UNIVERSITY OF CHICAGO

TOPOLOGY AND HETEROGENEITY AT THE RATE-LIMITING STEP OF THE PROTEIN FOLDING PATHWAY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY

BY
ADARSH D. PANDIT

CHICAGO, ILLINOIS
DECEMBER 2005

FOR MY PARENTS AND MY TEACHERS

TABLE OF CONTENTS

LIST OF FIGURES viii

LIST OF TABLES x

LIST OF EQUATIONS xi

ABBREVIATIONS xii

SUMMARY xv

1.0 Introduction 1

1.1 Historical Perspective 1

1.2 Timescales and kinetics 2

1.3 Physical processes in protein folding 5

1.3.1 Hydrophobic Collapse 5

1.3.2 Hydrogen bond formation 6

1.3.3 Conformational entropy 10

1.4 Examination of the rate-limiting step in protein folding 11

1.4.1 Two-state mechanisms and TS perturbation 11

1.4.2 Shortcomings 15

1.5 Heterogeneity 16

1.6 Contact order and folding 19

1.6.1 Plaxco/Baker to present 19

1.6.2 Alternative correlations 23

2.0 ψ-analysis As a Method to Determine TS Topology 25

2.1 Abstract 25

2.2 Introduction 25

2.3 Materials 27

2.3.1 Metal and buffer solutions 27

2.4 Methods 28

2.4.1 Engineering biHis sites 29

2.4.2 Equilibrium studies 30

2.4.3 Chevron Analysis 34

2.4.4 Metal-dependent folding kinetics 36

2.4.5 Testing for fast equilibrium 45

2.5 Data Analysis 46

2.5.1 Equilibrium denaturation profiles 46

2.5.2 Metal stabilization 46

2.5.3 Kinetic Chevron analysis 47

2.5.4 Obtaining ψ-values from the Leffler plot 48

2.5.5 Interpreting ψ-values 49

2.5.6 Correcting for the effects of the biHis site 57

2.5.7 Delineation between heterogeneous and homogeneous scenarios .. 58

2.6 Notes 59

2.6.1 Note 1 59

2.6.2 Note 2 60

2.6.3 Note 3 60

2.6.4 Note 4 60

2.6.5 Note 5 61

2.7 Conclusion 61

3.0 Pathway Heterogeneity in the Transition State of Ctacp 63

3.1 Abstract 63

3.2 Introduction 64

3.3 Materials and Methods 73

3.3.1 Expression and Purification 73

3.3.2 Kinetic folding measurements 73

3.3.3 ψ -analysis 73

3.4 Results 75

3.4.1 Engineering biHis Sites 75

3.4.2 ψ-analysis 76

3.5 Discussion 82

3.6 Conclusion 93

4.0 Native Topology at the Rate-Limiting Step in Folding 94

4.1 Abstract 94

4.2 Introduction 95

4.3 Results and Discussion 97

4.3.1 ψ-analysis and transition state topology 97

4.3.2 Alternative Correlations 103

4.3.3 Modeling TS using HX data 106

4.3.4 Topomer Search Model 126

4.4 Conclusions 128

5.0 Conclusions 129

5.1 Native-like transition state and pathway heterogeneity 129

5.2 Future work 131

5.3 Why topology? 133

References Cited 135

LIST OF FIGURES

FIGURE 1.1 - *Free Energy Diagram of Protein Folding* 3

FIGURE 1.2 - *Funnel model of protein folding*.. 7

FIGURE 1.3 - *Interpreting mutational effects on chevron behavior using ψ-analysis* 13

FIGURE 1.4 - *Relative contact order and folding rate correlation (Plaxco, Baker, and Simons)*.. 21

FIGURE 2.1 - *BiHis Metal Binding sites*.. 31

FIGURE 2.2 - *Chevron analysis as a function of metal concentration*................ 37

FIGURE 2.3 - *Metal-dependent folding kinetics "Metal Chevron"*.................... 41

FIGURE 2.4 - *Leffler Plot* 43

FIGURE 2.5 - *Metal induced equilibrium stabilization* 51

FIGURE 2.6 - *Heterogeneous vs. homogeneous situations*................................. 55

FIGURE 3.1 - *Classes of pathway heterogeneity* 67

FIGURE 3.2 - *Transition state ensemble in ubiquitin* 69

FIGURE 3.3 - *Structure of common-type acyl phosphatase*................................ 71

FIGURE 3.4 - *Schematic representation of biHis sites in ctAcP and ψ-values* ... 77

FIGURE 3.5 - *Multi-metal ψ-analysis of ctAcP-C* 79

FIGURE 3.6 - *Possible disjoint nuclei in ctAcP and Leffler plots indicating possible pathway partitioning*... 85

FIGURE 3.7 - *Proposed TS structures for ctAcP folding*.................................... 87

FIGURE 3.8 - *Contrasting ψ- and ϕ-values in ctAcP* 89

FIGURE 4.1 - *Contact Maps of ctAcP and Ubiquitin TS and native states* 101

FIGURE 4.2 - *Total Contact Distance vs. Folding Rate* 107

FIGURE 4.3 - *Percent local contacts vs. folding rate* 109

FIGURE 4.4 - *Long range order (LRO) vs. folding rate* 111

FIGURE 4.5 - *HX TS Modeling Results* 113

FIGURE 4.6- *RCO - TS vs. Folding Rate* 115

FIGURE 4.7 - *Mechanism of hydrogen bond isotope exchange as a result of unfolding events* 117

FIGURE 4.8 - *Free Energy Surface for global and local unfolding events in hydrogen exchange* 119

LIST OF TABLES

TABLE 2.1 - *Sample shot protocol for a denaturant chevron at high metal ion concentration* 40

TABLE 2.2 - *Sample shot protocol for measuring metal-dependent folding*....... 62

TABLE 3.1 - *Results from metal-binding studies on ctAcP* 81

TABLE 4.1 - *Alternative metrics applied to TS models* 125

LIST OF EQUATIONS

EQUATION 1.1 12

EQUATION 1.2 19

EQUATION 2.1 33

EQUATION 2.2 45

EQUATION 2.3 46

EQUATION 2.4 47

EQUATION 2.5 47

EQUATION 2.6 48

EQUATION 2.7 50

EQUATION 2.8 53

EQUATION 2.9 54

EQUATION 2.10 57

EQUATION 4.1 104

EQUATION 4.2 105

ABBREVIATIONS

Ub	ubiquitin
ctAcP	Common-type acyl phosphatase
TS	Transition State
U	Unfolded (denatured) state
N	Folded (native) state
m_f	Denaturant dependence of the rate of folding k_f (in $s^{-1}M^{-1}$)
m_u	Denaturant dependence of the rate of unfolding k_u (in $s^{-1}M^{-1}$)
m^o	Total denaturant dependence, sum of m_f and m_o
k_f	Single-exponential folding relaxation rate
k_u	Single-exponential unfolding relaxation rate
K_{eq}	Equilibrium constant of folding (k_f / k_u)
WT	Wild-type
biHis	bi-histidine
HX	Hydrogen Exchange
TCD	Total Contact Distance
LRO	Long Range Order
%local	Percent local contacts

ACKNOWLEDGEMENTS

I am thankful for the opportunity to study and research at an esteemed institution as the University of Chicago. The Department of Biochemistry and Molecular Biology has provided not only administrative and academic support without which this work would not have been possible, but also a scientifically stimulating environment as well. Particularly I would like to thank Lisa Anderson (guidance, haircuts), Lisa Alvarez (support, kind words), and Tony Kossiakoff (captain of the ship), as well as Wouter Hoff, Carl Correll, and Nancy Schwartz. I am also thankful for having such an amazing lab group over the years who have primarily provided scientific support, but also friendship, laughter, and pointless discussion. Their roster includes Xingwang Fang, Shane Dothager, Kevin Meisner, Kevin Shi, Muhammed Zaman, Andrés Colubri, Abishekh Jha, Gerra Bosco, Devin Strickland, Michael Baxa, Nathan Baird, and Emily Cukier. Special thanks to Andrés and Abishekh for assistance in modeling studies, simulations, and other calculations. I am especially thankful to Bryan Krantz both for serving as a mentor and also for the seminal work upon which the following research is based. Since first setting foot on campus, my boat has been tied to that of Ali Shandiz, in whom I have found a colleague, compatriot, intellectual sparring partner, and most importantly, a great friend. Committee meetings with Phoebe Rice, Tao Pan, and Jon Staley have been not only essential to my success and progress, but also have provided vibrant discussion as well as valuable perspective on my project. My appreciation for the discussions, advice, arguments, and guidance of my advisor Tobin Sosnick is beyond the ability of these pages to contain. I am not only lucky to have had an advisor with such amazing intellect and drive, but also with dignity, humor, and compassion. I am indebted to him for his guidance and humanity. I would not have become the scientist I am today without the science teachers in my life, including Shyam Vyas, Don Herbert, Dave Maile, Marcia Lizzio, John Campbell, Jim Gaut, Dave Ballou, Bruce Palfey, Phil Allen, Paul Janmey, and Rahul Kohli. The students of the Division of Biological Sciences have helped create a community of peers essential for scientific development and friendship for which I am thankful. I could not have finished this work without the support of my friends, especially Spencer Anderson, Jason Key, Erica Rosner, Sean Crosson, Jody Fletcher, Vic Huang, Adam Conway, Nels Elde, Matt Giorgianni, Shannon Byrd, Scott Robinson, Kim Sciarretta, Leslie King, Jon Hickson, Dave Leonard, Rabiah Mayas, Seema Bajaria, Sona Bajaria, Gavin Jackson, Maneesh Gulati, Nirav Shah, and DJ Silton. This work is also in memory of Drs. Ramu Pandit and Hansraj Bajaria, who have inspired me intellectually my whole life. In my own family, I am thankful to have had support and encouragement from Dr. Harshida Pandit, Bhupendra and Harsha Vyas, Neeti Vyas, Pranav Pandit, Ramu and Shirley Pandit, and Dr. Neha Pandit. Finally, I wish to thank my parents, my brother, and my sister-in-law for their love and support without which none of this would have been possible.

SUMMARY

Protein folding remains one of the last mysteries remaining in the field of molecular biology. Understanding the process of protein folding will likely be critical to design of novel proteins in the future. Two-state proteins only exist in either the unfolded or native states, so examination of the process of folding is targeted at the high-energy transition state barrier. The transition state of common-type acyl phosphatase (ctAcP) is characterized using ψ-analysis which identifies chain-chain contacts using engineered bi-histidine metal ion binding sites located throughout the protein. As with ubiquitin [1], the other globular protein extensively characterized using ψ-analysis, the transition state of ctAcP has very native-like topology. Using multiple-metal folding analysis, it was determined that very little pathway heterogeneity exists at the rate-limiting step, indicating the transition state ensemble of ctAcP has single consensus structure with a minor amount of optional elements. Using hydrogen exchange results, models of several transition states were generated and found to be very native-like in topology. Based on these results and previous correlations between topology and folding rate, we believe proteins who obey a linear relationship between topological complexity and folding activation energy are likely to retain ~80% of native topology in the transition state. This result experimentally confirms that arrangement of the polypeptide chain into a native-like configuration is the rate limiting step in protein folding.

1.0 Introduction

1.1 Historical Perspective

The manner in which proteins produced in the ribosome convert from disorganized chain of amino acids to a single globular structure is one of the most compelling and vigorously investigated mysteries of molecular biology. *In vivo*, proteins catalyze a dizzying array of cellular processes including chemical reactions, recognition, and ion translocation, making their study important to our understanding of the cell as a biomolecular machine. The so-called "protein folding problem" describes the inability to predict three-dimensional protein structure from one-dimensional sequence using first principles. This remains as the last unsolved piece of the central dogma of molecular biology, which describes the information transfer from primary DNA sequence to RNA messages which are then translated into protein sequence.

That unfurled polypeptide chains can adopt a distinct three dimensional structure based solely on the sequence of the amino acids is fascinating in and of itself, especially given only an alphabet of twenty chemical side chains to choose from. Christian Anfinsen first demonstrated the reversibility of folding in ribonuclease, by fully unfolding or denaturing the protein to disrupt catalytic activity, which could be recovered by returning the protein to refolding conditions [2]. More intriguing is the spontaneous nature of the reaction at room temperature, indicating that folding is energetically downhill and also occurs at on a time scale compatible with biological processes. Importantly, this result confirmed that structural information for a protein is fully encoded in the amino acid sequence. This seminal work ushered in the age of extensive studies of protein folding.

Early experimentalists noted the “all-or-none” conversion between folded and unfolded states, and began to describe the transition in energetic terms normally used for phase transitions. Formalism for simple chemical reactions was adopted using a reaction diagram with two states (denatured and native) separated by a single high-energy barrier, the graphical representation of which is known as a “free energy surface” [3]. (FIGURE 1.1) This concept gave experimentalists a strong foothold on which to base their observations regarding the protein folding energetics, especially kinetic behavior.

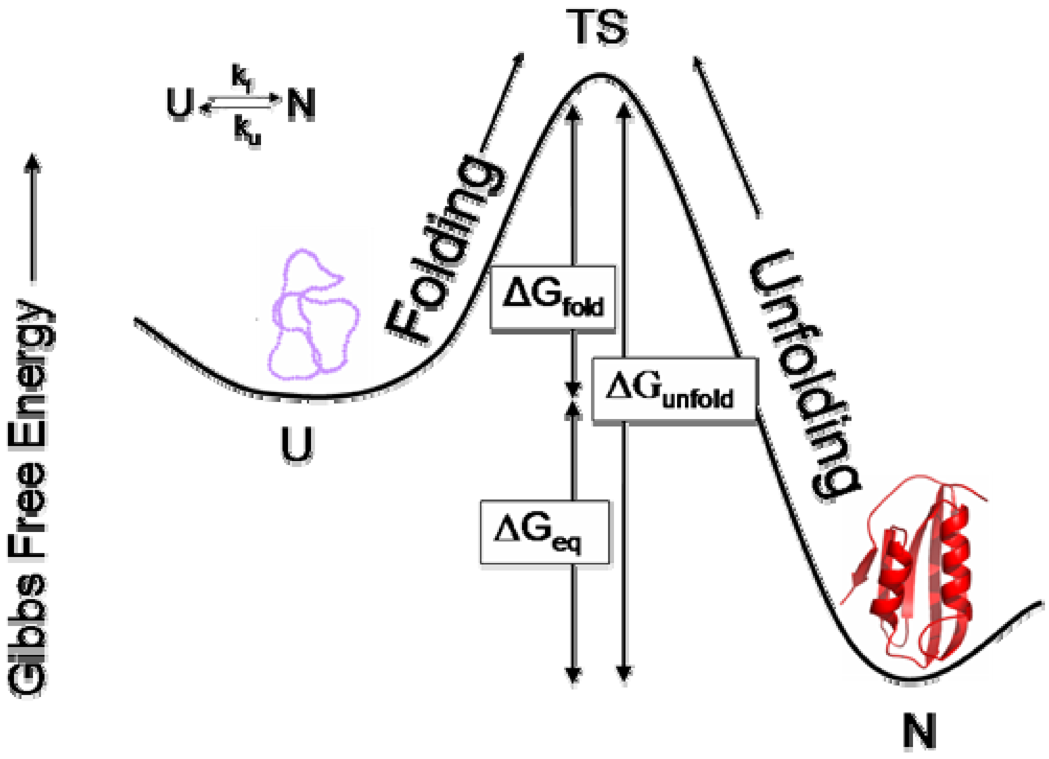
1.2 Timescales and kinetics

Speculation regarding the steps involved in protein folding began with a comparison of the timescales of chemical and biological reactions. In the cell, the response time to stimulus can be as rapid as a few seconds, so translation of a particular gene and the rate of protein folding must all occur on a similar order of magnitude. However, we know from *in vitro* experiments that folding rates span six orders of magnitude from microseconds to seconds and are in experimental agreement with the cellular timescales [4; 5].

However, theoretical simple predictions of folding rates are drastically slower, and this contrast between the measured and predicted rates gave rise to Levinthal’s paradox [6]. This concept states that if each residue can assume one of three backbone rotamers independent of it’s neighbors, then a 100-

FIGURE 1.1 - *Free Energy Diagram of Protein Folding*

A representation of the free energetic surface of the protein folding landscape. The polypeptide assumes any number of random coil configurations in the unfolded (U) state, and must transit the high energy transition state (TS) on the way to forming the singular native state (N). The activation energy for the folding process is ΔG_{fold}, for the unfolding process the energy required is indicated by ΔG_{unf}. The difference between the two quantities is the equilibrium stability of folding ΔG_{eq}.


TS
$U \underset{k_u}{\overset{k_f}{\rightleftarrows}} N$
Gibbs Free Energy
Folding
Unfolding
ΔG_{fold}
ΔG_{unfold}
ΔG_{eq}
U
N

amino acid chain would have 3^{100} possible conformations, and sampling them even very rapidly would take longer than the age of the universe. Clearly the answer to this paradox is that a full random search is not engaged, and the energy surface of the folding reaction somehow energetically guides the conformation to the native state [7].

Many ideas have been proposed to explain this discrepancy, the most popular of which describes the protein conformational search broken down into many smaller local searches instead of a large global search, which would be substantially faster [8]. Graphically, this concept can be roughly visualized using an energetic funnel, where each successive step toward the native state reduces the possible number of states and thus guides structure formation progressively [9; 10; 11]. (FIGURE 1.2)

1.3 Physical processes in protein folding

Protein folding rates experimentally span six orders of magnitude. To better understand the kinetics of folding, we can break the overall process into separate physico-chemical events, each with a distinct timescale dependent upon the processes involved.

1.3.1 Hydrophobic Collapse

Proteins are generally built with hydrophilic residues on the surface and hydrophobic residues in the internal core, shielded away from solvent. Energetically, the affinity of hydrophobic residues for one another, also known as the “hydrophobic effect”, is derived less from a mutual attraction than from a preference of water to self-associate. When hydrocarbons are exposed to water, a clathrate cage of specifically oriented water molecules forms around the nonpolar atoms to locally minimize the dielectric constant [12]. The energy of desolvation is the difference between a solvated hydrophobic residue surrounded by this entropically costly water cage and being buried among other nonpolar residues in the protein core. This process is thought to be the largest energetic driver in the folding process.

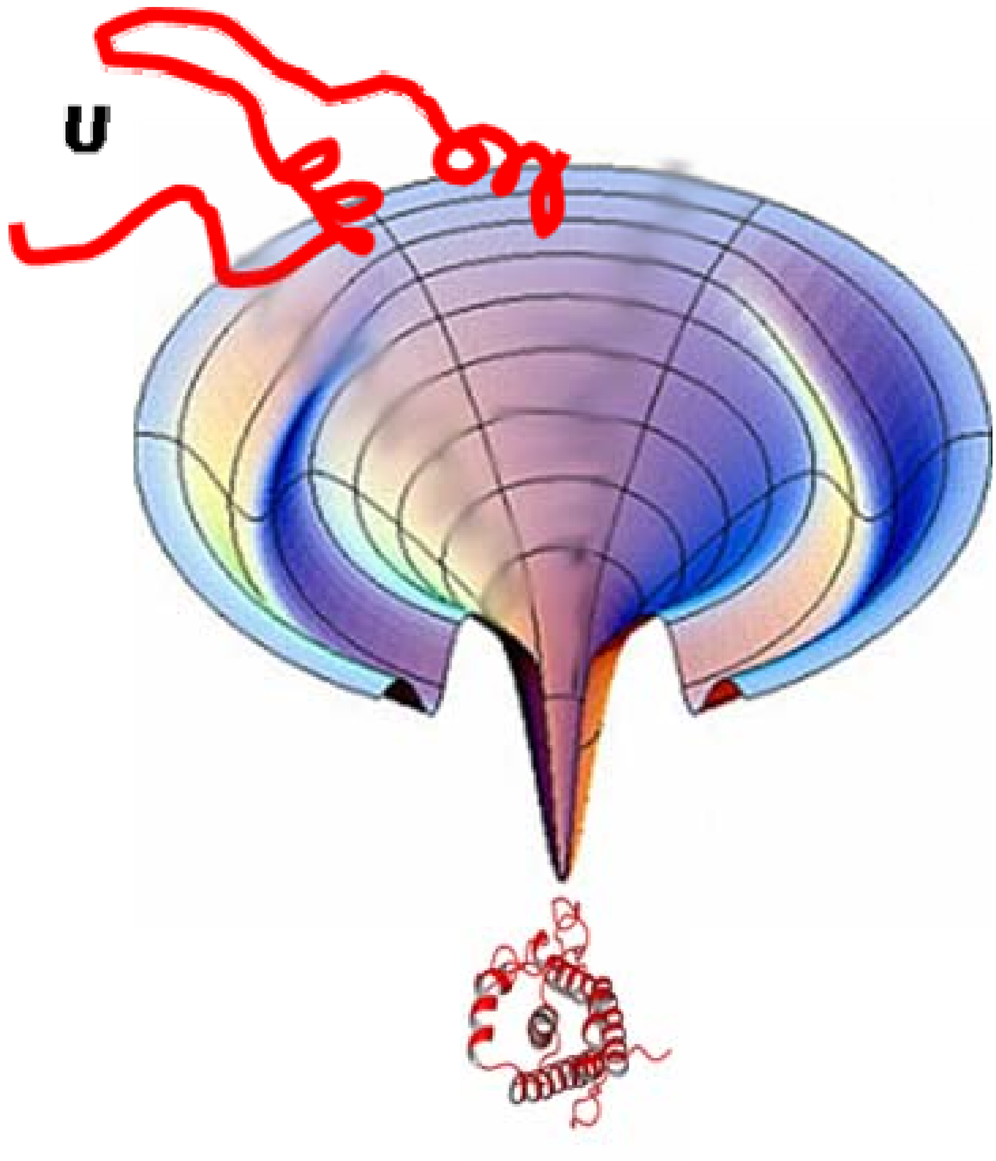
On the protein folding pathway, the hydrophobic residues must form the native residue-residue contacts, either early via hydrophobic collapse, or later in the pathway, for example, after secondary structure has formed. Some believe that the overall rate of folding is largely determined by the amount of hydrophobic content, suggesting that hydrophobic association itself is the rate-limiting step [13]. However other measurements indicate that collapse occurs concomitantly with formation of native topology and hydrogen bond formation [14]. Potentially, rates of hydrophobic collapse could be very rapid, as fast as tens of nanoseconds [15], or as long as 100μs [16], so it seems unlikely that collapse by itself could be the rate-limiting process.

1.3.2 Hydrogen bond formation

Polypeptide chains have two backbone hydrogen bonding groups, amide carbonyl and N-H moieties, between each side-chain. Having such regularly

FIGURE 1.2 - *Funnel model of protein folding*

A representation of the three-dimensional energetic surface described by the "funnel model" of protein folding. The outermost rim represents the large number of conformational states which the unfolded state can exist. On the way to the native state at the bottom of the funnel, a high energy barrier exists (TS) through which proteins must pass. As protein molecules more closely resemble the folded state, the number of possible configurations reduces to a nearly singular native state. Adapted from Ken Dill Laboratory website (*http://www.dillgroup.ucsf.edu*).


U

spaced polar groups in a polymer which must bury non-polar groups presents a unique puzzle – the backbone must have its hydrogen bonds paired to avoid energetic penalties associated with the burial of unsatisfied hydrogen bonds in the low dielectric environment of the core. Given this specific bonding requirement, along with the near-crystalline side-chain packing in the core of most proteins, the formation of hydrogen bonds in the folding process has remained a subject of much study.

Mirsky and Pauling were the first to suggest that hydrogen bonding was the dominant force stabilizing proteins and that a bond could form between the N-H and C=O groups of the backbone [17]. A solvated hydrogen bond in isolation is thought to be worth anywhere from 1 to 3 kcal/mol, depending on the context, based on solvent-transfer studies. However, these studies are subject to much debate (in ref. to a Rose paper, but not George, early, 70's or 80's), and the value extracted depends on the standard state reference concentration used in the calculation involving the dimeric reaction. A backbone hydrogen bond also is more stable in the core, not only due to the simple physics of electrostatic interactions in a low-dielectric media, but also because there are no available alternate bonding partners. When taken together, all hydrogen bonds formed in the folding process account for a great deal of the stability of the protein. However, a substantial amount of work is involved in desolvating the hydrogen bond partners. This work may even outweigh the energy derived from backbone hydrogen bonding in some cases.

Despite the importance of hydrogen bond formation in the folding process, the point on the pathway at which bond formation occurs varies greatly among models. In the extrema, the “hydrophobic collapse” model postulates rapid association of hydrophobic residues and then slow formation of native hydrogen bonds, whereas the “diffusion-collision” model suggests formation of isolated secondary structural elements and native hydrogen bonds as the initial step.

Work from our laboratory making use of the kinetic isotope effect examined hydrogen-bond formation in a library of proteins with a variety of secondary structural content [18]. The amide hydrogen will readily exchange for deuterium when exposed to D_2O solvent, which weakens the hydrogen bond by a small, but kinetically influential amount. Changes in kinetic rates as a result of this specific weakening of the hydrogen bond across a library of proteins indicate that hydrogen bond formation and surface area burial are concomitant processes, as opposed to sequential. Hence, neither nonspecific collapse [19] nor secondary structure formation [20; 21] are the first to occur on the folding pathway, the two processes seem to go hand-in-hand.

1.3.3 Conformational entropy

The energetic balance of the protein is stabilized on the one hand by hydrophobic association and destabilized by the large amount of backbone and sidechain conformational entropy. This entropy is reduced due to backbone rotamer preferences which are residue-dependent. In addition, the identity of a residue’s neighbors also influences its backbone preferences [22; 23; 24]. Between these forces, the conformer space is largely reduced. In spite of the ability of each individual rotamer to rapidly rotate [25; 26; 27; 28], alignment of the chain into a native-like arrangement can take a great deal of time, depending on the complexity of the topology.

Given the discrepancy between the rapid rate at which the individual electrostatic processes occur in isolation and the much slower measured rate of folding, it is clear that other more complex factors are at work in the folding process, which sum to more than the component energetic parts.

1.4 Examination of the rate-limiting step in protein folding

1.4.1 Two-state mechanisms and TS perturbation

Early experiments by Baldwin *et al* began to extend previous work on the stability of proteins to investigate the rates at which proteins fold under various conditions [29]. From a very early point, the discussion of the pathway of protein folding revolved around the number of steps between the unfolded (U) state and native (N) states, primarily divided between those who believed stepwise intermediates existed [30], versus those who saw the process as “two-state” with no intermediates [3]. The idea of sequential intermediate is intellectually appealing, in that the Levinthal Paradox could be overcome by formation of small amounts of structure, which would cooperatively induce more structure formation. However for most proteins under 120 amino acids, populated intermediates have been largely discredited as aggregation, proline isomerization, or slow disulphide bond formation. In some cases, for example Cytochrome C, the large prosthetic group does induce a stable folding intermediate, however the large heme group alters the chemical architecture in a complex fashion and this result is unlikely to represent a general phenomenon [31].

"Chevron analysis", developed by Goldenberg and Creighton as well as Matthews in the 1980s, was popularized in the 1990s principally by Fersht and coworkers, who saw point-mutations as a tractable perturbation through which changes in folding behavior could be readily interpreted [32; 33; 34]. Using experiments where the protein environment is rapidly changed from stabilizing to destabilizing conditions, the energies of the native and unfolded states can be determined, at least in relative terms. Plotting the kinetic rates of folding versus denaturant concentration gives an inverted chevron or "V"-shape and provides a wealth of kinetic and thermodynamic folding information for comparison between wild-type and mutant variants.

Importantly, chevron analysis provides a variety of information. Primarily, chevrons can be used to determine the intrinsic rate of folding which is proportional to the activation energy in

$$\Delta G \propto RT \ln k_f^{\mp}$$

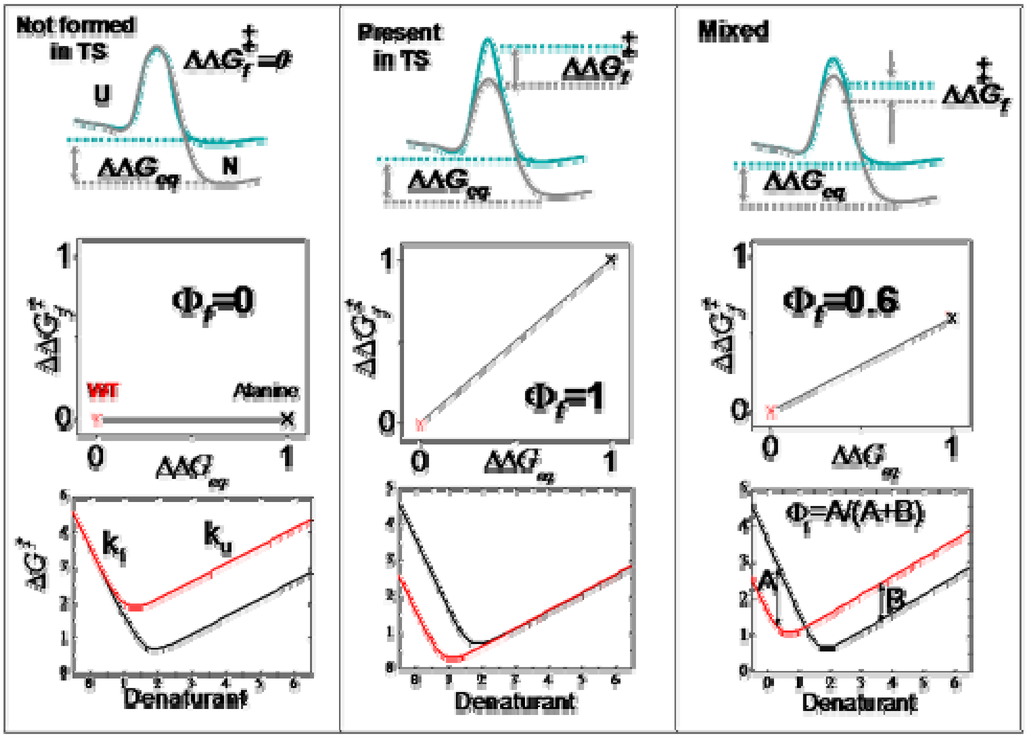
EQUATION 1.1

Frequently, there is a linear relationship between activation free energies of folding and denaturant concentration, where the proportionality between the two relates to the degree of surface area burial in the folding and unfolding processes [35]. Additionally, a protein can be verified as two-state when kinetically

FIGURE 1.3 - *Interpreting mutational effects on chevron behavior using ψ-analysis*

Mutational effects on folding behavior can be quantified using chevron analysis. Comparing wild type and point-mutant chevrons allow for identification of the degree of interactions the mutated residue makes in the transition state. The ratio of energetic change in the transition state to that of the native state is the ϕ-value ($\phi=\Delta\Delta G^{\ddagger}_{fold}/\Delta\Delta G_{eq}$). (Left) If the target residue makes no interactions in the transition state, then its modification through mutation will have no energetic effect on the rate of folding. The rate of unfolding (k_u), however will be faster or slower depending on the stabilizing or destabilizing nature of the interactions. (Center) Conversely, if the sidechain makes critical interactions for establishment of the transition state nucleus, mutations which eliminate this interaction, and destabilize the transition state and native state an equal amount (ϕ=1). (Right) The intermediate scenario results in a fractional ϕ-value which can signify either a partial interaction in the transition state, or pathway heterogeneity.

Quantifying effects of mutations: $\phi = \frac{\Delta\Delta G_{folding}}{\Delta\Delta G_{equil}}$



determined values for stability and the amount of surface area buried (*m*-value) agree with the equilibrium derived counterparts.

Measuring the effect of a single point mutation on the folding and unfolding rates can be used to quantify the participation of the residue in the transition state using ϕ-analysis, a formalism based on ideas from Leffler and Leffler [36; 37]. (FIGURE 1.3) For example, if a mutation has no effect on the stability of the transition state, but rather only on the native state, then one can conclude the residue in question does not participate energetically in the transition state. The ϕ parameter represents the ratio of energetic change in the TS to that of the ground state ($\phi=\Delta\Delta G^{\ddagger}/ \Delta\Delta G_{eq}$), such that zero indicates no participation in the TS while unity suggests the residue and its interactions are critical for establishing the folding nucleus. However, the application of ϕ-analysis can be as numerous and complex as the interactions each residue makes with its neighbors. [1; 38; 39; 40; 41; 42; 43; 44]

1.4.2 Shortcomings

Residues on the surface of a protein tend to make fewer interactions with other sidechains, as compared to those closely packed in the protein core. As a result, perturbations within the protein interior may report not only on the degree of the target residue's participation in the TS, but also on the changes in the complex web of interactions with other residues which it is participating in. In addition, the identity of various residues with different chemical properties makes not only the choice of original mutation important to the resultant ϕ-value, but also which amino acid the target is mutated to. In these ways, the design of the experiment can determine the outcome itself.

Those issues aside, there are interpretational difficulties regarding fractional ϕ-values. For example, if multiple TSs exists, a mutation which destabilizes the structure locally will also reduce the proportion of TS with this interaction by raising the energetic barrier through this pathway. As a result, lower energy pathways will be favored and the ϕ-value will be lower than otherwise predicted, as seen in our lab's study of the folding of a coiled-coil [45].

The shifting of the dominant transition state toward a less-native configuration is termed "Anti-Hammond behavior", or movement of the TS toward the unfolded state [46; 47; 48; 49; 50].Here we saw multiple nucleation positions which were inaccurately reported as low ϕ-values due to relative destabilization of the pathways [51]. Additionally, ϕ-values may underreport the degree of native structure in the TS due to relaxation of mutated regions in a nonnative fashion, as seen in nonnative packing of the four-helix bundle [52]. These and other results indicate that ϕ-analysis may grossly misdiagnose the degree of native structure or topology in the TS.

1.5 Heterogeneity

Additionally, fractional ϕ-values are difficult to interpret, as they could imply either a partial disruption of the TS interaction, or an elimination of the interaction in a heterogeneous folding populace [38]. In the latter case, for example, half of the protein population may have a full interaction in the TS whereas the other half does not, but superficially indicating a half-formed interaction. However kinetic measurements generally cannot distinguish between these scenarios using current methods.

A discussion of pathway heterogeneity must begin with a definition of what can be considered a separate pathway. Such a definition must necessarily be defined by the amount of structural difference and also the timescales involved. For example, at a very detailed level, it is very unlikely that every atom within a population of proteins is in an identical configuration. It is perhaps also unlikely that even side-chain residues will be in the same rotamer configurations in all cases. Many conformations are sure to exist as slightly different dynamic configurations which result from Boltzman thermal fluctuations. If we define heterogeneity in this fashion, then it seems likely that heterogeneous structure will be a ubiquitous phenomenon, albeit one which has less relevance.

It is perhaps more appropriate to think of protein folding pathway heterogeneity as a phenomenon wherein populations are grouped by significant amounts of non-overlapping structural dissimilarities. For example, one pathway could utilize a TS with one helix formed while molecules following another pathway lack this structure. Additionally, it is important to consider which elements are similar in the TS in addition to the differences. We believe pathway heterogeneity is clearly defined by “structurally disjoint” transition states, where multiple sets of structurally distinct TS populations exist, however we can imagine a good portion of the TS structure is shared in all pathways. At this level of detail, little evidence exists to support structural heterogeneity in the transition state in the folding of two-state proteins.

Results from theoretical folding studies often give results that indicate the presence of pathway heterogeneity [53; 54]. One study suggested that pathway heterogeneity was a mechanism through which proteins could fold faster [55]. However simulation algorithms produce individual trajectories which are rarely reproducible, so commonalities among large numbers of simulations are grouped together as “pathways”. As a result, extrapolation to experimental folding results becomes somewhat unclear.

More relevant experimental results in GCN4 using metal-binding studies clearly indicated multiplicity of pathways through the transition state. Using metal sites at either end of the coiled-coil, the amount of TS structure in the N- and C-termini was quantified. Additionally, the effect of changing the chain connectivity (introduction of a terminal disulfide crosslink) on nucleation in the TS was measured. The result suggested that local contacts are favored in the TS, and that several nucleation sites existed in the protein.

Other experiments in the β-sandwich protein titin have indicated a very diffuse transition state [56; 57]. Using ϕ-analysis results, the TS is modeled as having multiple pathways through which the protein may fold, which are considered on-pathway intermediates. While the results seem to be convincing enough, they are based on ϕ-analysis data, which is inherently biased toward diffuse transition states. This result, behind the GCN4 work, is considered the next best experimental representation of pathway heterogeneity. As such, a clear picture of the generalized behavior and multiplicity of transition states does not exist and requires additional investigation.

1.6 Contact order and folding

1.6.1 Plaxco/Baker to present

Many theoretical studies have attempted to predict protein folding rates based on properties other than the activation energy of folding. Studies have suggested that size [58], stability [59], and topology [60] all bear heavily on the rate at which a protein folds. However experimental results have not been very successful in demonstrating a relationship between folding rate and any of these properties [61].

In 1998, Plaxco, Simons, and Baker, analyzed existing folding data for a number of proteins at similar conditions and noticed a statistically significant linear anticorrelation between the logarithm of the folding rate and Relative Contact Order (log k_f vs. RCO) [62]. (FIGURE 1.4) RCO is defined as

$$RCO = \frac{1}{L \cdot N} \sum^{N} \Delta S_{i,j}$$

EQUATION 1.2

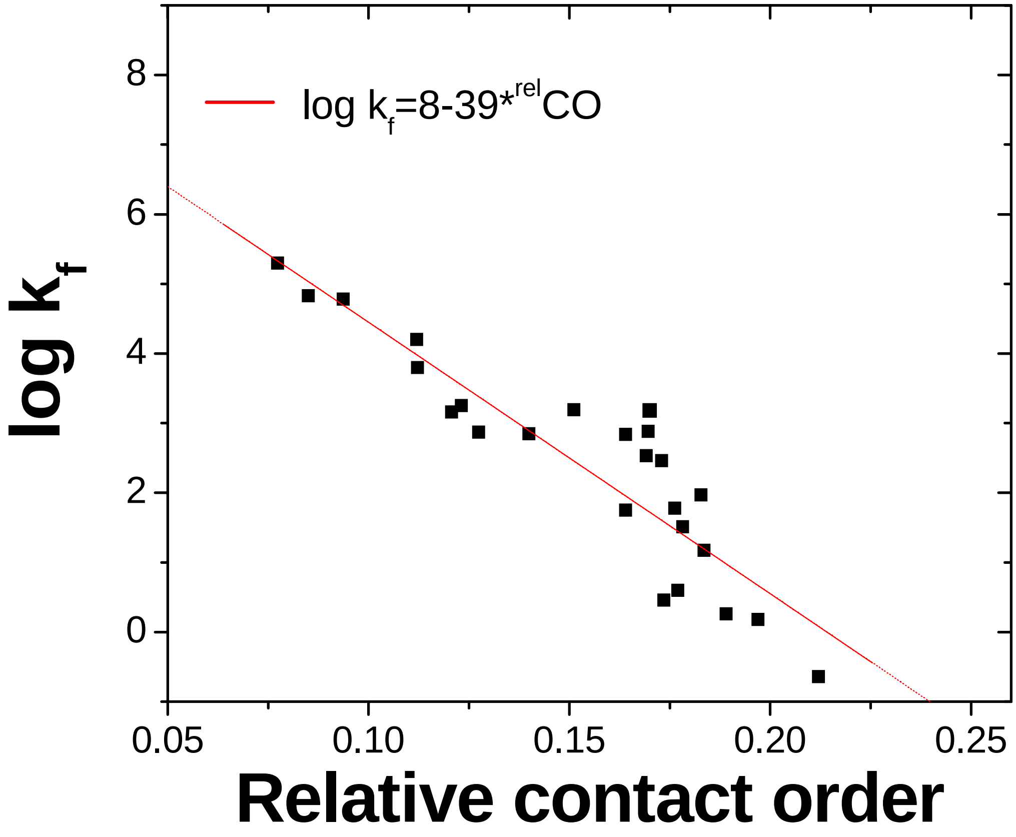
where L is the chain length, N is the number of residue-residue contacts, and $\Delta S_{i,j}$ is the sequence distance between any two residues *i* and *j* which are within some threshold distance of each other in three-dimensional space. This quantity represents the average sequence distance of all residue-residue contacts, normalized for number of contacts and chain length, or a calculation of the degree to which native contacts are long-range versus local. Taken broadly, RCO quantifies the difference between proteins which are topologically simple (α-helices, local contacts) and complex (β-sheets, long-range contacts). It is interesting to note the correlation becomes somewhat worse when not normalized for chain length and contacts (both of which scale nonlinearly with the three-dimensional size of a protein), suggesting that size is not a major determinant of folding rates.

The implication of a correlation between folding rate and the complexity of the native state is clear - the activation energy of folding, or reaching the TS from the unfolded state, is dependent upon arrangement of the protein chain into a native-like topology, after which folding is energetically downhill. Physically, the activation energy barrier then becomes strongly influenced by the loop closure entropy which establishes some large percentage of native topology, an idea which has been verified theoretically [63].

The correlation between log k_f and the RCO of the native state suggests that proteins tend to have so-called “late transition states”, or rate-limiting steps where a great deal of native structure is formed. These two pieces of information directly tie the TS structure to the native structure by way of a property which is more general than just specific native interactions - topology. Also, this meta-analysis confirms the idea first proposed by Sosnick *et al* regarding the nucleation of native topology as the rate-limiting step in folding [64].

FIGURE 1.4 - *Relative contact order and folding rate correlation (Plaxco, Baker, and Simons)*

Correlation of the log of the folding rate (k_{fold}) with Relative Contact Order, a parameter which quantifies protein topology along with a linear fit to the data. (adapted from Plaxco, Baker, and Simons, 1998 [62])


log k_f=8-39*^{rel}CO
log k_f
8
6
4
2
0
0.05
0.10
0.15
0.20
0.25
Relative contact order

1.6.2 Alternative correlations

The RCO correlation result was partially born of the simple observation zzthat helical proteins fold faster than those with β-sheet. Others have recapitulated the same result using different, yet related methodologies. The results which are most similar in results are alternative methods of quantifying residue-residue distances throughout a structure. A parameter called Long Range Order (LRO) developed by Gromiha and Selvaraj sums the number of inter-residue contacts which are further than 12 in length, normalized by the number of residues [65].

Mirny and Shakhnovich took the opposite approach and quantified the fraction of total contacts which were within four residues, which essentially limits the interactions to those in a helix or turn, again normalized for length [66]. Gong *et al* used regression analysis to determine what extent helices, sheets, and turns correlated to folding rates and assigned coefficients for each element, comparing various methods of secondary structure assignment [67]. Zhou and Zhou took an approach similar to that of Gromiha and Selvaraj's LRO but removed the normalization and got nominally similar results using a parameter called Total Contact Distance (TCD) [68].

The work of Bai *et al* sought to determine the size of the transition state topologically and found a correlation with folding rate [69]. In investigating the effect of how to appropriately model unfolded loops in partially folded structures, the authors sequentially delete the six amino acid stretch which has the least effect on TCD. The TS was determined to be the structure from which no residues could be deleted without a significant change in TCD. In this fashion, they aimed to create a structural map of the transition state using minimal effect on TCD as a limiting parameter. The results indicate a strong correlation between the size of the transition state, as measured by TCD, and folding rate. This result further bolsters the idea that a minimal number of native contacts is required for establishment of the transition state.

While many of these correlations perform better than the original topology vs. folding rate analysis, the general phenomenon of overall structural complexity is embodied in all of these analytical approaches. Quantifying the number of long range interactions versus secondary structure content, when α-helix gives a fixed low contact number and β-sheet gives a more variable, but on average higher value amounts to different approaches to the same idea. Importantly, the large amount of theoretical and analytical work, which has initially advanced ideas regarding topology as the determinant of folding rate, as of yet has no strong experimental demonstration. A method directly sensitive to topology is required to validate these results. This thesis explores the relationship between topological complexity and folding rate as well as the implication on the transition states of proteins in general.

2.0 ψ-analysis As a Method to Determine TS Topology

2.1 Abstract

We discuss the implementation of ψ-analysis for the structural characterization of protein folding transition states. In ψ-analysis, engineered bi-histidine metal ion binding sites at surface positions can locally stabilize structure with increasing divalent cation concentration. Increasing metal ion concentration stabilizes the interaction between the two known histidines in a continuous fashion. Measuring the ratio of transition state stabilization to that of the native state provides information about the presence of the binding site in the transition state. Because ψ-analysis uses non-invasive surface mutations and does not require specialized equipment, it can be readily applied to characterize the folding of many proteins. As a result, the method can provide a wealth of high resolution quantitative data for comparison with theoretical folding simulations. Additionally, investigations of other biological processes also may utilize such metal binding sites and ψ-analysis to detect conformational events during catalysis, assembly, and function.

2.2 Introduction

One of the main unanswered questions in molecular biology is how amino acid sequence codes for the three-dimensional structure of a protein. De novo attempts to determine structure from sequence have yet to succeed satisfactorily[70]. Elucidation of the energetic and structural steps on the protein folding pathway is likely to play a major role in the improvement of structure prediction algorithms.

Most small globular proteins fold in a kinetically two-state manner (U↔N) where intermediates do not populate [14; 71; 72; 73]. As a result, the transition state (TS) ensemble is the only point on the folding pathway readily amenable to characterization. Mutational ϕ-analysis has long been the accepted method for characterizing the structure of TS of the folding pathway [32; 74]. However, the interpretation of results, especially fractional ϕ-values, has become a subject of much debate [1; 38; 39; 40; 41; 42; 43; 44].

We have developed ψ-analysis, a method complementary to ϕ-analysis, in order to better characterize TS ensembles [1; 38; 75]. In ψ-analysis engineered bi-Histidine (biHis) metal ion binding sites are introduced at known positions on the protein surface to stabilize secondary and tertiary structures. The addition of increasing concentration of metal ions stabilizes the interaction between the two histidine partners in a continuous fashion. As a result, this method quantitatively evaluates the metal-induced stabilization of the TS relative to the native state, represented by the ψ-value. The translation of a measured ψ-value to structure formation is straightforward, because the proximity of two known positions is probed. Hence, the method is particularly well-suited for defining the topology and structure of TS ensembles. In the following sections, we will discuss the implementation of this method.

2.3 Materials

2.3.1 Metal and buffer solutions

In addition to the standard materials used in folding experiments, metal-containing denaturant and buffer solutions are required. The ψ-analysis method depends upon metal binding to the deprotonated form of histidine (intrinsic pKa ~ 6.5), so experiments should be conducted at pH 7.5 or above to maximize the binding stabilization energy, $\Delta\Delta G_{bind}$. Buffer selection depends primarily on proximity of the pKa to desired pH. However, testing of buffer-metal combinations before experimentation is recommended because some buffers such as sodium phosphate can lead to metal precipitation. We typically employ 50 mM Tris or HEPES at pH 7.5. Nevertheless, precipitation remains a persistent problem so solutions should be made fresh each day. We find it convenient to make a high metal-ion stock for diluting into metal-buffer solutions. All buffer solutions, especially metal-containing buffers, should be checked for pH changes immediately prior to use.

1. Zinc Chloride ($ZnCl_2$) stock solutions can be prepared at 0.25 M in 25 mM HCl and should be remade every month or two. Zinc chloride readily precipitates at high concentration and neutral pH, so buffers should be prepared with care. For example, when making a zinc-containing buffer, add the buffer stock solution, dilute to just under the final volume, and add the zinc stock solution. The addition of acidic metal solution often lowers the pH, so it should be rechecked immediately before use.

2. Cobalt chloride ($CoCl_2$) solutions appear burgundy red and stocks can be prepared at 1 M in water. Using buffers with cobalt at guanidine hydrochloride (GdmCl) concentration of ~5.5 M can cause metal precipitation and erratic kinetic results. Excessive amount of chloride ions, as present in high concentration GdmCl buffers, will cause cobalt solutions to appear blue, which may affect results.

3. Cadmium chloride ($CdCl_2$) solutions can be prepared at 1 M in water. Cadmium is a heavy metal which can be toxic. Please refer to material safety data sheets for safe handling procedures such as avoiding skin contact and inhalation. Solutions can be prepared in the same fashion as cobalt.

4. Nickel chloride ($NiCl_2$) solutions can be prepared at 0.25 M in 25mM HCl. We have found that nickel can be readily combined with buffers as this metal is not as susceptible to precipitation problems as the other metal ions. Additionally, Ni^{2+} adopts square planar orbital geometry as opposed to tetrahedral in other cases.

2.4 Methods

Overview: In both ϕ-analysis and ψ-analysis, the change in folding rate due to an energetic perturbation identifies the degree to which this interaction is present in the TS ensemble. In mutational ϕ-analysis, the perturbation is a single side-chain substitution. In ψ-analysis, biHis sites are individually engineered onto the protein surface and its stability is manipulated using divalent metal ions. Generally, after insertion of a biHis site at a region of interest, the denaturant dependence of folding rates ("chevron analysis" [32]) is conducted to characterize the overall folding behavior and identify a suitable metal ion. A Leffler [76] or Leffler [77] plot is then obtained, by determining the change in folding and unfolding rates as a function of metal concentration, to calculate the ψ-value.

2.4.1 Engineering biHis sites

The strategy of using engineered metal ion binding sites in biochemical studies has an extensive history [78; 79; 80; 81; 82; 83; 84; 85; 86; 87]. Our folding studies have used biHis sites located on the surface of the protein [84] rather than buried sites. For a surface site, the metal-induced stabilization is specific to a particular structural element, such as a helix or hairpin, and the protein can be folded in the absence of the metal. (FIGURE 2.1) These properties generally are not applicable to proteins with buried sites, such zinc finger proteins, where metal ions are required for the cooperative folding of the entire protein [81]. Buried sites typically have four side-chain ligands arranged in a precise geometry. The introduction of such sites often requires a substantial amount of protein design [88; 89; 90], which is not required for the surface biHis sites.

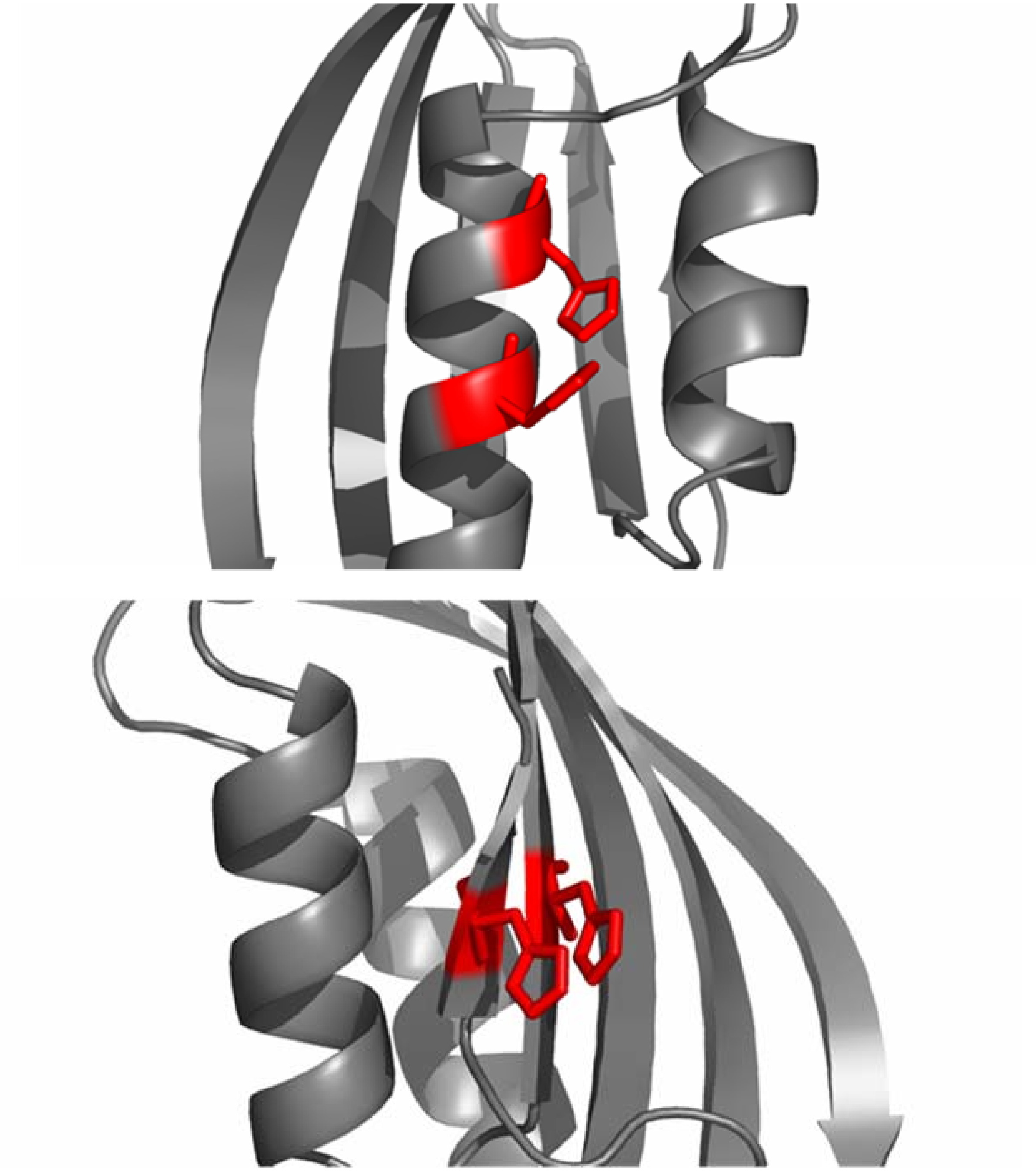
The placement of surface biHis metal binding sites can be accomplished by inspecting the protein structure. Generally, sites require the imidazole nitrogens (N_ε) of the histidines be located within 3-5 Å of each other, which can be accomplished using residues where the C_α-C_α distance is less than 13 Å [84]. The placement of histidine residues has a substantial effect on the binding affinity and the degree of energetic stabilization for each type of ion [84]. In helices, the histidines should be introduced four residues apart in *i* and *i+4* positions (HXXXH, SEE NOTE 2.6.1). Metal sites can be introduced across a hairpin or β-strand in either parallel or antiparallel orientation although site selection does not appear to be very stringent. For example, a histidine on one strand of a β-sheet an form a biHis site with residues on either adjacent strand (FIGURE 2.1). However, β-sheets often are quite twisted and care must be taken to use two positions where the side-chains are not angled away from each other. Previously, we have successfully replaced an inter-helical salt-bridge with a biHis site, suggesting that pre-existing side-chain interactions also may be good candidates for metal site replacement.

2.4.2 Equilibrium studies

The suitability of each biHis site must be investigated to verify that introduction of the site does not perturb the structure of the protein either in the presence or absence of metal. Verifying the lack of structural perturbation with and without metal ion using NMR methods is possible, albeit very time consuming. In most cases, surface histidine substitutions have been found to have little apparent effect on the structure or folding behavior of proteins, compared to the more common use of core residue substitutions in other folding studies. Nevertheless, the near-UV circular dichroism (CD) spectra can be measured with and without metal to confirm that no gross structural changes occur as a result of metal binding.

FIGURE 2.1 - *BiHis Metal Binding sites*

Examples of biHis metal-binding sites. (Top) A helical site with histidine mutations at *i*, and *i+4* positions across one helical turn in red, here shown in ctAcP. (Bottom) An example of a biHis site across a β-strand with histidine residues in red.



Equilibrium denaturant titrations should be conducted at several metal concentrations, including near-saturating concentrations (e.g. 1 mM). These data provide a number of useful quantities:

1. The energetic cost of inserting biHis mutation. Generally, biHis mutations are slightly destabilizing with reference to the wild type, usually $\Delta\Delta G_{eq}$ < 2 kcal mol^{-1}. This value is required later for correcting the ψ-value to account for the biHis mutation:

$$\Delta\Delta G_f = RT\ln\left(\left(1-\psi_o\right)+\psi_o e^{\Delta\Delta G_{bind}/RT}\right)$$

EQUATION 2.1

2. Change in denaturant m^o-value. This parameter, which reflects the amount of surface burial in the folding transition, should be largely unchanged upon the addition of metal. A decrease in the m^o-value may indicate the formation of residual structure in the denatured state, or a perturbation of the native structure.

3. Quantifying the maximal amount of metal-induced stabilization. The equilibrium change in free energy of metal binding under saturating conditions identifies the experimental limits of metal-induced stabilization and the sensitivity to minor pathways. Equilibrium values of $\Delta\Delta G_{bind}$ typically range from 0.5 – 3 kcal mol^{-1}. We have found that cobalt stabilizes α-helices to a greater extent while zinc and nickel tend to prefer β-sheet sites, although this correlation is inconsistent and several metals should be tested at the outset. (SEE NOTE 2.6.2)

The use of the most stabilizing ion increases the accuracy in which fractional ψ-values can be determined.

4. The value of $\Delta\Delta G_{bind}$ is to be compared to that obtained from chevron analysis to confirm that metal binding is in fast equilibrium during the kinetic measurements, a requirement for implementation of the method (SEE NOTE 2.4.5)

2.4.3 Chevron Analysis

Although ψ-analysis can be applied to more complicated reactions, for simplicity, its application is illustrated here with a kinetically two-state system. First a chevron [32] is measured on the biHis variant in order to compare the folding behavior of the mutant to that of the wild-type protein. The refolding and unfolding denaturant m_f- and m_u-values should be largely unchanged compared to the wild-type, signifying no major change in surface area burial during the course of the reaction. Just as in mutational ϕ-analysis, data from systems with changing *m*-values should be interpreted cautiously as the biHis substitution may have altered the conformation of the denatured, native, or transition states. From a comparison of the chevrons, a double-site $\phi^{\text{wt-biHis}}$ value can be calculated. However, the translation of this ϕ-value to structure in the TS may be difficult, as the perturbation due to the introduction of the biHis site often is unclear.

Additional chevrons should be measured under near-saturating metal conditions to determine if metal binding grossly alters the folding pathway of the biHis protein. (FIGURE 2.2) This experiment is performed as usual except with buffers that contain divalent cations (SEE NOTES 2.6.3, 2.6.4 AND TABLE 2.1). These kinetic studies can be used to calculate the maximal stabilization imparted by the binding of several different metal ions. Chevron data are acquired at a single high metal concentration for each cation, and the $\Delta\Delta G_{bind}$ is determined from the change in k_f and k_u (SEE NOTE 2.6.5). These values should match those obtained in the equilibrium measurements to confirm that cation binding is in fast equilibrium (SEE NOTE 2.6.5).

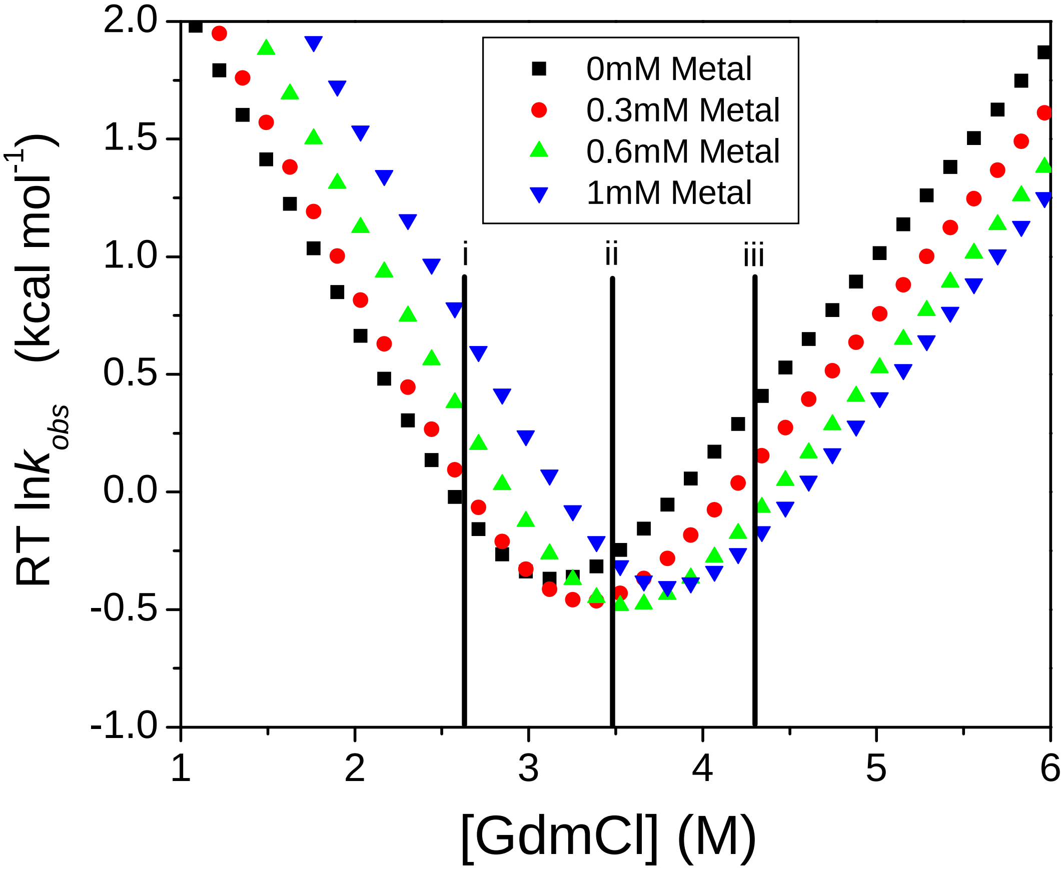
Once these chevrons have been acquired, a ψ_o-value can be estimated by comparing the chevron measured without metal to one at saturating metal concentration. If the high-metal chevron shifts only the folding arm up (folding is faster) while leaving the unfolding arm unchanged, then metal binding stabilizes both the TS and the native state equally and the ψ-value is unity. Conversely, if the presence of metal only shifts the unfolding rate down (unfolding is slower) then metal only stabilizes the native state, implying ψ~0. When both arms shift, the ψ-value is fractional. A ψ_o-value can be calculated using a two-point Leffler plot with data in the absence and presence of metal ion and fit to the single free parameter. More detailed metal-dependent folding measurements should be conducted to determine a more precise value of ψ_o.

2.4.4 Metal-dependent folding kinetics

Two strategies can be used to obtain more detailed information on the metal dependence of folding rates. Additional chevrons can be obtained, each at a fixed metal ion concentration traversing the range of interest (e.g. 0.0, 0.1, 0.2, 0.4, 1 mM [Me^{2+}]). (FIGURE 2.2) Alternatively, the denaturant concentration can be fixed, and numerous folding and unfolding measurements are conducted at multiple, finely-spaced metal concentrations (FIGURE 2.3). The folding measurements are conducted at a single denaturant concentration under strongly folding conditions (FIGURE 2.2, Line i), while the unfolding measurement is conducted at a denaturant concentration under strongly unfolding conditions (FIGURE 2.2) Line iii). When choosing the final denaturant concentration for the unfolding measurements, bear in mind that metal stabilizes the protein and shifts the chevron to the right. As a result, unfolding conditions in the absence of metal ions could become folding conditions with the addition of metal (FIGURE 2.2, Line ii). For this reason, it is advisable to measure a chevron under saturating metal conditions first so that appropriate folding and unfolding conditions can be chosen at the outset. This second strategy - varying metal concentration at a fixed denaturant value - effectively is a “vertical slice” through a multitude of denaturant chevrons in the presence of differing metal ion concentrations. The advantage of the second method is that many more points are obtained for the Leffler plot, which is used to calculate the ψ-value (FIGURE 2.4. The advantage

FIGURE 2.2 - *Chevron analysis as a function of metal concentration*

Kinetic analysis of biHis variant as a function of metal concentration. Sample chevron data representing the denaturant dependence of folding and unfolding rates. The left half of the curve represents a rapid jump from unfolding conditions to the denaturant concentration indicated on the x-axis (folding arm). The right half of the chevron represents the unfolding arm - kinetic relaxation rates under unfolding conditions. Chevron analysis of a biHis variant in the presence of 0 (square), 25uM (circle), 200uM (triangle), and 1mM metal ion concentration (star). The folding and unfolding rates in water are determined by extrapolating both arms to the y-axis, which change as a function of metal. The parallel arms indicate no change in the denaturant dependence (*m*-values) as a function of metal. The vertical lines indicate i) the final [GdmCl] where refolding experiments should be performed, iii) where unfolding should be performed, and ii) denaturant concentration apparently ideal for unfolding experiments, can become approach refolding conditions at higher metal ion concentrations



2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
1
2
3
4
5
6
RT lnk_{obs} (kcal mol^{-1})
[GdmCl] (M)
■ 0mM Metal
● 0.3mM Metal
▲ 0.6mM Metal
▼ 1mM Metal
i
ii
iii

of the first method is that one can monitor changes in the *m*-values as a function of metal ion concentration.The implementation of the second strategy depends on the nature of the equipment used. Our lab uses four-syringe Biologic SFM-4 and SFM-400 stopped-flow apparatuses (www.bio-logic.fr). The metal dependent folding measurements can be performed with this machine using a three-syringe mixing protocol where one syringe contains a protein/denaturant mixture. The two other syringes have identical denaturant concentration, but one syringe contains metal ions (TABLE 2.2). By varying the relative delivery volumes of these otherwise identical buffers, a large amount of finely-spaced data can be obtained over a range of ion concentrations using a single set of buffer solutions. A standard two

syringe, fixed volume ratio stopped-flow apparatus can be used as well, although the solutions must be changed for each metal ion condition.

Often stopped-flow apparatus have limited dilution ratios due to inaccuracies arising from low-volume delivery. In our experimental setup, one set of buffers provides data over approximately one decade in metal ion concentration. To cover a wider range, only the metal-containing buffer need be changed. This change readily can be accomplished using dilutions of the metal buffer using the no-metal buffer, e.g. lower ranges are obtained by five-fold serial dilutions of the metal-containing buffer, e.g. 5 mM, 1 mM, 200 μM, and 40 μM [Me^{2+}].

Once refolding data has been collected over the range of metal ion concentration, the process must be repeated for the unfolding measurements. The

TABLE 2.1 - *Sample shot protocol for a denaturant chevron at high metal ion concentration*

The metal concentration is designed to stay constant while the denaturant changes. For a three syringe protocol, Syringe 1 contains buffer, Syringe 2 contains 4 M GdmCl with buffer, and Syringe 3 contains Protein, 5 M GdmCl, 6 mM $MeCl_2$, and buffer.

		Buffer	4 M Gdm	Protein 4 M Gdm 6 mM Me^{2+}	
[Metal] μM	[GdmCl] M	Syringe 1 (μl)	Syringe 2 (μl)	Syringe 3 (μl)	Total vol (μl)
1000	0.83	250	0	50	300
1000	1.50	200	50	50	300
1000	2.17	150	100	50	300
1000	2.83	100	150	50	300
1000	3.50	50	200	50	300
1000	4.17	0	250	50	300

FIGURE 2.3 - *Metal-dependent folding kinetics "Metal Chevron"*

Kinetic folding and unfolding data as a function of divalent metal ion concentration. The circles represent the kinetic response of rapid folding measurements with a final concentration of 2.5 M GdmCl as the concentration of metal is increased from 0 to 1mM. Squares represent unfolding data taken at 4.5 M GdmCl final denaturant concentration.

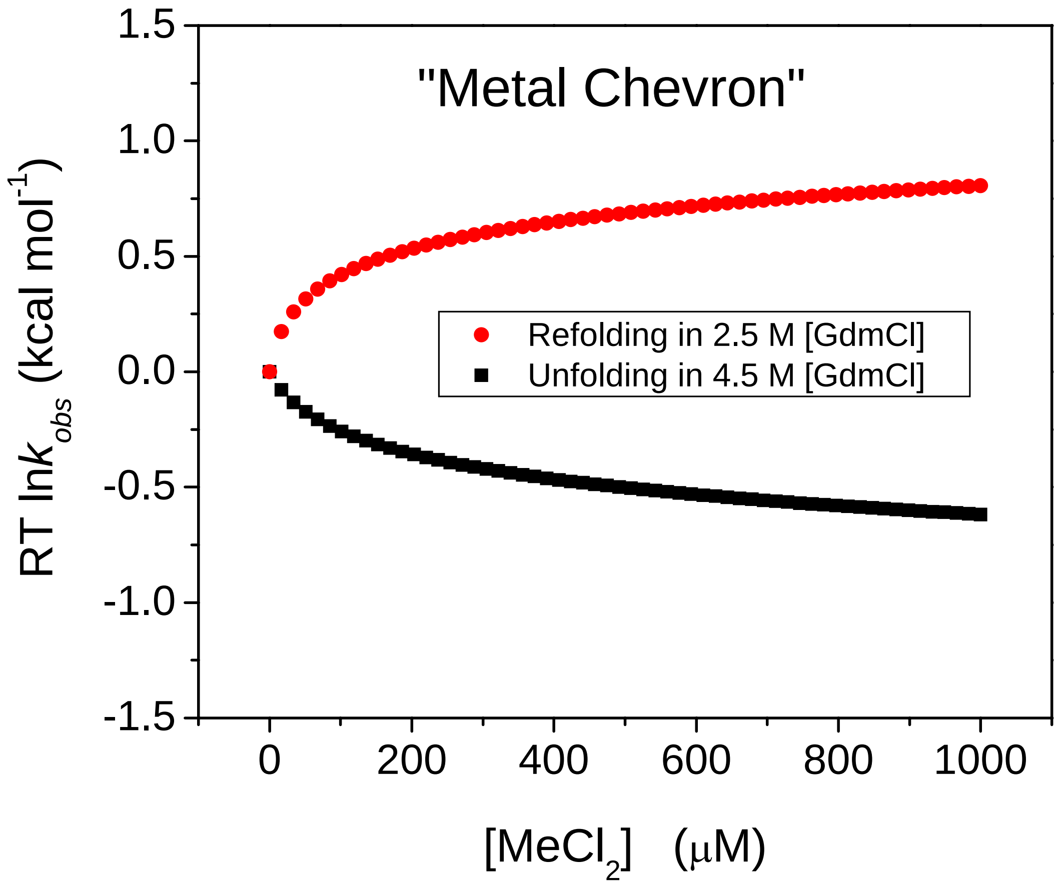
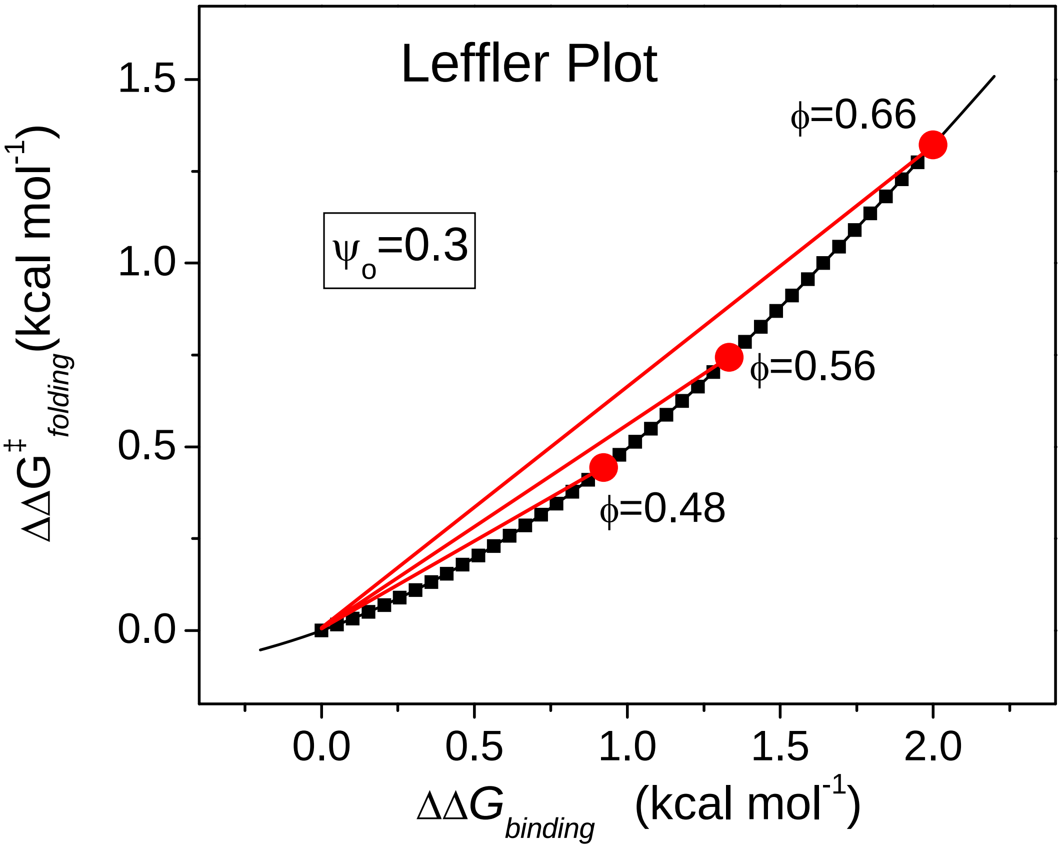

"Metal Chevron"
RT lnk_{obs} (kcal mol^{-1})
[MeCl$_2$] (μM)
● Refolding in 2.5 M [GdmCl]
■ Unfolding in 4.5 M [GdmCl]
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0
200
400
600
800
1000

FIGURE 2.4 - *Leffler Plot*

Leffler plot. Squares indicate change in free energy of folding as a function of binding energy. Each metal chevron data at a different metal ion concentration yields one data point. The parameter ψ_o is equal to the instantaneous slope at the origin (here shown with a value of 0.3) and is determined from fitting the data to the functional form given in Eq. 5. The slope approaches one with sufficiently high amount of stabilization. This plot is analogous to traditional ψ-analysis plots, although free energy relationships are generally assumed to be linear. The circles represent Leffler data generated from changes in relative free energy of folding and unfolding using denaturant chevrons at three different metal concentrations. These points can generate apparent ψ-values using the two-point fit from the origin, although underlying curvature can lead to systematic errors in ψ-value results.


Leffler Plot
ψ$_o$=0.3
φ=0.66
φ=0.56
φ=0.48
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
ΔΔG$^‡_{folding}$(kcal mol^{-1})
ΔΔ*G*$_{binding}$ (kcal mol^{-1})

data should be taken at the same metal concentrations as the folding data to allow calculation of $\Delta\Delta G_{eq}$ from the difference of $\Delta\Delta G_f$ and $\Delta\Delta G_u$ according to

$$\Delta\Delta G_{bind}([\mathrm{Me}^{2+}]) = \Delta\Delta G_f - \Delta\Delta G_u = RT \ln (k_f^{\mathrm{M2+}}/ k_f) - RT \ln (k_u^{\mathrm{M2+}}/ k_u)$$

EQUATION 2.2where R is the gas constant and T is the absolute temperature, $k_f^{\mathrm{M2+}}$ and $k_u^{\mathrm{M2+}}$ are the relaxation rates in the presence of the same concentration of cation, respectively. The folding rate at zero metal should also be well established as this rate serves as the reference point in the Leffler plot.

2.4.5 Testing for fast equilibrium

If the folding rates are fast compared to metal binding or release, then binding may not be in fast equilibrium during the course of the folding reaction. As a result, the folding and binding processes will be convoluted. For fast folding rates, metal ions may even no longer stabilize the TS as the biHis site is kinetically inaccessible for ion binding. Furthermore, if metal release rates are slower than unfolding rates, multiple populations will be observed in unfolding experiments; some molecules will unfold having metal bound while other molecules will unfold as if there is no metal ion present in solution. When binding is no longer in fast equilibrium, the ψ-analysis formalism will no longer be valid.

A required signature of metal binding being in fast equilibrium is the lateral translation of chevrons with increasing metal ion concentration. There should be no changes in the slope or kinks in the arms of plot, or additional kinetic phases. An additional test to confirm fast equilibrium is that the metal-induced stabilization determined from the equilibrium studies should match that from kinetic measurements. If metal binding is not in fast equilibrium, different metals can be tested as their binding properties may be more suitable. Also, folding rates can be manipulated by working at lower temperatures or higher denaturant concentrations.

2.5 Data Analysis

2.5.1 Equilibrium denaturation profiles

The folding transitions in the presence of metal ions are fit to the standard equation assuming a two-state equilibrium between the N(ative) and U(nfolded) states:

$$\mathrm{S}([den]) = \frac{\mathrm{S_U} + \mathrm{S_N}\mathrm{e}^{-(\Delta G + m[\mathrm{den}])/RT}}{1 + \mathrm{e}^{-(\Delta G + m[\mathrm{den}])/RT}}$$

EQUATION 2.3

where $\mathrm{S_U}$ and $\mathrm{S_N}$ are the signals of the U and N states, respectively. Parameters are fit using a non-linear least-squares algorithm (e.g. Microcal Origin software package).

2.5.2 Metal stabilization

The degree of stabilization due to metal binding, $\Delta\Delta G_{bind}$, depends upon difference in metal dissociation constants between the biHis sites in the native state, K_N, and in the unfolded state, K_U (FIGURE 2.5) the increase in protein stability upon the addition of metal is fit to a linked equilibrium expression [91]

$$\Delta\Delta G_{bind}(\mathrm{M}^{2+}) = RT\,\ln(1+[M^{2+}]/K_N) - RT\,\ln(1+[M^{2+}]/K_U)$$

EQUATION 2.4

2.5.3 Kinetic Chevron analysis

The kinetic data are analyzed using chevron analysis of the denaturant dependence of folding rate constants [32], where the standard free energy of folding, ΔG_{eq}, along with the standard activation free energy for folding, $\Delta G^{\ddagger}{}_{f}$, and unfolding, $\Delta G^{\ddagger}{}_{u}$, are linearly dependent on denaturant concentration

$$\Delta G_{eq}\,([\mathrm{Den}]) = \Delta G_{eq}{}^{\mathrm{H2O}} + m^{\mathrm{o}}[\mathrm{Den}] = -\,RT\ln \mathrm{K_{eq}} \qquad (2.5a)$$

$$\Delta G_{f}\,([\mathrm{Den}]) = -\,RT\ln \mathrm{k}_{f}{}^{\mathrm{H2O}} - m_f\,[\mathrm{Den}] + constant \qquad (2.5b)$$

$$\Delta G_{u}\,([\mathrm{Den}]) = -\,RT\ln \mathrm{k}_{u}{}^{\mathrm{H2O}} - m_u\,[\mathrm{Den}] + constant \qquad (2.5c)$$

EQUATION 2.5

The denaturant concentration dependencies (*m*-values) report on the degree of surface area burial during the folding process [35]. When equilibrium and kinetic folding reactions are two-state and are limited by the same activation barrier, the equilibrium values for the standard free energy and surface burial can be calculated from kinetic measurements according to ΔG_{eq}=ΔG_f -ΔG_u and m^{o} = m_u - m_f.

2.5.4 Obtaining ψ-values from the Leffler plot

The Leffler [77] plot is obtained from the change in the folding activation energy relative to the change in stability due to metal ion binding. (FIGURE 2.4) $\Delta\Delta G_f$ is derived from the ratio of the folding rate in the presence ($k_f^{\mathrm{M2+}}$) and absence of metal (k_f) using $\Delta\Delta G_f = RT \ln (k_f^{\mathrm{M2+}} / k_f)$. $\Delta\Delta G_{bind}$ can be obtained from the equilibrium data **(**EQUATION 2.4) or from the change in the folding and unfolding rates taken at identical metal ion concentrations. When multiple chevrons are obtained, each at a different metal ion concentration, one Leffler data point is obtain for each chevron. When folding and unfolding data are obtained at fixed denaturant concentrations at varying metal ion concentration, one Leffler data point is obtained for each metal concentration.

The plot will be either linear or curved depending upon the degree of structure formation and heterogeneity in the TS. In either case, the data can be fit with the same single parameter equation (EQUATION 2.1) where ψ_0 is the instantaneous slope at the origin where data is obtained in the absence of metal ions. The instantaneous slope (the ψ-value) at any point on the curve as a function of binding stability is given by

$$\psi = \frac{\partial \Delta\Delta G_f}{\partial \Delta\Delta G_{bind}} = \frac{\psi_o}{(1-\psi_o)e^{-\Delta\Delta G_{bind}/RT} + \psi_o}$$

EQUATION 2.6

2.5.5 Interpreting ψ-values

Once the value for ψ_0 has been determined, the next task involves understanding its significance. The interpretation generally is clear in the two cases where the plot is linear, $\psi_0 = 0$ or 1. For a value of unity, the biHis site is present (native-like) in the TS ensemble. For a value of zero, the site is absent (unfolded-like). In other cases, the Leffler plot will be curved as ligand binding continuously increases the stability of the TS ensemble [92]. The curvature can be due to TS heterogeneity, non-native binding affinity in a singular TS (D. Goldenberg, private communication), or a combination thereof. (see also Fersht [93] for a comparison of the ψ and ϕ analysis methods using an alternative kinetic model which focuses on unfolded state population shifts while omitting any consideration of TS binding).

The proper interpretation of fractional ψ-values involves an appreciation of the mathematical formalism behind Equations 5 and 6. ψ-analysis takes into account the shifts in the native, unfolded and TS state populations due to binding of the metal ion to each of these states. Folding rates are calculated assuming two classes of TSs depending on whether the biHis site is present ($k^{present}$) or absent (k^{absent}). In the first class, $\mathrm{TS}^{\mathrm{present}}$, the biHis site is present in a native or near-native geometry with a dissociation constant $K_{TS}^{present}$. In this case the associated backbone structure is folded, for example in a helical or β-sheet conformation. In the second class (TS_{absent}) the biHis site is essentially absent but may be considered to have an effective dissociation constant K_{TS}^{absent}, just as the unfolded state is allowed to bind metal with an dissociation constant K_U (EQUATION 2.4). As a result, the model contains two TS's, each having their own effective binding affinities, $K_{TS}^{present}$ and K_{TS}^{absent} .

As per Eyring Reaction Rate Theory [94], the over-all reaction rate is taken to be proportional to the relative populations of the TS and U ensembles, $k_f \propto$ [TS]/[U]. The net folding rate is the sum of the rates going down each of the two routes, $k_f = k^{present} + k^{absent}$ or

$$k_f \equiv k_o^{present} \frac{1+[M]/K_{TS}^{present}}{1+[M]/K_U} + k_o^{absent} \frac{1+[M]/K_{TS}^{absent}}{1+[M]/K_U}$$

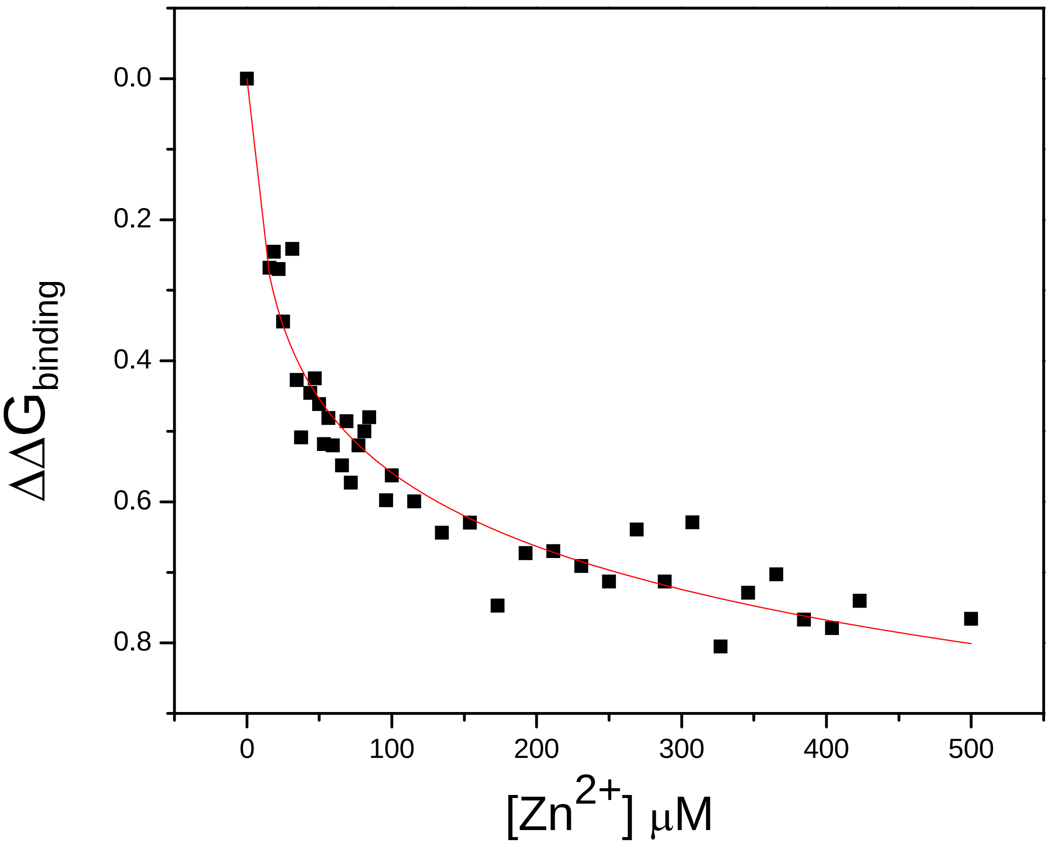
EQUATION 2.7

where $k_o^{present} \propto [TS_{present}]/[U]$ and $k_o^{absent} \propto [TS_{absent}]/[U]$ are the rates through each TS class prior to the addition of metal, with a ratio $\rho_o = k_o^{absent} / k_o^{present}$. By examining shifts in populations and assuming metal binding is in fast equilibrium (SEE SECTION 3.3.1), this treatment avoids any assumptions about possible pathways connecting each of the different bound and unbound states.

There are two scenarios where the Leffler slope is linear. In the first scenario, the slope is zero across all metal concentration (ψ=ψ_o=0). This behavior occurs when metal ion binding does not increase the population of the biHis site in the TS ensemble relative to U. In this case, the entire TS ensemble lacks the binding site, or more rigorously, the site has the same binding affinity as the unfolded state.

FIGURE 2.5 - *Metal induced equilibrium stabilization*

Increase in equilibrium stabilization as a function of divalent metal metal-ion concentration. In this sample site, $ZnCl_2$ is titrated into solution in an equilibrium experiment and the protein is stabilized by the change in free energy of binding $\Delta\Delta G_{bind}$.


0.0
0.2
0.4
0.6
0.8
$\Delta\Delta G_{binding}$
0
100
200
300
400
500
$[Zn^{2+}]$ μM

At the other limit, the slope is one (ψ=ψ_o=1) indicating that the entire ensemble has the binding site formed in a native-like manner. Otherwise, the Leffler plot will be curved as metal continuously increases the population of the biHis site in the TS ensemble. In such cases the curvature can be due to TS heterogeneity, non-native binding affinity in a singular TS, or a combination thereof. In the heterogeneous scenario, one can consider the simplified situation (FIGURE 2.6), where $TS_{present}$ has the biHis site present with native-like affinity ($K_{TS}^{present} = K_N$) while TS_{absent} has the site with the unfolded-like affinity ($K_{TS}^{absent} = K_U$). Here only the $TS_{present}$ state is stabilized upon the addition of metal ion. The height of the kinetic barrier associated with $TS_{present}$ decreases to the same degree as does the native state, $k^{present} = k_o^{present} e^{\Delta\Delta G_{eq}/RT}$. The instantaneous slope simplifies to the fraction of the TS ensemble which has the biHis site formed at a given metal ion concentration:

$$\psi = \frac{k^{present}}{k^{present} + k_o^{absent}}$$

EQUATION 2.8

In the simplified situation, the degree of pathway heterogeneity prior to the addition of metal ions is given by ψ_o, the slope at zero stabilization. The Leffler plot exhibits upward curvature as the ψ-value increases with added metal binding energy, which indicates the increasing fraction of the TS ensemble with the biHis site present. Generally, ψ-values continuously vary between 0 and 1 at the limits of infinite TS stabilization. When the ψ-value is 0.5, the site is formed half the time in the TS ensemble.

Curvature can also occur in a homogeneous scenario when the singular TS has non-native binding affinity in the TS. When the site has non-native binding affinity in the TS ($K_{TS}^{present} \neq K_N$), but maintains U-like affinity in the unfolded state ($K_{TS}^{absent} = K_U$), the initial slope is the degree of heterogeneity multiplied by an additional factor representing the differential binding affinity between $TS_{present}$ and N

$$\psi_o = \frac{K_N}{K_{TS}^{present}} \frac{K_{TS}^{present} - K_U}{K_N - K_U} \frac{k_o^{present}}{k_o^{present} + k_o^{absent}}$$

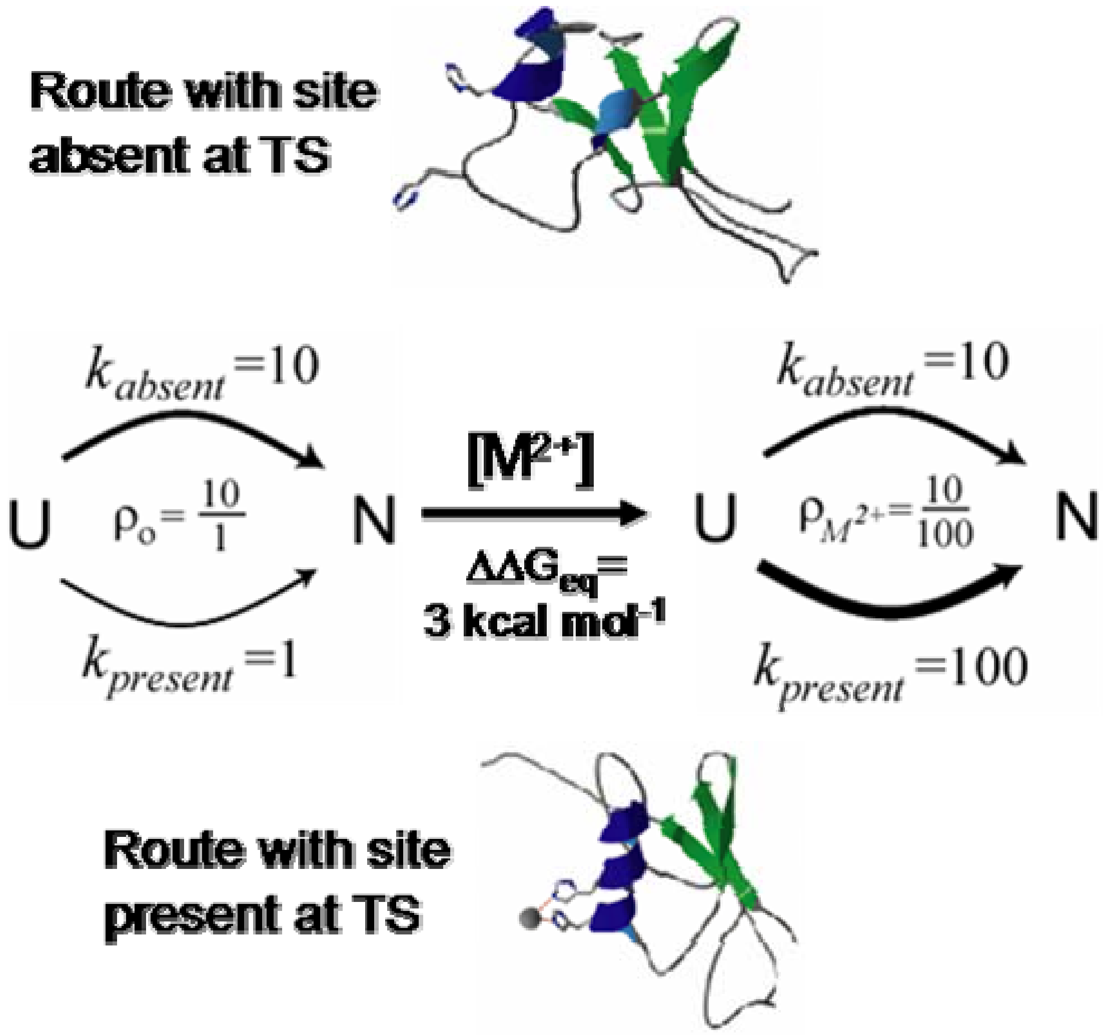
EQUATION 2.9

Here, the curvature reflects the stabilization of the single TS relative to U. Where the initial slope is $\psi_o = \frac{K_N (K_{TS} - K_U)}{K_{TS}(K_N - K_U)}$. It is important to note that for a homogeneous TS, the aforementioned interpretation of the two linear Leffler behaviors, ψ = 0 or 1, remains unchanged.

The analysis of metal binding presented here is slightly different than that presented in our earlier papers [1; 75] where curvature was associated only with the heterogeneous model. With the explicit inclusion of the binding affinities in the TS, $K_{TS}^{present}$ and K_{TS}^{absent}, the *f*-value ($f \equiv \Delta\Delta G_f / \Delta\Delta G_{eq}$) is no longer required.

FIGURE 2.6 - *Heterogeneous vs. homogeneous situations*

Application of ψ-analysis to a two route scenario with a helical site with native binding affinity which is formed on 9% of the pathways prior to addition of metal. The absent route contains a TS that has the same binding affinity as the U state. The folding rate for the route with the biHis site present ($k^{present}$, lower pathway) increases from 1 to 100 upon the addition of 2.86 kcal mol^{-1} of metal ion binding energy at 20 °C. This enhancement increases the flux down the metal ion stabilized route relative to all other routes (k^{absent}), from $\rho_o = k^{absent} / k_o^{present} = 10/1$, to metal-enhanced condition, $\rho_M = 10/100$. The corresponding ψ-values increase from $\psi_o = 0.1$ to $\psi_M = 0.9$. The binding energy required to stabilize a TS and switch a minor route to a major route identifies the barrier height for this route relative to that for all other routes. Reprinted from Sosnick *et al* 2004 [38].


Route with site absent at TS
k_{absent}=10
U
$\rho_o = \frac{10}{1}$
N
$k_{present}$=1
[M^{2+}]
$\Delta\Delta G_{eq}$=
3 kcal mol^{-1}
k_{absent}=10
U
$\rho_{M^{2+}} = \frac{10}{100}$
N
$k_{present}$=100
Route with site present at TS

It is generally not constant as it depends on metal ion concentration, except in the two linear scenarios where *f*=0 or 1.

2.5.6 Correcting for the effects of the biHis site

The introduction of the biHis substitution itself alters the stability of the native state by the amount $\Delta\Delta G_{eq}^{biHis}$. In the simplified, heterogeneous scenario where $K_{TS}^{present} = K_N$ and $K_{TS}^{absent} = K_U$, the ψ_o value should be corrected in order to account for this change in stability

$$\psi_o^{corr} = \frac{\psi_o}{\psi_o + e^{-\Delta\Delta G_{eq}^{biHis} / RT}(1-\psi_o)}$$

EQUATION 2.10

The resulting ψ_o^{corr} is the instantaneous Leffler slope at which the metal ion binding energy is exactly offset by the change in stability due to the biHis substitution. This correction is justified because both metal binding and the biHis substitution affect the same region of the protein. With this correction, the ψ-values for several different biHis variants can be combined to construct an accurate representation of the TS ensemble appropriate for the wild-type protein prior to mutation or metal binding.

2.5.7 Delineation between heterogeneous and homogeneous scenarios

The question of whether fractional ψ-values reflect TS heterogeneity or non-native binding affinity remains unresolved, and may be site dependent. We believe discrimination between these two models is possible through study of folding behavior of two different metal ions which have different coordination geometries. The two ions are likely to manifest the same ψ_o-value only in the case of TS heterogeneity, because the same fractional binding affinity is unlikely to be realized with both ions. However, ψ_o-values should depend on the type of metal ion if the site is distorted.

Another test for heterogeneity involves altering the relative stability of the TS structure with the site present, e.g. via mutation far from the biHis site. If the ψ-value responds accordingly, as observed in the dimeric GCN4 coiled coil [75], the heterogeneity model is the most parsimonious. For the coiled coil, the introduction of the destabilizing glycine (A24G) shifted the pathway flux away from this region so that most nucleation events occurred near the biHis site, which was located at the other end of the protein. As expected, the ψ_o value increased to 0.5, indicating that half of the nucleation events occurred with the biHis site formed.

A quantitative comparison indicated that the change in the degree of pathway heterogeneity recapitulated the destabilizing effect of the glycine substitution. The A24G mutation increased the amount of flux going through the N-terminal biHis site. The ratio of the heterogeneity in these two molecules reflected the loss in stability for this mutation, $\Delta\Delta G_{eq} = RT \ln(\rho_{\mathrm{Ala}}/\rho_{\mathrm{Gly}}) = 2.5$ kcal mol^{-1}. This shift was consistent with the decrease in stability for the mutation backgrounds (1.7- 2.4 ± 0.1 kcal mol^{-1}) [45]. Hence, in this case, ψ-analysis successfully quantified the level of TS heterogeneity. A homogeneous model with non-native binding affinity in the TS would require that the A24G mutation causes the biHis the site to acquire native-like binding affinity. This is an unlikely scenario given the distance between the substitution and the biHis site. Potentially, binding sites introduced into well-defined helices will have native-like binding affinities in the TS. In which case, fractional ψ-values will be due to TS heterogeneity.

2.6 Notes

2.6.1 Note 1

BiHis sites can be mutated sequentially using a common method, the QuikChange protocol from Stratagene. However, engineering a biHis site into a helix is possible in one step using a single primer with both mutations encoded. This strategy places the mutagenic codons nine nucleotides apart with 10-15 complementary residues on either end, resulting in a very long mutagenic primer (>40nt). Engineering in two point mutations with one step does save time but bear in mind the PCR reaction is less likely to be successful. When using this approach, it is best to lower the annealing temperature 5-10 degrees to improve the reaction efficiency.

2.6.2 Note 2

To quickly test the amount of stabilization imparted by each type of metal ion, the protein can be placed in a denaturant solution where ~20% of the molecules are folded (K_{eq}=[N]/[U]=1/4). This level of denaturant can be obtained from a denaturation profile. The addition of high concentrations of metal will renature a fraction of the molecules according to the degree of metal-induced stabilization. From the increase in the equilibrium constant, K'_{eq}, the stabilization can be calculated $\Delta\Delta G_{bind}$=- RT ln (K'_{eq}/K_{eq}),

2.6.3 Note 3

Metal ions can be introduced into the folding reaction by including them in the syringe containing the protein solution, taking into account the dilution factor of the final mix (TABLE 2.2).

2.6.4 Note 4

The addition of metal can stabilize a protein to the point where higher than convenient levels of denaturant are required to unfold the protein. Rather than adding the metal only to the protein solution, the same experiment can be performed with metal ions in all buffers at the desired concentration. Alternatively, the non-protein buffers both can contain metal at a concentration calculated to give the desired final value. The method suggested tends to allow for the reuse of denaturant buffers in other experiments and reduces waste.

2.6.5 Note 5

Several different cations can be readily tested using a configuration where metal ion is placed only in one syringe. The testing of different metal ions using this protocol requires only the changing the contents of the single metal-containing syringe. However, metal concentrations in this syringe will be higher than if the ion was placed in all syringes, and precipitation may become an issue.

2.7 Conclusion

The application of ψ-analysis can provide detailed, site-resolved information on TS structures of protein folding pathways as well as other conformational transitions. The use of this method only requires introduction of biHis sites on the surface of the protein and metal-dependent kinetic measurements, both of which are relatively undemanding. For the two limiting situations, ψ_o= 0 or 1, the region of the protein where the biHis site is introduced is either unfolded or folded, respectively. Fractional ψ-values indicate the biHis site is either fractionally populated and/or distorted with non-native binding affinity in the TS. With the introduction of sufficient number of biHis sites, the topology of the entire TS structure can be identified. When combined with mutational studies, modeling, and other information, a complete picture of the TS structure(s) can be determined.

TABLE 2.2 - *Sample shot protocol for measuring metal-dependent folding*

The shot protocol is designed to vary metal concentration while keeping denaturant concentration constant. Syringe 1 contains buffer, Syringe 2 contains buffer with 1 mM MeCl$_2$, and Syringe 3 contains Protein in 5 M GdmCl.

		0 M Gdm	0 M Gdm + 1 mM Me^{2+}	Protein 5 M Gdm	
[Metal] μM	[GdmCl] M	Syringe 1 (μl)	Syringe 2 (μl)	Syringe 3 (μl)	Total vol (μl)
0	0.83	250	0	50	300
166.7	0.83	200	50	50	300
333.3	0.83	150	100	50	300
500	0.83	100	150	50	300
666.7	0.83	50	200	50	300
833.3	0.83	0	250	50	300

3.0 Pathway Heterogeneity in the Transition State of Ctacp

3.1 Abstract

The identification of folding pathway multiplicity and transition state structures is a subject of much theoretical and experimental investigation. The transition state of common-type acyl phosphatase (ctAcP) is characterized using ψ-analysis. ψ-analysis identifies chain-chain contacts using engineered bi-histidine metal ion binding sites located throughout the protein. The majority of the protein is structured in the transition state with the exception of one of the five β-strands and one of the two helices. ψ-values are near zero or unity for all sites except the one site on the amino end of the structured helix, which has a fractional ψ-value of 0.34, providing the only indication of transition state heterogeneity. This ψ-value remains unchanged when multiple metals having varying coordination geometries are used, indicating this end of the helix undergoes fraying. As with ubiquitin [1], the other globular proteins extensively characterized using ψ-analysis, the transition state ensemble has a single consensus structure. Hence, the folding pathways of both proteins have essentially converged to a single transition state structure, albeit one which contains a minor amount of fraying around the periphery.

3.2 Introduction

Most small globular proteins fold in a kinetically two-state reaction (U↔N) where intermediates do not populate [14; 71; 72; 73]. However, even such simple kinetic behavior does not preclude the existence of multiple transition states (TSs). A meaningful discussion of TS heterogeneity requires the level of detail be identified. For example, at residue-level resolution, undoubtedly heterogeneity exists in the TS ensemble, just as it does in the native state, as evidenced by methods such as hydrogen exchange. Heterogeneity may be due to small-scale, Boltzmann-distributed conformational excursions, such as helical or sheet fraying involving a few residues. Hence, when considered at this fine level, TS heterogeneity can be considered a general phenomena.

It is useful to distinguish between this type of heterogeneity and one where members of the ensemble are significantly different at the level of structural elements, for example the presence or absence of entire helices or strands. Further, even at this more coarse level, it is important to distinguish between a scenario where the entire TS ensemble shares some common elements from one where there are subpopulations which are structurally disjoint. (FIGURE 3.1)

We believe this distinction is the most appropriate when discussing whether multiple pathways exist in protein folding. Minimal experimental evidence exists to support the existence of multiple pathways according to this more stringent definition. An exception is the dimeric GCN4 coiled coil which has nucleation sites throughout the molecule, although its cross-linked counterpart nucleates only at the tethered end [45; 75].

Previously, ubiquitin (Ub) was found to fold through a TS ensemble containing a large, consensus four-strand sheet and a helix. (FIGURE 3.2) Nevertheless, evidence existed that regions on the periphery of this structure were fraying in fast equilibrium. Hence, when considered at fine resolution, TS heterogeneity exists; however, when considered at the more meaningful level of structural disjointedness, only a single TS is demonstrably present.

Characterization of the GCN4 coiled coil and ubiquitin TS ensembles were accomplished using ψ-analysis [1; 38; 75]. Here, engineered bi-Histidine (biHis) metal ion binding sites are introduced at known positions on the protein surface to stabilize secondary and tertiary structures. The introduction of an increasing concentration of metal ions serves to stabilize the interaction between the two histidine partners in a continuous fashion. As a result, this method quantitatively evaluates the metal-induced stabilization of the TS relative to the native state, represented by the ψ-value. The translation of a measured ψ-value to structure formation is straightforward, because the proximity of two known positions is probed. Hence, the method is particularly well-suited for defining the topology and structure of TS ensembles. The mutational counterpart of this method, ϕ-analysis, reports on energetic effects of side-chain alteration, and can seriously under-report the amount of structure in the TS [1; 38; 75].

In order to better define the folding transition state and determine the degree of heterogeneity on the folding landscape, we apply ψ-analysis to human erythrocyte common-type acyl phosphatase (ctAcP) [95; 96]. The protein is a 97-residue muscle isozyme homolog composed of two helices packed against a fivestranded antiparallel sheet in an "open sandwich" fold [97; 98; 99]. (FIGURE 3.3) With a 1/3 more residues and one additional α-helix with respect to Ub, this system serves as an appropriate progression of our investigations into the existence of structurally disjoint TSs.

The folding pathway of ctAcP was previously studied using ϕ-analysis with multiple, simultaneous helical mutations [100]. These studies found that Helix 2 (residues 54-67) is formed in the TS while Helix 1 (residues 22-33) is unstructured. Small-angle scattering studies found that hydrophobic collapse does not occur prior to the rate-limiting step [101], as expected for a protein which folds in a kinetically two-state manner.

The ψ-analysis results from nine biHis sites located on the surface of ctAcP indicate the transition state ensemble has a consensus structure with native-like topology. Further, the ψ-values are typically near zero or one, indicating only a minimal degree of heterogeneity in the ensemble. (FIGURE 3.4) The sole exception is a ψ-value of 0.3-0.4 at one end of Helix 1, observed with multiple metal ions. This behavior is consistent with helical fraying in the ensemble, with the ψ-value representing the fraction of time the metal ion binding site adopts a native-like geometry. Therefore this protein, like Ub, folds via a native-like, consensus TS structure and provides no evidence for structurally-disjoint pathways.

FIGURE 3.1 - *Classes of pathway heterogeneity*

A description of the two types of pathway heterogeneity implied by fractional ψ- or ϕ-values. (Left) Folding may occur via a singular homogeneous TS with partial interactions or partially formed structures in the TS. White boxes represent unfolded residues, red boxes indicate residues absolutely required in a given nucleus, and orange either when involved in a fractional side-chain interaction (partly shaded orange) or when optional to the folding nucleus (completely shaded orange). (Middle) Folding may also occur through a structurally heterogeneous ensemble where some residues are critical for the folding nucleus in all cases (red) but different groups of optional structures may also exist at the TS. Here the TS exhibits some small amount of heterogeneity, but the folding nucleus is the same in all cases, generally speaking. (Right) Transition state nuclei may alternatively be comprised of structurally disjoint folding cores, employing multiple groups of TS structures which can be separated by distinct obligate TS residues (red) which would demonstrate true pathway heterogeneity. Adapted from Krantz *et al* [1].

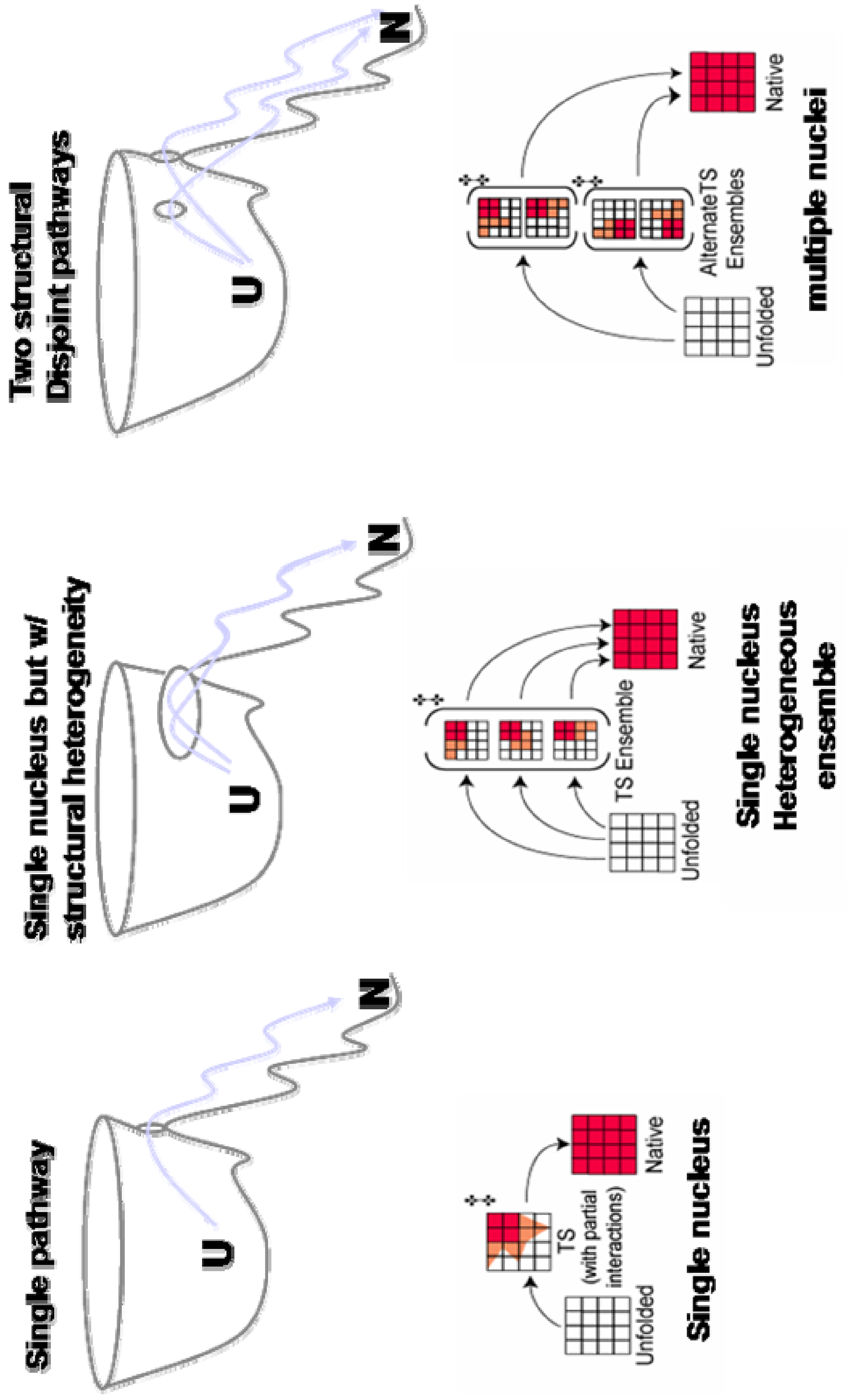
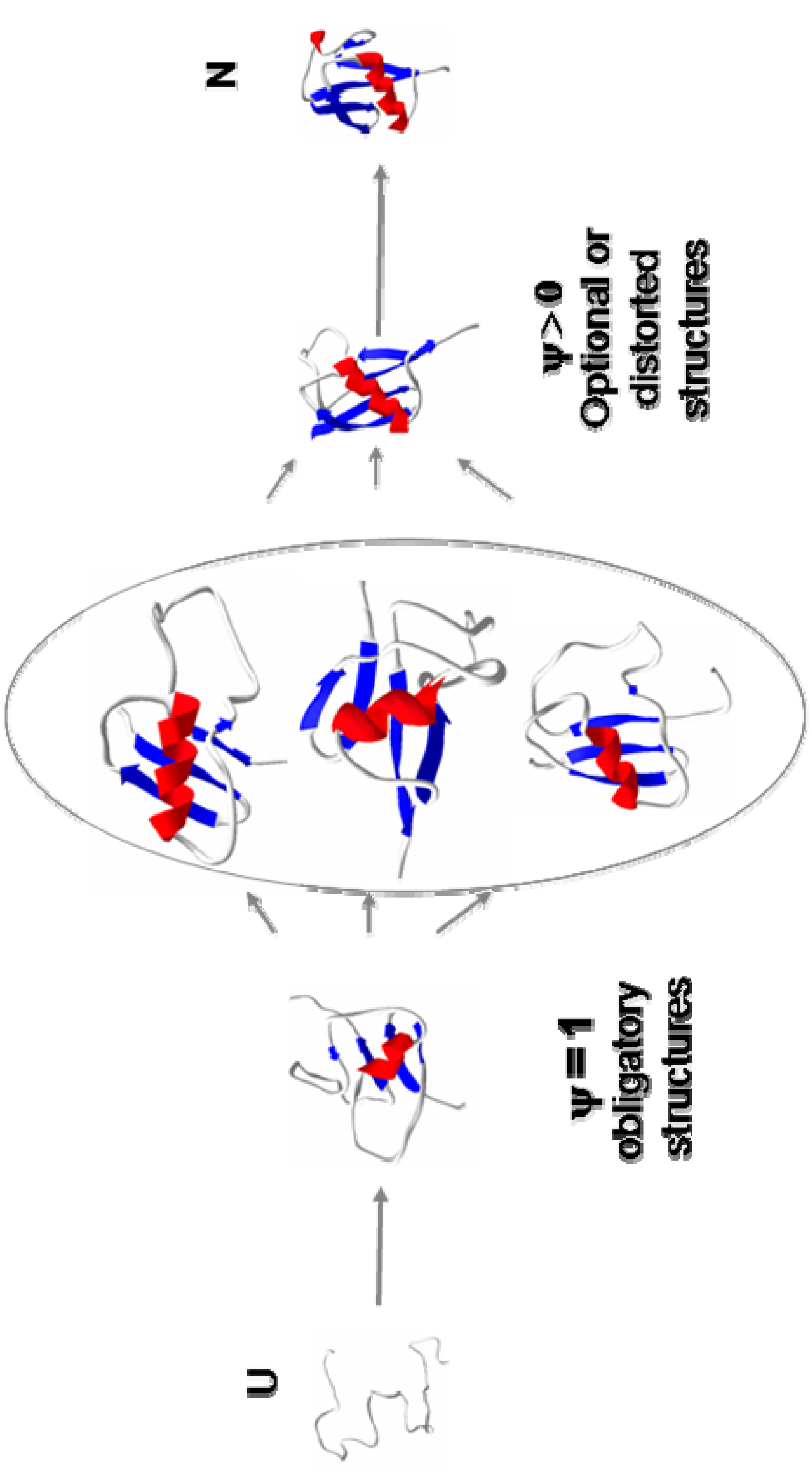

Single pathway
U
N
Unfolded
TS
(with partial
interactions)
‡
Native
Single nucleus
Single nucleus but w/
structural heterogeneity
U
N
Unfolded
TS Ensemble
‡
Native
Single nucleus
Heterogeneous
ensemble
Two structural
Disjoint pathways
U
N
Unfolded
AlternateTS
Ensembles
‡
‡
Native
multiple nuclei

FIGURE 3.2 - *Transition state ensemble in ubiquitin*

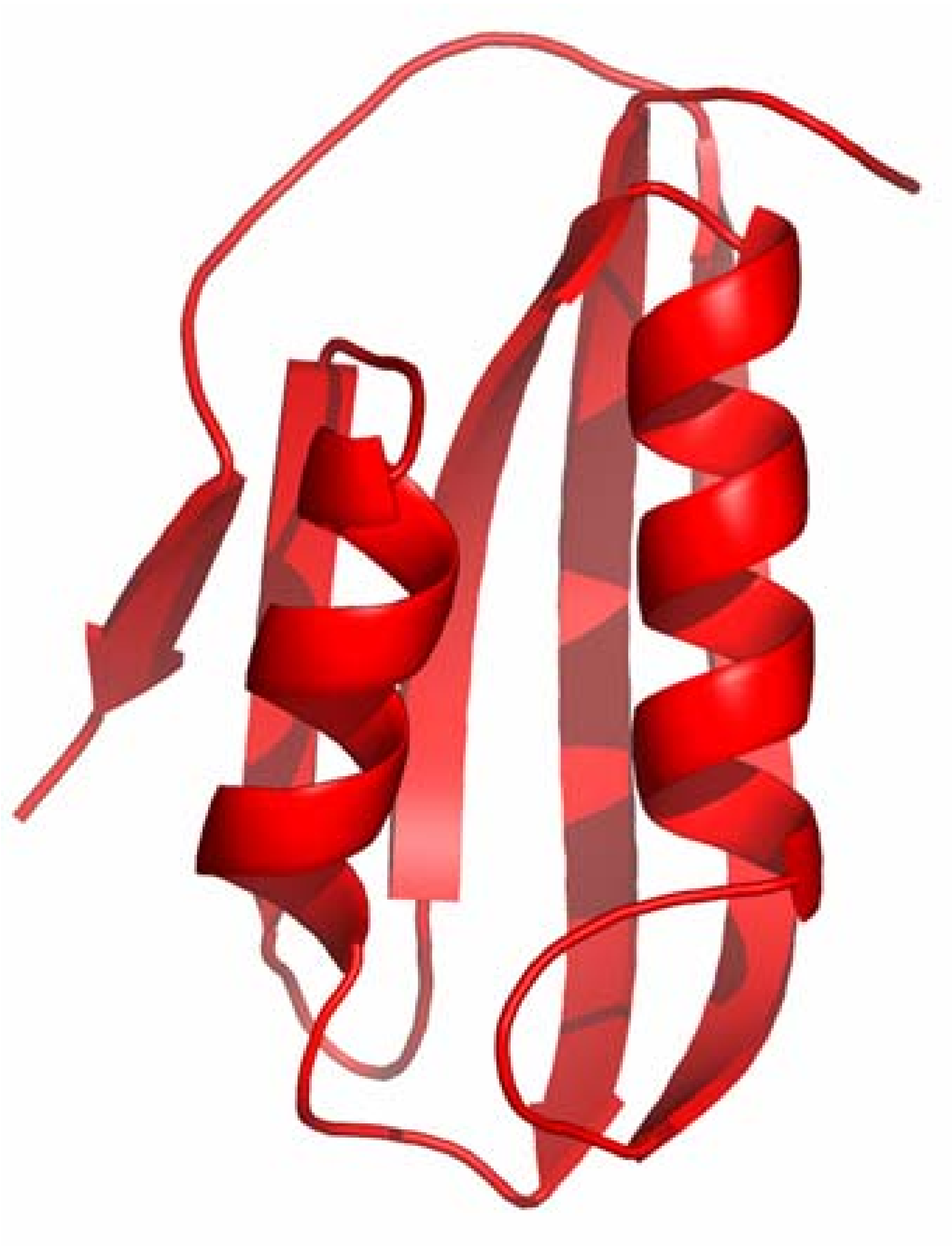
Ub's folding pathway is best described by a heterogeneous TS ensemble that emerges from a conserved nucleus. The TS ensemble along with the pre-TS and post-TS structures are identified using ψ-analysis. The initial step involves formation of the local B1–B2 hairpin, which populates native geometries at a low level, <20% [102]. Some NMR resonances change when the hairpin is extended to include residues in the helical region [102]. As the helix possesses low intrinsic helicity (<3% [103]), B1–B2 hairpin formation probably precedes helix formation. Strand B3 is adjacent to the nascent hairpin-helix nucleus and forms prior to B4. Beyond this pre-TS structure, the nucleus spreads in a number of possible directions reflecting the TS heterogeneity. The post-TS structure likely contains all optional TS elements and lacks only the one turn 3_{10} helix (H2) and strand B5. Hydrogen exchange data on native Ub [18], which report on the stability of hydrogen bonds, indicate that the 3_{10} helix likely folds before B5 on the major refolding pathway.



Krantz et al (JMB, 2004)

FIGURE 3.3 - *Structure of common-type acyl phosphatase*

The structure of 'common type' ACP from bovine testis determined by X-ray crystallography to a resolution of 1.8 A [98] (PDB 2ACY). Variant used was 89.9% identical to that used in crystal structure determination. Figure generated using PyMol (http://pymol.sourceforge.net).



3.3 Materials and Methods

3.3.1 Expression and Purification

The gene for ctAcP was synthesized commercially (Operon, Inc) and subcloned into a prSET expression vector. Bi-His double mutants were engineered sequentially using the QuikChange protocol (Stratagene) in a pseudo-WT triple-mutant background eliminating native histidines H25A, H60A, and H74A [104]. ctAcP was expressed and purified as previously described [73]. Protein identity was confirmed by electrospray TOF-MS using a Perkin-Elmer electrospray mass spectrometer.

3.3.2 Kinetic folding measurements

Rapid mixing experiments were conducted using SFM-4 and SFM-400 stopped-flow rapid mixing devices (Bio-logic) as previously described [72]. Tryptophan fluorescence spectroscopy used excitation wavelengths of 280–290 nm and emission wavelengths of >320 nm. Resulting kinetic data was fit to a single exponential using Bio-kine kinetic data fitting software. (Bio-logic) Data were analyzed using chevron analysis [32] with the free energy of equilibrium folding and the activation free energy for kinetic folding and unfolding dependent on denaturant according to the standard equations. Parameters were fit using a non-linear, least-squares algorithm with Microcal Origin.

3.3.3 ψ-analysis

The kinetic effect of divalent metal ions on folding behavior is measured in two ways. First, the dependence of folding rates on denaturant concentration is obtained in the absence and presence of high concentrations of metal ions (e.g. >1 mM $ZnCl_2$). The metal ions alter folding rates, resulting in movement of the folding and unfolding arms of the chevron plot. When the native state and TS are stabilized to the same extent (i.e. the site is native-like in the TS), the folding rates increase while the unfolding rate remains constant and ψ=1. Conversely, when the affinity in the TS is the same as in the unfolded state, the unfolding rate is slowed while the folding rate remains unchanged and ψ=0.

When both the folding and unfolding arms move in the presence of metal ions, the ψ -value is fractional, indicating the biHis site is fractionally formed or distorted in the TS, or a combination of thereof [38; 105]. A more detailed examination is conducted where the folding and unfolding rates are measured at dozens of metal ion concentrations, but at single folding and unfolding denaturant concentration. The resultant "metal chevron" data is used to generate a Leffler plot [37] of the change in folding activation free energy, $\Delta\Delta G_f$, versus the change in native stability, obtained from the kinetic data, $\Delta\Delta G_{bind}=\Delta\Delta G_u-\Delta\Delta G_f$. The data are fit to a functional form with a single free parameter, the ψ_o which is the initial slope at $\Delta\Delta G_{bind}=0$. (EQUATION 2.1)

3.4 Results

3.4.1 Engineering biHis Sites

To abrogate potential non-native metal binding, wild-type ctAcP is triply mutated removing native histidines (H25A, H60A, H74A), creating a pseudo-wild type variant on which all biHis mutants are based. Nine bi-histidine metal-ion binding sites then are engineered individually onto the surface of ctAcP. Two sites are located on each of the helices while five sites are within the β-sheet network. (FIGURE 3.4) Exogenously added divalent cations (e.g. Co^{2+}, Zn^{2+}, Ni^{2+}) bind the biHis sites and stabilize the protein for all except the site across strands 4 and 5, where the addition of cations had no effect on stability. The degree of stabilization varies among sites, presumably due in part to differences in preferred ion coordination geometry among histidine partners.

Chevrons of all biHis variants measured in the presence of saturating concentrations of divalent cations indicate that neither the biHis site and the associated metal-induced stabilization does not greatly affect the folding pathway. The amount of denaturant-sensitive surface area buried in the TS of biHis variants does not change appreciably as a result of mutation or metal-saturation of the binding site (TABLE 3.1). The ratio m_f/m^o indicates that 70-80% of the total denaturant sensitive surface area is buried in the TS. This invariance indicates that the histidine mutations and ion binding induce no qualitative changes in folding behavior.

3.4.2 ψ-analysis

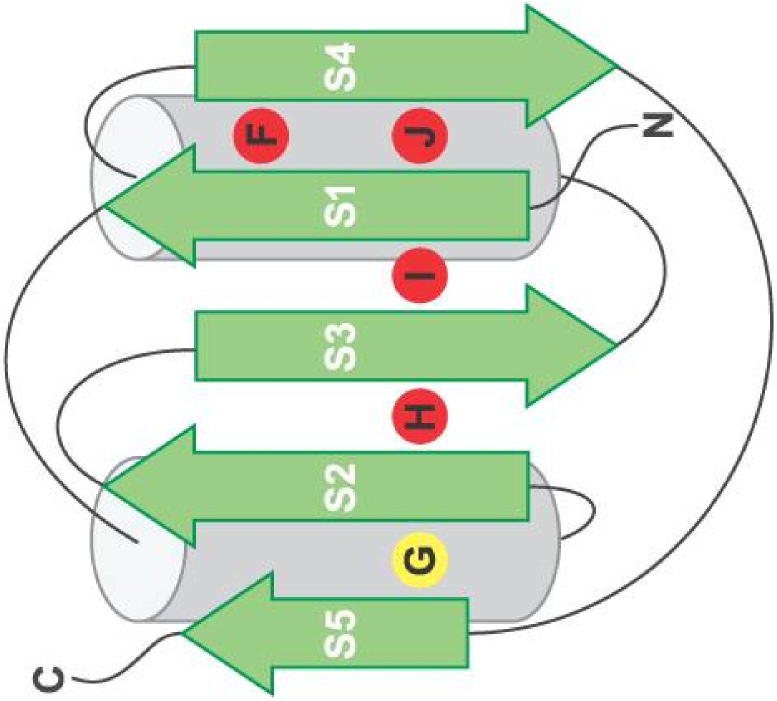
The biHis sites located in the four major strands of the β-sheet network are formed in the TS ensemble. For these sites bridging S1-S3, S1-S4, and S2-3, the ψ-value, or slope of the Leffler plot, is constant and nearly unity over the measured range of metal ion concentrations. ψ_0-values of one indicate the TS has native-like ion binding affinity. Hence, these four strands are arranged in a native-like topology at the rate-limiting step. Although the site across strands 4 and 5 was not stabilized by the addition of cations, we tentatively infer that strand 5 is absent in the TS because of its small size and long connecting region between the two strands.

Four sites are tested in the α-helices, Sites A and B in Helix H1 and Sites C and D in Helix H2. Sites A and B, located on two consecutive helical turns have very low ψ-values, despite H1 being a scaffold for several catalytic residues (FIGURE 3.4). Using denaturant chevrons at zero and high concentrations of $CoCl_2$, Site B is found to have a near-zero value of $\psi_o = 0.07$. For Site A, the metal dependence of relaxation rates produces an initial slope $\psi_o = 0.024$. There is no indication of any participation of H1 in the TS ensemble even after 0.2 kcal mol^{-1} of stabilization of the biHis site.

Chevron analysis of Site D in Helix H2 produced a ψ-value of unity (TABLE 3.1). The Leffler plot for Site C in Helix H2 is the only one which is significantly curved with an initial slope of ψ_o=0.37 ± 0.004, obtained using

FIGURE 3.4 - *Schematic representation of biHis sites in ctAcP and ψ-values*

A schematized representation of ψ-analysis results at various biHis sites in ctAcP. The two figures depict two views of the protein; the β-sheet network (S1-S5) and α-helices (H1 and H2). Each circle represents a biHis metal binding site across two secondary structure elements. Helical sites represent two histidines across one turn or four residues apart. Sheet sites bridge two strands with one histidine on either chain. Sites are color coded to represent resultant ψ-values; yellow indicates low or zero, red is 1, orange is 0.3.


S4
F
J
N
S1
I
S3
H
S2
G
S5
C

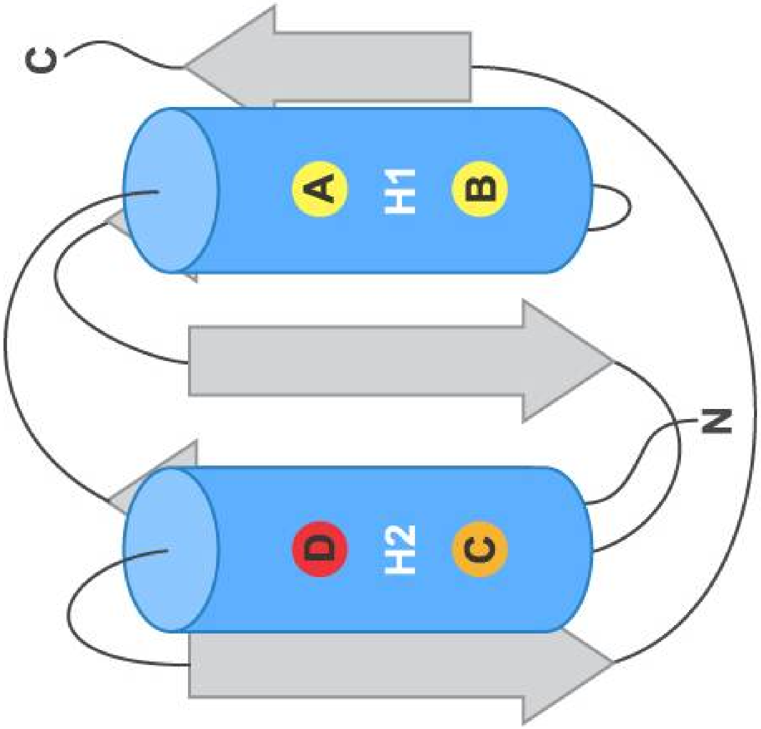

C
A
H1
B
N
D
H2
C

FIGURE 3.5 - *Multi-metal ψ-analysis of ctAcP-C*

Leffler plot of fractional Site C using several different divalent cations. Here $NiCl_2$ is in black, $CoCl_2$ in red, and $ZnCl_2$ in green.

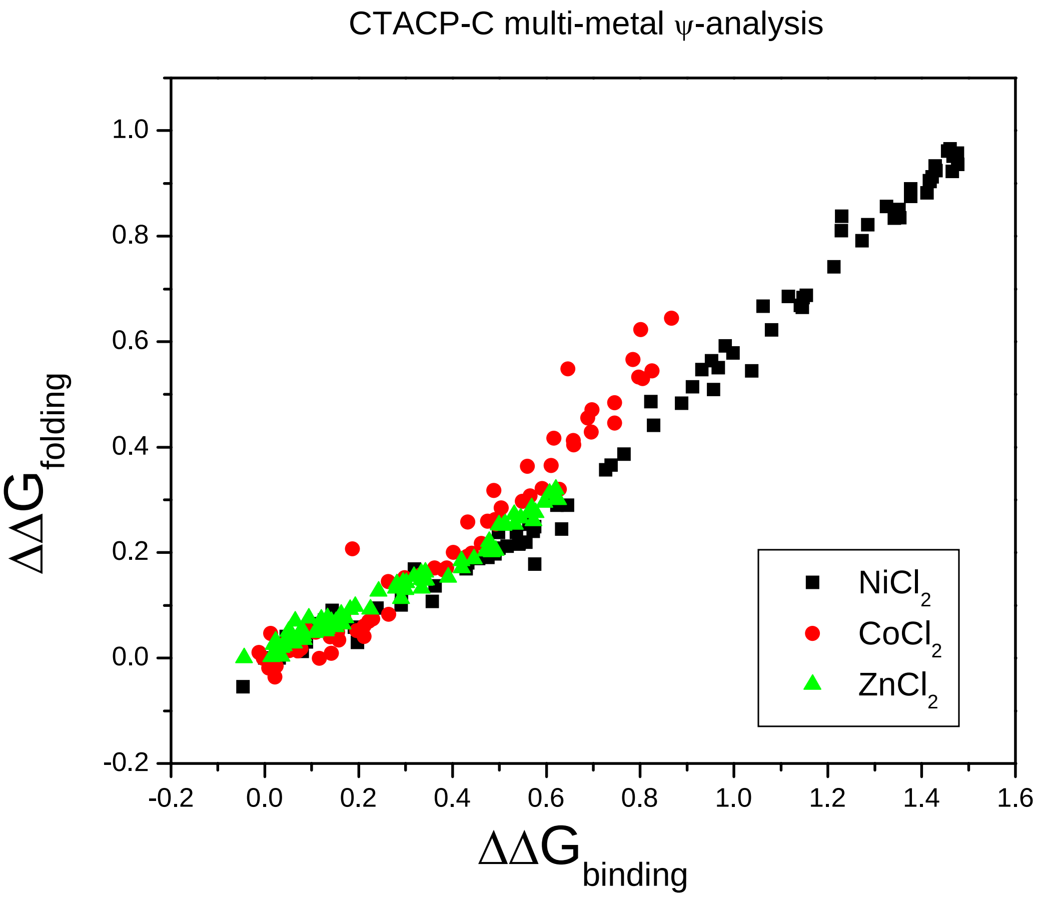

CTACP-C multi-metal ψ-analysis
$\Delta\Delta G_{folding}$
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
$\Delta\Delta G_{binding}$
■ $NiCl_2$
● $CoCl_2$
▲ $ZnCl_2$

TABLE 3.1 - *Results from metal-binding studies on ctAcP*

Abbreviations: PWT-Pseudo WT; ΔG_{eq} - Folding equilibrium stability; $\Delta\Delta G_{mut}$ - Change in equilibrium stability as a result of mutation; $\Delta\Delta G_{me}$ - Stabilization of metal binding; $\Delta\Delta G^{\ddagger}{}_{f}^{me}$ - Change in folding free energy as a function of metal; Δm^0 - Change in total *m*-value;Δ(mf / m0); ϕ - two-point ϕ-value.

Name	**Mutation**	ΔG_{eq}	$\Delta\Delta G_{mut}$	$\Delta\Delta G_{me}$	$\Delta\Delta G^{\ddagger}{}_{f}^{me}$	Δm_0	$\Delta(m_f / m^0)$	ψ	ϕ
WT	-	4.76	-	-					
PWT	H25A H60A H74A	5.57	0.80	-					
A	K24H A28H	4.82	-0.75	0.19	- 0.21	-0.29	-0.02	0.02	0.31
B	A28H K32H	4.68	-0.89	0.10	- 0.63	-0.54	-0.08	0.00	0.07
C	S56H H60H	4.70	-0.87	0.60	0.11	-0.26	-0.01	0.34	1.29
D	R59H E63H	4.69	-0.87	-0.05	0.17	-0.25	0.00	1.00	0.84
F	E12H N79H	3.78	-1.79	0.42	0.06	-0.80	-0.09	1.00	0.34
G	Q95H W38H	3.67	-1.89	-0.09	- 0.22	-0.37	-0.04	n/a	0.46
H	W38H Q50H	2.60	-2.96	0.31	0.56	-0.83	-0.04	1.21	0.38
I	Q50H D10H	4.34	-1.23	0.78	0.49	-0.80	-0.06	0.79	0.09
J	D10H N81H	3.33	-2.24	0.06	- 0.11	-0.61	-0.04	1.00	0.36
avg		4.08	-1.50	0.33	0.02	-0.56	-0.05		

$NiCl_2$ (FIGURE 3.4 AND FIGURE 3.5). The curve begins with a moderate slope and curves up slightly after 1.5 kcal mol^{-1} of stabilization. The use of two other metals gave similar results. $CoCl_2$ and $ZnCl_2$ imparted $\Delta\Delta G_{bind}$ = 0.9 and 0.7 kcal mol^{-1}, and generated ψ_o values of 0.46 ± 0.09 and 0.34 ± 0.01, respectively. The similarity of the three ψ_0-values is a probable signature that the biHis site is fractionally populated at the degree given by ψ_o-value, as we observed in the GCN4 coiled coil [75]. The alternative scenario, wherein the curvature is due to a distorted site with fractional binding affinity in a homogenous TS ensemble [38; 105], is doubtful as such fractional binding affinity is unlikely to be maintained with three ions having different coordination geometries and binding affinities. Hence, in the TS ensemble, the carboxy terminal portion of H2 is formed while the amino terminus undergoes fraying heterogeneously. In other words, at least two populations exist at the transition state, approximately 30% of proteins with the site formed and the remainder without.

3.5 Discussion

In choosing common-type acylphosphatase as our subject for ψ-analysis, we expected *priori* to find two major pathways, each comprising one of the two helices docked against three or more strands of β-sheet. Our previous results in ubiquitin showed proteins folding through TSs where the α-helix is associated four β-strands [106]. This helix/strand nucleus appears twice in ctAcP, so by extension, we initially believed folding would occur through two structurally disjoint folding nuclei. These two pathways would be defined by H1, S5, and S2 on the one hand and H2, S1, and S4 on the other. (FIGURE 3.5) The central strands of the β-sheet (e.g. S3) could potentially participate in both pathways, so it was predicted to form in both nuclei.

Previous experiments in ctAcP using traditional ϕ-analysis indicated uniform results in both helices (~0.3) and lower structure in the peripheral strands (S2, S4) [5]. More exhaustive work from the same group clarified these results and indicated the high degree of structure present in H2 compared to H1, suggesting H2 is much more critical to establishing the folding nucleus [107]. These mutational data agreed with our regional topological examination and led us to believe the nucleus containing H2 would be more likely to form in the TS than in H1, and these two structures would define separate, structurally distinct pathways through which the protein could fold.

Given these hypotheses, we have examined the folding TS topology of ctAcP using a host of biHis metal binding sites. Our examination of the α-helices returned results qualitatively similar to previous mutational data for the two helices.. Both sites in H1 have returned very low ψ -values indicating the lack of any structure in the TS. Sites in H2 have resulted in one high and another intermediate ψ-values indicating the presence of the helix with a partially frayed amino terminus. (FIGURE 3.7)

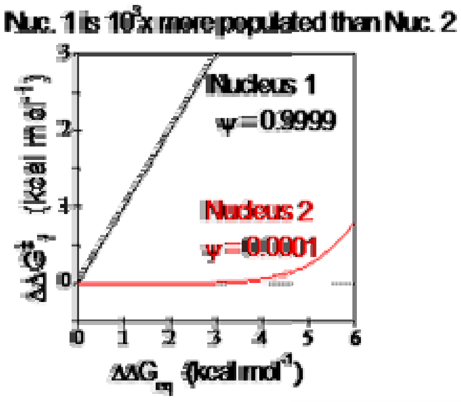
The m_f/m^0 values represent the percentage of native surface area buried at the transition state. In all histidine variants the transition state desolvates approximately 80% of the available surface area, indicating no large structural changes have occurred upon introduction of the metal site ligands. Additionally, chevrons with saturated metal sites (>1mM) show less than 2% change in surface area burial as compared to studies conducted in the absence of metal. This result verifies that local stabilization does not appreciably alter overall protein structure, or bias the transition state to some unnatural conformation.

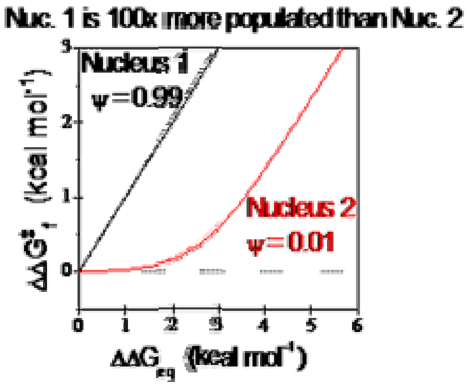
It is interesting to note that residues on the C-terminal strand are known to be important for stability and catalysis [108]. Additionally, the H1 biHis sites located opposite S5 (sites A, B) which position catalytic residues 15-21 show low ψ-values [109]. However, residues 42-45 at the S2/S3 hairpin turn region are also known to be catalytically important although the probe for hairpin formation (Site H) shows a high ψ-value. Enzymatic studies confirm that activity is recovered only after the major folding event has occurred [110]. The lack of correlation between TS structure formation and catalytic site formation further confirms the gap between folding and chemistry.

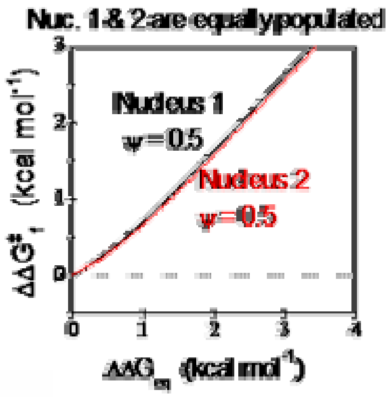
We have also generated two-point ϕ-values comparing the pseudo-wt ctAcP to each of the biHis mutations. These results were then compared with ϕ-analysis results seen previously and resulted in a correlation coefficient of 0.45. While these values may seem qualitatively similar, these two methods are clearly not representative of the same phenomenon. (FIGURE 3.8) Additionally, models were created of the protein transition state using the resultant ψ-values as constraints. Simulation software from A. Colubri was used to randomize torsional

FIGURE 3.6 - *Possible disjoint nuclei in ctAcP and Leffler plots indicating possible pathway partitioning*

Two possible structurally disjoint folding nuclei in ctAcP as predicted from experimental evidence. (Bottom Left) One possible folding nucleus, comprised of helix 1, strand 5, and strand 2. (Bottom Right) Another possible folding nucleus which is assembled from helix 2, strand 1, and strand 4. The central strand of the β-sheet (S3) could potentially participate in both pathways, so it is described as being shared. If two structurally disjoint folding nuclei were found to be present in the transition state, the participation of each pathway would be determined using ψ-analysis and a Leffler plot. (Upper Left) The results of sites in either nucleus if one TS ensemble was a vastly minor participant in the folding pathway, one part in ten-thousand. In this case the minor pathway may not be detectable using standard metal binding experiments. (Upper Center) If the minor pathway were participating on the 1% level, curvature would be detectable in all sites and the partitioning of proteins through the transition state would be quantifiable. (Upper Right) In the case of equal participation between both nuclei, the Leffler Plots would both show ψ-values of 0.5.


Nuc. 1 is 10^3x more populated than Nuc. 2
Nucleus 1
ψ=0.9999
Nucleus 2
ψ=0.0001
$\Delta\Delta G^{\ddagger}_{f}$ (kcal mol^{-1})
0
1
2
3
0 1 2 3 4 5 6
$\Delta\Delta G_{eq}$ (kcal mol^{-1})


Nuc. 1 is 100x more populated than Nuc. 2
Nucleus 1
ψ=0.99
Nucleus 2
ψ=0.01
$\Delta\Delta G^{\ddagger}_{f}$ (kcal mol^{-1})
0
1
2
3
0 1 2 3 4 5 6
$\Delta\Delta G_{eq}$ (kcal mol^{-1})


Nuc. 1 & 2 are equally populated
Nucleus 1
ψ=0.5
Nucleus 2
ψ=0.5
$\Delta\Delta G^{\ddagger}_{f}$ (kcal mol^{-1})
0
1
2
3
0 1 2 3 4
$\Delta\Delta G_{eq}$ (kcal mol^{-1})

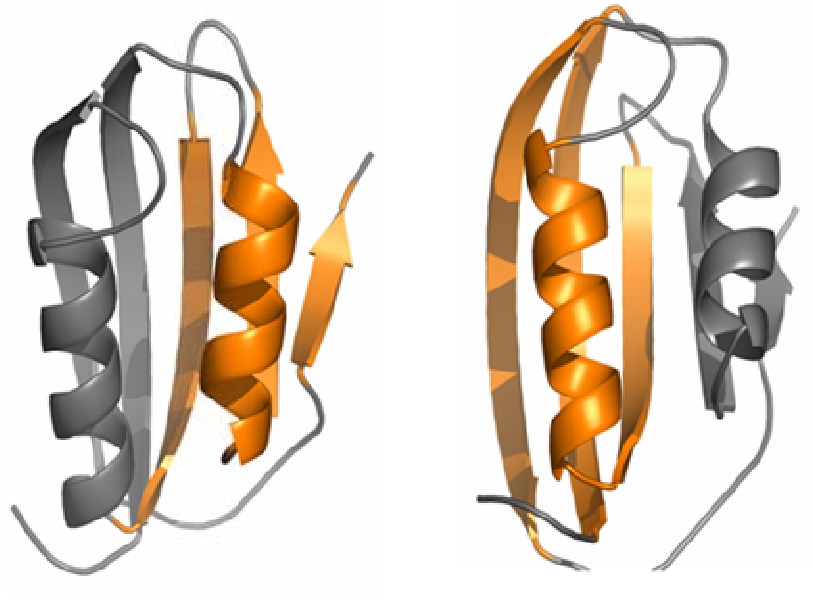


FIGURE 3.7 - *Proposed TS structures for ctAcP folding*

Structural models depicting steps on the folding pathway, as interpreted using ψ-value results. Regions colored blue are unstructured and red indicates native-like structure or topology. Beginning with the extended unfolded chain (U) the protein progresses to form the structures which show highest ψ-values, alignment of strands 1-4 of the sheet. At the transition state two pathways exist through structures identical aside from formation of a small amount of Helix 2 (Site D). Due to the low ψ-value of Site D, the population through this pathway is low enough to be considered a minor participant. The majority of molecules pass through a TS structure with only the four β-strands aligned. After the TS structure is formed, areas which showed low ψ-values are formed and the native state (N) is quickly reached.

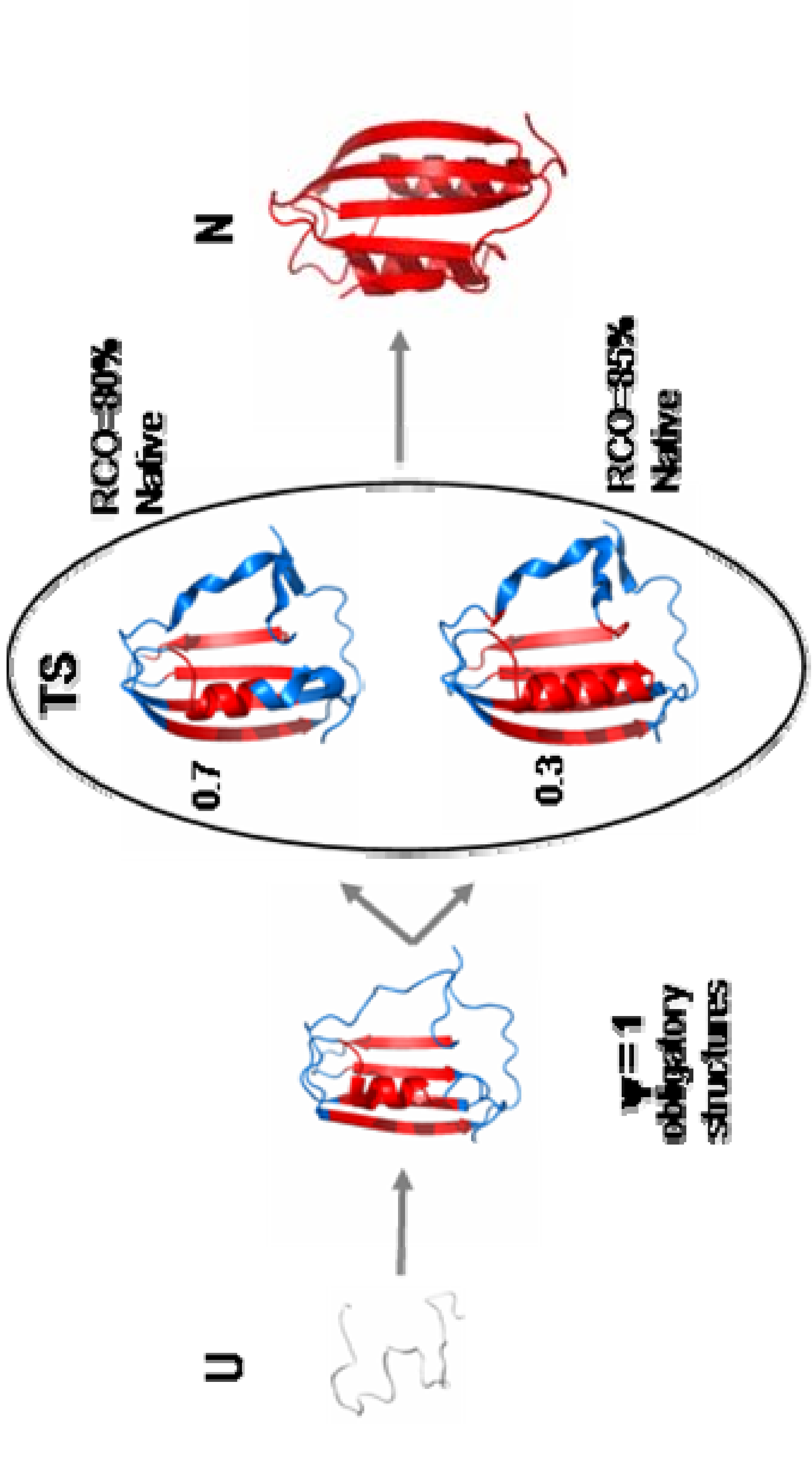
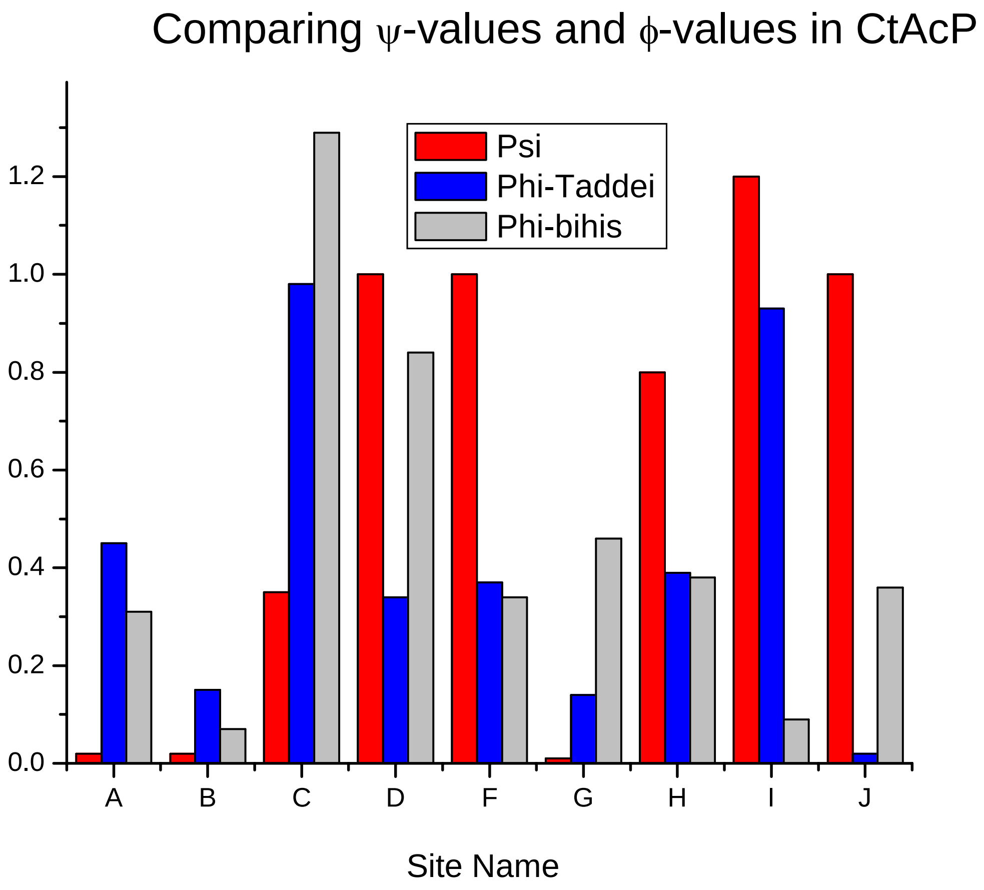

U
Ψ = 1
obligatory
structures
TS
0.7
0.3
RCO=80%
Native
RCO=85%
Native
N

FIGURE 3.8 - *Contrasting ψ- and ϕ-values in ctAcP*

Comparison of ϕ-values and ψ-values derived from two-mutation histidine variants and from previous work from Taddei *et al* [5]. The red bars indicate the ψ-values resultant from biHis metal binding kinetic studies in this work. The blue bars represent single-mutation ϕ data from previous analysis. Single mutations were assigned to various biHis pairs based on proximity to the binding pair. The grey bars indicate ϕ-values calculated from the change in folding rate ($\Delta\Delta G_f$) and stability ($\Delta\Delta G_{eq)}$ between the pseudo-WT and each biHis double mutation.


Comparing ψ-values and ϕ-values in CtAcP
Psi
Phi-Taddei
Phi-bihis
1.2
1.0
0.8
0.6
0.4
0.2
0.0
A
B
C
D
F
G
H
I
J
Site Name

angles in peripheral areas with low ψ-values while residues showing high ψ-values were maintained in their native configuration [111]. Resultant models were examined to verify a lack of steric clash and native-like backbone torsional angles. Transition state models were generated and the contact order was calculated as a function of native-state topology. Given the uncertainty in site C, several models were created which represented the maximally and minimally ordered structures. Results indicate that 73-79% of the native residue-residue contacts are maintained in the TS. Taken together with the very similar amount of native surface area burial from chevron analysis, ctAcP appears to bury a large number of residues and form native-like topology commensurately. This result is in good agreement with previous studies describing concomitant surface area burial and hydrogen bond formation in the folding process [18]. Additionally, these results both suggest the TS is very native-like both in the degree of surface area buried and topology.

As stated previously, binary ψ-values are clearly interpretable while intermediate or fractional ψ-values can indicate either distorted site formation with reduced binding affinity in the TS, or full site formation in a minor pathway which averages with the unformed population to appear as a fractional value. Fractional ϕ-values have also suffered from this ambiguity when attempting to discern fractional formation from pathway heterogeneity. However, ψ-values determined with several different metals offer a method through which heterogeneity can be determined. Mathematically, the only option aside from multiple fully formed pathways is for the biHis site to be distorted in the TS with reference to the native state.

If the site were distorted in geometry, then metals with different coordination geometries would stabilize the structure to different extents, and return very different ψ-values. In the present study, only one site showed a fractional ψ-value and Leffler plots measured with different metals were very similar aside from the saturating $\Delta\Delta G_{\text{binding}}$, which is a function of binding constant. As a result of the similarity of these results, we can conclude the site is not distorted in the TS, and each metal binds perhaps in a different configuration which affects the dissociation constant, but the responses from the TS and the native state are identical.

In summary, the ψ-value results for ctAcP exhibit values of either one or zero with the exception of one fractional site, which is believed to be representative of a partially populated native interaction in the transition state. Hence we can conclude ctAcP, much like ubiquitin, folds through a single very native-like transition state ensemble with a large amount of required native topology and some optional peripheral structures.

Generally, we believe that proteins fold through transition states which are very native-like in contacts, and which bury a significant amount of hydrophobic surface area while forming hydrogen bonds. Generally, the transition state is defined by accumulation of energetically costly long-range contacts, a phenomenon which is also observed in theoretical studies [112].

3.6 Conclusion

Using ψ-analysis provides a comprehensive quantitative assessment of the transition state on the folding pathway. Not only do we get a measurement of native topology using engineered histidine residues proximal in the native state, we can also utilize fractional ψ-values to determine if the protein folds through multiple structurally distinct pathways. In the case of ctAcP, we believe the protein is very native in topology as well as surface area burial at the rate-limiting step. We can conclude ctAcP folds through a homogenous transition state ensemble with a native-like consensus nucleus and minor optional structure in one helix.

4.0 Native Topology at the Rate-Limiting Step in Folding

4.1 Abstract

Two-state proteins move from an unfolded state to a single native conformation by passing through a high-energy transition state ensemble. The properties of the protein folding transition state ensemble, and the generality across different protein classes, is unclear. Topological characterization of the rate-limiting step on the folding pathway has been performed in common type acyl phosphatase (ctAcP) using ψ-analysis, which employs divalent metal ion binding sites to induce local stabilization. We find the transition state of ctAcP has very native-like topology much as seen in ubiquitin [1]. The topological complexity of the transition state, as quantified using relative contact order (RCO), is estimated to be approximately ~75% that of the native state in both proteins. As the transition states of both ubiquitin and acyl phosphatase have ¾ of the relative contact order of the native state, we propose that proteins which obey the known relationship between RCO and log k_f will have a transition state with a similar fraction of the native topology [62]. This conclusion places a very stringent constraint on possible transition state structures, notably offering evidence against highly polarized transition states. To test this hypothesis, models of the transition states of a dozen proteins are generated. Native-state hydrogen exchange data was used to discern residues whose hydrogen bonds only break upon complete protein unfolding, from residues which transiently unfold through "subglobal" openings.

In our modeling, the subglobal residues are locally deformed through randomization of the backbone torsional angles, and the RCO values of the TS models are calculated. For the proteins calculated, most were native-like with an RCO = 78 ± 12%. This result suggests other proteins are likely to be very native-like in transition state topology and that this result should be a general phenomenon for proteins obeying the folding rate-topology correlation.

4.2 Introduction

The structural states on the protein folding pathway are poorly understood. Protein folding can be modeled using a single reaction coordinate describing the conversion of random coil to native structure through a single high-energy transition state. Characterization of this transition state ensemble (TSE) is critical to understanding the underlying forces which drive the steps in the folding process. Extensive experiments have bolstered conceptually disparate models which describe transition states as having anywhere from very little regular structure (e.g. non-specific hydrophobic collapse [19]) to the early native-like formation of secondary structure elements in isolation [20; 21; 113; 114; 115].

In small, globular proteins, the conversion of a protein from random coil to the native structure is largely a "two-state" process, meaning that no intermediates accumulate on the folding pathway [71]. As a result, the only experimentally tractable species between the two populated states is the high-energy TS. Given this narrow foothold on the TS, much work has focused on the folding effects of protein structure, sequence, and connectivity perturbations to characterize the TS.

The idea of topology as a rate-limiting step in protein folding was proposed from work on equine Cytochrome C. Sosnick, Englander and coworkers observed that the equilibrium molten globule folded to the native state faster than experimentally measurable (< 1 msec), whereas folding from the chemically denatured state took tens of milliseconds [64]. In the molten globule, the three major helices are present in native-like arrangement, but this state lacks native-like side-chain packing and a solvent-excluded core [116]. Hence, packing and solvent exclusion are fast processes and cannot account for the slower rate of folding from the chemically denatured state. Rather, once the chain is organized into the native topology, folding can occur rapidly, implying that acquisition of the native topology is the rate-limiting step in folding from the fully denatured state.

Folding studies of Cytochrome C indicated that the rate-limiting step is an uphill conformational search for some minimal amount of native-like chain topology [64], suggesting the most difficult step in the folding process is formation of a sufficient number long-range native contacts such that subsequent steps are energetically downhill. These steps after the rate-limiting barrier are likely to involve the relatively fast folding of partially unfolded loop regions, via smaller-scale search processes.

Interestingly, a meta-analysis by Plaxco *et al* compared the folding rates and topological complexities of proteins which folding in a two-state manner [62]. They found a statistically significant correlation between log $k_{folding}$ and RCO for a dozen proteins, suggesting that the rate-limiting step in the folding pathway is directly dependent upon arrangement of the protein chain into a specific topological conformation (FIGURE 1.4).

In the original topology-folding rate correlation, 12 proteins were used in the test set which spanned topological content from completely α-helical (λ-repressor) to all β-sheet (Fyn-SH3 domain). To this sample set, we added the data compiled in a recent work which collected kinetic folding data for thirty proteins which were characterized under the same experimental conditions [117]. In an effort to further characterize the degree of native topology in the TS as well as the generality of this conclusion, we have characterized a topologically complex protein, common-type acyl phosphatase (ctAcP), and modeled the transition states of other proteins which obey the topology-folding rate correlation using published native-state hydrogen exchange data.

4.3 Results and Discussion

4.3.1 ψ-analysis and transition state topology

The ψ-analysis method probes for native residue-residue contacts using a two-partner histidine metal binding site. As a result, the method readily identifies the topology of the transition state. The methodology was applied to mammalian ubiquitin (Ub) and common-type acyl phosphatase (ctAcP) to identify the degree of intra-chain native contacts in the transition state. In both proteins, the TS ensemble had a single consensus ensemble involving the majority of the β-sheet network and a portion of an α-helix (FIGURE 3.2 AND 3.7). Around the periphery of the consensus structure, regions of the protein undergo fraying, for example at the end of the helix or between two strands. For Ub and ctAcP, the consensus TS structure has a very native-like topology, with values of the RCO of 80% and 71%, respectively. Contact maps, which graphically represent all residue-residue interactions in a protein, appear very similar between the TS models and the native state. (FIGURE 4.1)

The TS ensembles identified using ψ-analysis, in combination with the RCO trend suggests an intriguing proposition: For proteins which obey the known RCO correlation, their transition states will have 70-80% of the native RCO. The rationale for this proposition is as follows. The empirical correlation between folding rates and topology is log k_f = 8.3-39•RCO.(FIGURE 1.4) This relationship establishes a strong connection between the conformational properties of the TS and the ground state. It follows that the RCO of the TS should closely resemble that of the native state. Consistently, the transition state of both ubiquitin and ctAcP have a very native-like topology, with a $\mathrm{RCO}^{\mathrm{TS}}$~0.7-0.8•RCO. If the RCO correlation is to hold for a variety of proteins, their transition states likewise should have $\mathrm{RCO}^{\mathrm{TS}}$ ~ 0.7-0.8•RCO in order to appear correlated. That is, the transition state topology of ubiquitin and ctAcP serve to benchmark connection between $\mathrm{RCO}^{\mathrm{TS}}$ and RCO of the native state.

Another rationale is illustrated with a counter-example. If a protein only forms part of the native topology (e.g. $\mathrm{RCO}^{\mathrm{TS}}$ ~ 0.5•RCO), it would fold faster than expected based on the folding-topology trend due to the benefit of forming a simpler TS compared to the other proteins.

Inspecting the dispersion of the data around the observed trend, we expect all proteins to obey 0.8•RCO ≤ $\mathrm{RCO}^{\text{transition state}}$ ≤ 1.0•RCO. In addition, this relationship restricts the degree to which a TS can be small and polarized,[118; 119; 120; 121; 122; 123; 124; 125; 126] such as those proposed for CspB,[118] S6,[123] titan I27,[50] SH3,[119; 124; 127] Protein G,[128] and L.[129] From their ϕ-values, the TSs appear to have $\mathrm{RCO}^{\mathrm{TS}}$ much less than our result of 0.8. These results would seem at first glance to be incompatible with our proposed RCO relationship. However, there are at least two caveats in the identification of a small, polarized TS based solely upon medium to high ϕ-values on one side of the protein. The first is that ϕ-analysis can result in the incorrect assignment of a small, polarized TS, as we found in Ub[38]. Essentially, ϕ-values can under-report chain-chain contacts as ϕ-values reflect energies and not structures, as Schmid *et al* astutely noted "the TS of CspB folding is polarized energetically, but it does not imply that one part of the protein is folded and the other one is unfolded. Rather, it means that the positions that have reached a native-like energetic environment in the TS are distributed unevenly."[118] That is, *energetically* polarized does not necessarily mean *structurally* polarized.

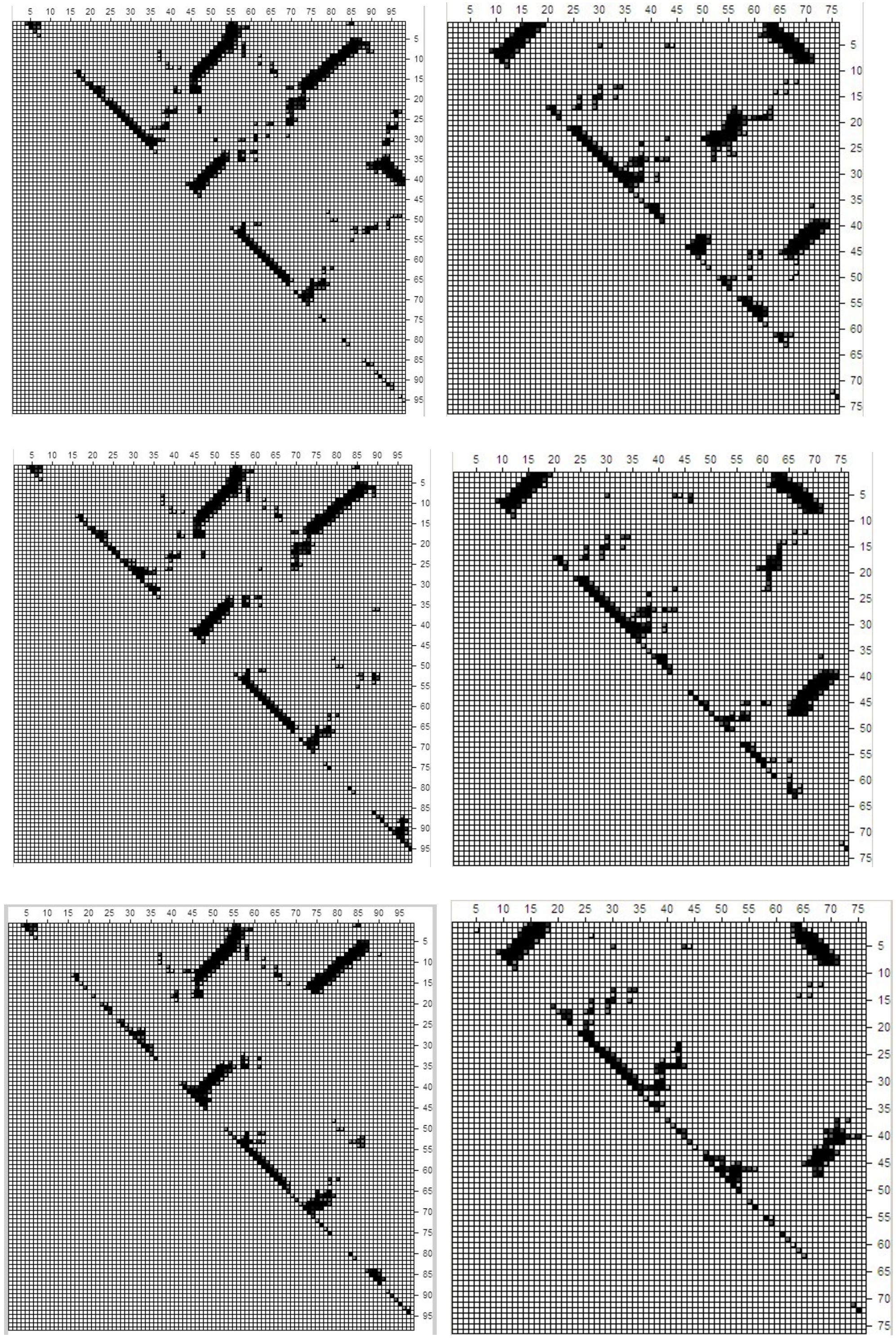
The second caveat is that many high ϕ-values in polarized transition states are associated with turn regions of a protein.[118; 119; 120; 124; 128] However, these results may not faithfully depict the picture of the transition state topology.

Serrano *et al* concluded three SH3 homologs, SSo7D, src- and α-spectrin, fold via different transition states based on different ϕ-values in the turn regions. [127] In response, one could state the over-all TS topology is similar in all three proteins, but the turns are only folded to the degree required for the chain to double back on itself. Not all turns have to be native-like in order for the chain to double-back on itself. If true, the sensitivity of the ϕ-values is more a reflection of the specific interactions of individual turn regions rather than the topology of TS. For example, the distal β-hairpin in src-SH3 with high ϕ-values is a tight turn [119] which is quite sensitive to mutation, whereas the corresponding turn in SSo7D contains three flexible glycines has low ϕ-values.[127] Hence, ϕ-values could be different for this turn in the two proteins despite them having similar TS topologies.

At its core, the correlation between folding rate and contact order speaks to the energetic difficulty of forming an adequate number of long-range contacts in the protein as the main hurdle of the folding process. Direct experimental measurement of the rate of folding as a function of loop extension and contraction demonstrated a linear relationship as seen previously, thus providing some suggestion for loop closure entropy as being a an important part of the energetic hurdle in folding [130].

FIGURE 4.1 - *Contact Maps of ctAcP and Ubiquitin TS and native states*

Contact maps graphically represent each residue-residue contact in a structure file. The axes represent the residue number and each contact is represented by a black square at the coordinate representing both residues. The plot contains an identity line at the diagonal, through which the results are mirrored; here the reflection in the lower left of each map is omitted. Groups of black squares parallel to the identity diagonal represent helical structures while those perpendicular represent β-hairpins. Contact maps on the left represent ctAcP while those on the right represent Ubiquitin structures. The top panels are the contact maps for the native states, the middle for the minimal TS models, and the lower panels represents the maximal sites. Here we can see the panels are very similar indicating the degree to which native contacts are preserved in our transition state model.



5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75

4.3.2 Alternative Correlations

Critics of the log(k_f) – RCO correlation have suggested other structural properties are better suited for predicting folding rates, such as short range [66] or long-range contacts [65], percentage of secondary structure elements [67], or the total contact distance [68]. These other approaches generate qualitatively similar results to the log(k_f)-RCO correlation and involve methods which are essentially variations on the original residue-residue distance quantification schemes, most of which are normalized for number of contacts as well as chain length [65; 66; 67; 68]. Surprisingly, normalization by chain length or number of contacts does not have much of an impact on the resultant correlations. The most plausible explanation for this lack of size-dependency is the distribution of candidate proteins is very tight, in which case normalization is inconsequential.

A rigorous statistical analysis of these models indicated most other schemes correlate about as well as the original work [131]. However, the conceptual difference between a topology parameter and one based on secondary-structure assignment is nontrivial, although both may generate a similar correlation. Other metrics differ largely in the definition of contact distance and normalization, although all confirm the importance of topology in determining folding rate. One parameter is Total Contact Distance, which sums the sequence distance between all residue-residue contacts using the following formula:

$$TCD = \frac{1}{n_r^2} \sum_{k=1}^{n_c} |i - j|$$

EQUATION 4.1

where n_r is the number of residues, n_c is the number of contacts. The summation is calculated for all residues within 5Å of each other through space, giving a correlation coefficient between TCD of the native state and folding rate of R=0.558 as compared to 0.602 with RCO. (FIGURE 4.2 TOP) This approach normalizes to chain length but not to the total number of contacts when compared with RCO. Additionally, the minimum cutoff sequence length between two contacts is variable, and has been shown to be constant up to 14, which lead the authors to conclude that long range contacts of at least that length were critical for establishment of the folding nucleus.

Another group developed a correlation with the percent of local contacts which gives a correlation of R=0.406 (FIGURE 4.3 TOP). This approach simply calculates the number of inter-residue contacts less than or equal to four residues apart in sequence. In effect, this calculates the amount of helical and turn content in the protein, and disregards a great deal of information by not considering longer range interactions, which likely explains the poor correlation. Additionally, the percent of local contacts is correlated with RCO itself (R=0.72) although this has no bearing on each parameter's correlation with folding rate.

The only correlative parameter which surpasses RCO is Long Range Order (LRO), which reports on the degree of interactions between residues which are more than 12 residues apart. (FIGURE 4.4 TOP) The formula is

$$LRO = \sum n_{ij} / N$$

EQUATION 4.2

where N is the number of residues and n_{ij}=1 if the distance between *i* and *j* is ≥ 12 and zero otherwise. While somewhat similar to the previous approach, LRO calculates the fraction of native contacts above a fixed sequence distance (default 12 residues) rather than below it. The correlation coefficient is R=0.729, which outdoes the RCO correlation. The implication here is simply that long range residue contacts are difficult to make and comprise the bulk of the energy barrier in the folding pathway, which is conceptually similar to other metrics evaluating topology.

While some correlations perform better than the original topology-folding rate correlation, the general phenomenon of overall structural complexity is embodied in many of these analytical approaches. Quantifying the number of long range interactions versus secondary structure content when an α-helix gives a fixed low contact number and β-sheets give a more variable but on average higher value, are essentially a different roads to the same place. Importantly, the large amount of theoretical and analytical work which has advanced topology as the determinant of folding rates has no strong experimental demonstration as of yet. A method directly sensitive to topology is necessary to be able to ascertain to what extent contact distance is important on the folding pathway.

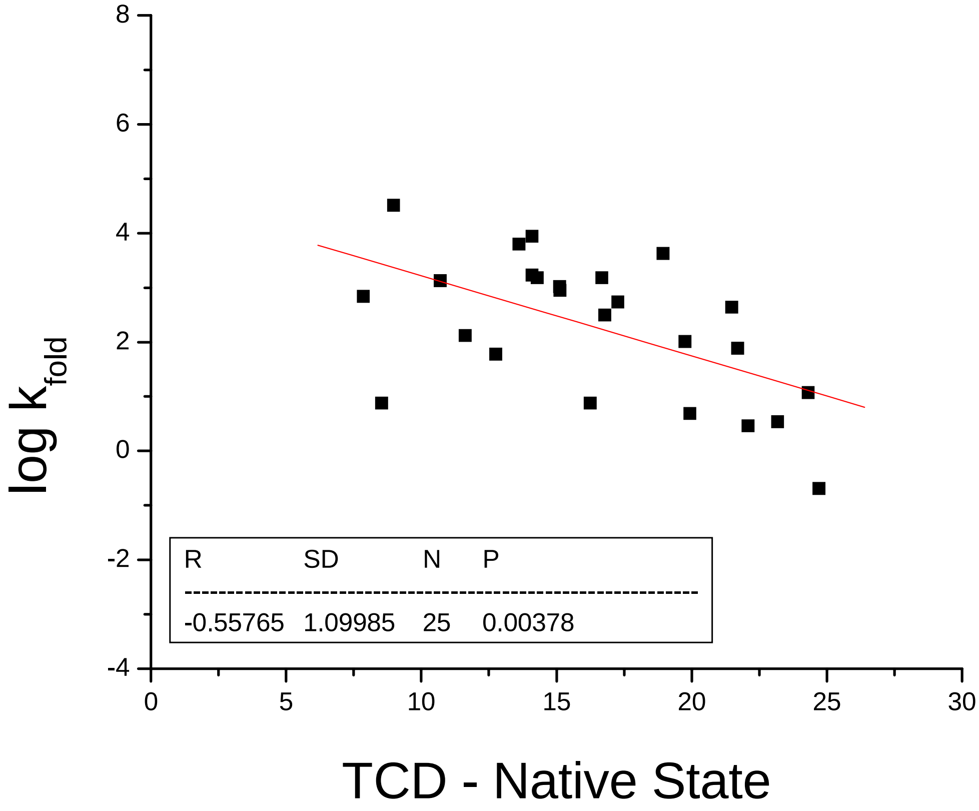
4.3.3 Modeling TS using HX data

In an effort to use experimental data to help model the transition state of several proteins in this correlation, we used native-state hydrogen exchange (HX) data due to its non invasive characterization of global and local unfolding events [132; 133]. Hydrogens on the main chain amides of proteins engage in continuous exchange with solvent protons when unstructured and exposed to solvent. In the native state, a protein may undergo spontaneous structural fluctuations which can lead to exposure of the amide moiety and hydrogen exchange. The rate of hydrogen exchange is dependent upon pH as well as intrinsic exchange rates of each residue.

Using H^1-NMR spectroscopy in D_2O, the exchange rate of individual residues are measured in the native state as proteins exchange for deuterons. When this rate is measured as a function of increasing denaturant, exchange rates increase as unfolding events are catalyzed. These events can be categorized as local unfolding events, which are largely denaturant independent, and global events, which exchange only when the entire protein is denatured, and are very denaturant dependent. (FIGURE 4.7) Using this hierarchy, a picture of the structural intermediates after the rate-limiting step on the folding pathway can be constructed. (FIGURE 4.8) Taken more simply, the global residues generally represent those which are buried or hydrogen bonded at the transition state, and can be used to construct a model of the structure at the rate-limiting step.

FIGURE 4.2 - *Total Contact Distance vs. Folding Rate*

The correlation between the log of the folding rate (k_{fold}) and the Total Contact Distance (TCD) in the native state (top). The red line represents the best linear fit to the data. The bottom panel represents the same correlation using transition state models generated as described in Section 4.3.3.


log k_{fold}
8
6
4
2
0
-2
-4
0
5
10
15
20
25
30
TCD - Native State
R SD N P
-0.55765 1.09985 25 0.00378

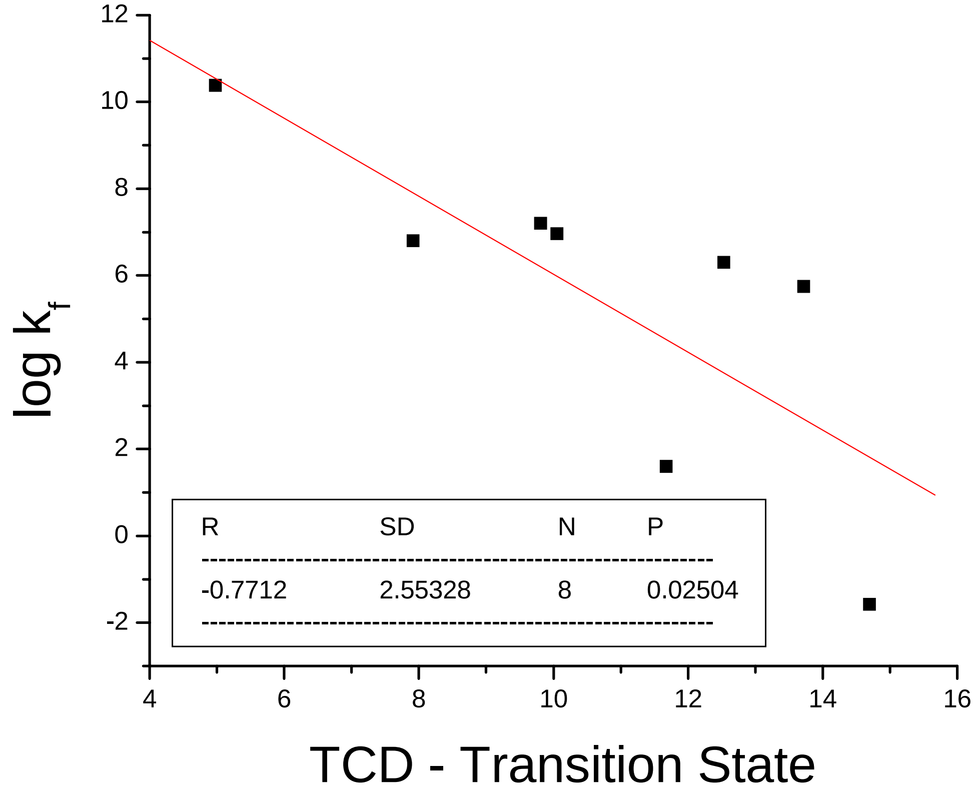
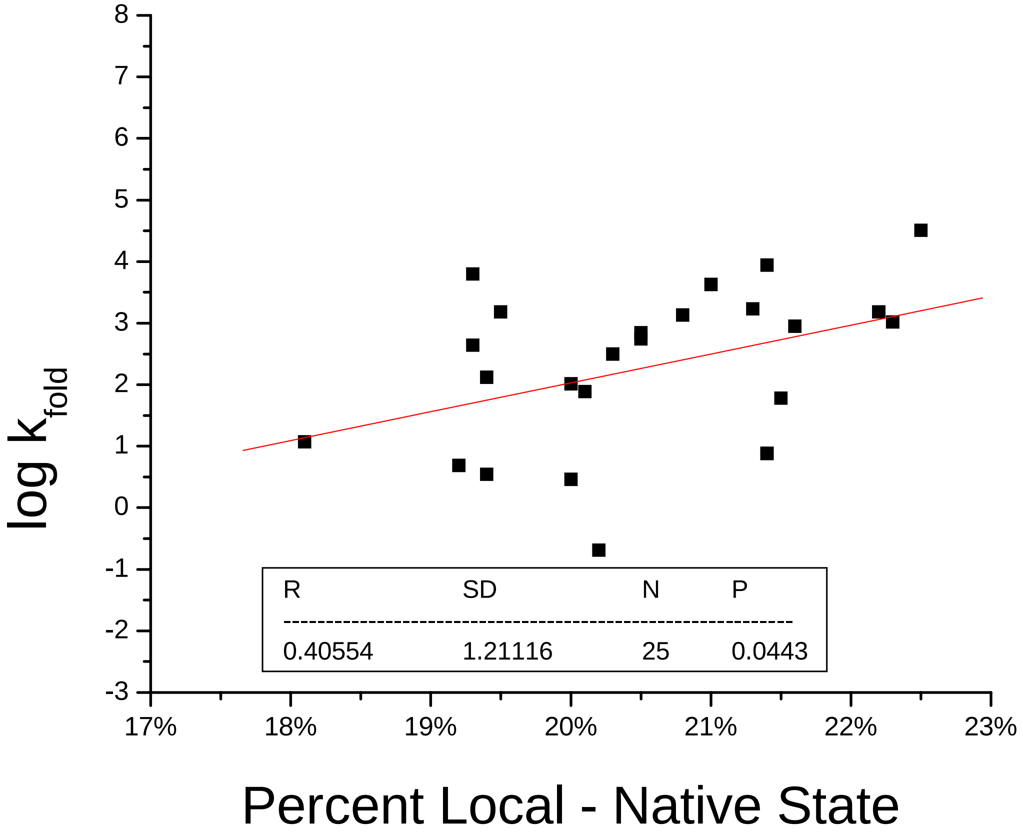

log k_f
12
10
8
6
4
2
0
-2
4
6
8
10
12
14
16
TCD - Transition State
R SD N P
-0.7712 2.55328 8 0.02504

FIGURE 4.3 - *Percent local contacts vs. folding rate*

The correlation between the log of the folding rate (k_{fold}) and the percentage of local contacts (top). The red line represents the best linear fit to the data. The bottom panel represents the same correlation using transition state models generated as described in Section 4.3.3.


log k_{fold}
8
7
6
5
4
3
2
1
0
-1
-2
-3
17%
18%
19%
20%
21%
22%
23%
Percent Local - Native State
R SD N P
0.40554 1.21116 25 0.0443

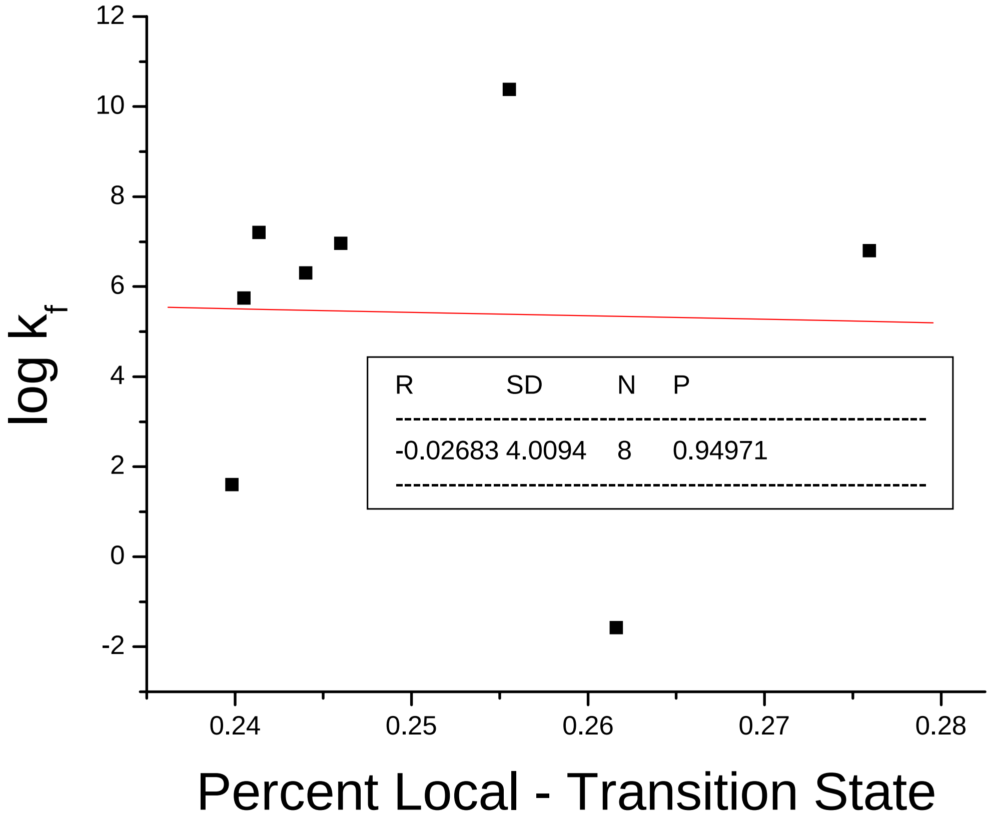
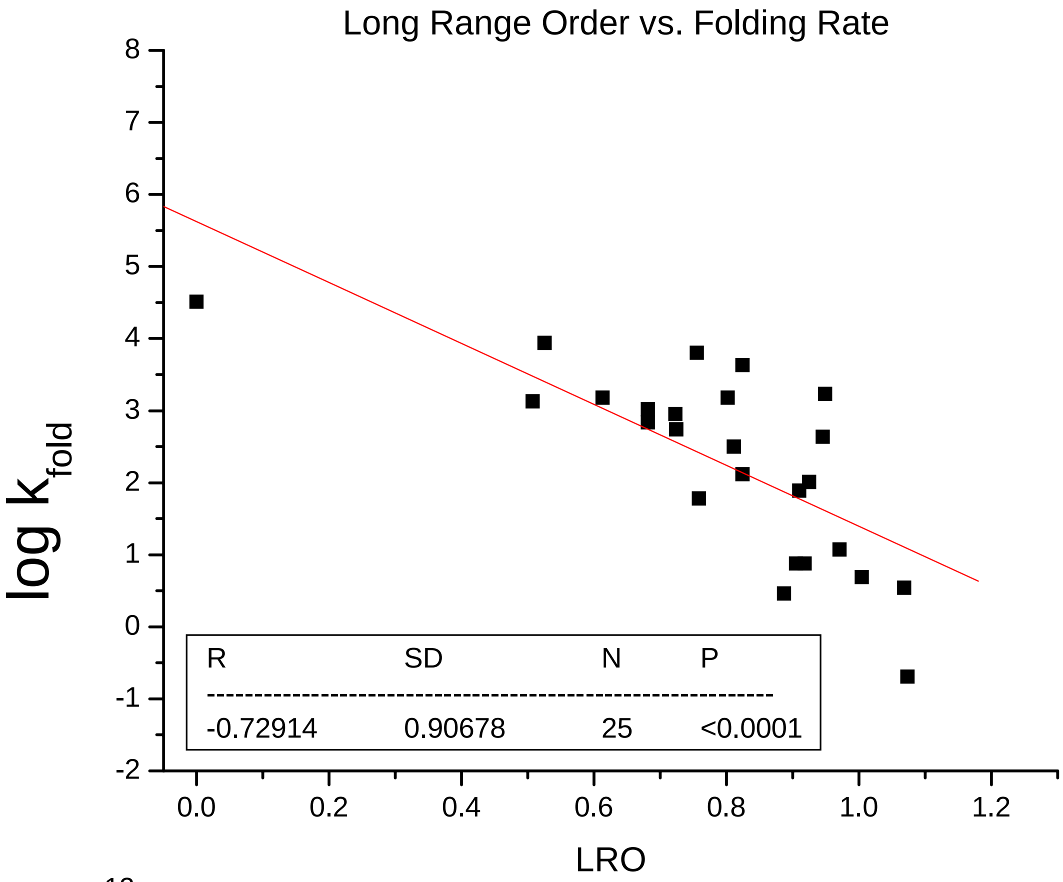

log k_f
12
10
8
6
4
2
0
-2
0.24
0.25
0.26
0.27
0.28
Percent Local - Transition State
R SD N P
-0.02683 4.0094 8 0.94971

FIGURE 4.4 - *Long range order (LRO) vs. folding rate*

The correlation between the log of the folding rate (k_{fold}) and the Long Range Order (LRO) top. The red line represents the best linear fit to the data. The bottom panel represents the same correlation using transition state models generated as described in Section 4.3.3.


Long Range Order vs. Folding Rate
log k_{fold}
8
7
6
5
4
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
LRO
R
SD
N
P
-0.72914
0.90678
25
<0.0001

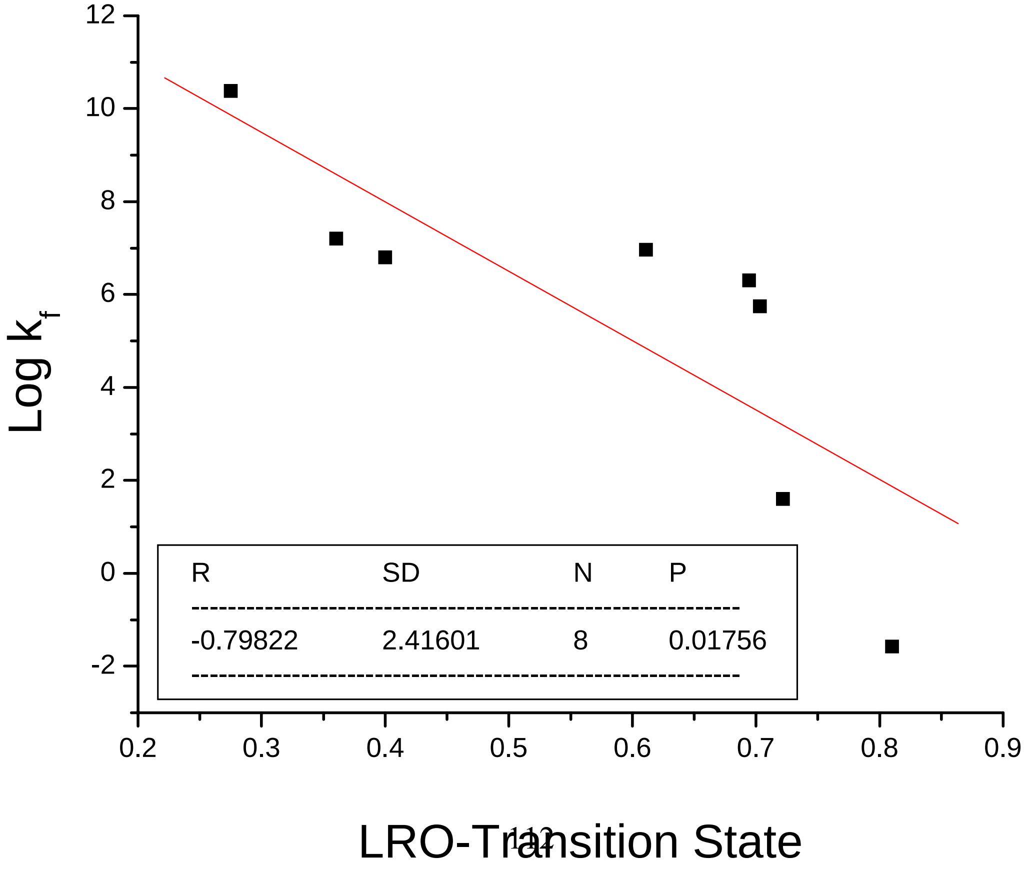
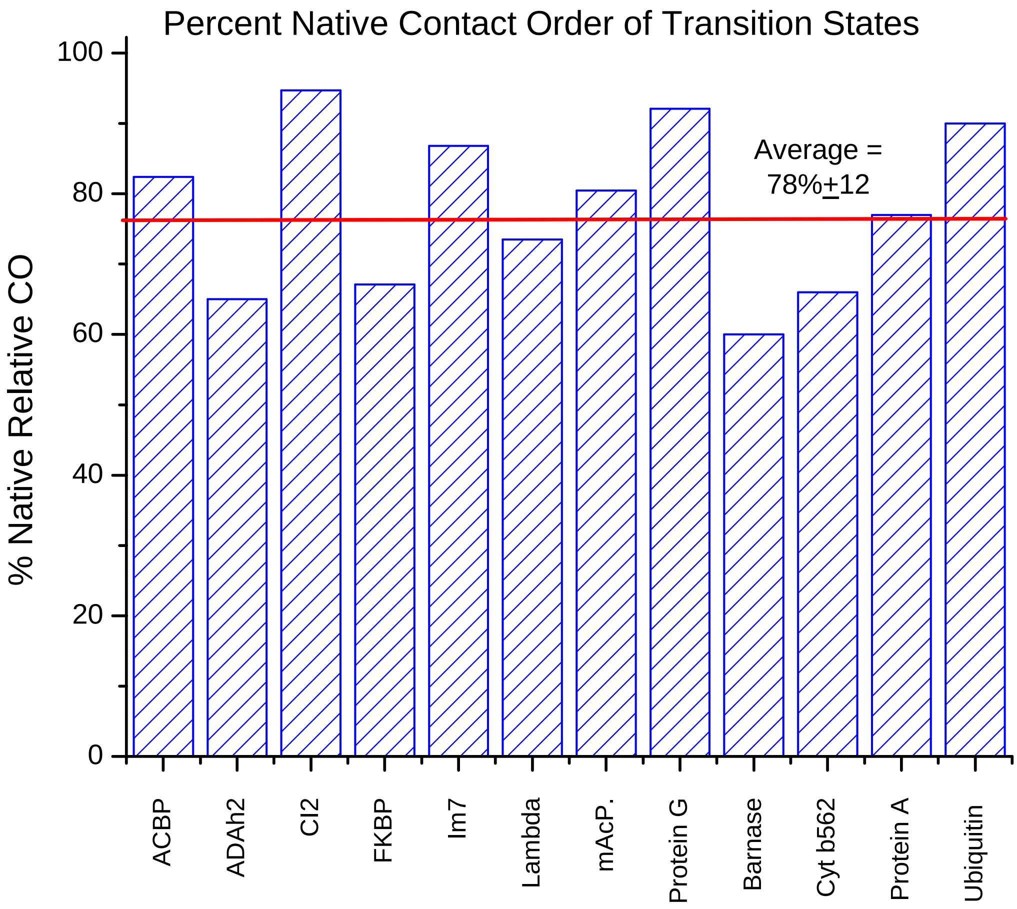

Log k_f
12
10
8
6
4
2
0
-2
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
LRO-Transition State
R
SD
N
P
-0.79822
2.41601
8
0.01756

FIGURE 4.5 - *HX TS Modeling Results*

A bar graph representing the degree of native topology as measured by RCO from models generated using hydrogen exchange data.



Percent Native Contact Order of Transition States
% Native Relative CO
0
20
40
60
80
100
Average =
78%±12
ACBP
ADAh2
CI2
FKBP
Im7
Lambda
mAcP.
Protein G
Barnase
Cyt b562
Protein A
Ubiquitin

FIGURE 4.6- *RCO - TS vs. Folding Rate*

Here we examine the correlation between folding rate and RCO for the transition state models created using experimental HX data.

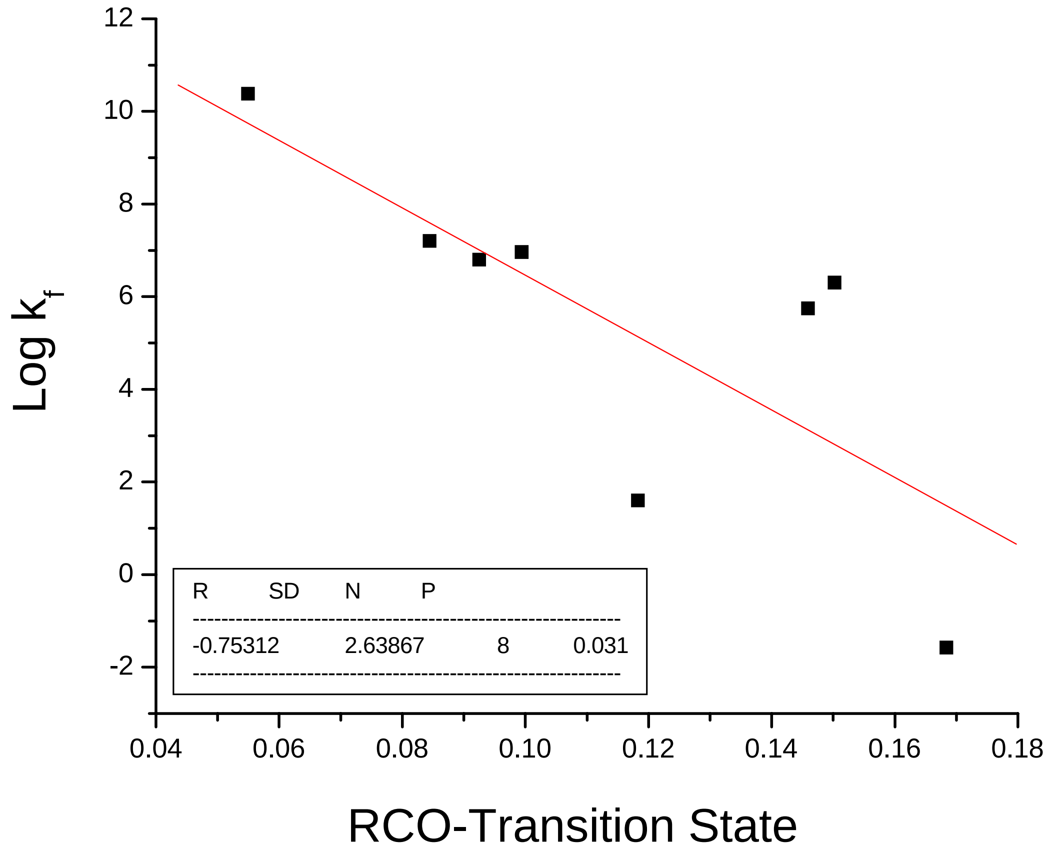

12
10
8
6
4
2
0
-2
Log k_f
0.04
0.06
0.08
0.10
0.12
0.14
0.16
0.18
RCO-Transition State
R SD N P
-0.75312 2.63867 8 0.031

FIGURE 4.7 - *Mechanism of hydrogen bond isotope exchange as a result of unfolding events*

The denaturant dependence of hydrogen exchange indicates the degree to which individual residues participate in local unfolding events (with equilibrium constant K_{eq}^{local}) and global unfolding events requiring total denaturation (K_{eq}^{local}). At very low denaturant (top) the majority of residues are in native configuration, with only a few participating in local unfolding events and thus hydrogen exchange. At intermediate denaturant concentrations (middle), the least stable local structures begin to unfold and their backbone hydrogens exchange for deuterons. At high concentrations of denaturant (bottom), all backbone hydrogens exchange with solvent.

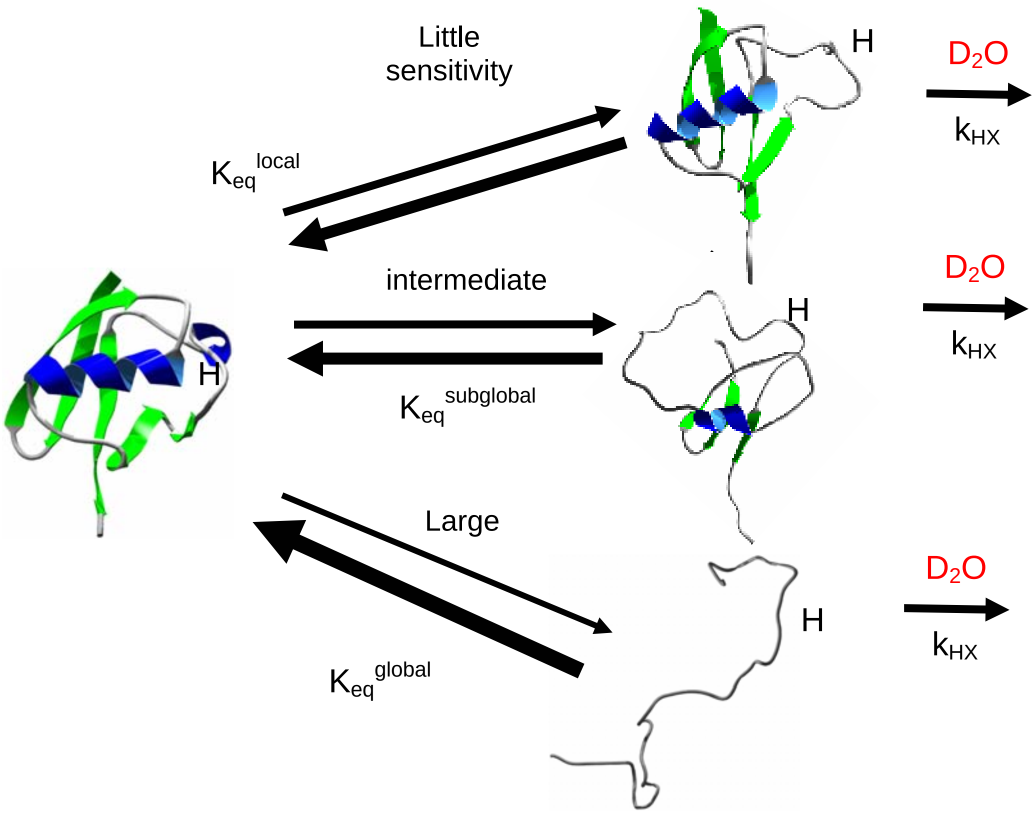
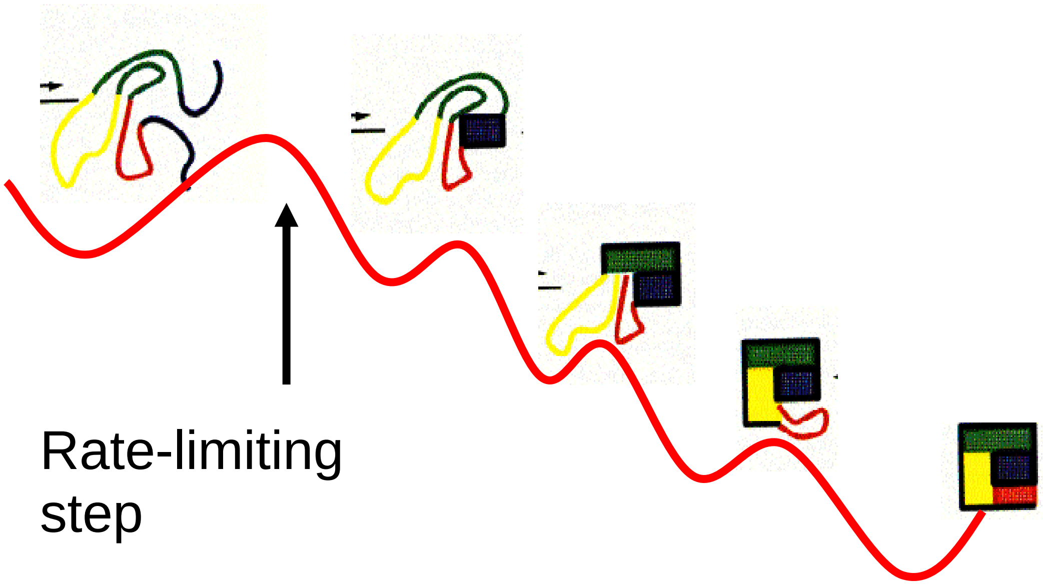

Little sensitivity
K_{eq}^{local}
H
D_2O
k_{HX}
intermediate
$K_{eq}^{subglobal}$
H
D_2O
k_{HX}
H
Large
K_{eq}^{global}
H
D_2O
k_{HX}

FIGURE 4.8 - *Free Energy Surface for global and local unfolding events in hydrogen exchange*

Using the denaturant dependence of hydrogen exchange, the specific folding events on the folding pathway can be determined. Beginning at the native state (right) denaturant is increased and the first set of amid hydrogens to exchange represent the final step on the protein folding pathway, as indicated by the red loop. In this fashion, all of the steps on the native end of the folding free energy surface can be structurally determined. The transition state structure can be determined from the residues last to exchange before complete denaturation. Here the folding pathway is schematized as occurring with the blue segment folding at the transition state, then after the rate-limiting barrier, the green, then yellow, then red segments adopt native configurations.


Rate-limiting step

By cross-referencing proteins whose folding behavior has been characterized kinetically with those upon which native state hydrogen exchange (HX) NMR experiments have been performed, we constructed a set of twelve proteins suitable for modeling using HX data and who obey the folding-topology correlation. (FIGURE 4.5)

The native-state HX data were parsed to interpret the resultant ΔG_{HX} values as either globals or locals. This was conducted by comparing the overall equilibrium unfolding energy (ΔG_{eq}) with the ΔG_{HX} values of the individual residues. Where the two appeared to agree, the residues were identified as global exchangers, and the remainder local residues. When extrapolating these residue-specific energetic parameters to segments of protein structure, residues lacking data were grouped with sequence-proximal residues where appropriate.

Using this classification of residues as either globally or locally exchanging allows us to look at a protein structure and determine which segments are likely to be unstructured as we begin to unfold a protein. Here, global residues requiring total unfolding to exchange and thus indicate residues which are ordered in the TS while locals are susceptible to transient localized unfolding events, suggesting a lack of stability, and thus late formation in the folding pathway. Grouping locals together gives a picture of which secondary structural elements are likely to be unfolded in the transition state.

Once the segments of proteins structure had been designated to be either folded or unfolded based on the HX data, those portions of the protein were locally deformed using a modeling and simulation program [111]. The algorithm deformed local segments of structure by randomizing backbone torsional ϕ,ψ angles to other conformers while maintaining native structure elsewhere. Additionally, deformed structures were discarded where atomic clash existed, using Van der Waals radii of atoms as the radial limits of atomic interaction. RCO was calculated by submitting PDB files for each model to the Baker Lab webpage, which enumerates all residue-residue contacts and sums the sequence distances as indicated in the RCO formula [62]. (http://depts.washington.edu/bakerpg/contact_order)

Generally, the protein regions protected from hydrogen exchange are closer to the core. Also, protection factors are high in α-helices and β-strands. For example, in CI2, the protected strand from residues 47-51 pairs with another strand from 65-70, suggesting the remaining helix and terminal strands are disordered in the last stable intermediate. In the case where only one β-strand is protected, the strand docking against the protected area is also assumed to be structured, as its presence is required for HX protection. In ACBP, the internal helices (res 40-54 and 17-28) are disordered, while the helices closer to the termini are well-protected.

Hydrogen exchange studies of ubiquitin provided an avenue for comparison between the TS picture determined with ψ-analysis and that from HX modeling [134]. The HX results indicate the TS is very native in shape, as most of the helix and the main strands of the sheet are protected. The least protected areas were the connective regions and the mini-helix. The fourth strand of the β-sheet has no data, but is assumed to be docked against the adjacent strand to protect from solvent exchange. The model generated from HX data retained roughly 90% of the native contact order, suggesting this may be a slight overestimate given the value from ψ-analysis is closer to 80%.

At times, the degree of native RCO in these models can often be misleading to the eye. In CI2, for example, only one pair of docked strands results in the retention of 95% of the native amount of topology. However, it is important to bear in mind that loss of α-helical structure raises the average level of RCO by eliminating very local contacts from calculation. Similarly, loss of β-strands can lead to very large decreases in overall RCO due to the generally long-range nature of this type of secondary structure.

We have applied the alternative topology metrics to the transition state models generated using HX data in an effort to examine what degree of native topology is present from viewpoints other than RCO. Alternative metric results were generated using a modified version of the RCO perl script from the Baker Lab website. (http://depts.washington.edu/bakerpg/contact_order/contactOrder.pl) Comparison of TCD values from TS models and native state resulted in 59.9 ± 20.4% native, suggesting a somewhat native TS but also a wide dispersion in the results. (TABLE 4.1) Percent local contacts was the only metric which increased in the TS with respect to the native state, on average 24 ± 19%. This phenomenon suggests our modeled transition states are losing proportionally more long-range contacts. TS LRO is on average 72 ± 14% native, suggesting longer-range interactions are also being disrupted to a roughly equal amount. Of the three, again LRO seems to be the closest to the results generated using TCO than other metrics, and also seems to have the tightest dispersion as well. From these results we can conclude that long-range contacts are proportionally more disrupted at the TS and summing sequence distance of contacts over 14 residues apart in sequence (LRO) is similar to complete summations.

If we take these alternative quantifications of chain topology and revisit the folding correlation using our TS structures, we can get an idea of how our TS models correlate with folding rate among the various metrics. TCO in the TS exhibits a better correlation with folding rate than TCO of the native state, R=-0.77 in the TS vs -0.56 in the native state. (FIGURE 4.2 BOTTOM) However when the folding rate is compared with the percent local contacts, the correlation goes from bad to terrible, from R=0.45 in the native state to essentially zero, R=-0.02. (FIGURE 4.3 BOTTOM) It is clear from this picture that the percentage of local structures is not an accurate reporter of the degree to topological complexity, as it ignores the difference between contacts 10 residues apart and those 50 apart.

The best correlation is from LRO, which has a correlation coefficient of 0.72 in the native state and 0.79 in the transition state models. (FIGURE 4.4) The relatively small change in correlation coefficient between the native and transition states using LRO indicates the difference lies in disrupting mostly local interactions. This is further evidenced by the drastic change in correlation between native and transition states when using percent local contacts.

When using the original metric of RCO to calculate the contact order of the transition state models, the correlation between the topological complexity

TABLE 4.1 **-** *Alternative metrics applied to TS models*

Abbreviations: TS - Transition state; PDB - Protein Data Bank ID; TCD - Total Contact Distance in the native state; %TCD - Percent of native TCD value found in the TS model. Loc - Percent of local contacts in the native state; %N-loc - percentage of native local contacts found in the transition state; LRO - Long Range Order in the native state; %LRO - percent of native Long Range Order found in the transition state.

Protein	**PDB**	**TCD**	**%TCD**	**Loc**	**%N-loc**	**LRO**	**%LRO**
Barnase	1A2P	24.71	22.5%	0.20	124.3%	1.073	54.5%
Cytb652	1APC	5.59	54.0%	0.24	120.1%	0.384	53.7%
mAcP.	1APS	24.71	59.5%	0.20	129.6%	1.073	75.5%
Im7*	1AYI	10.70	91.7%	0.21	116.3%	0.507	71.1%
FKBP	1FKF	19.93	58.6%	0.19	124.9%	1.004	71.9%
λ-Rep	1LMB	8.98	55.4%	0.22	113.7%	0.000	n/a
ACBP	1NTI	15.12	66.5%	0.22	110.5%	0.681	89.7%
ADAh2	1O6X	15.13	52.3%	0.22	127.5%	0.723	55.3%
Protein A	1SS1	10.85	77.3%	0.22	109.1%	0.447	70.5%
ubiquitin	1UBQ	16.67	27.1%	0.20	180.8%	0.802	66.7%
CI2	2CI2	16.79	81.7%	0.20	118.6%	0.811	86.7%
Protein G	3GB1	17.27	72.6%	0.20	119.1%	0.724	95.9%
AVG		**15.54**	**59.9%**	**0.21**	**124.0%**	**0.686**	**72.0%**
St De		**5.89**	**20.4%**	**0.01**	**18.9%**	**0.314**	**14.4%**

and folding rate is surprisingly good. (FIGURE 4.6) In most cases, regardless of the method of calculating the topological complexity, it's clear that there is a correlation between the folding rate and topology in both the native and transition states. This suggests that our estimation of very native-like topology being present in the TS is likely to be a general phenomenon for other proteins.

The average level of native-like structure using RCO in the TS models is ~ 3/4ths that of the ground state, as found in the measured systems of ubiquitin and ctAcP (FIGURE 3.2 AND FIGURE 3.7). We believe the modeling results to be indicative of the likely structures of these proteins to approximate the over-all structure present at the rate-limiting step on the folding pathway, especially given they are based on such non-invasive experimental methods. Since the modeled proteins all conform to the topology-rate correlation, it follows that other proteins in this sample set also are very native-like in topology at the transition state.

4.3.4 Topomer Search Model

Recently, models attempting to explain the process of protein folding have made use of topological elements more frequently. One of the most compelling new models is the topomer search model [135], which involves a statistical description of how possible backbone configurations are sampled until the native-like state is located for a majority of residues. A topomer here is defined as a set of structurally disjoint backbone configurational angles which constitute a unique overall protein chain conformation, only a few of which are close to the native state.

This idea was proposed first by Debe and Goddard, and is aimed at reducing the large number of possible chain configurations as defined in the Levinthal Paradox to a sample size compatible with folding and sampling rates. [136]. The original estimation of possible chain configurations is roughly 3^{100} or 10^{48} given three positions for each residue, where the number of possible disjoint topomers formed by a chain of similar length is approximately 10^7. The number of topomeric states is calculated largely by eliminating protein configurations which contain steric clash and also grouping several topologically similar chain conformers together as a single topomer. The reduced sample set combined with a reasonable statistical approximation of sampling rates using Gaussian chain simulations as a benchmark, lead to calculation of folding rates of 100 ms. However, this number serves as a benchmark for the upper bound of protein folding rather than an average rate.

The topomer search model also posits that topology and folding rate are correlated, but are not directly causative of one another per se. Rather, the topology bears on a proxy variable, Q_D the statistical probability of the reaching the native topomer, which is calculated through ennumerating the number of sequence-distant native residue pairings. It is this probability and the rate of sampling which dictate the folding rate itself. Recent work from Marqusee tested the effect of circular permutation on folding rate and discovered that drastic alterations in the chain connectivity have minimal effect on the folding rate, which seems counterintuitive given the topology-rate correlation [137]. However the topomer search model proposes a formalism which calculates similar Q_D values for circular permutants since the number of sequence-distant pairs which define the topology is still constant. As such, the model would predict very similar folding rates among circular permutants of the same protein.

4.4 Conclusions

Experimental evidence generated using ψ-analysis for Ub and ctAcP has found that ~3/4's of the native RCO is attained in the TS. Using existing HX exchange data to identify core regions of the protein which are likely to be present in the TS, we have generated models of the TS for twelve proteins. These model TSs also have RCO between 60-90% of the native value indicating this property of TS is likely to be general.

5.0 Conclusions

5.1 Native-like transition state and pathway heterogeneity

For many years, experimentalists and theorists alike have attempted to understand the chemical and structural determinants of the rate-limiting step in two-state folding. In recent years, work has begun to focus on alignment of native-like topology as the slowest and thus rate-determining step in the protein folding process. However a topology-sensitive methodology was lacking until introduction of ψ-analysis using metal-binding biHis sites. This new probe for topology allows rough characterization of transition state ensembles and categorization of structural elements as critical or optional to the folding nucleus. Additionally, extension of this method using multiple metals can help determine if there are multiple structurally disjoint routes from the unfolded state to the native state.

This thesis has focused on the structural properties and pathway heterogeneity of the rate-limiting step in the folding of two-state globular proteins. We have examined the transition state of the topologically complex protein, common-type acyl phosphatase (ctAcP), using ψ-analysis to identify native intra-chain contacts. The results from this study indicated a very native-like topology where the TS has relative contact order which is ~70% of the native, as has been observed in previous studies of ubiquitin. These proteins both align well with the existing correlation between topological complexity and folding rate, and likely serve as a benchmark indicating this level of native topology is general in transition states of other proteins.

In our examination of ctAcP, ψ-values are near zero or unity for all sites except one fractional result on the amino end of the structured helix. This result provides the only indication of transition state heterogeneity in this study. The ψ-value remains unchanged when multiple metals of varying coordination geometries are used. The lack of metal preference suggests multiple pathways through this site, some of which are metal-stabilized, and others which are not. As with ubiquitin [1], the other globular protein extensively characterized using ψ-analysis, the transition state ensemble has single consensus structure. Despite this small amount of heterogeneity, the remaining results indicate a singular transition state ensemble with just a small amount of optional structure, not the structurally disjoint pathways indicative of heterogeneity. Hence, the folding pathways of both protein have essentially converged to a single transition state structure, albeit one which contains a minor amount of fraying around the periphery.

Using native-state hydrogen exchange data, models of several other protein transition states were generated. These models suggest other proteins also utilize transition states with a large degree of native-like topology as measured by fraction of the native RCO attained in the TS. Taken together, we believe that all proteins who obey the log(k_f)-RCO correlation will exhibit native-like topology in the transition state, with an RCO ~3/4's that of the native state. Ergo, the rate-limiting step in protein folding is likely to be defined by a large amount of native topology as a general phenomenon of protein structure and dynamics.

5.2 Future work

The transition state of ctAcP has been characterized to a large extent, both in topology and pathway heterogeneity. While the main secondary structure has been probed with biHis sites, the degree of structure in the connecting loop regions is unclear and could change the degree of native-like topology in the transition state as much as 10%. Characterizing these possibly unstructured regions would be difficult using biHis sites, as the results may be uninterpretable due to variability in the native state metal site configuration.

Instead, simple glycine or alanine mutations on surface residues which significantly distort the backbone torsional angle preferences would be informative as to whether or not the turn regions were structured. In other words, using a residue's ϕ/ψ configuration and also backbone torsional preferences of other residues, a mutation can be made where the backbone angles are significantly distorted to determine the importance in the TS. For example, Gly45's position on the Ramachandran map is at ϕ=-100, ψ=+25, while a mutation to alanine would push it to ϕ=-40, ψ=-60. When chosen to minimize changes in side-chain interactions, this type of mutation can be a reliable mechanism to use ϕ-analysis to probe connective structure in the transition state.

Additionally, ψ-analysis can be performed between secondary structural elements, for example across two helices. When arranged in this fashion, the site probes the alignment of formed structural elements and can indicate if tertiary structure is formed in the transition state. Sites across the two helices, or between a helix and a strand can clarify to what extent structural alignment occurs and if formation of long-range topology includes contacts across secondary structural elements.

The lone fractional site has been measured using several metals and the similarity of Leffler plots suggests pathway heterogeneity through this site alone. However, other sites have not been investigated with multiple metals, especially those with ψ-values of one or zero. While the interpretation of the fractional result is suggestive of a small degree of heterogeneity, there is a paucity of data regarding biHis site response to different divalent metals. Additional investigation of this phenomenon is necessary for complete development of ψ-analysis as a tool to detect pathway heterogeneity, not only in ctAcP and ubiquitin, but other systems as well.

Another mechanism for testing pathway heterogeneity involves using a biHis site in one segment of the protein and introducing destabilizing mutations in a distal part of the molecule. The site in question must have a low ψ-value to begin with, and the change in ψ-value indicates the degree to which the dominant pathway has been destabilized and now the probed site becomes favored. Experiments of this type, coupled with the previously described multiple-metal analysis would be able to definitively answer the question of if there is pathway heterogeneity in the transition state in protein folding, even down to the sub-1% level.

5.3 Why topology?

Many forces are involved in establishment of protein structure, and many energetic and conformational requirements must be satisfied before reaching the native state. These processes include hydrophobic surface area burial, hydrogen bond formation, establishment of electrostatic side-chain interactions, packing of the core, and assumption of native topology. Over the years nearly all of these processes have been suggested to occur in different orders and estimation of energetic benefits or costs has varied.

Based on the evidence presented in this work and others, we believe topology is the largest energetic hurdle on the pathway from random coil to native state. We see the folding process as a large-scale random search for arrangement of long and short-range residues into roughly the native orientation, burying surface area and forming hydrogen bonds concomitantly. Rather than describing the folding process as a sequence of these critical processes, we believe that while the establishing gross topology is the most energetically costly step, other interactions present in the native state occur concomitantly.

However, the reason behind topology being rate-limiting is not necessarily straightforward, nor is the fashion in which the native topology is arrived at. Additionally, proteins which contain longer-range interactions tend to fold slower, indicating a longer search process before location of native topology. It would seem upon examination of the two TS pictures we have established that the critical contacts for the folding nucleus are those which allow for a significant amount of desolvation. Beyond this requirement, it seems as if the selection of native interactions is somewhat stochastic as long as the appropriate amount of topology is established. Overall, the critical step in protein folding seems to be the coarse sorting of topology into a native-like arrangement in order to facilitate more local downhill conformational searches on the way to the native state. Future studies of the kinetic, thermodynamic, and structural phenomenon surrounding protein folding will likely shine light on this subject.

References Cited

1. Krantz, B. A., Dothager, R. S. & Sosnick, T. R. (2004). Discerning the structure and energy of multiple transition states in protein folding using psi-analysis. *J. Mol. Biol.* 337, 463-75.

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

3. Lumry, R. & Biltonen, R. (1966). Validity of the "two-state" hypothesis for conformational transitions of proteins. *Biopolymers* 4, 917-44.

4. Munoz, V., Thompson, P. A., Hofrichter, J. & Eaton, W. A. (1997). Folding dynamics and mechanism of beta-hairpin formation. *Nature* 390, 196-9.

5. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* 6, 1005-9.

6. Levinthal, C. (1968). Are there pathways for protein folding. *J. Chim. Phys.* 65, 44-45.

7. Karplus, M. (1997). The Levinthal paradox: yesterday and today. *Folding & Design* 2, 69-75.

8. Imoto, T. (2001). Effective protein folding in simple random search. *Biopolymers* 58, 46-9.

9. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold Des* 1, 441-50.

10. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992). Protein folding funnels: A kinetic approach to the sequence-structure relationship. *PNAS* 89, 8721-8725.

11. Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1997). Structural correlations in protein folding funnels. *Proc Natl Acad Sci U S A* 94, 777-82.

12. Teeter, M. M. (1992). Order and disorder in water structure of crystalline proteins. *Dev Biol Stand* 74, 63-72.

13. Calloni, G., Taddei, N., Plaxco, K. W., Ramponi, G., Stefani, M. & Chiti, F. (2003). Comparison of the folding processes of distantly related proteins. Importance of hydrophobic content in folding. *J Mol Biol* 330, 577-91.

14. Jacob, J., Krantz, B., Dothager, R. S., Thiyagarajan, P. & Sosnick, T. R. (2004). Early Collapse is not an Obligate Step in Protein Folding. *J. Mol. Biol.* 338, 369-82.

15. Sadqi, M., Lapidus, L. J. & Munoz, V. (2003). How fast is protein hydrophobic collapse? *Proc. Natl. Acad. Sci. U S A* 100, 12117-22.

16. Ladurner, A. G. & Fersht, A. R. (1999). Upper limit of the time scale for diffusion and chain collapse in chymotrypsin inhibitor 2. *Nat Struct Biol* 6, 28-31.

17. Mirsky, A. E. & Pauling, L. (1936). *Proc. Natl. Acad. Sci. U.S.A.* 22, 439.

18. Krantz, B. A., Srivastava, A. K., Nauli, S., Baker, D., Sauer, R. T. & Sosnick, T. R. (2002). Understanding protein hydrogen bond formation with kinetic H/D amide isotope effects. *Nature Struct. Biol.* 9, 458-63.

19. Robson, B. & Pain, R. H. (1976). The mechanism of folding of globular proteins. Equilibria and kinetics of conformational transitions of penicillinase from Staphylococcus aureus involving a state of intermediate conformation. *Biochem J* 155, 331-44.

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

21. Karplus, M. & Weaver, D. L. (1979). Diffusion collision model for protein folding. *Biopolymers* 18, 1421-1438.

22. Jha, A. K., Colubri, A., Zaman, M. H., Koide, S., Sosnick, T. R. & Freed, K. F. (2005). Helix, Sheet, and Polyproline II Frequencies and Strong Nearest Neighbor Effects in a Restricted Coil Library. *Biochemistry* 44, 9691-702.

23. Avbelj, F. & Baldwin, R. L. (2004). Origin of the neighboring residue effect on peptide backbone conformation. *Proc Natl Acad Sci U S A* 101, 10967-72.

24. Jha, A., Colubri, A., Zaman, M. H., Freed, K. F. & Sosnick, T. R. (submitted). Helix, sheet, and Polyproline propensities and strong nearest neighbor effects in a restricted coil library.

25. Brutscher, B., Bruschweiler, R. & Ernst, R. R. (1997). Backbone dynamics and structural characterization of the partially folded A state of ubiquitin

by 1H, 13C, and 15N nuclear magnetic resonance spectroscopy. *Biochemistry* 36, 13043-53.

26. Yang, D. & Kay, L. E. (1996). Contributions to conformational entropy arising from bond vector fluctuations measured from NMR-derived order parameters: application to protein folding. *J Mol Biol* 263, 369-82.

27. Shaw, G. L., Davis, B., Keeler, J. & Fersht, A. R. (1995). Backbone dynamics of chymotrypsin inhibitor 2: effect of breaking the active site bond and its implications for the mechanism of inhibition of serine proteases. *Biochemistry* 34, 2225-33.

28. Schneider, D. M., Dellwo, M. J. & Wand, A. J. (1992). Fast internal main-chain dynamics of human ubiquitin. *Biochemistry* 31, 3645-52.

29. Tsong, T. Y., Baldwin, R. L. & P., M. (1972). A sequential model of nucleation dependent protein folding kinetic studies of rnase a. *J mol biol* 63, 453-475.

30. Kim, P. S. & Baldwin, R. L. (1990). Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.* 59, 631-660.

31. Bai, Y., Sosnick, T. R., Mayne, L. & Englander, S. W. (1995). Protein folding intermediates studied by native state hydrogen exchange. *Science* 269, 192-197.

32. Matthews, C. R. (1987). Effects of point mutations on the folding of globular proteins. *Methods Enzymol.* 154, 498-511.

33. Goldenberg, D. P. & Creighton, T. E. (1985). Energetics of protein structure and folding. *Biopolymers* 24, 167-82.

34. Matouschek, A., Kellis, J. T., Jr., Serrano, L., Bycroft, M. & Fersht, A. R. (1990). Transient folding intermediates characterized by protein engineering. *Nature* 346, 440-5.

35. Myers, J. K., Pace, C. N. & Scholtz, J. M. (1995). Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* 4, 2138-48.

36. Fersht, A. R., Leatherbarrow, R. J. & Wells, T. N. C. (1986). Quantitative-Analysis of Structure-Activity-Relationships in Engineered Proteins by Linear Free-Energy Relationships. *Nature* 322, 284-286.

37. Leffler, J. (1953). Parameters for the Description of Transition States. *Science* 117, 340-341.

38. Sosnick, T. R., Dothager, R. S. & Krantz, B. A. (2004). Differences in the folding transition state of ubiquitin indicated by phi and psi analyses. *Proc. Natl. Acad. Sci. U S A* 101, 17377-82.

39. Feng, H., Vu, N. D., Zhou, Z. & Bai, Y. (2004). Structural examination of Phi-value analysis in protein folding. *Biochemistry* 43, 14325-31.

40. Sanchez, I. E. & Kiefhaber, T. (2003). Origin of unusual phi-values in protein folding: evidence against specific nucleation sites. *J. Mol. Biol.* 334, 1077-85.

41. Bulaj, G. & Goldenberg, D. P. (2001). Phi-values for BPTI folding intermediates and implications for transition state analysis. *Nature Struct. Biol.* 8, 326-330.

42. Ozkan, S. B., Bahar, I. & Dill, K. A. (2001). Transition states and the meaning of Phi-values in protein folding kinetics. *Nature Struct. Biol.* 8, 765-9.

43. Fersht, A. R. & Sato, S. (2004). Phi-Value analysis and the nature of protein-folding transition states. *Proc Natl Acad Sci U S A* 101, 7976-81.

44. Raleigh, D. P. & Plaxco, K. W. (2005). The protein folding transition state: what are phi-values really telling us? *Protein Pept. Lett.* 12, 117-22.

45. Moran, L. B., Schneider, J. P., Kentsis, A., Reddy, G. A. & Sosnick, T. R. (1999). Transition state heterogeneity in GCN4 coiled coil folding studied by using multisite mutations and crosslinking. *Proc. Natl. Acad. Sci. USA* 96, 10699-10704.

46. Hammond, G. S. (1955). A Correlation of Reaction Rates. *J. Amer. Chem. Soc.* 77, 334-338.

47. Matouschek, A. & Fersht, A. R. (1993). Application of physical organic chemistry to engineered mutants of proteins: Hammond postulate behavior in the transition state of protein folding. *Proceedings of the National Academy of Sciences of the United States of America* 90, 7814-8.

48. Dalby, P. A., Oliveberg, M. & Fersht, A. R. (1998). Movement of the intermediate and rate determining transition state of barnase on the energy landscape with changing temperature. *Biochemistry* 37, 4674-9.

49. Fowler, S. B. & Clarke, J. (2001). Mapping the folding pathway of an immunoglobulin domain: structural detail from Phi value analysis and movement of the transition state. *Structure (Camb)* 9, 355-66.

50. Wright, C. F., Lindorff-Larsen, K., Randles, L. G. & Clarke, J. (2003). Parallel protein-unfolding pathways revealed and mapped. *Nature Struct. Biol.* 10, 658-62.

51. Sosnick, T. R., Jackson, S., Wilk, R. M., Englander, S. W. & DeGrado, W. F. (1996). The role of helix formation in the folding of a fully alpha-helical coiled coil. *Proteins* 24, 427-432.

52. Feng, H., Takei, J., Lipsitz, R., Tjandra, N. & Bai, Y. (2003). Specific non-native hydrophobic interactions in a hidden folding intermediate: implications for protein folding. *Biochemistry* 42, 12461-5.

53. Englander, S. W., Sosnick, T. R., Mayne, L. C., Shtilerman, M., Qi, P. X. & Bai, Y. (1998). Fast and Slow Folding in Cytochrome C. *Accts. of Chem. Res.* 31, 737-744.

54. Settanni, G., Rao, F. & Caflisch, A. (2005). Phi-value analysis by molecular dynamics simulations of reversible folding. *Proc Natl Acad Sci U S A* 102, 628-33.

55. Das, P., Matysiak, S. & Clementi, C. (2005). Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. *Proc Natl Acad Sci U S A* 102, 10141-6.

56. Wright, C. F., Steward, A. & Clarke, J. (2004). Thermodynamic characterisation of two transition states along parallel protein folding pathways. *J Mol Biol* 338, 445-51.

57. Wright, C. F., Lindorff-Larsen, K., Randles, L. G. & Clarke, J. (2003). Parallel protein-unfolding pathways revealed and mapped. *Nat Struct Biol* 10, 658-62.

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

59. Sali, A., Shakhnovich, E. & Karplus, M. (1994). How does a protein fold? *Nature* 369, 248-51.

60. Dill, K. A., Fiebig, K., M. & Chan, H. S. (1993). Cooperativity in protein-folding kinetics. *Proc. Natl. Acad. Sci. USA* 90, 1942-1946.

61. Munoz, V. & Serrano, L. (1996). Local versus nonlocal interactions in protein folding and stability--an experimentalist's point of view. *Fold. Des.* 1, R71-7.

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

63. Oztop, B., Ejtehadi, M. R. & Plotkin, S. S. (2004). Protein folding rates correlate with heterogeneity of folding mechanism. *Phys Rev Lett* 93, 208105.

64. Sosnick, T. R., Mayne, L. & Englander, S. W. (1996). Molecular collapse: The rate-limiting step in two-state cytochrome c folding. *Proteins* 24, 413-426.

65. Gromiha, M. M. & Selvaraj, S. (2001). Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J Mol Biol* 310, 27-32.

66. Mirny, L. & Shakhnovich, E. (2001). Protein folding theory: from lattice to all-atom models. *Annu Rev Biophys Biomol Struct* 30, 361-96.

67. Gong, H., Isom, D. G., Srinivasan, R. & Rose, G. D. (2003). Local secondary structure content predicts folding rates for simple, two-state proteins. *J. Mol. Biol.* 327, 1149-54.

68. Zhou, H. & Zhou, Y. (2002). Folding rate prediction using total contact distance. *Biophys J* 82, 458-63.

69. Bai, Y., Zhou, H. & Zhou, Y. (2004). Critical nucleation size in the folding of small apparently two-state proteins. *Protein Sci.* 13, 1173-81.

70. Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. (2003). Assessment of progress over the CASP experiments. *Proteins* 53 Suppl 6, 585-95.

71. Jackson, S. E. (1998). How do small single-domain proteins fold? *Fold. Des.* 3, R81-91.

72. Krantz, B. A. & Sosnick, T. R. (2000). Distinguishing between two-state and three-state models for ubiquitin folding. *Biochemistry* 39, 11696-701.

73. Krantz, B. A., Mayne, L., Rumbley, J., Englander, S. W. & Sosnick, T. R. (2002). Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding. *J. Mol. Biol.* 324, 359-71.

74. Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* 224, 771-782.

75. Krantz, B. A. & Sosnick, T. R. (2001). Engineered metal binding sites map the heterogeneous folding landscape of a coiled coil. *Nature Struct. Biol.* 8, 1042-1047.

76. Brønsted, J. N. & Pedersen, K. (1924). The catalytic decomposition of nitramide and its physico-chemical applications. *Z. Phys. Chem.* A108, 185-235.

77. Leffler, J. E. (1953). Parameters for the description of transition states. *Science* 107, 340-341.

78. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2003). Computational design of a Zn2+ receptor that controls bacterial gene expression. *Proc Natl Acad Sci U S A* 100, 11255-60.

79. Liu, H., Schmidt, J. J., Bachand, G. D., Rizk, S. S., Looger, L. L., Hellinga, H. W. & Montemagno, C. D. (2002). Control of a biomolecular motor-powered nanodevice with an engineered chemical switch. *Nat Mater* 1, 173-7.

80. Goedken, E. R., Keck, J. L., Berger, J. M. & Marqusee, S. (2000). Divalent metal cofactor binding in the kinetic folding trajectory of Escherichia coli ribonuclease HI. *Protein Sci* 9, 1914-21.

81. Kim, C. A. & Berg, J. M. (1993). Thermodynamic beta-sheet propensities measured using a zinc-finger host peptide. *Nature* 362, 267-70.

82. Webster, S. M., Del Camino, D., Dekker, J. P. & Yellen, G. (2004). Intracellular gate opening in Shaker K+ channels defined by high-affinity metal bridges. *Nature* 428, 864-8.

83. Lu, Y., Berry, S. M. & Pfister, T. D. (2001). Engineering novel metalloproteins: design of metal-binding sites into native protein scaffolds. *Chem. Rev.* 101, 3047-80.

84. Higaki, J. N., Fletterick, R. J. & Craik, C. S. (1992). Engineered metalloregulation in enzymes. *TIBS* 17, 100-4.

85. Morgan, D. M., Lynn, D. G., Miller-Auer, H. & Meredith, S. C. (2001). A designed Zn2+-binding amphiphilic polypeptide: energetic consequences of pi-helicity. *Biochemistry* 40, 14020-9.

86. Jung, K., Voss, J., He, M., Hubbell, W. L. & Kaback, H. R. (1995). Engineering a metal binding site within a polytopic membrane protein, the lactose permease of Escherichia coli. *Biochemistry* 34, 6272-7.

87. Vazquez-Ibar, J. L., Weinglass, A. B. & Kaback, H. R. (2002). Engineering a terbium-binding site into an integral membrane protein for luminescence energy transfer. *Proc Natl Acad Sci U S A* 99, 3487-92.

88. Benson, D. E., Wisz, M. S. & Hellinga, H. W. (1998). The development of new biotechnologies using metalloprotein design. *Curr. Opin. Biotechnol.* 9, 370-376.

89. Dwyer, M. A., Looger, L. L. & Hellinga, H. W. (2003). Computational design of a Zn2+ receptor that controls bacterial gene expression. *Proc. Natl. Acad. Sci. U S A* 100, 11255-60.

90. Regan, L. (1995). Protein design: novel metal-binding sites. *Trends Biochem. Sci.* 20, 280-5.

91. Sharp, K. A. & Englander, S. W. (1994). How much is a stabilizing bond worth? *Trends Biochem Sci* 19, 526-9.

92. Sancho, J., Meiering, E. M. & Fersht, A. R. (1991). Mapping transition states of protein unfolding by protein engineering of ligand-binding sites. *J. Mol. Biol.* 221, 1007-14.

93. Fersht, A. R. (2004). ϕ value versus ψ analysis. *Proc. Natl. Acad. Sci. U S A.* 101, 17327-8.

94. Eyring, H. (1935). The activated complex in chemical reactions. *J. Chem. Phys.* 3, 107-115.

95. Liguri, G., Camici, G., Manao, G., Cappugi, G., Nassi, P., Modesti, A. & Ramponi, G. (1986). A new acylphosphatase isoenzyme from human erythrocytes: purification, characterization, and primary structure. *Biochemistry* 25, 8089-94.

96. Stefani, M. & Ramponi, G. (1995). Acylphospate phophohydrolases. *Life Chemistry Reports* 12, 271-301.

97. Krantz, B. A. & Sosnick, T. R. (2001). Engineered metal binding sites map the heterogeneous folding landscape of a coiled coil. *Nat Struct Biol* 8, 1042-7.

98. Thunnissen, M. M., Taddei, N., Liguri, G., Ramponi, G. & Nordlund, P. (1997). Crystal structure of common type acylphosphatase from bovine testis. *Structure* 5, 69-79.

99. Saudek, V., Boyd, J., Williams, R. J., Stefani, M. & Ramponi, G. (1989). The sequence-specific assignment of the 1H-NMR spectrum of an enzyme, horse-muscle acylphosphatase. *Eur J Biochem* 182, 85-93.

100. Taddei, N., Chiti, F., Fiaschi, T., Bucciantini, M., Capanni, C., Stefani, M., Serrano, L., Dobson, C. M. & Ramponi, G. (2000). Stabilisation of alpha-helices by site-directed mutagenesis reveals the importance of secondary structure in the transition state for acylphosphatase folding. *J Mol Biol* 300, 633-47.

101. Jacob, J., Krantz, B., Dothager, R. S., Thiyagarajan, P. & Sosnick, T. R. (2004). Early collapse is not an obligate step in protein folding. *J Mol Biol* 338, 369-82.

102. Zerella, R., Chen, P. Y., Evans, P. A., Raine, A. & Williams, D. H. (2000). Structural characterization of a mutant peptide derived from ubiquitin: implications for protein folding. *Protein Sci.* 9, 2142-50.

103. Munoz, V. & Serrano, L. (1997). Development of the multiple sequence approximation within the AGADIR model of alpha-helix formation: comparison with Zimm-Bragg and Lifson-Roig formalisms. *Biopolymers* 41, 495-509.

104. Braman, J., Papworth, C. & Greener, A. (1996). Site-directed mutagenesis using double-stranded plasmid DNA templates. *Methods Mol Biol* 57, 31-44.

105. Krantz, B. A., Dothager, R. S. & Sosnick, T. R. (2004). Erratum to Discerning the structure and energy of multiple transition states in protein folding using psi-analysis. *J. Mol. Biol.* 347, 889-1109.

106. Krantz, B. A., Dothager, R. S. & Sosnick, T. R. (2004). Discerning the structure and energy of multiple transition states in protein folding using psi-analysis. *J Mol Biol* 337, 463-75.

107. Taddei, N., Chiti, F., Fiaschi, T., Bucciantini, M., Capanni, C., Stefani, M., Serrano, L., Dobson, C. M. & Ramponi, G. (2000). Stabilisation of alpha-helices by site-directed mutagenesis reveals the importance of secondary structure in the transition state for acylphosphatase folding. *J. Mol. Biol.* 300, 633-647.

108. Taddei, N., Magherini, F., Chiti, F., Bucciantini, M., Raugei, G., Stefani, M. & Ramponi, G. (1996). C-terminal region contributes to muscle acylphosphatase three-dimensional structure stabilisation. *FEBS Lett* 384, 172-6.

109. Taddei, N., Chiti, F., Magherini, F., Stefani, M., Thunnissen, M. M., Nordlund, P. & Ramponi, G. (1997). Structural and kinetic investigations on the 15-21 and 42-45 loops of muscle acylphosphatase: evidence for

their involvement in enzyme catalysis and conformational stabilization. *Biochemistry* 36, 7217-24.

110. Chiti, F., Taddei, N., Giannoni, E., van Nuland, N. A., Ramponi, G. & Dobson, C. M. (1999). Development of enzymatic activity during protein folding. Detection of a spectroscopically silent native-like intermediate of muscle acylphosphatase. *J Biol Chem* 274, 20151-8.

111. Colubri, A. (2004). Prediction of protein structure by simulating coarse-grained folding pathways: a preliminary report. *J Biomol Struct Dyn* 21, 625-38.

112. Shmygelska, A. (2005). Search for folding nuclei in native protein structures. *Bioinformatics* 21 Suppl 1, i394-i402.

113. Islam, S. A., Karplus, M. & Weaver, D. L. (2002). Application of the diffusion-collision model to the folding of three-helix bundle proteins. *J. Mol. Biol.* 318, 199-215.

114. Bashford, D., Weaver, D. L. & Karplus, M. (1984). Diffusion collision model for the folding kinetics of the phage lambda repressor operator binding domain. *J biomol struct dyn* 1, 1243-1256.

115. Bashford, D., Cohen, F. E., Karplus, M., Kuntz, I. D. & Weaver, D. L. (1988). Diffusion-collision model for the folding kinetics of myoglobin. *Proteins struct funct genet 4 (3)* 4, 211-227.

116. Jeng, M. F., Englander, S. W., Elove, G. A., Wand, A. J. & Roder, H. (1990). Structural description of acid-denatured cytochrome c by hydrogen exchange and 2D NMR. *Biochemistry* 29, 10433-7.

117. Maxwell, K. L., Wildes, D., Zarrine-Afsar, A., De Los Rios, M. A., Brown, A. G., Friel, C. T., Hedberg, L., Horng, J. C., Bona, D., Miller, E. J., Vallee-Belisle, A., Main, E. R., Bemporad, F., Qiu, L., Teilum, K., Vu, N. D., Edwards, A. M., Ruczinski, I., Poulsen, F. M., Kragelund, B. B., Michnick, S. W., Chiti, F., Bai, Y., Hagen, S. J., Serrano, L., Oliveberg, M., Raleigh, D. P., Wittung-Stafshede, P., Radford, S. E., Jackson, S. E., Sosnick, T. R., Marqusee, S., Davidson, A. R. & Plaxco, K. W. (2005). Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci.* 14, 602-16.

118. Garcia-Mira, M. M., Boehringer, D. & Schmid, F. X. (2004). The folding transition state of the cold shock protein is strongly polarized. *J. Mol. Biol.* 339, 555-69.

119. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* 5, 714-720.

120. Gruebele, M. & Wolynes, P. G. (1998). Satisfying turns in folding transitions. *Nature Struct. Biol.* 5, 662-5.

121. Guo, W., Lampoudi, S. & Shea, J. E. (2004). Temperature dependence of the free energy landscape of the src-SH3 protein domain. *Proteins* 55, 395-406.

122. Klimov, D. K. & Thirumalai, D. (2001). Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins* 43, 465-75.

123. Lindberg, M., Tangrot, J. & Oliveberg, M. (2002). Complete change of the protein folding transition state upon circular permutation. *Nature Struct. Biol.* 9, 818-22.

124. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol.* 6, 1016-1024.

125. Weikl, T. R. & Dill, K. A. (2003). Folding kinetics of two-state proteins: effect of circularization, permutation, and crosslinks. *J. Mol. Biol.* 332, 953-63.

126. Yi, Q., Rajagopal, P., Klevit, R. E. & Baker, D. (2003). Structural and kinetic characterization of the simplified SH3 domain FP1. *Protein Sci* 12, 776-83.

127. Guerois, R. & Serrano, L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* 304, 967-82.

128. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of beta-hairpin formation in protein G folding. *Nature Struct. Biol.* 7, 669-673.

129. Kim, D. E., Fisher, C. & Baker, D. (2000). A Breakdown of Symmetry in the Folding Transition State of Protein L. *J. Mol. Biol.* 298, 971-984.

130. Fersht, A. R. (2000). Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc Natl Acad Sci U S A* 97, 1525-9.

131. Kuznetsov, I. B. & Rackovsky, S. (2004). Class-specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors. *Proteins* 54, 333-41.

132. Englander, S. W., Mayne, L. C., Bai, Y. & Sosnick, T. R. (1997). Hydrogen exchange: the modern legacy of Linderstrom-Lang. *Protein Sci.* 6, 1101-1109.

133. Ferraro, D. M., Lazo, N. D. & Robertson, A. D. (2004). EX1 hydrogen exchange and protein folding. *Biochemistry* 43, 587-94.

134. Sivaraman, T., Arrington, C. B. & Robertson, A. D. (2001). Kinetics of unfolding and folding from amide hydrogen exchange in native ubiquitin. *Nature Struct. Biol.* 8, 331-3.

135. Makarov, D. E. & Plaxco, K. W. (2003). The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.* 12, 17-26.

136. Debe, D. A., Carlson, M. J. & Goddard, W. A., 3rd. (1999). The topomer-sampling model of protein folding. *Proc Natl Acad Sci U S A* 96, 2596-601.

137. Miller, E. J., Fischer, K. F. & Marqusee, S. (2002). Experimental evaluation of topological parameters determining protein-folding rates. *Proc Natl Acad Sci U S A* 99, 10359-63.