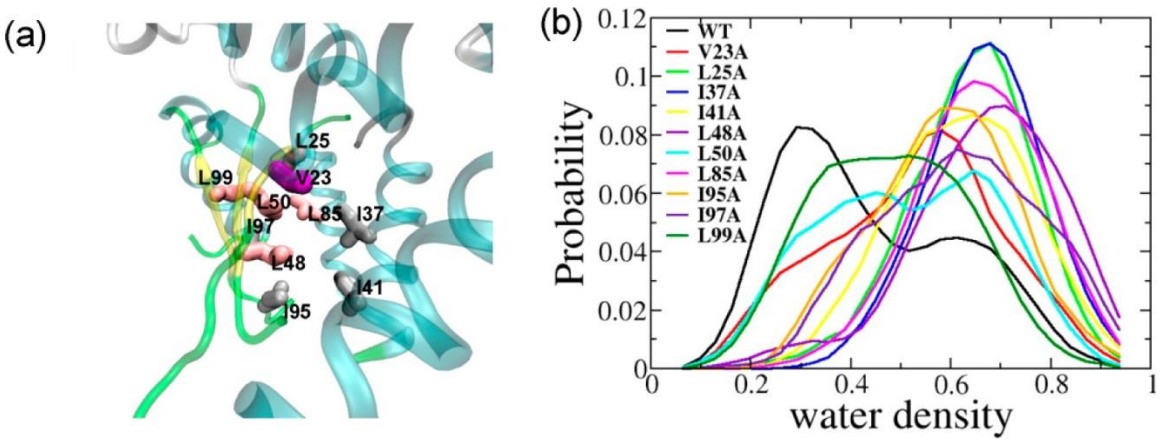


Figure 2. 3 (a) Ribbon diagram of the interior of cluster 1, with the ILV side chains selected for alanine-scanning mutagenesis portrayed in space-filling format. (b) Histograms of water density inside cluster 1 for the wild-type protein and the 10 alanine variants.

Figure 2.3b plots the histograms showing the probability of the water density within of the cavity for Cluster 1 of the wild-type protein and all ten alanine mutants. The histogram for the wild type protein shows a bimodal distribution, confirming that the cavity undergoes transitions between a dry (water density ~0.3) and a wet state (water density ~0.65), the dry state being more probable over the wet state. Apart from V23A, L50A and L99A, the cavity in the mutated proteins experienced complete or nearly complete loss of dewetting (Figure 2.3b). The histograms of water density for these mutants are gaussian in nature with peaks centered at a water density of 0.6-0.7, showing that the cavity

largely wets upon alanine substitution. These findings show that Cluster 1 in wild-type protein has evolved to favor a dry cavity to provide stability. Subtle changes in the surface topography and chemistry (e.g. single mutation I/L to A) can tip the balance of the cavity to a more wet state, potentially lowering the stability of the cluster and protein. In contrast, alanine replacement to residues V23, L50, and L99 resulted in partial loss of dewetting, with the L99A mutation being most resistant to wetting. The histograms of these three mutants show a considerable population of the low water density states (Figure 2.3b). These β-strand residues are centrally located in the cluster, facing the helical shell and are surrounded by neighboring ILV residues (Figure 3a). To further 'wet' the cavity inside Cluster 1, we performed a more radical perturbation on the hydrophobic surface by substituting L50 and L99 with their isosteric and polar counterpart, asparagine. The water density histograms of L50N and L99N mutant proteins illustrate that the introduction of a polar side chain at positions 50 and 99 results in a complete or significant loss of dewetting (Figure 2.4b-c).

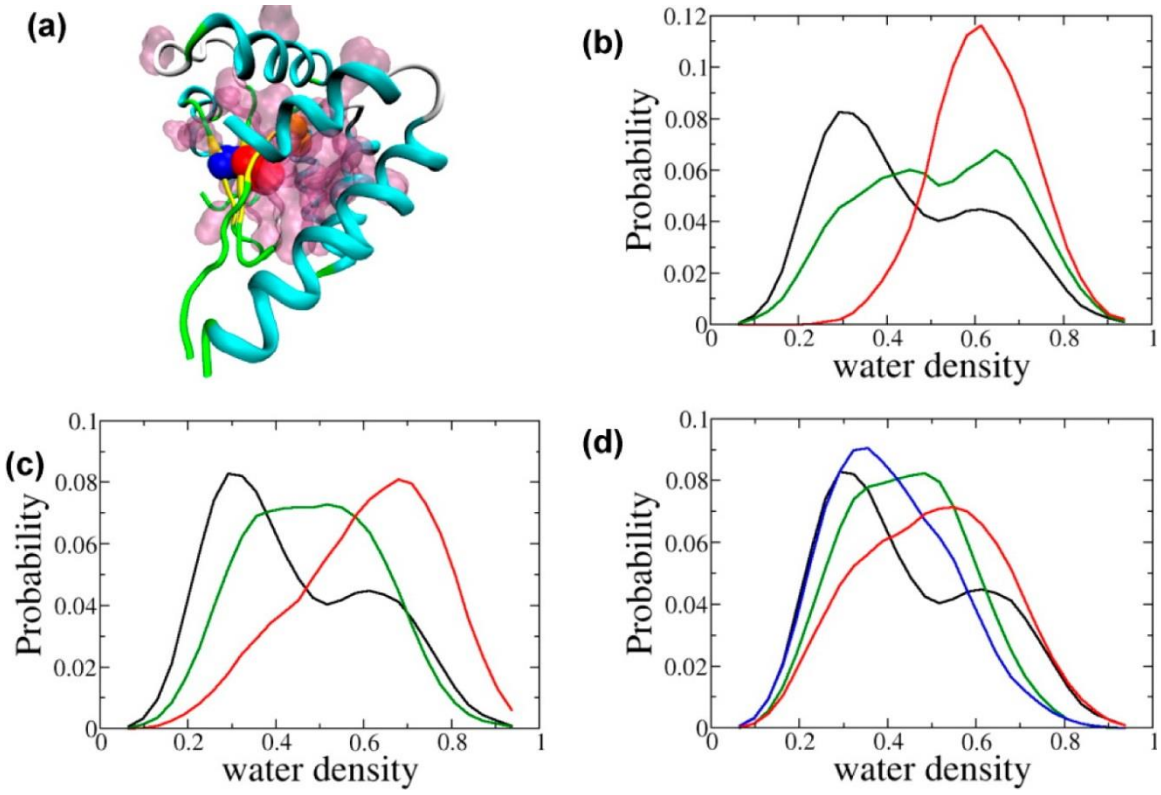


Figure 2. 4 (a) Position of Leu50 (red), Cys81 (orange), and Leu99 (blue) within ILV cluster 1. Mutation sites are shown as van der Waals spheres. (b–d) Histograms of water density inside the cavity of cluster 1: (b) L50 with WT in black, L50A in green, and L50N in red; (c) L99 with WT in black, L99A in green, and L99N in red; (d) C81 with WT in black, C81I in green, C81V in blue, and C81N in red.

The histograms for those two asparagine mutants have a water density peak centered at 0.6-0.7; the shoulder near 0.4 for L99N shows a limited propensity for dewetting. As will be confirmed experimentally, these results led to the expectation that native states and the folding intermediates for the L50N and L99N variants of αTS would be substantially destabilized relative to their wild-type counterpart.

Probing Non-ILV Positions for Effects on Hydration of Cluster 1. Inspection of the simulations for wild-type αTS found residual water density near C81, which is adjacent to Cluster 1 in helix α_2 and near the C-termini of strands β_3 and β_4 (Figure 2.4a). To discover the effect of side chains proximal to Cluster 1 on hydration, C81 was substituted with isoleucine, valine, and asparagine, respectively. The simulated water density distributions of C81I and C81V mutants (Figure 2.4d) showed a cavity that fluctuates between wet and dry states, C81V making the cavity noticeably drier compared to the wild-type protein. These results suggest that C81V mutation would be the best candidate to further dewet Cluster 2. Unfortunately, the larger steric bulk of valine versus cysteine precludes an unambiguous experimental test of this conjecture. In contrast, the cavity in the C81N variant favors the wet state with the maximum of water density probability around 0.6 (Figure 2.4d). These findings illustrate the sensitivity of water probability inside the cavity to the local environment.

Hydration in Clusters 2 and 3. We also compared the water density fluctuations of the two other ILV clusters of αTS for the separation distance of 4 Å

(Figure 2.2d). The ~ 500 Å^3 cavities for Cluster 2, created by displacing the α_5 and α_6 helices and the β_5, β_6 and β_7 strands, experienced strong fluctuations in water density during the 16 ns simulation time. However, this cluster did not experience extended periods of dehydration at a separation distance of 4 Å. The ~700 Å^3 cavities in Cluster 3 was created by pulling the α_0 helix away from the β_1 and β_8 strands. This cluster showed hydration even at a separation of 4 Å for most of the simulation time (Figure 2d); thus, Cluster 3 was not considered further. In contrast to the behavior of Cluster 1, the cavity in Cluster 2 primarily remained in the wet state even at a small separation distance of 4 Å. Closer inspection showed that drying is more favored toward the C-terminus of β-strands. In particular, residue L176 protrudes from β_6 toward the α_5 and α_6 helices, acting as a barrier inside the cavity (Figure 2.5a).

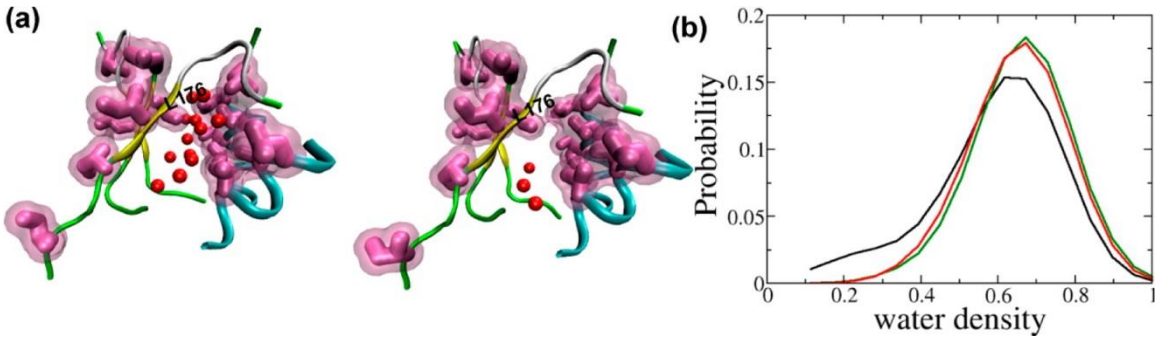


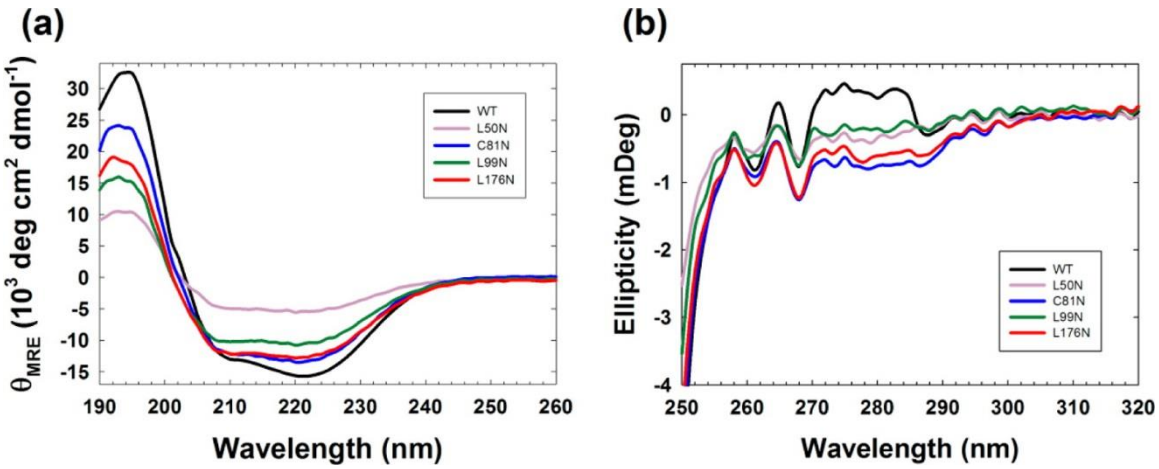
Figure 2.5 Dewetting of ILV cluster 2. (a) Snapshots of typical wet (left) and dry (right) states of cluster 2 populated during the simulation. The cluster is oriented in such a way that the bottom of the figure is the N-terminus of the β-strands. The separation distance was 4 Å. The maximum number of water molecules within the observation volume of ~500 Å3 was 17. (b) Histograms of water density inside ILV cluster 2 for the wild-type protein (black) and its two mutants, L176A (green) and L176N (red)

To check the sensitivity of this position to mutation in terms of dewetting, we performed two in silico mutations, L176A and L176N. The water density distribution of the wild type protein indicated that the cavity in Cluster 2 prefers the wet state; however, there was a small but not insignificant probability for drying (Figure 2.5b). Both mutations resulted in enhanced wetting of the cavity, as the water density distributions shift to the high-density side with a peak around 0.7 (Figure 2.5b). The distributions of the two mutants appear almost identical, suggesting that an alanine substitution is sufficient at position 176 to further wet the cavity. This behavior contrasts with that for positions L50 or L99 in Cluster 1, where a much stronger perturbation, such as mutation to asparagine, is needed.

Experimental Analysis of Structure and Stability for Hydration Mutations in ILV Clusters. The effects of the asparagine hydration mutations on the structural properties of αTS were determined by CD spectroscopy. The far-UV CD spectra of the Cluster 1 variants L50N, C81N, L99N, and the Cluster 2 variant L176N all display a broad negative minimum between 222 nm and 208

nm and a positive band at 195 nm (Figure 26a), indicative of α-helix and β-sheet contributions.

Figure 2. 6 (a) Far-UV CD spectra of wild-type αTS and the L50N, C81N, L99N, and L176N variants from 190 to 260 nm with protein concentrations ranging from 3 to 7 μM. (b) Near-UV spectra of wild-type αTS and the L50N, C81N, L99N, and L176N variants from 250 to 320 nm.



However, the reductions in the ellipticities at 195 and 222 nm for the variants show the introduction of a polar side at all four positions disrupts the secondary structure to varying degrees. The C81N and L176N mutations decrease the ellipticity at 222 nm by 20%. Surprisingly, the L50N and L99N mutations have a more dramatic effect, decreasing the ellipticity by 70% and 40%, respectively. The near UV-CD spectra, which provide insight into the chiral packing of aromatic side chains, reveal that all of the hydration variants have altered tertiary structures (Figure 2.6b). The positive band observed for tyrosines between 270 and 285 nm for wild-type αTS becomes negative for C81N and L176N and is eliminated for L50N and L99N. The phenylalanine bands between 255 and 270

nm are present for all of the variants, however, the bands at 265 nm are comparably reduced in magnitude vs. wild-type αTS for L50N and L99N.

The effects of the mutations on the thermodynamic properties were determined by monitoring the far-UV CD spectrum as the proteins were denatured with urea. The equilibrium unfolding transitions, as illustrated by the changes in ellipticity at 222 nm, are shown in Figure 2.7.

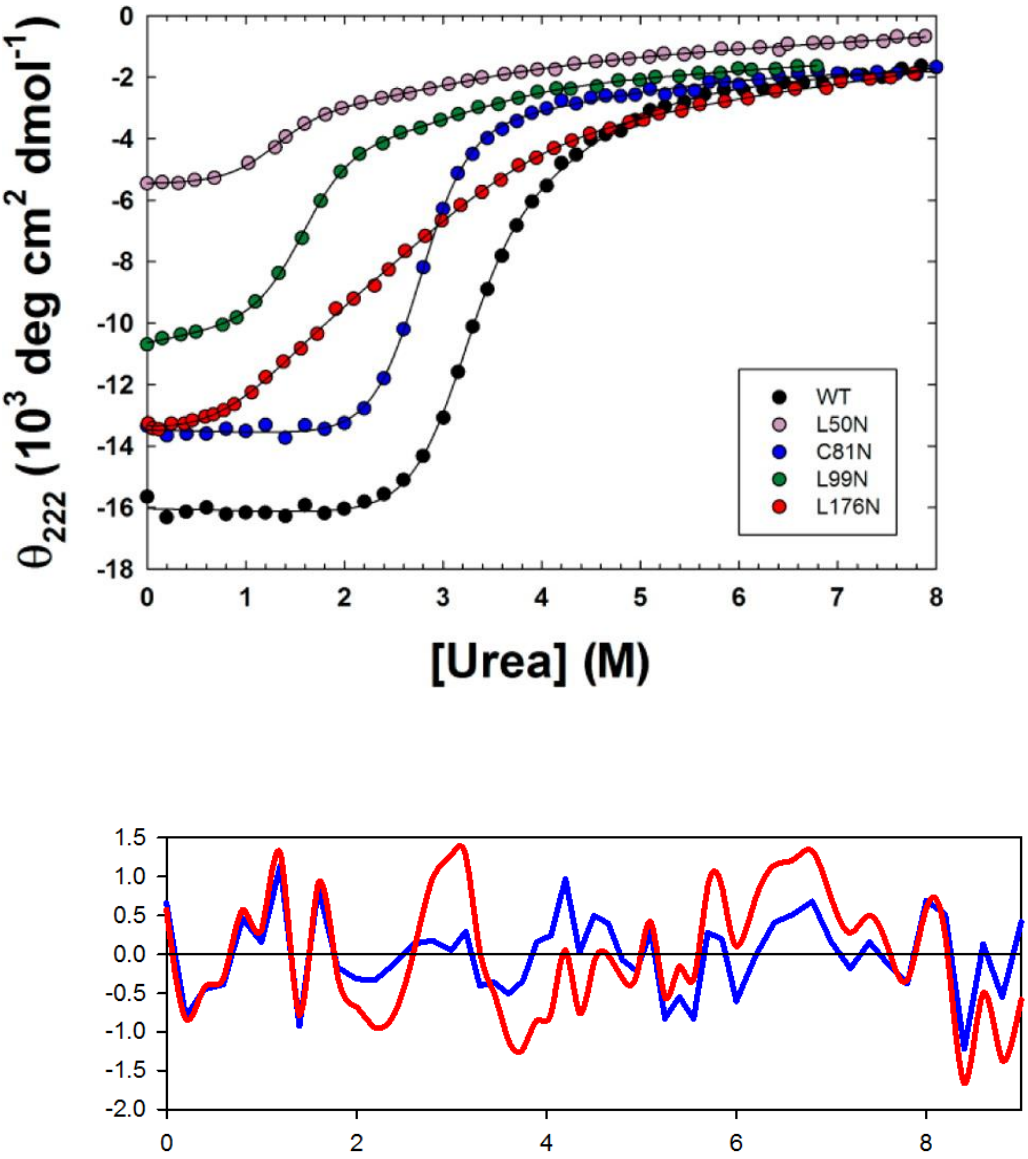


Figure 2. 7 (A) Urea-induced equilibrium unfolding profiles for wild-type αTS and the hydration variants L50N, C81N, L99N, and L176N. The continuous lines represent fits of the data to a three-state model. (B) The residuals for the C81N fit of the raw data before conversion to MRE are shown in the for the two state (red) and three state (blue) models. The large deviation from 2 M urea onward in the 2 state fit indicates a poor fit to the 2 state model.

All the variants display a nearly urea-independent baseline indicative of a thermodynamically stable state in the absence of denaturant. As previously observed for wild-type αTS[35], the equilibrium unfolding reactions of the four hydration variants are well-described by a 3-state model, N ⇌ I ⇌ U, to fit the CD data. For the L176N variant with a limited native baseline, the stability of the N ⇌ I transition was determined by measuring the amplitude of the rate-limiting N → I unfolding phase as a function of the initial urea concentration while jumping to the same final urea concentration. Because the amplitude is proportional to the

Table 2.1: Thermodynamic parameters for urea-induced unfolding of wild-type and hydration variants of αTS[a]

	$\Delta G_{NI}°(H_2O)$[b,c]	$-m_{NI}$	$Cm_{(NI)}$	$\Delta G_{IU}°(H_2O)$	$-m_{IU}$	$Cm_{(IU)}$	Z-value	$\Delta G_{(total)}°(H_2O)$	$-m_{total}$
WT	6.60±0.10	2.05±0.03	3.22±0.06	4.59±0.54	1.09±0.09	4.21±0.60	0.68±0.05	11.19±0.55	3.14±0.10
L50N	3.01±0.13	2.30±0.09	1.31±0.07	1.90±0.25	0.72±0.06	2.64±0.41	0.58±0.05	4.91±0.28	3.02±0.11
C81N	6.47±0.16	2.29±0.06	2.83±0.10	2.30±0.19	0.57±0.05	4.04±0.48	0.80±0.00	8.77±0.25	2.86±0.07
L99N	3.77±0.05	2.36±0.03	1.59±0.03	3.27±0.21	1.03±0.05	3.17±0.26	0.75±0.02	7.04±0.22	3.39±0.05
L176N	0.65±0.19[d]	0.72±0.11[d]	0.90±0.30	3.13±0.12[e]	0.94±0.03[e]	3.33±0.15	0.86±0.05[e]	3.78±0.22	1.66±0.11

[a] The equilibrium unfolding data were fit to a three-state model, N⇌ I⇌ U. $\Delta G°(H_2O)$, m, and C_m represent the free energy of unfolding in the absence of urea, the urea dependence of the free energy of unfolding and the concentration of urea at the midpoint of transition, respectively.
[b]Units are as follows: $\Delta G°(H_2O)$, kcal mol^{-1}; m, kcal mol^{-1} (M urea)$^{-1}$; Cm, M (urea).
[c]Errors for $\Delta G°(H_2O)$ and m are standard errors from the fits. Errors in Cm, $\Delta G_{(total)}°(H_2O)$ and $-m_{total}$ were obtained by standard error propagation of the equation: $C_m=\Delta G°(H_2O)/m$; $\Delta G_{(total)}°(H_2O) = \Delta G_{NI}°(H_2O)+ \Delta G_{IU}°(H_2O)$; $-m_{total}= -m_{NI} + -m_{IU}$
[d]Values obtained from fitting the amplitudes of the N → I kinetic unfolding reaction to a two state model.
[e]Values obtained from equilibrium data with the N ⇌ I transition constrained by the values found from the kinetic unfolding experiment.

fraction of the native state at the initial urea concentration, the fit of the amplitude to a 2-state model yields the desired thermodynamic parameters.

The free energy differences for the N ⇌ I and I ⇌ U transitions for the variants are shown in Figure 2.8A and Table 2.1. The stability of the N state vs. the I state is substantially decreased for the L50N, L99N and L176N mutations,

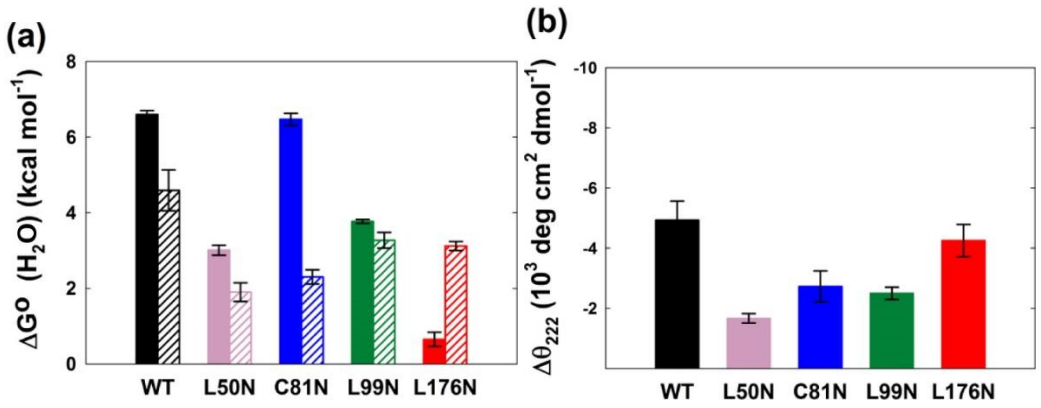


Figure 2. 8 (a) Bar graph showing the free energy differences for the N ⇄ I (solid bars) and I ⇄ U (hatched bars) transitions obtained by fitting the data for the hydration variants to a three-state model. (b) Bar graph showing the differences between the mean residue ellipticities of the intermediate (I) and unfolded (U) states at 222 nm (ΔMRE) for wild-type αTS and the L50N, C81N, L99N, and L176N variants.

however, the C81N mutation leaves the stability virtually unchanged. In addition to the stability, the fits also provide the m-value, a measure of the sensitivity of the folding free energy to the denaturant concentration that is proportional to the change in buried surface area. The average of the m-values of the N ⇌ I transition for the L50N, C81N and L99N variants, <m> = 2.32 ± 0.04 kcal mol^{-1} M^{-1}, is larger than for the wild-type αTS, 2.05 ± 0.03 kcal mol^{-1} M^{-1}, suggesting that all three are less well-folded in the I state (Table 2.1). The smaller m-value

for the N ⇌ I transition for L176N, 0.72 ± 0.11 kcal mol^{-1} M^{-1}, could reflect a less compact folded state or the existence of additional intermediates in the conversion of N to I. If present, the additional species would lead to an overestimation of the perturbation in stability for the native state.

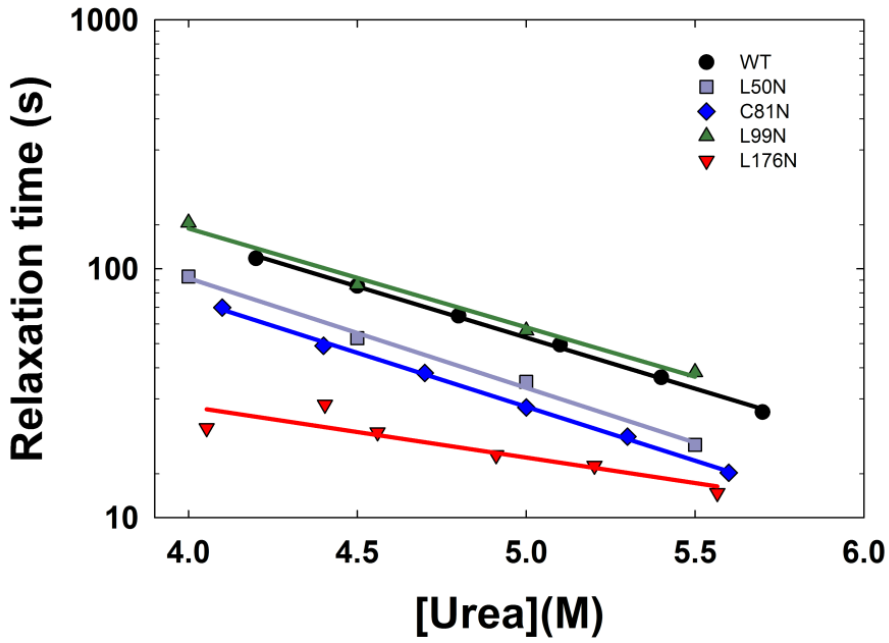
The stability of the I state vs. the U state is reduced for all four αTS variants, indicating that both clusters are sensitive to the state of hydration of the mutated side chains. The reductions in the m-values for the I ⇌ U transition for L50N and C81N are very similar in magnitude to the increases seen for the N ⇌ I transition (Table 2.1), again suggesting a less well-folded I state. The L99N variant, however, does not display this behavior and, with a total m-value of 3.39 ± 0.05 kcal mol^{-1} M^{-1} versus an average of 2.94 ± 0.07 kcal mol^{-1} M^{-1} for the wild-type, L50N and C81N variants, may experience a disruption of residual nonpolar structure in the U state. The unfolding of the intermediate for the L176N variant exposes a comparable amount of buried surface area as the wild-type protein.

Insights into the impact of the mutations on the secondary structure of the intermediate can be calculated from the Z parameter employed in the 3-state fits to the equilibrium unfolding data (Materials and Methods). The Z parameter reflects the normalized change in ellipticity of the intermediate state relative to the unfolded state and is defined as $Z = (\theta_I - \theta_N) / (\theta_U - \theta_N)$. Rearranging to extract θ_I, the ellipticities of the intermediates vs. their respective unfolded states for all three polar replacements in or near Cluster 1 were found

to decrease by 2-3 fold (Figure 2.8b). By contrast, the ellipticity of the intermediate for the L176N variant is very similar to its wild-type counterpart.

Although all the asparagine replacements retain a 3-state unfolding profile, the disruption of as much as 70% of the ellipticity in the L50N variant raises the possibility that their folded states no longer reside in the native basin for wild-type αTS. To explore this issue, denaturant jumps from the native state to the

Figure 2. 9 Semi-log plot of the urea-dependence of the observed unfolding phase monitored by manual mixing CD for WT αTS and the hydration variants. The slopes are within 10% for the mutants in the large N-terminal cluster.



unfolded state were employed to monitor the rate limiting, N → I, unfolding reaction. All of the variants display a slow unfolding reaction whose relaxation times differ less than a factor of 5 from that for the wild-type protein and, similar

to wild-type αTS, decrease exponentially with increasing denaturant concentration (Figure 2.9). The denaturant dependence of the observed relaxation times for the L50N, C81N and L99N variants varies less than 10% from that for wild-type αTS (Table 2.2).

Table 2.2: Unfolding kinetics *m*-values (kcal mol^{-1} (M urea)$^{-1}$)	
Wild type	-0.56±0.01
L50N	-0.60±0.03
C81N	-0.57±0.01
L99N	-0.54±0.04
L176N	-0.28±0.07

The minimal perturbation of the N →I unfolding dynamics for these L → N variants, all in or near Cluster 1, vs. the wild-type protein, implies that the mutations have a very similar effect on the energies of the native state and the transition state ensemble (TSE). The similar denaturant dependences for these variants imply the exposure of a comparable amount of buried surface to access the TSE**.** By contrast, the L176N variant has a 50% reduction in the denaturant dependence for its unfolding reaction (Table 2.2). When considered with the more than 50% decrease in the m-value for the N → I reaction at equilibrium (Table 2.1), the folded state of L176N must be less compact than for wild-type αTS. The retention of the 3-state unfolding model, a similar degree of compaction implied by the total of the m-values for the two transitions (with the exception of the L176N variant) (Table 2.1) and the same barrier to

unfolding all argue that the variants occupy the same native basin as wild-type αTS.

Discussion

We have characterized the relationship between nanoscale dewetting transitions and the stability and structure of αTS, a TIM barrel protein, using a combined molecular dynamics simulations and experimental approach. Simulations reveal that cavities created inside the two large hydrophobic ILV clusters of αTS undergo either intermittent or strong water density fluctuations, depending on the size and composition of the cluster. The largest ILV cluster (Cluster 1) was found to be optimized in terms of dehydration. Substituting selected ILV residues with alanine was found to weaken or completely diminish dewetting, which strongly depends on the local environment of the mutation site. Our simulations also showed that the replacement of buried leucines in both clusters with asparagines is sufficient to completely wet their cavities.

The experiments performed in this study suggest a folding mechanism in which segments of large ILV clusters adopt folded-like conformations prior to final collapse and expulsion of water. In our simulations, we introduced a cavity by separating α-helices from β-strands of the ILV clusters in their native structures to capture how single amino acid substitutions to the critical ILV residues affect this final stage of folding. Additionally, we did not observe any significant conformational changes upon mutation from ~100 ns long “folding”

simulations of the wild-type protein and the L50N and L99N mutants, which was not unexpected because the conversions of the U state to the I state and the I state to the N state occur on the millisecond time scales[39].

As found in our simulations, ILV Cluster 1 experiences frequent transitions between a wet and a dry state, whereas Cluster 2 sits on the wet side. Previous studies have shown that small perturbations, such as single amino acid substitutions, can shift the protein from a dry state to a wet state[73]. For example, a single I2A or I2V mutation can tip the protein melittin tetramer channel from dry to wet. Along this line, recent simulation studies by Garde and coworkers also show that water near protein surfaces can be sensitive to subtle changes in surface conformation, topology, and chemistry, and small changes can tip the balance from dry to wet or vice versa. That is, the protein can be "sitting at the edge" of the dewetting transition[75]. For example, melittin sits on the dry side of a dewetting transition, while another protein BPHC on the wet side. It is possible to tip the balance to the other side for both melittin and BPHC proteins by introducing additional perturbations, e.g. point mutations. Taken together, these findings by Garde and coworkers and our current results suggest that biomolecules often sit at the edge of dewetting transitions and are sensitive to perturbations[75]. We further show that such sensitivity to perturbations can be readily manipulated by protein engineering, which allows the TIM barrel protein to fine-tune its stability and folding.

Experimental analysis of leucine to asparagine mutations in the N-terminal ILV cluster in αTS not only demonstrated dehydration in both native and intermediate states but also revealed that the introduction of polarity substantially decreased the stabilities and had a dramatic effect on the structures of both states. The substitution of an acetamide group for an isobutyl group at L50 and L99 reduced the secondary structure of the native state by 40-70% and appeared to mobilize the tyrosine side chains. The secondary structures of the corresponding intermediate states were also greatly diminished for these variants. The results are consistent with the prediction that the interior of this cluster strongly prefers to dewet in a TIM barrel configuration and the conclusion that this configuration also exists for the intermediate state. What is very surprising, however, is that these mutations do not simply destabilize the TIM barrel fold or its folding intermediate. Rather, the presence of the polar side chain leads to distinct high-energy thermodynamic states in the native basin on the TIM barrel folding free energy surface. In the case of Cluster 1, there appears to be sufficient driving force from the need to sequester the remaining 30 aliphatic side chains from solvent to populate these alternative states. The substantial decrease in the CD signal at 222 nm and loss of signal at 280 nm could reflect a highly dynamic α-helical shell that enables the partial exposure of the asparagine side chains at positions 50 and 99 to water while retaining buried surface area.

Interestingly, the C81N mutation in helix α_2 and adjacent to Cluster 1 had a lesser effect on the secondary structure and left the stability of the native versus the intermediate state virtually unchanged (Table 2.1). Although the midpoint of the urea-induced transition, 2.83 M, is lower than wild-type, 3.22M, the larger m-value for C81N results in a stability that is coincidently the same as wild-type. As noted above, the increased m-value reflects a less well folded I state. The ready adaptation to the polar side chain at position 81 is similar to the previously-described response of the L85A mutation, also in helix α_2, reflecting the conformational adaptability of a surface helix[35]. The stability of the I state versus the U state and the m-value, however, were markedly reduced. The lower inherent stability of the I state apparently does not provide sufficient driving force to accommodate the asparagine side chain and maintain the secondary structure and compactness for the C81N variant.

Although an asparagine mutation in Cluster 2 had a lesser effect on the secondary structure in the native state and little or no effect on the intermediate, the L176N variant could not achieve the same stability or degree of compactness in the native state as the wild-type protein. The distinct changes in the near-UV CD spectrum might reflect perturbations in the packing of the adjacent Y173 and Y175 inside the β-barrel as well as more global effects accompanying the decreased packing efficiency. In contrast to the predictions of the simulations, the wet state favored for the interior of this cluster was not capable of supporting the presence of the polar side chain in

the native conformation. The contradiction may reflect the smaller size of Cluster 2, leading to only marginal drying of the cavity in the simulations (Figure 2.5). The limited effect of the L176N mutation on the stability and secondary structure of the intermediate state, in contrast to the substantial effects on the native state, suggests that the side chain is only partially dehydrated at this stage of folding. All of these findings are consistent with the previous conjecture that the region encompassing Cluster 1 is well packed in the intermediate state while Cluster 2 is best described as a loosely-folded, molten globule-like structure[46,76].

It was surprising that the MD simulations for the entire set of 10 ILV → A mutations in Cluster 1 resulted in the wetting of the cavity. One might have expected that the removal of 2-3 carbons from a cluster of 31 branched aliphatic side chains would have little effect on the propensity of water to occupy the exposed nonpolar volume. However, the sensitivity of drying to the composition and/or structure of the cavity may be the explanation for the previous experimental observation that these same alanine replacements substantially reduce the stability of the intermediate in αTS[35]. The simulations suggest that the enhanced propensity of the alanine variants in Cluster 1 to wet, i.e., favor a less well-folded state, is, along with the loss of packing interactions, a mechanism for destabilizing the intermediate. The tendency of the cavity in Cluster 2 to wet in the wild-type αTS would mitigate any enhanced

hydration from alanine replacements and minimize the perturbation on the stability of the intermediate, as observed.

An unanticipated outcome of creating the L50N and L99N variants was the discovery of discrete thermodynamic states that have substantially disrupted secondary structure and the apparent loss of tight packing around the 7 tyrosines with a compactness comparable to the wild-type protein. Although further experiments are required to rule out the coincidental cancellation of positive and negative bands for the tyrosines and confirm their putative dynamic properties, native-like compactness with mobile side chains are characteristics of the “dry molten globule”[77]. The dry molten globule was initially proposed as a model for folding transition states or to arise in a membrane environment[40] and, subsequently, as a discrete state in the native basin[78–80]. The putative dry molten globule states for the L50N and L99N variants of αTS, however, do not unlock all of the phenylalanines and have a substantially altered secondary structure compared to the canonical TIM barrel. Further studies are required to determine if the folded states of the L50N and L99N variants are indeed dry molten globules, as envisioned by Shakhnovich and Finkelstein[40], or represent related high energy states in the native basin. Intriguingly, the existence of such states might provide a path for the evolution of the sequence to produce TIM barrels with alternative locations for their ILV sequences[49].

The results of this combined experimental-simulation study on αTS demonstrate the critical role of dehydration in a large hydrophobic ILV cluster in determining the stability and structure of a TIM barrel fold and a critical folding intermediate. ILV clusters are common in the other (βα)-repeat motifs, such as the flavodoxin-fold and the Rossmann-fold families[81], and they also define coiled coils[82,83], repeat-sequence proteins[84–86], β-sandwich motifs[87], and anti-parallel β-sheet arrays found in amyloidogenic peptides[88]. Thus, our findings may provide useful insights into the link between hydrophobicity, dewetting, and stability of a large number of protein motifs.

Materials and Methods

Molecular Dynamics Simulations. The initial structures of the ILV clusters were taken from the crystal structure deposited in the Protein Data Bank (PDB ID code 1BKS). The clusters were determined using the same protocol described previously[35]. Three ILV hydrophobic clusters are found within the αTS native structure: (i) a large external-to-the-barrel cluster spanning the N- and the C-termini (Cluster 1); (ii) a second external-to-the-barrel cluster in the C-terminal region (Cluster 2) and (iii) an internal-to-the barrel cluster (Cluster 3) (Figure 2.1a). The helical parts of the clusters were pulled 4-6 Å away from the β-sheet region to create the cavity for investigating dewetting. The system was solvated in a box of TIP3P water. The initial state for the cavity for all systems was set to be wet, unless otherwise stated. The

resulting systems were minimized for 10000 steps followed by a 16 ns MD simulation at 310 K and 1 atm. During this simulation, the protein heavy atoms were constrained, whereas water molecules were free to move. The particle-mesh-Ewald (PME) method was used for the long-range electrostatic interactions, while the van der Waals interactions were treated with a cutoff distance of 12Å. The CHARMM (c32b1 parameter set) force field was used and simulations were performed using NAMD2 molecular modeling package with a 2 fs time step. At least 15 different trajectories were run for Cluster 1 and Cluster 2 of the wild type protein and its C81, L99, and L176 mutants (in silico variants). For all other systems, at least three different trajectories were run. The total aggregate simulation time was about 5 μs. Additionally, we performed ~100 ns long “folding” simulations, in which both main chain and side chains atoms were free to move.

Site-directed Mutagenesis. The codon-optimized αTS WT gene was synthesized by Genscript in pUC 57 and re-cloned into a modified pGS-21a vector with an N-terminal 6X His tag and TEV protease site using EcoRV and BamHI restriction sites. Various hydration mutations were made using mutagenic oligonucleotides purchased from Integrated DNA Technologies using the Stratagene Quick-change site-directed mutagenesis kit and mutations were confirmed using DNA sequencing. The pGS-21a plasmid DNA was transformed into BL21 (DE3) pLysS cells for protein expression and purification.

Equilibrium and Kinetic Unfolding Experiments. The thermodynamic properties of both wild-type αTS and various hydration mutants were determined by urea titrations on a Jasco J-810 spectropolarimeter. Samples at varying urea concentrations were prepared using a Hamilton 540B automatic titrator and were incubated overnight at 25 °C for complete equilibration. Data were collected using a 2 mm pathlength quartz cuvette and a 2.5 nm bandwidth. The spectra were recorded at every 1 nm in the wavelength range from 215 nm to 260 nm with a scan speed of 50 nm min-1 and an eight second averaging time. The denaturant dependence of the ellipiticities for αTS and its variants was fit to a three-state model using Savuka, an in-house nonlinear least squares program, and assuming a linear dependence of the free energy of unfolding on the denaturant concentration[39]. These fits provided the free energy differences between the three thermodynamic states, the denaturant dependences of these free energy differences and the Z parameter required to estimate the ellipticity of the intermediate[44].

The manual-mixing kinetic unfolding jumps began in the absence of denaturant and ended between 4.0 M to 6.0 M urea, with the final protein concentration ranging from 3-5 μM. Data were collected at 222 nm and at 25 °C in a 1 cm pathlength cuvette. The relaxation times were obtained by fitting the kinetic traces to a single exponential function in Savuka[39].

Chapter III: The role of ILV clusters during the early events of folding

This chapter presents results from a collaboration with Dr. Brooks' Lab at the University of Michigan and Dr. Srinivas Chakravarthy and Dr. Tom Irving at the BioCAT beamline at APS. The simulations were performed by Yanming Wang and Dr. Karunesh Arora while I performed the experiments with the help from the team at BioCAT and Dr. Osman Bilsel.

Introduction

Protein folding energy landscapes have been driven by evolution to minimize the energetic and topological frustration experienced during the folding reaction[26,89]. Small, two-state folding proteins, whose folding rates vary inversely with the complexity of their topologies, are a very good example of this evolution at play[90,91]. However, larger proteins are often found to have intermediates populated along their energy landscape, some of which may be misfolded or off-pathway[39,60,92–95].

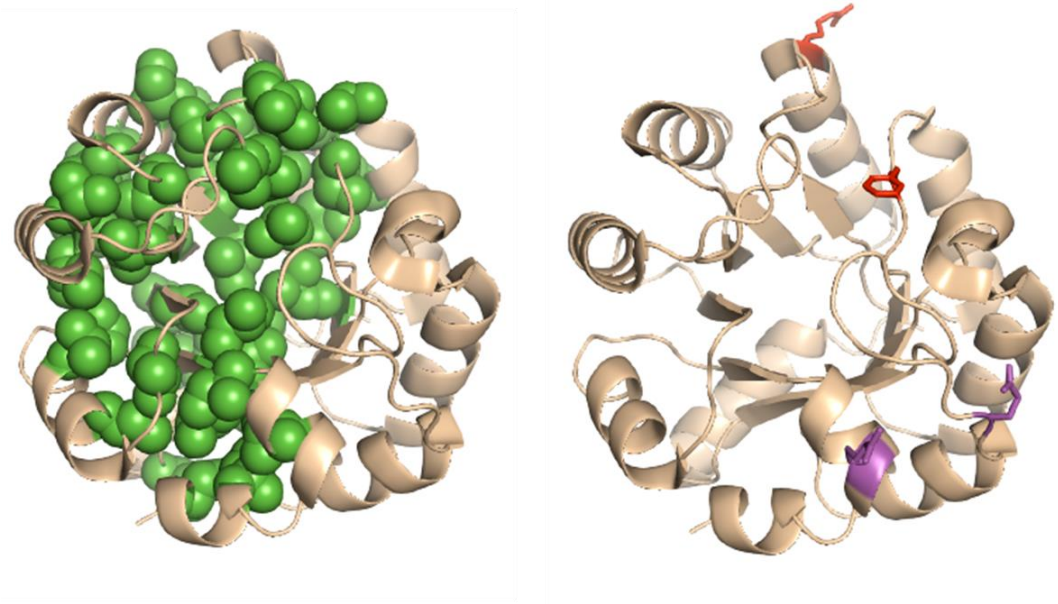
Simulations have shown that premature formation of structures can lead to topological frustration. As a result, such proteins are unable to reach their proper transition state and must at least partially unfold in order to continue proceeding on their kinetic pathway to the native state. Experiments and simulations have revealed that these intermediates may be native-like in secondary structure but contain nonnative interactions[96] or structural elements not found in the native structure [97]. Experimentally, early folding intermediates are difficult to study as the timescales associated with their formation are typically in the time range of 10s of microseconds or faster [46,92,98–100].

$(\beta\alpha)_8$ TIM barrel proteins, one of the most common motifs in biology[1], are one such class of proteins to have a misfolded intermediate [36,46]. Previous folding studies on several homologs TIM barrels with low sequence identity have shown that the general folding mechanism is conserved across the barrel architecture indicating the topology of the protein determines the general free

energy landscape [37,39,49]. Hydrogen exchange experiments on several barrels have highlighted the important role individual sequences play in determining the structures formed along the landscape as the regions of strong protection formed during the folding reaction vary from barrel to barrel [37,45,50]. One hypothesis is that large, sequence-local clusters of isoleucine, leucine, and valine (ILV) residues that vary in location from protein to protein drive the formation of the intermediates, including the initial off-pathway intermediate [32].

To continue our assessment of the formation of structure throughout the folding of TIM barrels, especially at early times, we are expanding our studies on the indole-3-glycerol phosphate synthase from *S. solfataricus* (sIGPS). The folding pathway of sIGPS has previously been shown to follow the expected mechanism of TIM barrel proteins with the formation of an off-pathway kinetic trap followed by two on pathway intermediates before the native state[49]. Stopped-flow circular dichroism experiments revealed that within 5 milliseconds the protein has acquired roughly two thirds of its native CD signal. This early intermediate is predicted to have a stability of 3.5 kcal mol^{-1}. However, stopped-flow fluorescence experiments showed that this early intermediate was a kinetic trap as there was a “rebound” reaction on the timescale of 100s of milliseconds[36]. The structure of the kinetic trap and on-pathway intermediates were probed by quench-flow hydrogen exchange experiments. The early kinetic trap displayed strong protection in the central region of the protein (βα4) within the 75 ms dead time of the experiment. Longer time points showed a

progression of protection out to the N and C termini of the protein with $(\beta\alpha)_1$ and $(\beta\alpha)_8$ being the last to show strong protection. This would suggest the protein folds via a nucleation and condensation mechanism [21] with the central module of the 4-fold pseudo-symmetry $(\beta\alpha)_{3\text{-}4}$ serving as the nucleation site of structure formation.



$$\mathbf{I_{BP} \rightleftharpoons U \rightleftharpoons I_A \rightleftharpoons I_B \rightleftharpoons N}$$

Figure 3. 1 Ribbon diagram of sIGPS with the side chains of the large ILV cluster highlighted with green spheres. The two sets of FRET pairs are highlighted in the right structure with the 63-238 pair in purple and 112-140 pair in red. The FRET pairs were chosen to look at the role of the ILV cluster as well as the barrel closure process. The consensus folding mechanism for the TIM barrel is below the structures.

The formation of the kinetic traps is not well understood in protein folding and with advances in mixing techniques over the past few years[101], it is now possible to probe these intermediates with multiple techniques on the sub-millisecond timescale. To monitor the formation of structure in the kinetic trap in sIGPS we have performed continuous flow kinetic experiments interfaced with time resolved small angle x-ray scattering (trSAXS) and time resolved Förster resonance energy transfer (trFRET). Gō model simulations have provided insights into potential structures formed throughout the folding reaction while providing insights into the cause of frustration in the early off-pathway intermediate. The combined experimental and computational approach towards probing the folding mechanism of one of the most common protein folds in biology has revealed shared and distinct features that enable detailed insights into the potential sources of frustration in folding in TIM barrel proteins.

Results

Measuring global dimensions by small angle x-ray scattering (SAXS)

To obtain structural insights on a global level of the intermediates, SAXS profiles were obtained under equilibrium and kinetic refolding conditions. At equilibrium, the native state of the protein has a calculated radius of gyration (Rg) of approximately 18 Å. The unfolded state, under native conditions, extrapolates to about 46 Å (Figure 3.2). Unfortunately, due the high protein concentrations required for the scattering experiment, insights into the

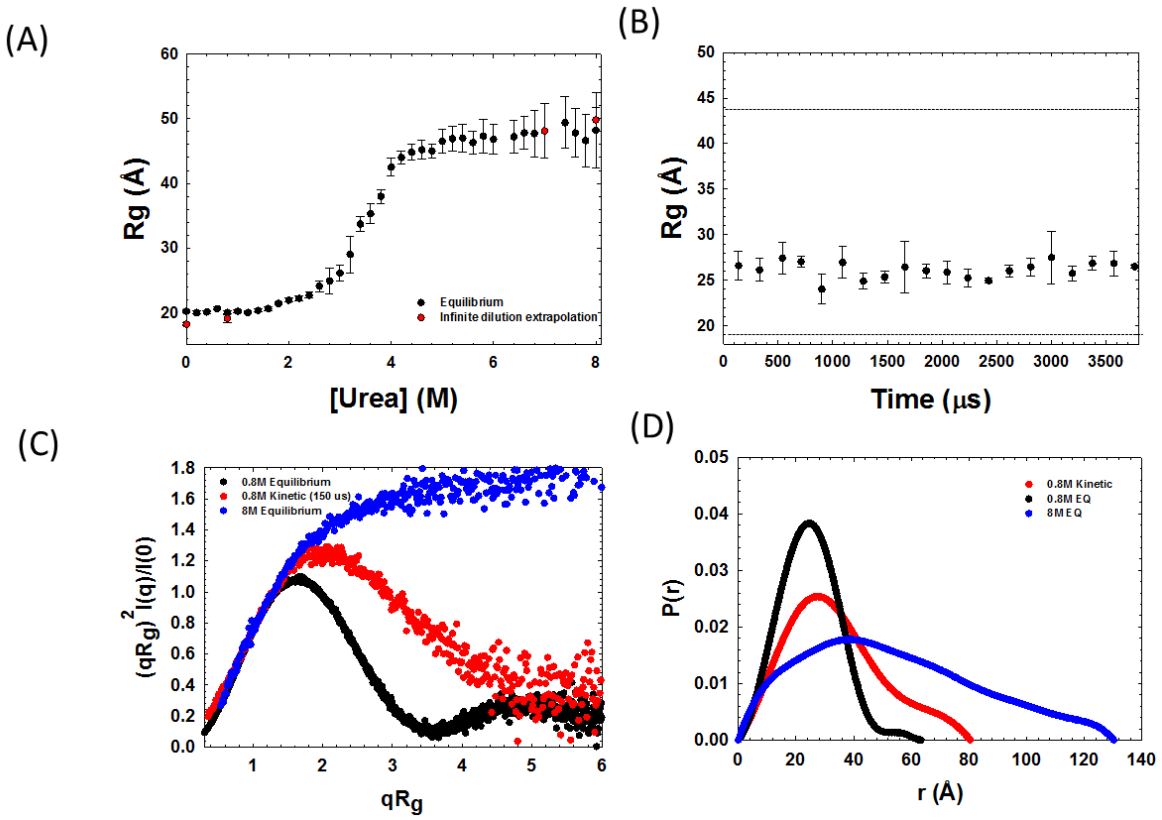


Figure 3. 2 (A) Rg as a function of [urea] shows a single transition from the native to the unfolded state. (B) Calculated Rg values from the continuous flow data set. Dashed lines at ~45A and 18A represent the unfolded and native state Rgs. (C) Dimensionless Kratky plot of the unfolded (blue), I_{BP} intermediate (red) and native state (black). (d) The P(r) of the unfolded (blue), I_{BP} intermediate (red) and native state (black). The Kratky and P(r) for the burst phase intermediate show partial globularity.

equilibrium intermediate state was not possible due to dimerization of the I_A intermediate, which was previously known to dimerize[49].

Previous stopped-flow CD experiments revealed a significant amount secondary structure, ~60% of the native signal, was formed within the 5 ms deadtime. The large amount of CD signal indicates that the polypeptide chain has undergone a large-scale contraction to form a significant amount secondary structure. To test the dimensions of the burst phase intermediate were measured by inducing ten-fold dilution kinetic jumps initiated from 8 M urea using custom, single piece microfluidic mixers. The scattering profiles starting from as early as ~150 μs were transformed to calculate the Rg, the Kratky plot and the P(r) function for the off-pathway burst phase intermediate from ~150 μs to ~4 ms (Figure 3.2). Based upon Rg, the protein chain collapses from the denatured, unfolded Rg of ~46 Å down to 26 Å within the dead time of the mixer. Unfortunately, no kinetics were observed in the mixer on the experimental timescale (~150 μs- 4 ms) as no change in Rg was measured (Figure 3.2B).

As no kinetic phases were seen over the experimental timescale, the measured collapse in the chain may be due to the change in solvent and not the formation of a discrete thermodynamic state. If the chain was collapsing solely due to the rapid change in solvents, the Rg would increase if weaker refolding jumps were performed due to swelling of the chain. However, weaker refolding jumps initiated from 8 M to a final of 1.2 M and 1.6 M urea show no

measurable change in Rg, 26 Å, when compared to the 0.8 M final refolding jump. This would indicate that the collapse of the chain to a Rg of 26 Å is due to the formation of a thermodynamically stable state. Now, because no kinetic phase was seen during the experiment, it can be inferred that the reaction of the unfolded state to the I_{BP} intermediate is much faster than the 150 μs dead time.

Transforming the 0.8 M urea equilibrium scattering curve to a dimensionless Kratky plot shows the typical globular parabolic shape with the maximum at $(\sqrt{3}, 1.1)$ as expected by Guinier's approximation (Figure 3.2C). The 8 M equilibrium sample shows an extended random coil like profile with the expected hyperbolic plateau shape. However, the curve from the continuous flow refolding kinetic jump to 0.8 M urea shows the I_{BP} intermediate has a peak shift on the x-axis to a qRg of approximately 2 with a maximum of 1.25 within 150 μs. This indicates a deviation from Guinier's approximation and that the protein has regions that are not yet fully globular. The P(r) distribution for the I_{BP} intermediate confirms that there is a large collapse of the chain as the peak of the distribution is at 26 Å. The maximum distance between a set of any two atoms, D_{max}, also decreases from 130 Å to 80 Å, however, there is a significant shoulder to the curve (Figure 3.2D). The Rg as calculated from the P(r) matches that of the Guinier analysis. Taken together, the backbone contracts significantly from the unfolded to the I_{BP} state, however, the intermediate is not yet fully globular or there are still extended regions of the protein.

Pair-wise dimensional analysis by time resolved FRET

To complement the global structural data obtained by SAXS, two sets of pair-wise distances were measured by time resolved tryptophan-AEDANS Förster resonance energy transfer (trFRET). By using trFRET, we are able to measure the lifetimes of the tryptophan donor and use that information to build distributions of distances, unlike total intensity FRET where only single distances can be measured. The first FRET pair was positioned at 63 and 238 (α_1 and α_8) to monitor the N- and C-termini and barrel closure (Figure 3.1). Based upon the ILV cluster map and the HX-MS data set [36], the second pair was positioned at 112 and 140 to cover the α_3-α_4 region (Figure 3.1). The $(\beta\alpha)_4$ region showed the strongest protection against exchange within 75 ms and contains a very hydrophobic stretch of amino acids with 11 out of 22 residues being I, L, or V.

The average Trp lifetime for the 63-238 pair as measured under equilibrium conditions shows no significant FRET in the denatured unfolded state at 8 M urea as expected for a Trp-AEDANS FRET pair (R_0 = 22 Å) that is 175 residues apart. A similar microfluidic mixer from the SAXS experiments was used in the trFRET experiments to assess the distances between FRET pairs in the burst phase intermediate. The continuous flow trFRET data for the 63-238 pair shows a rapid change to a non-native-like lifetime, 4.5 ns for the donor only sample and 3.3 ns for the donor-acceptor, within the dead time of

the mixer (~50 μs) for refolding jumps to 0.8 M final urea. (Figure 3.3). As was the case in the Rg measurements, there are no significant changes in lifetimes within both the donor only and donor-acceptor samples measured during the experiment from ~50 μs out to ~1 ms. This confirms that the I_{BP} intermediate forms much faster than the 50 μs dead time of the experiment. However, because the burst phase intermediate does show a difference in lifetimes between the donor only and donor-acceptor samples, measurable FRET is taking place.

The 112-140 FRET pair in the α_3-α_4 region, showed limited FRET signal in the unfolded state. During the rapid mixing refolding jumps to 0.8 M urea, non-native-like lifetimes are measured for the donor only and donor-acceptor, 4.6 ns and 3.7 ns respectively. Once more, the continuous flow kinetic experiment shows all the changes in the lifetimes for the donor only and donor acceptor pair takes place during the dead time of the experiment. Like the 63-238 pair, the donor only and donor acceptor samples for the 112-140 pair show differences in their average lifetimes during the experiment indicating FRET is taking place.

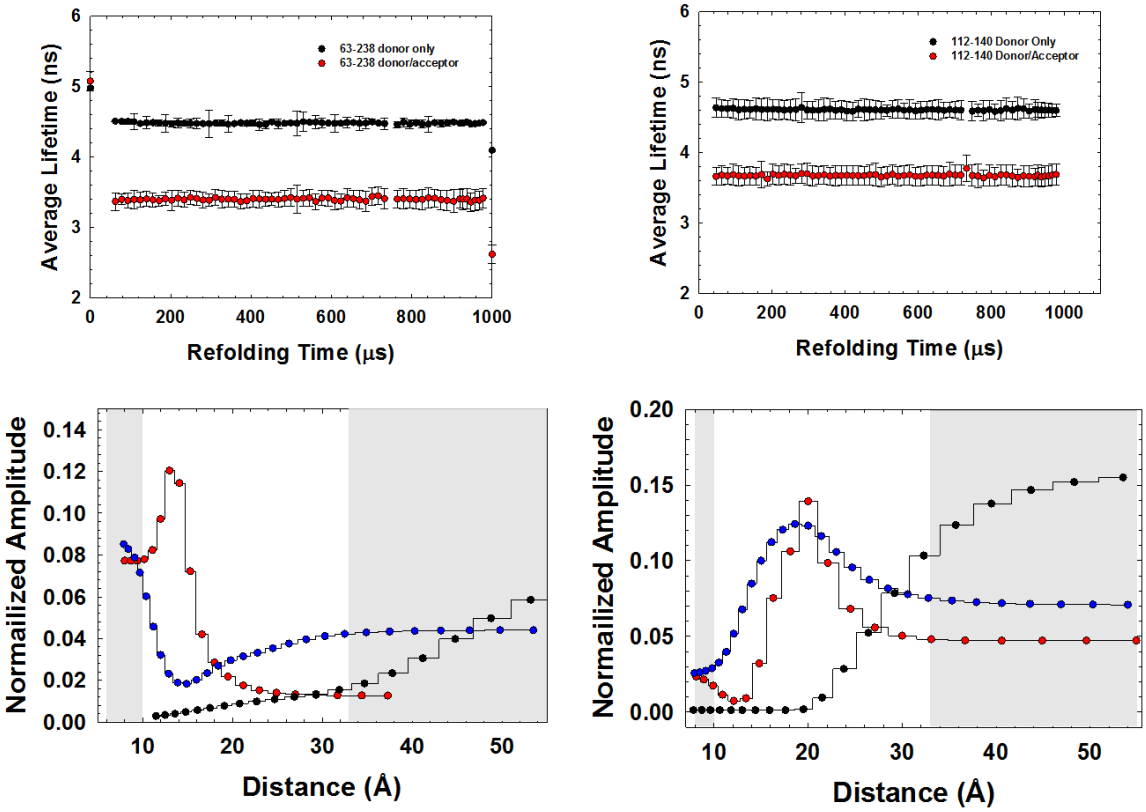


Figure 3. 3 (A/B) The average tryptophan lifetimes for the 63-238 and 112-140 FRET pairs during the continuous flow experiments. The kinetics of the unfolded state to I_{BP} were detected (C/D). The distance distributions from the MEM analysis for the unfolded (black), I_{BP} Intermediate (blue), and the native state (red). Areas in the shaded regions represent distances outside the optimal distances for the Trp-AEDANS fret pair. The 112-140 pair appears to have native-like distances within the deadtime of the experiment while the 63-238 pair has a compact and an extended conformation present.

Maximum Entropy Modeling

Obtaining distance distributions from simultaneous analysis of donor-only and donor-acceptor time-resolved fluorescence decays when multiple sub-populations are present can be challenging with established analytical approaches[102,103]. A limitation of these approaches is that they often assume a functional form for the distance distribution, assume the donor lifetime to be the same for all sub-populations or impose the same distribution on all sub-populations. These assumptions are not broadly applicable and the limitations of these assumptions may not be readily apparent in the analysis. The breakdown of these assumptions is of particular concern when tryptophan is used as the donor owing to the multiple rotamers often present in collapsed states. The multiple rotamers will have different lifetimes for each rotomeric state and cause difficulty in fitting the data to a single functional form. Other fluorophores also exhibit donor excited state decay rates that are sensitive to changes in the local environment[104,105].

The two-dimensional maximum entropy (2D-MEM) approach[46] overcomes many of these limitations by fitting the donor-only and donor-acceptor excited state decay traces using a two-dimensional grid of donor-rates and energy transfer rates. No assumptions are made about the number of donor rates and sub-populations or the functional forms of the donor rate

distribution and the distance distribution. The distributions are regularized by maximum entropy, ensuring that sub-populations are identified and associated with a corresponding energy transfer rate only if warranted by the data. The 2D-MEM approach avoids over-parameterization and identifies regions of the two-dimensional rate space (i.e., k_{donor} *vs*. k_{ET}) where the information content is limiting[106].

Additionally, in cases where the donor exhibits multiple populations and multiple excited state decay rates, the association between the donor lifetimes and energy transfer rates provides the necessary information to correct for variations in quantum yield when generating a distance distribution. Distance distributions can be constructed for specific sub-populations separated by their donor decay rate. Alternatively, a distance distribution can be obtained which contains contributions from all sub-populations, appropriately weighted by their relative quantum yields.

Both sets of FRET pairs were analyzed for the unfolded state (8 M urea), the I_{BP} intermediate (CF kinetics), and the native state (0 M urea) (Figure 3.3). The I_{BP} intermediate state kinetics were binned in 50 μs time bins to have sufficient photons in the decay traces for fitting.

The unfolded state for both the 63-238 and the 112-140 FRET pairs show very little amplitude from 12 to 35 Å, the distances most sensitive to FRET for the Trp-AEDANS pair. This was expected as the distances from donor and acceptor are predicted to average 115 Å and 40 Å apart if the protein

is expected to behave as a self-avoiding random coil in 8M urea. The native state for the 112-140 FRET pair has a major peak around 19 Å which corresponds well to the expected distance bewtween Cβs in the crystal structure, when accounting for the additional length of the EADANS moiety attached to the cysteine residue. The native protein for 63-238 pair measured about 13 Å, also is in agreement with the expected distance of 12 Å from the crystal structure.

The 112-140 pair continuous flow data shows a single major distribution around 20 Å that closely matches that of native protein (Figure 3.3D). The difference between the I_{BP} and native states comes in the widths as the I_{BP} state has a broader distribution. This is not surprising as the I_{BP} species is a transient kinetic intermediate. The tryptophan at 112 or the AEDANS at 140 may be more dynamic than in the native state and therefore contribute to the broader distribution of distances. The 63-238 pair during the kinetic experiment shows two distributions, one very high FRET state with a distance at approximately 10 Å and a second much broader distribution. This indicates that there are two species present during the kinetic refolding experiment. The two species might reflect the presence of the I_{BP} intermediate and the I_A intermediate.

Ensemble Averaged Folding Properties from Simulations

Using a native-centric coarse-grained Gō model, extensive refolding simulations were carried out with an unfolded initial structure just below the folding temperature. One hundred independent trajectories were generated and used for further analysis. Conventional folding reaction coordinates, radius of gyration (Rg) and fractional native contacts (Q), were used to monitor the folding process. Not surprisingly, multiple intermediates were observed to form as characterized by plateaus with various magnitudes of fluctuations (Figure 3.4). Interestingly, there appears to be three major plateaus (Q = 0.5-0.6, Q = 0.7-0.8 and Q ~0.85) before the native state is achieved during the trajectories which matches experimental results with the formation of three intermediates, I_{BP}, I_A, and I_B. When looking at all one hundred trajectories, it became clear that not all reached the native state and plateaued at an Rg just over 20 Å and Q of about 0.8 (Figure 3.4). Another set of trajectories were started from this point and allowed to go for another 8000 time steps with the majority of the trajectories still not reaching the native state. It is not clear if the presence of this “trapped” intermediate is a result of the I_{BP} intermediate having native like structures despite being off-pathway or if the long lifetime of the “stuck” intermediate during the simulations reflects the experimentally known rate limiting step in folding of the I_A to I_B taking 100s of seconds.

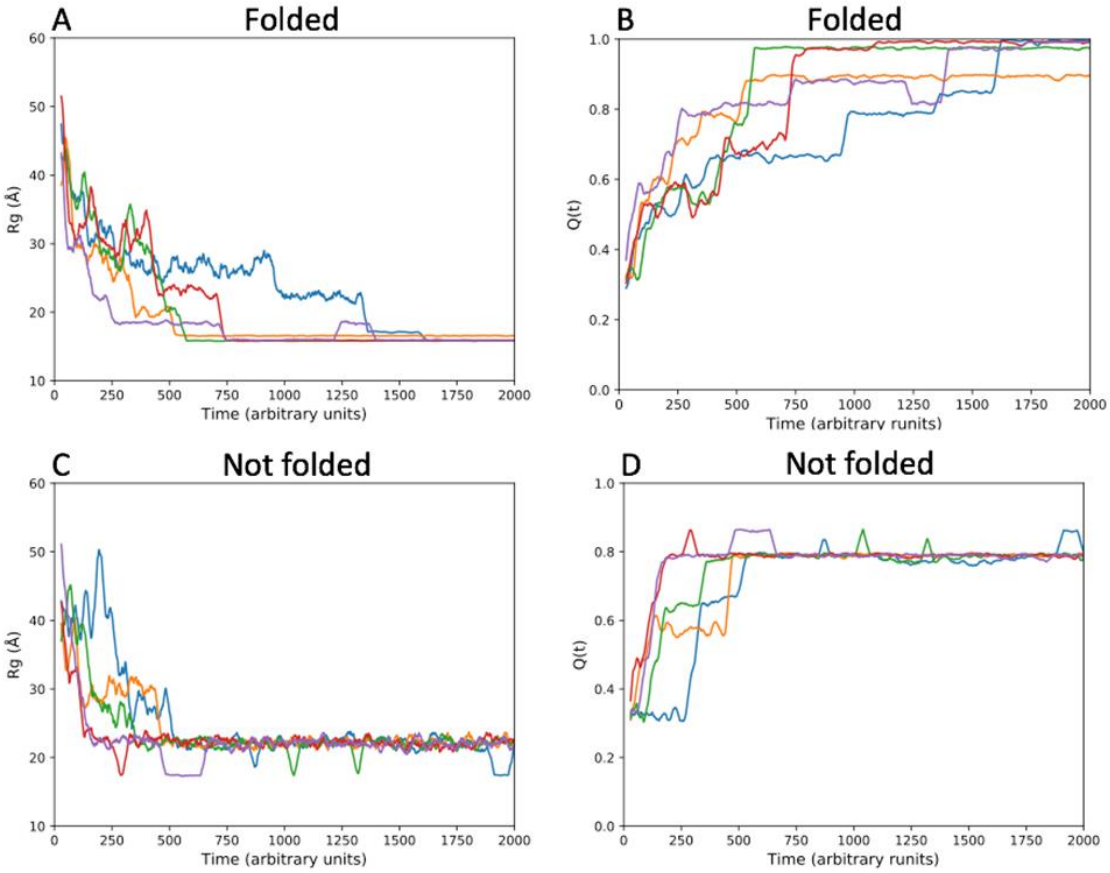


Figure 3. 4 Time evolution of Rg (A and C) and fractional native contact (B and D) from representative folding trajectories of sIGPS. For clarity, kinetic traces are shown as moving averages of 30 successive snapshots. The leveling off at various Rg and Q values indicates multiple intermediates are formed during the simulations.

The probability of pair-wise distances, P(r), for the alpha carbons were calculated based on the Gō model simulations and binned into a histogram for different fractional times (Figure 3.5). In addition, to monitor the sequence of the assembly of secondary structure units contact maps averaged over all

trajectories at the same time steps were examined (Figure 3.6). At time 0, most long-range native contacts were not formed and the protein was in the unfolded

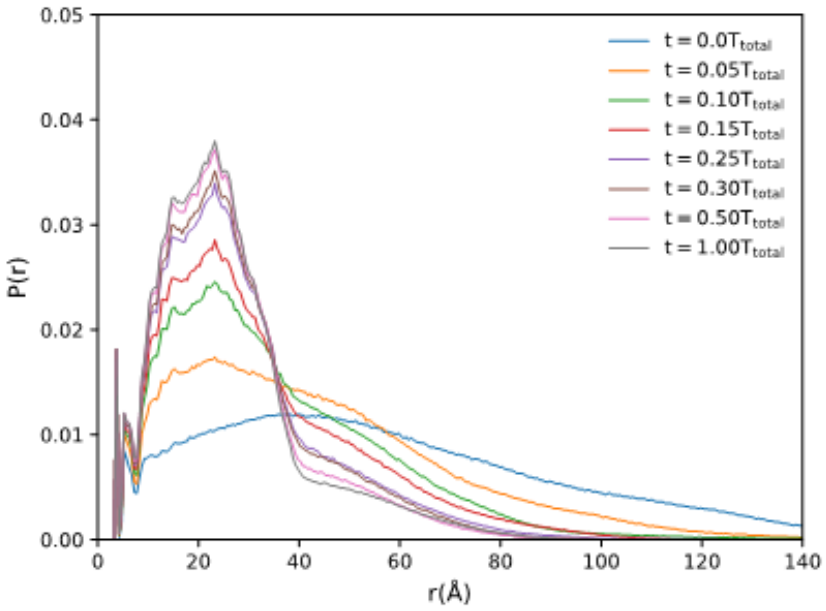


Figure 3.5 pair distribution functions at different time calculated from MD trajectories shows a progressive contraction of the r max with time. The decrease in amplitude of the shoulder of the distribution as the simulation time progresses indicates the formation of globular structure.

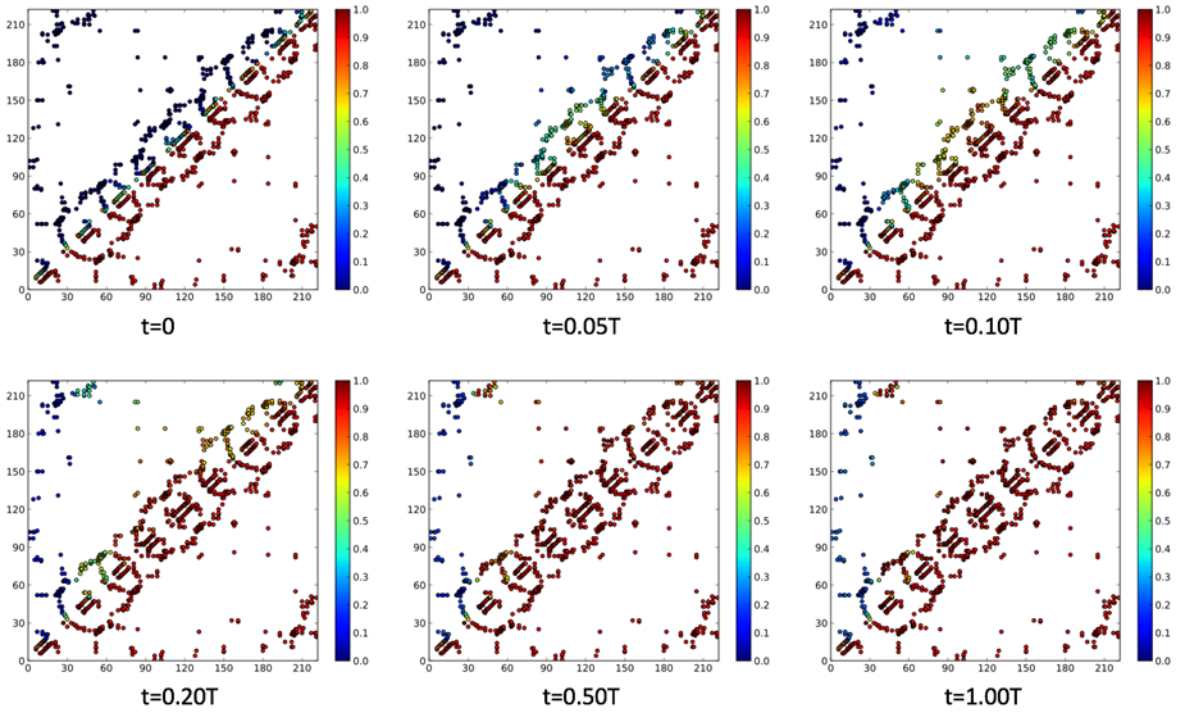


Figure 3.6 Ensemble averaged contact maps at different time relative to the total simulation time T. The red points on the lower right side of each subplot indicate the native contacts and the possibilities of forming native contacts are indicated by colors on the upper left side of each subplot. Formation is found to occur with the central region of the protein and progress outwards. The N-terminal helix α0 is the last structure to form.

state. At time 0.05T, where T is simulation time, possibilities of forming native contacts began to increase in the $(\alpha\beta)_{2\text{-}5}$ region indicating the central region of the chain folds first. This small change in contact probability is apparent in the P(r) as a shift to a more distinct peak at smaller values of r. As time increases, 0.1T to 0.3T, more native contacts formed in the central region and expanded outwards from the $(\beta\alpha)_{3\text{-}4}$ region. The calculated P(r) at this time matches what

is seen in 150 μs by the continuous flow SAXS experiments with a peak at an r just over 20 Å with a tail extending out to longer distances due to the N and C termini not having formed their proper native contacts. The ensemble folds the $(\beta\alpha)_{2\text{-}8}$ region within 0.50 of the total simulation time. The contact map did not show significant changes from 0.50T to 1.00T, because a large portion of the trajectories were trapped "stuck" intermediate. From the contact maps, however, we can conclude the $\alpha_0\beta_1$ region is last to fold.

Simulations reveal frustration in folding

The formation of native structure was also measured by plotting the fractional native contacts within a βαβα module[107] against the total fractional native contacts (Figure 3.7). As was the case in other βα repeat proteins[108], back tracking or a loss in fractional native contacts within a segment of the protein chain can be seen throughout the folding reaction. Though the loss of native contacts in the simulations is counterproductive, if the loss is paired with a gain of contacts in another critical region, i.e. regions responsible for the transition state, it will drive the folding reaction forward. The first back tracking event takes place just prior to a Q_{total} of 0.4, when the first, third, and fourth βαβα modules lose contacts and the second module, $(\beta\alpha)_{3\text{-}4}$, quickly gains native contacts. This would argue that it is the second module that is important for the formation of the initial intermediate state. A second back tracking event follows in the first module at a Q_{total} of ~0.5 which allows the two C-terminal

modules to achieve an increase in native contacts. The two C-terminal modules then experience back-tracking at ~0.6 and 0.75 Q_{total} allowing the first module to catch up in the folding reaction. Throughout these events, the $(\beta\alpha)_{3\text{-}4}$ module experiences only minor back tracking events in comparison to the other three modules. This matches the experimental FRET dataset which saw the 112-140 FRET pair reach native like distances within 50 μs. Also, it appears that proper barrel closure is an important step for reaching the native state and that the process of barrel closure leads to frustration and back tracking events in the simulations.

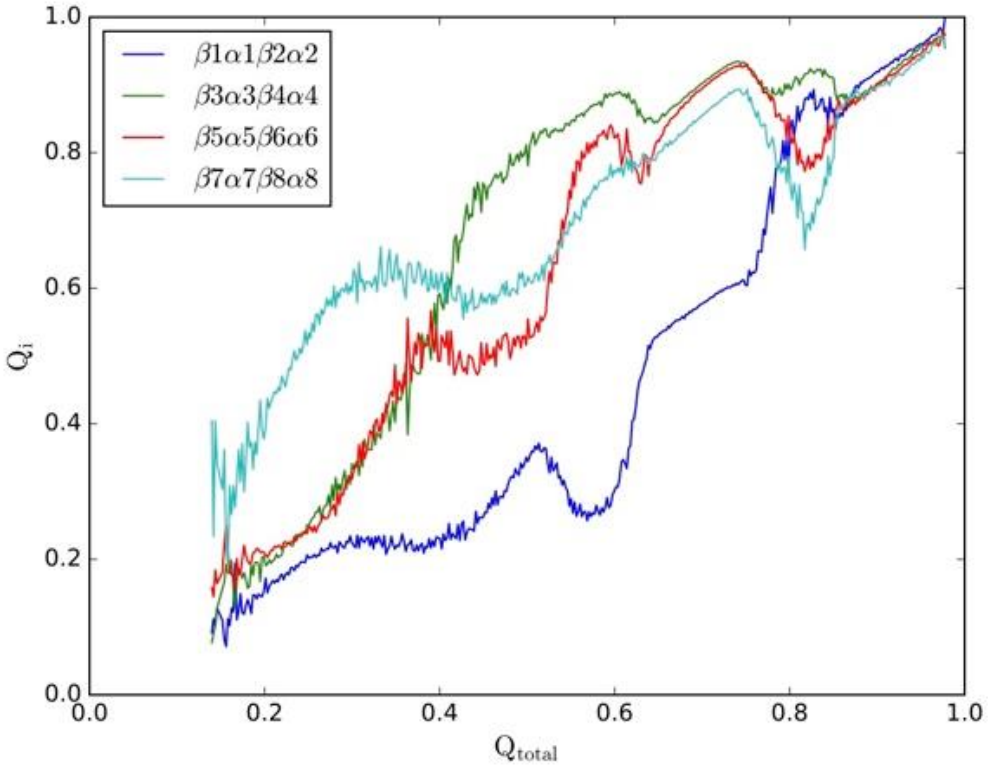


Figure 3. 7 Fractional native contacts of the four βα modules plotted vs. total fractional native contacts. Frustration events are seen in the modules by the decrease in fractional native contacts. Each frustration event is correlated with a strong folding event in a different module.

Folding mechanism inferred from the simulations

After examining each individual trajectory, the one hundred trajectories could be used to propose a general folding mechanism based upon the assembly of secondary structural units (Figure 3.8). All trajectories adopted a similar folding pathway in the early stage before entering an intermediate state, I_1, with a folded $(\beta\alpha)_{3\text{-}6}$ unit. After the maturation of $(\beta\alpha)_{3\text{-}6}$, there were two major folding pathways depending upon the folding order of $\alpha_0\beta_1$. A trapped

intermediate, I_{trap} would form if the $\alpha_0\beta_1$ did not fold earlier than the C-terminal module. If the C-terminal module, $(\beta\alpha)_{7\text{-}8}$, folds and docks on the intermediate I_1, the barrel prematurely closes and locks β_1 out of the barrel. Upon additional sampling, α_0 could occasionally dock on the bottom of the barrel, I_{trap2}, and as a result β_1 could slip though the helical shell and form proper contacts with β_2 and β_8. However, this was a rare event and most simulations would end with the β_1 locked out of the barrel. If the $\alpha_0\beta_1$ folds before the C-terminal module, the protein always achieved the native state. The $\alpha_0\beta_1$ could either dock on the I_1 intermediate followed by the C terminal module folding and closing the barrel or the $\alpha_0\beta_1$ could form native contacts with the C-terminal module forming a half barrel. The two halves then come together to form the native state. In either case, it is evident again that proper barrel closure is the critical step in folding and the competition between the N and C-termini cause frustration in the simulations.

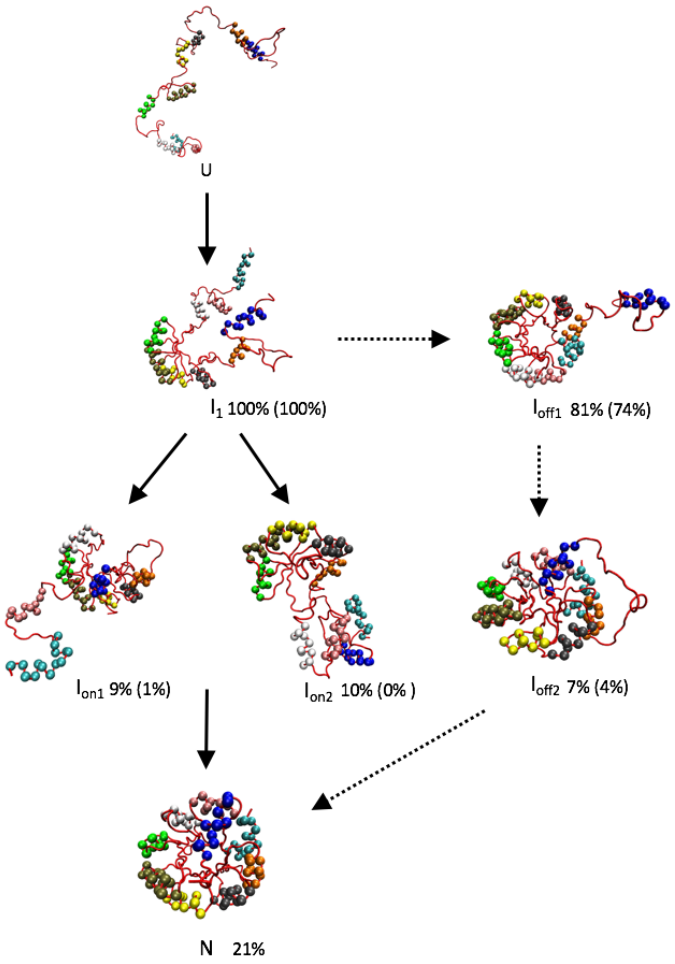


Figure 3. 8 Folding mechanism of sIGPS from Gō model simulations. From unfolded state (U) to native folded state (N), all trajectories first have a folded $(\alpha\beta)_4$ unit (I_1). The pathway then branches into two folding pathways: 1. folding pathway with less frustration through two possible intermediates I_{on1} or I_{on2} with $\alpha_0\beta_1$ folded earlier. 2. folding pathway by first forming a kinetic trap intermediate I_{off1} with unfolded $\alpha_0\beta_1$ and then folded to native state through an intermediate I_{off2}.

Discussion

The combined experimental and simulation results presented here provide insight into folding of one of the most common motifs in biology. Rapid mixing experiments using microfluidic devices revealed the formation of a structured intermediate within the dead time the experiments, 50 μs. The simulations meanwhile find extensive competition between βα modules at the N and C-termini that could be leading to the formation of the off-pathway intermediate.

Globally, the trSAXS data set revealed the rapid collapse of the backbone as seen by the large decrease in the Rg. Interestingly, the reaction of the unfolded state forming the I_{BP} state was too fast to be detected. In combination with the trFRET data we can infer that the reaction must be happening faster than ~15 μs, as a tail of an exponential would have been detectable if the reaction was any slower. Meanwhile, the dimensionless Kratky plot of the I_{BP} intermediate shows a decrease after a qRg of 1.5 indicating that the collapse is due to specific structure formation, as a collapsed unfolded state would have higher values at larger qRg, more like the unfolded state in 8 M urea (Figure 3.1). This intermediate is thermodynamically stable, as the Rg is insensitive out to 1.6 M urea. The P(r) would also have been a broader distribution with less of a distinct peak at low distances if the collapse was due to contraction of the unfolded state in a poor solvent. The trFRET data indicates

the structure formed in the I_{BP} intermediate is centered around the central $(\beta\alpha)_{3-4}$ segment of sIGPS with the near-native distances at 50 µs.

The $(\beta\alpha)_{3-4}$ segment is part of a large ILV cluster that spans both the $(\beta\alpha)_{3-4}$ and $(\beta\alpha)_{5-6}$ modules and contains an extremely hydrophobic stretch of 11 out of 22 residues in $(\beta\alpha)_4$. It is most likely that this very local hydrophobic core is driving the rapid exclusion of water and thus forming a strong hydrogen bond network allowing the burst phase intermediate to form in under 15 µs. The rapid formation of a segment in a TIM barrel and the role of an ILV cluster driving that formation has been seen before in the N-terminal half of the α subunit of tryptophan synthase. The $(\beta\alpha)_{2-4}$ region was found to be well-formed within 50 µs[46]. This region in αTS is dominated by a large ILV cluster made up of 31 ILV residues spanning the $(\beta\alpha)_{1-4}$ region. Alanine mutations in a subset of the 31 ILV residues eliminated the off-pathway intermediate indicating the important role of properly packing the ILV cluster to the formation of the burst phase intermediate[35].

It was surprising to not observe any kinetic phases within the experimental timeframe of 50µs to 5 milliseconds. Although small two state folding proteins have been found to have kinetic phases on the sub 50 µs timescale [100,109] it was surprising for a protein over 200 amino acids to have such a fast-kinetic phase and then no other phase until the 100s of milliseconds time frame [36]. The local-in-sequence, local-in-structure nature of TIM barrels creates a relatively low contact order for the protein. The low contact order in

combination with the high density of I, L, and V residues in the central two βα modules $(\beta\alpha)_{3\text{-}6}$ most likely is creating a nucleation site for the formation of the I_{BP} intermediate [21]. A folding rate of about 10 μs is expected for a protein of ~100 amino acids and an absolute contact order of 8.3[30]. Despite this rapid structure formation, something about the structure is non-native and must at least partially unfold.

The 63-238 FRET pair shows two distributions, one with a short distance and a second more expanded distance. Due to the time scales of the experiments, we are unable to detect any transitions between the states leading to the differing distance distributions. The transition would have needed to be faster than the approximately 15 ns to be detected on the time correlated single photon counting (TCSPC) timescale or slower than the 50 μs dead time of the fluorescent experiments for us to detect the transition between the populations. The presence of multiple species early in the folding reaction has been seen before in TIM barrels, although they were attributed to various proline isomerization states [39,46].

The decrease in dead time for kinetic experiments with the use of microfluidic mixers has revealed complex folding free energy landscapes for multiple proteins [99,109–111]. Markov state models (MSMs) built from all atom MD simulations are also able to capture complexity early in folding [109,112,113]. Although the MSMs built from the simulations generate many more states than the experiments can observe, the multiple states can be reconciled by the fact

that the experiments are blind to faster timescales than the rate limiting folding timescale [28]. Because of this, it is important when studying the early folding events to combine multiple techniques to probe the protein chain in various ways.

Unfortunately, the size of sIGPS precludes it from study with all-atom MD simulations. In this work, a coarse-grained native-centric Gō model was applied to simulate the folding of sIGPS. The Cα only model has previously shown robustness in studying protein folding while keeping a very low computational cost [108,114]. This model was designed based on the principle that the protein folding code is mainly embodied in side chain solvation interactions. One important feature of this model is that it explicitly treats the hydrophobic effect by adding a desolvation penalty to the side-chain interactions described by Lennard Jones potential[115]. When this desolvation penalty was removed or the solvent effect was turned off, the simulations could no longer capture the intermediates and multiple folding pathways. The simulations could detect several intermediates during the folding simulations of sIGPS and were able to confirm the FRET data and the HDX-MS data set[36] that structure first appears in the central region of the protein. Based upon the kinetic model of sIGPS (I_{BP} ⇌ U ⇌ I_A ⇌ I_B ⇌ N), the transition from I_A to I_B is the rate limiting step of folding and the simulations, from the calculated contact maps, would suggest that this step involves the closing of the barrel. The HDX data support this as the very little protection was seen in the N and C-termini for the I_A intermediate and was

followed by an increase in protection at these locations with the I_B intermediate[36].

Despite the simulations being native centric and unable to truly populate the off-pathway intermediate, they can provide insight into possible structural features of the intermediate. There were several backtracking events that took place during the simulations which in the case of sIGPS potentially relate to the off-pathway intermediate. The backtracking always involved the N terminal and the C terminal βα modules. This would suggest that forming the proper contacts between the two modules, and in particular between β1 and β8, is critical for proper barrel formation. With the nucleation of folding taking place in the central region, the low contact order, and high hydrophobicity of the region due to the ILV cluster, it is possible that the rapid structure formation causes the N and C termini to be out of register and unable to fold properly. Therefore, the chain must at least partially unfold to allow the folding reaction to continue. This unfolding event is not observable in the simulations due to the setup, however, it may be related to the I_{stuck} occasionally achieving the native state over the long timescale. ILV clusters leading to frustration in simulations and off-pathway intermediates appears to be common in βα repeat proteins[46,98].

The work presented here provides insight into some of the earliest folding events of sIGPS. Experimentally, interfacing microfluidic mixers with SAXS and FRET allowed us to observe the folding reaction on the microsecond timescale. Surprisingly, we did not observe any kinetic phases but could

determine structure first appears the central βα modules of the protein. The simulations independently confirm this with a computationally inexpensive model. They also allow gain insights into the cause of the off-pathway intermediate. ILV clusters appear to be a major driving force in early folding events for proteins and in the case of βα repeat motifs such as TIM barrels and flavodoxin folds may lead to the population of an off-pathway intermediate due to their low contact order.

Materials and Methods

Site-Directed Mutagenesis

The codon-optimized sIGPS gene was synthesized by Genscript in pUC 57 and recloned into a modified pGS-21a vector with an N-terminal His_6 tag and tobacco etch virus (TEV) protease site using EcoRV and *Bam*HI restriction sites. Cysteine and tryptophan mutations were made with mutagenic oligonucleotides purchased from Integrated DNA Technologies using the Stratagene QuikChange site-directed mutagenesis kit. The pGS-21a plasmid DNA was transformed into BL21 (DE3) pLysS cells for protein expression and purification.

Protein Expression and Purification

Cells were grown to an OD of 0.7-0.9 in Terrific Broth (Fisher); induced with 1 mM IPTG and harvested at 5000 rpm after 4 hours of induction. The cells were resuspended in buffer (2 ml per gram cell pellet) containing 25 mM Tris pH 8.0,

8 M urea, 10 mM imidazole, sonicated for about 5 minutes with 30 second pulses and centrifuged at 18,000 rpm for 1 hour to remove cell debris. The supernatant was filtered through a 0.22 μm filter and bound onto the His-60 Ni SuperflowTM resin (Clontech). The column was washed with 10 column volumes of buffer containing 25 mM Tris, 8 M urea and 20 mM imidazole and eluted in 10 column volumes of buffer containing 25 mM Tris 8 M urea and 300 mM imidazole. The protein was refolded by dialysis into buffer containing 25 mM Tris and 1 mM βME at pH 8.0. The His tag was cleaved overnight using 6X His tagged TEV protease at 10:1 molar ratio at 4 °C and reloaded onto His-60 Ni resin to trap the TEV protease. The protein was eluted and buffer exchanged into 10 mM KPi, 0.2 mM K_2EDTA, 1 mM βME at pH 7.8 and it was further purified by loading onto a DEAE Sepharose Fast Flow column. The column was washed with 2 column volumes of wash buffer and bound protein was eluted with ten column volumes of a linear gradient from 0-750 mM KCl. Eluted protein was pooled and run over a Hi-Prep Sephacryl S100 column to ensure purity. Purified protein fractions were pooled, concentrated and dialyzed against 10 mM KPi pH 7.8, 0.2 mM K_2EDTA and 1 mM βME.

Protein Labeling

Purified cysteine mutants were fully reduced with 1 mM TCEP for 1 hour at room temperature. Protein was then labeled with IAEDANS (ThermoFischer/Molecular Probes). A 10-fold mole excess of dye was added at

room temperature and allow to react for 2 hours. A second 10-fold mole excess of dye was added and allowed to react overnight at 4 ºC. Excess dye was removed by filtration and dialysis over the course of several dyes. Labeling efficiency was calculated using a Cary 100 UV/VIS spectrophotometer. Typical results were ~90% labeling efficiency.

Small angle x-ray scattering

Small-angle x-ray scattering measurements were performed at the BioCAT beamline at the Advanced Photon Source, Argonne, IL. Equilibrium SAXS measurements were performed by interfacing an autosampler running custom software to the standard glass sample capillary [98]. Kinetic experiments were performed by interfacing Harvard syringe pumps with single piece quartz mixers from Translume (Ann Arbor, MI)[99].

Time Correlated Single Photon Counting.

Details of the TCSPC apparatus equipped with a microsecond continuous- flow mixer have been described previously[46]. Flow to the microchannel mixer was provided by two syringe pumps (Isco) operating at a combined flow rate of 4 to 8 ml min^{-1}. Excitation at 293 nm with a repetition rate of 3.8 MHz was provided by the vertically polarized third harmonic of a Ti/sapphire laser. The mixer flow channel was aligned to yield excitation power that was uniform along the flow channel within 5%. This variation in excitation intensity was corrected by using

a standard N-Acetyl-L-tryptophanamide (NATA) as described. Separate instrument responses were recorded for each channel by recording a scattered light signal or by numerical deconvolution from the NATA decay curve.

MEM

Our 2D-MEM package, coded in LabVIEW 8.2 (National Instruments), incorporates procedures described by Kumar et. al. [106]. The distribution p(kd,kET) was represented as a grid of rates in logarithmic rate space. In the MEM optimization the 2D grid of amplitudes was collapsed into a 1D array. The same amplitudes were used for the donor and donor– acceptor data, with additional terms for labeling efficiency and for normalization of protein concentration. An instrument response for each decay trace was taken into account by aperiodic convolution with the decay rate matrix.

Gō model simulations

System preparation and model: The sIGPS was modeled on the crystal structure of a truncated version of IGPS (Protein Data Bank ID code: 2C3Z). The protein-folding simulations were performed using a Cα-only Gō-like model developed by Karanicolas and Brooks[115]. The native interactions were explicitly favored by adding a modified Lennard-Jones like potential with a desolvation penalty to enhance folding cooperativity. Simulations were performed using the CHARMM package. The equation of motion was propagated using Langevin

dynamics with a friction coefficient of 1.36ps^{-1} and a time step of 22 fs. The virtual bonds lengths between two nearest Cα s were fixed using the SHAKE algorithm. The folding temperature Tf was first estimated as a temperature corresponding to the peak in the heat capacity curve Cv(T) calculated from replica exchange simulations. Then 100 independent folding simulations were each performed for 2×10^8 dynamics steps at 0.72 Tf.

Chapter IV: Probing cores of stability in the higher energy states of sIGPS

This chapter is a collaborative effort with Dr. Francesca Massi's laboratory. I have worked with Dr. Asli Ertekin in set up of the NMR for the HDX experiments.

Introduction

Kinetic studies of protein folding reactions reveal the presence of transient intermediates along the folding pathway from the denatured, unfolded state to the native state. These transient intermediates are difficult to study due to their short life-times, marginal stabilities and dynamic properties [46,110,116]. However, studying the intermediates offers the possibility to better understand the relationship between sequence and folding by following structure formation along the free energy landscape. Equilibrium studies can be used to study the higher energy states[44,50], but requires the addition of a perturbant to the system, i.e., a chemical denaturant, high temperature or a change in the pH. The intermediate must then be more stable than the native state and the unfolded state under these conditions to significantly populate the state. One must also consider the perturbant and its role in the higher energy state as the protein is no longer in its biologically relevant solvent system.

Under conditions similar to the biologically relevant solvent system that favors the native state, higher energy states are populated, but their marginal stabilities are very low compared to the native state[56,117]. To detect the higher energy states, high resolution techniques are required. One such technique is to monitor the protection of main chain amide hydrogens from exchange with solvent deuterium. Protection of the amide hydrogen from exchange reflects the presence of significant hydrogen bonding with carbonyl oxygens due to secondary structure being present[118,119]. If monitored over time, the protection

patterns reveal the hydrogen bond networks present in the rare high energy states that are in equilibrium with the native state.

The folding mechanism of $(\beta\alpha)_8$ TIM barrels has been extensively studied due to the conserved, but complex mechanism[37,39,49]. The large sequence variation between barrels allows the opportunity to study the effect of sequence on the folding and the intermediate structures that form during the folding reaction. Clusters of branched aliphatic residues have been shown to acts as cores of stability for the higher energy states[32] in TIM barrels through the exclusion of water and stabilizing the hydrogen bond network of the high-energy states. The clusters of isoleucine, leucine, and valine residues change in size and location from barrel to barrel due to the changes in sequence of the barrels. Hydrogen exchange monitored by NMR has been used previously to map the hydrogen bond networks of both the native[37] and high-energy states[45] of TIM barrels. In both cases, the ILV clusters of the barrels acted as cores of stability.

The indole-3-glycerol phosphate synthase from *S. solfataricus* (sIGPS) presents a unique opportunity to study both the native state as well the higher energy states due its moderate kinetic stability[49]. The extrapolated unfolding of the native state under native conditions is expected to be several hours which allows monitoring of the exchange within the native manifold before the higher energy states have been sampled. However, by monitoring the protein for a week, the protection patterns of the higher energy states can be revealed,

providing insight to the structures of these states. Obtaining insights into the structures will allow for better insights into the progressive development that occurs during spontaneous folding.

Results

Thermodynamic and kinetic studies

Previous thermodynamic and kinetic studies were carried out at room temperature, however, the NMR studies required an increase in temperature to allow for better resolution of the peaks and increase the rate of the hydrogen exchange reaction. Therefore, the thermodynamics and kinetics at 35ºC must be studied prior to the NMR experiments. The equilibrium titration with guanidine hydrochloride at 35ºC is shown in Figure 4.1. The protein showes a denaturant independent native baseline and like the room temperature profile, fit to a 3-state model, $N \rightleftharpoons I \rightleftharpoons U$, with the equilibrium intermediate state assumed to be a mixture of both the I_A and I_B kinetic intermediates. The free energy differences for the $N \rightleftharpoons I$ and $I \rightleftharpoons U$ transitions were 6.6 ± 0.16 kcal mol^{-1} and 2.88 ± 0.11 kcal mol^{-1}, respectively. The model also provides the dependence of the free energy on denaturant concentration, the *m*-value, which is proportional to the amount of surface area buried during the transition[120]. The *m*-value for the N to I transition is 4.5 ± 0.12 kcal mol^{-1} $[M]^{-1}$, GdnHCl while the I to U transition is 1.16 ± 0.03 kcal mol^{-1} $[M]^{-1}$, GdnHCl. The fit to the three-state model also provides the Z-parameter, the normalized change in signal of the

intermediate relative to the unfolded state with $Z = (\theta_I - \theta_N) / (\theta_U - \theta_N)$. With a Z-value of 0.42, the equilibrium intermediate state has

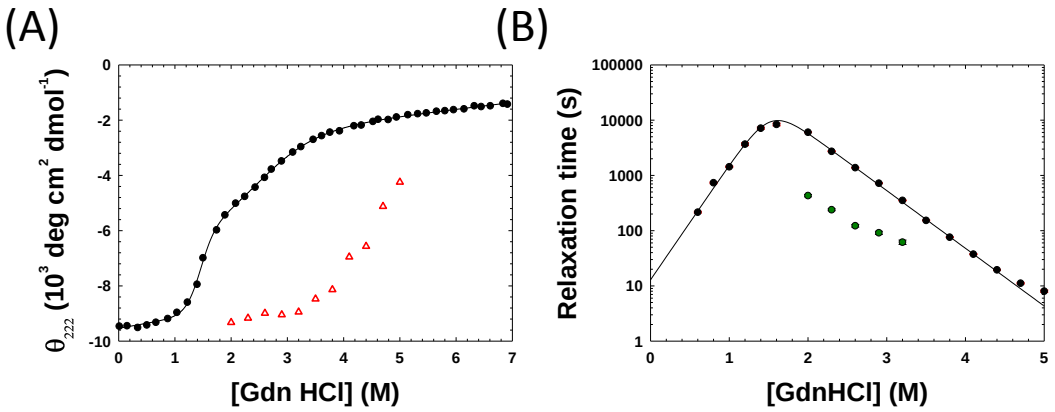


Figure 4. 1 (A) The GdnHCl induced equilibrium unfolding profile of sIGPS at 35°C. The solid continuous curve represents the fit of the data to a three-state model. The red triangles are the extrapolated initial signal from the manual mixing kinetic traces. (B) Semi-log plot of the time constants from kinetic refolding and unfolding experiments. The unfolding kinetics show multiple phases from 2M to 3.2M (green circles).

a relative MRE signal of approximately -6,600 deg cm^2 dmol^{-1}. This would indicate that majority of the CD signal is coming from the U to I transition.

The kinetics of both refolding and unfolding were monitored by manual mixing CD experiments. All the refolding kinetic jumps, from 6 M GdnHCl to various final denaturant concentrations, show a single exponential phase (Figure 4.1). From the chevron analysis, the extrapolated refolding time is 13 ± 4 seconds. The majority of native CD signal is recovered within the dead time of the experiment at 35°C, a feature that was also seen during room

temperature kinetic experiments[49]. The unfolding kinetic jumps show a more complex response with multiple phases present when jumps to final GdnHCl concentrations between 2 M and 3.2 M were performed. The faster kinetic phase is unable to be fit past 3.2M as the amplitude associated with the phase disappears due to a burst phase, with only approximately 30 percent of the signal from N to U being seen under strongly denaturing conditions. The faster unfolding phase extrapolates to approximately 6,000 seconds in the absence of denaturant, while the slower phase extrapolates to 750,000 seconds.

The two unfolding kinetic phases have different denaturant dependencies with the slow phase having an *m* value of -1.02 kcal mol^{-1} M^{-1} GdnHCl and the fast phase having an m value of -0.61 kcal mol^{-1} M^{-1} GdnHCl. The two different slopes for the kinetic phases may explain why we see a small roll over of the time constants for jumps to high denaturant. This would suggest that the rate limiting step in unfolding changes when jumps to the high denaturant occur. Altogether, the fast kinetic phase in unfolding might be reflective of the N to I_B reaction while the slow phase is the I_B to I_A reaction.

NMR Hydrogen Exchange

The protein backbone has been partially assigned providing an opportunity to explore the protection patterns for the higher energy states of sIGPS. Unfortunately, difficulties in the assignment process have precluded any residues in either β_5 or β_8 from being assigned. The exchange process was initiated by diluting a concentrated stock of protein into deuterated buffer.

The first spectrum was collected after 35 minutes and the reaction was followed for 8 days. The sample began to slowly aggregate after the 8 day period and caused the amide peaks to begin to shift limiting our analysis to the first 8 days of the reaction.

The exchange kinetics of the amide protons fall into 3 major categories, with representatives of the two slowest categories shown in figure 4.3. Class I protons were those that exchanged rapidly and were fully exchanged before the first spectrum or within the first few spectra acquisitions, making it difficult to get an accurate decay fit. The 49 protons in Class I were primarily located within the loops and helical shell of the protein. Class II protons were those that exchanged at an intermediate rate with rate constants, between 10^{-5} and 10^{-6} s^{-1}. The rate constants can be found in Table 4.1. The 43 protons in Class II were located at the ends of β-strands and the helical shell with the majority falling in α3, α4, α5. Class III protons were those that did not exchange more than 40% over the 8 day experiment which prohibited the exchange rate from being determined. The 24 amide protons of Class III fall in the β-barrel, α0, α4, and α5. The 3 different exchange classes have been mapped back to the crystal structure in Figure 4.2.

Table 4.1 List of Exchange Parameters

Residue	k_{obs} (s^{-1})	$\Delta G°_{HX}$ (kcal mol^{-1})
L38	2.71E-06	7.6
E39	9.95E-06	7.5
F40	3.14E-06	8.8
N41	0.000101	7.8
I45	0.000115	6.1
K71	0.000113	7.2
F72	2.29E-06	9.2
E74	9.07E-05	6.1
A77	5.24E-06	8.9
I82	2.73E-05	7.3
I113	9.8E-05	6.1
V114	3.74E-05	6.3
K115	0.00012	6.7
Q118	1.39E-05	4.0
I119	1.04E-05	7.7
D120	1.83E-06	10.6
D121	2.06E-05	7.4
Y123	1.12E-06	10.9
N124	3.15E-06	12.1
V130	1.55E-06	10.6
I133	2.02E-06	9.9
K135	0.000109	6.8
L137	1.88E-05	6.9
L142	4.92E-05	6.5
E143	3.1E-05	6.8
L145	2.01E-06	10.9
E147	2.01E-06	10.3
Y148	7.41E-06	8.0
S151	7.68E-05	8.1
Y152	1.22E-05	8.5
I168	3.66E-05	6.4
L170	1.39E-06	7.8
R171	0.000118	6.8
A174	9.79E-05	6.9
F176	3.24E-06	11.2
I179	3.31E-06	10.5
I198	7.08E-05	5.9
V206	4.42E-05	6.4
K207	0.000105	6.8
E210	4.6E-05	2.8

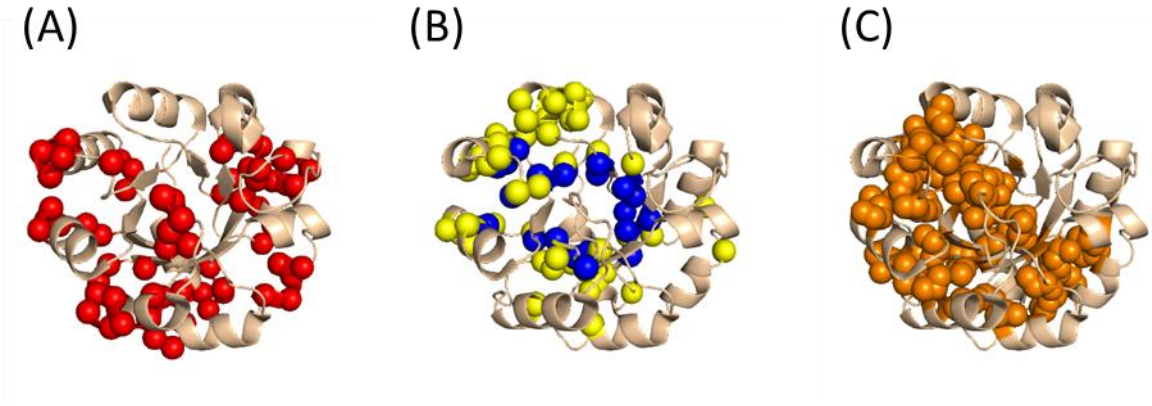


Figure 4. 2 (A) Ribbon diagram of sIGPS with the Class I protons that exchange rapidly marked by red spheres. (B) The Class III (blue spheres) protons map primarily to the β-barrel as well as α3, α4, α5. The class II protons that exchange over the 8-day experiment are highlighted in yellow. (C) The ILV cluster is highlighted in orange. Together, the Class II and Class III protection patterns map closely to the ILV cluster (panel B and C).

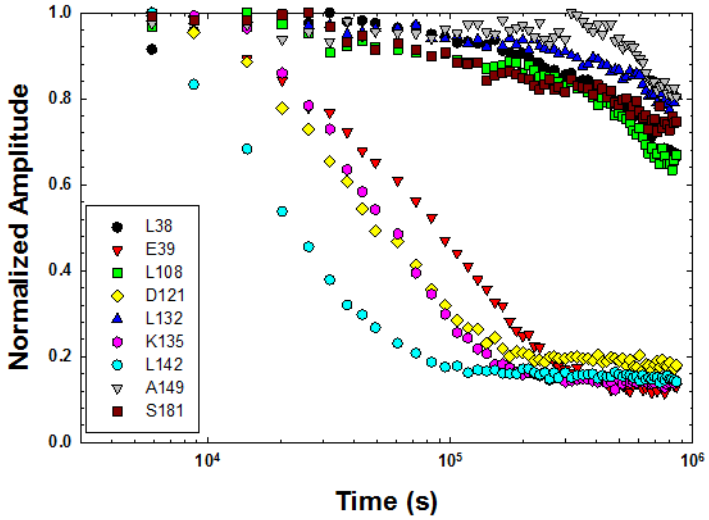


Figure 4. 3 Example amide proton decays for members of Class II and Class III. There was insufficient exchange in the Class III protons to allow the data to be fit due to less than 20 percent of the initial signal decaying. Class II residues typically exchange out by several days.

To assess the exchange in terms of either the thermodynamics or kinetics of the high energy states, the exchange mechanism (EX1 or EX2) for the Class II amide protons was determined by looking at the effect of pH on the exchange rates[117]. The EX1 mechanism is dependent on the exchange reaction being much faster than the refolding from the exchange competent state back to the incompetent state. A change in pH would therefore have no effect on the rate of exchange as the limiting reaction would be the unfolding event. The EX2 mechanism is limited by the relative stability between the non-exchange competent and the exchange competent state. Therefore, it will have an off-shift on the y-axis of the log-log plot due to the acceleration of exchange with increasing pH[56]. When the stability differences of sIGPS at pH 7.2 and 7.8 are accounted for the Class II amide protons fall on the EX2 limit (Figure 4.4).

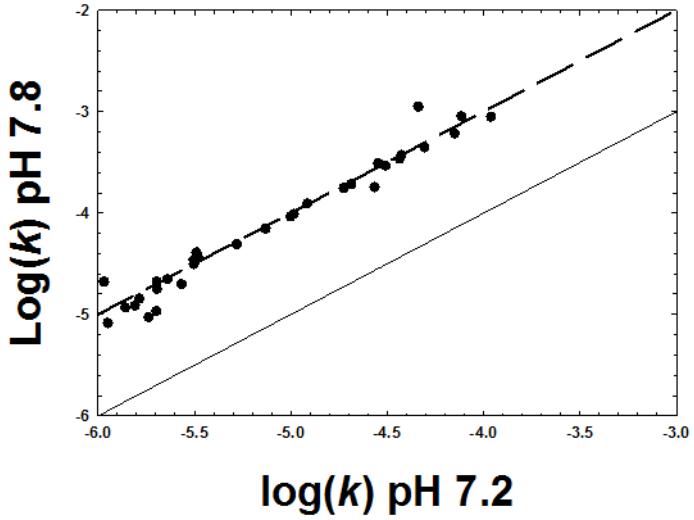


Figure 4. 4 Comparison of the exchange rates at pH 7.2 and 7.8 on a log-log plot for Class II protons. The solid line represents the EX1 limit and the dashed line represents the EX2 limit. The Class II protons fall on the EX2 line, therefore $k_{cl} > k_{int}$ and $\Delta G_{HX} = -RT \ln(k_{obs}/k_{int})$

Since the exchange is occurring via the EX2 mechanism, the relative stability between the exchange competent and noncompetent state can be extracted through the relationship $\Delta G^{o}_{HX} = -RT \ln(k_{obs}/k_{int})$. The calculated values for the Class II protons are found in Table 4.1. From the chevron, the fast unfolding phase of N to IB would be ~6,000 seconds in the absence of denaturant. For exchange to be taking place on an EX2 basis, the refolding reaction from I_B to N must be faster than 80 milliseconds, based on the intrinsic rate of exchange as calculated from Sphere. The roughly 10^5 difference in the refolding and unfolding rates would account for approximately 6 kcal mol^{-1} of stability, which closely matches the majority of the calculated ΔG^{o}_{HX} values.

Due to the experiment being limited to 8 days and the extrapolated unfolding of the I_B to I_A also being 8 days, we are unable to determine whether the Class III amide protons are exchanging via EX1 or EX2. The Class III protons were primarily found to be in the β-barrel. Although we have no assignments from either β5 or β8 we would expect them to be highly protected as well because of the hydrogen bond network holding the barrel together. The network has hydrogen bonds from β_4, β_6, β_7, and β_1 that are involved with β_5 and β_8 and it would be expected that residues involved in hydrogen bonding would have similar exchange properties. This would indicate that the barrel structure is intact for the IB intermediate (Figure 4.5).

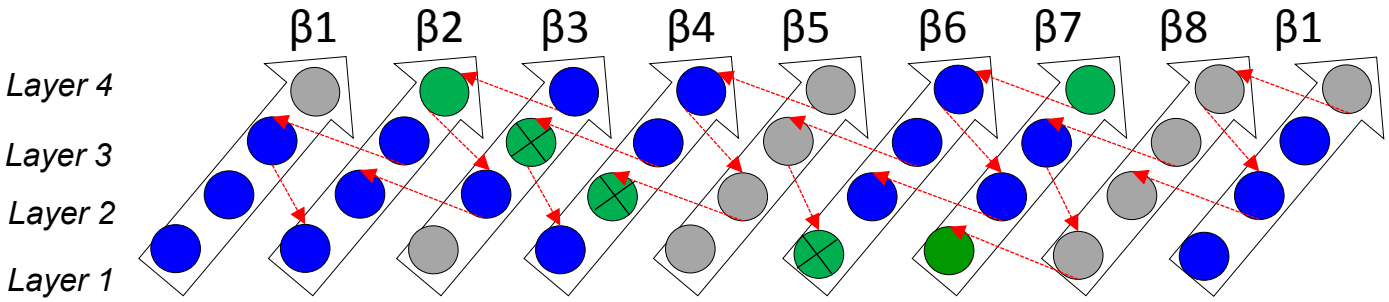


Figure 4.5 The protection classes from the fits are mapped to the β-barrel structure (Green class II and blue Class III). The hydrogen bond network of the barrel is shown by red arrows. The majority of the barrel is part of Class III. The class II residues in β3,4,6 are part of a group that exchanges at a much higher energy state (ΔG ~11 kcal mol^{-1})

Discussion

By monitoring the exchange properties of amide protons with solvent deuterium, protection patterns have mapped structural features of the high-energy states of globular proteins that are in equilibrium with the native state[45,56,118]. Application of the technique to sIGPS has revealed the hydrogen bond network that stabilizes the higher energy states.

The 49 Class I amide protons that exchange rapidly were found to reside in loops and the helical shell. Apart from residues I99 through I105 of α_2 and Q194 through I201 of α_6, the Class I residues found in helices were located at the N-terminus of the helix. The lack of protection at the N-terminal end of a helix is common and most likely reflects the lack of intrahelical N-H donors for the first four residues of the helix[121]. While the loss of protection at the C-terminal end of α_2 and α_6 might be due to the lack of intrahelical hydrogen bond

acceptors, another possibility is that the two helices are the terminal helix in a βαβα module. It has been thought that the modules have arisen from gene duplications which eventually led to the barrel structure[122] with the initial βαβ serving as the minimum stability unit[107,123]. If the last helix in the module acts as a linker between modules, it is possible that the fast exchange properties of α_2 and α_6 is due to the helix acting as a linker connecting one module to the next.

The Class II amide protons display a range in exchange rates on the order of 10^{-5} to 10^{-6} s^{-1} (Table 4.1). The pH dependence test indicates an EX2 mechanism. Because the exchange is dependent on the population of the higher energy state, the rate of exchange was used to calculate the $\Delta G^{o}HX$ (Table 4.1). The stability values for the higher energy state were found to average ~7 kcal mol^{-1}. Based on the thermodynamics and kinetics of sIGPS at 35ºC, it is believed that exchange takes place through the I_B intermediate. Due to range of narrow range of stabilities, it can be inferred that there is an ensemble of microstates that are loosely coupled that together make up the thermodynamic I_B intermediate. The amide hydrogens associated with these states in Class II are primarily located in α_0, α_1, α_3, α_4, α_5, α_6, α_7, and α_8. This indicates that the helical shell of sIGPS is loosely packed in the I_B intermediate. Within this class, over half are in α_3, α_4, and α_5. With the underlying β-strands strongly protected as well, it suggests that this region of the protein is acting as the core of stability for the intermediate state. With α_1, α_2, α_6, α_7, and α_8 being

loosely packed, it may explain the low m-value in unfolding of the N to I_B step as very little surface area is exposed during the reaction[120].

The exchange rates for the Class III protons could not be quantified due to the limited exchange that took place during the experiment. When mapped back to the crystal structure, Class III residues fall within the β-barrel, α_3, α_4, and α_5. Due to limitations of the experiment, it cannot be determined if the exchange is taking place out of the I_A or the I_{BP} intermediate. The TIM barrel is thought to contain 4 layers[1], with each layer containing the side chains from either the odd or the even strands. When the protection is mapped back to the layers (figure 4.5), the majority of protons fall into Class III. The slow exchange of these residues would suggest that the barrel is fully intact in the I_B intermediate. The rate limiting step in folding for sIGPS is thought to be the I_A to I_B reaction. If barrel formation is happening at the transition state during this step, it would explain the slow exchange of the amide protons of the β-barrel as the timeframe of the experiment does not allow us to monitor over the barrier between I_B and I_A. The barrel being properly formed in the first higher energy state above native has previously been seen in the α-subunit of Trp synthase[45].

The BASiC hypothesis states that the native state as well as higher energy states are stabilized by the formation of a network of hydrogen bonds and van der Waal interactions due to the exclusion of water from large clusters of branched aliphatic residues [32]. Previous hydrogen exchange experiments have revealed the role these clusters play in acting as stability cores for other

TIM barrels [36,45,50]. The calculated ILV cluster map of sIGPS reveals one large cluster that runs from α_3 through α_8, including the underlying β-barrel (Figure 4.2), with a total of 46 ILV residues being a part of the cluster (http://biotools.umassmed.edu/ccss/ccssv2/basic.cgi).

It is known that the ILV cluster prediction algorithm will over predict the amide protons protected against exchange[32], however, the algorithm is useful for providing insight into the role of the amino acid sequence in the higher energy states. Although the large ILV cluster of sIGPS includes 46 residues, there is a varying degree of side chain contacts and burial within the cluster. When looking at the sequence and the cluster map, a stretch of 11 out of 22 residues being I, L, or V within the $(\beta\alpha)_4$ region becomes apparent. Leucine 132, centrally located on β4 with its side chain pointing out towards the helical shell, makes contacts with 5 other ILV side chains. Over the 8-day experiment, the amide proton of L132 only exchanges approximately 20% and is a member of Class III. The surrounding protons in three-dimensional space, including non-ILV residues, are also strongly protected and belong to Class II or Class III. The high density of the ILV network in this region is most likely leading to the strong hydrogen bond network that is providing stability in the I_B and the I_A/I_{BP} intermediates.

The only helices to contain Class III protons are α_3, α_4, and α_5. Based upon the ILV cluster map, this is not surprising as the region is tightly packed with a high density of ILVs. The microsecond kinetics (Chapter III), revealed

native-like distances between α_3 and α_4 within 50 µs. Together, this would support the hypothesis that it is the ILV network within the $(\beta\alpha)_{3\text{-}4}$ module that is acting as the stability core for the early folding intermediates.

The native state exchange study has revealed the hydrogen bond networks of the partially folded high energy states that are in equilibrium with the native state. The experiment provided insights into the energetics of some of the higher energy states with results matching well with previous studies. Structurally, it was determined that barrel formation is complete by the I_B intermediate. A subgroup of ILV residues within the ILV cluster of sIGPS show persistent resistance to exchange over the 8-day experiment. The burial of the hydrophobic side chains seems to be driving early folding and then providing sufficient stability to allow the protein to development structure in a multi-step manner.

Methods

Protein Purification

WT sIGPS was purified as previously described (Chapter III). For the NMR samples, uniformly ^{15}N-labeled protein was obtained by growing the *E. coli* in M9 media containing ^{15}N-ammonium chloride. Due to the slow growth rate in M9, after induction with IPTG, the growth temperature was dropped to 30 ºC and the cells were harvested after 12 hours of induction.

Thermodynamic and Kinetic Studies

The thermodynamic properties at 35 ºC were determined by guanidine hydrochloride titrations on a Jasco J-810 spectropolarimeter. Samples at varying guanidine hydrochloride concentrations were prepared using a Hamilton 540B automatic titrator and were incubated overnight at 35 °C for complete equilibration. Data were collected using a 2 mm pathlength quartz cuvette and a 2.5 nm bandwidth. The spectra were recorded at every 1 nm in the wavelength range from 215 nm to 260 nm with a scan speed of 50 nm min-1 and an eight second averaging time. The denaturant dependence of the ellipiticities was fit to a three-state model using Savuka, an in-house nonlinear least squares program, and assuming a linear dependence of the free energy of unfolding on the denaturant concentration. These fits provided the free energy differences between the three thermodynamic states, the denaturant dependences of these free energy differences and the Z parameter required to estimate the ellipticity of the intermediate.

The manual-mixing refolding kinetic jumps began in 6 M GdnHCl and ended between 0.6 M and 2 M while the unfolding jumps started in the absence of denaturant and ended between 4.0 M to 6.0 M GdnHCl. The final protein concentration ranged from 3-5 μM. Data were collected at 222 nm and at 35 °C in a 1 cm pathlength cuvette. The relaxation times were obtained by fitting the kinetic traces to a single or double exponential function in Savuka.

Exchange Studies

NMR spectra were collected on a Varian 600 MHz spectrometer. The hydrogen exchange reaction was started by dilution of concentrated protein with deuterated buffer. The sample was filtered and transferred to a NMR tube. After tuning and shimming the magnet (about 10 minutes), TROSY 2D ^{15}N-^{1}H spectra were collected periodically over 8 days. The peak intensities were than fit to an exponential decay to obtain k_{obs}. ΔG^{o}_{HX} values were calculated from $\Delta G^{o}_{HX} = -RT\ln(k_{obs}/k_{int})$ with k_{int} obtained from the SPHERE program. http://landing.foxchase.org/research/labs/roder/sphere/

Chapter V: Conclusion and Future Directions

Summary

As a protein chain spontaneously folds, many different interactions between the backbone and side chain atoms help direct the chain to form the proper contacts required for the native state. Due to the diversity of three-dimensional structures, it has been impossible, as Kendrew noted, to determine how a protein folds from the structure alone. Over the years there has been a great interest in the role that hydrophobic residues play in folding of proteins[20,35,110]. The ability to exclude water from large clusters of branched aliphatic residues allows for a local drop in the dielectric constant, thus allowing for tighter hydrogen bond formation of the backbone and tighter packing of side chains[34]. It has been hypothesized that the clusters of branched aliphatic residues provide cores of stability to the higher energy states of proteins[32].

To investigate the role of large clusters of branched aliphatic residues on the folding free energy landscapes of proteins, the $(\beta\alpha)_8$ TIM barrel proteins was chosen as a model system. The TIM barrel family is one of the most common motifs in biology and has a complex folding free energy landscape that includes an initial off-pathway intermediate as well as two on-pathway intermediates before the native state is achieved[39,49]. The ability to populate various high-energy states allows us to test the importance of ILV clusters at multiple stages during the folding reaction.

In Chapter II, molecular dynamic simulations and experimental approaches demonstrated that properly dewetting the hydrophobic ILV clusters in the alpha subunit of Trp synthase (αTS) is important during multiple stages of folding. The simulations showed strong dewetting transitions in the large N-terminal cluster of αTS. Replacement of ILV residues with alanine, a less hydrophobic and smaller side chain, weakened the dewetting transition. A previous study showed that the alanine mutations caused a loss in stability for all states, including potentially destabilizing the off-pathway intermediate to the extent that it no longer was significantly populated during folding[35]. Replacement of leucine with asparagine, which is nearly isosteric, disrupts the dewetting transition and promotes water being present in the cavity. Experimental results examining the stability, secondary structure, and compactness of the intermediates showed dramatic decreases in stability, secondary and tertiary structure. Comparison of the polar mutations within a second ILV cluster, located at the C-terminus, proved to be less responsive to the mutation in the unfolded to intermediate state reaction. The results highlight the different roles ILV clusters can have on the folding free energy landscape of TIM barrels.

In Chapter III, to gain insights into the early misfolding reaction in TIM barrels, we expanded our studies to the indole-3-glycerol phosphate synthase (sIGPS). With the use of custom, single-piece microfluidic chips we were able to study sub-millisecond folding reactions by small angle x-ray scattering and

fluorescence. Small angle x-ray scattering showed a rapid collapse of the chain in 150 μs to a compact, but not fully globular intermediate that was resistant to urea denaturation. To complement the SAXS dataset, pair-wise distance measurements were performed using time-resolved FRET on donor/acceptor pairs. Within the deadtime of the experiment, 50 μs, the $(\beta\alpha)_{3\text{-}4}$ region was found to have near native-like distances. The high density of ILV residues (11 out of 22) are thought to be driving this early reaction. Surprisingly, the N- and C-termini pair had multiple species present and may reflect both the I_{BP} and I_A intermediate being populated.

To provide additional insights, coarse grain Gō model simulations were performed to probe the folding landscape. The simulations revealed progressive contraction of the chain as the protein samples partially folded states along the folding reaction coordinate, with folding initiating in the same $(\beta\alpha)_{3\text{-}4}$ region of the ILV cluster. These simulations also show significant frustration in the N- and C-termini throughout folding. While frustration has been seen in Gō model simulations[98] it is believed the frustration seen here is likely related to the rate-limiting closing of the barrel during folding.

To probe site specifically the higher energy states of sIGPS, native state hydrogen exchange was employed in Chapter IV. Protection of amide hydrogens from deuterated solvent revealed the hydrogen bond network present in the high-energy states that are in equilibrium with the native state. Due to the limited time scales that we could monitor the exchange reaction

over, only the energetics of the I_B intermediate could be determined. The protection patterns revealed the barrel is well formed in the I_B state with the helices being dynamic. Strong protection in α_3, α_4, and α_5 and the underlying β-strands matched well with the ILV cluster in that region and supports the hypothesis that it is the ILV cluster that acts as a stability core for the higher energy states.

Discussion

One of the major outstanding questions in the protein folding field is, what is happening at multiple locations in the chain during the early folding event. One method to study the unfolded ensemble and the early events is through simulation. The rapid growth in computational power has led to the ability to sample longer timescales during folding simulations[124]. However, the ability to sample longer timescales has generated a problem: a large flux of data with no way to systematically analyze the data. Markov State Models (MSMs) have been applied to folding trajectories to analyze the data in a systematic and quantitative manner [28,109,112,113]. The MSMs that have resulted from various folding studies show high energy states that are made up of many microstates.

Experimentally, accessing the earliest folding events and obtaining multi-dimensional information has been difficult because of the timescales associated with the folding events[99,110,111,125]. This is evident from the results in Chapter III where even with a dead time of 50 μs there was a burst phase and the reaction

of U to I_{BP} for sIGPS was missed. While it may be possible to develop mixers with shorter dead times, another possibility is to move to single molecule experiments. To perform correlation spectroscopy experiments, more photostable dyes must be used than those used in this dissertation. However, due to limitations in chemistry, typically only cysteine chemistry has been used to label proteins which creates issues in site-specifically labeling proteins with multiple dyes.

Recently, oxaziridine derivatives have been shown to site specifically label the sulfur of methionine, even in the presence of cysteine[126]. Through modification of the oxaziridine with click chemistry, one can site specifically label proteins with dyes and other small ligands at methionine residues. Following the protocol from Lin et. al., oxaziridine probes have been synthesized[126] and sIGPS has successfully been labeled with ALEXA-594. With the ability to now label site specifically multiple locations within one protein chain, more robust distance distributions can be measured from single molecule experiments with the potentially to perform 3-color experiments as well.

For the TIM barrel family, one of the outstanding questions in the early folding reaction is why the protein misfolds initially. Despite the large sequence variation seen in the family, the barrels studied to date all display the formation of the off-pathway intermediate on the microsecond timescale. The low contact order of the barrel architecture, and the high number of aliphatic residues are

most likely responsible for the misfolding taking place. There is still the question however about what in the protein is misfolded. With the potential to form helices on the nanosecond timescale[121], and the high content of hydrophobic residues, one potential method for burying the hydrophobes found in the strands quickly is to initially form helices[45].

Interestingly, when secondary structure predictors, such as JPred (http://www.compbio.dundee.ac.uk/jpred/), are used on the sequence for sIGPS, the β_4 strand is predicted to form a helix instead of a strand. The potential use of helices in the burst phase intermediate has been proposed as in αTS based in NMR hydrogen exchange data[45]. Experiments using the new methionine labeling strategy will allow for us to look for the formation of the helix during the early folding events of sIGPS. On a broader scale, a more informatics approach looking at location of ILV clusters within TIM barrels and secondary structure propensity will elucidate the potential role of burying hydrophobic residues through the rapid formation of helices resulting in improperly folded intermediates that must partially unfold to reach their native state.

Perspective

A combined experimental and computational approach probing the folding landscape of one of the most common protein folds in biology, the $(\beta\alpha)_8$ TIM barrel, has revealed shared and distinct features within the protein family

that enable detailed insights into the role clusters of branched aliphatic side chains play in the energy landscapes. The results presented suggest that certain clusters of ILV residues are critical for forming stable cores for these high-energy states. However, these clusters are known to change size and move within a protein family, indicating a general role for the sequence of a protein in determining the energy landscape of a protein. Understanding the role sequence in defining these clusters in three-dimensional space will help develop more efficient design principles for the many new classes of do novo designed proteins, including *de novo* TIM barrels. High energy states of proteins are also of great interest in the biomedical field, as aggregation of proteins can lead to serious medical disorders. By understanding the forces that stabilize high energy states, the potential to develop effective therapeutics against the various diseases.

References

1. Thornton, J. M., Orengo, C. A., Todd, a E. & Pearl, F. M. Protein folds, functions and evolution. *J. Mol. Biol.* **293,** 333–42 (1999).

2. Korn, E. D. Actin polymerization and its regulation by proteins from nonmuscle cells. *Physiol. Rev.* **62,** 672–737 (1982).

3. Hamm, H. E. The many faces of G protein signaling. *J. Biol. Chem.* **273,** 669–72 (1998).

4. Hartl, F. U. & Hayer-Hartl, M. Converging concepts of protein folding in vitro and in vivo. *Nat. Struct. Mol. Biol.* **16,** 574–581 (2009).

5. Kayatekin, C., Zitzewitz, J. A. & Matthews, C. R. Disulfide-Reduced ALS Variants of Cu, Zn Superoxide Dismutase Exhibit Increased Populations of Unfolded Species. *J. Mol. Biol.* **398,** 320–331 (2010).

6. Selkoe, D. J. Cell biology of protein misfolding: The examples of Alzheimer's and Parkinson's diseases. *Nat. Cell Biol.* **6,** 1054–1061 (2004).

7. Leandro, J., Simonsen, N., Saraste, J., Leandro, P. & Flatmark, T. Phenylketonuria as a protein misfolding disease: The mutation pG46S in phenylalanine hydroxylase promotes self-association and fibril formation. *Biochim. Biophys. Acta - Mol. Basis Dis.* **1812,** 106–120 (2011).

8. Sato, S., Ward, C. L., Krouse, M. E., Wine, J. J. & Kopito, R. R. Glycerol reverses the misfolding phenotype of the most common cystic fibrosis mutation. *J. Biol. Chem.* **271,** 635–8 (1996).

9. KENDREW, J. C. *et al.* A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181,** 662–666 (1958).

10. SELA, M., WHITE, F. H. & ANFINSEN, C. B. Reductive Cleavage of Disulfide Bridges in Ribonuclease. *Science (80-.).* **125,** 691–692 (1957).

11. Anfinsen, C. B. & Haber, E. Studies on the Reduction and Re-formation of Protein Disulfide Bonds. *J. OP Biol. Chem.* **236,** (1961).

12. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181,** 223–30 (1973).

13. Levinthal, C. ARE THERE PATHWAYS FOR PROTEIN FOLDING ? *Extr. du J. Chim. Phys.* **65,** (1968).

14. Mirsky, A. E. & Pauling, L. On the Structure of Native, Denatured, and

Coagulated Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **22,** 439–47 (1936).

15. Bolen, D. W. & Rose, G. D. Structure and Energetics of the Hydrogen-Bonded Backbone in Protein Folding. **77,** (2008).

16. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The Protein Folding Problem. *Annu. Rev. Biophys.* **37,** 289–316 (2008).

17. Yang, J. S., Chen, W. W., Skolnick, J. & Shakhnovich, E. I. All-Atom Ab Initio Folding of a Diverse Set of Proteins. *Structure* **15,** 53–63 (2007).

18. Kauzmann, W. Structural factors in protein denaturation. *J. Cell. Comp. Physiol.* **47,** 113–131 (1956).

19. KAUZMANN, W. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14,** 1–63 (1959).

20. Tanford, C. **Contribution of Hydrophobic Interactions to the Stability of the Globular Conformation of Proteins**. *J. Am. Chem. Soc.* **84,** 4240–4247 (1962).

21. Fersht, A. R. Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7,** 3–9 (1997).

22. Kim, P. S. & Baldwin, R. L. Specific Intermediates in the Folding Reactions of Small Proteins and the Mechanism of Protein Folding. *Annu. Rev. Biochem.* **51,** 459–489 (1982).

23. Karplus, M. & Weaver, D. L. Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci.* **3,** 650–68 (1994).

24. Kuwajima, K. & Sugai, S. Equilibrium and kinetics of the thermal unfolding of alpha-lactalbumin. The relation to its folding mechanism. *Biophys. Chem.* **8,** 247–54 (1978).

25. Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nat. Struct. Mol. Biol.* **4,** 10–19 (1997).

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

27. Chan, H. S. & Dill, K. A. Protein folding in the landscape perspective: Chevron plots and non-arrhenius kinetics. *Proteins Struct. Funct. Genet.* **30,** 2–33 (1998).

28. Lane, T. J. & Pande, V. S. A Simple Model Predicts Experimental Folding Rates and a Hub-Like Topology. *J. Phys. Chem. B* **116,** 6764–6774 (2012).

29. Gromiha, M. M. Importance of Native-State Topology for Determining the

Folding Rate of Two-State Proteins †. *J. Chem. Inf. Comput. Sci.* **43,** 1481–1485 (2003).

30. Ivankov, D. N. *et al.* Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* **12,** 2057–62 (2003).

31. Radzicka, A. & Wolfenden, R. Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. *Biochemistry* **27,** 1664–1670 (1988).

32. Kathuria, S. V., Chan, Y. H., Nobrega, R. P., Özen, A. & Matthews, C. R. Clusters of isoleucine, leucine, and valine side chains define cores of stability in high-energy states of globular proteins: Sequence determinants of structure and stability. *Protein Sci.* **25,** 662–675 (2016).

33. Simonson, T. & Perahia, D. Internal and interfacial dielectric properties of cytochrome c from molecular dynamics in aqueous solution. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 1082–6 (1995).

34. Schell, D., Tsai, J., Scholtz, J. M. & Pace, C. N. Hydrogen bonding increases packing density in the protein interior. *Proteins Struct. Funct. Bioinforma.* **63,** 278–282 (2005).

35. Wu, Y., Vadrevu, R., Kathuria, S., Yang, X. & Matthews, C. R. A Tightly Packed Hydrophobic Cluster Directs the Formation of an Off-pathway Sub-millisecond Folding Intermediate in the α Subunit of Tryptophan Synthase, a TIM Barrel Protein. *J. Mol. Biol.* **366,** 1624–1638 (2007).

36. Gu, Z., Rao, M. K., Forsyth, W. R., Finke, J. M. & Matthews, C. R. Structural analysis of kinetic folding intermediates for a TIM barrel protein, indole-3-glycerol phosphate synthase, by hydrogen exchange mass spectrometry and Go model simulation. *J Mol Biol* **374,** 528–546 (2007).

37. Gangadhara, B. N., Laine, J. M., Kathuria, S. V, Massi, F. & Matthews, C. R. Clusters of branched aliphatic side chains serve as cores of stability in the native state of the HisF TIM barrel protein. *J. Mol. Biol.* **425,** 1065–81 (2013).

38. Urfer, R. & Kirschner, K. The importance of surface loops for stabilizing an eightfold βα barrel protein. *Protein Sci.* **1,** 31–45 (2008).

39. Osman Bilsel, Jill A. Zitzewitz, Katherine E. Bowers, and & Matthews*, C. R. Folding Mechanism of the α-Subunit of Tryptophan Synthase, an α/β Barrel Protein: Global Analysis Highlights the Interconversion of Multiple Native, Intermediate, and Unfolded Forms through Parallel

Channels†. (1998). doi:10.1021/BI982365Q

40. Shakhnovich, E. I. & Finkelstein, A. V. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers* **28,** 1667–1680 (1989).

41. Carstensen, L. *et al.* Conservation of the Folding Mechanism between Designed Primordial $(\beta\alpha)_8$ -Barrel Proteins and Their Modern Descendant. *J. Am. Chem. Soc.* **134,** 12786–12791 (2012).

42. Wu, Y., Vadrevu, R., Kathuria, S., Yang, X. & Matthews, C. R. A tightly packed hydrophobic cluster directs the formation of an off-pathway sub-millisecond folding intermediate in the alpha subunit of tryptophan synthase, a TIM barrel protein. *J Mol Biol* **366,** 1624–1638 (2007).

43. Dunn, M. F., Niks, D., Ngo, H., Barends, T. R. M. & Schlichting, I. Tryptophan synthase: the workings of a channeling nanomachine. *Trends Biochem. Sci.* **33,** 254–264 (2008).

44. Gualfetti, P. J., Bilsel, O. & Matthews, C. R. The progressive development of structure and stability during the equilibrium folding of the alpha subunit of tryptophan synthase from Escherichia coli. *Protein Sci* **8,** 1623–1635 (1999).

45. Vadrevu, R., Wu, Y. & Matthews, C. R. NMR analysis of partially folded states and persistent structure in the alpha subunit of tryptophan synthase: implications for the equilibrium folding mechanism of a 29-kDa TIM barrel protein. *J Mol Biol* **377,** 294–306 (2008).

46. Wu, Y., Kondrashkina, E., Kayatekin, C., Matthews, C. R. & Bilsel, O. Microsecond acquisition of heterogeneous structure in the folding of a TIM barrel protein. *Proc Natl Acad Sci U S A* **105,** 13367–13372 (2008).

47. Wu, Y. & Matthews, C. R. Proline replacements and the simplification of the complex, parallel channel folding mechanism for the alpha subunit of Trp synthase, a TIM barrel protein. *J. Mol. Biol.* **330,** 1131–1144 (2003).

48. Schneider, B. *et al.* Role of the N-Terminal Extension of the $(\beta\alpha)_8$ -Barrel Enzyme Indole-3-glycerol Phosphate Synthase for Its Fold, Stability, and Catalytic Activity †, ‡. *Biochemistry* **44,** 16405–16412 (2005).

49. Forsyth, W. R. & Matthews, C. R. Folding Mechanism of Indole-3-glycerol Phosphate Synthase from Sulfolobus solfataricus: A Test of the Conservation of Folding Mechanisms Hypothesis in (βα)8 Barrels. *J. Mol. Biol.* **320,** 1119–1133 (2002).

50. Gu, Z., Zitzewitz, J. A. & Matthews, C. R. Mapping the structure of folding cores in TIM barrel proteins by hydrogen exchange mass spectrometry:

the roles of motif and sequence for the indole-3-glycerol phosphate synthase from Sulfolobus solfataricus. *J Mol Biol* **368,** 582–594 (2007).

51. Bartlett, A. I. & Radford, S. E. Desolvation and Development of Specific Hydrophobic Core Packing during Im7 Folding. *J. Mol. Biol.* **396,** 1329–1345 (2010).

52. Fernandez-Escamilla, A. M. *et al.* Solvation in protein folding analysis: combination of theoretical and experimental approaches. *Proc Natl Acad Sci U S A* **101,** 2834–2839 (2004).

53. Brun, L., Isom, D. G., Velu, P., García-Moreno, B. & Royer, C. A. Hydration of the Folding Transition State Ensemble of a Protein †. *Biochemistry* **45,** 3473–3480 (2006).

54. Nishiguchi, S., Goto, Y. & Takahashi, S. Solvation and desolvation dynamics in apomyoglobin folding monitored by time-resolved infrared spectroscopy. *J Mol Biol* **373,** 491–502 (2007).

55. Kimura, T. *et al.* Dehydration of main-chain amides in the final folding step of single-chain monellin revealed by time-resolved infrared spectroscopy. *Proc Natl Acad Sci U S A* **105,** 13391–13396 (2008).

56. Englander, S. W. Protein folding intermediates and pathways studied by hydrogen exchange. *Annu Rev Biophys Biomol Struct* **29,** 213–238 (2000).

57. Wintrode, P. L., Rojsajjakul, T., Vadrevu, R., Matthews, C. R. & Smith, D. L. An obligatory intermediate controls the folding of the alpha-subunit of tryptophan synthase, a TIM barrel protein. *J Mol Biol* **347,** 911–919 (2005).

58. Miranker, A., Robinson, C. V, Radford, S. E., Aplin, R. T. & Dobson, C. M. Detection of transient protein folding populations by mass spectrometry. *Science (80-.).* **262,** 896–900 (1993).

59. Jones, B. E. & Matthews, C. R. Early intermediates in the folding of dihydrofolate reductase from Escherichia coli detected by hydrogen exchange and NMR. *Protein Sci* **4,** 167–177 (1995).

60. Kathuria, S. V., Day, I. J., Wallace, L. A. & Matthews, C. R. Kinetic Traps in the Folding of βα-Repeat Proteins: CheY Initially Misfolds before Accessing the Native Conformation. *J. Mol. Biol.* **382,** 467–484 (2008).

61. Lum, K., Chandler, D. & Weeks, J. D. Hydrophobicity at Small and Large Length Scales. *J. Phys. Chem. B* **103,** 4570–4577 (1999).

62. Huang, X., Margulis, C. J. & Berne, B. J. Dewetting-induced collapse of

hydrophobic particles. *Proc Natl Acad Sci U S A* **100,** 11953–11958 (2003).

63. ten Wolde, P. R. & Chandler, D. Drying-induced hydrophobic polymer collapse. *Proc Natl Acad Sci U S A* **99,** 6539–6543 (2002).

64. Zhang, F. *et al.* Epitaxial growth of peptide nanofilaments on inorganic surfaces: Effects of interfacial hydrophobicity/hydrophilicity. *Angew. Chemie-International Ed.* **45,** 3611–3613 (2006).

65. Hua, L., Zangi, R. & Berne, B. J. Hydrophobic Interactions and Dewetting between Plates with Hydrophobic and Hydrophilic Domains. *J. Phys. Chem. C* **113,** 5244–5253 (2009).

66. Hummer, G., Rasaiah, J. C. & Noworyta, J. P. Water conduction through the hydrophobic channel of a carbon nanotube. *Nature* **414,** 188–190 (2001).

67. Li, J. *et al.* Hydration and dewetting near graphite-CH(3) and graphite-COOH plates. *J Phys Chem B* **109,** 13639–13648 (2005).

68. Li, X., Li, J., Eleftheriou, M. & Zhou, R. Hydration and dewetting near fluorinated superhydrophobic plates. *J Am Chem Soc* **128,** 12439–12447 (2006).

69. Zhou, R., Huang, X., Margulis, C. J. & Berne, B. J. Hydrophobic collapse in multidomain protein folding. *Science (80-.).* **305,** 1605–1609 (2004).

70. Hua, L., Huang, X., Liu, P., Zhou, R. & Berne, B. J. Nanoscale dewetting transition in protein complex folding. *J Phys Chem B* **111,** 9069–9077 (2007).

71. Young, T. *et al.* Dewetting transitions in protein cavities. *Proteins-Structure Funct. Bioinforma.* **78,** 1856–1869

72. Krone, M. G. *et al.* Role of water in mediating the assembly of Alzheimer amyloid-beta a beta 16-22 protofilaments. *J. Am. Chem. Soc.* **130,** 11066–11072 (2008).

73. Liu, P., Huang, X., Zhou, R. & Berne, B. J. Observation of a dewetting transition in the collapse of the melittin tetramer. *Nature* **437,** 159–162 (2005).

74. Yu, N. & Hagan, M. F. Simulations of HIV capsid protein dimerization reveal the effect of chemistry and topography on the mechanism of hydrophobic protein association. *Biophys J* **103,** 1363–1369 (2012).

75. Patel, A. J. *et al.* Sitting at the edge: how biomolecules use hydrophobicity to tune their interactions and function. *J Phys Chem B*

116, 2498–2503 (2012).

76. Wu, Y., Vadrevu, R., Yang, X. & Matthews, C. R. Specific structure appears at the N terminus in the sub-millisecond folding intermediate of the alpha subunit of tryptophan synthase, a TIM barrel protein. *J Mol Biol* **351,** 445–452 (2005).

77. Baldwin, R. L., Frieden, C. & Rose, G. D. Dry molten globule intermediates and the mechanism of protein unfolding. *Proteins Struct. Funct. Bioinforma.* **78,** 2725–2737 (2010).

78. Kiefhaber, T., Labhardt, A. M. & Baldwin, R. L. Direct NMR evidence for an intermediate preceding the rate-limiting step in the unfolding of ribonuclease A. *Nature* **375,** 513–515 (1995).

79. Kiefhaber, T. & Baldwin, R. L. Kinetics of hydrogen bond breakage in the process of unfolding of ribonuclease A measured by pulsed hydrogen exchange. *Proc Natl Acad Sci U S A* **92,** 2657–2661 (1995).

80. Jha, S. K. & Udgaonkar, J. B. Direct evidence for a dry molten globule intermediate during the unfolding of a small protein. *Proc Natl Acad Sci U S A* **106,** 12289–12294 (2009).

81. Bueno, M., Campos, L. A., Estrada, J. & Sancho, J. Energetics of aliphatic deletions in protein cores. *Protein Sci.* **15,** 1858–1872 (2006).

82. Munson, M. *et al.* What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. *Protein Sci* **5,** 1584–1593 (1996).

83. Dalal, S., Canet, D., Kaiser, S. E., Dobson, C. M. & Regan, L. Conservation of mechanism, variation of rate: folding kinetics of three homologous four-helix bundle proteins. *Protein Eng Des Sel* **21,** 197–206 (2008).

84. Yoder, M. D., Lietzke, S. E. & Jurnak, F. Unusual structural features in the parallel beta-helix in pectate lyases. *Structure* **1,** 241–251 (1993).

85. Zitzewitz, J. A., Bilsel, O., Luo, J., Jones, B. E. & Matthews, C. R. Probing the folding mechanism of a leucine zipper peptide by stopped-flow circular dichroism spectroscopy. *Biochemistry* **34,** 12812–12819 (1995).

86. Main, E. R., Lowe, A. R., Mochrie, S. G., Jackson, S. E. & Regan, L. A recurring theme in protein engineering: the design, stability and folding of repeat proteins. *Curr Opin Struct Biol* **15,** 464–471 (2005).

87. Lappalainen, I., Hurley, M. G. & Clarke, J. Plasticity within the obligatory folding nucleus of an immunoglobulin-like domain. *J Mol Biol* **375,** 547–

559 (2008).

88. Hills Jr., R. D. & Brooks 3rd, C. L. Hydrophobic cooperativity as a mechanism for amyloid nucleation. *J Mol Biol* **368,** 894–901 (2007).

89. Brooks, C. L., Gruebele, M., Onuchic, J. N. & Wolynes, P. G. Chemical physics of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 11037–8 (1998).

90. Jackson, S. E. How do small single-domain proteins fold? *Fold. Des.* **3,** R81–R91 (1998).

91. Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins11Edited by P. E. Wright. *J. Mol. Biol.* **277,** 985–994 (1998).

92. Rosen, L. E., Kathuria, S. V., Matthews, C. R., Bilsel, O. & Marqusee, S. Non-Native Structure Appears in Microseconds during the Folding of E. coli RNase H. *J. Mol. Biol.* **427,** 443–453 (2015).

93. Whittaker, S. B.-M., Spence, G. R., Günter Grossmann, J., Radford, S. E. & Moore, G. R. NMR Analysis of the Conformational Properties of the Trapped on-pathway Folding Intermediate of the Bacterial Immunity Protein Im7. *J. Mol. Biol.* **366,** 1001–1015 (2007).

94. Bollen, Y. J. M., Kamphuis, M. B. & van Mierlo, C. P. M. The folding energy landscape of apoflavodoxin is rugged: hydrogen exchange reveals nonproductive misfolded intermediates. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 4095–100 (2006).

95. Baldwin, R. L. On-pathway versus off-pathway folding intermediates. *Fold. Des.* **1,** R1–R8 (1996).

96. Nishimura, C., Dyson, H. J. & Wright, P. E. Identification of Native and Non-native Structure in Kinetic Folding Intermediates of Apomyoglobin. *J. Mol. Biol.* **355,** 139–156 (2006).

97. Matsumura, Y. *et al.* Transient Helical Structure during PI3K and Fyn SH3 Domain Folding. *J. Phys. Chem. B* **117,** 4836–4843 (2013).

98. Nobrega, R. P. *et al.* Modulation of frustration in folding by sequence permutation. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 10562–7 (2014).

99. Kathuria, S. V. *et al.* Microsecond Barrier-Limited Chain Collapse Observed by Time-Resolved FRET and SAXS. *J. Mol. Biol.* **426,** 1980–1994 (2014).

100. Davis, C. M. & Dyer, R. B. WW Domain Folding Complexity Revealed by Infrared Spectroscopy. *Biochemistry* **53,** 5476–5484 (2014).

101. Kathuria, S. V. *et al.* Advances in turbulent mixing techniques to study microsecond protein folding reactions. *Biopolymers* **99,** 888–896 (2013).

102. Orevi, T., Lerner, E., Rahamim, G., Amir, D. & Haas, E. in 113–169 (Humana Press, Totowa, NJ, 2014). doi:10.1007/978-1-62703-649-8_7

103. Zhang, X. *et al.* Direct visualization reveals dynamics of a transient intermediate during protein assembly. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 6450–5 (2011).

104. Konstantinos N. Aprilakis, ‡,§, Humeyra Taskent, ‡,§ and & Daniel P. Raleigh*, ‡,∥. Use of the Novel Fluorescent Amino Acid p-Cyanophenylalanine Offers a Direct Probe of Hydrophobic Core Formation during the Folding of the N-Terminal Domain of the Ribosomal Protein L9 and Provides Evidence for Two-State Folding†. (2007). doi:10.1021/BI7010674

105. Chen, H., Ahsan, S. S., Santiago-Berrios, M. B., Abruña, H. D. & Webb, W. W. Mechanisms of quenching of Alexa fluorophores by natural amino acids. *J. Am. Chem. Soc.* **132,** 7244–5 (2010).

106. Kumar, A. T. N., Zhu, L., Christian, J. F., Demidov, A. A. & Champion, P. M. On the Rate Distribution Analysis of Kinetic Data Using the Maximum Entropy Method: Applications to Myoglobin Relaxation on the Nanosecond and Femtosecond Timescales. doi:10.1021/jp0101209

107. Zitzewitz, J. A., Gualfetti, P. J., Perkons, I. A., Wasta, S. A. & Matthews, C. R. Identifying the structural boundaries of independent folding domains in the alpha subunit of tryptophan synthase, a beta/alpha barrel protein. *Protein Sci.* **8,** 1200–9 (1999).

108. Hills, R. D. & Brooks, C. L. Subdomain Competition, Cooperativity, and Topological Frustration in the Folding of CheY. *J. Mol. Biol.* **382,** 485–495 (2008).

109. Lapidus, L. J. *et al.* Complex pathways in folding of protein G explored by simulation and experiment. *Biophys. J.* **107,** 947–55 (2014).

110. Reddy Goluguri, R. & Udgaonkar, J. B. Microsecond Rearrangements of Hydrophobic Clusters in an Initially Collapsed Globule Prime Structure Formation during the Folding of a Small Protein. (2016). doi:10.1016/j.jmb.2016.06.015

111. Sen, S., Goluguri, R. R. & Udgaonkar, J. B. A Dry Transition State More Compact Than the Native State Is Stabilized by Non-Native Interactions during the Unfolding of a Small Protein. *Biochemistry* **56,** 3699–3703 (2017).

112. Voelz, V. A., Bowman, G. R., Beauchamp, K. & Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J. Am. Chem. Soc.* **132,** 1526–8 (2010).

113. Bowman, G. R. & Pande, V. S. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 10890–5 (2010).

114. Finke, J. M. & Onuchic, J. N. Equilibrium and Kinetic Folding Pathways of a TIM Barrel with a Funneled Energy Landscape. *Biophys. J.* **89,** 488–505 (2005).

115. Karanicolas, J. & Brooks, C. L. Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J. Mol. Biol.* **334,** 309–25 (2003).

116. Eaton, W. A. *et al.* Fast Kinetics and Mechanisms in Protein Folding. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 327–359 (2000).

117. Englander, S. W. & Mayne, L. PROTEIN FOLDING STUDIED USING HYDROGEN-EXCHANGE LABELING AND TWO- DIMENSIONAL NMR. *Annu. Rev. Biophys. Biomol. Struct* **21,** 243–5 (1992).

118. Parker, M. J. & Marqusee, S. A statistical appraisal of native state hydrogen exchange data: evidence for a burst phase continuum?11Edited by P. E. Wright. *J. Mol. Biol.* **300,** 1361–1375 (2000).

119. Parker, M. J. & Marqusee, S. A kinetic folding intermediate probed by native state hydrogen exchange11Edited by A. R. Fersht. *J. Mol. Biol.* **305,** 593–602 (2001).

120. Myers, J. K., Pace, C. N. & Scholtz, J. M. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci* **4,** 2138–2148 (1995).

121. Presta, L. G. & Rose, G. D. Helix signals in proteins. *Science* **240,** 1632–41 (1988).

122. Gerstein, M. A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.* **274,** 562–76 (1997).

123. Yang, X., Kathuria, S. V., Vadrevu, R. & Matthews, C. R. βα-Hairpin Clamps Brace βαβ Modules and Can Make Substantive Contributions to the Stability of TIM Barrel Proteins. *PLoS One* **4,** e7179 (2009).

124. Shi, J. *et al.* Atomistic structural ensemble refinement reveals non-native structure stabilizes a sub-millisecond folding intermediate of CheY. *Sci.*

Rep. **7,** 44116 (2017).

125. Bhuyan, A. K. & Udgaonkar, J. B. Two structural subdomains of barstar detected by rapid mixing NMR measurement of amide hydrogen exchange. *Proteins* **30,** 295–308 (1998).

126. Lin, S. *et al.* Redox-based reagents for chemoselective methionine bioconjugation. *Science (80-.).* **355,** 597–602 (2017).