ABSTRACT

Title: **BARRIER HEIGHTS AND DIFFUSION COEFFICIENTS IN PROTEIN FOLDING**

Athi Narayanan Naganathan, Ph.D., 2007

Directed By: Associate Professor Victor Muñoz, Department of Chemistry & Biochemistry

A widely held view with respect to the folding of single-domain proteins is that they are two-state. In other words, it is seemingly sufficient to invoke just two thermodynamic macrostates – folded and unfolded – to explain the experimental data with a transition-state like picture. Unfortunately, a chemical two-state model and the resulting conventional analyses do not estimate the barrier height which is essential in determining whether protein folding can be approximated as a two-state, all-or-none transition. However, the energy landscape theory of protein folding predicts small and even zero folding free energy barriers (downhill folding) because of partial or complete compensation between large enthalpic and entropic terms as folding proceeds. They have been recently validated by the thorough experimental characterization of proteins that fold globally downhill (BBL) and those that fold over marginal free energy barriers.

In light of these findings, the question of whether this observation is an exception or merely the tip of the iceberg assumes primary importance. Analyzing the

experimental data on previously characterized proteins with statistical mechanical models, it is shown here that the barrier to folding are indeed small and the folding phase space can be quantitatively classified into four regimes – global downhill, marginal barrier, twilight-zone and two-state like. The average effective diffusion coefficient to folding (D_{eff}) is predicted to be strongly temperature dependent changing from 1/(20-25 μs) at 298 K to 1/(2 μs) at ~330-340 K. The activation term on the D_{eff} is found to scale linearly with the protein size while the folding rates themselves scale inversely with the square root of protein length. This work further highlights the importance of baselines and proposes additional thermodynamic and kinetic signatures of downhill folding. A comprehensive experimental and theoretical characterization of PDD, a structural and functional homolog of BBL is also presented. The results indicate that PDD folds downhill at 298 K while crossing a marginal barrier at the apparent T_m. The evolutionary conservation of downhill folding indirectly suggests that this folding behavior has a functional consequence. In short, this work underlines the need for a fundamental shift towards physical models in characterizing protein folding processes.

BARRIER HEIGHTS AND DIFFUSION COEFFICIENTS IN PROTEIN FOLDING

By

Athi Narayanan Naganathan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2007

Advisory Committee:
Associate Professor Victor Muñoz, Chair
Professor George Lorimer
Associate Professor David Fushman
Assistant Professor Daniel Kosov
Professor Marco Colombini

Dedication

To Raghu

Acknowledgements

I thank my advisor, Dr. Victor Muñoz for offering me an opportunity to work in his lab, for laying the foundations to my understanding of protein folding and guidance. I have benefited immensely from the many insightful discussions we have had and his unique, original and meticulous approach to science.

I would like to acknowledge the contributions of my colleagues and collaborators - Dr. Urmi Doshi for the development of the statistical mechanical model (DM model), Dr. Li Peng for the kinetic experiments, and Drs. Jose M. Sanchez-Ruiz and Raúl Perez-Jimenez from the University of Granada, Spain for the calorimetry experiments.

I am also greatly indebted to the other past and current members of the Muñoz group, particularly, Mourad, Tanay, Luis, Jianwei, Rani, Murugan, Christina, Adam, and Raquel, for providing a wonderful work atmosphere and making my stay thoroughly enjoyable. Special thanks to my good friends Nishanth, Anil, Tanay, and Satish for putting up with me over these years.

Many thanks to the University of Maryland Graduate School, Department of Chemistry and Biochemistry and Dr. Herman Kraybill for the fellowships, the committee members for their comments and suggestions, and the staff for assistance with oft-cumbersome paperwork and orders.

Finally, I would like to thank my parents Bagyam and Naganathan, my brother Sriram, my cousin Prasanna, my grandmother, and the rest of my family for their constant encouragement and support. Thank you for everything.

Table of Contents

Symbols and Abbreviations

D_{eff}	Effective diffusion coefficient
τ_{min}	Minimal folding time; $1/D_{eff}$
k_f, k_u	Folding and unfolding rate constants
τ_f, τ_u	Folding and unfolding time constants
$\Delta G^{\dagger}$, β	Barrier height
N	Protein length/size
T	Temperature
T_m	Apparent midpoint temperature
T_0	Characteristic temperature
H	Enthalpy
ΔH_m, ΔS_m	Equilibrium enthalpy and entropy change at the T_m
ΔH_{Cal}, ΔH_{vH}	Calorimetric and vant Hoff enthalpy
C_p	Heat capacity
ΔC_p	Heat capacity change upon unfolding
C_m	Chemical midpoint
ε	Extinction coefficient
f	Asymmetry factor
m_{kin}, m_{eq}	Sensitivity to chemical denaturation from kinetics and equilibrium
[θ]	Mean-residue ellipticity
RC	Reaction coordinate
DSC	Differential scanning calorimetry
CD	Circular dichroism
UV	Ultra-violet
FRET	Förster resonance energy transfer
IR	Infra-red
FTIR	Fourier transform infra-red
NMR	Nuclear magnetic resonance
FCS	Fluorescence correlation spectroscopy
VB	Variable Barrier
DM	Doshi-Muñoz
QY	Quantum yield
NALA	Naphthyl alanine
E_T	FRET efficiency
ASA	Accessible surface area
BLAST	Basic Local Alignment Search Tool
TFA	Tri-fluoro acetic acid
T-jump	Temperature-jump

1. Introduction and Research Objectives

1.1 Introduction

Ever since Anfinsen's seminal work on RNase[1], one of the biggest unsolved problems in science is the prediction of the three-dimensional structure of a protein from its amino-acid sequence. This has assumed even more importance in the 'post-genomic era' with sequences being churned out at an astonishing rate. However, this is far from being a trivial problem. The building blocks of proteins – amino acids – vary significantly in their in their size and chemical nature. Apparently unrelated sequences are therefore able to fold to the same final structure. In other words, Nature utilizes this chemical diversity to choose one among the various ways to combinatorially pack residues while at the same time satisfying functional, geometric, thermodynamic and kinetic constraints.

A direct offshoot of this complexity is the need to understand the various physico-chemical forces that guide the folding of a protein. Identifying the basic rules also enables the development of *ab-initio* methods for protein structure prediction purely based on physical principles rather than the widely used 'knowledge-based' potentials[2]. In this aspect, deciphering the mechanistic details of the folding process has been an area of intense research. But even a moderately sized protein spans ~300 residues and in many cases possesses distinct domains that fold independent of one another. Therefore to simply the experimental signals and analysis, significant attempts have been made to dissect the factors that determine the stability and folding kinetics of smaller globular proteins or individual domains of much larger proteins

that typically span a size range of 30-150 residues. But the sheer number of conformations a given sequence can adopt immediately highlights the magnitude of the problem. This is encompassed in the so-called Levinthal's paradox[3] that states that a protein cannot fold to its thermodynamic free energy minimum within a biologically relevant time if it samples all possible conformations randomly. However, protein folding rates span about 9 orders of magnitude from microseconds to minutes clearly suggesting that the search process is not entirely random.

1.1.1 Energy Landscape Theory

An attempt to answer the Levinthal's paradox led to a series of groundbreaking papers from the group of Peter Wolynes in the late 1980s and early 1990s. Deriving concepts from condensed matter physics, they envisaged the folding process to occur in a hyper-dimensional space; the dimensions correspond to the available degrees of freedom of backbone and sidechain atoms of the constituent residues of the protein chain [4-6]. This landscape can be visualized in three dimensions using two effective degrees of freedom - radius of gyration and the degree of similarity to the folded structure, for example. The resulting landscape of a protein would be funnel shaped with the width representing the conformational entropy and height the solvent averaged free energy. The bottom of the funnel is populated by an ensemble of structures characterized by conformations with low energy and entropy and large degree of similarity to the fully folded state. Partially folded structures occupy successively higher energy tiers of the funnel while the completely unfolded state marked by structures with highest entropy and energy sit at the top.

This treatment partially solves the Levinthal's paradox as here the unfolded protein does not search for its native state at random. Every stabilizing interaction takes an unfolded or partially folded molecule closer to the folded state on an average thus effectively guiding the search process. In other words, folding can be visualized as a stochastic thermal energy driven process in which the unfolded molecules 'flow' down the funnel with the loss in conformational entropy being partially or fully compensated by the gain in energy. A given protein sequence can then find its thermodynamic minimum by choosing any of the innumerable microscopic routes from the top to the bottom in contrast to chemical reactions that typically involve a well-defined pathway. This is turn suggests that protein folding can be appropriately described only when ensembles are considered as any hyper-dimensional plane would reveal molecules with varying degrees of structure. In fact, the fundamental reason for an 'ensemble view' derives itself from the statistical nature of the protein molecule.

The rate of folding is determined by two factors: the average free energy gradient of the funnel and the degree of 'frustration' in the protein molecule. Interpretation of the effect of a gradient on the rate of a process is straightforward - a higher slope would speed up the search for the minimum and *vice versa*. The idea of frustration is a concept borrowed from the theory of glasses and polymer systems. As a protein folds it repeatedly makes and breaks a number of non-covalent interactions. A native contact (defined as those interactions present in the fully folded structure) will push the molecule down the funnel, but any non-native interaction will place the molecule at a relatively higher energy subspace. Folding is therefore impeded as the

molecule has to reconfigure to break the non-native contact. This slowing down due to internal friction effects and the competition between native/non-native interactions is termed 'frustration'. It can be thought of as bumps on the 3-D representation; ruggedness and roughness are two other terms that convey the same meaning. Along these lines, one of the predictions of energy landscape theory is the 'principle of minimal frustration' that emphasizes that the folding landscape of natural proteins have been evolutionary selected to reduce the level of roughness to enable folding within biologically relevant times. Evidence for this primarily comes from lattice models of proteins with random heteropolymers showing a high degree of frustration and a non-unique thermodynamic minimum. Recent experiments on a designed protein Top7 by Baker and co-workers revealed high degree of kinetic complexity with multiple phases in contrast to traditional single-exponential kinetics observed in natural proteins, thus lending strong support to the idea of minimal frustration[7].

1.1.2 Predictions from Energy Landscape Theory

1.1.2.1 Reaction Coordinates

The high dimensionality of folding landscapes however poses a problem. It is challenging to analyze experimental data using multi-dimensional free energy surfaces. However, energy landscape theory predicts that it is possible to resolve folding mechanisms as a function of few appropriately chosen reaction co-ordinates (RC) especially since proteins are minimally frustrated. Therefore, attempts have been made to characterize folding process with simple one-dimensional RCs. The ability of a single RC to capture to the essential features of folding was first demonstrated in the analysis of cubic lattice simulations of protein-like

heteropolymers by Onuchic and co-workers[8]. Such one-dimensional surfaces have also been successful in predicting the folding rates of proteins from 3-D structures[9], explaining complex kinetic behavior of helix-coil transitions[10] and β-hairpin kinetics[11] and the results of protein engineering experiments[12]. Thermodynamically or structurally motivated RCs like the fraction of native contacts (Q), radius of gyration (R_g), and number of ordered residues (N) are the preferred RCs in molecular dynamics (MD) simulations and statistical mechanical models of proteins (see below). P_{fold}, a kinetic RC defined as the probability of a particular conformation to reach the folded state before reaching the unfolded state has also been widely used[13]. However, the applicability of P_{fold} is restricted to MD simulations and requires exhaustive sampling. Therefore, the discussion below will pertain to single but well-defined structural/thermodynamic reaction co-ordinates.

1.1.2.2 Folding Mechanisms – Free Energy Barriers

The landscape theory emphasizes that the folding/unfolding barriers in low-dimensional projections are bound to be small compared to the activation terms of the order of *few hundred* kJ mol^{-1} common to chemical reactions[5]. This is primarily due to large compensations between stabilization energy and conformational entropy along the reaction co-ordinate as a protein folds. Effectively, it predicts two folding scenarios – global two-state and downhill to two-state transitions. A global two-state process is one in which the protein folds over a significant free energy barrier under all degrees of denaturational stress that includes temperature, chemical denaturants, pressure, pH *etc* (Figure 1.1A). Under these conditions, the population is always well

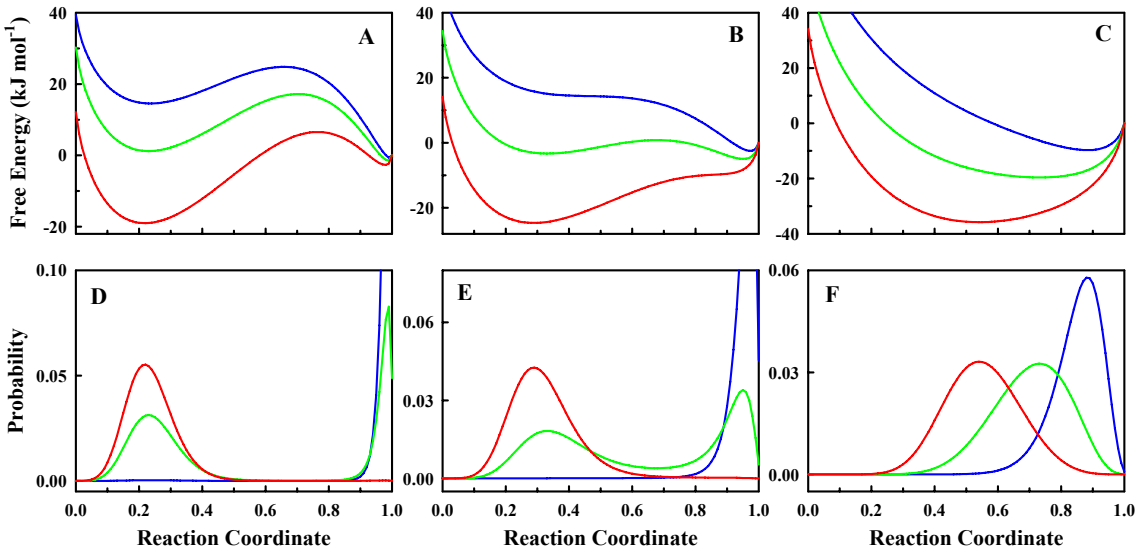


Figure 1.1 Simulated free energy profiles and probability densities for the three folding mechanisms at temperatures below (blue), at (green) and above (red) the apparent T_m. A & D) Two-State, B & E) Marginal Barrier, C & F) Global Downhill.

separated into folded and unfolded ensembles (i.e. bimodal distribution) with no accumulation of partially folded structures under equilibrium (Figure 1.1D). The second scenario is more complex and it suggests that under conditions of extreme native bias (low temperature, no denaturant etc) some proteins might not encounter any significant barrier to folding and the process proceeds 'downhill' driven by the gradient in free energy (Figure 1.1B). The population distribution is unimodal under these conditions and is located at larger values of the order parameter. Bimodal distribution (i.e. a barrier) is restored under denaturational stress which again shifts to unimodal at higher stress but with the population concentrated at smaller values of the order parameter (Figure 1.1E). The barriers in this case are bound to be much smaller than a typical two-state system (folding over marginal barriers).

On a parallel front, considerable advances had been made in the application of structure-based statistical mechanical models to protein folding[9,11,14]. Apart from

possessing a significant predictive power, these models have the distinct ability to quantitatively explain experimental results. A related statistical model that does not incorporate any structural information was developed by Zwanzig to explain the general kinetic and thermodynamic properties of two-state protein folding[15]. Using a variant of this model Muñoz and co-workers proposed an additional folding scenario – global downhill or one-state folding[16]. This is mechanistically different from either of the processes discussed above. In global downhill folding the population distribution is necessarily unimodal at all native bias with the statistical ensemble shifting continuously from high degree of order under folding conditions to low degrees under unfolding conditions (Figures 1.1C & 1.1F). In other words, the various partially folded sub-ensembles that determine the folding to a specific structure are sufficiently populated at one condition or the other and hence this situation is diametrically opposite to two-state folding.

1.1.2.3 Dynamics

Given the success of one-dimensional free energy surfaces it is then possible to explain protein folding kinetics as diffusion along one such RC while employing a transition-state like expression

$$k = D_{eff} \exp(-\Delta G^{\dagger} / RT) \qquad (1.1)$$

where k is the observed rate constant at a temperature T, $\Delta G^{\dagger}$ is the activation free-energy to folding, R is the gas constant and D_{eff} the pre-factor also known as pre-exponential or the effective diffusion coefficient (also $\tau_{min} = 1/D_{eff}$, the minimal folding time). Equation 1.1 is applicable only when there is a barrier. For downhill

folding systems $\Delta G^{\dagger}$ equals zero and hence the observed rate constant would directly correspond to the effective diffusion coefficient to folding.

In chemical reaction kinetics the pre-factor can be derived from first principles as $k_B T/h$ or $1/(0.2\ ps)$ and corresponds to the fundamental frequency of bond vibrations thus allowing for a direct estimate of the activation barriers. Application of equation 1.1 to protein folding however requires a precise knowledge of the elementary motions and roughness that determine the rate of folding. The fundamental motions include peptide bond rotations and large scale concerted movement of residues that are intricately linked to the making and breaking of multitude of non-covalent native and non-native interactions (*i.e.* roughness). All of these are further significantly influenced by frictional solvent collisions and hence temperature dependent. The resultant highly damped motions of the protein chain imply that there will be multiple re-crossings of the barrier (when there is one) unlike chemical reactions that assume an attempt frequency of unity. The multiple re-crossings have in fact been observed in cubic lattice simulations[8]. This in turn underlines the fact that pre-factors to protein folding reactions are complex functions of not only these temperature-dependent motions and interactions but also the reaction co-ordinates used in the analysis. This is because as a protein folds or gets more compact the reconfiguration time correspondingly increases due to the large number of interactions that has to be broken. All of these can be thought of as hyper-dimensional bumps on a multi-dimensional surface as earlier discussed; but along a single reaction co-ordinate these effects are lumped into the effective diffusion

coefficient. More concisely, D_{eff} has been predicted and shown (from lattice simulations) to exhibit a super-Arrhenius scaling with temperature[8]

$$D_{eff} = k_0 \exp(-\Delta E^2 / (RT)^2) \tag{1.2}$$

where ΔE^2 corresponds to the variance in energy or the roughness and k_0 is an elementary rate constant, compared to the Arrhenius dependence for simple activated processes.

1.2 Results from Experiments

1.2.1 Intractability of D_{eff}

Convenient as equation 1.1 may be, the inherent correlation between D_{eff} and $\Delta G^{\dagger}$ however poses a considerable problem - one needs to know the magnitude of $\Delta G^{\dagger}$ or D_{eff} *a priori* to estimate the other. Inspired by chemical reaction kinetics, initial attempts at estimating barrier heights relied on rate measurements of elementary processes in view of setting physical bounds to D_{eff}. Unfortunately, such experimental predictions of D_{eff} vary by over 5 orders of magnitude. They include 1/(1-10 ns) for peptide bond rotations extracted from mechanistic analysis of α-helix and β-hairpin kinetics[11,17], 1/(7 – 250 *n*s) for end-to-end contact formation in disordered peptides[18-22] and 1/(0.1 μs) for BBL[23] to 1/(40 μs) for cytochrome C[24] collapse processes, respectively. There is also ambiguity over whether such piecewise estimates are good approximations of D_{eff} as the roughness, sequence- and size-dependent effects have to be taken into account. Several empirical estimates also reveal pre-factors in the range of 1/(0.1 – 5 μs) [25,26] with single-molecule

measurements setting an upper limit of 1/(200 μs) for CspB[27]. The broad range of predicted D_{eff} values translate to a large uncertainty in $\Delta G^{\dagger}$ of ~30 kJ mol^{-1} severely limiting the applicability of these numbers[28]. Thus, it precludes any unequivocal characterization of the statistical nature of the transition.

1.2.2 Paradoxical Nature of Apparent Two-State Folding

As a result, the experimental folding literature is dominated by observations of two-state folding with a simple chemical model [29-31]

$$F \rightleftharpoons U \tag{1.3}$$

where F and U stand for the native and unfolded states, respectively, seemingly sufficient in explaining experimental data. Kinetically, the barriers separating the grounds states are assumed to be large with the maximum corresponding to an apparent structurally defined transition state in analogy to chemical reactions. This chemical picture is therefore contradictory to energy landscape theory. Moreover, there is a large emphasis on protein engineering experiments to enable comparison of relative rates and stabilities thereby conveniently eliminating the uncertainty in D_{eff} and hence the barrier height[32]. In other words, the experimental data are forced to comply with a two-state model. Evidence for a two-state mechanism stems from observations of sigmoidal unfolding transitions (as a function of temperature or denaturant) in equilibrium experiments indirectly suggesting that they could be represented as a linear combination of folded and unfolded populations whose signals are arbitrarily defined as baselines. Multiple spectroscopic probes reveal identical melting temperatures (T_m; the temperature at which folded and unfolded states are

equally populated in a two-state system) but true signals in the absence of baselines are rarely reported. Kinetic experiments characterized by single-exponential relaxations have been traditionally interpreted as signature of barrier crossing events. However series of recent papers indicate that the presence of a barrier guarantees a single-exponential but not *vice versa*[16,23,33]. Single-molecule experiments employing freely diffusing protein molecules labeled with FRET (Förster Resonance Energy Transfer) pairs have been partially successful in providing evidence to the 'two-state' unfolding nature of a few proteins – two peaks corresponding to the folded and unfolded states' FRET distribution are evident typically with significant overlaps at the chemical midpoint[34]. Though informative, it does not give an estimate on the barrier height as even small barriers might results in two distributions with overlaps (see the probability distribution in Figure 1.1E). In the absence of denaturants a unimodal FRET distribution is observed which is mainly due to the fact that under folding conditions there are not enough molecules sampling the unfolded state. In other words, even these experiments fail to provide information on the nature of the ensemble under functionally relevant conditions.

This universality of two-state protein folding evidenced by its ability to explain most experimental data with a chemical two-state model is not only at odds with theory that predicts small barriers, heterogeneous folding and various folding mechanisms, it is also quite unexpected as the constituent secondary structures themselves reveal a high degree of thermodynamic and kinetic complexity. More specifically, helix-coil transitions are non-two-state with a distribution of helix lengths populating any equilibrium condition[35,36]. Site specific ^{13}C labeling studies

using Fourier Transform Infrared Spectroscopy (FTIR) studies further indicate that the center of helices differ in their T_m by more than 5 K compared to the termini[37]. Temperature jump experiments on helix-coil kinetics reveal two phases with the slower phase corresponding to the equilibration between the folded and unfolded ensembles over a free energy barrier of ~8 kJ mol^{-1} and the faster phase representing the diffusive redistribution of helix lengths within the folded well. Interestingly, at the protein level, native state hydrogen exchange experiments that monitor protection factors reveal a wide range of time-scales, stabilities and denaturant dependent changes for several proteins [38] in contrast to a single value expected for a true two-state system. The authors however interpret them as equilibrium intermediates separated by large barriers though alternate explanations based on fluctuations within a single harmonic potential have also been proposed[39]. Also, ^{19}F, ^{15}N NMR and FCS (Fluorescence Correlation Spectroscopy) experiments on IFABP[40-42], time-resolved FRET measurements on Barstar[43], residual dipolar coupling measurements on GB1[44], and FCS measurements on cytochrome C[45] report on significant conformation heterogeneity under equilibrium conditions and the lack of agreement between multiple residue/atom-level probes. In many cases the unfolding has been interpreted qualitatively as 'sequential unfolding'. It is also of interest to note that many of these proteins had been previously labeled as two-state folders.

The results from various experiments indicate that even folded proteins are better defined as ensembles that are in dynamic equilibrium with one another and there is considerable disagreement even between experimentalists as to the nature of the unfolding transition for certain proteins. The discrepancy seems more apparent

higher the experimental resolution at which a particular process is monitored[46]. Therefore, a chemical two-state picture should be rather viewed as an approximation than anything else. The fine line between the various mechanisms has become even more apparent with the recent thorough characterization of proteins that fold over marginal/negligible barriers.

1.3 Global Downhill Folding and Folding Over Marginal Barriers

Muñoz and co-workers working on a 40-residue independently folding helical domain BBL from *Escherichia coli* observed disparate temperature induced unfolding behaviors in equilibrium when followed by techniques that monitor different structural features like Differential Scanning Calorimetry (DSC), far-UV Circular Dichroism (CD), fluorescence and FRET[47]. The apparent melting temperatures were found to vary from 295 K – 335 K clearly illustrating the non-two-state nature of the transition. Exploring the origins of the complex behavior by a structure based statistical mechanical model, they obtained downhill free energy profiles under all conditions (i.e. global downhill folding) thus providing the first unequivocal evidence to the possibility of one-state folding. The complex unfolding behavior (i.e. the spread in T_ms) is therefore a result of the varied partially-folded sub-ensembles that populate at different temperatures. It is worthwhile to note that they did see sigmoidal unfolding behaviors in their experiments strongly suggesting that this should not be used a criterion for two-stateness. Extending this approach to Nuclear Magnetic Resonance (NMR) experiments, they tracked changes in the chemical environment of 156 protons in BBL as a function of temperature[48]. It again revealed a conformationally rich behavior with apparent T_ms spanning more than 50 K.

Interestingly, the T_ms were normally distributed implying that smaller the number of experimental probes the more probable that it falls close to the average T_m for the entire transition. The implication is that multiple structural probes have to be employed to assess the nature of transition unlike a few as is traditionally done in equilibrium measurements. The average atomic unfolding behavior was strikingly similar to that of far-UV CD, underlining the fact that unfolding curves of low resolution experiments (like CD, fluorescence *etc.*) are highly simplified representations of a more complex behavior; this observation further answers the unfolding complexity seen in high resolution experiments of the apparent two-state-like proteins discussed in the previous section. The authors were also able to map the thermodynamic interaction network in BBL providing an unprecedented view on the nature and relative magnitude of interactions that dictate folding in this protein. These results are further supported by a simple Variable Barrier (VB) analysis of DSC thermograms based on Landau model of phase transitions[49]. This model is based on a one-dimensional description with enthalpy as the reaction coordinate. It predicts zero barrier height for BBL at the T_m consistent with the statistical mechanical model. Global downhill folding in BBL has also been computationally validated in coarse-grained and native-centric off-lattice models[50,51]. Double-perturbation experiments involving urea and temperature reveal crossing baselines and non-unique T_{max} (temperature of the maximum signal upon cold denaturation) highlighting the deviation from two-state behavior[52]. To summarize, Muñoz and co-workers have cataloged a set of equilibrium criteria to distinguish between the various mechanisms and particularly for global downhill folding systems[53].

Gruebele and co-workers have been instrumental in developing corresponding kinetic signatures of downhill folding[54]. The kinetics of fast-folding mutants of λ-repressor, an 80-residue α-helical protein, reveals two phases[55]. The amplitude of the slow phase decreased continuously upon addition of co-solvents that stabilize the folded state, and was replaced by increasing amplitude of the fast phase[56]. Taken together, they provide clear evidence to the origins of these phases - the fast phase corresponds to the diffusive downhill motion of activated species (i.e. population at the top of the barrier) and the slow phase to barrier-crossing in analogy to helix-coil transitions. The rate of the fast phase ~1/(2 μs) at 340 K then provides a direct estimate of the D_{eff} for this protein. Plugging this number into equation 1.1 they predicted the barrier height at 340 K to be on the order of ~1.5 RT – the first example of folding over marginal barriers. Similar to the equilibrium signature of probe-dependent T_m reported by Muñoz and co-workers, they observe probe-dependent kinetics when monitored by fluorescence and infra-red T-jump experiments at temperatures lower than the T_m suggestive of downhill folding[57,58]. They have also been successful in engineering λ-repressor to fold globally downhill[59]. However, the probe dependency of kinetics processes and the origins of the fast phase have been challenged in recent works[60,61]. These observations also reveal that equilibrium criteria are more robust in discerning the folding mechanisms than the kinetic counterparts.

1.4 Research Objectives and Chapter Summary

In light of these findings, it is clear that the three folding mechanisms – two-state, downhill to two-state, and global downhill – are prevalent in proteins. The fact

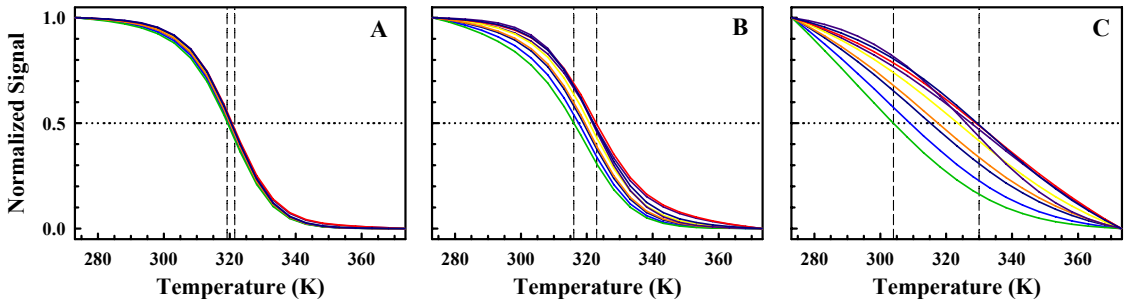


Figure 1.2 Simulated ensemble signals for various probes. A) Two-State, B) Marginal Barrier and C) Global Downhill. The dotted lines correspond to the average signal of 0.5 while the dashed lines represent the spread in apparent T_ms.

that slower folding proteins can be engineered to fold downhill or even globally downhill suggests that the folding barrier of two-state-like proteins is small[5,62]. But why has this situation not been observed before? This is partially a consequence of the nature of ensemble experiments – they report on average properties and not on the distribution of structural features. The ability of these classical experiments to distinguish between different folding scenarios is highlighted in Figure 1.2 where the changes in average signal as a function of temperature are shown for the three different mechanisms discussed in Section 1.1.2.2. A global downhill folding protein produces a spread in apparent T_m of 26 K for various assumed probes. However, as the barrier height increases the spread in observed T_m decreases exponentially to 10 and 1.3 K for barriers of 4 and 15 kJ mol^{-1} at ~320 K, respectively. Since the populations are exponentially sensitive to the free energy (*i.e.* $p \propto \exp(-G/RT)$), the spread in T_ms themselves show a similar relation. Unfortunately, the use of arbitrary baselines in traditional analysis precludes an unequivocal estimate of the apparent T_ms, providing a possible reason for the paucity of examples of proteins that fold over marginal or zero barriers. The exponential decay of the populations with free energy

also suggests that any disagreement between different experimental probes and/or the non-compliance to two-state models in ensemble experiments should be treated with extreme caution[53].

These considerations therefore raise a number of questions. Are there other experimental signatures that could better discern the various mechanisms when employing the same classical techniques? How sensitive are the thermodynamic parameters to the definition of baselines and how much do they influence the results? When can proteins be classified as folding over large or marginal barriers or more precisely, what are the limits in terms of barrier heights? How well do the results of calculations employing different one-dimensional RCs compare against each other? Such questions have assumed even more importance with the recent characterization of a number of fast folding proteins (folding time in the order of microseconds) to extract dynamic and energetic contributions to folding[25]. Effectively, estimates of D_{eff} and $\Delta G^{\dagger}$ without a priori assumptions on the folding mechanism is the only way out. My work presented here will attempt to answer the questions raised above with a quantitative outlook. The chapter organization is as follows.

Chapter 2 provides a general introduction to the experimental techniques commonly used to monitor the structural changes accompanying protein folding/unfolding. It also introduces the basic thermodynamics, notations, conventions and parameters typically employed in a two-state treatment of thermal and chemical denaturation experiments from the perspective of both equilibrium and kinetic measurements.

In **Chapter 3** the importance of protein length in determining the rate of folding is analyzed. The rate is found to scale sublinearly as the inverse of the square root of chain length in the range of 16 - 396 residues with a significant correlation of 0.74. The scaling law is consistent with an earlier prediction from polymer physics arguments[63]. The origin of this scaling is explained here with a simple thermodynamic parameter (n_σ) that in turn provides an indirect estimate of the barrier height to folding. With the folding rate and $\Delta G^\dagger$ available, a D_{eff} value of 1/(1 μs) at 335 K is predicted. The size consideration alone further hints the possibility of downhill folding for proteins of size < 50 residues[64].

In **Chapter 4** the VB model analysis of DSC profiles that had been earlier used to distinguish between global downhill and two-state folding in BBL and thioredoxin, respectively, is introduced[49]. This chapter also broaches on the validity of chemical two-state approximation in protein folding from the view of calorimetry and the importance of baselines. The thermodynamic barrier height for a set of 15 proteins is calculated using VB model. The predicted barrier heights are small in agreement with theory, and are found to vary between -3 to 18 kJ mol^{-1} for this dataset. Moreover, they scale with the rates at ~298 K producing a high correlation of 0.95. An average D_{eff} of 1/(25 μs) at 298 K is computed from this analysis. A clear threshold in folding times of 1 *m*s at 298 K is evident – proteins that fold faster than this time scale are bound to have smaller barriers and chemical two-state treatment of folding breaks down[65].

Recent fast-folding data on proteins are being successfully explained with a chemical two-state model. But these proteins by definition should fold at or near the

downhill regime. **Chapter 5** discusses this apparent paradox using a simple one-dimensional description that has its roots in the Zwanzig's statistical mechanical model (Doshi-Muñoz model – DM model). A systematic deviation from a true two-state behavior is observed upon analyzing the chemical and thermal denaturation data in equilibrium and kinetics of many fast-folding proteins. Additional experimental signatures - the ratio of sensitivity to chemical denaturation in kinetics and equilibrium and the shape of temperature versus relaxation rate plot - are proposed to distinguish between the various folding mechanisms. This theoretical treatment also provides the individual D_{eff} and $\Delta G^{\dagger}$. Moreover, the D_{eff} is predicted to have an activation term that scales linearly as ~1 kJ mol^{-1} with protein length. As a result the D_{eff} at 298 K (~1/(20 μs)) is calculated to be about an order of magnitude lower than the values at ~330-340 K (1/(2 μs)). This analysis also reveals all the regimes predicted by theory – two-state, downhill and global downhill – with precise limits in terms of barrier heights[66].

There have been suggestions that BBL folds anomalously possibly due to the presence of hydrophobic dye used in fluorescent experiments, low ionic strength experimental conditions and shorter sequence constructs[67,68]. **Chapter 6** provides strong evidence that these factors do little to affect the mechanism of folding implying that global downhill folding is a robust property of BBL. It also highlights the ability of baselines to skew the results in chemical and thermal denaturation experiments of small fast-folding proteins. The thermodynamic parameters of BBL from a pseudo-two-state analysis are also found to be consistent with the various thermodynamic scaling laws proposed earlier. It further affirms the fact that two-state

and downhill folding are just extremes of the folding mechanisms predicted by theory[69].

Chapter 7 presents a detailed experimental characterization of PDD, the structural and functional homolog of BBL. Experimental probes that include DSC, far- and near-UV CD, fluorescence, FRET, FTIR and IR T-jump kinetic studies indicate that PDD folds in a non-two-state fashion. Most of the spectroscopic probes show steep pre-transition baselines signaling structural changes even at the lowest temperature explored. Furthermore, they qualitatively suggest that PDD has a marginal barrier at the T_m in comparison with BBL.

In **Chapter 8**, the experimental data of PDD is analyzed with a structure based statistical mechanical model that had earlier been used to explain the complex thermodynamic behavior in BBL. Most of the data are explained without employing arbitrary baselines. It attributes the steep pre-transition slopes to the gradual melting of helices in the protein. The theoretical treatment of unfolding using three different models indicate that PDD indeed folds over a marginal barrier of 2 ± 2 kJ mol^{-1} at 320 K while folding downhill at 298 K. This renders a D_{eff} of ~1/(116 ± 32 μs) at 298 K and ~1/(41 ± 26 μs) at 320 K. The conservation of downhill folding together with a simple phylogenetic analysis indirectly suggests that this folding behavior has a functional implication.

2. Methods, Materials and a Primer on Two-State Analysis

2.1 Methods and Materials

2.1.1 Differential Scanning Calorimetry (DSC)

DSC is a simple yet powerful technique to characterize the temperature-induced changes in partial molar heat capacity of proteins, and hence the global conformational transition. A typical calorimeter has two cells, one for the buffer and one for the protein solution. The cells are simultaneously heated at a constant rate of ~0.5 – 1.5 K min^{-1} while maintaining a zero temperature difference between them. But since the heat capacity of the protein is different from that of the buffer a certain power required to achieve this. The ratio of this power difference (J s^{-1}) to the scanning rate (K s^{-1}) then directly corresponds to the apparent heat capacity of the protein-buffer system, i.e. $\Delta C_p^{app} = C_p^{sol} - C_p^{solv}$ where C_p^{sol} and C_p^{solv} correspond to the absolute heat capacities of the solution (protein + buffer) and solvent (buffer), respectively in units of J K^{-1}. A buffer-buffer baseline with buffer in both the cells is usually measured before and after the scan and is subtracted from the apparent heat capacity of the protein-buffer system to correct for instrumental effects.

A quantity of more relevant interest is the partial molar heat capacity of the protein (C_p^{Prot}) that can be calculated from the expression:

$$C_p^{Prot} = \frac{\Delta C_p^{app}}{C.V_o.10^{-6}} + \frac{V^{Prot}}{V^{solv}} C_p^{solv} \qquad (2.1)$$

where C is the concentration of the protein in *m*M, V_o is the volume of the calorimetric cell in *m*L, and V_{Prot} and V_{solv} are the molar volumes of the protein and solvent, respectively. The latter values are obtained from well-recognized works in the literature[70]. Measuring precise values of C_p^{Prot} is not trivial as it is sensitive to the concentrations used. This is further compounded by the high concentrations (in the *m*M range) used in DSC experiments that might result in non-specific aggregation. To overcome this problem, several scans at different temperatures and protein concentrations are typically done to precisely estimate the *absolute* heat capacity of the protein[71,72]. In future discussions C_p^{Prot} is simply represented as $<C_p>$.

<u>*Current Work*</u> The DSC thermograms shown in Chapters 6 and 7 were measured in collaboration with the group of Prof. Sanchez-Ruiz, University of Granada, Spain. The experiments were carried out with a VP-DSC calorimeter from MicroCal (Northamton, MA) at a scan rate of 1.5 K min^{-1}. Proteins solutions were prepared by exhaustive dialysis against the buffer. Calorimetric cells of volume ~0.5 *m*L were kept under excess pressure of 30 psi to prevent degassing during the scan. All values are reported in absolute heat capacity units. The protein concentrations were in the range of 0.2 – 0.8 *m*g mL^{-1}.

2.1.2 Circular Dichroism (CD)

CD or more precisely electronic CD measures the difference in the absorption of left and right circularly polarized light as it passes through an optically active solution. A protein far-ultraviolet (far-UV) CD spectrum is measured in the wavelength range of 190-250 nm where electronic transitions from the peptide bonds

(amide groups) dominate. The shape and intensity of the bands in this wavelength range are therefore sensitive to the conformation of the protein chain, i.e. on the degree of alignment of the amide transition dipoles with one another. An α-helical spectrum is characterized by three bands: a negative at ~222 nm that corresponds to the excitation of the non-bonding electrons of the carbonyl oxygen to the anti-bonding p orbital (n-π*) and a negative and positive couplet at 208 and 193 due to p-π* parallel and perpendicular components (as a result of exciton coupling) from the delocalized electrons of the amide group. A disordered protein chain has a positive band at ~230 nm and a negative band at ~195 nm from the n-π* and π-π* transitions, respectively. Other secondary structures like β-sheets, turns and loops have their own characteristic signals and are not discussed.

The instrumental output (ellipticity; θ_{obs}) in millidegrees can be expressed as

$$\theta_{obs}(\lambda) = 32.98\Delta A = 32.98(\varepsilon_L - \varepsilon_R).l.C \tag{2.2}$$

where ΔA is the difference in absorbance as a function of the wavelength λ, ε_L and ε_R are the molar absorptivities of the left and right circular polarized light in units of M^{-1} cm^{-1}, l is the pathlength of the quartz cuvette in centimeters (cm) and C is the protein concentration in moles L^{-1}. For comparison between proteins of different lengths and concentrations, the mean residue ellipticity ([θ]) in units of deg cm^2 $dmol^{-1}$ is reported

$$[\theta] = \frac{\theta_{obs}(\lambda)}{n_{pb}.10.l.C} \tag{2.3}$$

where n_{pb} is the number of peptide bonds in the protein.

Aromatic residues (tyrosine and tryptophan) in asymmetric environments, i.e. buried within the protein core or in the vicinity of other asymmetric groups, give rise to near-UV signals in the wavelength range of 250-300 nm. The signal from phenylalanine is weak. The shape, magnitude and sign of the spectrum depend on the degree of hydrophobic burial, coupling to amide transition dipoles and the identity of the chromophore. This provides an additional probe to monitor the conformational changes in the protein at localized environments as opposed to the global nature of DSC and far-UV CD. The signal is reported as in equation 2.3 with n_{pb} replaced by the number of aromatic residues.

Current Work The far-UV CD spectra reported in Chapters 6 and 7 were collected with a 1 mm pathlength quartz cuvette in a Jasco-810 Spectropolarimeter coupled to a Peltier system. Protein concentrations were usually ~50 μM unless stated otherwise. The typical acquisition parameters were: scanning mode – continuous, scanning rate - 10 nm min^{-1}, response time – 16 seconds, and bandwidth – 2 nm. The temperature slope was 3 K min^{-1} with a sample equilibration time of 2 minutes. The near-UV CD spectra were collected in the same instrument with the same parameters, but using a pathlength of 1 cm and protein concentration of ~100 μM.

2.1.3 Fluorescence and Förster Resonance Energy Transfer (FRET)

Fluorescence is the spontaneous emission of a photon from the ground vibrational level of the excited singlet to any of the vibrational energy levels of the ground electronic state. A molecule that absorbs light can fluoresce, but the intensity or whether it fluoresces depends on the nature of absorption and the lifetime of excited state. The latter is determined by competing non-radiative loss of energy due

to collision with solvent molecules or quenching as a result of dipole-dipole interaction with a nearby fluorophore or proton/electron-transfer reactions. Typically, π-π* absorptions result in a strong fluorescence ($\tau \sim 10^{-9}$ s) while n-π* are weak due to the longer lifetime of the excited state ($\tau \sim 10^{-6}$ s). The side-chains of tyrosine and tryptophan residues are conjugated systems with delocalized π electrons, thus resulting in a significant absorption and fluorescence.

In the presence of a large overlap between the emission wavelengths of one fluorophore (donor) and the absorption wavelengths of another (acceptor), and if they are within a certain distance (r) the excited donor can transfer its energy to the acceptor based on a dipole-dipole coupling mechanism. This results in a quenching of donor fluorescence and a sensitized emission of the acceptor. The transfer efficiency (E_T) decays as the 6th power of the intervening dye distance

$$E_T = \frac{1}{1+(r/R_0)^6} \tag{2.4}$$

where R_0 is a characteristic of a donor-acceptor pair and corresponds to the distance at which transfer efficiency is 0.5. R_0 (nm) in turn depends on the quantum yield of the donor (QY_D), refractive index of the medium between the dyes (n = 1.33 for water), the orientation factor (κ^2) and the overlap integral in the region of donor emission and acceptor absorbance (J) as

$$R_0 = 2.11\times10^{-2}(\kappa^2.n^{-4}.J.QY_D)^{1/6} \tag{2.5}$$

with

$$J = \int F_D^{norm}(\lambda)\varepsilon_A(\lambda)\lambda^4 d\lambda \tag{2.6}$$

where F_D^{norm} is the normalized fluorescence of the donor and ε_A is the extinction coefficient of the acceptor. The distance *r* can then be calculated combining equations 2.4, 2.5, 2.6 and

$$E_T = 1 - \frac{QY_{DA}}{QY_D} \tag{2.7}$$

where QY_{DA} is the quantum yield of the donor in the presence of acceptor.

Current Work PDD was labeled with a donor-acceptor pair of naphthyl alanine and dansyl lysine at the C- and N-terminus, respectively. The fluorescence and FRET measurements presented in Chapter 7 were collected with a Flurolog-3 Spectrofluorimeter (Jobin Yovin, Inc.) coupled to a Peltier system using a 1 cm pathlength quartz cuvette. Protein concentrations were ~5 μM. The donor was excited at 288 nm and the fluorescence was collected at 90° to the incident radiation. The excitation and emission slit widths were 2 nm with an integration time of 0.25 seconds. The protein solution was equilibrated for 2 minutes at each temperature before data acquisition. The quantum yield of NATA at pH 7.0 and 298 K – 0.13 – was used a reference for the calculation of donor quantum yields. A κ^2 value of 2/3 was assumed in all calculations.

2.1.4 Fourier Transform Infrared (FTIR) Spectroscopy

FTIR monitors the stretching and bending vibrations of the atoms constituting the protein molecule. A vibration should produce a change in the dipole moment of the constituent bonds to result in IR absorption. They typically involve transitions from the ground vibrational level to higher vibrational levels in the ground electronic

state. The intensity and the frequency at which these molecular motions occur are sensitive to nature of atoms constituting the bonds, the presence or absence of secondary structures (i.e. the alignment of the dipoles) and hydrogen-bonding. The experiments are typically performed in deuterated buffer to reduce the absorbance from water O-H stretch. The final absorbance spectra are calculated using $A = -\log_{10}(I/I_0)$ where I and I_0 are the transmission intensities of the protein and buffer solutions, respectively.

The amide I' region (1600 – 1700 cm^{-1}) is dominated by C=O stretching with minor contributions from C-N stretch and N-H bend (prime denotes the frequency of the deuterated groups) while amide II' regions (1480 – 1575 cm^{-1}) are dominated by C-N stretch and C-N-H deformations. Typically, FTIR spectra of proteins are collected around the amide I' region as it has the strongest intensity. In an α-helix, the backbone carbonyl is involved in hydrogen-bonds with the N-H groups thus giving rise relatively more intense bands compared to other wavenumbers. Though the bands are intense at these wavenumbers, the change is intensity with external perturbants like temperature is small. The spectra are usually represented in reference to some low-temperature spectrum to highlight the changes. The characteristics of the amide I' spectrum is discussed in greater detail in Chapter 7.

Current Work The FTIR spectra recorded in an Excalibur FTS-3000 Spectrometer (BioRad). CaF_2 windows divided by a 50 μm teflon spacer was used as the sample cell. The buffers were prepared in 99.9 % D_2O. The exchangeable protons in the protein sample were substituted with deuterium by double heating-lyophilization cycle. Protein concentration was 2.5 mM.

2.1.5 Kinetics

Proteins that fold in the millisecond time-scale are characterized using the familiar stopped-flow techniques that have a dead-time of ~1 ms. Continuous flow setups have managed to reduced the dead-time to ~10 μs thus enabling the study of faster folding proteins. But the more preferred method for fast folding proteins is the laser temperature jump (T-jump) pump-probe setup. Here, a pump-beam from one laser is used to heat up the solution within a few nanoseconds thus perturbing the equilibrium. The relaxation of the system to the new equilibrium is then monitored by the probe beam. Explained below is the infra-red T-jump setup. The principle is essentially similar for fluorescence kinetics.

IR Kinetics A Continuum Surlite I-10 Nd-YAG laser with 7 ns pulse-width was used as the pump beam. The fundamental of the YAG laser (1064 nm) was shifted by a Raman cell (Lightage, 1 m path length and filled with mixed Argon and Hydrogen with overall pressure of 1000 psi) to ~1900 nm that corresponds to the vibrational absorption of the water bending mode. Heating pulse with ~20 mJ power was used to generate T-jumps in the range of 8 – 10 K. A continuous wave (CW) lead salt diode laser purchased from Laser Components was used as probe beam. A MCT detector from Kolmar Technologies with 50 MHz bandwidth was used to monitor changes in transmission intensity at 1631.8 cm^{-1} as the system relaxes to the new equilibrium. D_2O buffer was used as an internal thermometer to determine the magnitude of the jump. The sample preparation is identical to that of the equilibrium FTIR experiment. Protein concentrations were ~2.5 mM. CaF_2 windows divided by a 50 μm teflon spacer was used as the sample cell. The temperature of the sample cell

was determined by an Aluminum bath controlled by a Thermoelectric (Peltier) Cooling system with ± 0.2 K precision.

2.1.6 Buffer Solutions and Concentration Measurements

The pH 7.0 and pH 3.0 experiments were carried out in 20 mM sodium phosphate and 5 mM Glycine-HCl buffers, respectively. The ionic strength of the protein solutions were corrected to the required values (Chapter 6) using NaCl. Concentrations were determined using the following extinction coefficients (in units of M^{-1} cm^{-1}): $\varepsilon_{280}^{7.0,3.0}$ = 5526 and $\varepsilon_{266}^{7.0}$ = 3595 for naphthyl, $\varepsilon_{280}^{7.0,3.0}$ = 1280 for tyrosine, $\varepsilon_{280}^{7.0}$ = 1571, $\varepsilon_{280}^{3.0}$ = 6517 and $\varepsilon_{266}^{7.0,3.0}$ = 4528 for dansyl, where the superscripts and subscripts denote the pH and wavelength in nm.

2.2 Two-State Analysis

This subsection provides a primer on the basics of two-state analysis of protein folding apart from introducing the various terms that will be heavily used in the forthcoming chapters.

2.2.1 Characterization of a DSC Thermogram

The simulated heat capacity profile of a two-state-like protein is shown in Figure 2.1 (blue circles). The details of the model used in producing the DSC thermogram are presented in Chapter 6. The thermogram is single-peaked with a maximum at ~ 320 K and apparent baselines in the pre- (< 295 K) and post-transition regions (> 345 K). There is a positive heat capacity change upon unfolding suggesting an unfolded state that has a higher heat capacity than the folded state. Since DSC

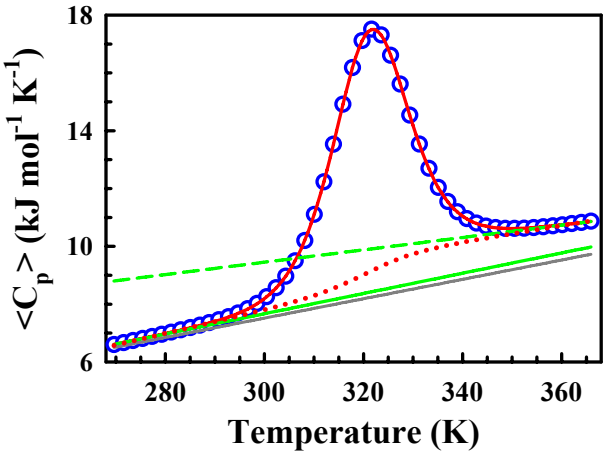


Figure 2.1 Simulated DSC profile of a two-state-like 50-residue protein (blue circles) assuming a native baseline shown in dark gray. Fit to a two-state model is plotted in red along with the folded (continuous green line), unfolded (dashed green line) and chemical baseline (dotted red curve).

measures the changes in heat capacity which in other words is the derivative of enthalpy, it provides a direct access to the partition function of the system under study[73]. A general treatment for a system with an arbitrary number of macrostates or species (I) is presented below followed by the more common two-state analysis. For a N macrostate system

$$I_1 \rightleftharpoons I_2 .....I_{i-1} \rightleftharpoons I_i \rightleftharpoons I_{i+1} .....I_{N-1} \rightleftharpoons I_N \tag{2.8}$$

the partition function (Q) can be written as:

$$Q = \sum_{i=1}^{N} w_i = \sum_{i=1}^{N} \exp\left(-\frac{\Delta G_i}{RT}\right)$$

$$= \sum_{i=1}^{N} \exp\left(\frac{\Delta S_i}{R}\right) \exp\left(-\frac{\Delta H_i}{RT}\right) \tag{2.9}$$

where T is the temperature, w_i the statistical weight and $\exp(\Delta S_i/R)$ and ΔH_i the equivalent of density of microstates and the enthalpy in the traditional statistical mechanical representation of partition function all referenced to a particular state. The temperature-dependent probability (p_i) of each of the states or species can be calculated from

$$p_i = \frac{w_i}{Q}$$

The excess heat capacity $< C_p^{ex} >$ of the system can then be expressed as:

$$< C_p^{ex} >= \frac{d < \Delta H >}{dT} = \frac{d(\sum_{i=1}^{N} \Delta H_i \, p_i)}{dT} \tag{2.10}$$

This function is termed excess heat capacity as it refers to any heat capacity change in excess of the reference state and hence the Δ sign. Differentiating, we get

$$< C_p^{ex} >= \sum_{i=1}^{N} \left(p_i \frac{d\left(\Delta H_i\right)}{dT} + \Delta H_i \frac{dp_i}{dT} \right)$$

$$= \sum_{i=1}^{N} \left(p_i \Delta C_p^i + \frac{\Delta H_i^2 p_i - (\Delta H_i \, p_i)^2}{RT^2} \right)$$

$$=< C_p^i > + \frac{< \Delta H_i^2 > - < \Delta H_i >^2}{RT^2} 2 \tag{2.11}$$

$$=< C_{p,i}^{\text{int}} > + < C_{p,i}^{tr} > \tag{2.12}$$

The first part of above expression is called the intrinsic heat capacity (or chemical baseline) while the second part the transition heat capacity. The intrinsic heat capacity corresponds to changes in the probability weighted heat capacity values of the different species as a function of temperature when the system moves gradually from one state to the other. The transition heat capacity refers to the temperature dependence of the denaturation equilibrium. The area enclosed between the chemical baseline and the heat capacity curve is therefore the total enthalpy realized during the transition and is referred to as the calorimetric enthalpy (ΔH_{Cal}) of the system. It is

independent of any assumption on the number of species involved, but is sensitive to the definition of baselines (see below).

For a chemical two-state system $i = 2$ with folded (F) and unfolded (U) macrostates:

$$F \underset{k_f}{\overset{k_u}{\rightleftharpoons}} U \tag{2.13}$$

where k_f and k_u are the folding and unfolding rate constants. The temperature dependent equilibrium constant ($K(T)$) with the folded state as a reference is

$$K(T) = \frac{[U]}{[F]} = \frac{k_u}{k_f} = e^{\left(\frac{-\Delta G(T)}{RT}\right)} \tag{2.14}$$

The corresponding partition function and folded and unfolded probabilities (p_f and p_u) can be calculated from the following equations

$$Q = 1 + K(T)$$

$$p_f = \frac{1}{1+K(T)} \quad \text{and} \quad p_u = \frac{K(T)}{1+K(T)} \tag{2.15}$$

The intrinsic, transition and observed heat capacity changes are therefore

$$< C_p^{\text{int}} >= \Delta C_p p_u$$

$$< C_p^{tr} >= \frac{\Delta H_{Cal}^2 p_u - (\Delta H_{Cal} p_u)^2}{RT^2} \tag{2.16}$$

$$< C_p >= C_p^f + < C_p^{\text{int}} > + < C_p^{tr} >$$

where ΔC_p is the difference in heat capacity between the folded and unfolded state baselines ($C_p^u - C_p^f$) that are assumed to be linear. Crucial to the estimation of

probabilities is the change in Gibbs free energy of the system as a function of temperature. From the Gibbs-Helmholtz relation,

$$\Delta G(T) = \Delta H_m + \Delta C_p (T - T_m) - T[\Delta S_m + \Delta C_p \ln(T / T_m)] \qquad (2.17)$$

Here ΔH_m and ΔS_m are the reference enthalpy and entropy changes, H_u-H_f and S_u-S_f, respectively, at the denaturation mid-point (T_m). T_m is defined as the temperature at which $p_f = p_u$ and $\Delta G(T_m) = 0$. Two-state characterization of a system then requires the estimation of only ΔH_m, T_m and ΔC_p, as ΔS_m can be directly calculated as $\Delta H_m / T_m$. ΔH_m is also referred to as the van't Hoff enthalpy (ΔH_{vH}) as it is based on a two-state assumption.

DSC is one of the preferred techniques for characterizing a system as all of the above thermodynamic parameters can be unambiguously estimated apart from being able to determine the nature of transition from the so-called 'calorimetric criterion'[74]. From equations 2.10 to 2.17, it is clear that ΔH_{vH} determines the probabilities of the folded and unfolded species as a function of temperature while ΔH_{Cal} determines the area under the peak in excess of the chemical baseline. Therefore, the ratio $\Delta H_{Cal}/\Delta H_{vH}$ can be used to determine the presence or absence of intermediates, or the apparent 'cooperativity' of the unfolding transition. A $\Delta H_{Cal}/\Delta H_{vH}$ ratio of 1 is always interpreted as the hallmark of a two-state system and hence termed co-operative. The ratio greater than or less than one corresponds to either the presence of intermediates or the possibility of a higher order reaction (*e.g.*, aggregation), respectively.

A two-state fit to the profile shown in Figure 2.1 can be carried out in two different ways: by allowing both ΔH_{Cal} and ΔH_{vH} to float or by fixing the ratio

$\Delta H_{Cal}/\Delta H_{vH}$ to 1. Restraining the ratio to 1 is more rigorous and is akin to testing the adherence of the profile to a two-state model while the former can result in different numbers for ΔH_{vH} and ΔH_{Cal}. The latter fit therefore requires 6 parameters: $\Delta H_m = \Delta H_{vH} = \Delta H_{Cal}$, T_m and two parameters each for the folded and unfolded states' linear heat capacity baselines. The fit obtains the baselines by essentially extrapolating the pre- and post-transition regions into and beyond the transition region. The temperature dependence of enthalpy and entropy is ignored in the transition region, so that the ΔH_m can be defined at a single point, i.e. at the T_m. The result of the 6-parameter fit is shown as a red curve in Figure 2.1 while the area between the chemical baseline (dotted red circles) and fit corresponds to ΔH_{Cal}. The folded and unfolded baselines (continuous and dashed green lines) are reasonable with the folded baseline agreeing well with that initially used to simulate the DSC profile (dark gray line). Moreover, the inflection point for the chemical baseline agrees well with the maximum of the thermogram. All of the above results point to a perfect two-state scenario, resulting in the term 'first-order phase transition' to be widely used in describing protein folding reactions.

What is the origin of positive ΔC_p? Model hydrophobic compound transfer studies from apolar to polar solvents as a function of temperature also show a positive ΔC_p. This trend has been successfully explained by the 'ice-berg model' proposed by Frank and Evans[75]. It is based on the idea that solvent molecules are ordered around apolar surfaces especially in their first hydration shell resulting in smaller entropy and large negative enthalpy (because of increased hydrogen bonding) at lower temperatures. As the temperature is increased the 'cage' melts resulting in an increase

in entropy (due to increased solvent fluctuations) and enthalpy (as they are less and less probable to be hydrogen bonded), thus leading to a positive ΔC_p. This result is also supported by statistical mechanical models of water, particularly the Mercedes-Benz (MB) model[76]. As protein unfolding is pictured as the exposure of hydrophobic groups to water that are otherwise buried within the core, the ice-berg model has been extended to these polymer systems to explain the positive ΔC_p.

2.2.2 Equilibrium

2.2.2.1 Thermal Denaturation

In addition to DSC, the thermal unfolding can also be followed by CD, fluorescence, FRET and FTIR. The two-state equilibrium characterization for these techniques is encompassed in the equations 2.13 to 2.15 and 2.17. A typical unfolding curve, i.e. ensemble signal (<*S*>) versus temperature (*T*), is shown in Figure 2.2 highlighting the three transition regimes. It is almost always sigmoidal though the sharpness of transition can vary drastically between proteins and experimental probes employed. In fitting to a two-state model, arbitrary free-floating linear baselines are assumed for the folded (S_f) and unfolded signals (S_u), irrespective of the degree of pre-transition slope. They are meant to represent the temperature dependence of the folded and unfolded signals in either of these wells (non-population weighted). A two-state fit (shown in red) then requires 6 parameters as in DSC: ΔH_m, T_m, and 2 parameters each for the folded and unfolded baselines, with the final signal calculated

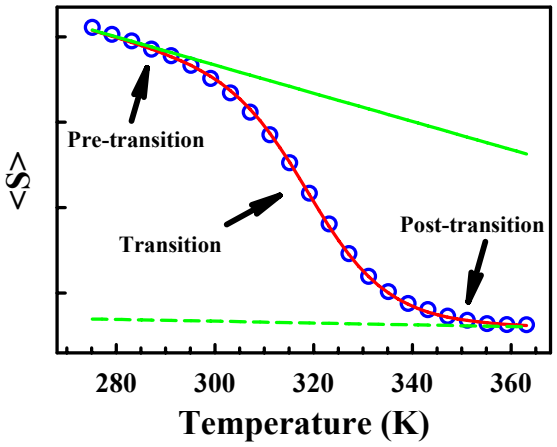


Figure 2.2 Thermal unfolding as monitored by a classical experimental probe like CD, fluorescence or FTIR (blue circles). The fit to a two-state model is plotted in red together with the folded (continuous green line) and unfolded (dashed green line) baselines.

as:

$$< S > (T) = S_f p_f + S_u p_u \tag{2.18}$$

There is however little information for determining the heat capacity change associated with unfolding (ΔC_p) from a denaturation curve shown in Figure 2.2. It is therefore usually estimated from one of the following procedures.

a) Characterizing the system by a DSC experiment provides ΔC_p as a function of temperature from the difference between folded and unfolded heat capacity baselines used in the two-state fit ,

b) Measuring ΔH_m under various stability conditions by changing the pH or ionic strength enables the direct estimation of ΔC_p from the relation

$$\Delta C_p = \frac{\Delta H_m}{T_m}$$

c) A positive ΔC_p curves the plot of stability versus temperature as,

$$\frac{d^2 \Delta G}{dT^2} = -\frac{\Delta C_p}{T}$$

Therefore, the stability decreases at lower and higher temperatures with a maximum value at a temperature in which the entropic contribution to the free

energy vanishes. This phenomenon of low temperature destabilization is termed cold denaturation and is one of the reasons for the widely accepted view of a dominant hydrophobic effect in proteins folding and stability. The cold denaturation temperature is much lower than 0°C under physiological conditions for most proteins. However, it can be the increased by decreasing the stability of proteins through the addition of chemical denaturants. The resulting curvature in the plot of <*S*> versus temperature enables a precise estimation of ΔC_p as all of the parameters are well determined.

d) Empirical analyses that relate ΔC_p to the change in accessible surface area upon unfolding (ΔASA) based on model compound transfer studies or protein datasets also provide an indirect estimate[77]. A linear relation has been observed in such calculations. But the coefficients are notoriously sensitive to the choice of the protein/compound dataset and the algorithm used to estimate the ASA of the unfolded state. Since ΔASA is highly correlated with the protein length (N), ΔC_p also scales linearly with N.

Methods (b), (c) and (d) assume ΔC_p to be independent of temperature.

2.2.2.2 Chemical Denaturation

This is the most widely employed experimental technique to characterize the folding behavior of proteins. In an equilibrium chemical denaturation experiment, the protein ensemble signal (<*S*>, circular dichroism or fluorescence, for example) is measured at various increasing concentrations of either urea or GuHCl ([*D*]). The typical concentration range spans between 0 and 10 M. In a two-state system

represented by equation 2.13, the equilibrium constant (K_{eq}) and corresponding fractions can be extracted from a chemical denaturation curve as:

$$K_{eq}([D]) = \frac{[U]}{[F]} = \frac{k_u}{k_f} = e^{\left(-\frac{\Delta G_{eq}([D])}{RT}\right)}$$

$$p_f = \frac{1}{1 + K_{eq}([D])} \quad \text{and} \quad p_u = \frac{K_{eq}([D])}{1 + K_{eq}([D])}$$

As with thermal unfolding, linear baselines are assumed for folded and unfolded state signals. A number of models have been proposed to explain the changes in stability as a function of denaturant concentration ($\Delta G_{eq}([D])$), the prominent being: denaturant binding model and solvent-exchange model.

The denaturant binding model assumes that there are specific but independent sites (n) on the protein molecule (folded or unfolded) to which the denaturant binds with an effective binding constant k[78]. The equilibrium shifts towards the unfolded state at high denaturant concentrations as it has more binding sites for the denaturant relative to the folded state (Δn). In other words, an increase in the number of potential binding sites exposed in the unfolded state is seen as the reason for denaturation transitions. An elementary treatment results in the following functional form for the stability:

$$\Delta G_{eq}([D]) = \Delta G_{eq}^{H_2O} - \Delta nRT \ln\left(1 + k[D]\right) \tag{2.19}$$

where $\Delta G_{eq}^{H_2O}$ is the stability in water in kJ mol^{-1}. Recent simulation studies by Thirumalai and co-workers support this model[79]. The solvent exchange model (also called the 'weak binding' model or 'selective solvation' model) of Schellman invokes

the idea of an equilibrium between the water molecules bound to independent sites on protein and the denaturant molecules in solution. It has the form:

$$\Delta G_{eq}([D]) = \Delta G_{eq}^{H_2O} - \Delta nRT \ln\left(1 + \left(K - 1\right) X_D\right) \tag{2.20}$$

where K is the equilibrium constant for the exchange reaction and X_D is the mole-fraction of the denaturant in solution[80]. This model addresses the question of whether denaturant molecules actually bind to the protein or they *seem* to be bound just because of the high volume fraction (~20-30 %) used in experiments, i.e. non-specific effects – and hence the term 'weak binding'. Apart from these, other models that take into account the changes in structure of water (solvent) upon the addition of co-solvents have also been proposed. The success of these models that are based on entirely different physical principles suggest the possibility of all the three mechanisms in action simultaneously, but their relative contribution to co-solvent induced denaturation is still unclear.

Intuitively, the difference in the number of binding sites between the folded and unfolded states is directly proportional to the differences in the accessible surface area. This forms the basis for the so-called Linear Energy Model (LEM) which assumes a simple linear dependence of stability on the denaturant concentration[77,81]. The resulting slope of the plot of stability versus the denaturant concentration is called the *m*-value. In pure mathematical terms, *m*-value is the derivative of the change in stabilization free energy upon the addition of denaturant. However, a strong correlation between the accessible surface area (ASA) exposed upon unfolding, i.e. difference in the ASA between the unfolded and folded state of the studied protein

(ΔASA), and the *m*-value has been reported by Pace and co-workers[77]. In view of this observation, the *m*-values are typically interpreted as being proportional to the ΔASA. This 'model' is widely used in interpreting co-solvent induced denaturation. It has the general form:

$$\Delta G_{eq}([D]) = \Delta G_{eq}^{H_2O} - m_{eq}[D] \tag{2.21}$$

and hence

$$\Delta G_{eq}^{H_2O} = m_{eq}[D]_{50\%} \tag{2.22}$$

where $[D]_{50\%}$ or C_m is the denaturation midpoint, *i.e.* the denaturant concentration at which $p_f = p_u$, and m_{eq} is the equilibrium *m*-value in units of kJ mol^{-1} M^{-1}.

2.2.3 Kinetics

2.2.3.1 Thermal Denaturation

The kinetics as a function of temperature can be followed by stopped-flow or laser T-jump techniques for millisecond and microsecond folding proteins, respectively. The relaxation is single exponential for most proteins apparently suggestive of barrier-limited folding (see Chapter 5 for a more detailed discussion). The observed relaxation rate (k_{obs}) for a reversible chemical two-state system can be obtained by solving the time-dependent differential equation:

$$\frac{d[U]}{dt} = k_u[F] - k_f[U]$$

with the constraint [U] + [F] = constant, resulting in

$$k_{obs} = k_f + k_u \tag{2.23}$$

Figure 2.3A shows a simulated temperature dependent relaxation plot of a two-state protein (blue circles). The observed relaxation rate (k_{obs}) has a peculiar behavior wherein it shows a parabolic dependence at temperatures less than the T_m (~350 K) and a linear dependence at higher temperatures, in contrast to a linear Arrhenius dependence observed in chemical reactions involving activated rates. This non-Arrhenius dependence is typically attributed to the temperature dependence of the hydrophobic interaction, i.e. arising out of a large change in heat capacity in going from the unfolded to the transition state[82]. This is the kinetic analogue of cold denaturation observed in equilibrium. The implication is that the degree of hydrophobic burial in the transition state is intermediate to that of the ground states. A typical two-state fit therefore employs a transition-state like treatment of the folding and unfolding reactions using the Eyring's relation that assumes an instant equilibration of populations between ground and transition states:

$$k_{obs}(T) = D_{eff}\left(e^{\left(-\frac{\Delta G^{\dagger-F}}{RT}\right)} + e^{\left(-\frac{\Delta G^{\dagger-U}}{RT}\right)}\right) \tag{2.24}$$

with

$$\Delta G^{\dagger-X}(T) = \Delta H_{\mathrm{m}}^{\dagger-X} + \Delta C_{\mathrm{p}}^{\dagger-X}.(T - T_{\mathrm{m}}) - T[\Delta S_{\mathrm{m}}^{\dagger-X} + \Delta C_{\mathrm{p}}^{\dagger-X}.\ln(T/T_{\mathrm{m}})] \tag{2.25}$$

Here † refers to the transition state, and $\Delta G^{\dagger-X}$, $\Delta H_m^{\dagger-X}$, $\Delta S_m^{\dagger-X}$ and $\Delta C_p^{\dagger-X}$ are the activation terms for free energy, enthalpy, entropy and heat capacity in the folding (X = U) and unfolding (X = F) directions, respectively, while D_{eff} is the effective diffusion coefficient or the pre-exponential. The activation term corresponding to the

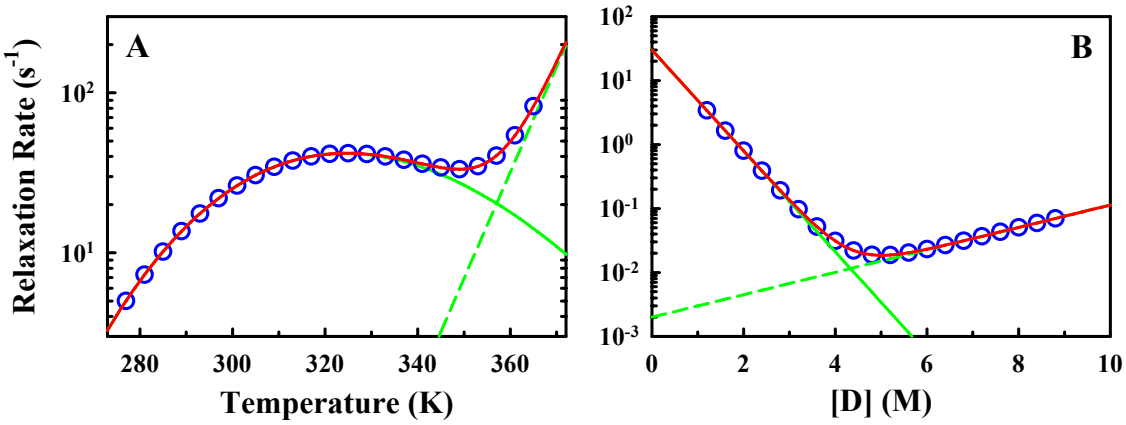


Figure 2.3 Simulated relaxation rates for a two-state-like protein as a function of temperature (A) and denaturant (B), respectively. The continuous and dashed green lines correspond to the folding and unfolding rate constants.

temperature dependence of water viscosity (~16 kJ mol^{-1}) is embedded in $\Delta H_m^{\dagger-X}$. D_{eff} is assumed to be independent of temperature and fixed to a value anywhere in the range between 10^6-10^{10} s^{-1} (see Chapter 1). Therefore the estimated folding and unfolding barriers are singularly dependent on the magnitude of the pre-exponential. Moreover, the fitting procedure is not trivial as the enthalpic and entropic activation parameters are highly correlated. The result of the 6-parameter two-state fit (red curve) with the corresponding folding and unfolding rates (continuous and dashed green lines) are shown in Figure 2.3A. From the fit it is clear that folding rate dominates the curvature. The activation terms are related to their equilibrium counterparts as:

$$\Delta Y_{eq} = \Delta Y^{\dagger-F} - \Delta Y^{\dagger-U}$$

where Y = G, H, S or C_p, thus providing a criterion for assessing the two-stateness of the transition.

An alternative non-committal way is to solve for k_f and k_u using equations 2.14 and 2.23. This analysis is highly error-prone as it employs the equilibrium

populations of folded and unfolded states that are sensitive to the description of baselines. However, this is the preferred scheme for a two-state characterization of kinetic data. In other words, the presence of a large free energy barrier is pre-assumed and the data is forced to comply with two-state expressions. It is important to note that neither of these methods provides an estimate of the barrier height or the pre-exponential to the folding reaction.

2.2.3.2 Chemical Denaturation

The relaxation rate (k_{obs}) is measured at various denaturant concentrations by stopped flow or T-jump apparatus. The resulting plot of k_{obs} versus [D] is usually 'V'-shaped and hence called chevrons (Figure 2.3B; blue circles). This has been traditionally seen as a sign of two-state behavior, though unsubstantiated. In analyzing the chevron plot by a two-state model, the natural logarithm of the folding and unfolding rates (k_f and k_u) is assumed to depend linearly on [D]. Hence,

$$\ln(k_f) = \ln(k_f^{H_2O}) - m_f[D]$$

and

$$\ln(k_u) = \ln(k_u^{H_2O}) + m_u[D] \qquad (2.26)$$

where $k_f^{H_2O}$ and $k_u^{H_2O}$ are the folding and unfolding rates in the absence of denaturant in units of s^{-1}. m_f and m_u are the slopes of folding and unfolding limbs of the chevron in units of M^{-1}. The observed relaxation rate can then be calculated as the sum of the two rates. The fit (red) to this phenomenological two-state model is shown in Figure 2.3B along with the extrapolated folding and unfolding rates. Furthermore, it is possible to estimate the stability from kinetic data for comparison with equilibrium measurements,

$$\Delta G_{kin}([D]) = -RT\ln(k_u / k_f)$$

Therefore, m_{kin} defined as

$$m_{kin} = RT(m_f + m_u) \tag{2.27}$$

in energy units of kJ mol^{-1} M^{-1} should equal m_{eq} for a strict two-state system, thus providing a direct test for conformity to two-state behavior.

2.2.3 Criteria for Two-state Folding

The criteria for identifying two-state folding can therefore be summarized in the following observations:

a) Sigmoidal unfolding transitions upon thermal and/or chemical denaturation,

b) Coincidence of equilibrium unfolding transitions when monitored by various techniques, i.e. identical T_ms,

c) Single-peaked DSC thermograms satisfying the calorimetric criterion of $\Delta H_{Cal}/\Delta H_{vH} = 1$,

d) Single-exponential kinetics under various stability conditions,

e) Agreement between the thermodynamic parameters (ΔH_m, ΔS_m and ΔC_p) from thermal unfolding equilibrium and kinetics,

f) Chevron plots with linear folding and unfolding limbs, and identical sensitivity to the denaturant from equilibrium and kinetics, *i.e.* $m_{kin}/m_{eq} = 1$.

3. Scaling of Folding Times with Protein Size

3.1 Introduction

Protein folding times vary by 9 orders of magnitude. What determines this large spread? One common theme prevalent in the folding literature is that the topological complexity of the protein fold, i.e. the organization of secondary structures and their interconnectivity, determines this variability. It has crystallized into the idea of contact orders[83] and topomer search models[84] to explain the kinetics of folding. Intuitively, the length of a protein should scale with the folding times as longer the length the longer it is bound to search for a given native contact. Predictions from polymer theory in fact suggest that the folding times should scale as the square root of the protein length[63]. The analysis presented in this chapter strongly suggests that the elementary determinant of the folding rate is the protein length with the effects of sequence, structure and topology an outcome of this dependence rather than the cause. A thermodynamic origin of the square root length dependence is also proposed. This enables the estimation of barrier height and pre-exponential to the folding process.

Section 3.2 discusses some of the more successful rate predictors in protein folding and their impact on the field. Section 3.3 revisits the length dependence of folding originally proposed by Thirumalai along with the results from the analysis of 69 proteins. Section 3.4 discusses the relation between fluctuations and heat capacity and hence the thermodynamic parameter n_σ. Section 3.5 provides an estimate of the

folding barrier height for the various proteins, the average diffusion coefficient, and the implications.

3.2 A Brief History

A numbers of models and predictors have been proposed to explain the observed spread in the folding rates. They can be broadly classified into two categories: structure-based predictors that employ structural information derived from X-ray crystallography or NMR and non-structure based predictors that are based on considerations of the protein size and/or sequence. This section provides a brief introduction to representative examples from each category along with an assessment of their impact on the field.

3.2.1 Relative and Absolute Contact Order

Relative contact order (*RCO*) is defined as the average sequence separation between all contacting residues (< 6 Å) in the native structure normalized to the protein length (*N*) [83],

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

where C is the total number of contacts and $\Delta S_{i,j}$ is the separation in sequence between the contacting residues *i* and *j*. Therefore, proteins that have more local contacts should fold faster than those whose structure is dominated by non-local contacts. It is important to note that all non-native interactions that are formed and broken during the folding of a protein are ignored in such a calculation as *RCO* is solely based on the structure of the fully folded protein. Using a database of 12 two-

state folding proteins, Plaxco *et al.* showed that *RCO* provides a significant correlation coefficient (*r*) of 0.81 with the folding rates in water at 298 K[83]. This suggested that the rate of folding is primarily determined by the topological complexity of the fold, i.e. the arrangement of secondary structures, as the authors found no significant correlation with protein length.

However, when applied to a much larger database of 51 proteins that included 27 two-state and 24 multi-state folders, *RCO* failed to predict the folding rates as well resulting in an insignificant correlation of 0.1[85]. But, absolute contact order (ACO) which is essentially the relative contact order corrected for protein length, i.e. $ACO = RCO \times N$, produced a highly significant correlation of 0.74, thus questioning the validity of the conclusions previously made. The high correlation might just have been an artifact of the limited dataset. A number of variants of RCO have also been proposed but possess similar predictive abilities.

3.2.2 Effective Protein Length

While contact order and its variants are based on structure, Ivankov and Finkelstein proposed another measure just based on sequence considerations alone[86]. They proposed that the logarithm of the folding time in water should scale with the effective length of a protein (N_{eff}) as:

$$\log(\tau_f) \sim (N_{eff})^{\gamma}$$

where

$$N_{eff} = N - N_H + a \cdot n_H$$

where N, N_H and n_H correspond to the protein length , number of residues in helical conformation and number of helices, respectively, while a represents the nucleus size of a helix (~4 residues). The exponent γ is a scaling parameter. In other words, this model resorts to the idea of the presence of folding units or foldons in the assumption that helices form much more rapidly (τ ~200 ns) than most of the other secondary structural elements thus requiring the need to factor out their contribution to N. Using a dataset of 64 two-state and multi-state folding proteins, they report a correlation as high as 0.82. They find that the scaling parameter γ can vary anywhere between 0 and 0.5. Moreover, this predictor needs information on N_H and n_H which is obtained from secondary-structure prediction algorithms like PSIPRED.

The success of ACO and N_{eff} is predicting the rates highlight the crucial role of protein size in determining the folding rates. Furthermore, though N_{eff} produces a high correlation, it loses its intuitive appeal when extended to beta-sheet structures as $n_H = N_H = 0$ in which case only the scaling with length is considered. Moreover, the prediction of folding times of multidomain proteins requires additional considerations of protein length. This therefore raises an important question: how much does the folding rate depend on protein length alone? Answering this question provides a much needed yardstick to quantify the effect of topological complexity and so-called foldons in influencing protein folding rates.

3.3 Scaling with Protein Length

The earliest prediction of length dependence of protein folding times (or rates) was made by Thirumalai based on extrapolation from analytic theory of glasses, where he proposed that,

$$\log(\tau_f) \sim N^{\gamma} \tag{3.1}$$

with γ = 0.5[63]. Theoretical treatment of the length dependence by Finkelstein and Badredtinov[87] and later by Wolynes with foundations in the capillary theory of protein folding[88] placed the estimate of γ at 2/3. Results from off-lattice and lattice simulations of Go-like models of proteins lend support to the above arguments thus placing γ anywhere in the range between 1/2 and 2/3[26,89,90].

Figure 3.1 shows a plot of experimentally determined folding times versus $N^{1/2}$ for 69 proteins/peptides in native conditions at 298 K. This dataset is much larger than that previously used to investigate this effect, with protein lengths varying from 16 to 396 residues and incorporates both two-state and three-state folding proteins. Proteins from all structural classes α, β, α+β, α/β are well-represented including the de-novo designed helix bundle α_3D (Table I). It produces a strong correlation of 0.74 with an exponent of 0.5 and approaches ~0.78 when $\gamma \rightarrow 0$. An important implication is that it is possible to predict the folding times to within ~1.1 time decades by just considering the length effects. Interestingly, the obtained correlation values are comparable to that estimated by considering the effective protein length or absolute contact order. Therefore, it suggests that the effect of structure, sequence or the topological complexity on the folding rates is very minimal. The large spread in rates is therefore the result of shorter/longer protein lengths. This observation also debunks the idea of hierarchical folding extrapolated from contact order calculations that local contacts form first followed by long range non-local contacts and so on along a specific pathway. The essence of the contact order however does hold true – local contacts are bound to form faster than non-local ones; but individual protein

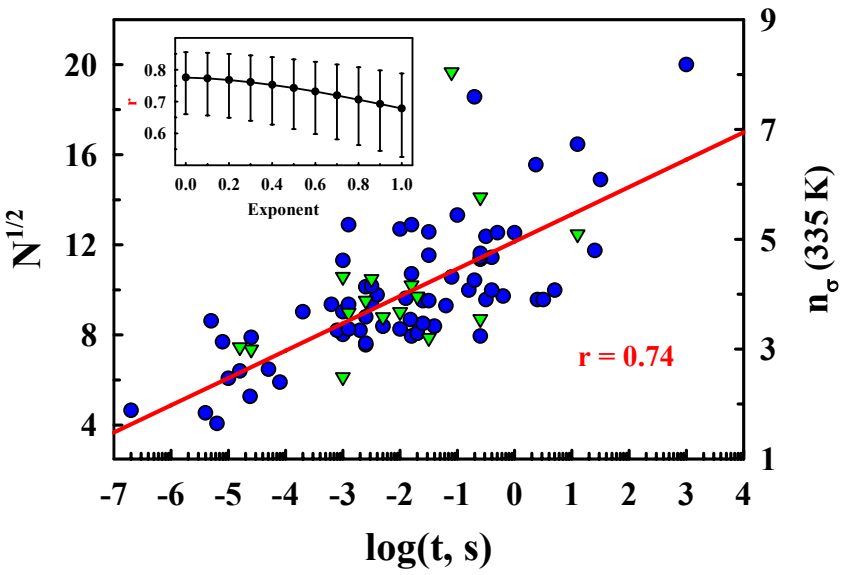


Figure 3.1 (Blue circles) Rate data from 69 proteins plotted as a function of the square root of protein length. The red line corresponds to the linear fit to the data. (Green triangles) n_σ calculated for proteins with thermodynamic data available against the folding times. Inset plots the dependence of correlation coefficient on the magnitude of the exponent to N. (Rate data from published works)

molecules take widely different pathways to reach the folded minimum as predicted by the energy landscape theory.

3.4 $\sqrt{N}$ Dependence from Thermodynamic Arguments

The above scaling provides a ruler to estimate the increase in barrier height per residue on an average. But to have a handle on the magnitude of the barriers an independent estimate on the value of the pre-exponential should be known or vice-versa. In fact, as discussed in Chapters 1 & 2, there have been a number of estimates of the pre-exponential to the folding reaction that vary from 10^5 to 10^{10} s^{-1}. A simple solution would then be to assume a range of pre-exponentials to calculate the barrier height. However, it is possible to do even better using the available thermodynamic

data on proteins and an alternative interpretation of the origins of the positive ΔC_p change upon unfolding.

3.4.1 Revisiting the Origins of Positive ΔC_p

The observation of positive ΔC_p upon protein unfolding is seen as a result of the solvent exposure of hydrophobic groups in the unfolded state (Chapter 2). This solvation view and corresponding correlation to ΔASA upon unfolding, though widely used, has its deficiencies. In a landmark paper Spolar and Record have shown that the change in the accessible surface area upon DNA-binding for a number of proteins is insufficient to explain the observed changes in heat capacity[91]. They attribute the large heat capacity change to a sum of two terms: the change in accessible surface area and any changes in the conformational flexibility of the system upon binding to DNA. Interestingly, DSC studies on α-helical cyclic peptides[92] and β- hairpins[93] show a positive ΔC_p upon unfolding in spite of the complete exposure of their hydrophobic groups to solvent in the native state, as initially noted by Cooper[94]. These observations underline the surprisingly neglected issue of the ability of heat capacity functions to estimate the degree of conformational flexibility. It is more readily apparent if one considers, for example, the change in heat capacity for solid ice to water phase transition. A positive heat capacity change upon temperature increase is observed in this system in spite of no changes in exposure or burial of the water molecules (if these terms can ever be used to describe such transitions). The increase in heat capacity is in fact the result of a larger degree of freedom in the water phase that is able to partition the added heat to different conformational states (more

Table 3.1 Proteins Used in Scaling Analysis

Index	Protein Name	PDB ID	N	$\log(k_f)$	N_σ (335 K)	$\Delta G^\dagger$ (kJ mol^{-1})
1	β-hairpin	1PGB	16	5.2	1.64	3.8
2	Trp-cage	1L2Y	20	5.4	1.84	4.7
3	α-helix	-	21	6.7	1.88	4.9
4	FSD-1	1FSD	27	4.6	2.14	6.3
5	Pin WW domain	1PIN	34	4.1	2.40	8.0
6	Villin headpiece	1VII	36	5	2.47	8.5
7	BBL	2CYU	40	4.8	2.60	9.4
8	PDD	2PDD	41	4.3	2.63	9.6
9	NTL9	1DIV	56	2.6	3.08	13.2
10	Protein G	1PGB	57	2.6	3.10	13.4
11	BdpA	1BDD	58	5.1	3.13	13.6
12	Engrailed	1ENH	61	4.6	3.21	14.3
13	α-spectrin SH3	1SHG	62	0.6	3.23	14.6
14	Protein L	1HZ6	62	1.8	3.23	14.6
15	DNA-binding protein	1C8C	63	3	3.26	14.8
16	src SH3	1SRL	64	1.7	3.29	15.0
17	CI2	2CI2	64	1.7	3.29	15.0
18	CspB (*B. caldolyticus*)	1C9O	66	3.1	3.34	15.5
19	CspB (*T. maritima*)	1G6P	66	2.7	3.34	15.5
20	fyn SH3	1SHF	67	2	3.37	15.7
21	CspB (*B. subtilis*)	1CSP	67	2.9	3.37	15.7
22	Photosystem I accessory protein	1PSF	69	1.4	3.41	16.2
23	CspA	1MJC	69	2.3	3.41	16.2
24	Cro protein	2CRO	71	1.6	3.46	16.7
25	Tendamistat	2AIT	72	1.8	3.54	17.4
26	α_3D	2A3D	73	5.3	3.51	17.2
27	Ubiquitin	1UBQ	76	2.6	3.58	17.9
28	λ repressor	1LMB	80	3.7	3.68	18.8
29	Activation domain of procarboxypeptidase A2	1AYE	80	3.0	3.68	18.8
30	Hpr	1POH	85	1.2	3.79	20.0
31	ACBP	2ABD	86	2.9	3.81	20.2
32	Im9	1IMQ	86	3.2	3.81	20.2
33	Im7	1CEI	87	2.5	3.83	20.5
34	Twitchin Ig repeat 27	1TIT	89	1.6	3.88	20.9

Index	Protein Name	PDB ID	N	$\log(k_f)$	N_σ (335 K)	$\Delta G^\dagger$ (kJ mol^{-1})
35	Barstar	1BRS	89	1.5	3.88	20.9
36	Fibronection 9th FN3	1FNF	90	-0.4	3.90	21.2
37	Tenascin (short form)	1TEN	90	0.5	3.90	21.2
38	SH3 (PI3 Kinase)	1PNJ	90	-0.5	3.90	21.2
39	HypF-N	1GXT	91	1.9	3.92	21.4
40	Twitchin	1WIT	93	0.2	3.96	21.9
41	Fibronection 10th FN3	1FNF	94	2.4	3.99	22.1
42	muscle AcP	1APS	98	-0.7	4.07	23.0
43	common-type AcP	2ACY	98	0.4	4.07	23.0
44	CD2, 1st domain	1HNG	98	0.8	4.07	23.0
45	S6	1RIS	101	2.6	4.13	23.8
46	U1A	1URN	102	2.5	4.15	24.0
47	FKBP12	1FKB	107	0.7	4.25	25.2
48	Barnase	1BNI	110	1.1	4.31	25.9
49	Suc1	1SCE	113	1.8	4.37	26.6
50	Villin 14T	2VIK	126	3	4.61	29.6
51	1LBP	1EAL	127	0.6	4.63	29.9
52	CheY	3CHY	129	0.4	4.67	30.3
53	Lysozyme		129	0.6	4.67	30.3
54	IFABP (rat)	1IFC	131	1.5	4.71	30.8
55	CRBP II	1OPA	133	0.6	4.74	31.3
56	CRABP I	1CBI	136	-1.4	4.79	32.0
57	Apomyoglobin	1A6N	151	0.5	5.05	35.5
58	GroEL apical domain	1AON	155	0.3	5.12	36.4
59	Ribonuclease HI	2RN2	155	0	5.12	36.4
60	P16 Protein	2A5E	156	1.5	5.13	36.7
61	DHFR	1RA9	159	2	5.18	37.4
62	Cyclophilin A	1LOP	164	2.9	5.26	38.6
63	T4 Lysozyme	2LZM	164	1.8	5.26	38.6
64	N-terminal domain PGK	1PHP	175	1	5.44	41.2
65	C-terminal domain PGK	1PHP	219	-1.5	6.08	51.5
66	7-repeat ankyrin protein	1OT8	239	-0.37	6.36	56.2
67	Tryptophan synthase β2-subunit (truncated)	1QOP	268	-1.1	6.73	63.0
68	VlsE	1L8W	341	0.7	7.59	80.2
69	Tryptophan synthase β2-subunit	1QOP	396	-3	8.18	93.1

possibility of bond stretching, bending and breaking due to lesser hydrogen-bonding ability) thus requiring more enthalpy for a unit temperature increase. Also, Cooper has shown that any system that undergoes order to disorder transition, especially those that possess hydrogen-bonded networks (proteins, for example), will show a positive heat capacity change irrespective of the presence or absence of hydrophobic groups[94]. The discussion presented in Chapter 1 also points to a significant structural flexibility inherent to protein systems as earlier predicted by Cooper[95,96]. In fact, the Variable Barrier Model developed to explain the barrierless folding in BBL is based on similar principles (a more detailed discussion of this is presented in Chapter 4) [49].Therefore, hydrophobic surface exposure alone might not be the sole contributor to heat capacity changes.

3.4.2 n_σ

These observations and interpretations comply with the familiar statistical mechanical description of heat capacity (C_p) as the fluctuation in energy or enthalpy (H),

$$C_p = \frac{< H^2 > - < H >^2}{RT^2} \tag{3.2}$$

In protein folding, the calculated ΔC_p assuming a two-state system can also be written as

$$\Delta C_p = \frac{< \Delta H^2 > - < \Delta H >^2}{RT^2} \tag{3.3}$$

where ΔH is the difference in enthalpy between the folded and unfolded states at the temperature T. The function $\sqrt{RT^2 \Delta C_p}$ (from equation 3.3) then corresponds to the

enthalpy fluctuations in the unfolded state in excess of the folded state. This treatment implicitly assumes that the heat capacity of the folded state is primarily a result of non-structural enthalpy fluctuations (see Chapter 4). Therefore, the dimensionless parameter n_σ defined as

$$n_\sigma = \frac{\Delta H(T)}{\sqrt{RT^2 \Delta C_p}} \tag{3.4}$$

signals the frequency at which the enthalpy fluctuations of the unfolded state match the total enthalpy difference, i.e. when they reach the folded state. A small value of n_σ then corresponds to a system whose unfolded states' equilibrium fluctuations are of the same order as the unfolding enthalpy suggesting a marginal or zero free energy barrier. In other words, n_σ is directly proportional to the free energy barrier of the system under consideration. Furthermore, empirical correlations by Robertson and Murphy using a dataset of 49 large proteins report a significant linear correlation of ΔH and ΔC_p with size[97]. Extrapolating this observation to equation 3.4 indicates that n_σ scales as $N^{1/2}$, thereby providing a thermodynamic interpretation for the observed scaling behavior.

3.5 Calculation of Barrier Heights

The scaling of n_σ with size and its direct connection to the equilibrium fluctuations provides the required parameter to extract barrier heights as explained below. However, the temperature dependence of unfolding enthalpy and hence n_σ poses a challenge – at what temperature should n_σ be calculated? For this calculation,

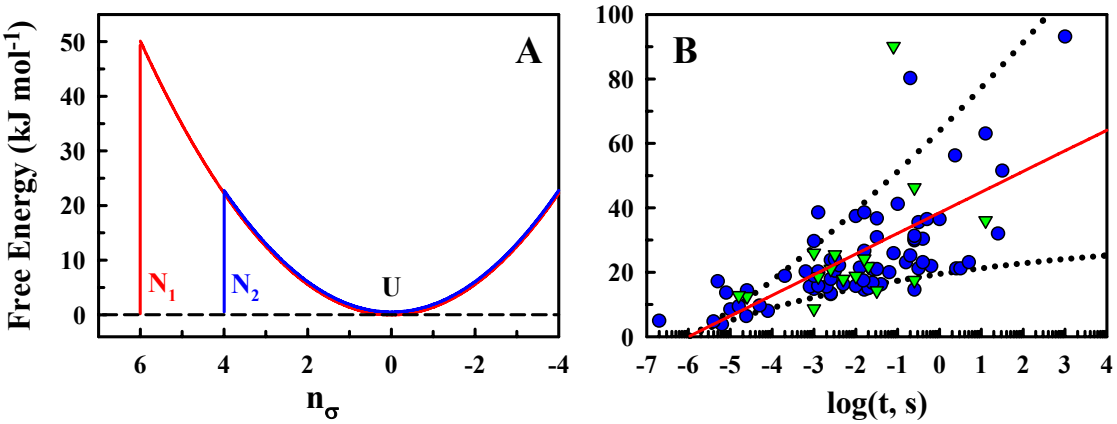


Figure 3.2 A) The one-dimensional harmonic approximation used in the calculation of barrier heights at 335 K. N – folded state and U – unfolded state. B) Free energy barriers versus the folding times for the various proteins shown in Figure 3.1. The red line plots the folding times calculated with a pre-exponential of 1 μs and barrier heights as a function of protein size obtained using ΔH (333 K) = 2.92 kJ mol^{-1} residue^{-1} and ΔC_p = 58 J mol^{-1} K^{-1} residue^{-1}. The dotted line represents the uncertainty in measuring barrier heights one standard deviation above and below the red line. (Rate data from published works).

the length dependence of folding times can be used a ruler. The green triangles in Figure 3.1 show the calculation at 335 K for proteins with both rate and thermodynamic data are available. At this temperature, the slope from n_σ agrees with that of the size correlation slope. At temperatures lower than 335 K the n_σ calculation under-estimates the slope while over-estimating the same at higher temperatures. It suggests that at 335 K n_σ approximates the scaling behavior, possibly as a result of the cancellation of solvent contributions to ΔH and ΔC_p at this temperature[97]. This provides a simple way to compute barrier heights with a mean-field approach employing n_σ as the reaction coordinate. The unfolded state is approximated as a harmonic well and the native state as an infinitely sharp potential with no structural fluctuations (Figure 3.2A) at n_σ standard deviations from the unfolded minimum. The barrier height can then be directly estimated from the point of intersection of the two potentials. Figure 3.2B plots the folding times versus the n_σ calculated directly using

the experimental thermodynamic parameters (green triangles) and those calculated from empirical size-scaling law assuming ΔH (333 K) = 2.92 kJ mol^{-1} per residue and ΔC_p = 58 J mol^{-1} K^{-1} per residue. There is a good agreement between the two numbers. More importantly, the predicted barriers are small with more than 90 % of the dataset resulting in barriers less than 40 kJ mol^{-1}. Proteins of size less than 55 residues are also predicted to fold over marginal barriers.

In this respect, the plot of n_σ versus protein length is more informative (Figure 3.3). Here, the n_σ are values are plotted for proteins with T_m values near 333 K. The solid line plots the average n_σ at 333 K using the parameters above while the dashed lines is the calculation for the spread of T_m values in the plot (318 – 348 K). It is evident in this figure that $n_\sigma \sim 3$ signals a threshold differentiating proteins that fold over marginal/zero barriers from two-state-like proteins. This is because the proteins that lie below the threshold fold in the microsecond time scale (these two statements will be vindicated in the forthcoming chapters). It is also interesting to note that the designed protein, α_3D, lies well below the expectation compared to its natural counterparts of the same length.

An independent estimate of barrier heights in turn enables the calculation of the pre-exponential to the folding reaction. This necessitates the need to compare barriers calculated at 335 K and the rates at 298 K. A previous work by Akmal and Muñoz that employs a structure-based thermodynamic approach to dissect the kinetic data of 6 two-state-like proteins, predicts marginal temperature dependence

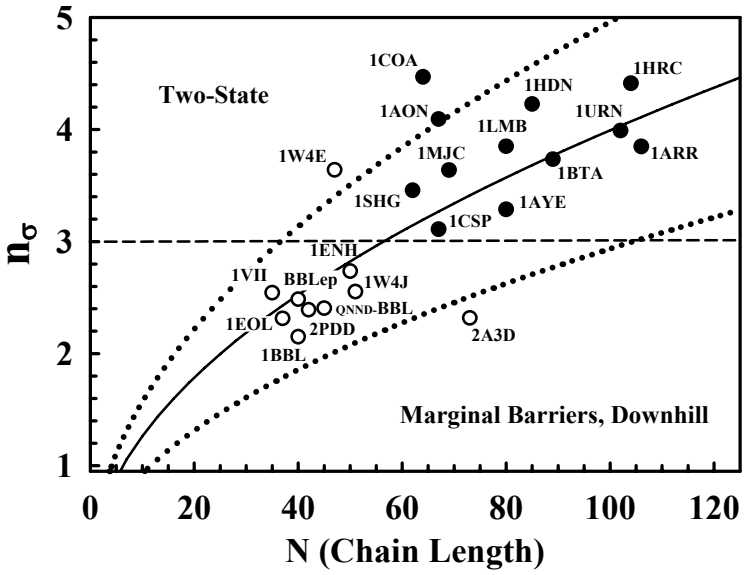


Figure 3.3 n_σ versus the chain length for several proteins. (Open circles) Predicted marginal barrier/downhill proteins. (Filled circles) Predicted two-state-like proteins. The continuous line plots the average n_σ at 333 K while the dotted lines plot the same at 318 and 348 K. See the main text for more details. (Thermodynamic data from published works).

and small folding barriers. The authors conclude that this is an effect of enthalpy-entropy compensation due to a positive ΔC_p associated with unfolding [62]. Moreover, assuming the following relation for temperature dependence of rates

$$\tau_f = \tau_{\min} \exp(\Delta G^\dagger / RT) \tag{3.5}$$

that is equivalent to

$$\Delta G^\dagger = 2.303RT[\log(\tau) - \log(\tau_{\min})] \tag{3.6}$$

it is possible to compare the energy scales, i.e. the average barrier dependence on the folding times should have a slope of 2.303*RT*. The red line in Figure 3.2B plots the dependence with a slope of 0.9 x 2.303RT indicating that such a comparison is indeed valid. This in turn provides an average estimate of $\tau_{\min}$ (i.e. $1/D_{eff}$) to the folding reaction of ~ 1 μs at 335 K consistent with experimental and empirical estimates.

3.6 Conclusions

The analysis of a large dataset of 69 proteins suggests that the logarithm of the folding times scale sublinearly as $N^{1/2}$ with a significant correlation coefficient of 0.74, in agreement with the original predictions of Thirumalai[63]. The definition of n_σ and its fundamental connection to equilibrium fluctuations provides an indirect estimate of the folding barrier heights that range from almost zero barriers to an average maximum of 40 kJ mol^{-1}. The small barriers are consistent with theory. The pre-exponential estimate of 1 μs at 335 K is in accordance with various theoretical and experimental estimates[25]. Moreover, the agreement between these parameters together with the high correlation indicates that the effects of sequence, structure and topology on the folding rates are minimal. It should however be possible to tune the folding mechanism by just redesigning the strength and distribution of contacts within a particular structure. This analysis also emphasizes that downhill folding or marginal barriers should not be uncommon observations as smaller and faster folding proteins are studied. Though encouraging, it is important to note that the pre-exponential of 1 μs is an average value for the entire dataset as proteins are bound to have individual diffusion coefficients. The barrier heights are also rough estimates because of the uncertainty in accurately determining ΔH and ΔC_p. Furthermore, this predictor does not take into account any changes in rate arising out of changes in solvent conditions or mutations. The ideal conditions to calculate the folding rates would be temperature or chemical midpoints as any effects of stability on folding rates are cancelled when comparing different proteins. However, the lack of data at the denaturation midpoint

for fast folding proteins and at the temperature midpoint for the slower counterparts precludes such an analysis.

4. Direct Measurement of Barrier Heights in Protein Folding

4.1 Introduction

Size-scaling arguments presented in the previous chapter suggest that barriers to protein folding are significantly smaller compared to chemical reactions, in accordance with theory and empirical estimates. Given the degree of conformational flexibility observed in proteins the prediction of small folding barriers is not entirely unexpected. However, the surprising issue is the neglect of heat capacity estimates that in fact provide a direct and a precise measure of the equilibrium energy fluctuations (see section 3.4.1) than any other techniques, save MD simulations. This was recognized by Muñoz and Sanchez-Ruiz that led to the development of the Variable Barrier (VB) model[49]. Based on Landau theory of phase transitions it extracts barrier heights and residual fluctuations in the native state ensemble of proteins by analyzing DSC thermograms – the first of its kind in physical biochemistry. It had been earlier employed to distinguish between global downhill and two-state like folding in BBL and thioredoxin, respectively. This chapter employs the VB model to extract barrier heights from the DSC thermograms of previously published proteins. They are then compared with the corresponding rates to estimate the pre-exponential to the folding reaction at 298 K.

Section 4.2 highlights the disadvantages of a chemical two-state analysis of DSC profiles and discusses the extent of its applicability in protein folding. Section 4.3 provides an introduction to the Variable-Barrier Model developed to analyze DSC

thermograms with the ability to extract folding barrier heights. Sections 4.4, 4.5, 4.6, & 4.7 discuss the quantitative calorimetric characterization of a collection of proteins and the corresponding results and implications.

4.2 Chemical Two-State Approximation - Perspectives from Calorimetry

The apparent agreement of the DSC profiles of many proteins to the so-called stringent calorimetric criterion ($\Delta H_{Cal}/\Delta H_{vH} = 1$) has been one of the major selling points for a chemical two-state model. This need not mean that the assumption of a two-state situation is correct. For example, consider the scenario shown in Figure 4.1 (blue circles). The DSC profile is much broader than the one shown in Chapter 2 (Figure 2.1), with no apparent pre-transition baseline. The lowest temperature point of the DSC profile is higher in heat capacity units with respect to the native baseline used to simulate the profile (dark gray line). The fit to a two-state model by fixing $\Delta H_{Cal}/\Delta H_{vH} = 1$ is very good (red curve) with the final thermodynamic parameters being: ΔH_m = 115 kJ mol^{-1} and T_m = 320.6 K. But the inflection point resulting from the chemical baseline is only 311.4 K. More importantly, the baselines cross within the transition region with the slope of the native baseline (green line) absurdly higher than that used to simulate the profile (dark gray line). Can this system be still referred to as two-state just from the calorimetric criterion? The answer is no. In fact, a closer look at the assumptions involved along with a number of works published recently, question not only the validity of using a calorimetric criterion to test the nature of a transition but also the idea of two-stateness. They are as listed below.

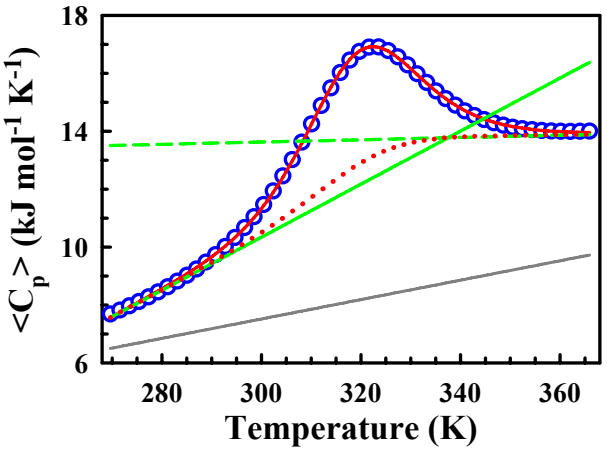


Figure 4.1 Simulated heat capacity profile of a 50-residue protein (blue circles) employing the native baseline shown as a dark gray line. The fit to a two-state model constraining $\Delta H_{Cal}/\Delta H_{vH} = 1$ is shown together with the folded (continuous green line) and unfolded baselines (dashed green line) and the chemical baseline (dotted red curve).

a) The calorimetry profile of even the most two-state-like systems show a finite width in the transition region. In other words, the transition from completely folded to a completely unfolded protein occurs over a finite range of temperatures, and not at a single temperature. This is in contrast to solid-liquid or liquid-gas phase transitions (water to water vapor, for example) that occur at a single temperature resulting in a discontinuity (infinitely sharp) in their heat capacity profiles. In fact, the term first order phase transition or two-state is more appropriate for such systems. Therefore, protein folding can be at best *approximated* (when applicable) as pseudo-two-state reactions or pseudo-first-order phase transitions. This is evident in the two-state fit to a DSC profile where the temperature dependence of enthalpy and entropy (i.e. the change in heat capacity, ΔC_p) is ignored in the transition region *just so that the entire enthalpy change can be assumed to occur at a single temperature* for comparison with ΔH_{Cal}. This is the case in spite of the fact that ΔC_p can be directly estimated from baselines used in the fit! This observation suggests that width of the transition monitored by DSC is critical in approximating the unfolding as an all-or-none process. If the width is sufficiently narrow the typical two-state approximation of

assuming a constant enthalpy and entropy within the transition region probably holds true (for example, Figure 2.1). But what determines the width and how to quantify it? Moreover, the broadness of the transition can be easily trimmed by assuming unphysical baselines; this brings up the next question.

b) What is the meaning of heat capacity baselines? The folded and unfolded baselines used in two-state calorimetry fits signify the temperature dependent changes in the enthalpy fluctuations of the corresponding states. They are supposed to have non-structural origin with contributions from vibrational modes and due to hydration of side chain groups (see (d) as well). The two-state fitting procedure however can choose baselines that trim the wings of the heat capacity curve thereby eliminating data incompatible with a two-state model. In fact, the DSC thermogram shown in Figure 4.1 is for a hypothetical 50-residue protein with zero barrier height at the midpoint, i.e. globally downhill. It can be perfectly fit to a two-state model largely because of the steep native baseline that attributes larger enthalpy fluctuations to the 'native' state than what is actually seen (dark gray line). Moreover, Karplus and co-workers have also shown that a DSC thermogram simulated by assuming a 3-state system can be fit by a 2-state model by choosing appropriate baselines[98]. Therefore, extreme care should be exercised in assessing the nature of the transition just based on the $\Delta H_{Cal}/\Delta H_{vH}$ ratio alone.

As of now, there are no first principle calculations predicting the magnitude of the heat capacity of folded proteins, leave alone the temperature dependence. This is a complex problem as all the possible non-structural relaxations like bond vibrations, bending, stretching etc. including the structure of the solvent have to be taken into

account in modeling. One reasonable way around this limitation and to assess the quality of baselines from fits is to compare them to empirical standards. Freire and co-workers have shown that the heat capacity of the native state and its dependence with temperature strongly correlates with the size of the protein[99] that can be represented as:

$$C_p^f = [1.323 + 6.7 \times 10^{-3}(T - 273.15)]M_r \text{ J K}^{-1} \text{ mol}^{-1} \quad (4.1)$$

where T is the temperature in K and M_r is the molecular weight of the protein in g mol^{-1}. This has been shown to be a reasonable approximation of the native baseline for DSC experiments reporting absolute heat capacities. The fitted native baselines should therefore be compared to the Freire's baseline, at least at the level of the temperature slope obtained.

c) The above discussion questioning the general approximation of protein folding as a chemical two-state reaction and the importance of native baselines assumes even more significance with the recent characterization of a number of fast-folding proteins. From size-scaling arguments (Chapter 3) it is clear that most of the fast folding proteins tend to fold over small/negligible barriers and hence the traditional chemical two-state model would be of little physical significance. Such fast-folding proteins are also expected to have broad unfolding transitions. Therefore the native baseline cannot be extrapolated from low-temperature points as has been done for a number of slow folding proteins, necessitating the need for an alternate model.

d) Recent theoretical analyses with elementary statistical mechanical[47] and polymer-chain models[100] have shown that single-peaked DSC thermograms are also observed

in continuous unfolding transitions (*i.e.* downhill). Therefore, the observation of a single-peaked thermogram also does not guarantee a two-state system.

e) The inability to precisely estimate the equilibrium fluctuation contribution to the observed positive ΔC_p in proteins has resulted in the popularity of hydrophobic solvation (sections 2.2.1 and 3.4.1) models. It is clear that this assumption makes any protein folding problem intrinsically two-state-like. This is because it attributes the positive ΔC_p change entirely to the population or de-population of two states: folded state (with buried apolar groups) and the unfolded state (with exposed apolar groups), whose properties are defined by the baselines. It therefore neglects any contribution that could arise from the difference in conformational flexibility of these states or even more importantly, *the possible ensemble of structures that could exist at any given temperature*.

4.3 Variable Barrier Model

The issues outlined in the previous section highlight one of the major drawbacks in the field of protein folding - the absence of an ensemble based approach to characterize folding reactions to distinguish between the various folding scenarios. This was recognized by Muñoz and Sanchez-Ruiz that led to the development of the Variable Barrier Model to analyze DSC profiles. This model is the first of its kind in physical biochemistry that enables the extraction of folding barrier heights from equilibrium measurements. It is essentially based on the fundamental relation between the heat capacity profile of a protein and its partition function (Q):

$$Q = \int \rho(H) \exp\left(-\frac{H}{RT}\right) dH \tag{4.2}$$

where $\rho(H)$ is the density of enthalpy microstates along a suitable enthalpy scale H, with the crucial assumption being the enthalpy scale and the enthalpy of the microstates are fixed and independent of temperature. This specific representation of the partition function enables the characterization of a system based on continuous distribution of conformational microstates. Therefore changes in the density of the conformational microstates accounts for the changes in entropy as a function of temperature while heat capacity defines the temperature dependence of average enthalpy. This is in contrast to a two-state approximation where the temperature dependence of enthalpy and entropy is attributed to a difference in heat capacity arising out of non-conformational effects (solvation), thus ignoring the possible distribution of microstates. The probability of finding the protein in a microstate of enthalpy H, $p(H|T)$, in the current representation is simply

$$p(H \mid T) = \frac{\rho(H) \exp\left(-\frac{H}{RT}\right)}{Q} \tag{4.3}$$

The probability can also be expressed in reference to some characteristic temperature T_0

$$p(H \mid T) = C \cdot p(H \mid T_0) \cdot \exp(-\lambda H) \tag{4.4}$$

where

$$\lambda = \frac{1}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)$$

and C is normalization constant. Enthalpy moments can be calculated as:

$$\left\langle H^n \right\rangle = \int H^n p(H \mid T) dH$$

where n=1,2…, and the excess heat capacity (C_p^{ex}) referenced to the native state, as

$$C_p^{ex} = \frac{d\left\langle \Delta H \right\rangle}{dT} = \frac{\left\langle \Delta H^2 \right\rangle - \left\langle \Delta H \right\rangle^2}{RT^2}$$

with

$$\left\langle \Delta H \right\rangle = \left\langle H \right\rangle - H_F$$

As can be seen from above equations, the probability density can be directly extracted from the DSC profile by performing an inverse Laplace transform of the partition function. But such a transformation has a non-unique solution with several different assumptions of $\rho(H)$ giving identical results[101,102]. This severe limitation was cleverly overcome by Muñoz and Sanchez-Ruiz by assuming that the probability density at T_0 can be represented as

$$p(H \mid T_0) = C' \exp\left(-\frac{G_0(H)}{RT_0} \right)$$

where C' is a normalization constant and $G_0(H)$ is the shape of the free energy functional that defines the probability density at T_0. The free energy functional was expressed as a 4th other polynomial, similar to the Landau theory of phase transitions

$$G_0(H) = -2\beta \left(\frac{H}{\alpha} \right)^2 + |\beta| \left(\frac{H}{\alpha} \right)^4 \tag{4.5}$$

where β and α have a physical meaning as shown below. Setting $dG_0(H)/dH = 0$ and evaluating d^2G_0/dH^2 for $\beta > 0$ leads to two minima at $H = \pm\alpha$ and a maximum at $H =$

0. This corresponds to a two-state scenario with β representing the barrier height separating the folded (+ α) and unfolded (- α) macrostates. $\beta < 0$ results in a single macrostate thus mimicking a downhill folding situation. To account for the fact that folded states have smaller enthalpy fluctuations than unfolded state, a parameter α_N was introduced for $H < 0$ and α_P for $H > 0$. For convenience in fitting, these are represented as

$$\alpha_N + \alpha_P = \sum \alpha$$

$$\alpha_N = \sum \alpha \cdot f / 2 \quad \text{and} \quad \alpha_P = \sum \alpha \cdot (2 - f) / 2 \tag{4.6}$$

where $0 < f < 1$. Assuming a two-state scenario, $f = 1$ corresponds to a situation where the probability density has equal widths for the folded and unfolded states, while $f < 1$ results in an asymmetric distribution of the probabilities with folded shape being more sharper. Analyzing a DSC profile with the Variable Barrier Model therefore requires 4 parameters: β (barrier height), T_0 (characteristic temperature), $\Sigma\alpha$ (enthalpy at T_0) and f (asymmetricity factor). Apart from these, a reliable estimate of the native baseline is required as the model reproduces only the excess heat capacity. In other words, all the non-structural contributions to the heat capacity in the folded state have to eliminated, which is made possible by subtracting the native baseline from the measured heat capacity curve.

To summarize, the variable-barrier model analysis of a DSC profile gives an estimate of the barrier height close to T_0 with 2 less parameters than a typical two-state fit. The fit is highly constrained as it does not employ free floating baselines thus minimizing data trimming. Moreover, as the barrier height is itself a parameter it

serves as a more stringent test (compared to $\Delta H_{Cal}/\Delta H_{vH}$) to characterize the statistical nature of the transition.

4.4 Sensitivity of the Model

The Variable Barrier Model has been successful in differentiating the barrier-less and barrier-limited unfolding observed in BBL and thioredoxin, respectively[49]. This raises an important question - are the barrier heights extracted by this method absolute? If so, then it is possible to independently estimate the activation free energy to folding from which the elusive dynamic term in the rate expression (D_{eff}) can be extracted. But, to apply this model to estimate absolute barrier heights, the intrinsic limitations of the model and its sensitivity to the range of barrier heights have to be evaluated. This is because of the implicit approximation that the free energy surface of natural proteins is smooth and that it can be represented as a Landau polynomial.

A simple procedure to ascertain the sensitivity range is to simulate DSC profiles from free energy surfaces of known barrier heights and then characterize them by the variable barrier model. The comparison between the two barrier estimates then provides a direct tool to test the model. This methodology can also be used to investigate the effect of native baseline approximation (*i.e.* changes in slope and intercept) in influencing the final barrier estimates. The one dimensional free energy surfaces were calculated using the DM model (discussed in detail in Chapter 5). The plot of theoretical barrier height from the DM model's known free energy surface versus the estimated barrier height from the Landau model (β) is shown in Figure 4.2. It shows that the variable barrier model is able to accurately predict barrier

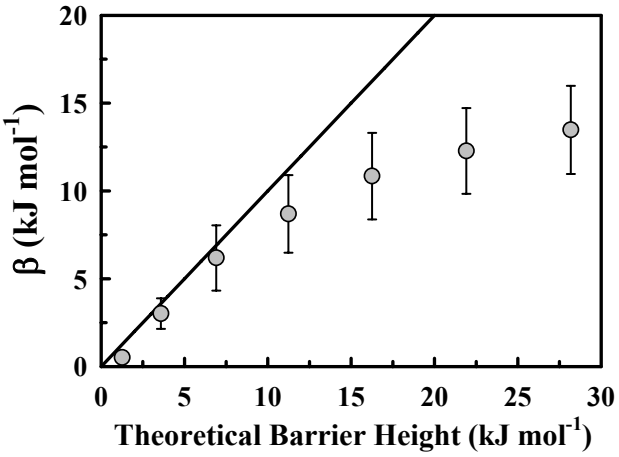


Figure 4.2 Comparison between the theoretical barrier heights directly calculated from the DM model and the barrier heights extracted from the variable-barrier analysis of the simulated DSC thermograms (β). The continuous line plots the expected 1:1 correlation.

heights when they are small but progressively under-estimates at higher barrier heights. The reason for this observation is simple. The method is highly sensitive only when the population is maximal at H = 0 kJ mol^{-1} or at the top of the barrier when there is one. This can be seen from the very negligible errors at low barrier heights that get continuously larger at higher barriers. The estimated barrier height data tend asymptotically to ~15 kJ mol^{-1} (~6 *RT*), indicating the sensitivity limit of the technique. This corresponds to ~0.25% population at the top of the barrier. In fact, the barrier heights agree surprisingly well until 10 kJ mol^{-1} (~4 RT or 1.8%). It is still possible to estimate relative differences in barrier height between 10 and 15 kJ mol^{-1}. In other words, the variable barrier model can be used to differentiate barrier heights in the range of 0 – 15 kJ mol^{-1} within the smooth Landau approximation of protein folding free energy surfaces. The magnitude of the barrier heights was also found to be slightly sensitive to the initial native baseline approximation; but the resulting errors are within those shown in Figure 4.2.

4.5 Proteins studied

Is this sensitivity range useful to characterize natural proteins? There is no direct answer to this question because there are no alternate methods to estimate absolute barrier heights. This is because the barriers are typically estimated by assuming a pre-exponential in the rate expression from kinetic studies on fast folding proteins or from elementary reconfiguration steps like loop formation or unfolded state dynamics (see Chapter 1). But the results of size-scaling analysis together with the estimates of n_σ indicate that there are many proteins within the sensitivity range of this model. As a control, the extracted barrier heights are also compared to the corresponding folding rates in water at 298 K for a set of proteins for which both the DSC and kinetic data are available. Such a comparison (similar to the n_σ versus τ_f detailed in the previous chapter) provides a ruler to gauge the accuracy of the results. Table I shows the database of 15 proteins used in this analysis. They are quite uniform in size (64 ± 15 residues), but include representatives from the three main structural classes: α, β, and $\alpha + \beta$. The DSC profiles of these proteins showed no signs of irreversibility. Moreover, the folding rates at 298 K span almost 4 orders of magnitude from 10 (α-spectrin SH3) to 10^5 s^{-1} (BBL).

4.5.1 Native Baseline Determination

To predict accurate barrier heights, a reliable estimation of native baseline is essential. If the heat capacity units are absolute, the ideal candidate for native baseline

is the Freire's relation (equation 4.1). However, the available DSC data is heterogeneous requiring different native baseline estimation schemes to be devised.

Method I Accurate absolute heat capacity values are available for 1BBL and 2PDE. Therefore, Freire's relation was directly used to calculate the native baseline.

Table 4.1 Proteins Studied

Index	Protein Name	PDB ID	Length (N)	Struc. Class	log (k) (298 K)	Baseline Estimation Method
A	BBL	2CYU	40	α	4.8	I
B	PDD	2PDD	42	α	4.3	I
C	Engrailed	1ENH	52	α	4.6	II
D	PDD (F166W)	1W4E	42	α	4.1	II
E	Horse Cytochrome C	1HRC	104	α	3.4	II
F	Sso7d	1SSO	64	β	3.0	III
G	CspB (*B. subtilis*)	1CSP	67	β	2.9	III
H	CspB (*T. maritima*)	1G6P	66	β	2.7	III
I	CspA	1MJC	69	β	2.3	III
J	Fyn-SH3	1SHF	67	β	2	III
K	α-spectrin-SH3	1SHG	62	β	0.9	III
L	α –spectrin-SH3 (D48G)	1SHG	62	β	1.6	III
M	Tendamistat	2AIT	74	β	1.8	IV
N	Hpr	1HDN	85	$\alpha + \beta$	1.2	IV
O	Protein G	1PGB	57	$\alpha + \beta$	2.6	V

Method II Calorimetric data for proteins 1ENH, 1W4E and 1HRC are not in absolute heat capacity units. To determine the native baseline, a baseline was initially calculated from the low temperature point of the respective calorimetry profiles using Freire's empirical temperature dependence. The baseline was allowed to up- or downshift in the fitting procedure keeping the temperature dependence constant. Such fitted baseline was used in the grid-analysis (see below). It is of interest to note that

two-state fits to these 5 proteins resulted in significant baseline crossing within the transition region.

Method III For proteins F-L the same procedure outlined in Method II was employed, but by fixing the baseline throughout the fitting procedure. This makes the DSC profile as two-state-like as possible while reducing the errors in concentration determination. Two-state fits for all these proteins provided good fits without baseline crossing.

Method IV The heat capacity of the native state of 2AIT has lower temperature dependence than that predicted by the Freire's relation; probably due to the presence of two disulfide bridges in the protein. In this case the native baseline was directly extrapolated from the low-temperature points of the calorimetry profile. The same procedure was followed for 1HDN as it had an unusually large difference in heat capacity between low and high temperatures. Aggregation problems at higher temperatures were reported for this protein.

Method V No native baseline estimation was required for Protein G (index O) as the excess heat capacity data is directly reported in the literature.

4.5.2 Fitting Procedure and Error estimation

The parameters β and f are intrinsically correlated. Therefore, to avoid any erroneous results a grid-analysis was carried out on β with a spacing of 1 kJ mol^{-1} and ranging from -15.5 to 65.5 kJ mol^{-1}. The fit then requires just 3 parameters f, $\Sigma\alpha$ and T_0, as the native baselines are fixed (except for the group II). All fits to the model were performed in Matlab 6.5 using inbuilt non-linear least square fitting routines.

The results from a grid-analysis also enable estimation of errors in the measured barrier height. The χ^2 value, defined as

$$\chi^2(\beta) = \sum_{i=1}^{n_d} \left(f_i - d_i \right)^2$$

where *f*, *d* and n_d correspond to the fit, data and number of data points , respectively, was normalized to the best fit. The resulting plot of χ^2_{red} versus β (χ^2_{red}/β plot) was then interpolated to obtain χ^2_{red} values every 0.01 kJ mol^{-1} β spacing. Approximating this high density χ^2_{red}/β plot to a Gaussian curve, the 68 % confidence interval (one standard deviation) can be simply obtained from the intersect of the residual plot to the value corresponding to

$$\chi^2_{red}(\beta_{\min} \pm \sigma) = \frac{n_d - n_p + 1}{n_d - n_p}$$

where

$$\chi^2(\beta_{\min}) = n_d - n_p$$

$\beta_{\min}$ is the barrier height corresponding to the least χ^2 and n_p is the number of parameters (n_p = 3 for this calculation). Though the χ^2_{red}/β plots were not truly Gaussian, the approximation works well close to $\beta_{\min}$. The final $\beta(T_0)$ values are calculated as weighted sum of the reduced χ^2 within the confidence interval:

$$\beta\left(T_0\right) = \frac{\sum_i \frac{1}{\chi^2_{red,i}} \beta_i}{\sum_i \frac{1}{\chi^2_{red,i}}}$$

4.6 Results

The model provided fits of comparable quality to two-state fits. The obtained characteristic temperatures (T_0) span a range of ~60 K. For a two-state system, T_0 is the temperature at which the folded and unfolded states have the same free energy. However, in a two-state analysis $T_m = T_0$ as the difference in widths of the wells due to the difference in the conformational fluctuations are ignored. For proteins with narrow folded wells ($f < 0.5$; protein indices I-O and G), T_0 is therefore higher than the T_m (data not shown) or the maximum of the DSC thermogram, and $\Sigma\alpha$ approximates ΔH_m. For proteins with large asymmetricity values ($f > 0.5$; protein indices A, B, D-F and H) T_0 approximates the maximum of the thermogram. It corresponds to the temperature at which the conformational fluctuations are maximal thus resulting in a peak in the DSC profile. The parameter T_m is not applicable for this subset as baselines cross in a two-state fit. The thermogram of engrailed homeodomain (protein index C) has a low temperature slope higher than that predicted by the Freire's relation thereby producing an unusually small *f*-value. The high slope is probably a result of pre-equilibration artifacts or due to the presence of a long unstructured tail in the protein.

Figure 4.3 shows two examples of the fits – that of a downhill folding protein BBL and the two-state-like protein CspB (*Bacillus subtilis*). Since absolute heat capacity values are available for BBL, it provides a unique opportunity to compare the degree of fluctuation already present in the folded state. No pre-transition baselines are evident in the thermogram thereby producing crossing baselines in a two-state fit. Also, the native baseline for BBL is downshifted by almost 2 kJ mol^{-1}

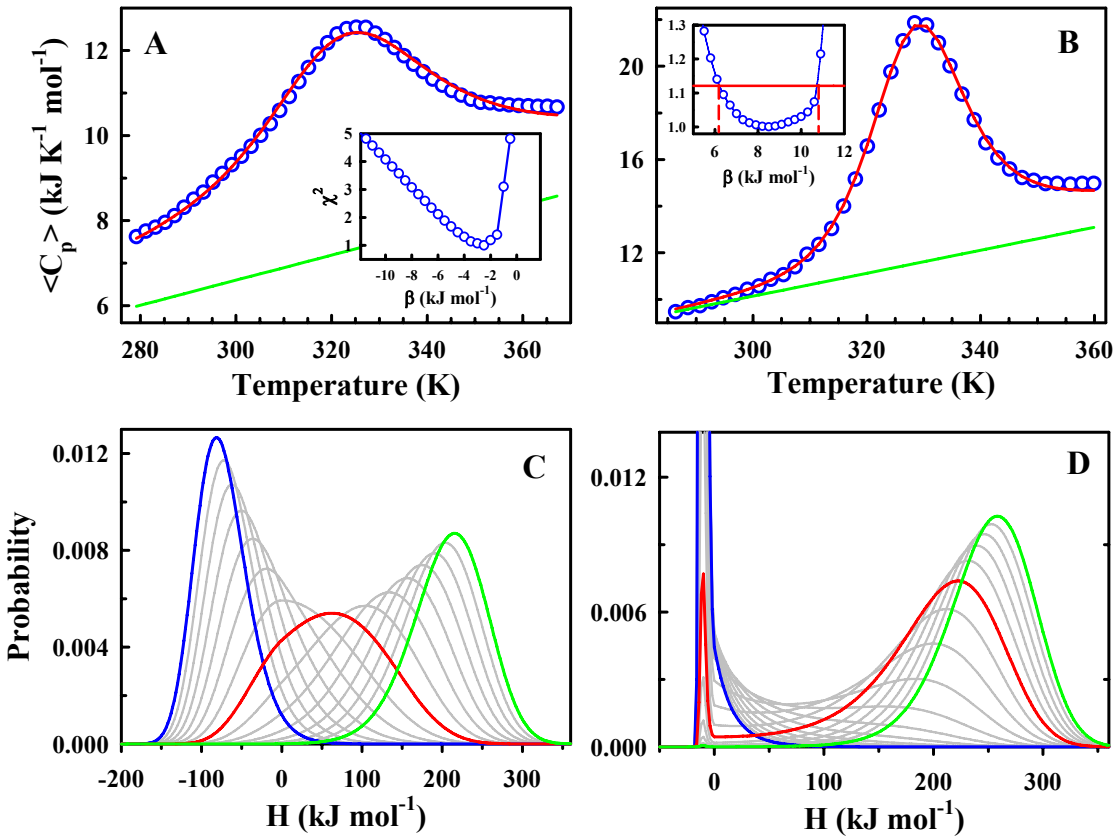


Figure 4.3 A) DSC thermogram of BBL (2CYU)(blue circles) and fit to the VB model (red line) assuming the baseline shown in green. The inset corresponds to the χ^2_{red}/β plot. The steep plot results in negligible errors in β. B) Same as panel A but for CspB (1CSP) with red lines in the inset signaling 95% confidence intervals. C & D) The extracted probability density as a function of the one-dimensional reaction co-ordinate of enthalpy for the profiles in panels A and B, respectively. The probability distribution at the lowest temperature (blue), T_0 (red) and the highest temperature (green) are highlighted. (DSC data from published works).

compared to the lowest temperature point thus indicating significant fluctuations even in the folded state (Figure 4.3A). The asymmetry factor (f) of 0.69 is an evidence for this observation. The fit resulted in a β of -2.7 ± 0.5 kJ mol^{-1} at a T_0 of ~ 317 K. The χ^2_{red}/β plot has a sharp minimum (Figure 4.3A inset) resulting in a negligible error in the estimation of β. The probability density plotted as a function of enthalpy (the reaction co-ordinate) shows the hallmark behavior of global downhill folding proteins

(Figure 4.3C). They are unimodal at all temperatures with the peak position shifting progressively towards the unfolded ensemble at increasing denaturational stress.

Table 4.2 Parameters from the Variable Barrier Model Analysis

Index	Protein Name	β (kJ mol^{-1})	$\Sigma\alpha$ (kJ mol^{-1})	T_0 (K)	f	Baseline Shift (kJ mol^{-1})
A	BBL	-2.7	294.4	317.1	0.69	-
B	PDD	0.5	133.1	322.8	1.00	-
C	Engrailed	1.1	163.5	324.2	0.10	0.12
D	PDD (F166W)	4.0	216.2	330.9	0.94	0.57
E	Horse Cytochrome C	3.8	364.7	335.8	0.94	0.83
F	Sso7d	11.9	292.2	373.4	0.58	-
G	CspB (*B. subtilis*)	8.5	235.3	339.8	0.08	-
H	CspB (*T. maritima*)	11.4	315.1	363.4	0.65	-
I	CspA	9.3	228.9	343.5	0.08	-
J	Fyn-SH3	13.2	275.5	356.7	0.06	-
K	α-spectrin-SH3	16.3	239.6	355.3	0.06	-
L	α-spectrin-SH3 (D48G)	15.7	256.5	359.5	0.05	-
M	Tendamistat	18.1	350.1	377.4	0.03	-
N	Hpr	14.5	372.5	341.4	0.10	-
O	Protein G	14.1	332.2	367.4	0.04	-

Unfortunately, the thermogram of CspB is not in absolute heat capacity units requiring the use of the first temperature point and the temperature slope from Freire's relation (Method III) to determine the native baseline. This precludes the estimation of any residual structural fluctuation in the folded state. Therefore, this procedure makes the thermogram look more two-state like and the extracted barrier heights are upper estimates. The resulting probability density from the fit (Figure 4.3B) is two-state-like with a sharp peak (and hence an f of 0.08) at the lowest temperature and bimodal distribution at ~340 K (T_0) resulting in a β of 8.5 ± 2.1 kJ

mol^{-1} (Figure 4.3D). The error estimates are higher than that for the downhill folding protein. This is mainly because of the fact that the χ^2_{red}/β curve broadens (compare the insets of Figures 4.3A and 4.3B) at higher values of barrier height signaling the approach to the sensitivity limit of this method.

Table 4.2 also lists the extracted barrier heights at T_0. They range from globally downhill for BBL ($\leq$ 0 kJ mol^{-1}) to two-state-like for Tendamistat (~18 kJ mol^{-1}). Direct comparison of the barrier heights and rates however presents a problem as they correspond to different temperatures. But the presence of enthalpy-entropy compensation in protein folding as discussed in Chapters 2 and 3 (and reference 17) suggests that the rate and barrier height can be compared directly by just correcting for the temperature effects of stability between different proteins, i.e. log(k) versus β/T_0. This approximation is necessary as the rate at T_0 is not available for most of the slow folding proteins. Assuming that under folding conditions (i.e. low temperature) the kinetics is entirely dominated by the magnitude of the folding barrier and using a transition-state like expression, the rate at 298 K (k_{298}) can be written as,

$$k_{298} = D_{eff} \exp(-\beta / RT_0)$$

where D_{eff} is the effective diffusion coefficient. Rearranging,

$$\beta / T_0 = a(\log D_{eff} - \log k_{298})$$

where a = 2.303R. In other words, a plot of log(k_{298}) versus β/T_0 should have a slope of a and an intercept that corresponds to a.log(D_{eff}), i.e. the pre-exponential to the folding reaction, if there is exact agreement. Figure 4.4 shows the correlation between β/T_0 and the logarithm of the folding rates at 298 K. The obtained correlation

coefficient (r) is 0.95 with a slope of 0.8a. A slope below 2.303R and the under-estimation at higher barrier heights (apparent from the non-linearity of curve), are in agreement with the theoretical calculations (Figure 4.1). The correspondence is striking as there is no significant correlation between folding rates and protein size (r^2 < 0.2) or unfolding enthalpy (r^2 = 0.25) for this dataset. Moreover, the correlation with β values directly is of similar quality (r^2 = 0.86, slope ~0.9a), in accordance with results of Akmal and Muñoz[62].

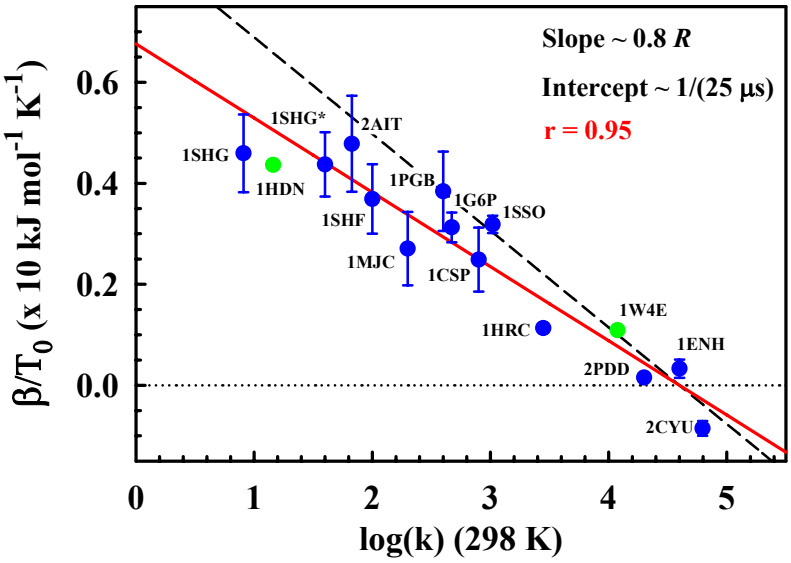


Figure 4.4 Correlation between folding rates at 298 K and the ratio between barrier height (β) and characteristic temperature (T_0). The dashed line shows the expectation for a slope of R while the correlation line is shown in red. For 1W4E, the folding rate at 298 K was obtained by scaling the available rate at 325 K for the changes in water viscosity (a factor of ~2 decrease). For 1HDN though aggregation was reported the model produced a reasonable fit. Neither of these two proteins was included in the correlation. (Rate data from published works).

4.7 Implications

Proteins with smaller barriers have large asymmetry values highlighting the crucial link between equilibrium fluctuations and barrier height (Table 4.2). Baseline crossing in two-state fits is evident for proteins whose relaxation rate is > 1000 s^{-1} (τ

<1 ms) (protein indices A-E). The predicted barrier heights are also negligible/marginal for these proteins. In general, this observation strongly suggests the breakdown of a two-state approximation when k > 1000 s^{-1} at 298 K. Interestingly, this specific subset of proteins belong to the α-helical structural class. The structure of small single domain alpha helical proteins is almost entirely dominated by local i→i+4 H-bonding and i→i+3 and i→i+4 electrostatic and hydrophobic interactions. It is also known that isolated alanine-based alpha helices are non-two-state like with a distribution of helical lengths populating any given temperature[36]. Taken together, these observations indicate that there is lesser long-range influence in small α-helical proteins compared to those dominated by β or $\alpha + \beta$ structures, thus increasing the possibility of decoupled unfolding and hence marginal barriers. In fact, Zuo et al. have identified a structural descriptor they define as N_N - the average number of non-local interactions per residue - that is apparently able to distinguish between two-state-like and marginal barrier/downhill scenarios based on PDB structures alone[50].

The predicted barrier heights from DSC experiments agree well with results from size-scaling arguments (see previous chapter) or empirical folding speed-limits. The barrier height of 11.4 ± 1.1 kJ mol^{-1}(~4.6 *RT*) for the cold-shock protein *T. maritima* (1G6P) is within the estimates from single molecule spectroscopy (4 $RT < \beta < 11$ *RT)* [27]. The model is also able to discern differences in folding barrier heights between homologous cold-shock proteins from *B. subtilis* and *T. maritima* (1CSP and 1G6P, respectively). The procedure is even able to detect changes induced by point mutations. For example, PDD wild-type a structural and functional homolog of the

downhill folding BBL, has a marginal barrier of 0.5 kJ mol^{-1} (~0.2 *RT*). A single point mutation of F→W (non-conservative) on this domain[103] produces a small barrier of ~4 kJ mol^{-1}. From these results, it appears that single domain proteins can be classified in three distinct groups: with marginal or no barriers (<2 *RT*, or 0.017 in Figure 4.4), two-state-like (>4 *RT*, or 0.033), and twilight zone proteins (<4 *RT* and >2 *RT)*. This classification provides the much needed quantitative tool to compare the widths of DSC transitions as discussed in section 4.2. The nature of the folding ensemble for the first two groups is evident from the names. The third group – twilight zone proteins – corresponds to those proteins whose widths are intermediate between downhill and two-state. They have significant barriers but not high enough to be labeled two-state, suggesting that these proteins are bound to be highly sensitive to mutational effects and other perturbations. A member of this group, the cold-shock protein from *B. subtilis* indeed shows remarkable sensitivity to even simple deletion mutations evidenced by large changes in *m*-values[104].

The most interesting result however is the intercept for zero barrier in Figure 4.4. As discussed before, this corresponds to an average pre-exponential for this dataset. This analysis yields a value of 40,000 s^{-1} at 298 K. Surprisingly, this value is ~10 times slower than other empirical and experimental estimates (see Introduction and Chapter 3). The reason for this observation and its implications will be discussed in greater detail in Chapter 5. This value as a pre-exponential is applicable only to proteins with significant barriers. In the absence of barriers, a higher free energy gradient will speed up folding or it might even slow it down due to residual roughness. Correcting for the temperature dependence of viscosity the D_{eff} scales to

~100,000 s^{-1} at the average temperature of T-jump experiments (~330–340 K). The results therefore suggest that it is possible to estimate the individual diffusion coefficients to folding by combining the rate data and the VB model analysis of DSC profiles.

4.8 Conclusions

The results presented above are consistent with previous analyses that downhill folding proteins can result in single-peaked thermograms, though much broader than their two-state counterparts. They also highlight the importance of reporting data in absolute units that enables the extraction of the degree of residual thermodynamic fluctuation in the native state. The estimated barrier heights are more accurate than those from size-scaling contributions as the latter employs average thermodynamic parameters. The barrier heights are predicted to be small and in the range of 0 - 18 kJ mol^{-1} for this dataset. The remarkable agreement between measured barrier heights and kinetic relaxation rates further suggests that the one-dimensional reaction co-ordinate of enthalpy can be a suitable scale for characterizing protein folding reactions. This observation is in accordance with other works where one-dimensional reaction co-ordinates have been highly successful in reproducing complex helix-coil kinetics and behaviors of lattice polymers. The results convincingly suggest that there are significant thermodynamic fluctuations in proteins molecules even under native conditions (low temperature and absence of denaturants), signaling the need for a fundamental shift in the approach in characterizing protein folding reactions.

5. Protein Folding Kinetics: Barrier Effects in Chemical and Thermal Denaturation Experiments

5.1 Introduction

The variable-barrier model analysis of DSC thermograms reveals that the barriers of slow folding proteins - folding time of the order of a millisecond or higher at the T_m - are small (~10 – 20 kJ mol^{-1}). The proteins that fold in the microsecond range should then have even smaller barriers and fold in the global downhill to marginal barrier range. This situation is strongly supported by the corresponding observations in BBL and mutants of λ-repressor, respectively, both of which fold in the microsecond time range[55,62]. However, most of the experimental data on fast folding in the current literature is analyzed using a chemical two-state model. It is in fact able to explain the data reasonably well to a first approximation. The justification for employing a two-state model for these proteins stems from the observation of single-exponential kinetics. But it is important to note that even folding over marginal barriers or global downhill folding produces single-exponential decays as evidenced in a number of experiments and simulations[16,23,33,54]. Therefore, the concept of 'fast-folding' and the use of a two-state model to explain the data are at odds with one another. This fast folding paradox then raises an interesting question: are there any distinct observations or results of two-state treatments of these proteins that signal the presence of marginal barriers? This question is addressed here using a simple variant of the Zwanzig's one-dimensional statistical mechanical model of protein folding[15] – the Doshi-Muñoz (DM) model. The results from a quantitative analysis of chemical

and thermal denaturation experiments on previously published proteins indicate that they do indeed fold over marginal barriers at the C_m (chemical midpoint) or T_m while folding downhill under native conditions.

Section 5.2 outlines the various experimental observations that suggest a clear deviation from two-state behavior for the fast folding proteins. Section 5.3 introduces the statistical mechanical model followed by the treatment of the chemical and thermal denaturation data in sections 5.4 and 5.5, respectively.

5.2 Experimental Observations - Deviations from *bona fide* Two-State Behavior

a) *Broad Equilibrium Chemical Denaturation Curves* The chemical denaturation curves of a number of microsecond folding proteins are broad without any apparent pre- and post-transition slopes. For example, the width of the transition spans ~ 4 M GuHCl for FBP28 WW domain[105] and mutants of the BBL family [68]compared to the typical width of 1-2 M in millisecond folding proteins muscle AcP[106] and yeast ACBP[107]. The experimental temperature is usually 298 K for such experiments; so the broadness is not the result of temperature effects. This observation is similar to the steep pre-transition slopes observed in DSC experiments. The width is traditionally interpreted as arising out of the small size of the protein that would result in a small change in ΔASA. However, size-scaling arguments suggest that small size is also correlated to a smaller barrier height. In other words, the phenomena are interlinked and the physical reason behind the origin of broadness is unexplored.

b) *Different Sensitivities to Chemical Denaturation from Equilibrium and Kinetics* The m_{kin} value for these microsecond folding proteins also has been observed to be

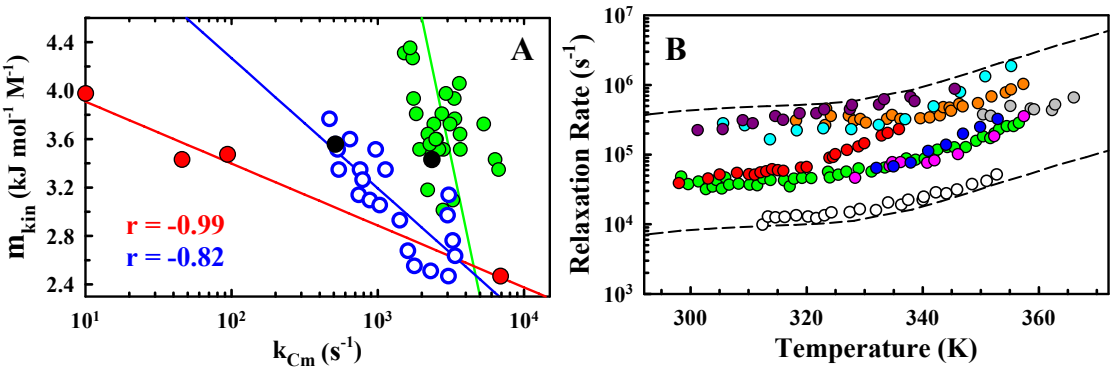


Figure 5.1 Fast-folding experimental data. A) Plot of m_{kin} versus k_{Cm} for the engrailed family [108](red), mutants of PDD F166W[103] (blue), and mutants of FBP28 WW domain[105]. Wild-type proteins are shown as filled black circles. Red and blue lines represent linear regression fits while the green line is shown to guide the eye. B) Relaxation rate versus temperature for microsecond folding proteins. FBP WW domain* (DNDC Y11R-W30F FBP WW[109]; light green), Pin WW domain[110] (white), Villin N27H[111] (cyan), Villin HP36[112] (purple), albumin binding domain[113] (1prb$_{7\text{-}53}$ K5I; gray), engrailed homeodomain[114] (red), B-domain of staphylococcal protein A[115] (BdpA; pink), α_3D[116] (orange), and $\lambda_{6\text{-}85}$ D14A[117] (dark blue). (m-values and rate data from published works).

significantly lower compared to the m_{eq}. The ratio of m_{kin}/m_{eq} is as low as 0.65 in the case of some of the fastest folding proteins[68,105,108], suggesting that these proteins are non-two state (see Chapter 2). In literature, this effect has been qualitatively attributed to the presence of an on-pathway intermediate that has a different rate dependence on the denaturant concentration[108]. In spite of this speculation only two-state models have been used to characterize such chevrons thus yielding no information on the nature of the intermediate. This, therefore, still remains an open issue.

c) *Correlation Between m_{kin} and k_{Cm}* A more striking phenomenon is the observation of deceasing m_{kin} values as the k_{Cm} increases for a given set of homologous proteins and/or mutant series (Figure 5.1A). m-values are known to depend on the size, structure and sequence of a protein. Given this fact, the decrease in m_{kin} is even more apparent as there is expected to be no great differences in m_{kin} between mutants or

homologous proteins. Not only does the m_{kin} decrease but also shows a very high correlation: -0.99 for the engrailed family and -0.82 for the E3BD F166W pseudo-wildtype. In other words, this suggests a chevron that gets flatter (smaller m_{kin}) upon an increase in k_{Cm} by mutation. This is a clear example of a deviation from two-state behavior as the slope of the chevrons are supposed to remain invariant in a two-state system upon mutation and further should show no correlation with the corresponding k_{Cm}. Interestingly, the slope of the plot of m_{kin} versus k_{Cm} increases with the average magnitude of k_{Cm}. The slope becomes so pronounced that a correlation analysis becomes inappropriate for the mutant series of FBP28 WW domain. This suggests that the faster the mutant series the smaller the observed change in k_{Cm}, though the m_{kin} values change by as much as 30 %. This not only questions the validity of a two-state treatment but also the degree of mechanistic information one can extract by performing a mutation analysis.

d) Linear Temperature Dependence of Relaxation Rates Below and Above the T_m

More common is the characterization of the temperature dependence of rates of fast folding proteins. Figure 5.1B shows the data for nine fast folding proteins studied by T-jump experiments. This database includes proteins of length ranging from 32 to 80 residues, α-helical and β-proteins apart from the de-novo designed protein α_3D and vary in midpoint (T_m) rates by almost 2 orders of magnitude. In spite of these differences, the proteins show a remarkably similar dependence with temperature. The dependence is marginal at low temperatures, and becomes more pronounced upon crossing the T_m (here, T_m is more or less the point at which the slope changes) resembling a stretched-L. This is in striking contrast to that of the slower two-state

folding proteins (see Figure 2.3A) that show opposing dependencies with temperature close to the T_m. The rates at the T_m show no correlation with the length or absolute contact order in contrast to two-state folding proteins. The relative contact order does marginally better with a correlation coefficient of ~0.6. Moreover, there are clear disagreements between the thermodynamic T_m reported in the literature for these proteins and inflection point of the kinetic curves.

e) *Probe Dependent Kinetics* The temperature dependence of villin wildtype (purple) followed by fluorescence and the corresponding N27H mutant (cyan) monitored by infra-red are markedly disparate. The reported apparent T_ms are very similar (~ 335 K) ruling out any possible stability effects. A single point mutation might change the magnitude of the rate but it is highly unlikely to affect the shape of the temperature dependence plot for a two-state protein, given that they have similar T_ms. Probe-dependent kinetics have been previously reported for mutants of lambda repressor that fold over marginal barriers[58] suggesting that this behavior is possibly a manifestation of the same.

The observations outlined above suggest a distinct deviation from two-state behavior that, remarkably, has gone unexplained. The following sections analyze these issues quantitatively with a phenomenological statistical mechanical model.

5.3 Doshi-Muñoz (DM) Model

5.3.1 Theory and Model Parameterization

A variant of Zwanzig's one-dimensional free energy surface model that has been previously employed to explain the complex thermodynamic behavior of BBL is used. Zwanzig's model uses the number of residues in incorrect conformation (S) as

the reaction coordinate. Each residue can be either in a correct or incorrect conformation, and the entropy is directly obtained from all the possible combinations for each value of S. Instead, this model uses a property - nativeness (n) - as reaction coordinate. n is defined as the average probability of finding any residue in native-like conformations. It is a continuous version of the parameter (N-S)/N in Zwanzig's model (with N being the total number of residues and S the number of residues in incorrect conformation). The definition of n as a probability allows for straightforward calculation of the conformational entropy ($\Delta S^{conf}(n)$) using the Gibbs entropy formula:

$$\Delta S_{res}^{conf}(n) = -R\left[n\ln(n) + (1-n)\ln(1-n)\right] + n\Delta S_{res}^{n=1} + (1-n)\Delta S_{res}^{n=0} \text{ for } 0<n<1 \quad (5.1)$$

with

$$\Delta S_{res}^{conf}(0) = \Delta S_{res}^{n=0} = S_{res}^{n=0} - S_{res}^{n=1}$$

$$\Delta S_{res}^{conf}(1) = \Delta S_{res}^{n=1} = 0$$

$$\Delta S^{conf}(n) = N\Delta S_{res}^{conf}(n) \quad (5.2)$$

$\Delta S_{res}^{n=0}$ reflects the difference in conformational entropy between a residue that is populating all possible non-native conformations and the same residue in the fully native conformation.

The folding stabilization energy ($\Delta H^0(n)$) is modeled an exponential function of n:

$$\Delta H^0(n) = N\Delta H_{res}^0\left[1 + (\exp(k_{\Delta H}n - 1)/(1 - \exp(k_{\Delta H}))\right] \quad (5.3)$$

where ΔH^0_{res} is the stabilization energy per residue. The one-dimensional free energy surface for folding is then directly obtained from:

$$\Delta G(n) = \Delta H^0(n) - T\Delta S^{conf}(n) \tag{5.4}$$

In this model, the free energy barrier for folding arises from the non-synchronous decay of conformational entropy and stabilization energy, consistent with energy landscape descriptions of protein folding. Quadratic or higher order functionals can also be used for $\Delta H^0(n)$ as long as there is a sufficient difference between the folded and unfolded states energies (the so-called 'stability gap' hypothesis invoked to explain apparent two-state behavior[15]). But using an exponent simplifies the calculation with the ability to easily tune the shape and sharpness by just changing $k_{\Delta H}$, resulting in free energy profiles that vary from two-state to globally downhill. In fact, the energies of conformations generated in lattice and off-lattice when projected onto a single reaction co-ordinate have a similar dependence[8,118].

To model the effect of temperature on protein folding a heat capacity functional ($\Delta C_p(n)$) is defined that also decays exponentially with *n*:

$$\Delta C_p(n) = N\Delta C_{p,res}\left[1 + (\exp(k_{\Delta C_p} n) - 1)/(1 - \exp(k_{\Delta C_p}))\right] \tag{5.5}$$

$\Delta C_p(n)$ increases linearly with protein size as it has been observed empirically. The exponent determines the curvature of the heat capacity functional, which controls the value of the heat capacity at the top of the barrier for a two-state protein. Using the entropy convergence temperature (385 K) of Robertson and Murphy[97] as the temperature at which solvation terms to the entropy cancel out, the total entropy (conformational plus solvation) can be represented as:

$$\Delta S(T,n) = \Delta S^{conf}(n) + \Delta C_p(n)\ln(T/385) \tag{5.6}$$

The folding stabilization energy (equation 5.3) is then defined at the midpoint temperature leading to the following expression for the total changes in enthalpy as a function of temperature and *n*:

$$\Delta H(T,n) = \Delta H^0(n) + \Delta C_p(n)(T - T_m) \tag{5.7}$$

It is then straightforward to obtain the one dimensional folding free energy surface as:

$$\Delta G(T,n) = \Delta H(T,n) - T\Delta S(T,n) \tag{5.8}$$

This treatment of the temperature dependence for folding complies with existing empirical descriptions of thermal protein denaturation.

Chemical denaturation effects are modeled as changes in the total free energy of folding that depend linearly on denaturant concentration following:

$$\Delta G(F_D,n) = \Delta H^0(n) - T\Delta S(n) - mF_D \tag{5.9}$$

where $\Delta H^0(n)$ corresponds to the folding stabilization energy at the experimental temperature (equation 5.3), and $\Delta S(n)$ corresponds to the entropy functional at the experimental temperature (calculated using the conformational entropy equations 5.1 and 5.2). In this model, *m* describes the dependence of the chemical destabilization free energy on nativeness, defined phenomenologically as:

$$m = 1 - \left[(1+C)(n^j/(n^j + C))\right] \tag{5.10}$$

where *C* and *j* are phenomenological parameters. *m* goes from 1 for $n = 0$ to 0 for $n = 1$ and partitions the chemical destabilization free energy between the folding and unfolding sides of the barrier for two-state proteins in ratios that are consistent with empirical measurements of m_f / m_{eq}.

The relaxation kinetics arising from perturbations in the free energy surface are treated as diffusive following a Kramers-like treatment. Diffusive kinetics is simulated by employing a discrete representation of the free energy surface and the matrix method for diffusion kinetics of Lapidus *et al.*[119]. The effective diffusion coefficient is defined as:

$$D_{eff}(T) = k_0 \exp(-NE_{a,res} / RT) \tag{5.11}$$

For simplicity k_0 is assumed temperature independent. All the temperature effects arising from changes in solvent viscosity and internal friction from the protein (or landscape roughness) are embedded in the activation energy per residue ($E_{a,res}$).

5.3.2 Calculation of Free Energy Barrier Heights

Barrier heights are calculated from the free energy surface using a dividing line located at 2/3 of the distance in nativeness between the fully unfolded and native minima. The transition state ensemble is defined as the area centered in the dividing line and with width of 0.12 (for chemical denaturation) or 0.22 (for thermal denaturation) nativeness. Barriers are then obtained from the ratio between the weighted probability of the ground state (unfolded or native) and the transition state. The width of the transition state ensemble was calibrated to the specific shape of the free energy surface at the chemical and temperature midpoints to maximize agreement between folding and unfolding barrier heights and populations on both sides of the barrier.

5.4 Barrier Effects in Chemical Denaturation Experiments

5.4.1 Simulation and Model Predictions

To model the effect of chemical denaturants, a chain length of 80 residues (N) and a temperature (T) of 298 K was assumed. An entropic cost ($\Delta S_{res}^{n=0}$) of 10 J mol^{-1} K^{-1} per residue, consistent with the empirical estimates of Robertson and Murphy[97] at 298 K was used. The resulting entropic component of the free energy (calculated from equation 5.1) is shown in blue in Figure 5.2A. The enthalpic contribution for exponents ($k_{\Delta H}$) ranging from 0.1 to 4.5 is also shown (black curves). The sensitivity to denaturants along the reaction co-ordinate (m; equation 5.10) was modeled with the parameters j = 8 and C = 0.04 (blue curve in Figure 5.2B). This specific combination of coefficients produces chevron plots with three-fourths of the m in the folding limb and one-fourth in the unfolding limb for the high barrier cases. The ¾ - ¼ portioning is about the average seen for a number of two-state-like proteins. The signal decay along the reaction coordinate was modeled as a switching function at 65 % nativeness that goes from 0 to 1. This sharp change is similar to a fluorescence signal that can typically monitor only two conformations (solvent exposed and buried) and is the most common experimental probe. Simulations by Ma and Gruebele show that such step functions are reasonable estimates of the signal in one-dimensional free energy surface analyses[54].

In short, the magnitude of entropy determines the position and width of the entropy curve while free energy shapes are determined by the interplay between enthalpy and entropy. $k_{\Delta H}$ >1.5 result in steep enthalpy functionals and produce two-state-like free energy surfaces at all denaturant concentrations. For smaller values of

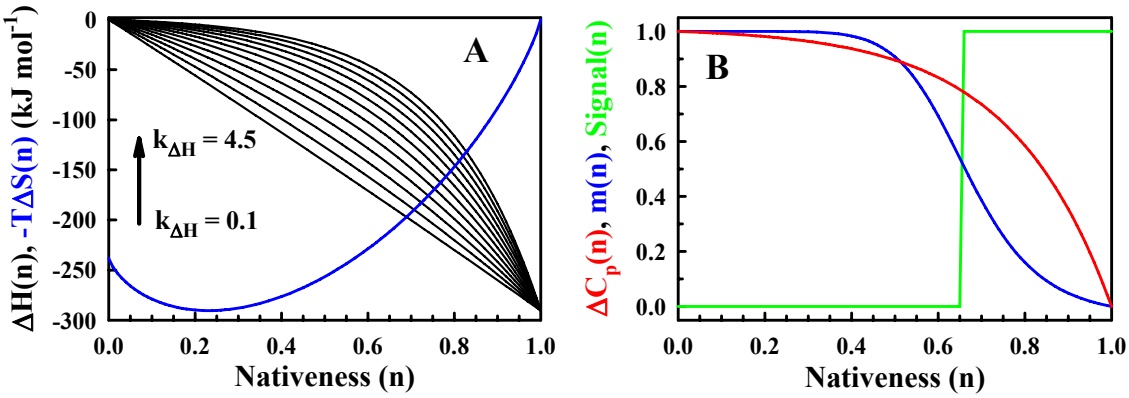


Figure 5.2 Functionals employed in the free energy surface analysis. A) Entropic (blue) and enthalpic (black curves) contributions to the free energy. B) Normalized heat capacity (ΔC_p (n); red), *m*-value (blue), and fluorescence signal (green) as a function of nativeness.

$k_{\Delta H}$ or shallower enthalpy functionals, the model generates surfaces that are either globally downhill (zero barrier heights at all denaturant concentrations) or switch from downhill to two-state. Figure 5.3A shows the calculated free energy surfaces at the chemical midpoint for different values of $k_{\Delta H}$, with the midpoint barrier heights (β_{Midpoint}) ranging from -2 to ~ 40 kJ mol^{-1}. The resulting macroscopic destabilization energy (ΔG_{eq}), calculated as the integrated probability on either side of a dividing line at 65 % nativeness, is linear at all values of $k_{\Delta H}$ (or the midpoint barrier height) consistent with experimental observations (Figure 5.3B). The population weighted signal (Figure 5.3C) as a function of induced chemical destabilization is sigmoidal for all cases. Interestingly, the slope of the plot of ΔG_{eq} versus F_D (m_{eq}) and the apparent cooperativity (or the width of transition) of the equilibrium unfolding curves are insensitive for β_{Midpoint} values > 10 kJ mol^{-1}. But the transition width gets broader upon decreasing barrier height highlighting the fundamental connection between the broadness and barrier height. The chevron plots simulated by performing diffusive

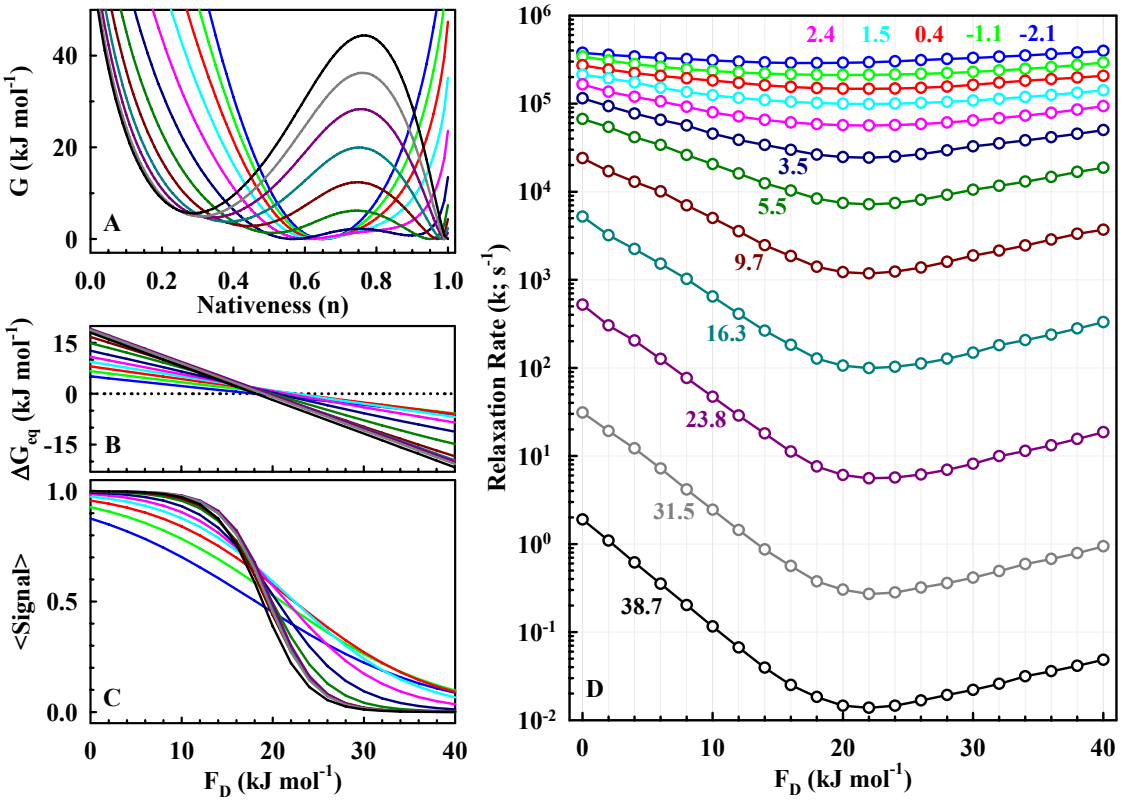


Figure 5.3 Simulation of chemical denaturation experiments. The coloring scheme corresponds to $\beta_{Midpoint}$ values ranging from -2.1 to 38.7 kJ mol^{-1} (labels in Figure 5.3D) and is maintained throughout the figure. A) Free energy profiles at the chemical midpoint. B, C, & D) Macroscopic stabilization free energy, population weighted signal, and chevron plots, respectively, as a function of the microscopic destabilization free energy (F_D).

kinetics on the generated free energy surfaces are shown in Figure 5.3D. For two-state-like scenarios (black curve), the slopes of the chevron are steep with the magnitude of the observed rate changing by more than two orders of magnitude from zero destabilization energy to the chemical midpoint. The plots are still v-shaped for folding over marginal barrier or globally downhill scenarios. Therefore, the observation of a chevron does not guarantee a two-state system. However, the chevrons flatten (or m_{kin} gets smaller) for smaller barrier heights suggesting that the degree of shallowness might have information on the nature of the transition.

5.4.2 Chemical Two-State Treatment

A more quantitative analysis can be performed by fitting each of the individual equilibrium unfolding curves and chevrons to a two-state model (Section 2.2.3.2) to extract m_{eq} and m_{kin}, as is traditionally done for such experiments. Figure 5.4A plots the results of such a fitting procedure. The m_{eq} value is insensitive to $\beta_{\text{Midpoint}} > 10$ kJ mol^{-1} but decreases abruptly for smaller barrier heights (blue curve). m_{kin} shows a similar but more pronounced dependence on barrier heights (red curve). This decrease in m_{kin} with decreasing β_{Midpoint} (or an increase in k_{Cm}) explains the observed negative correlation shown in Figure 5.1A, suggesting that the midpoint barrier heights for these proteins are smaller than 10-15 kJ mol^{-1}. What is the origin of the decrease in *m*-values with barrier height? The answer can be found in the free energy plots shown in Figure 5.3A. As the β_{Midpoint} decreases both the folded and unfolded minima move closer, with the movement more pronounced for the unfolded minimum. m_{kin} and the position of the unfolded minima show a perfect correlation with β_{Midpoint}, indicating that structured denatured states automatically result in smaller folding barriers. This can be rationalized by the fact that as the enthalpy functional gets shallower (for $k_{\Delta H} < 1$) there is a larger compensation between enthalpic and entropic contributions to free energy along the reaction coordinate. This results in the unfolded minima getting more structured (higher values of nativeness and a smaller free energy) while at the same time decreasing the folding barrier height. Any free energy surface analysis would result in such an observation indicating that this in fact could be used as an alternate way to estimate the barrier heights. This is also supported by the observations of structured denatured states in

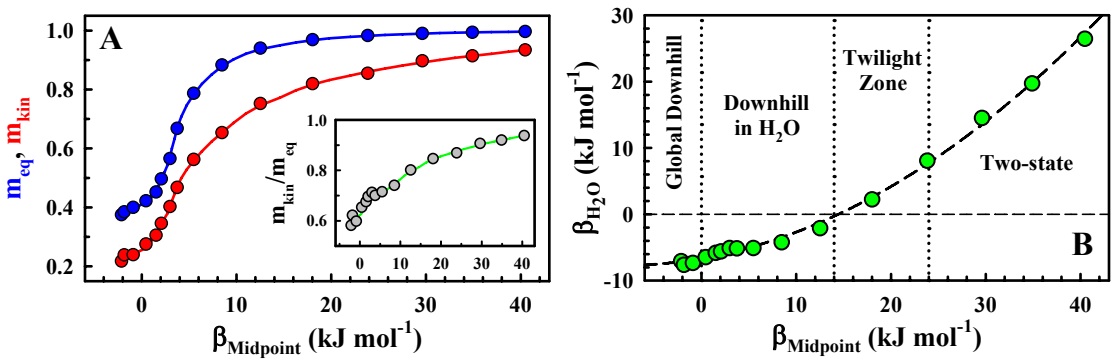


Figure 5.4 Barrier effects in chemical denaturation experiments. A) Dependence of m_{kin} and m_{eq} on midpoint barrier height ($\beta_{Midpoint}$). The inset plots the m_{kin}/m_{eq} ratio. B) Plot of the barrier height in water (β_{H20}) versus $\beta_{Midpoint}$ showing the four folding regimes.

fast-folding proteins[105] and recent analyses that attribute the changes in *m*-values to the changes in unfolded states' structure[120].

5.4.3 Protein Folding Phase Diagram

This suggests that the above treatment can be extended to estimate the midpoint barrier heights of proteins independent of the diffusion coefficient based on the m_{kin}/m_{eq} ratio. The inset to Figure 5.4A shows the ratio as a function of $\beta_{Midpoint}$. The values are smaller than one throughout as m_{kin} has a stronger dependence on the barrier height (Figure 5.4A main panel). The plot indicates that m_{kin}/m_{eq} gets as low as 0.6 when proteins fold globally downhill and approaches one upon increasing barrier height. However, extreme caution should be taken in the analysis of this ratio as the numbers are highly error prone. In spite of these caveats, there are data available in the literature with ratios well below the error threshold, *i.e.* $m_{kin}/m_{eq} < 0.9$ (or $\beta_{Midpoint} < 25$ kJ mol^{-1}). Specifically, m_{kin}/m_{eq} values of 0.89 for CspB[104] (no error reported), 0.74 ± 0.06 for engrailed homeodomain[108], 0.68 ± 0.04 for FBP28 W30A WW domain[105], and 0.74 ± 0.09 for BBL H166W[68] correspond to midpoint barrier heights

of 24, 6.7 ± 5, 2.1 ± 2.5, and 6.4 ± 8 kJ mol^{-1}, respectively. Folding over marginal barriers is predicted for the latter three proteins at the chemical midpoint, i.e. even when the rate is the slowest. Though informative, chemical midpoints conditions are rather artificial while the biologically relevant situation is the absence of denaturants. Figure 5.4B shows the plot of the barrier height in water (β_{H20}) versus the $\beta_{Midpoint}$ estimated from the free energy surface analysis. This provides a unique opportunity to characterize the protein folding phase diagram into four regimes. Globally downhill folding proteins (also known as one-state) fold over zero (or negative) barriers in water as well as at the midpoint (BBL for example). Two-state proteins are those that have a significant barrier > 9 kJ mol^{-1} (~3RT) even in native conditions, resulting in pronounced higher barrier heights (> 24 kJ mol^{-1}) at the chemical midpoint. Downhill folding proteins have zero (or negative) barrier heights in native conditions but fold over marginal barriers at the chemical midpoint ($0 < \beta_{Midpoint} < 14$ kJ mol^{-1}). Similar to the results of variable-barrier model analysis, twilight zone proteins can be classified as those that fold over marginal barriers in water ($0 < \beta_{Midpoint} < 9$ kJ mol^{-1}) and moderate barriers at the chemical midpoint ($14 < \beta_{Midpoint} < 24$ kJ mol^{-1}).

5.4.4 Comparison with Experiments

To estimate the barrier heights of proteins the magnitude of the effective diffusion coefficient should be known (D_{eff}). Does the mutant data provide any clue in this aspect? Figure 5.1A suggests that the mutants or homologues of a protein have a specific dependence (i.e. varying slopes of the plot) with the rate at the chemical midpoint (k_{Cm}). Therefore, one could superimpose these mutant data onto an appropriate segment of the theoretical m_{kin} curve. In other words, their slopes could be

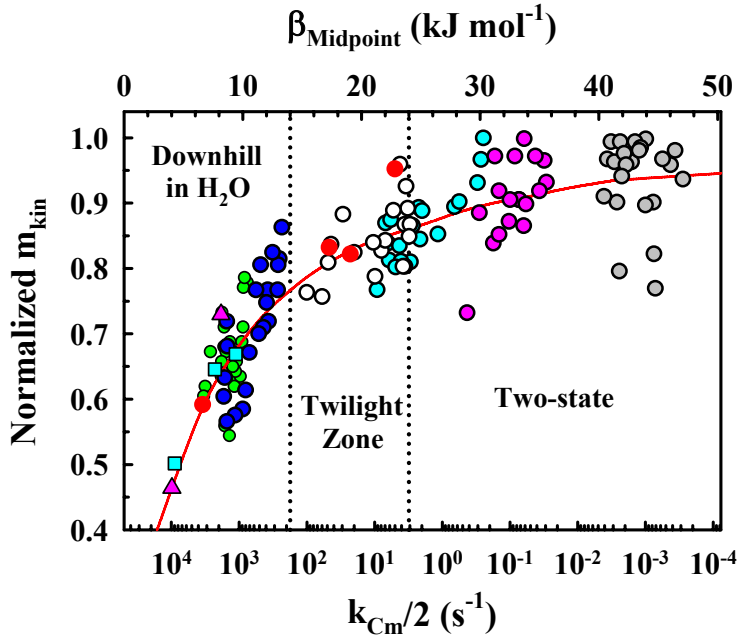


Figure 5.5 Superimposition of the theoretical m_{kin} curve and normalized experimental data for engrailed family (red circles), BBL-related variants (pink triangles), WW domain family (cyan squares), PDD F166W (dark blue circles), FBP28 WW domain (green circles), CspB (white circles), yeast ACBP (cyan circles), L23 (pink circles), and muscle AcP (gray circles). The abscissa on the top represents the midpoint barrier heights calculated with a pre-exponential of 1/(20 μs). (*m*-value and rate data from published works).

matched by assuming a specific value for the diffusion coefficient to convert the barrier heights into rates. This would enable the estimation of the pre-exponential for every mutant series. Unfortunately, the experimental accuracy in *m*-values is too low for such an exercise. An alternative is to combine the experimental data from several proteins spanning a large range in midpoint rates by using an average diffusion coefficient. But this comparison presents a problem. *m*-values of proteins depend on the size, structure and composition, thus entailing a normalization procedure. The position on the x-axis for each protein dataset was obtained by converting its average rate at midpoint to a free energy barrier using a pre-exponential factor of 1/(20 μs) at 298 K. The experimental *m*-values for each protein dataset were then normalized

using the expression: $\left(m_{kin}^{i} / \langle m_{kin} \rangle\right) y$, where y is the y-axis value in the theoretical curve that corresponds to the average barrier height of the mutant series. The result of such a procedure is shown in Figure 5.5 with data from two BBL-related variants (BBL H166W and PDD F166W), engrailed homeodomain family, three WW domains (YAP, Prototype and FBP28 W30A; mutants of FBP28 WW domain), two millisecond folding proteins (CspB and yACBP) and two slow folding proteins (L23 and mAcP). The slope of each of the mutant series agrees remarkably well with the theoretical curve. Furthermore, the barrier heights for CspB, engrailed homeodomain, FBP28 WW domain and BBL H166W agree with the independent estimates obtained from the m_{kin}/m_{eq} ratio.

The free energy surface analysis is therefore able to quantitatively explain the observed deviations from true two-state behavior and providing strong evidence that these are indeed the manifestations of folding over marginal/zero barriers. But it also leads to a number of intriguing conclusions with the prominent one being: why is the average pre-exponential value of 1/(20 μs) at 298 K an order of magnitude lower than that estimated by other groups (~1/(2 μs)). The current estimate is similar to average pre-exponential (1/(25 μs)) obtained from the comparison of barrier heights from Variable Barrier analysis and the rates at 298 K[65]. It is however important to note that the pre-exponential values reported by other groups correspond to temperatures of ~330-340 K, suggesting the possibility of a temperature dependent diffusion coefficient. To answer this question, the experimental data of proteins shown in Figure 5.1B is analyzed with the same model by incorporating temperature effects.

5.5 Barrier Effects in Thermal Denaturation Experiments

5.5.1 Simulation and Model Predictions

In the previous section, the experimental data was not directly fit to the model. The trends were explained by invoking an average pre-exponential. However, individual characterization of the temperature dependent rates is more challenging requiring reasonable estimates of the entropic cost of fixing a residue in native conformation ($\Delta S_{res}^{n=0}$) and the change in heat capacity per residue upon unfolding ($\Delta C_{p,res}$). The empirical estimate of Robertson and Murphy was therefore used providing a $\Delta S_{res}^{n=0}$ of 16.5 J mol^{-1} K^{-1} per residue at the convergence temperature of 385 K[97]. The heat capacity functional (equation 5.5) was parameterized by fitting the calorimetry profiles of 14 proteins (used in the Variable Barrier analysis) to this model. This resulted in a $\Delta C_{p,res}$ of 50 J mol^{-1} K^{-1} per residue and a $k_{\Delta C_p}$ of 4.3. The fitted $\Delta C_{p,res}$ value is very similar to the empirical estimates of ~58 J mol^{-1} K^{-1} per residue. The final values of the entropic cost and the heat capacity change were then scaled by the size of the corresponding proteins. A thermodynamic description of the system can thus be obtained by fitting just two parameters: T_m, the thermal midpoint and $k_{\Delta H}$, the parameter determining the curvature of the enthalpy functional. The values of stabilization energy per residue (ΔH_{res}^{0}) were manually adjusted for every protein to avoid convergence problems (similar to a grid analysis). To describe the kinetics, two additional parameters are required: k_0, the temperature independent fundamental rate constant and $E_{a,res}$, the activation energy per residue. The complete description (thermodynamic + dynamic) therefore requires only 4 parameters,

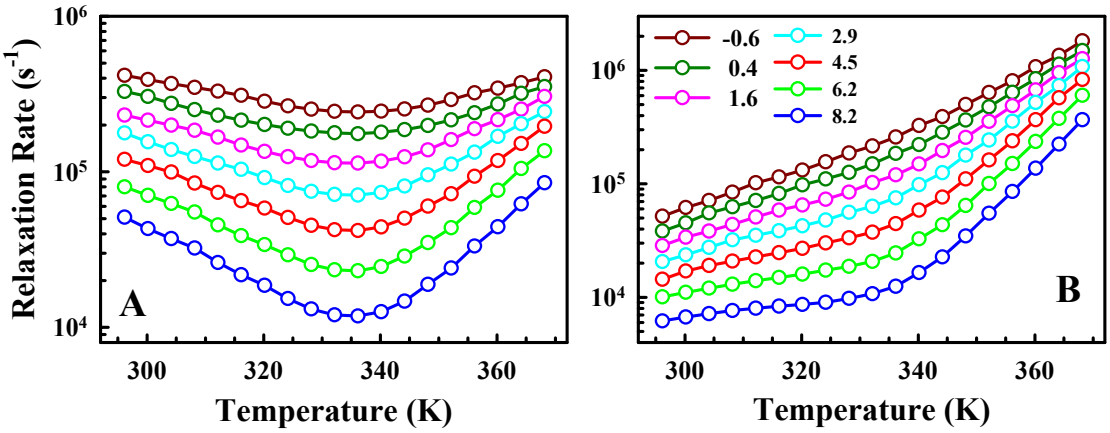


Figure 5.6 Barrier effects in thermal denaturation experiments. A & B) Simulated relaxation rate versus temperature for examples of 50-residue proteins with midpoint barrier heights ($\beta(T_m)$) ranging from -0.6 (dark red) to 8.2 (blue) kJ mol^{-1} in the absence (A) and presence (B) of an activated diffusion coefficient of 0.9 kJ mol^{-1} per residue.

compared to the 6 required in a two-state model. This model is far more superior to a two-state treatment as it not only provides the free energy surface at various temperatures but also the diffusion coefficient.

Figures 5.6A and 5.6B illustrate the dependence of relaxation rates on free energy surfaces with midpoint barrier heights varying from -0.6 to 8.2 kJ mol^{-1} generated by the model for a 50-residue (N) protein with a T_m of 335 K. The size and the T_m agree with the average numbers for this protein dataset. The simulation in Figure 5.6A is performed without incorporating a temperature dependent diffusion coefficient and hence the shape is entirely determined by the thermodynamic properties of the system. The relaxation rate plots are roughly V-shaped with a minimum at the T_m and speeding up at lower and higher temperatures. In the absence of any heat capacity effects, the plots should be perfectly V-shaped. The incorporated heat capacity effects can be seen from the slight downward curvature of the folding limbs – due to cold denaturation as can be visualized in plots of stability versus

temperature. This curvature is less apparent than that of two-state folding proteins because of the limited temperature range. The unfolding limbs are almost linear with temperature as the heat capacity effects are less pronounced and constant. As observed in chemical denaturation experiments, the shapes of the plots flatten with decreasing barrier heights. If the diffusion coefficient is temperature independent the rate versus temperature plots should therefore look V-shaped rather than the 'stretched-L' dependence shown in Figure 5.1B. This convincingly suggests that the diffusion coefficient is indeed temperature dependent.

But what determines the temperature dependence? The first candidate is the temperature dependence of water viscosity that contributes to ~ 16 kJ mol^{-1} to the activation term. In addition to this simple effect there are higher order effects arising out of barriers to peptide bond rotation and due to breaking non-native interactions as the folding proceeds[5,121]. These would contribute to bumps on a higher dimensional free energy surface, but are lumped into a single effective diffusion coefficient in a one-dimensional representation. Furthermore, activated terms arising out of crossing these microbarriers should scale with protein size as folding dynamics involve concerted motions of the entire polypeptide chain. Lattice simulations also suggest co-ordinate dependent diffusion coefficient and super-arrhenius dependence (see Chapter 1; equation 1.2). However, to simply the analysis a simple Arrhenius temperature dependence is considered. Figure 5.6B shows the effect of introducing a temperature dependent diffusion coefficient with an activation energy of 45 kJ mol^{-1} or $E_{a,res}$ of 0.9 kJ mol^{-1}. The resulting plots are remarkably similar to the experimental data. For larger barriers (blue and green curves), the rate dependence is typically L-

shaped showing a slight downward curvature at temperatures < T_m due to the combined effect of the activation term and the positive heat capacity. As the barrier height decreases the plots tend to get almost linear with temperature, i.e. more downhill the protein the more linear it gets. These results suggest that the shape of the rate versus temperature plot has information on the barrier height as well as the diffusion coefficient.

5.5.2 Reproducing Experimental Relaxation Rate Plots

The 4 parameter fit of the experimental data to this model is shown in Figure 5.7. The quality of fits is very similar to the original two-state fits produced by the authors. The results are summarized in Table 5.1. It shows that the activation energy scales by ~1 kJ mol^{-1} per residue except for the designed α_3D that shows a markedly weak dependence. Fits performed by fixing the activation term to just the viscosity dependence of water failed to reproduce the high slopes seen in the unfolding limbs of the plot. Super-Arrhenius fits (not shown) were only marginally better than the Arrhenius fits. This is because of the limited temperature range of the available data and the absence of amplitude information for any of the proteins. It is also interesting to note that the size-scaling of the activation energy agrees with that estimated for a 20-residue α-helix[17]. The estimated barrier heights at the T_m are small (Table 5.1). 1prb appears to fold globally downhill in agreement with simulation results[122]. The barrier height at T_m is in good agreement with independent computational estimates for λ-repressor[123] and Pin WW domain[124]. The predicted midpoint barrier height of 5.8 kJ mol^{-1} for engrailed homedomain is consistent with the observation of a faster

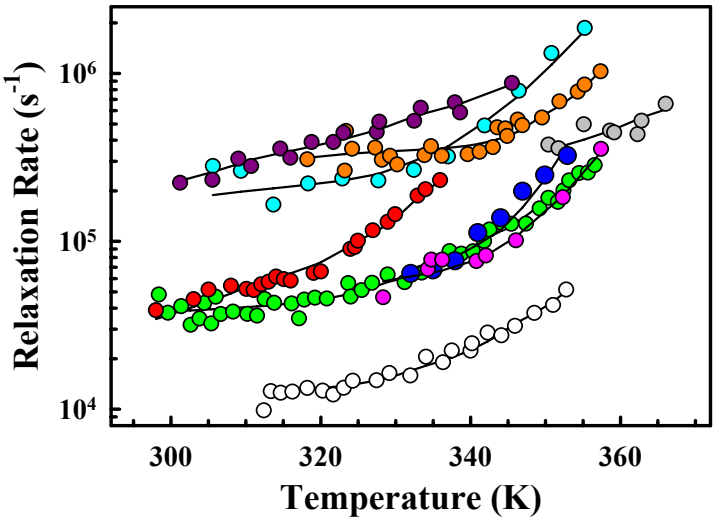


Figure 5.7 Fits (black curves) to the experimental data for the nine microsecond folding proteins shown in Figure 5.1B (coloring scheme is maintained). (Rate data from published works).

phase in kinetic experiments that are diagnostic of folding over marginal barriers[125]. However, it is important to note that barrier heights and activation energies reported in Table 5.1 are upper estimates. This is because $\Delta C_{p,res}$ is directly correlated to height of the barrier and the value of 50 J mol^{-1} K^{-1} per residue is likely to be an over-estimate (see Chapters 3, 6 and 7). A higher $\Delta C_{p,res}$ also results in a higher curvature in the folding limb of temperature versus rate plots, thus requiring more activation to account for the flat low temperature dependence seen in experiments. These effects suggest the need for a better thermodynamic description of protein folding to accurately determine parameters of physical significance.

Table 5.1 lists the individual diffusion coefficients to folding as a function of temperature. The median value of diffusion coefficient at the T_m ($\langle T_m \rangle$ = 335 K) is ~ 1/(2.5 μs) similar to the recent empirical estimates [25]. The speed limits at T_m for the fast-folding mutant of lambda repressor[55] and for the N27H mutant of Villin

headpiece[126] are in close agreement with estimates made by the authors with independent methods. The model is also able to explain the disparate rate behaviors of villin N27H mutant

Table 5.1 Parameters from the free energy surface analysis of thermal denaturation kinetics in microsecond folding proteins

Protein	**Length** (N)	$k_{\Delta H}$	$E_{a,res}$	T_m (K)	β (T_m)	τ_{min} (T_m)	β_F (298)	β_U (298)	τ_{min} (298)
FBP WW domain*	32	1.54	0.77	327	7.7	2.2	4.1	10.1	5.5
Pin WW domain	34	1.52	0.76	332	6.8	9.8	2.5	10.0	32.9
Villin N27H	35	1.28	1.22	334	6.9	0.5	0.7	11.8	4.7
Villin HP36	36	0.52	0.88	335	0.5	2.8	-3.3	3.0	4.5
$1prb_{7\text{-}53}$ K5I	47	1.08	1.15	369	-3.1	1.5	-12.7	3.3	30.8
Engrailed	52	0.75	1.07	325	5.8	2.5	1.4	8.9	17.2
BdpA	58	1.46	1.07	346	5.6	2.8	-4.5	12.6	57.2
α_3D	73	1.05	0.55	346	1.6	2.6	-8.6	9.2	5.4
$\lambda_{6\text{-}85}$ D14A	80	1.36	0.99	346	5.9	2.3	-6.6	14.5	80.4

Minimum folding times (τ_{min}) are in microseconds and $\beta/E_{a,res}$ values in kJ mol^{-1}.

followed by fluorescence and villin wildtype monitored by FTIR, resulting in barrier heights and minimal folding times at T_m that are widely different. Moreover, results from *m*-value analysis suggest that proteins folding over marginal barriers are highly sensitive to mutational changes (see Figure 5.1A for example). The rate behavior of villin and the model's prediction are therefore consistent with this observation. In fact, there are significant differences in the rate behavior of several single-point mutants of villin N27H[111,126]. It is also of interest to note that Variable Barrier model analysis of the PDD wildtype resulted in a barrier height of ~0.5 kJ mol^{-1} while a single point mutation of F to W increased the barrier height by ~4 kJ mol^{-1}. These drastic changes upon a single point mutation are also suggestive of folding over marginal barriers in which the tradeoff between energetic and dynamic contributions

to the relaxation rate is delicate. Intriguingly, the minimal folding times of FBP and Pin WW domain at the T_m differ by almost of factor of 5 in spite of resulting in similar barrier height estimates. They are homologues with a very high sequence similarity. These results were validated in a recent mutant analysis of Pin WW domain in which the fastest and supposedly downhill folding mutant relaxed at a rate of 10 μs at the midpoint[127].

Biologically relevant quantities are the barrier heights and minimal folding times at 298 K. Table 5.1 shows that most of the proteins from this dataset fold downhill under native conditions. A notable exception is the truncated version of FBP WW domain that folds over a marginal barrier of 4.1 kJ mol^{-1} (~1.6 RT). Downhill folding at 298 K for engrailed homeodomain was also predicted by *m*-value analysis and Variable Barrier analysis. The estimated minimal folding times at 298 K differ from those at T_m by an order of magnitude with a median value of 17 μs. The intrinsic errors are larger at 298 K due the long extrapolation from T_m for stable proteins that have no data points at lower temperature, especially for BdpA $1prb_{7\text{-}53}$ K5I and $\lambda_{6\text{-}85}$ D14A. But the predicted value of 80 μs for $\lambda_{6\text{-}85}$ D14A is consistent with the rates obtained for many other mutants of this protein[117]. Importantly, the median value is strikingly similar to the estimations from *m*-value and Variable Barrier analysis. Also, a value of 17 μs at 281 K was predicted by Sabelko and Gruebele in their renaturation analysis of cold denatured PGK[128]. The excellent agreement between four fundamentally independent estimates therefore suggests that a value of 20-25 μs is therefore a sound estimate of the magnitude of the average minimal folding time at 298 K.

5.6 Conclusions

The results from a quantitative treatment of the chemical and thermal denaturation experiments on fast folding proteins reveal that they do indeed fold over marginal barriers at around the T_m/C_m. More importantly, the proteins are predicted to fold downhill under native conditions, i.e. the absence of denaturants and 298 K. This observation therefore highlights the need to exercise caution in a two-state treatment of protein folding data far from the denaturation midpoint. The predicted barriers and pre-exponential to the folding reaction are consistent with one another and in agreement with independent estimates of folding speed limits and with the thermodynamic barriers extracted from DSC data. This analysis further indicates that the pre-exponential includes an activation term that scales linearly with protein size as ~1 kJ mol^{-1} residue. A direct consequence of this effect is that the average pre-exponential at 298 K (~1/(25 μs)) is much smaller than the estimates at 330-340 K (~1/(1-5 μs)), suggesting an increasing roughness with a decrease in temperature. The theoretical treatment also outlines two other experimental signatures in classical experiments to distinguish between two-state and downhill folding mechanisms: the m_{kin}/m_{eq} ratio and the shape of temperature versus relaxation rate plots.

6. Robustness of Downhill Folding: Guidelines for the Analysis of Equilibrium Folding Experiments on Small Proteins

6.1 Introduction

BBL is the first independently folding domain that has been experimentally shown to fold globally downhill. It is a small 40-residue all-helical sub-domain, a part of a much larger E2 subunit from the 2-oxoglutarate multi-enzyme complex of *Escherichia coli*. The global downhill behavior was initially identified by studying the equilibrium thermal unfolding using multiple structural probes and characterizing the complex thermodynamics by an elementary statistical mechanical model (for a description see Chapter 8) [47]. These results were further confirmed by investigating the coupling between temperature and chemical denaturation in BBL[52], by the variable barrier model that extracts barrier heights from DSC thermograms[49] (Chapter 4) and by studying the temperature induced chemical-shift perturbation of 158 individual protons in the structure[48]. The conclusions from these widely different techniques were self-consistent suggesting a scenario wherein BBL unfolds gradually with different structural ensembles populating the various stability conditions. Most of the experiments, including the CD, DSC and NMR, were performed in a protein encompassing residues 111 to 150 of the E2 subunit, in which alanine 111 was substituted by naphthyl-alanine (hereafter named Naf-BBL) [129]. The FRET and fluorescence measurements were carried out in another variant with an additional C-terminal probe of Dansyl-lysine (named Naf-BBL-Dan). Both of these proteins were synthesized with the ends free.

In spite of the careful spectroscopic and quantitative characterization of the downhill folding behavior in BBL, considerable criticism has been raised questioning the validity of the initial assessment. It has been claimed that these observations are an artifact due to the following reasons:

a) the use of hydrophobic fluorescence probes that apparently perturb the folding behavior,
b) employing short boundaries from the E2 construct and thus deleting potentially important interactions (tail effects), and
c) the low stability experimental conditions (i.e. low ionic strength).

Particularly, Fersht and co-workers have reported an investigation of the equilibrium unfolding behavior of another variant of BBL[68]. This version incorporates four additional residues at the N-terminus, has no fluorescent labels and has been produced recombinantly (hereafter named QNND-BBL). Their experiments were also carried out under higher ionic strength conditions. They find their version to be ~8 K more stable and that it complies with two-state folding criteria.

This chapter deals with the issues raised above with particular emphasis on the absolute characterization of signals, the interpretation of baselines, and the validity of employing the criteria used to distinguish between two-state and three-state folding on small fast-folding proteins that have broad unfolding transitions. It will be shown that,

a) Naf-BBL and Naf-BBL-Dan have identical thermodynamic properties and undergo reversible thermal unfolding (Section 6.2),

b) QNND-BBL does not fold in a two-state fashion and that it has a folding behavior similar to Naf-BBL, albeit with a higher stability (Section 6.3),

c) a variant of Naf-BBL with the ends protected, i.e. acetylation and amidation at the N- and C-termini, respectively, (termed Ac-Naf-BBL-NH2) folds with a thermodynamic stability similar to QNND-BBL (Section 6.4),

d) The stabilities of the variants can be easily tuned by ionic strength (Section 6.5), and

e) Ac-Naf-BBL-NH_2 shows all the equilibrium signatures for downhill folding (Section 6.6).

Naf-BBL: NH_3^+-NafA-LSPAIRRLLAEHNLDASAIKGTGVGGRLTREDVEKHLAK-COO^-

Naf-BBL-Dan: NH_3^+-NafA-LSPAIRRLLAEHNLDASAIKGTGVGGRLTREDVEKHLA-DanK-COO^-

QNND-BBL: NH_3^+-QNNDALSPAIRRLLAEHNLDASAIKGTGVGGRLTREDVEKHLAKA-COO^-

Ac-Naf-BBL-NH_2: CH_3CONH-NafA-LSPAIRRLLAEHNLDASAIKGTGVGGRLTREDVEKHLAK-$CONH_2$

Figure 6.1 Sequences and names of the four BBL variants. NafA, naphthyl alanine; DanK, dansyl lysine; Ac, acetyl.

6.2 Singly and Doubly Labeled BBL Unfold Reversibly and with the Same Thermodynamic Properties

The unfolding behavior of Naf-BBL is extremely reversible even in millimolar protein concentrations typically employed in DSC experiments. Figure 6.2A shows the raw DSC thermograms of a series of four heating-cooling scans of Naf-BBL after baseline subtraction. The maximum of the thermogram is identical for all scans at ~ 324 K. There is no decrease in the amplitude of the scans upon successive reheating that would otherwise accompany any irreversible aggregation effects. The small low temperature artifact during the first heating is probably a result

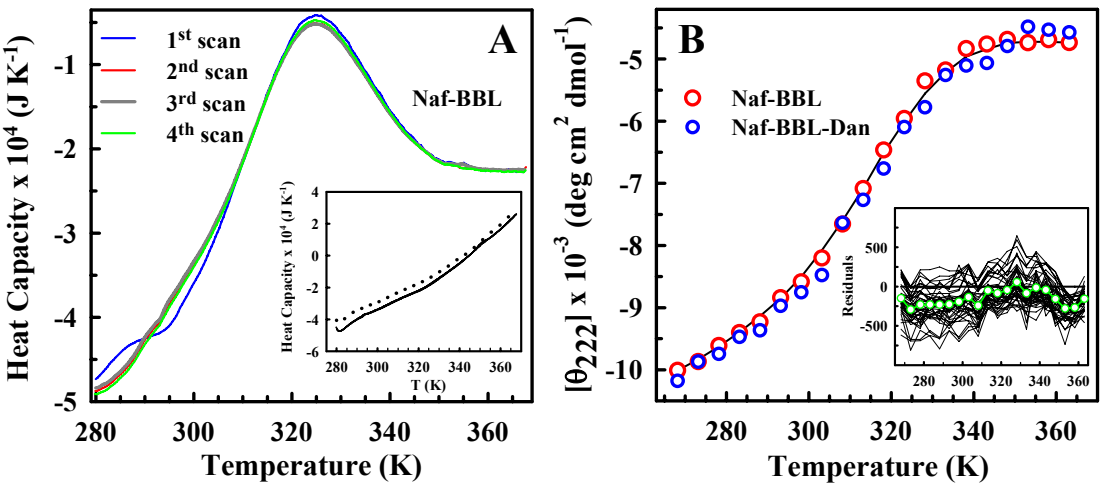


Figure 6.2 A) Reversibility of Naf-BBL thermal unfolding monitored by DSC. Thermograms are shown in raw heat capacity units and baseline corrected. (inset) Baseline reproducibility; (continuous lines) six subsequent baselines measured before measuring the protein, (dotted line) baseline upon refilling the calorimeter after the protein scan. B) Thermal unfolding curves for Naf-BBL (red) and Naf-BBL-Dan (blue), at 50 mM and 12.7 mM protein concentrations, respectively, and in 5 mM sodium phosphate buffer, pH 7.0. The continuous line is shown to guide the eye. (inset) Residuals between the data for Na-BBL-Dan and Naf-BBL for different wavelengths. The green circles represent the wavelength-averaged residuals. (far-UV CD experiments by Naganathan AN; DSC experiments by Perez-Jimenez R & Sanchez-Ruiz JM).

of structural changes induced by lyophilization of the protein samples[130,131]. DSC experiments could not be carried out on Naf-BBL-Dan as it aggregates at millimolar concentrations because of the hydrophobic dansyl group. This by no means suggests that the folding behavior is perturbed. However, the unfolding could be followed by CD measurements as they require significantly smaller concentrations. It has also been shown earlier that Naf-BBL-Dan unfolding is reversible under low ionic strength buffers[23]. In view of these considerations, the thermal unfolding was monitored by CD in 5 mM sodium phosphate buffer (ionic strength ~11 mM) at pH 7.0. The thermal unfolding of Naf-BBL-Dan is 100 % reversible under these conditions (not shown). Figure 6.2B shows the signal at 222 nm in molar ellipticity

units for Naf-BBL-Dan as a function of temperature (blue circles). The data from Naf-BBL at 50 μM protein concentration and under these conditions is shown for comparison (red circles). The Naf-BBL-Dan data is relatively noisier because of the low protein concentrations used. The inset shows the residuals between the CD signals of Naf-BBL and Naf-BBL-Dan at the different wavelengths, together with the wavelength-averaged residuals (green circles). The small magnitude and the lack of any apparent trend in the residuals suggest that the spectrum of these proteins is essentially the same at all temperatures.

The unfolding of Naf-BBL-Dan monitored at 222 nm is identical to Naf-BBL as the curves pretty much overlay on top of each other. A quantitative description can also be obtained by fitting the curves to a two-state model to estimate the thermodynamic parameters. Though unphysical for these downhill folding systems, it provides a simple common ground to compare the transitions. As discussed in Chapter 2, a two-state model bases all its description of a system on just two parameters: ΔH_m and T_m (there is no information for estimating ΔC_p here). Thus, fitting the data shown in Figure 6.2B requires 6 parameters in total including the linear folded and unfolded baselines. However, such a fit for signals with steep pre-transition slopes is highly sensitive to the description of baselines (indicating non-two-state behavior). This problem can be overcome by performing a simple statistical analysis as described below. The 6 parameter two-state fit for the relatively less noisy Naf-BBL data produces a ΔH_m of 91 kJ mol^{-1} and a T_m = 321 K. This value is ~3 K lower than the maximum of the DSC thermogram, consistent with the probe dependent T_m previously reported for BBL[47]. Fitting the unfolding curve of Naf-BBL-

Dan by floating the baselines, but fixing the ΔH_m and T_m to the values from Naf-BBL, produces a fit with a sum of least squares (SLS) just 9 % higher than the SLS of its best unconstrained fit. This value is lower than the 20 % increase in SLS expected if the discrepancy between the two sets of parameters is due to experimental noise alone. These numbers were estimated by generating noise-free two-state curves and adding statistical noise to match the magnitude of the observed experimental noise. These curves were then fit by constraining their T_m and ΔH_m to their original values by allowing just the baselines to float. Such fits rendered SLS values that were 20 % higher on average compared to the unconstrained fit that employs 6 parameters. Therefore, the statistical analysis indicates that Naf-BBL and Naf-BBL-Dan have identical thermodynamic parameters within experimental error. Moreover, these results also suggest that the anomalous folding behavior observed by Fersht and co-workers[67] is due to the ill-advised choice of experimental conditions that promoted aggregation in Naf-BBL-Dan.

6.3 QNND-BBL is Not a Two-State Folder

One of the main conclusions of Fersht and co-workers is that their QNND-BBL that is 4 residues longer at the C-terminus and has no fluorophores folds in a two-state fashion with a higher stability[67]. This sections deals with the reason for this misclassification.

6.3.1 Wavelength Dependent T_m by far-UV CD

A simple experimental criterion that has been proposed to identify downhill folders is the probe-dependence of T_m especially when they monitor different

structural features. This has been validated both computationally based on a statistical mechanical model and experimentally on BBL. This is one of the reasons for incorporating the fluorescent probes at the ends of BBL so that both fluorescence and end-to-end distance changes can be monitored, apart from DSC and CD. In fact, it is possible to discern between a simple two-state folding and a more complex conformational behavior by just collecting temperature dependent CD spectra. This is because of the fact that the shape and magnitude of CD spectra are dependent on the length and straightness of the helix. It is a consequence of exciton effects as a CD signal essentially arises out of the number and degree of alignment of peptide bond dipoles with the α-helix axis. Work from several groups has shown that changes in the magnitude and shape are apparent when monitoring the specific wavelengths of 193, 208 and 222 nm (see Chapter 2), as they correspond to the characteristic alpha helical bands. In the original report on downhill folding, these wavelengths showed drastically different T_ms and pre-transitions. They are shown here in Figure 6.3A for comparison. The normalized signal at 200 nm is also shown as it monitors the population of the coil signal. For a two-state system the temperature dependence of the normalized signal should be identical, as they there are only two structural species present: folded and unfolded state. However, a downhill folding system or any α-helix for that matter will have significant fraying at the ends resulting in a population with varying helix lengths co-existing in equilibrium. The equilibrium distribution further changes with temperature. This would in turn affect the shape of the α-helical spectrum resulting in non-coincident unfolding transitions as shown in Figure 6.3A.

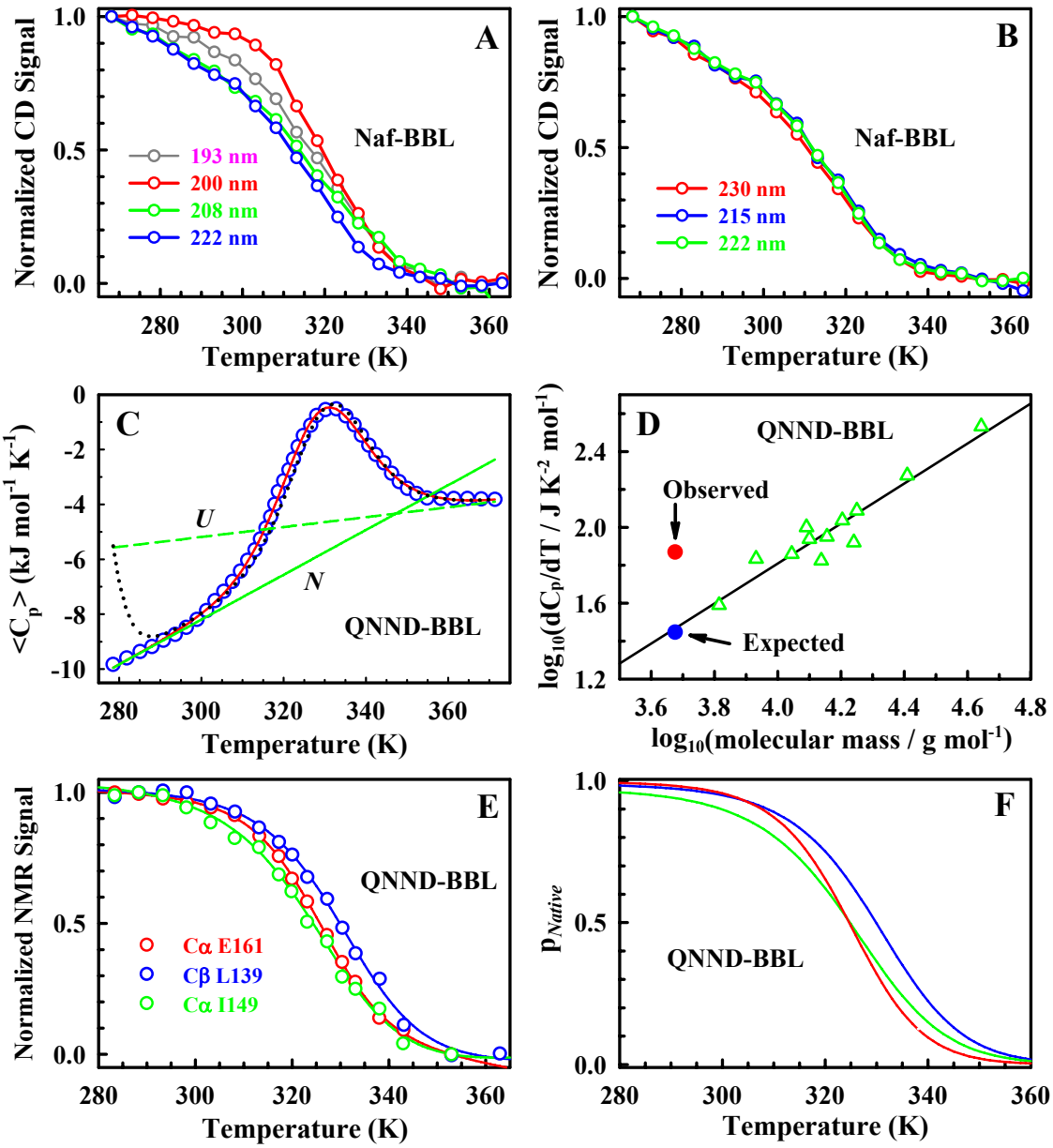


Figure 6.3 A) Normalized thermal unfolding transitions of Naf-BBL monitored by CD at the wavelengths used in Garcia-Mira et al. B) Normalized thermal unfolding transitions of Naf-BBL monitored by CD at the wavelengths used by Fersht and co-workers. C) DSC thermogram of QNND-BBL: (blue circles) data of Fersht and co-workers, (red curve) fit to a two-state model enforcing $\Delta H_{Cal}/\Delta H_{vH}$ of unity, (green lines) folded and unfolded baselines, and (dotted curve) thermogram predicted by calculating ΔC_p from the baselines. D) Empirical correlation between dC_p/dT and molecular mass; (green triangles) experimental data from the 12 proteins originally used in Freire's correlation, (blue circle) estimation for QNND-BBL, (red circle) measured for QNND-BBL from the 280 – 300 K data in panel C. E) Normalized ^{13}C chemical shifts as a function of temperature for QNND-BBL as measured by Fersht and co-workers and the two-state fits (continuous lines). F) QNND-BBL native state

probabilities calculated from the fits shown in panel E. (far-UV CD experiments by Naganathan AN; QNND-BBL experiments from Fersht group's published data).

Interestingly, Fersht and co-workers report that QNND-BBL shows no wavelength-dependent unfolding transition suggesting the apparent compliance to two-state folding. They in fact monitor wavelengths in range of 215-230 nm which is quite different from that originally used in studying BBL. The question of importance is then: what does the wavelength range of 215-230 nm monitor? In proteins with only helical and coil segments, this reports entirely on just one of the three a-helical bands (i.e. 222 nm band). Therefore, it is not surprising that they find identical unfolding transitions. This is demonstrated in Figure 6.3B that shows the normalized CD signals of Naf-BBL at three wavelengths of 215, 222 and 230 nm spanning the range originally used by Fersht and co-workers. The three curves are identical with respect to the pre-transition slopes, the T_m and the post-transitions. This argument can in fact be generalized to other spectroscopic probes typically used to ascertain the mechanism of folding. It is highly unlikely to observe differences in the unfolding transition of a protein when monitored by fluorescence, absorbance, near-UV CD and NMR of a single chromophore, even if it folds in a non-two-state fashion.

6.3.2 Crossing Baselines in a Two-State Analysis of DSC

Fersht and co-workers also claim that the DSC thermogram of QNND-BBL shows a weak low temperature dependence (pre-transition) apart from complying to the calorimetric criterion of $\Delta H_{Cal}/\Delta H_{vH} = 1$. However, they do not give any quantitative information on the temperature dependence of the pre-transition or the fitted calorimetry baselines. The significance of both these estimates has been

discussed in Chapter 4. The baselines correspond to the fluctuations in the low and high temperature ensembles. Particularly, the slope of the low temperature baseline (dC_p/dT) has been shown to scale with the size of the protein by Freire and co-workers[99]. Therefore, any slope value higher than that predicted by this Freire's baseline (equation 4.1) is suggestive of a non-two-state situation as it suggests fluctuations that cannot be explained by a unique 'native' state. The DSC data of QNND-BBL was not reported in absolute units thus eliminating the possibility of comparing the intercept of the baseline with that predicted by Freire's. However, it is still possible to measure the slopes below 300 K. Such a calculation renders a value of ~74 J mol^{-1} K^{-2}, which corresponds to a protein of ~11.5 kDa instead of the 4.7 kDa QNND-BBL (Figure 6.3D). In fact, the slope measured for QNND-BBL is even higher than the slope of Naf-BBL (~55 J mol^{-1} K^{-2}) indicating a possible equilibration artifact in the calorimeter. A two-state fit for the QNND-BBL thermogram enforcing $\Delta H_{Cal}/\Delta H_{vH} = 1$ and ignoring the temperature dependence of ΔH and ΔS within the transition region, is very good with ΔH_m = 129 kJ mol^{-1} and a T_m = 329 K (Figure 6.3C). However, it produces baselines that cross in the middle of the transition. This is clearly unphysical as it suggests a native state whose fluctuations are much higher than that of the unfolded state at higher temperatures, apart from the high pre-transition slope. It also results in unrealistically high ΔC_p at low temperatures that further changes sign in the middle of the transition. Indeed, a simple calculation of the DSC thermogram with those parameters, but now taking into account the ΔC_p obtained from the difference between the baselines predict a significant degree of

cold denaturation which is not experimentally observed. Fits by assuming a constant ΔC_p or fixing the folded baseline also result in baseline crossing.

6.3.3 Non-coincidental Unfolding Transitions by NMR

NMR is a powerful tool to study the chemical environment of individual atoms in proteins. Apart from providing structural information it could also be used to study the changes in the chemical shift as a function of temperature for multiple atoms thereby providing a direct evidence of the complexity of unfolding process. To obtain a high resolution picture or to ascertain the mechanism, the unfolding has to be followed for multiple atoms. However, Fersht and co-workers report on the ^{13}C chemical shift as a function of temperature of the Cα or Cβ of just six different residues in QNND-BBL. In spite of this drawback, individual two-state fits to these unfolding curves vary significantly with ΔH_m varying from 92 to 134 kJ mol^{-1} and T_m varying from 324 to 329 K[67]. Furthermore, the ^{13}C chemical shifts are independent of temperature and have very small baseline effects[132]. Errors in the determination of temperature are common to all of the points and cancel out in a direct comparison as the data is recorded simultaneously for all atoms from the same sample. Therefore, the apparent differences in these parameters immediately suggest a non-two-state behavior. But they interpret these differences as a byproduct of experimental noise propagating into uncertainty in the fitted parameters.

A simple way to estimate the experimental noise (*i.e.* the error in the determination of chemical shift values) is to fit each of the curves to a 6-parameter two-state model and estimate the resulting residuals. Such an analysis indicates that the error is between 1-2% of the total amplitude of the unfolding curve. If this error is

the source of parameter discrepancies, then the deviations between the normalized data and fit for one probe should be similar in magnitude to those of the other probes. In other words, the spread in the data points should span at least 5 K as this is the maximum difference in T_m reported for these six residues. Figure 6.3E shows the normalized chemical shift versus temperature for three of the six probes that have the maximum changes in ΔH_m and T_m as identified by the two-state analysis. It is apparent that the differences between the unfolding curves are much larger than the experimental noise. It can also be seen that each of the 12 points corresponding to the unfolding transition of Cα I149 is consistently ~6K lower than the equivalent point of Cβ L139. These differences can be quantified by performing the statistical error analysis discussed in the previous section. The general idea is to investigate the compatibility of the T_m and ΔH_m estimated from a two-state fit of one probe with that of the other. Thus, each of the unfolding curves is fit to a series of two-state models in which the baselines floated while ΔH_m and T_m are fixed to the values corresponding to each of the other probes (obtained from unconstrained fits). The SLS of the constrained fits is 5.5 ±1.9 times higher than that of the unconstrained fit compared to a difference of 0.2 expected if they arise out of random noise. Therefore, the probability that the probes analyzed by Fersht and co-workers share the same unfolding behavior is statistically negligible. In a follow up paper[68] they report the chemical shift changes of nine additional probes that again show a difference in individual ΔH_m varying between 38 and 250 kJ mol^{-1} with T_m between 323 and 329 K. They were able to explain all these discrepancies with a global two-state model, i.e. a model with a single ΔH_m = 134 kJ mol^{-1}, T_m = 324 K and floating baselines for

the 15 curves (60 + 2 parameters). Interestingly enough, they do not show the baselines for the global two-state fit. In fact, a closer look suggests that the difference in individual T_ms is efficiently suppressed by steep unphysical baselines in the global fit thus forcing the system to comply with a two-state model (data not shown). This observation is similar to the perfect two-state fit of the DSC thermogram of QNND-BBL with $\Delta H_{Cal}/\Delta H_{vH} = 1$, but with steep, crossing and unphysical baselines.

6.4 Ac-Naf-BBL-NH_2 and QNND-BBL have the Same Thermodynamic Properties

A quantitative analysis of the thermal unfolding of QNND-BBL described in the previous section indicates that the system does not fold in a two-state fashion as previously claimed. However, the question of the ~8 K higher stability of QNND-BBL compared to Naf-BBL (or Naf-BBL-Dan) remains. Fersht and co-workers claim that either the presence of fluorescent probes or the lack of QNND N-terminal tail perturbs the folding behavior of BBL. The second explanation is unlikely as the N-terminal tail is unstructured in the NMR structure. The dansyl label at the C-terminal can also be excluded as it is not present in Naf-BBL and its presence does not affect the folding of BBL when introduced in Naf-BBL-Dan. The only remaining possibility is the presence of N-terminal naphthyl-alanine which they claim to be at the edge of the hydrophobic core in their structure that might affect the folding. But even this interpretation is not possible as the quantum yield of the incorporated naphthyl is identical to that of free naphthyl-alanine[47]. There are then only two other possible explanations for the increased stability: lack of N- and C-terminal protection in Naf-BBL or the lower ionic strength of the buffers used in experiments on Naf-BBL. The

lack of protection by acetylation and amidation at the N- and C-terminus, respectively, of Naf-BBL possibly induces a repulsive interaction with the end charges and the macrodipole of the helix. This repulsion is particularly strong at the N-terminus and decreases with increasing sequence separation between the charge and start of the helix[133]. In Naf-BBL, the N-terminal charge is just two residues from the N-cap of helix 1, while in QNND-BBL it is separated by 6 residues. Experiments were therefore carried out on ends-protected version of Naf-BBL (Ac-Naf-BBL-NH_2) to test this interpretation and whether its presence affects the folding behavior.

6.4.1 Far-UV CD

Figure 6.4A compares the CD signal of Naf-BBL and Ac-Naf-BBL-NH_2 at 222 nm in molar ellipticity units. Both the variants have similar pre-transition, transition and post-transition slopes. Also, the Naf-BBL data overlays perfectly on Ac-Naf-BBL-NH_2 when shifted to the right by 9 K (not shown). Interestingly, the ends protected version has much higher signal (*i.e.* more negative) at the lowest temperature and in the high temperature post-transition region compared to Naf-BBL. The increase in signal at the lower temperature is of particular importance as it suggests that the helical content of Ac-Naf-BBL-NH_2 is higher than that of Naf-BBL. This effect is surprising considering that Ac-Naf-BBL-NH_2 has two additional disordered peptide bonds. The implication is that the folded ensemble of BBL is highly malleable and increases its nativeness in response to an increased energy gradient. In this case, an increase in nativeness is a result of decreased repulsion between the end-charges and the macrodipole of helix 1.

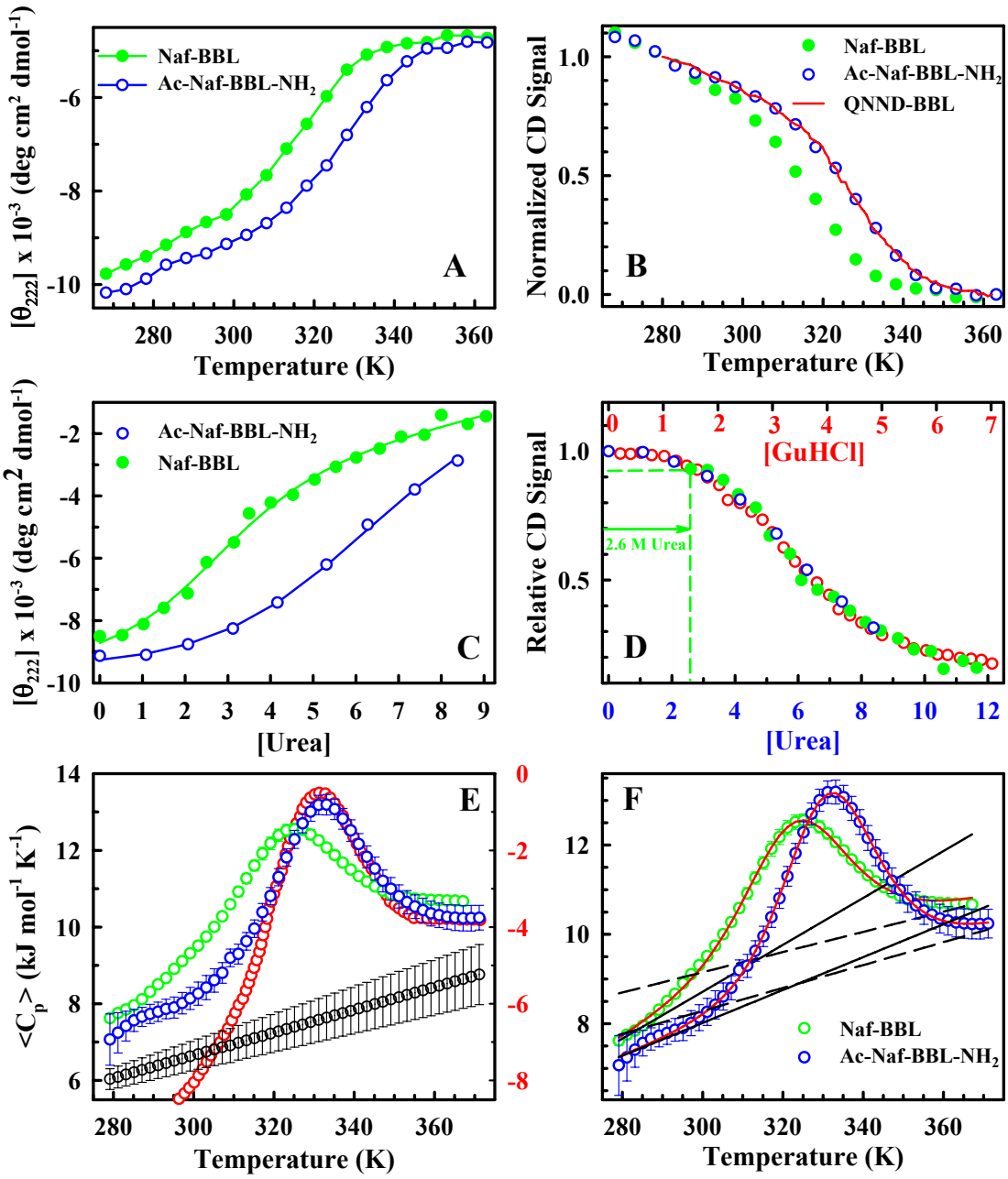


Figure 6.4 The data of Naf-BBL, Ac-Naf-BBL-NH_2 and QNND-BBL are shown in green, blue and red, respectively. A) Thermal unfolding transitions monitored by far-UV CD at 222 nm. B) Normalized CD unfolding transitions at 222 nm for comparison with QNND-BBL. C) Urea denaturation at 298 K. The continuous curves are fit to a two-state model. D) Super-imposition of the urea-induced unfolding of Naf-BBL and Ac-Naf-BBL-NH2 and GuHCl-induced unfolding data of QNND-BBL. The data for Naf-BBL and Ac-Naf-BBL-NH_2 is shown relative to the Ac-Naf-BBL-NH_2 signal at 0M. E) DSC thermograms of the three variants together with the native baseline predicted by Freire (black circles). F) Two-state sits (red curves) to the DSC thermograms showing the crossing of folded (continuous black line) and unfolded (dashed black line) baselines. (far-UV CD experiments by Naganathan AN; Naf-BBL

and Ac-Naf-BBL-NH_2 DSC experiments by Perez-Jimenez R & Sanchez-Ruiz JM; QNND-BBL experiments from Fersht group's published data).

The data for QNND-BBL was not reported in absolute units, thus requiring normalization. Also, the experiments on QNND-BBL were carried out at a ionic strength of 200 mM compared to the experiments on Naf-BBL and Ac-Naf-BBL-NH_2 that were performed at 43 mM ionic strength (i.e. 20 mM sodium phosphate buffer). However, the normalized CD signal at 222 nm for QNND-BBL (red curve) superimposes perfectly on the Ac-Naf-BBL-NH_2 curve (Figure 6.4B). The Naf-BBL data is also shown for comparison. This leads to the interesting conclusion that the degree of stabilization induced by the protection of ends in Naf-BBL is equivalent to the addition of the N-terminal tail of QNND together with an excess ~160 mM ionic strength. This together with the fact that the tail QNND is unstructured indicates that the extra stability of QNND-BBL reported by Fersht and co-workers is primarily the result of the higher ionic strength used in their experiments. The tail might induce a small difference in stability due to end-effects but is bound to be small. It is rather fortuitous that the protection of ends and an ionic strength of 43 mM match the BBL variant and the experimental conditions employed by them. It is also clear from the similar pre-transition, transition and post-transition slopes that the three proteins share similar thermodynamic properties. A two-state analysis with floating baselines produces ΔH_m ~115 kJ mol^{-1} for Ac-Naf-BBL-NH_2 and QNND-BBL and ΔH_m ~96 kJ mol^{-1} for Naf-BBL. These numbers are very similar to that expected for proteins of size 40-44 residues based on the scaling of thermodynamic parameters with size for much larger proteins. As ΔC_p is found to scale by ~58 J mol^{-1} K^{-1} per residue and ΔH at 333 K by ~2.92 kJ mol^{-1} per residue, this analysis predicts ΔH = 115 kJ mol^{-1} at

330 K (the apparent T_m of Ac-Naf-BBL-NH_2 and QNND-BBL) and ΔH = 93 kJ mol^{-1} at 321 K (the apparent T_m of Naf-BBL). Once more, a simple calculation invalidates the assertion by Fersht and co-workers that the thermodynamic properties of Naf-BBL are inconsistent with other similar sized proteins.

6.4.2 Chemical Denaturation

Figure 6.4C shows the urea unfolding curves of Naf-BBL and Ac-Naf-BBL-NH_2 at 298 K followed by CD at 222 nm. The sensitivity to urea induced unfolding is very low for these proteins as evidenced by the broad transitions. Naf-BBL shows little or no pre-transition. This is probably because of the fact that at 298 K it is significantly unfolded (compare the CD signal at 298 K with that at 268 K). But it shows a significant post-transition baseline. On the other hand, Ac-Naf-BBL-NH_2 shows a pre-transition but the transition is not complete within the experimentally accessible range. Though they seem quite disparate the transitions overlay on each other when the Naf-BBL data is shifted by 2.6 M urea (Figure 6.4D). This exercise also provides the much needed native baseline for Naf-BBL and the unfolded baseline for Ac-Naf-BBL-NH_2. The relative CD signal is obtained by using the molar ellipticity of Ac-Naf-BBL-NH_2 at 0 M urea as the reference. The response to urea-induced unfolding can be quantified by performing a simple two-state fit of chemical denaturation using the linear energy model (Chapter 2). This model necessitates the use of 6 parameters: ΔG_{H20}, m_{eq} and two parameters each for the native and unfolded baseline. Fitting the composite curve (green and blue circles of Figure 6.4D) to this model produces a m_{eq} value of 1.7 kJ mol^{-1} M^{-1} and a $[Urea]_{1/2}$ of ~5.6 M. Two-state fits to the individual curves by fixing the native baseline of Naf-BBL and the

unfolded baseline of Ac-Naf-BBL-NH_2 results in the same m_{eq} value of 1.7 kJ mol^{-1} M^{-1}, and a $[Urea]_{1/2}$ of ~3 and ~5.6 M, respectively. The baselines are constrained for the individual fits as there is not much information for them, as discussed above. Also the GuHCl induced denaturation of QNND-BBL is identical to the composite urea denaturation curve when the scales are corrected for the corresponding sensitivities. The resulting ratio of the higher sensitivity of GuHCl to urea is ~1.75 consistent with that observed for much larger proteins[77]. A similar ratio is obtained from independent two-state fits, which render ~2.9 kJ mol^{-1} M^{-1} for the GuHCl unfolding curve of QNND-BBL versus the ~1.7 kJ mol^{-1} M^{-1} for the composite curve.

6.4.3 DSC

Figure 6.4E compares the DSC thermograms of Naf-BBL and Ac-Naf-BBL-NH_2 in absolute heat capacity units along with the associated standard errors. The native baseline is shown in black and is calculated from the Freire's equation. DSC of Naf-BBL displays all the characteristics as previously discussed Chapter 4. Particularly, the thermogram is broad with no apparent pre-transition region. The transition therefore spans more ~70 K with a maximum at ~324 K. The lowest temperature point is more than 1.5 kJ mol^{-1} K^{-1} higher than that of the native baseline suggesting significant enthalpy fluctuations even in the 'native' state. The ends protected version is sharper with a maximum at ~332 K showing a hint of the true baseline at low temperatures. However, the width of the transition is comparable to that of Naf-BBL. The sharpness is merely a result of the higher T_m for this variant that produces a higher enthalpy of unfolding. Interestingly, the pre- and post-transition regions of Ac-Naf-BBL-NH_2 have a lower absolute heat capacity compared to Naf-

BBL consistent with an increased nativeness predicted by the CD analysis. In spite of the increased structure or lesser enthalpy fluctuations, the lowest temperature point of Ac-Naf-BBL-NH_2 is still ~1 kJ mol^{-1} K^{-1} higher than the native baseline. These two thermograms can still be fit to a two-state model enforcing $\Delta H_{Cal}/\Delta H_{vH} = 1$ (Figure 6.4F). However, the baselines cross in the middle of the transition with the fitted native baseline showing steep temperature dependence. The parallel two-state baselines and the matching ΔH_m values (Naf-BBL, ~100 kJ mol^{-1} at 322 K; Ac-Naf-BBL-NH_2, ~125 kJ mol^{-1} at 331 K) emphasize that these proteins have slightly different stability but the same overall thermodynamic behavior.

The thermogram of QNND-BBL is not reported in absolute units and is therefore shown in Figure 6.4E with a separate scale on the right. The scale was adjusted to match its post-transition baseline with that of Ac-Naf-BBL-NH_2. The sharpness, width and temperature maximum of QNND-BBL thermogram is almost identical to that of Ac-Naf-BBL-NH_2 suggesting that these two proteins have similar thermodynamic properties in agreement with the conclusions from CD analysis. But it shows a marked deviation in low temperatures with a steep pre-transition that not only has a higher dependence than either of the proteins but also crosses the Freire's baseline. This is probably a result of equilibration artifacts in the calorimeter that was surprisingly ignored by Fersht and co-workers.

6.5 Tuning the Stability with Ionic Strength

The agreement between the CD curves of QNND-BBL at 200 mM ionic strength and Ac-Naf-BBL-NH_2 at 43 mM ionic strength suggests that the stabilities of these systems could be tuned by simply changing the buffer composition. In other

words, the stability of QNND-BBL should be similar to Naf-BBL when compared under identical conditions. But there is no available data for QNND-BBL at 43 mM ionic strength. This problem can be easily overcome by repeating the experiments on Naf-BBL at 200 mM ionic strength and checking for agreement between the corresponding thermodynamic parameters. Figures 6.5A and 6.5B plot the CD signal

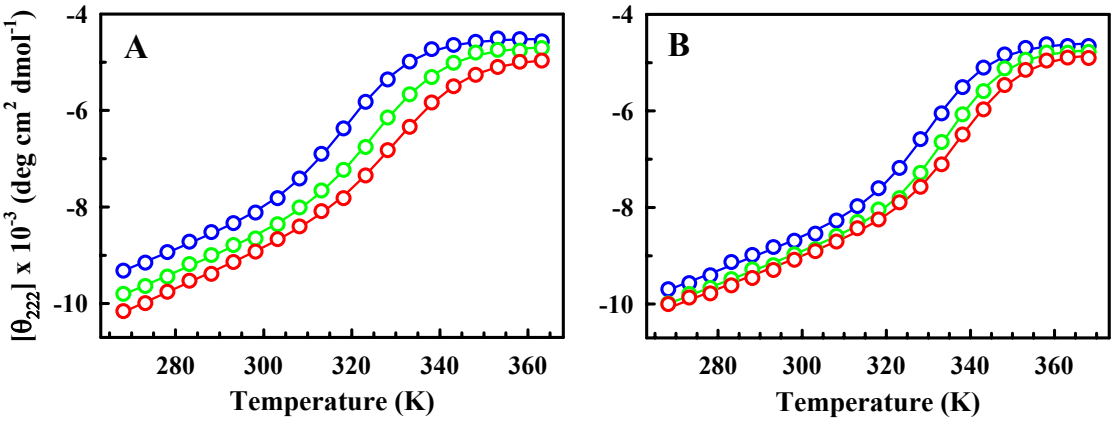


Figure 6.5 Thermal unfolding transitions of Naf-BBL (A) and Ac-Naf-BBL-NH_2 (B) at 43 (blue), 200 (green), and 400 (red) mM ionic strengths, respectively, as monitored by far-UV CD at 222 nm. (Experiments by Naganathan AN).

at 222 nm for Naf-BBL and Ac-Naf-BBL-NH_2 as a function of temperature at ionic strength values of 43, 200 and 400 mM, respectively. The signals are reported molar ellipticity units and hence directly comparable. Naf-BBL shows a significant increase in stability as a function of ionic strength with two state fits producing apparent T_m values of ~322, 329 and 335 K at the different ionic strength values, respectively. The signals have similar pre-transition, transition and post-transition slopes. As noted in the comparison between Naf-BBL and Ac-Naf-BBL-NH_2, the nativeness increases progressively both in the folded and unfolded ensemble. Ac-Naf-BBL-NH_2 shows a similar sensitivity to salt though the increase in nativeness is not as pronounced as in Naf-BBL. This is not surprising as the ends protected variant has already been

stabilized by ~9 K compared to Naf-BBL under similar conditions. However, the stability does increase with salt producing apparent T_ms of ~332, 337 and 341, respectively, when analyzed by a two-state model. In fact, the data for Naf-BBL at 400 mM ionic strength overlays well on Ac-Naf-BBL-NH_2 data at 43 mM with small differences in the pre- and post-transition regions (data not shown). This also indicates that the difference in stability induced by the addition of the QNND tail in Naf-BBL is not more than 2 K. To summarize, the low stability conditions employed in the original experiments of Naf-BBL do not affect the degree of 'cooperativity' of folding. The effects of salt and ends-protection also provide ample evidence to the conformational plasticity of the BBL's ensemble. In a simple two-state analysis any increase in nativeness as a function of salt would be simply incorporated in the baselines, with only the fraction native shown as a function of temperature (see below), *i.e.* such an analysis would make the curves look two-state like. This observation further highlights the importance of reporting the spectroscopic signal in absolute units.

6.5.1 Physical Meaning of far-UV CD Baselines

Most of the unfolding curves presented above have been analyzed by simple two-state models with free floating baselines. But, how much does a baseline affect the resulting thermodynamic parameters? And, is it apt to employ a two-state treatment for signals with high pre-transition slopes? This sub-section deals with these questions.

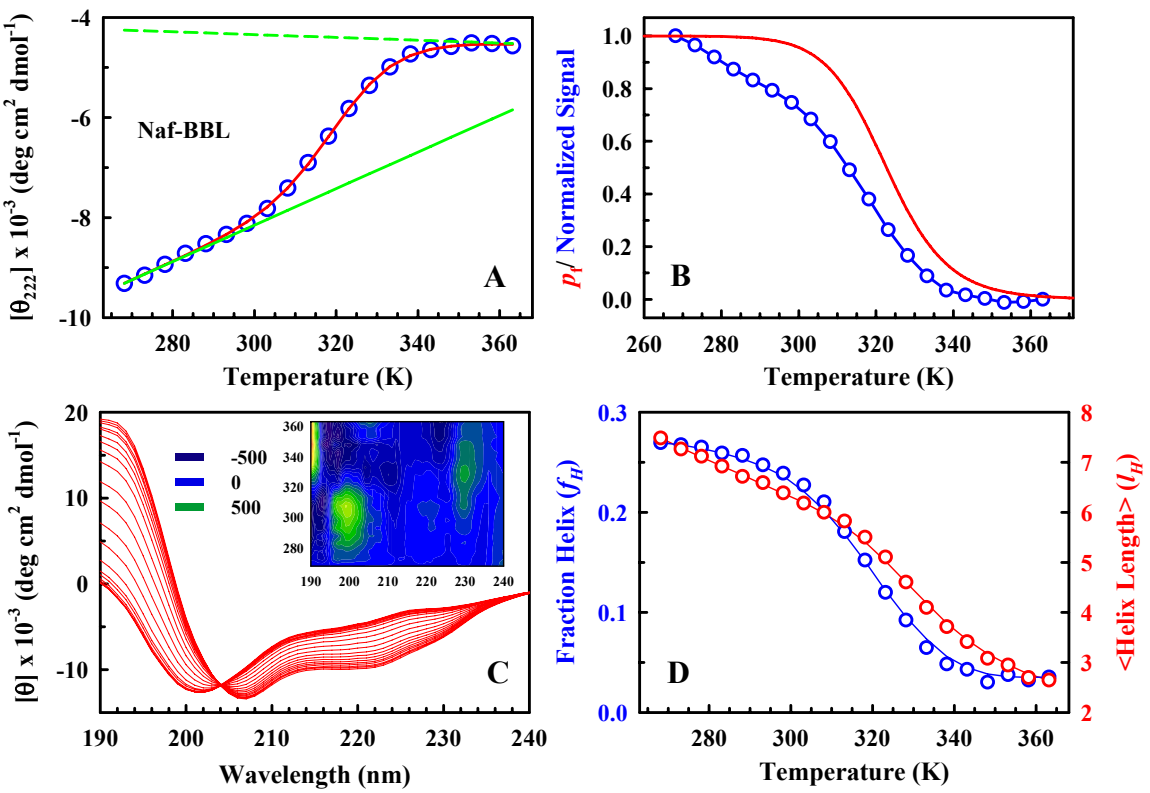


Figure 6.6 A) A two-state analysis (red) of Naf-BBL far-UV CD signal at 222 nm (blue), showing the folded (continuous green line) and unfolded (dashed green line) baselines. B) Plot of the normalized signal (blue) together with the probability derived from a two-state analysis (red). C) Calculated CD spectra of Naf-BBL from Chen's model with the inset showing a contour map of the difference between the data and fit. D) The predicted fraction helicity (blue) and helix lengths (red) as a function of temperature for Naf-BBL. (Experiments by Naganathan AN).

The baselines in a two-state DSC analysis correspond to the enthalpy fluctuations of the respective ensembles. In case of CD, they represent the intrinsic temperature dependencies of the signals. This intrinsic dependence is rather small for a CD signal and proteins with almost zero slopes in their native baselines have been reported in literature[53], suggestive of little structural change in the folded state as a function of temperature. In fact, such unfolding curves are a feature of two-state-like proteins. Figure 6.6A shows the data of Naf-BBL at 43 mM ionic strength (blue circles) along with the 6-parameter fit to a two-state model (red curve). The fit is very

good but produces a steep native baseline. The high slope indicates that at 363 K (the final temperature point) the 'folded' state is only ~63 % structured compared to 268 K. Alternately, the signal of the folded state at 363 K is only 23 % more than that of the unfolded state while at 268 K it more than double that number. In other words, a steep native baseline provides strong evidence that the structure of Naf-BBL unfolds continuously with temperature. This is entirely against the spirit of a two-state model though it has been used to arrive at these conclusions!! This apparent paradox highlights the effects of baselines in forcing the system to comply with a two-state criterion. This is all the more evident when the fraction folded from a two-state fit (p_f) is compared to that of the corresponding normalized signal (Figure 6.6B). They look as disparate as unfolding curves from two different proteins with no overlap throughout the transition, though the probability (red curve) has been estimated from the blue curve. In fact, a model-free first derivative analysis of the normalized signal produces an apparent T_m of ~318 K while the two-state fit results in ~322 K. This calculation immediately suggests a possible reason as to why Fersht and co-workers could not identify any differences in the T_m between different experimental probes. Moreover, most of the thermal/chemical unfolding data in literature is reported only in terms of fraction folded, thereby making the curves look more two-state like (compare blue and red curves).

The tendency of baselines to skew the thermodynamic parameters highlights the need to interpret the data structurally. A two-state fit provides no information on the helix length (l_H) or the fraction of residues that are in a helical conformation (f_H). However, a CD spectrum provides much more information when analyzed

appropriately. In particular, the rotational strength of each of the far-UV transitions, n-π^*, π-π^* parallel and π-π^* perpendicular components centered at around 222, 208 and 193 nm, respectively, is sensitive to the length of a-helix as discussed before. Recognizing this, Chen *et al.* developed an empirical equation to account for the changes in helical band shape and intensity as a function of helix length. It can be represented as[134]:

$$\theta(\lambda, l_H) = \theta(\lambda, \infty)\left[1 - \frac{k(\lambda)}{l_H}\right] \tag{6.1}$$

where $\theta(\lambda, l_H)$ and $\theta(\lambda, \infty)$ are the mean residue ellipticities at wavelength λ of an α-helix of l_H residues and infinite length, respectively. The wavelength dependent parameter $k(\lambda)$ accounts for the end-effects as the last four residue of an α-helix are not hydrogen-bonded. Typically, $1 < k(\lambda) < 4$, with an average of ~2.5. Therefore, it is possible to characterize the temperature dependent CD spectra of any protein using the relation:

$$[\theta](\lambda, T) = f_H \cdot \theta(\lambda, \infty) + (1 - f_H) \cdot \theta^{coil}(\lambda) \tag{6.2}$$

where $\theta^{Coil}(\lambda)$ is the coil basis and f_H is temperature dependent. In analyzing the BBL data with equations 6.1 and 6.2, the helical infinite length basis $\theta(\lambda, \infty)$ is obtained from Chen *et al.* while the spectrum at pH 3.0 and the highest temperature (363 K) is used as the unfolded basis. The values of $k(\lambda)$ is estimated by fitting the lowest temperature spectrum to fine tune the structural details.

The result of such a fit to the wavelength-temperature data of BBL is shown in Figure 6.6C with the inset representing the contour plot of the difference spectra. The contour map shows that the spectra are reproduced almost perfectly with a mean

absolute wavelength-temperature residual of ~200 deg cm^2 $dmol^{-1}$. The resulting parameters are plotted as a function of temperature in Figure 6.6D. It shows that the helical content of BBL decays sigmoidally with temperature while the average helix length decreases almost linearly. This complex helix length dependence with temperature is the basis for the observed differences in the melting curves at various wavelengths (Figure 6.3A). The same fitting procedure for a two-state-like protein produces a helix length that remains constant across the entire temperature range. Moreover, the predicted alpha helical content of 27 % and the average helix length of ~7.5 at the lowest temperature are consistent with the NMR structure. This analysis therefore provides a simple yet physical explanation for the steep pre-transition slopes observed by CD at 222 nm for Naf-BBL and variants. They indeed correspond to changes in helix length and are diagnostic of folding over marginal/zero barriers.

The spectra collected at various ionic strengths for the different BBL variants could in principle be analyzed with this model. However, the added salt (NaCl) absorbs significantly at lower wavelengths restricting the data collection to only 205 nm, thus eliminating valuable information from the 193 nm band. In spite of this drawback, the basic interpretation could still be extrapolated to the other unfolding curves as they essentially have the same pre-transition slopes and overall behavior. In Naf-BBL that shows the largest increase in nativeness with salt concentration, this interpretation would suggest that the average helix content increases to values higher than 27 % at the lowest temperature thus providing a quantitative picture for the observed changes in signal.

6.6 Ac-Naf-BBL-NH_2 Shows All the Thermodynamic Signatures of Global Downhill Folding

The analysis presented in the previous sections show beyond doubt that QNND-BBL and Ac-Naf-BBL-NH_2 have identical thermodynamic properties that are tunable by ionic strength. This section deals with the compliance of Ac-Naf-BBL-NH_2 data to the previously reported thermodynamic signatures of downhill folding.

Figure 6.7A plots the CD signal of Ac-Naf-BBL-NH_2 at the diagnostic wavelengths of 193, 200, 208 and 222 nm. As seen for the globally downhill Naf-BBL, the transitions have different apparent T_ms, pre-transition slopes and differ throughout the temperature range. The differences in T_m and pre-transitions are more readily observed by calculating the first derivative of the thermal unfolding curves (Figure 6.7B). It produces apparent T_m values ranging from 326 K (200 nm) to 332 K (193 and 208 nm). The steep pre-transition observed at 222 nm has the same interpretation provided in the previous section, i.e. the helices unfold progressively with temperature indicating a globally downhill folding transition.

Figure 6.7C plots the data from a double-perturbation experiment in which the CD signal is measured at 222 nm as function of temperature at various urea concentrations. This is a powerful technique to distinguish between downhill and two-state folding transitions and has been used earlier to highlight the complex thermodynamic behavior of Naf-BBL. Chemical denaturation at a specific temperature (Figure 6.4C for example) is an example of a single perturbation experiment which provides information on the first moments of the folding ensemble. These experiments produce sigmoidal unfolding curves irrespective of the nature of

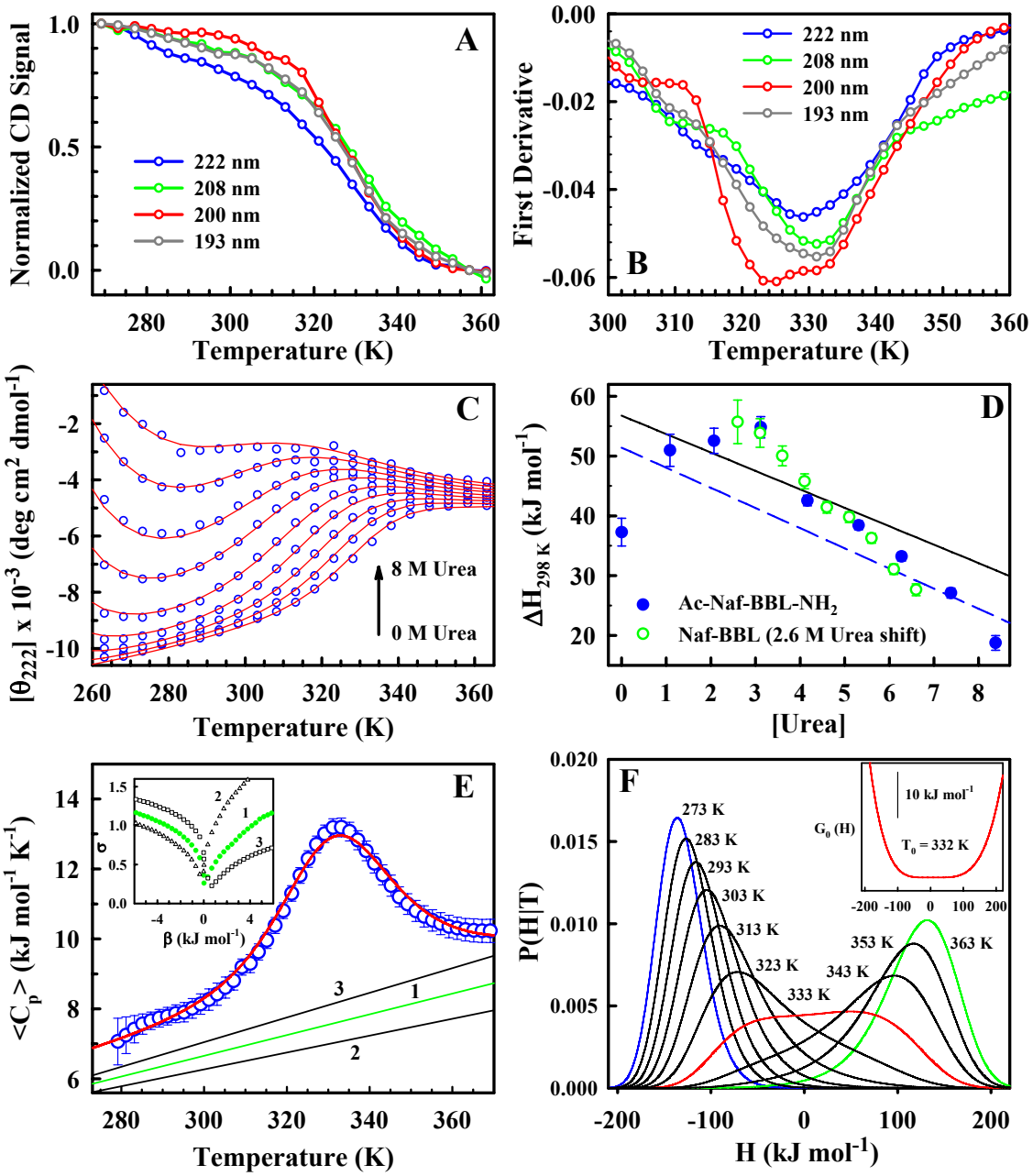


Figure 6.7 A) Normalized thermal unfolding transitions of Ac-Naf-BBL-NH_2 monitored by CD at different wavelengths. B) First derivative of the curves shown in panel A. C) Equilibrium unfolding of Ac-Naf-BBL-NH_2 induced by temperature and urea (0 – 8 M urea in steps of 1 M) monitored by CD at 222 nm. The continuous red lines correspond to a global two-state fit. D) Urea-dependence of the apparent ΔH for folding at 298 K obtained from individual two-state fits to the data shown in panel C; (continuous black line) linear dependence of ΔH_{298K} measured by Felitsky and Record in lac-repressor and (dashed blue line) linear regression of the composite data for Naf-BBL and Ac-Naf-BBL-NH_2. E) DSC of Ac-Naf-BBL-NH_2 data (blue circles) and fit (red curve) to the variable-barrier model using baseline 1 (green). (inset) Plots

of the standard deviation (in kJ mol^{-1} K^{-1}) versus the parameter β of the fits to the variable-barrier model using baselines 1 to 3. F) Probability density of Ac-Naf-BBL-NH_2 as a function of temperature calculated from the fit of panel E. (inset) Free-energy profile at the characteristic temperature (T_0 = 332 K). (far-UV CD experiments by Naganathan AN; DSC experiments by Perez-Jimenez R & Sanchez-Ruiz JM).

the transition, though the broadness of transition might vary. It is difficult to quantify the broadness when two-state models are used as it can significantly trim the information content using baselines. This difficulty can be overcome by performing a double perturbation with two denaturing agents with differing mechanism of action. Temperature and chemical denaturants are the most widely used denaturing agents whose mechanism of unfolding and the resulting effect on the folding properties are entirely different (Chapter 2). In a two-state system, though the different denaturants might elicit a different response the observed signal can always be represented as a linear combination of the folded and unfolded states. However, in a global downhill folding scenario where different structural ensembles populate at each value of native bias, the response will not be linear thereby providing a direct access to the underlying nature of transition[52].

The effect of the double-perturbation on Ac-Naf-BBL-NH_2 is similar to that of Naf-BBL. Specifically, the curves are sigmoidal with pronounced low temperature curvatures at higher denaturant concentration signaling the onset of cold denaturation (Figure 6.7C). The unfolded ensemble gets progressively less structured upon addition of urea. Of particular importance is the parameter T_{max}. This corresponds to the temperature at which the signal shows a maximum, i.e. the most negative in case of CD signal at 222 nm. In a two-state system this temperature varies very little as there is a unique folded ensemble population at any combination of denaturants.

However, in a downhill folding system since the ensemble gets progressively less structured the T_{max} shifts to the right with increasing chemical denaturant concentration. This effect is evident in Figure 6.7C where the T_{max} changes by ~20 K going from 2 to 8 M urea.

A phenomenological two-state fit to the curves is also shown in Figure 6.7C as red curves. The fit has been performed by assuming a linear dependence of ΔH and ΔS on urea ([D]) as has been empirically observed for a number of proteins [135,136], *i.e.*

$$\Delta H([D]) = \Delta H_m^0 + a[D] \text{ and } \Delta S([D]) = \Delta S_m^0 + b[D] \tag{6.3}$$

and

$$\Delta S_m^0 = \Delta H_m^0 / T_m \tag{6.4}$$

ΔH_m^0 and ΔS_m^0 correspond to the change in enthalpy and entropy at T_m in the absence of denaturant. These expressions together with equations 2.14, 2.15 and 2.17 from Chapter 2, provide a complete description of the thermodynamics as a function of both urea and temperature. The baselines as a function of temperature and urea are again unknown. Keeping the spirit of two-state models, they were estimated as

$$S_f = a_1 + a_2 T$$

and

$$S_u = a_3 + a_4 T + a_5 T^2 + a_6[D] + a_7[D] \cdot T$$

where the folded signal (S_f) has a linearly dependence on temperature (T) while the unfolded signal (S_u) has a quadratic temperature and linear denaturant dependence. Apart from this the unfolded signal has an additional parameter (a_7) that accounts for the coupling between the denaturants. Therefore, the fit requires a total of 12

parameters: ΔH_m, T_m, ΔC_p, a, b and the 7 parameters describing the baselines. It is important to keep in mind that this exercise is carried out to merely investigate the compliance of the double-perturbation data to a two-state model and to estimate ΔC_p. The heavy reliance on baselines to explain any signal change is illustrated by the need to employ 7 parameters to explain the signals which is 2 more than that required for the thermodynamics. The global fit is very good (Figure 6.7C). However, a closer look suggests that the model is unable to precisely reproduce the changes induced by urea in the ΔH_m and T_m, i.e. underpredicts the T_m at high and low urea and overpredicts it in the mid-range.

The systematic deviations between the data and global fit are small mainly because of the large number of parameters employed. This problem can be overcome by performing individual two state fits to the thermal unfolding curves at different urea concentrations by fixing the native baseline and ΔC_p to 50 J mol^{-1} K^{-1} per residue obtained from the global fit. Such a fit is highly constrained as it requires just 4 parameters: T_m, ΔH_m and the unfolded baseline. Figure 6.7D shows the apparent enthalpy change (ΔH) at 298 K versus urea obtained from such a fit. The plot for Ac-Naf-BBL-NH_2 is highly curved with a maximum between 2 and 3 M urea (blue circles). For a two-state system such a plot should be linear as the ΔH is defined by the difference in sensitivity between just two states (folded and unfolded). For a downhill folding system since the structural ensembles themselves are different at each urea concentration, the sensitivities are dissimilar thus producing a non-linear dependence. The plot for ΔS versus urea is similarly curved while the plot for ΔG is S-shaped (data not shown). The data for Naf-BBL overlays on Ac-Naf-BBL-NH_2

when shifted by 2.6 M urea (green circles) illustrating once again that these proteins have similar thermodynamic properties. It is also interesting to note that the magnitude of ΔH and its average sensitivity to urea for Ac-Naf-BBL-NH2 (dashed blue line) is similar to that reported for two-state like lac-repressor protein (continuous black line).

The thermogram of Ac-Naf-BBL-NH_2 was also analyzed by the variable barrier model[49] to estimate the barrier height. Figure 6.7E shows the result of such a fit using baseline 1 (green line) estimated by the Freire's relation. The fit is very good (red curve) and comparable to that of a two-state fit with residuals within experimental noise. Baselines 2 and 3 correspond to one standard error in determining Freire's baseline. The parameters of the best fit are $\beta = 0$ kJ mol^{-1}, $\Sigma\alpha = 58$ kJ mol^{-1} and T_0 =332 K, thus suggesting a barrierless unfolding transition at all values of native bias. The inset shows the residual plot obtained from the grid analysis for the three baselines. All of them produce zero barriers with negligible errors as seen from the sharpness of the curves. The probability distribution as a function of temperature is unimodal at all temperatures while having a maximal width at T_0 (Figure 6.7F). The inset shows the free functional at T_0 clearly showing the absence of a barrier. Therefore, CD, double-perturbation and DSC experiments provides a strong evidence to the global downhill folding behavior of Ac-Naf-BBL-NH_2.

6.7 Conclusions

The quantitative spectroscopic analysis of Ac-Naf-BBL-NH_2 and to a minor extent Naf-BBL, suggests that these proteins have the same overall behavior and identical to that of QNND-BBL. Furthermore, BBL folds globally downhill

irrespective of the presence of fluorescence probes or tail and the results are not a byproduct of aggregation. Also, higher ionic strength merely increases the stability without affecting the nature of transition as evidenced by the presence of steep pre-transition slopes in all variants under all conditions. All of these results suggest that downhill folding is a robust property of this protein. This investigation also reveals the reason for the erroneous classification of QNND-BBL as a two-state folder. The presence of sigmoidal unfolding curves and single-peaked thermograms that result in a $\Delta H_{Cal}/\Delta H_{vH}$ of 1 cannot be used a criterion for two-state folding. An unbiased analysis of the baselines from two-state fits is required to estimate the degree of compliance to a two-state model as little or not information is available on them for most experiments.

An increase in the degree of nativeness with increasing salt when monitored by CD and the ability to quantify the degree of fluctuations already present in the native state by DSC were possible only by reporting the data in absolute units. Such careful characterization of folding is necessary as it provides a number of hints to the plasticity of the ensemble which otherwise are incorporated into baselines in a traditional two-state analysis. This assume even more importance with the recent characterization of a number of fast folding proteins that have been shown to fold at near the folding speed limit (i.e. downhill folding) (see chapter 5). They have broad unfolding transitions when monitored with little baseline information. For these proteins, a two-state analysis is clearly inappropriate. Furthermore, the thermodynamic parameters from a pseudo-two-state analysis of the BBL variants match those of much larger proteins; particularly the scaling of ΔH_m with size and

temperature, the dependence of ΔH (298) on urea and the absolute and relative sensitivity to denaturants. This supports the idea that downhill and two-state folding does not originate as a result of drastically different thermodynamic parameters but are in fact the extremes of a spectrum of folding behaviors. Which regime dominates is dependent on a number of factors including size (Chapter 3), structure and experimental conditions.

7. Evolutionary Conservation of Downhill Protein Folding: 1. Experimental Characterization of PDD

7.1 Introduction

Having identified and quantitatively characterized a multitude of theoretical and experimental criteria for downhill folding, the more relevant question is whether such a folding behavior has any functional significance. The absence of a free energy barrier immediately suggests a broad underlying distribution of structural ensembles that could in principle be exploited for a variety of functions. In other words, global downhill folding proteins (or those with marginal barriers) could be thought of as a separate functional class similar to what has been described for allosteric and natively unfolded proteins. In the former, ligand binding results in a structural transition while in the latter the unfolded state heterogeneity is believed to help in identifying a binding partner and thus folding. In the case of downhill folding, a ligand/protein could bind to any one of the partially structured species that 'fits' best resulting in the equilibrium shifting towards that state. Thus evolutionary selection of downhill folding would at the same time enable a protein to bind a number of ligands satisfying different structural and orientation restraints. The equilibrium will be determined by the immediate cellular conditions such as pH, ionic strength, presence or absence of another competing ligand/protein etc. This could well be the case for proteins or domains involved in regulation (*e.g.* domains of transcription factors) that bind to multiple effectors. Such a folding scenario also partially removes the threat of proteases inside a cell giving them an added advantage over natively unfolded

proteins. An innate structural flexibility also enables proteins to find their binding site on substrates that would otherwise take much longer times when having a unique structure (DNA-binding domains for example). Thus, downhill folding behavior offers significant functional capabilities for small domains over their sturdier two-state-like counterparts. That downhill folders could act as 'molecular rheostats', had already been proposed in the experimental characterization of the one-state BBL[47].

The simplest and the most unequivocal way to investigate the downhill folding requirement of a specific function is to introduce mutations that result in significant folding barriers (> 3*RT*) in proteins that fold downhill. Functional analysis could then be carried out on the mutated protein. But such an approach is challenging as this would involve multiple perturbing mutations while at the same time maintaining the functional requirements like binding site charge and surface complementarity. This is additionally complicated by the fact that many of the proteins shown to fold downhill (see Chapter 5) are not more than 60 residues in length restricting the window for experimentation. An alternate, a much easier and non-invasive solution would be to identify a distant homolog of a downhill folding protein and characterize its folding behavior. If the homolog does fold downhill then there is a good chance that downhill folding has significant functional implications. This chapter together with chapter 8 will attempt to do precisely that.

Sections 7.2 introduces PDD, the homolog of BBL along with its functional role and previous experimental characterization. Section 7.3 provides an in-depth experimental characterization of the equilibrium folding behavior of PDD followed by kinetics of folding. Section 7.4 summarizes the folding behavior of this domain.

7.2 PDD

BBL, shown to fold globally downhill (Chapters 4 and 6), is a domain from the E2 subunit of 2-oxoglutarate dehydrogenase complex from *Escherichia coli*. One of its homologs whose structure has been solved is the 42-residue PDD domain (also called the E3BD in literature) from the E2 subunit of pyruvate dehydrogenase multi-enzyme complex from *Bacillus stearothermophilus*[137] (recently Fersht and co-workers solved the structure of another homolog from an extremophile[68]). Thus, these two proteins are as far apart from one another with respect to the organisms (one is a mesophile and the other a thermophile) in which they function and the enzyme subunits involved, though they perform similar functions (see below). BBL and PDD

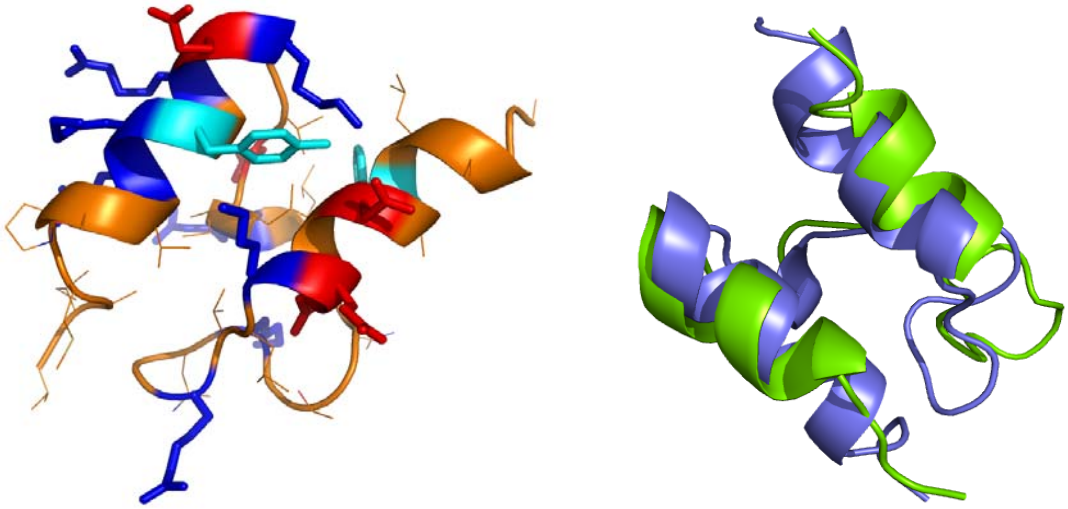


Figure 7.1 A) Structure of PDD – the acidic and basic residues are highlighted in red and blue, respectively, with the tyrosine and phenylalanine in cyan. B & C) Structural and sequence alignment of BBL (green) and PDD (violet).

are very similar in structure with two parallel alpha helices together with a long structured loop and share 30 % sequence identity (Figure 7.1).

7.2.1 Swinging Arm Mechanism

The function of these peripheral subunit binding domains (PSBD) as they are sometimes referred is discussed below with PDD as an example. The general features of the mechanism hold true for BBL as well.

The end-product of glycolysis - pyruvate - enters the Citric Acid cycle as acetyl-CoA. Oxidative decarboxylation of pyruvate is catalyzed by the pyruvate dehydrogenase multi-enzyme complex (PDH). This complex has multiple copies of 3 enzymes: pyruvate decarboxylase (E1), dihydrolipoyl transacetylase (E2) and dihydrolipoyl dehydrogenase (E3) and also requires the presence of 5 coenzymes: thiamine pyrophosphate (ThDP), lipoamide (Lip), coenzyme A (CoA), FAD and NAD^+. The PDH complex from *Bacillus stearothermophilus* is built around an icosahedral core (diameter of ~ 240 Å) containing 60 copies of E2[138]. The individual E2 subunits associate to form trimers at the 20 vertexes of the icosahedron. E2 has a multi-domain structure consisting of 3 independently folded domains: an N-terminal lipoyl domain (~80 residues) that binds lipoic acid, peripheral subunit binding domain (PSBD ~ 40 residues) that binds E3 and E1 and a C-terminal domain (~250 residues) which harbors the catalytic activity and forms the core of the enzyme. The domains are connected to one another by flexible linkers rich in alanine, proline and charged residues which are often called the ‘swinging arms’ for the reason explained below. To this central core bind multiple copies of E1 (60 $\alpha_2\beta_2$ tetramers) and E3 (6-10 dimers) giving a total molecular weight of about 9 MDa. The E1 and E3 molecules

are located on the outside of the core leaving a 90 Å gap between itself and the outer periphery of the core. This annular space is occupied by the linker regions and the PSBD of E2.

As per the recent models[138], the entire activity of the complex is dependent on the PSBD which acts as a hinge in moving the substrate between different active sites (substrate channeling) - moving the lipoyl domain towards E1, then to the transacetylase active site situated at the core of the complex and back to the E3 for the regeneration of oxidized lipoyl domain - with the length of the linkers enabling distances of the order of 100 Å to be covered with relative ease, and hence the name 'swinging arms'. Also, the position of the lipoyl domain at the N-terminus of E2 with the PSBD at its C-terminus requires it to bend not more than 180° to access the active sites.

7.2.2 Previous Studies

PDD has been characterized earlier as a two-state folder by Raleigh and co-workers[139-142]. This was based on the apparent coincidence of unfolding curves monitored by far UV CD at 222 nm, near UV CD at 280 nm and NMR chemical shifts of Tyr9 C^{δ} proton as a function of temperature. However the authors do observe significant differences in the pre-transition slopes of the above techniques. To re-phrase from reference 139 " The T_m value obtained from the curve fits is identical for all the three techniques, but ΔH is very sensitive to the way the pre-transition region is defined. The values of ΔH obtained from the various spectroscopic techniques range from 26 to 42 kcal mol^{-1}." This is precisely what is expected of a non-two-state folding transition. When the pre-transition slopes are steep, the free baselines

typically employed forces all the error in the ΔH value resulting in similar T_ms – the so-called differences in the apparent cooperativity. A two-state transition should result in identical enthalpies of unfolding and T_m values (within error) from various techniques. Clearly the ~50 % change in ΔH_m from one technique to another observed by Raleigh and co-workers does not represent such a situation. As with BBL, the steep pre-transition slope could correspond to structural changes rather than a 'baseline effect'.

From a functional standpoint, a two-state behavior raises more questions as this 40-residue domain then has only two accessible 'states' – a rigid folded state as a chemical two-state view entails and an unfolded state with no regular structure – to bind to two different partners, coordinate the movement of the swinging arms and regulate the activity of the entire complex. More intriguingly, the enzyme complex has it maximal activity at ~328 K (the growth temperature of *B. stearothermophilus*) that is close to the apparent T_m for this domain observed by Raleigh and co-workers. For the domain to act as a hinge, a highly stable structure is a pre-requisite and would have been evolutionary selected. The experimental observation is on the contrary indicating that only 50 % of the domains are completely folded on an average thus reducing the efficiency of the enzyme if the proposed folding and functional models are true. Apart from these inconsistencies, the domain's high structural and sequence similarity to BBL that is known to fold globally downhill indicates that the folding mechanism of PDD has to be revisited.

7.3 Experimental Characterization of PDD

The temperature-induced unfolding of PDD was studied in our laboratory by various equilibrium spectroscopic techniques like DSC, far-UV CD, near UV CD, FTIR (Fourier Transform Infrared Spectroscopy), FRET (Forster Resonance Energy Transfer) and fluorescence from extrinsic fluorophores. The kinetics of this domain was measured by Infrared laser temperature jump (IR T-jump) setup. Two PDD variants were synthesized – one in which the C-terminal is labeled with the fluorescence donor naphthyl alanine and the other with both the donor and an acceptor (dansyl lysine) at the N-terminus.

7.3.1 Differential Scanning Calorimetry (DSC)

DSC was carried out only on the donor labeled protein as the doubly labeled protein aggregates at the millimolar concentrations typically required for these experiments. Figure 7.2A plots the heat capacity profile of PDD at pH 7.0 and 20 mM sodium phosphate buffer (blue circles). As in BBL, the DSC profile is broad with an unfolding process that spans almost 70 K. It has a single peak at ~322 K and a steep pre-transition slope. The profile is also shifted to higher heat capacity values than expected from size-scaling arguments alone as represented by the Freire's slope (black circles in Figure 7.2A) suggestive of significant enthalpy fluctuations in the 'native state' of PDD. The pH 3.0 data is shown for comparison (green circles). It agrees well with the pos-transition region of the pH 7.0 data suggesting that the pH 3.0 data is a good approximation for the unfolded state of PDD at high temperatures.

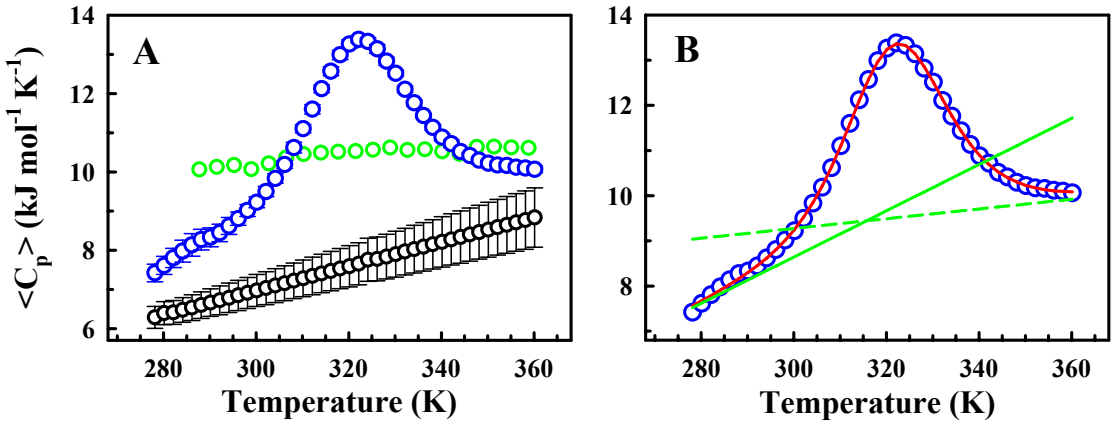


Figure 7.2 A) DSC thermograms of PDD at pH 7.0 (blue) and 3.0 (green) together with the Freire's baseline (black). B) Two-state fit (red curve) to the data in panel A highlighting the crossing of folded (continuous green line) and unfolded (dashed green line) baselines. (Experiments by Perez-Jimenez R & Sanchez-Ruiz JM).

As discussed in Chapters 5 and 6, the width of the DSC profile is directly related to the underlying probability density. Moreover, since DSC monitors the global unfolding process it is independent of any probe-specific details. Thus a rigorous test of the folding behavior can be performed by a two-state analysis of the DSC profile. A two-state fit with free baselines (red curve in Figure 7.2B) results in an apparent melting temperature (T_m) of 323 K and an enthalpy of unfolding at the T_m (ΔH_m) of ~113 kJ.mol^{-1}. Though the model fits the experimental data perfectly, it is clearly unphysical as the baselines cross close to the midpoint of the transition. This would mean a ΔC_p that changes its sign from positive to negative as the temperature is increased. Also, the slope of folded baseline obtained from the fit is ~1.6 times higher than what is expected from Freire's baseline - 51 J mol^{-1} K^{-2} compared to the expected 31 J mol^{-1} K^{-2}. These observations readily suggest a non-two-state transition in PDD as observed by DSC. Does this hold true for other spectroscopic probes?

7.3.2 Far-ultraviolet Circular Dichroism (far-UV CD)

In contrast to DSC, far UV-CD is sensitive to the peptide bond conformation and hence reports on the secondary structure of the protein. Figure 7.3A shows the spectra of PDD at various temperatures in the wavelength range of 190 to 250 nm, collected at pH 7.0 and 20 mM sodium phosphate buffer. The spectrum of a 100 % α-helix has two minima at 222 nm and 208 nm and a maximum at 193 nm with a magnitude of -40,000 deg cm^2 $dmol^{-1}$ at 222 nm[134]. Using this value as a reference, the fraction helicity is calculated to be ~22 % for PDD. This value is much smaller than that measured from the structure - ~43 %. An estimate from NMR structures is erroneous as it does not take helix fraying into consideration or any possible distribution of structures. Furthermore, Baldwin and co-workers have experimentally measured the effect of tyrosine on α-helical spectrum by employing a host-guest approach[143]. They have shown that tyrosine has a positive band at 222 nm and estimate the error in determining the fraction helicity to be ~13 % on the lower side. But in a protein, the magnitude of the signal obviously depends on the environment of the tyrosine residue and its degree of coupling to the peptide bond dipole. Since PDD has a tyrosine right at the center of helix 1 and protruding into the hydrophobic core, the calculated value of fraction helicity from the intensity alone is probably an under-estimate. The true value of helicity is thus somewhere between the two numbers. Also, the magnitude of the signal alone does not give any information on the helix length, and needs a more detailed analysis to arrive at the true numbers (see Chapter 8).

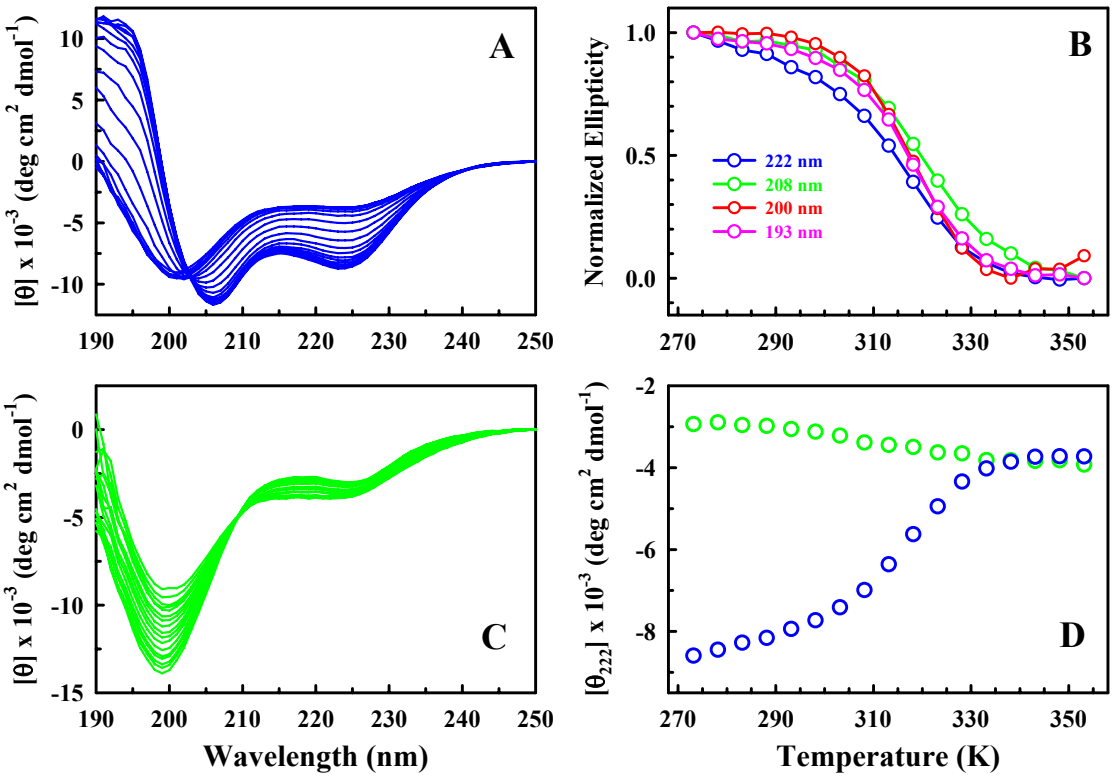


Figure 7.3 A & C) Temperature dependent far-UV spectra at pH 7.0 and 3.0, respectively. B) Wavelength dependent unfolding at pH 7.0. D) Mean residue ellipticity 222 nm at pH 7.0 (blue) and pH 3.0 (green), respectively, as a function of temperature. (Experiments by Naganathan AN).

The intensity of the pH 7.0 spectra decreases with temperature indicating a loss of helical content and a simultaneous population of the unfolded state. The loss in signal alone does not provide any information on the nature of the unfolding behavior, i.e., whether it is downhill or two-state-like. In fact, there are two possible ways of looking at this. One could think of a single structure (with specific helixlengths) whose population changes with temperature (two-state-like) or a scenario in which the helix length continuously decreases resulting in the loss of signal (downhill folding) or a situation between the two. At present, none of these can be ruled out. The signal at 222 nm at the highest temperature is ~ -4000 deg cm^2 $dmol^{-1}$ indicative of a significant amount of structure in the unfolded state. The

spectra also show an isodichroic point at ~ 203 nm. On numerous occasions in the literature, this has been shown as a sign of two-state behavior. Unfortunately, all an isodichroic point suggests is that only two structural species are (de)populated as the protein unfolds, namely, alpha helical and coil state, but provides no information on the nature of the underlying transition. The spectra were also measured at pH 3.0 to study the effect of temperature on the unfolded state (Figure 7.3B). It shows a pronounced minimum at 222 nm, indicating the presence of residual structure in agreement with results from pH 7.0 data. Figure 3C plots the raw signals at 222 nm at pH 7.0 (blue) and pH 3.0 (green). They overlay very well at high temperatures consistent with results from DSC.

An advantage of collecting spectra as a function of temperature as opposed to a single wavelength is that the temperature dependencies at multiple wavelengths can be directly compared to test the idea of two-state behavior. Figure 7.3D shows such a plot. The four wavelengths monitor specific aspects of the structure (see Chapter 6 for a detailed explanation). All of them show clear differences in the apparent melting temperatures and in pre- and post-transition slopes. But the differences seem less prominent when compared with BBL, possibly indicating the presence of a barrier. The signal at 222 nm which has been traditionally seen as in indicator of the 'cooperativity' of transition, has the most prominent pre-transition slope (Figure 7.3D). Fitting the temperature dependencies individually to a two-state model results in very similar T_ms clustered around 320 K and ΔH_m of ~120 kJ mol^{-1}. A first-derivative analysis for a model-free determination of the inflection points, results in a small spread of T_ms ranging from 318 to 320 K.

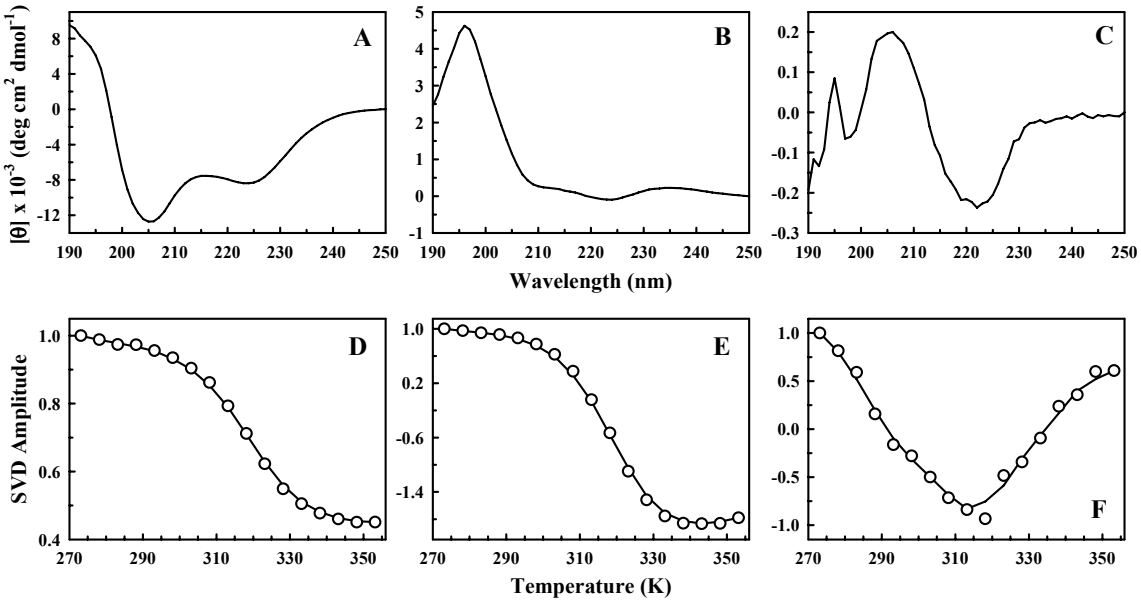


Figure 7.4 SVD components (A, B, & C) and their corresponding amplitudes (D, E, & F) extracted from pH 7.0 far-UV CD data. Lines in panels D, E, & F are shown to guide to eye.

A better description of the structural species involved in (un)folding process can be obtained by a Singular Value Decomposition (SVD) analysis of the temperature-wavelength spectra. The result of such an analysis is shown in Figure 7.4. The first component is the average spectrum and constitutes about 70 % of the total basis (Figure 7.4A). The corresponding amplitude is an indicator of the average decrease in intensity of this spectrum as a function of temperature (Figure 7.4D). The second component in a SVD usually accounts for changes in spectra. In this case, the second component is a mixed helix-coil spectrum accounting for ~27% of the total basis (Figure 7.4B). The shape of the spectrum, with the minima at 222 nm deeper than 208 nm suggests that short helices are involved in the transition[47]. The amplitude shows the depopulation of the helix with temperature and accumulation of the coil, as evidenced by the changes in sign (Figure 7.4E). This component gives an apparent T_m

of 318-319 K, ~4 K lower than that monitored by DSC. For a strict two-state transition, only two components are required to reproduce the original spectra and the rest should be contributions from noise or any lamp-specific/wavelength-dependent effects. But in the case of PDD as with many other helical proteins studied in out laboratory (unpublished results), there is a third component with opposing signs at 222 and 208 nm (~1% of the total basis; Figure 7.4C). The amplitude has a peculiar behavior - increasing with temperature, reaching a maximum at ~315 K and decreasing again (Figure 7.4D). The possible origins and its connection to the mechanism of unfolding are discussed later in this chapter.

7.3.3 Near-ultraviolet Circular Dichroism (near-UV CD)

Aromatic residues in asymmetric environments give rise to signals in the near-ultraviolet region. PDD has a tyrosine at position 9 and a phenylalanine at position 37, both protruding into the core of the protein, thus providing specific probes to monitor the immediate vicinity of either of these residues or the core region. Figure 7.5A shows the near UV spectra collected at various temperatures at pH 7.0 and 20 mM sodium phosphate buffer. The ellipticity is positive, with a peak at around 280 nm and dominated by tyrosine. Upon increase in temperature, the signal decreases reaching steady-state value of around 3000 deg cm^2 $dmol^{-1}$, hinting at the presence of residual unfolded structure in compliance with the results from far-UV CD.

The ellipticity monitored at 280 nm as a function of temperature shows a distinct pre-transition followed by a major transition (Figure 7.5B; blue circles). The

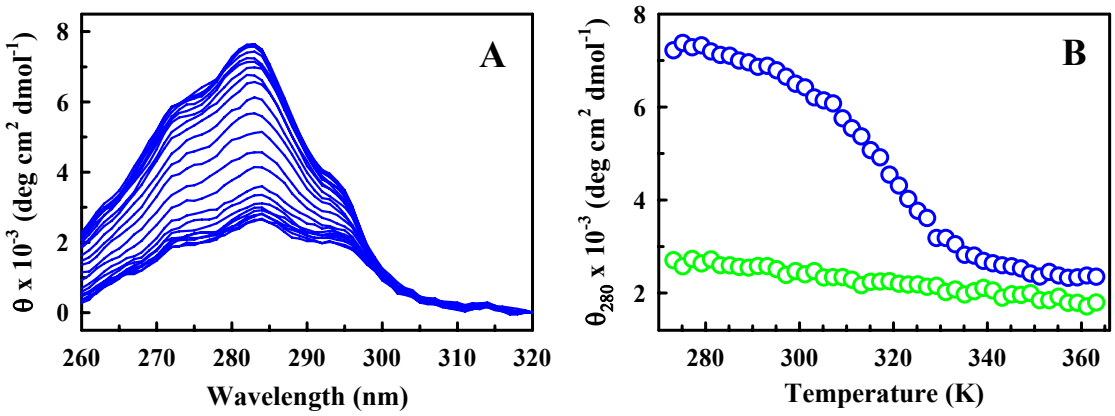


Figure 7.5 A) Temperature dependent near-UV spectra at pH 7.0. B) Comparison of the signals at 280 nm from pH 7.0 (blue) and pH 3.0 (green) data. (Experiments by Naganathan AN).

latter phase has an apparent T_m of about 322 K by first-derivative analysis. A two-state analysis results in a T_m of ~320 K and a ΔH_m of 109 kJ mol^{-1}. The difference in T_ms resulting between the two techniques is an indicator of baselines skewing the obtained results. The signal at pH 3.0 and 280 nm is shown for comparison (green circles in Figure 7.5B). SVD analysis of the temperature dependent pH 7.0 spectra reveals 3 components. The first component (~ 88 % of the total) is the average spectrum (Figure 7.6A) with temperature dependent amplitude very similar to the raw signal at 280 nm (Figure 7.6E). This is not surprising as the first component accounts for 88 % of the total change. The second component probably corresponds to a red shift of the tyrosine spectrum upon unfolding revealed by the opposing signs of the peaks at 295 and 275 nm, accounting for 6.2 % of the total basis (Figure 7.6B). Interestingly, the amplitude decays continuously with temperature with no evident pre-transition slope (Figure 7.6E). Inspection of the amplitude reveals that the second component has an apparent transition midpoint lower by at least 5 K with respect to

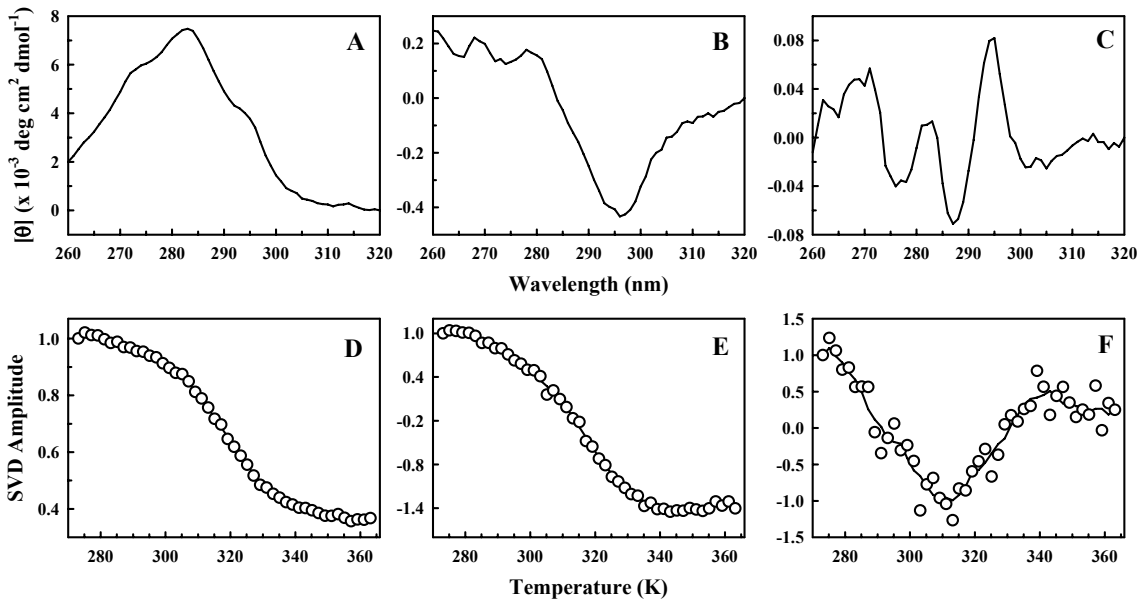


Figure 7.6 SVD components (A, B, & C) and their corresponding amplitudes (D, E, & F) extracted from pH 7.0 near UV-CD data. Lines in panels D, E & F are shown to guide to eye.

the first. In fact, a gradient analysis of the second component results in a T_m of around 315-316 K. But, a two-state analysis results in an apparent T_m of 321 K and a ΔH_m of 95 kJ mol^{-1}, further highlighting the effect baselines. The amplitude of the third component has a behavior similar to that from far-UV CD with a maximum at around 312-315 K, contributing ~ 0.6 % to the total change (Figure 7.6C and 7.6F).

7.3.4 Fourier Transform Infrared Spectroscopy (FTIR)

The pD 7.0 FTIR absorbance spectra of PDD in the amide I' region is shown in Figure 7.7A. Though the bands are intense, the change in intensity as a function of temperature is relatively weak. The signal variation is more significant in the range of 1600-1660 cm^{-1} while the maximum at ~1672 cm^{-1} shows little movement. A simple deconvolution of the lowest temperature spectrum at 278.3 K reveals 4 bands at 1580, 1632, 1652, and 1672 cm^{-1}, respectively (Figure 7.7B). The band at 1580 signals the

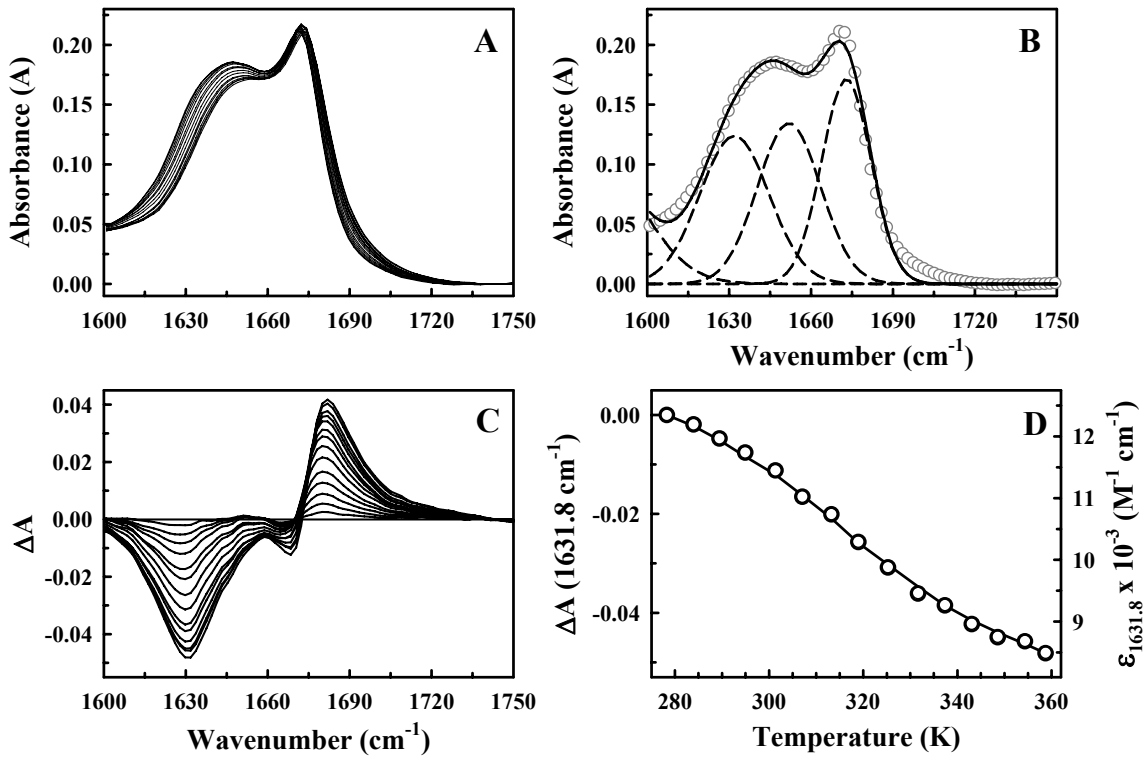


Figure 7.7 A & C) FTIR raw and difference spectra in the amide I' region, respectively, at pD 7.0. B) Deconvolution of the lowest temperature spectrum highlighting the various structural bands. D) The signal at 1631.8 cm^{-1} as a function of temperature. (Experiments by Naganathan AN & Li P)

asymmetric carboxylate stretching modes of aspartate side-chains (not shown). The 1632 cm^{-1} band is dominated by H-bonded carbonyl stretching modes in α-helix with minor contributions from C-N stretch and N-H bend. This band overlaps significantly with the peak at 1652 cm^{-1} that corresponds to the unfolded populations and helical carbonyls that are not H-bonded. The 1672 cm^{-1} peak is the result of the carboxylate stretching modes of TFA. The deconvolution is shown as a mere illustration. This was not attempted at all temperatures as the peak positions, widths and amplitudes of the respective bands change with temperature. This in turn requires a large number of adjustable parameters and the solution is not unique (data not shown).

Interestingly, the difference FTIR spectra ($\Delta A = A\text{-}A_{278.3}$) of PDD at pD 7.0 shows just two bands at 1630 and 1680 cm^{-1} (Figure 7.7C). The intensity of the band at 1630 cm^{-1} increases with temperature (in the raw spectrum it decreases), thus signaling the loss of helical content. But the peak corresponding to the accumulation of unfolded population is not evident as the bands of opposite signs at ~1680 and 1667 cm^{-1} originate from a movement of the TFA peak to higher wavenumbers coupled to a decrease in intensity[144]. This is because of a rather fortuitous compensation between the spectra of TFA and unfolded conformations as explained below. In a pure TFA difference spectrum the 1667 cm^{-1} peak is ~3 times more intense than its counterpart at 1680 cm^{-1}. However, in the difference spectra of PDD the magnitude of the intensity change is flipped. This clearly suggests that a positive peak in the range of 1650-1660 cm^{-1} compensates for this effect. This peak indeed corresponds to the unfolded conformations whose intensity in the difference spectra should increase with temperature.

The amide I' region of larger alpha-helical proteins (> 60 residues) typically reveal two bands[115] at 1632 and 1652 cm^{-1} with coil population evident beyond 1660 cm^{-1}. To identify the origin of these two bands, Vanderkooi and co-workers separately labeled (C^{13}) the buried and solvent-exposed peptide carbonyls of specific residues in the alpha-helical, dimeric GCN4 coiled-coil[145]. They observed a 20 cm^{-1} shift to lower wavenumbers in the peaks of the solvent-exposed carbonyls. Based on this evidence, the presence of two peaks at 1632 and 1652 cm^{-1} is usually seen as an evidence for a structure with buried and solvent-exposed helices. In fact, a single band at 1632 cm^{-1} is also seen in a number of α-helical peptides whose side chains

are predominantly exposed to the solvent. These observations suggest that the helices of PDD are solvent exposed with little burial of the backbone carbonyls or the side chains. The implication is that the change in heat capacity due to solvation effects alone should be minimal for this protein.

The difference spectra signal at 1631.8 cm^{-1} decreases almost linearly with temperature (Figure 7.7D). No pre- and post-transition baselines are evident. Not surprisingly, this curve can still be fit to a two-state model with a T_m of 321 K and an apparent enthalpy of ~75 kJ mol^{-1}. Though the T_m agrees with estimates from two-state treatments of far- and near-UV CD, the ΔH_m is almost 50 kJ mol^{-1} lower than that from far-UV CD and DSC. Also, the baselines from the fits are steep suggesting a complex helix unfolding mechanism. A first-derivative analysis of the same resulted in a T_m of ~316-317 K significantly different from a two-state fit and that is much lower than all other spectroscopic techniques. This observation is a result of the sensitivity of the FTIR technique to local conformations, i.e. H-bonding that typically spans just a few residues in an all α-helical protein. This particular observation also highlights the non-two-state nature of the transition as in a two-state system the change in signal should be identical irrespective of the level of structural detail probed by the technique.

SVD analysis of temperature dependent raw spectra in the range of 1500 - 1750 cm^{-1} reveals 4 components. The first is the average signal accounting for ~92 % of the total basis (Figure 7.8A). The peak at 1580 cm^{-1} corresponds to the carboxylate stretching mode mentioned before with large intensity band centered at ~1500 is mixture of the tyrosine C-C stretching and O-D stretching vibrations from HDO. The

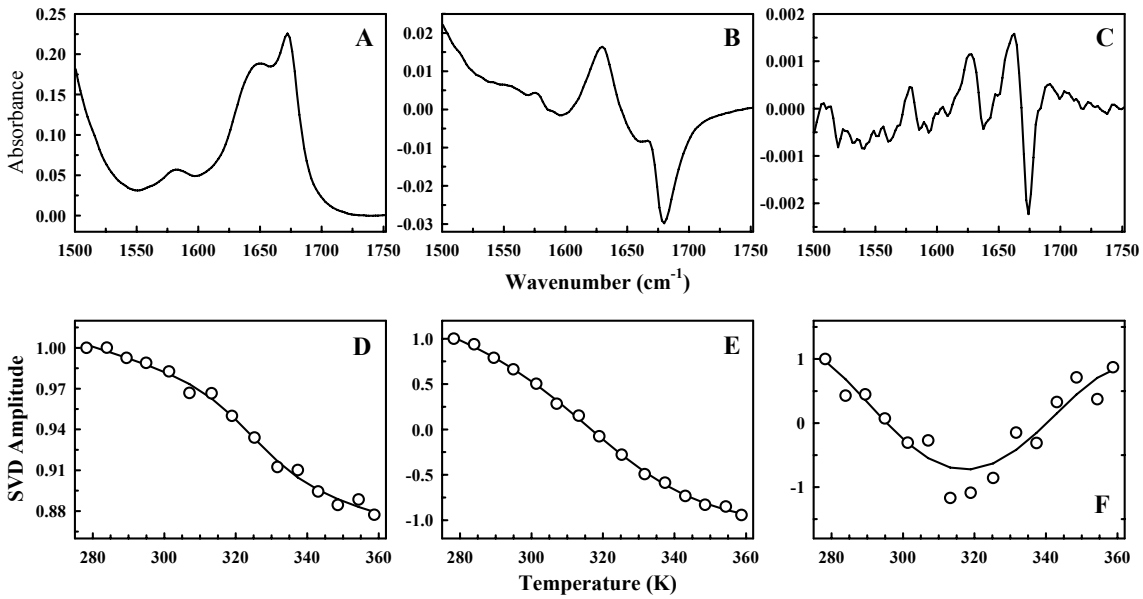


Figure 7.8 SVD components (A, B, & C) and their corresponding amplitudes (D, E, & F) extracted from pD 7.0 FTIR data. Lines in panels D, E & F are shown to guide to eye.

corresponding amplitude for this basis spectrum shows an apparent transition that is still not complete but that does show a minor pre-transition slope (Figure 7.8D). A two-state fit to this curve results in a T_m of 326 K and a ΔH_m of 109 kJ mol^{-1}. Though there is very little change in the overall amplitude of this component, the T_m agrees well the estimates from DSC. The second component accounts for ~ 7% of the total basis (Figure 7.8B). The basis spectrum shows peaks of opposing signs for the helical and TFA/unfolded modes of the amide I' band, thus indicating a change in signal involving those regions. The amplitude of the second component (Figure 7.8E) is similar to that of the signal dependence at 1632 cm^{-1} of the difference spectra (mathematically they are equivalent). The change in sign indicates that a loss of helical signal is accounted for an increase in the unfolded state population. But a two-state fit results in a T_m of 325 K and ΔH_m of ~60 kJ mol^{-1}. Both the enthalpies and the T_m are different compared to that obtained from FTIR difference spectra further

highlighting the effect of baselines. The third component has noisy temperature dependence and is not shown. On the other hand, the amplitude of the fourth component (~ 0.4 % of the total basis) has a behavior similar to that seen in 3rd components of far- and near-UV CD (Figure 7.8F). In the absence of spurious signals resulting from noise this would have constituted the third component. Intriguingly, the 4th basis spectrum (Figure 7.8C) clearly shows all the major amide I' and II' bands from Tyr (1510 cm^{-1}), COO^- stretch (1580 cm^{-1}), solvent-exposed helix (1630 cm^{-1}) and the TFA/disordered regions (1650 - 1680 cm^{-1}).

7.3.5 Possible Origins of the 'Third component'

SVD analysis of the spectral contributions to far-UV, near-UV CD and FTIR reveals that more than two basis spectra are necessary to describe the unfolding. The third component has similar temperature-dependent shapes but seemingly contribute very little to the overall process (between 0.4 to 1 %). This does not mean that the process is insignificant. All it means that the change in signal arising out of the specific process is small. This is because of the fundamental drawback of ensemble measurements that report on the average signal with little information on the possible distribution. However, a deeper insight can be gleaned by understanding the molecular origin of signals in each of these experiments. The tyrosine is located at the center of helix 1 and a change in its environment is directly linked to the melting of helices thus giving a near-UV signal; far-UV CD is sensitive to long-range dipolar coupling of peptide bonds while FTIR monitors the H-bonding of carbonyls and is sensitive to local structure. The 3rd basis from far-UV CD has peaks of opposite signs at 222 and 208 nm that are sensitive to the changes in the length of the helix (see

chapter 6). The third component of FTIR has bands corresponding to helix and tyrosine while that of the near-UV CD produces a signal resembling a red-shift of tyrosine. Taken together, it is clear that these three experiments monitor the mechanism of helix unfolding.

Keiderling and co-workers have observed similar temperature dependent amplitudes in alanine-based helical peptides by vibrational CD (VCD) and FTIR[146]. They qualitatively interpret it as the formation of helix-coil junctions whose numbers increase with temperature and then decrease. Extending this view to the observations in PDD, it would mean that the helices do not unfold in a two-state mechanism between a fully folded and fully unfolded structure but starts melting at various points thus resulting in a change in the number of helix-coil junctions. Such a melting is more probable from the ends of the helix, as far less interactions are broken while at the same time gaining a significant amount of conformational entropy. The amplitude could also be interpreted as arising from changes in helix length or differences in the alignment of peptide-bond dipoles with temperature. Mechanistically, all of the above interpretations point to considerable helix fraying and the presence of helices of different lengths. It would also explain the steep pre-transition slopes observed by far-UV CD at 222 nm, near-UV CD 2nd component, FTIR at 1632 cm^{-1} and DSC.

7.3.6 Double Perturbation Experiment

Double perturbation studies were performed on PDD as a function of temperature and urea/GuHCl. Figure 7.9A plots the results from urea. Upon successive addition of urea, the ensemble signal at the 298 K decreases almost continuously (Figure 7.9B) as observed in Naf-BBL and Ac-Naf-BBL-NH_2 (Chapter

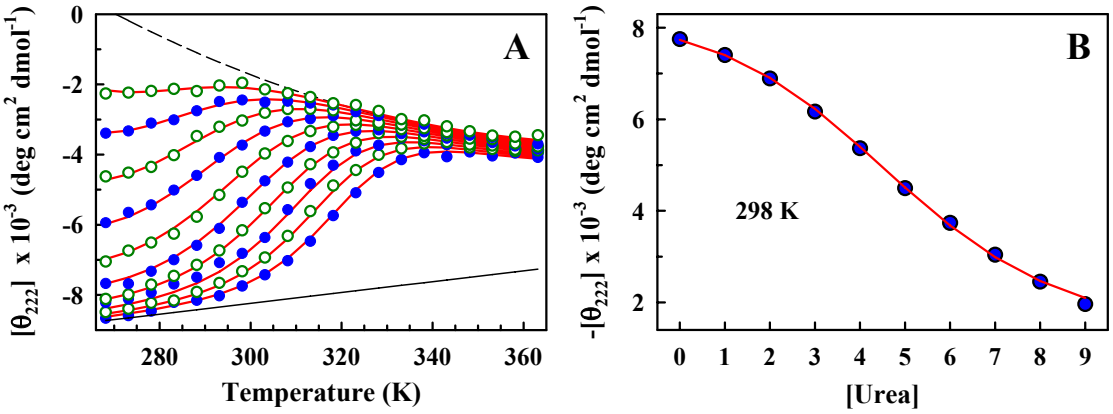


Figure 7.9 A) Double perturbation experiment (blue filled and open green circles) as a function of temperature and 0 to 9 M urea (in steps of 1 M) together with a global two-state fit (red curves). The folded (continuous black line) and unfolded (dashed black line) from the two-state model are also shown. The unfolded baseline corresponding to only 9 M urea is shown for the sake of clarity. B) The ellipticity at 222 nm and 298 K as a function of urea (blue circles) with the two-state fit (red curve) shown in panel A. (Experiments by Naganathan AN).

6). Surprisingly, very little cold denaturation is observed - the temperature of maximum signal (T_{max}) is not evident even at high urea concentrations. This was further confirmed by repeating the experiments with GuHCl (data not shown) thus eliminating any denaturant-specific effects. The absence of cold denaturation induced curvature in the signal complicates the extraction of the heat capacity change arising out of solvation. Figure 7.9A also plots the results from a two-state fit (red curves) assuming a quadratic and linear dependence of the unfolded state on temperature and urea, respectively, and a linear temperature dependence on the folded state. The linear energy model was used as it directly provides an estimate on m_{eq} (Chapters 2 and 6). The final thermodynamic parameters from the fit are: ΔH_m = 105.1 (± 3.1) kJ mol^{-1}, T_m = 319.6 (± 0.6) K, m_{eq} = 1.33 (± 0.05) kJ mol^{-1} M^{-1} and ΔC_p = 1.64 (± 0.08) kJ mol^{-1} K^{-1}. Though the fitting errors in ΔC_p are small, the magnitude is highly sensitive to the description of baselines. In fact, two-state fits employing different baseline

assumptions including denaturant dependence of heat capacity and/or folded and unfolded states and a coupling term for temperature/denaturant produce a heat capacity change in the range of ~0.6 – 2.3 kJ mol^{-1}. This translates to per residue ΔC_p values of 15 – 55 J mol^{-1} K^{-1}. Comparable results have also been reported by Raleigh and co-workers. The m_{eq} values have similar sensitivities to baselines with values anywhere between ~1 - 1.7 kJ mol^{-1} M^{-1}.

The observed dependence of the magnitude of the thermodynamic parameters on the baselines clearly emphasizes that a two-state approximation is not valid. Though Raleigh and co-workers report a similar dependence[141], they attribute these uncertainties to the smaller size of the protein that would in other words produce broader transitions. However, results from size-scaling arguments also point to a correspondingly smaller barrier height with a decrease in size.

7.3.7 Fluorescence of Naphthyl Alanine

The naphthyl alanine (NALA) label at the C-terminus is sensitive to the unfolding process, as evidenced by the sigmoidal changes in quantum yield (QY) as function of temperature (Figure 7.10A). Free NALA has a QY of ~ 0.11 with a small and negative intrinsic temperature dependent slope (data not shown). NALA tagged to the protein has a QY of 0.13 at the lowest temperature and ~ 0.07 at the highest temperature. The slopes at higher temperatures are much more than that expected from the intrinsic temperature dependence alone. These observations suggest that there is a stimulation of fluorescence at lower temperatures probably as a result of interaction with either the hydrophobic core or the residues of 2nd helix. At higher temperatures the fluorescence is quenched compared to free NALA indicating that the

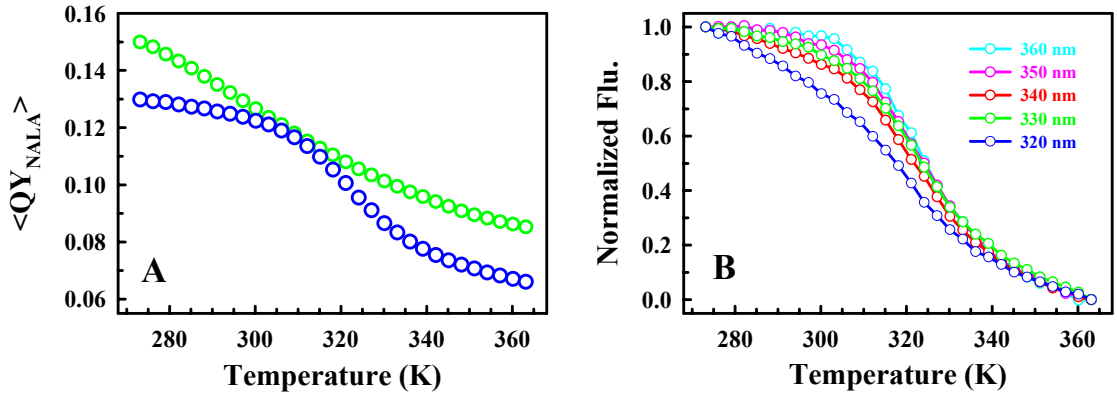


Figure 7.10 A) The temperature dependent naphthyl quantum yield at pH 7.0 (blue) and pH 3.0 (green). B) Wavelength-dependent normalized unfolding curves at pH 7.0. (Experiments by Naganathan AN).

unfolded state has some residual structure in agreement with other spectroscopic probes. A simple two-state fit results in a T_m of 322 K and a ΔH_m of 112 kJ mol^{-1}. QY of NALA at pH 3.0 has a significant temperature dependent slope - and is similar magnitude to the high temperature pH 7.0 data - perhaps indicating that the region around tagged NALA is involved in the structured unfolded state. The QY temperature dependence at pH 3.0 shows no sigmoidal behavior further confirming that the pH 7.0 data monitors a conformational transition. As with far-UV CD the NALA fluorescence shows a wavelength dependent unfolding transition resulting in apparent two-state T_ms in the range of 321-325 K (Figure 7.10B). The differences in the pre-transitions are also clearly evident from being flat at 360 nm to highly sloped at 320 nm.

7.3.8 Förster Resonance Energy Transfer (FRET)

The addition of dansyl lysine at the N-terminus stabilized the doubly-labeled protein by ~3 K compare to the singly-labeled variant. Therefore, the stabilities were matched by decreasing the ionic strength of the buffer from 43 mM (*i.e.* 20 mM

sodium phosphate buffer) to 21 mM (10 mM buffer) for the experiments on doubly-labeled PDD. Changes in mean end-to-end distance (<r>) of PDD were then followed by monitoring the NALA QY changes in singly- and doubly-labeled protein in buffer concentrations of 20 mM and 10 mM, respectively, as a function of temperature. The average changes in FRET efficiency ($<E_T>$) was calculated using equation 2.7. The plot of such a calculation at pH 7.0 is shown in Figure 7.11A (blue circles). One could convert the changes in $<E_T>$ to <r> by using the experimentally determined R_0 of the free NALA-Dansyl pair as a function of temperature. But, the QY of the donor tagged to the protein is different from the free-dye and changes considerably with temperature (Figure 7.10A). To account for these changes, R_0 was calculated from the $<QY_{NALA}>$ and the extinction coefficient of the free dansyl. Ideally, this calculation should be done with the extinction coefficient of the dansyl attached to the protein. But since there is large overlap between the absorbance bands of dansyl and naphthyl groups, the absorbance spectrum of the free dansyl was used. The resulting value of <r> (green circles in Figure 7.11A) shows a constant end-to-end distance value of 2.2 nm till 310 K after which it decreases sharply to a final value of 1.85 nm at 363 K. A two-state fit resulted in a T_m of 324 K and a ΔH_m of ~149 kJ mol^{-1}, clearly much larger than the numbers obtained from other techniques. The pH 3.0 data also show a significant change in end-to-end distance as a function of temperature (green circles in Figure 7.11B). Though the apparent transfer efficiency is smaller, the <r> values are similar to the numbers at pH 7.0 because of the pH dependence of R_0. As with NALA QY the absence of a sigmoidal transition is evidence to the fact that not all the distance changes observed at pH 7.0 are the result of unfolded state effects.

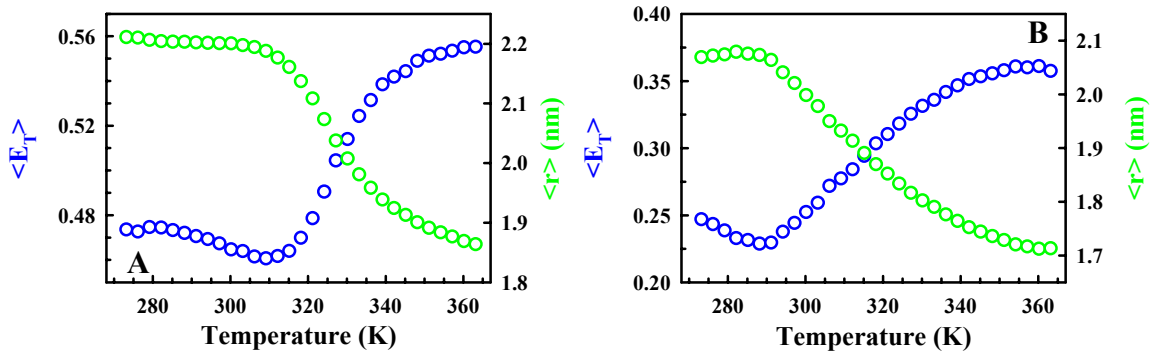


Figure 7.11 A & B) Transfer efficiency (<E_T>) and the end-to-end distance (<r>) at pH 7.0 and pH 3.0, respectively. (Experiments by Naganathan AN).

A surprising result is the decrease in end-to-end distance with an increase in temperature suggesting a collapse of the polypeptide chain. What causes the protein to collapse? Model hydrophobic compound studies indicate that the transfer free energy from the pure phase to water has a parabolic dependence similar to that of proteins[147]. Molecular dynamics simulations of pure hydrophobic homopolymers also reveal an analogous dependence[148]. The corresponding temperature versus free energy plots show a maximum at > 380 K signaling the temperature at which the exposure of non-polar groups to the solvent becomes the most unfavorable beyond which it become favorable again. These observations imply that the increase in conformation entropy upon unfolding can be compensated by an increasing hydrophobic 'force' (within this temperature range of 273-363 K) thus promoting the collapse of the polypeptide. This does not mean that there is little conformational heterogeneity in the unfolded state. MD simulations by Pande and co-workers on several small proteins reveal that the mean geometry of the unfolded state (from thousands of simulations) is similar to that of the native structure, though the individual members of the unfolded state are themselves quite different from the

native structure[149]. They also observed a collapse of the structures upon unfolding with a mean radius of gyration (and hence end-to-end distance) similar to that of the folded ensemble, consistent with the results from PDD. Therefore the apparent paradox of smaller end-to-end distance of an 'unfolded state' and a total end-to-end distance change of just 4 Å can be resolved by recognizing that ensemble experiments report on the average and not the distribution. In the ensemble view of protein folding, a distribution of structural features is the norm, and this is particularly relevant for changes in end-to-end distance. Such distributions have been invoked to explain the apparent random-coil behavior of unfolded states using the Gaussian chain model where excluded volume and intra-chain interactions are ignored. But do the folded states have a fixed end-to-end distance? There is no definitive answer to this question as there is no concrete evidence for or against this statement. Single molecule experiments do reveal a significant width of the folded sub-population; but a quantitative interpretation is complicated by shot noise contributions to the width[34]. Interestingly, Monte Carlo simulations by Fitzke and Rose reveal that it is possible for proteins with significant native structures and connected by flexible linkers to exhibit random coil statistics, though they did exclude interaction effects[150]. All of these results merely highlight the difficulty in interpreting the end-to-end distance data.

Furthermore, the FRET experiment also provides subtle evidences on the magnitude of other thermodynamic parameters, specifically the heat capacity change upon unfolding and the barrier height. A collapsed unfolded state should have a significant burial of hydrophobic surface area. Therefore, the net change in solvation

going from folded to unfolded states is also bound to be lower and hence a smaller ΔC_p. The small changes in <r> are also suggestive of the minimal movements of the unfolded wells (from the folded well) in one-dimensional free-energy profiles of marginal barrier or downhill folding proteins.

7.3.9 IR Kinetics

The relaxation kinetics of PDD was studied by infrared laser temperature-jump (IR T-jump) technique. This experiment is particularly advantageous as it directly monitors the amide I' region of the infra-red spectrum (1632 cm^{-1} in this case) thus reporting on the changes in secondary structure content with temperature. The set up and the details of the experiment are explained in Chapter 2.

All the kinetic relaxation transients had 3 phases with only one corresponding to that of the protein (Figure 7.12A). A fast phase at around 100-200 ns has been reported for a number of fast-folding proteins and has been attributed to the formation of helical structures[61,115]. A fast phase is clearly seen in this experiment as well. However, its origin in our experimental set-up is not clear. This is because of the observation of this phase even in buffer solutions with amplitude that increases linearly with temperature. It is possibly a result of cavitation artifacts; therefore the rate of this phase or its amplitude was not analyzed further. The slowest phase is the result of thermal energy diffusing within the pump-probe volume as can be seen by the negative amplitude of this component. The intermediate phase corresponds to that of the protein relaxation and is a single exponential within the signal to noise ratio of the experiment. Thus, the decays were fit to a sum of three exponentials with the rate of the cooling (~3 ms) and total amplitude fixed. The resulting relaxation rates

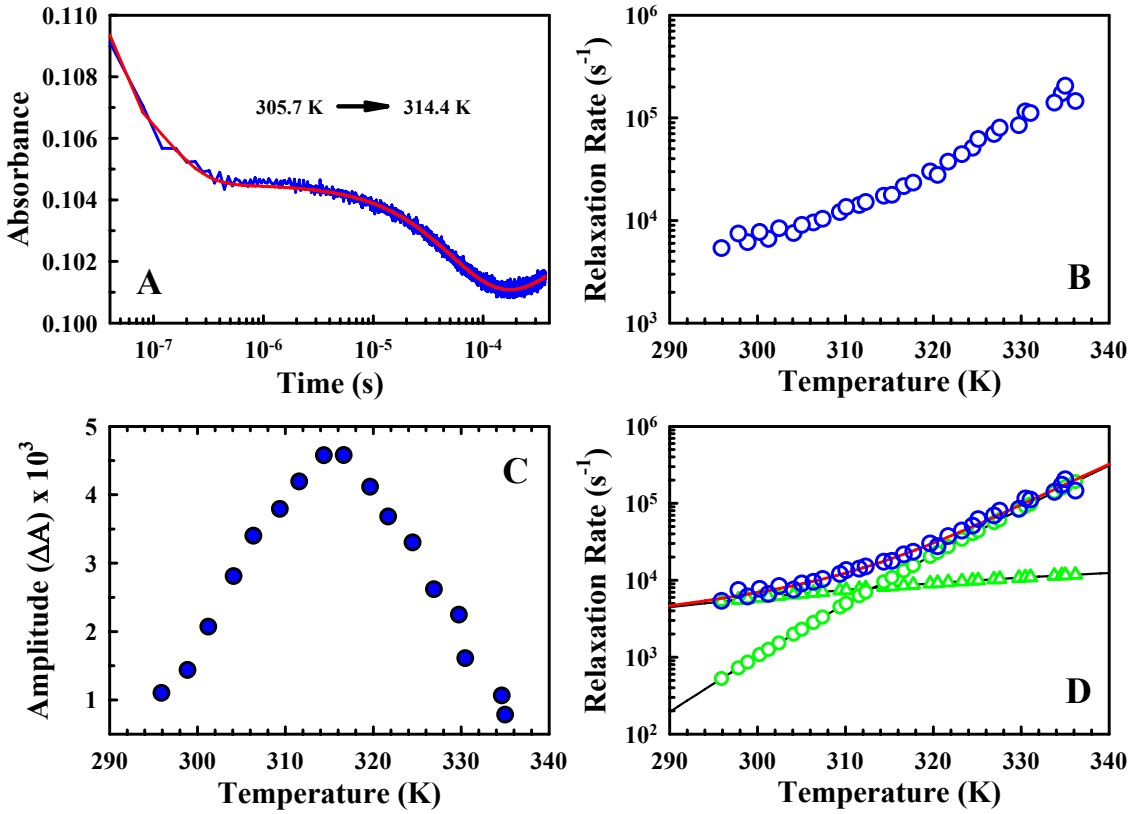


Figure 7.12 A) Kinetic relaxation curve (blue) monitored at 1632 cm^{-1} and the triple exponential fit (red) at 314.4 K. B) The plot of relaxation rate versus temperature. C) The kinetic amplitude for one continuous experiment in absolute units. D) Two-state fit (red curve) to the data shown in panel B with the folding (green triangles) and unfolding (green circles) rate constants. (Experiments by Li P & Naganathan AN).

(Figure 7.12B) have a steep temperature dependence changing from ~5000 s^{-1} (τ = 200 μs) at 296 K to ~200,000 s^{-1} (τ = 5 μs) at 335 K. Though there is a positive heat capacity change upon unfolding (see DSC section), it is not evident in the temperature dependence of the relaxation rate. It shows no apparent downward curvature at lower temperatures which is otherwise seen in a number of two-state-like proteins (see Figure 2.3A for example). At a first glance the rate versus temperature plot is also reminiscent of the low-barrier curves shown in Figure 5.6, with the rate shifted down by about an order of magnitude.

The temperature versus amplitude of the protein relaxation phase is broad with no pre- or post-transition slopes hinting at a continuous structural change within the experimental temperature range (Figure 7.12C). It shows a maximum at ~ 315-316, in agreement with the model-free first-derivative analysis of the equilibrium signal at 1632 cm^{-1}. However, there is one difference. The equilibrium first derivative is significantly broader than it kinetic counterpart suggesting that the equilibrium signal has a temperature dependence (not shown). The origin of this dependence and an estimate are presented in Chapter 8. The apparent T_m estimate from the maximum of the amplitude is similar to that reported by from the first-derivative analysis of far-UV CD at 222 nm (~318 K). Interestingly, the maximum of the amplitude also compares well with the numbers obtained from the temperature dependent amplitude of the 3rd basis spectrum from various techniques (Figures 7.4F, 7.6F, & 7.8F).

The temperature versus rate data was fit to a two-state model using the equations 2.24 and 2.25 to extract the thermodynamic parameters and check for the consistency with the equilibrium FTIR experiment. For the following analysis, the value of k_0 was taken to be 1.5 x 10^8 so that the D_{eff} at 333 K corresponds to ~2 μs (see chapter 5). The T_m was used as the reference temperature and was fixed to 316 K from the maximum of kinetic amplitude. The resulting fit and the parameters are shown in Figure 7.12D and table 7.1, respectively.

Table 7.1 Parameters from a two-state fit of the Figure 7.12B

Parameter	†-*U*	†-*F*	*U*-*F*
ΔH (kJ mol^{-1})	0.37	105.00	104.64
ΔS (kJ mol^{-1} K^{-1})	-0.03	0.30	0.33
ΔC_p (kJ mol^{-1} K^{-1})	-0.18	0.00	0.18

The quality of the fit is highly sensitive to the starting numbers. Therefore, the above parameters correspond to a fit that gave the least residuals. They reveal a situation wherein the enthalpy of the transition state is higher (i.e., positive $\Delta H^{\dagger\text{-}U}$ and $\Delta H^{\dagger\text{-}F}$) than both of the ground states while the entropy and heat capacity of activation is intermediate. The higher enthalpy of the transition state has been traditionally seen as an origin of the barriers to protein folding – and hence the popularity of 'enthalpic barriers' in the field. For PDD, the folding activation enthalpy is just ~0.3 % of the total enthalpy change suggesting little or no enthalpic barriers even at the T_m; but the assumption of a fixed pre-exponential produces a much larger folding *free-energy barrier* (~9.5 kJ mol^{-1}). However, there is a fundamental misconception in this analysis as noted by Akmal and Muñoz[62]. Using a structure based description of temperature dependent kinetic data from two-state-like proteins they noted that the positive enthalpic barrier is a result of the non-inclusion of the entropic free energy of solvation that is also stabilizing. Upon including this term they find that the apparent enthalpy of the transition state is in between that of the ground states consistent with the other two parameters. The barrier height will then be determined by the compensation between stabilizing enthalpic and destabilizing conformational entropic terms. Akmal and Muñoz's treatment did include a range of pre- exponentials to arrive at this conclusion. However, the analysis presented here uses a single pre-exponential necessitating the need to independently estimate the folding free energy barriers. This exercise also highlights the inability of the simple two-state model to estimate the folding barriers or the precise meaning of the thermodynamic parameters.

The total enthalpy change from kinetic fit is ~30 kJ mol^{-1} higher than that obtained from a two-state fit to the 1632 cm^{-1} signal. As noted in Chapter 2, this disagreement between equilibrium and kinetics is strong evidence to the non-two-state nature of PDD. Interestingly, the parameters also suggest a transition state whose solvation properties are identical to the fully folded state (as $\Delta C_p^{\dagger-F} = 0$) and marginally different from that of the fully unfolded state. In other words, the change in heat capacity between the folded and unfolded states is very small in agreement with results from FTIR and chemical denaturation. Due to high degree of correlation between the parameters, the total change in heat capacity can vary between zero to 1 kJ mol^{-1}, i.e. a maximum change in heat capacity per residue of ~ 20 J mol^{-1} K^{-1}, but these numbers are much smaller than that expected by size scaling arguments alone (~ 50-58 J mol^{-1} K^{-1} per residue).

7.4 Conclusions - The Unfolding of PDD is Not Two-State

The experimental characterization of PDD reveals a number of observations inconsistent with a two-state picture. Importantly,

a) Two-state fits to the DSC profile render crossing baselines. As outlined in the previous chapters, this outcome is a clear signature of the non-two-state nature of the transition. Moreover, the lowest temperature point of the DSC thermogram is ~1 kJ mol^{-1} higher than that of Freire's baseline implying substantial thermodynamic fluctuations.

b) A spread of apparent ΔH_m and T_m is observed, varying between 60-150 kJ mol^{-1} and 316-325 K. The normalized experimental signals are shown in Figure 7.13A, with changes in apparent enthalpy evident from varying pre-

and post-transition slopes. In a classical two-state interpretation the above differences are attributed to the temperature dependence of signals. This is probably true in the case of fluorescence and to a minor extent for FTIR as they have intrinsic temperature dependence. But the steep pre-transition slopes evident in DSC, far-UV and near-UV are clearly structural in origin. This result is also consistent with the interpretation of the molecular origins of the third component in the spectroscopic data. Eliminating the 'baseline effects' with a two-state model still produces a dispersion in T_m (Figure 7.13B).

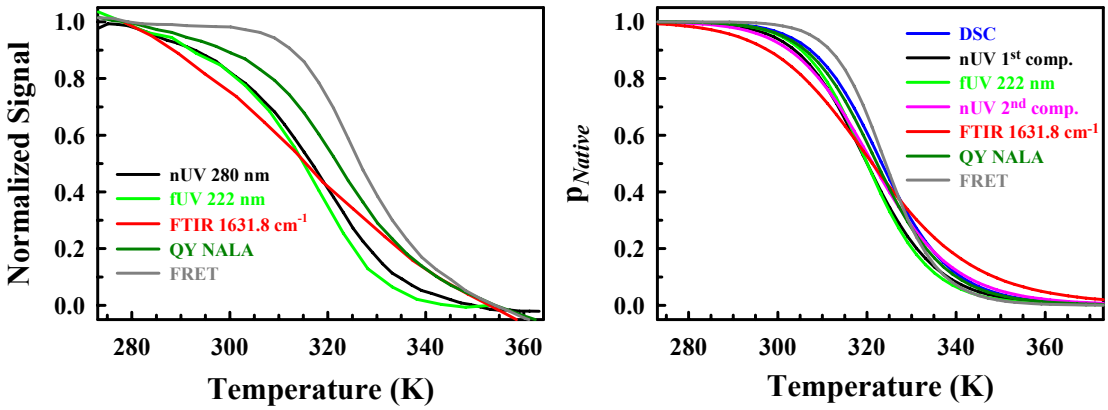


Figure 7.13 A) Normalized raw signals for the experimental probes indicated. B) The native state probability for the various probes as derived from a two-state fit.

c) Results from FTIR, chemical denaturation, kinetic fit to a pseudo-two-state model and to a lesser extent FRET hint at small changes in heat capacity arising out of solvation effects that is probably in the range of 0 – 1.2 kJ mol^{-1}. However, the DSC thermogram shows a much larger change of ~3 kJ mol^{-1} between the lowest and highest temperatures. This observation also emphasizes that enthalpy fluctuations contribute significantly to the heat capacity changes suggestive of folding over marginal/zero barriers.

d) The shape of the kinetic relaxation rate versus temperature is similar to that reported for folding over marginal barriers (Chapter 5) albeit shifted to lower values.

e) The thermodynamic parameters extracted from equilibrium and kinetics are inconsistent with one another.

Furthermore, a two-state picture certainly does not provide an independent estimate of the pre-exponential and hence the barrier-height to the folding reaction. In spite of the above arguments, one could still propose a hypothetical 'two-state' situation wherein there is an ensemble of structures constituting the folded state, whose structures change with temperature but separated from the unfolded state by a large barrier, sufficient to invoke a transition-state-like treatment. But downhill folding or folding over marginal barriers is mechanistically different from either one of them. A simple way to distinguish these different scenarios is to analyze the data with a structure-based model that directly incorporates the physical details of the folding reaction. This is the topic of Chapter 8.

8. Evolutionary Conservation of Downhill Protein Folding: 2. Statistical Mechanical Modeling of Equilibrium and Kinetic Signals

8.1 Introduction

This chapter attempts at reproducing the equilibrium and kinetic data of PDD quantitatively with a structure-based statistical mechanical model. This model has been earlier used to explain the complex thermodynamic and kinetic behavior of alpha helices and beta hairpin[11] and to predict protein folding rates from three-dimensional structures[9]. It has also been highly successful in describing the thermodynamics of BBL unfolding resulting in barrier-less free-energy profiles at all temperatures explored[47]. Section 8.1 introduces model; this is followed by a comprehensive treatment of the DSC thermograms, spectroscopic and kinetic signals in sections 8.2, 8.3, and 8.4, respectively. It predicts the folding over a marginal barrier (2 ± 2 kJ mol^{-1}) for this domain at ~320 K – Section 8.5. Section 8.6 provides an evolutionary view of folding in the entire PSBD family with implications in function.

8.2 Structure–based Statistical Mechanical Model

The model is entirely based on the structure of the protein under consideration. It is Ising-like in the sense that each residue is assumed to have only two conformations: folded (native) and unfolded (non-native) [14]. It accounts for the statistical nature of folding while at the same time greatly reducing the conformational space explored by employing a single- or double-sequence

approximation - single sequence approximation allows for only one stretch of residues in native conformation while a double-sequence allows for two native stretches and so on.

8.1.1 Parameterization

The enthalpic contribution to the free energy per residue ($\Delta H_{\mathrm{res}}^{\mathrm{T_m}}$) and the cost in conformational entropy of fixing a residue in native conformation ($\Delta S_{\mathrm{res}}^{\mathrm{T_{385}}}$) are assumed to be the same for all residues and conformations, respectively, in spirit of mean-field models. Invoking a single-sequence approximation to represent the instantaneous ensemble of the 42-residue (N) PDD results in 904 species, including the reference unfolded state (903 + 1). The structure of the species is directly obtained by editing the PDB file. Each species can then be defined by just two numbers: position of the first native residue (m) and the length of the native stretch (n) (for example, the native state of PDD is represented as (1, 42)). The individual probabilities ($p^{(\mathrm{m,n})}$) of the structured species and the unfolded state (p^{U}) are calculated from:

$$p^{(\mathrm{m,n})}(T) = \frac{w^{(m,n)}(T)}{1 + \sum_{m=1}^{N} \sum_{1}^{N-m+1} w^{(m,n)}(T)} \tag{8.1}$$

and

$$p^{\mathrm{U}}(T) = \frac{1}{1 + \sum_{m=1}^{N} \sum_{1}^{N-m+1} w^{(m,n)}(T)} \tag{8.2}$$

where the denominator is the partition function of the system. The statistical weight $w^{(m,n)}$ is defined as

$$w^{(\mathrm{m,n})}(T) = \exp\left(\frac{-\Delta G^{(\mathrm{m,n})\text{-}\mathrm{U}}(T)}{RT}\right)$$

where

$$\Delta G^{(\mathrm{m,n})\text{-}\mathrm{U}}(T) = n.\Delta H_{\mathrm{res}}^{\mathrm{T_{ref}}} + \Delta C_{\mathrm{p}}^{(\mathrm{m,n})\text{-}\mathrm{U}}(T - T_{ref}) - T[n.\Delta S_{\mathrm{res}}^{\mathrm{T_{385}}} + \Delta C_{\mathrm{p}}^{(\mathrm{m,n})\text{-}\mathrm{U}} \ln(T / T_{385})] \qquad (8.3)$$

$\Delta C_{\mathrm{p}}^{(\mathrm{m,n})\text{-}\mathrm{U}}$ represents the change in heat capacity arising out of solvation upon unfolding (see Chapter 2), referenced to the unfolded state (*U*). ΔC_p is assumed to depend linearly on the difference in accessible surface areas (ΔASA) between the structured species and the fully unfolded state, as empirically observed in a number of proteins[77]:

$$\Delta C_p^{(\mathrm{m,n})\text{-}\mathrm{U}} = a\Delta ASA^{(m,n)-U} \qquad (8.4)$$

where *a* is the proportionality constant in units of J mol^{-1} K^{-1} $Å^{-2}$. ASA for an unfolded residue (X) is calculated from the standard tri-peptide model, Gly-X-Gly, the assumption here being the unfolded state of PDD is ideal with no residual structure. For reasons described in Chapter 5, *385 K* is used as the reference temperature for calculating the cost in conformational entropy and T_{ref} is the reference for the enthalpic part. To summarize, the model requires just 3 thermodynamic parameters ($\Delta H_{\mathrm{res}}^{\mathrm{T_{ref}}}$, $\Delta S_{\mathrm{res}}^{\mathrm{T_{385}}}$ and *a*) in addition to an average about 2 spectroscopic parameters per experiment (see below). The thermodynamic parameters essentially describe the probability of the ensemble of conformations as a function of temperature. This is much simpler than a typical two-state model (that needs at least 2

thermodynamic and 4 spectroscopic parameters per experiment) while at the same providing information on the barrier height to folding and the possible origin of the observed pre-transition slopes in spectroscopic measurements.

8.2 Analysis of DSC Thermogram

8.2.1 Variable Barrier Model

The calorimetry profile of PDD was first characterized by the variable barrier model[49] (see Chapter 4). As an accurate estimate of the absolute heat capacity is available, Freire's native baseline (slope ~31 J mol^{-1} K^{-2}) was directly used. It resulted in a very good fit (Figure 8.1A) with a similar quality of a two-state analysis, providing a barrier height estimate (β) of 0.15 (± 0.04) kJ mol^{-1} at 322.9 (± 0.4) K (T_0). It is worthwhile to note that a two-state fit resulted in crossing baselines with a much higher apparent slope of ~51 kJ mol^{-1} K^{-2}. The resulting apparent enthalpy at T_0 ($\Sigma\alpha$) is 105.2 (± 6.7) kJ mol^{-1} with an asymmetry factor of 1. The high asymmetry factor suggests a broad distribution of states even at the lowest temperatures explored. The low barrier and a high asymmetry factor are consistent with the steep pre-transition heat capacity slope observed for this protein.

The uncertainty in native baseline determination and its effect on the calculated barrier heights was explored by assuming various baselines and then checking for the resulting quality of fits. The best fit overall was for a folded baseline up-shifted by 0.5 kJ mol^{-1} with respect to the Freire's predicting a barrier height of ~1.1 kJ mol^{-1} at the T_0. Any further up-shifting of the baseline resulted in poorer fits while at the same time increasing the barrier heights. These estimates together with other baseline assumptions place the barrier height of PDD at 2 (± 2) kJ mol^{-1} at T_0.

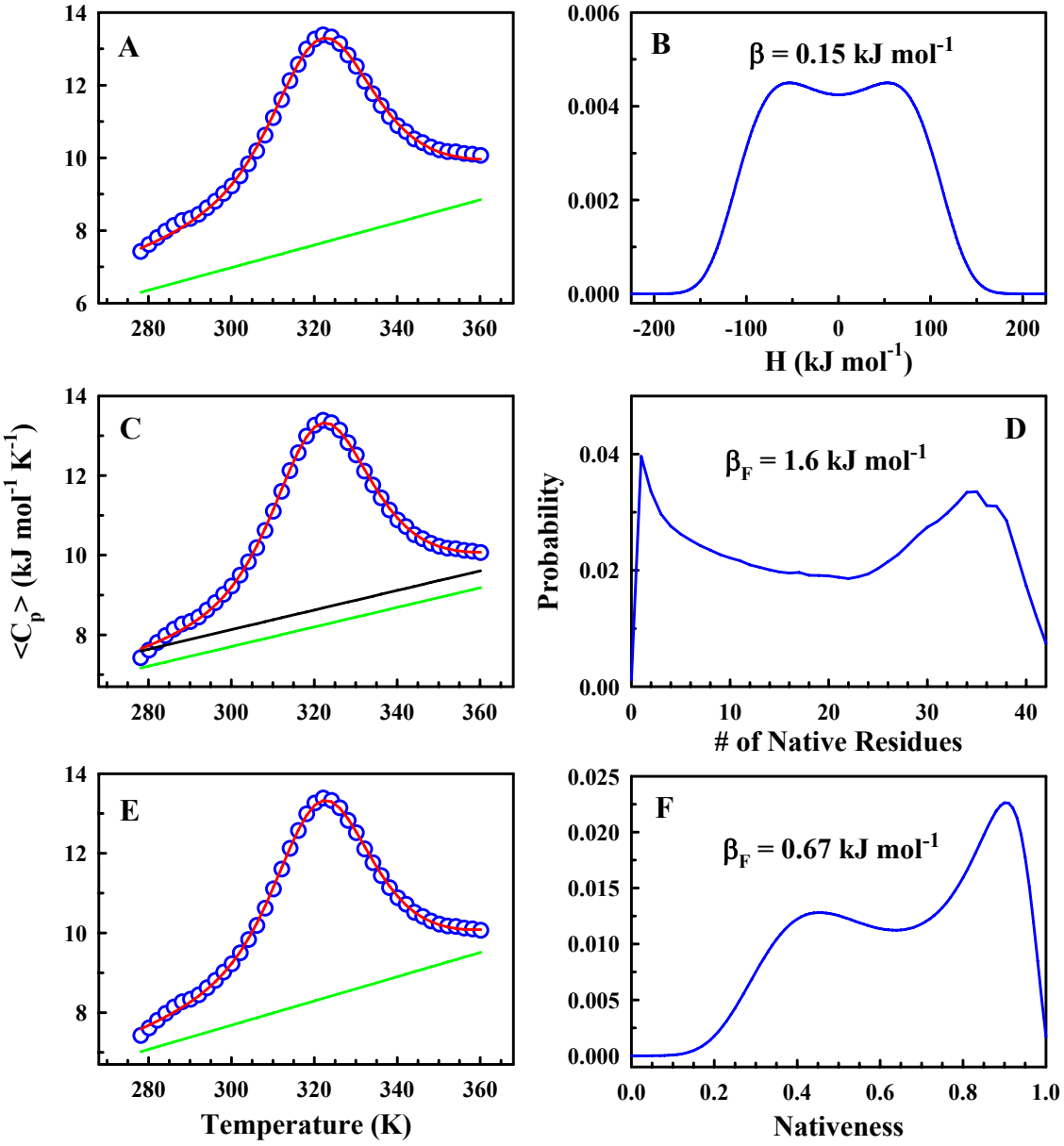


Figure 8.1 Fits (red curve) to the DSC thermogram (blue circles) using various models and the corresponding probability density. β stands for the barrier height at the T_m values noted in the main text. A & B) Variable Barrier Model. The green line is the Freire's baseline. C & D) Structure-based model. The green and black lines are the predicted folded and unfolded baselines, respectively. E & F) DM Model. The green line is the predicted native baseline assuming a $\Delta C_p = 0$.

The limits can be thought of as 95 % confidence intervals. As discussed in Chapter 4, this model attributes all the changes in heat capacity to difference in conformational fluctuations between the folded and unfolded ensemble, while ignoring solvation

effects. How much does the incorporation of solvation terms affect the barrier height estimates?

8.2.2 Structure-based Statistical Mechanical Model

To answer the above question, the DSC profile was analyzed by the structure-based statistical mechanical model that directly incorporates the changes in heat capacity arising out of solvation. The fit required a total of 5 parameters, the three thermodynamic parameters plus an additional two for the unfolded state heat capacity (C_p^U) that is assumed to vary linearly with temperature (T):

$$C_p^U = C_p^{T_o} + b(T - T_0)$$

where $C_p^{T_o}$ is the heat capacity of the unfolded state at T_0. Assuming that the heat capacity of the structured and unfolded states' have the same temperature dependence, the heat capacity of individual species can be calculated from:

$$C_p^{(m,n)} = C_p^U - \Delta C_p^{(m,n)-U}$$

The value of T_0 was fixed to 273.15 K to enable direct comparison with the Freire's slope. The reference temperature for $\Delta H_{\text{res}}^{T_{\text{ref}}}$ was fixed to 324 K which is the maximum of the heat capacity peak. The intrinsic and transition heat capacities were calculated as described in Chapter 2. The number of parameters required by this model to fit the DSC profile is one less than that used by a typical two-state fit.

The model fits the DSC data very well (Figure 8.1C). The final thermodynamic parameters from the fit are: $\Delta H_{\text{res}}^{T_{\text{ref}}}$ = - 5.15 (± 0.22) kJ mol^{-1}, $\Delta S_{\text{res}}^{T_{385}}$ =

-17.47 (± 0.19) J mol^{-1} K^{-1}, a = 0.15 (± 0.04) J mol^{-1} K^{-1} Å^{-2}, $C_p^{T_0}$ = 7.47 (± 0.22) kJ mol^{-1} K^{-1} and b = 24.60 (± 1.80) J mol^{-1} K^{-2}. The parameters are consistent with various size-scaling arguments discussed in Chapter 4. Specifically, the cost in conformational entropy per residue is in close agreement with the numbers estimated by Robertson and Murphy[97] (-17.3 J mol^{-1} K^{-1}). Also, the slope of the heat capacity baseline obtained from the fit (24.6 J mol^{-1} K^{-2}) is similar to that predicted by Freire (31.1 J mol^{-1} K^{-2}). The total change in heat capacity, *i.e.* the difference between folded and unfolded heat capacity baselines (green and black lines, respectively, in Figure 8.1C), is just 10 J mol^{-1} K^{-1} per residue (ΔC_p^{res};compared to 58 J mol^{-1} K^{-1} obtained for larger proteins). However, this value is consistent with the small heat capacity change predicted by a two-state fit to the kinetic data and also evidenced by the minimal cold denaturation upon chemical denaturation. The probability distribution of the species obtained from the fit can be projected on to a single reaction co-ordinate – the number of native residues – to generate one-dimensional free energy profiles as a function of temperature. The resulting free energy profiles are downhill at lower and higher temperatures (data not shown) while predicting a marginal folding barrier (β_F) of 1.6 kJ mol^{-1} at the apparent T_m of 321 K (Figure 8.1D). T_m here is defined as the temperature at which the probability weighted reaction co-ordinate (<n>) is 21, *i.e.* the temperature at which the mean native stretch is half the protein length (N/2). It is of interest to note that the same model produced downhill folding profiles at the apparent T_m for BBL.

8.2.3 DM Model

To check for the robustness of the calculated value of barrier height and its degree of sensitivity to ΔC_p^{res}, the calorimetry profile was characterized by the DM model for protein folding described in Chapter 5. The same energy, entropy and heat capacity functionals were used. The fit required only 4 parameters: $k_{\Delta H}$, ΔH_{res}^0 and the linear baseline for the folded state. A grid analysis was performed by fixing ΔC_p^{res} to a particular number and floating the rest. The fit assuming ΔC_p^{res} of zero J mol^{-1} K^{-1} is shown in Figure 8.1E together with the predicted native baseline (green) of slope 30.4 (± 0.9) J mol^{-1} K^{-2}. It renders a folding barrier height of 0.67 kJ mol^{-1} at the estimated T_m of 326 K (T_m here is defined as the temperature at which <nativeness> = (n_U + n_F)/2)) in agreement with the results from variable-barrier model that employs the same assumption (Figure 8.1F). The approximation ΔC_p^{res} = 0 J mol^{-1} K^{-1} results in the best quality of fit while getting progressively worse for higher values. The barrier height at the T_m also increases up to ~5 kJ mol^{-1} for a ΔC_p^{res} = 30 J mol^{-1} K^{-1}, while the predicted native baseline becomes more flat and starts deviating significantly from Freire's baseline. This behavior is entirely expected and the exercise clearly indicates the difficulty in interpreting the possible origin of the observed changes in heat capacity. Another drawback is the assumption of a constant heat capacity difference between the folded and unfolded states at all temperatures that need not be true. This problem could in principle be overcome by fixing the unfolded baseline to that predicted by Makhatadze and Privalov[151]; however, various treatments of the folded and unfolded baseline failed to produce reasonable fits to the thermogram. But the

barrier heights were insensitive to the value of $k_{\Delta C_p}$. The above analysis predicts a barrier of ~1.5 kJ mol^{-1} at the T_m for a ΔC_p^{res} of 10 J mol^{-1} K^{-1}, further confirming the results from the statistical mechanical model. This investigation also highlights the superiority of the Variable Barrier Model over other models.

The calculated thermodynamic barrier heights from the three different models are very similar despite the fact that they use entirely different reaction co-ordinates, namely, enthalpy, number of native residues and nativeness. The striking agreement between these treatments clearly indicates that PDD folds in a downhill fashion at lower temperatures while crossing a marginal barrier of at most 4 kJ mol^{-1} close to the apparent T_m. The range of barrier heights from the Variable Barrier Model for the various baseline assumptions are in fact within the magnitude predicted by other models that incorporate solvation effects. This indirectly suggests that contribution of solvation to the observed heat capacity change is indeed small for these proteins. The prediction from analyzing the shape of relaxation rate plot versus temperature is presented later in this chapter.

8.3 Spectroscopic Characterization

Extracting the probability density from a DSC profile offers distinct advantage over a spectroscopic technique like fluorescence. The latter monitors only the local environment of a probe and is inherently two-state-like as it is difficult to discern in an ensemble measurement the varying degrees of solvent exposure/burial upon unfolding. The same can be said of CD and FTIR, though to a lesser extent, as there are distinct helix length dependent signals that could in principle be derived from the

raw data. DSC in contrast monitors the global properties of a system and provides a more direct and robust measure of the underlying distribution of states. This is because of the fact that enthalpy, entropy and particularly heat capacity (see above) can be more accurately determined from characterizing a DSC profile, as other spectroscopic probes might have temperature dependent signals super-imposed on top of conformational changes. Particularly, the probability density from the structure-based statistical mechanical model has more information than the other two models. By assigning signals to structured species one could then reproduce the apparent slopes and varying T_ms as monitored by fluorescence, FRET, FTIR and CD, as demonstrated below.

8.3.1 Far-UV CD

Since PDD is an alpha-helical protein the far-UV CD spectrum of the 903 structured species can be simply calculated as a linear combination of helical and random coil basis spectra (see equations 6.1 and 6.2). The spectral range spanning 190-240 nm was modeled as there is no information in the 241-250 nm range. Equation 6.1 was used to calculate the ellipticity of helices of varying lengths while the infinite length basis spectrum was taken from Chen *et al*. The far-UV spectrum at pH 3.0 and 348 K was used as the random coil basis (θ_U^{fUV}). The values of $k(\lambda)$ were modified to reproduce the lowest temperature pH 7.0 spectrum. The assignments of N- and C-terminal helices were taken from the PDB structure file, i.e. 5-13 and 31-39, respectively. Helix nucleation has a significant entropic barrier as it requires 4 successive peptide bonds to be in an α-helical conformation without any stabilizing

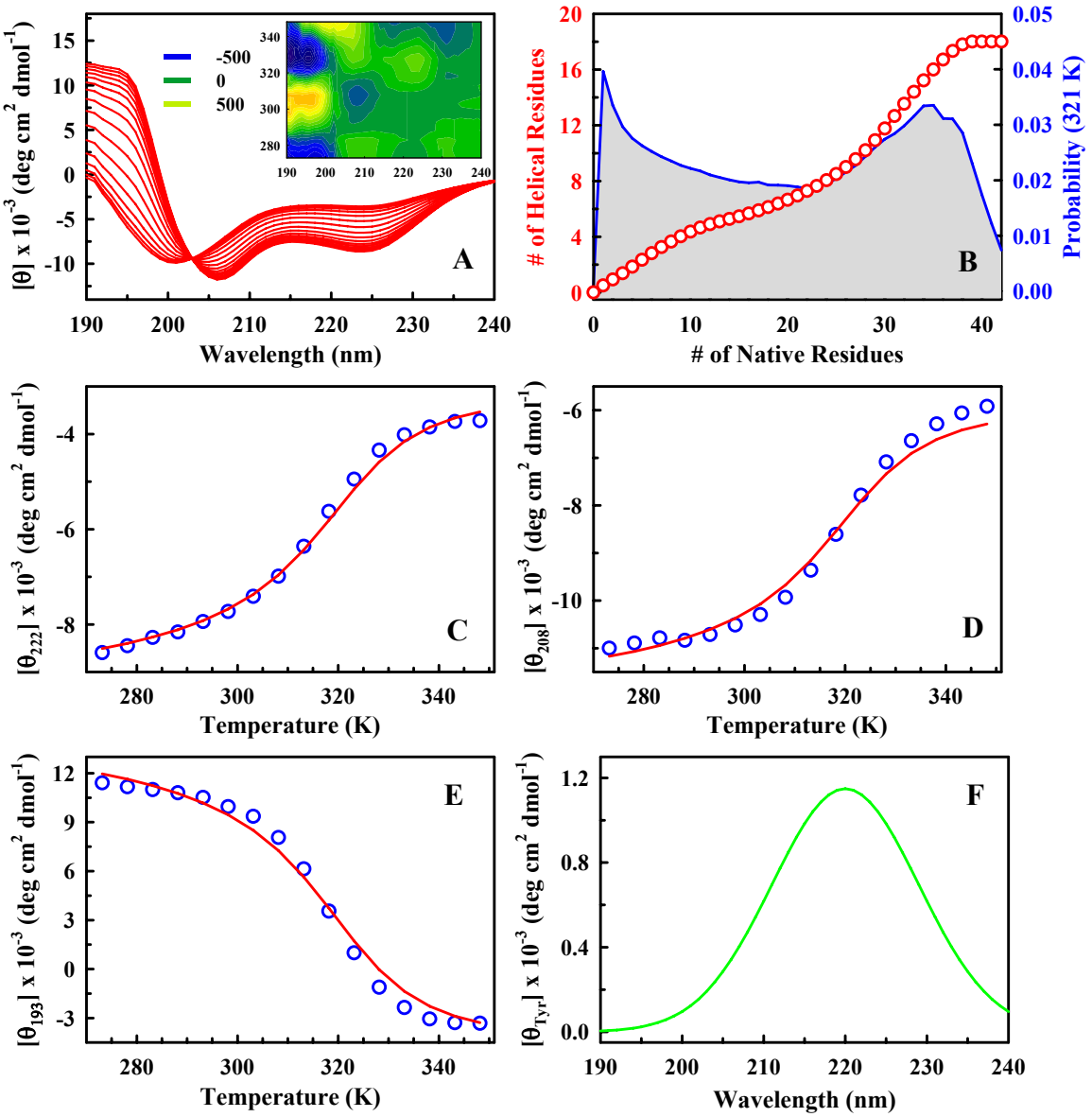


Figure 8.2 A) Calculated pH 7.0 spectra together with the contour map of the difference between the data and fit. B) The number of helical residues superimposed on the probability density at 321 K. C, D & E) Data (blue circles) and fit (red curve) to the molar ellipticity values at 222, 206 and 193 nm. F) The predicted tyrosine spectrum.

interaction[35,36]. A helical nucleus (*nuc*) of 4 was therefore assumed. A helix of length > 4 would then give rise to a length-dependent helical signal while resulting in a proportional coil spectrum otherwise. The far-UV CD signal is calculated from:

$$\theta^{fUV}(m,n,\lambda) = \frac{1}{N}\left[\left(l_H - nuc\right)\times\theta\left(\lambda, l_H\right) + nuc\times\theta_U^{fUV}(\lambda)\right] \quad \text{if } nuc > 4$$

and

$$\theta^{fUV}(m,n,\lambda) = \frac{1}{N}\left[nuc\times\theta_U^{fUV}(\lambda)\right] \qquad \text{if } nuc \le 4$$

The basis spectra were kept constant during the fitting procedure. The changes in magnitude and shape of the spectrum are reproduced by the changes in probabilities of each of the species as a function of temperature:

$$\theta_{Calc}^{fUV}(\lambda,T) = \sum_{m=1}^{N}\sum_{1}^{N-m+1} p^{(m,n)}(T)\times\theta(m,n,\lambda) + p^{U}(T)\times\theta_U^{fUV}(\lambda) \tag{8.5}$$

The fit required no additional parameters apart from the probability density obtained by analyzing the DSC profile. It reproduced the signal at lower wavelengths (<200 nm) while significantly over-predicting at higher wavelengths (not shown). Changing the helical assignments by reducing helical lengths failed to account for the observed discrepancy. One possible reason for this could be the presence of tyrosine which is known to produce a positive band around 222 nm. Though the magnitude of the signal has been previously estimated[143], it is bound to be sensitive to the location and environment of the tyrosine residue and hence protein-specific. To account for this effect, the tyrosine band (θ_{Tyr}^{fUV}) was modeled as a Gaussian function independent of temperature and by fixing the mean to 220 nm. The new fit thus required a total of two parameters accounting for the magnitude and width of the tyrosine band (σ_{Tyr}). The simulated far-UV spectrum is shown in Figure 8.2A. The contour graph inset shows that it reproduces the raw data very well with an mean absolute error of ~ 200 deg cm^2 dmol^{-1} (~5 % of total amplitude across all wavelengths). The fits at

individual wavelengths of 222, 206 and 193 nm, and the predicted tyrosine spectrum are shown in Figures 8.2C-8.2F. σ_{Tyr} was calculated to be ~9 nm, strikingly similar to that measured by Baldwin and co-workers[143] (~10 nm).

It is important to note that no baselines are assumed in this fit. Therefore, the pre-transition slope observed specifically at 222 nm have a straightforward interpretation - they correspond to the gradual melting of helices. This is shown graphically in Figure 8.2B, where the number of helical residues (red circles) is projected onto the reaction co-ordinate along with the probability density (shaded area). At the apparent T_m, the number of helical residues spans from none to 18 (fully folded), highlighting the broad distribution of structural species. This is precisely what is expected from a non-two-state folding process that would otherwise predict only two species – fully folded and fully unfolded. This result is consistent with the non-conformity of DSC thermogram to a two-state model and the marginal barriers predicted by the models employed.

8.3.2 Near-UV CD

The presence of tyrosine offers the unique opportunity to monitor the melting of region between the two helices of PDD. As discussed in the previous chapter, an SVD of the near-UV CD spectra reveals two temperature-dependent components – one signaling the average change in intensity and the other a red-shift of tyrosine. These basis spectra were utilized in simulating the near-UV CD spectral changes. The change in intensity of the first component ($\theta_1^{nUV}(\lambda)$) was modeled to arise from the melting of the core region of the protein, i.e. the native stretch from residue 12-33.

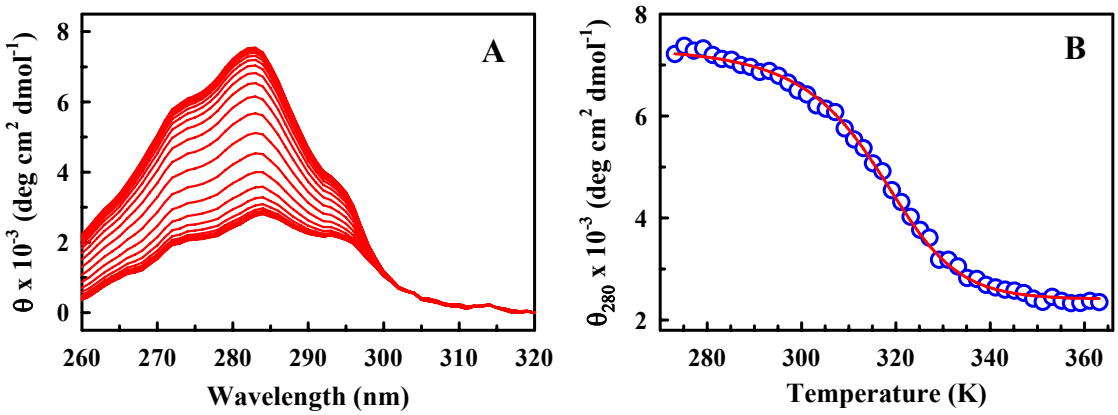


Figure 8.3 A) Calculated near UV-CD spectra. B) The representative signal at 280 nm (blue circles) together with the fit (red curve).

Since the tyrosine is partially buried this would go in hand with the change in its asymmetric environment. Species with intact hydrophobic core were assigned the lowest temperature signal of the first basis ('folded' signal - 1) while the species with unfolded core were assigned the highest temperature signal ('unfolded' signal – 0.37; $C_1^{(m,n)}$). The observed red-shift in the second component ($\theta_2^{nUV}(\lambda)$) is possibly a result of a change in the ASA of tyrosine. The continuous nature of the amplitude of this component together with the fact that tyrosine is located at the center of helix 1 indicates that the change in ASA is possibly connected to the melting of helix 1. To incorporate this effect, the ASA of tyrosine residue in all species ($ASA_{Tyr}^{(m,n)}$) was first calculated. The ASA of tyrosine in species with residues 8-31 folded ($ASA_{Tyr}^{(8,24)}$) was then used as the reference and normalized to the ASA of the fully folded structure ($ASA_{Tyr}^{(1,42)}$). The decrease in amplitude would thus correspond to the change in ASA as a result of the melting of the first turn of helix 1. It can be represented as:

$$C_2^{(m,n)} = \frac{ASA_{Tyr}^{(m,n)} - ASA_{Tyr}^{(8,24)}}{ASA_{Tyr}^{(1,42)} - ASA_{Tyr}^{(8,24)}}$$

As with far-UV CD, the basis spectra are kept constant and the changes in spectra ($\theta_{Calc}^{nUV}(\lambda,T)$) are reproduced by the changes in the probability distribution with temperature:

$$\theta_{Calc}^{nUV}(\lambda,T) = \sum_{i=1}^{2}(\sum_{m=1}^{N}\sum_{1}^{N-m+1} p^{(m,n)}(T)\times C_i^{(m,n)})\times\theta_i^{nUV}(\lambda) + p^{U}(T)\times\theta_U^{nUV}(\lambda) \qquad (8.6)$$

where $\theta_U^{nUV}(\lambda)$ is the unfolded state spectrum and was fixed to the pH 7.0 highest temperature spectrum (363 K).

The fit required no additional parameters. The near-UV spectra thus simulated is shown in Figure 8.3A along with the fit at 280 nm (Figure 8.3B). The assignments of signals reproduce the data very well with a mean absolute difference between the data and fit of ~ 60 deg cm^2 dmol^{-1} (~ 4 % of total amplitude averaged over all wavelengths; contour map not shown). It is of interest to note that the melting of helix 1 monitored by the change in ASA of tyrosine is identical to far-UV CD signal at 222 nm. This indicates that either the assignment of the second component to tyrosine red-shift (and the corresponding modeling) is correct or that near-UV CD is influenced by that of the far-UV CD transitions. Since the tyrosine is at the center of helix 1 it is difficult to distinguish between the two scenarios and neither can be ruled out. Amplitude analysis of far- and near-UV CD thus point to a mechanism where the helices unravel gradually (apparent T_m is also lower) followed by the melting of the hydrophobic core of the protein.

8.3.3 FTIR

Extracting quantitative structural information from FTIR spectra is challenging as the positions, widths, and amplitudes of the various bands change with temperature. As discussed before, the deconvolution procedure produces non-unique solutions. The presence of TFA absorption bands in PDD further complicates the analysis. Therefore spectral reproduction is not attempted here. Taking a cue from traditional analysis far-UV CD it should be possible to reproduce the temperature induced intensity changes at a single wavenumber – 1632 cm^{-1} – that monitors the carbonyl stretches of α-helices. In fact, previous spectroscopic characterization of α-helix unfolding has attempted the same[152]. It is then informative to represent the signals in terms molar extinction coefficient units for comparison with published results.

The signal at 1632 cm^{-1} is dominated by hydrogen-bonded carbonyls in helical conformation (*hhc*) with significant contributions from non-hydrogen bonded helical carbonyls (*nhc*) and to a lesser extent from carbonyls in other conformations that includes turns, loops and coils (*oc*) all of which are length dependent. The structural assignment of the helix lengths and positions are directly taken from far-UV modeling results. The average extinction coefficient for a 'folded' species (ε^F) of helix length l_H can be represented as

$$\varepsilon^F(m,n,T) = \varepsilon_T^{nhc}.nuc + \varepsilon_T^{hhc}.(l_H - nuc)$$

and that of the unfolded species (ε^U) as

$$\varepsilon^U(m,n,T) = \varepsilon_T^{oc}.(N - l_H)$$

where *nuc* stands for the helical nucleus and is fixed at 4 residues, and *N* is the protein length. ε_T^{nhc}, ε_T^{hhc} and ε_T^{oc} are the temperature dependent extinction coefficients of the *nhc, hhc* and *oc* carbonyls,

$$\varepsilon_T^{hhc} = \varepsilon^{hhc} + \varepsilon_T.T \text{ and } \varepsilon_T^{nhc} = \varepsilon_T^{oc} = \varepsilon^{nhc,oc} + \varepsilon_T.T$$

where ε_T is temperature slope and was assumed to be identical for the different conformations. The changes in the overall extinction coefficient can then be written similar to the equations 8.5 and 8.6.

The result of a 3-parameter fit is shown in Figure 8.4A (red curve) with the following final parameters: ε^{hhc} = 630.0 (± 9.4) and $\varepsilon^{hhc,oc}$ = 419.4 (± 52.6) in units of M^{-1} cm^{-1} per peptide carbonyl, and ε_T = -0.61 (± 0.15) M^{-1} cm^{-1} K^{-1} per carbonyl. The value of ε_T^{hhc} at 298 K is ~447 M^{-1} cm^{-1} and is comparable to the 460 M^{-1} cm^{-1} estimated by Trushina and co-workers[153], thus validating the assumptions employed here. The overall features of the unfolding curve are well reproduced even in the absence of the temperature slope except for the pre- and post-transition baselines necessitating its need (gray curve). What factors contribute to the temperature dependence in ε? It is a well-known fact that the frequency of carbonyl motions is temperature dependent; it shifts to higher frequencies due to decreased hydrogen bonding ability[145]. In PDD, the frequency change is not evident as the helices and carbonyls are already solvent exposed even at the lowest temperature. Therefore, the degree and strength of hydrogen-bonding is closely coupled to the solvent vibrational modes that increase with temperature. This in turn reduces the hydrogen-bonding ability of the carbonyls with the N-H backbone and their alignment and hence the

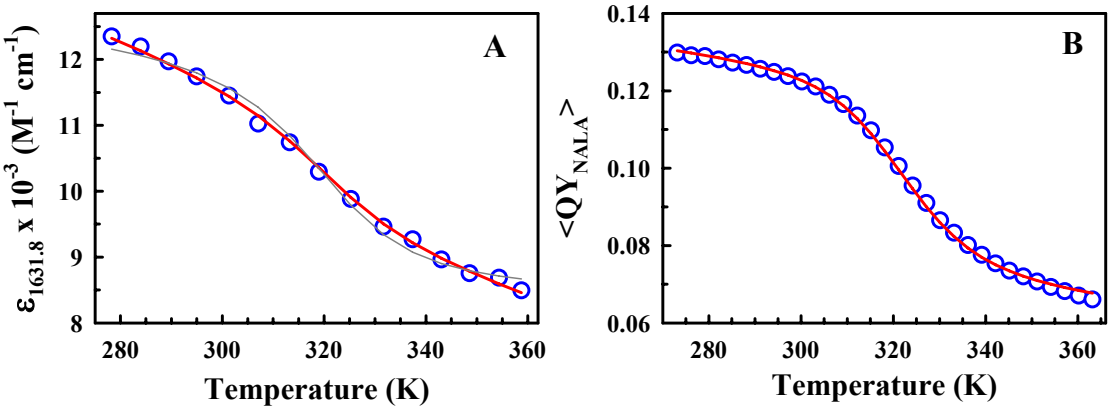


Figure 8.4 A) Fit to the FTIR signal at 1631.8 cm^{-1} assuming a temperature dependent (red) and independent (light gray) extinction coefficient. B) Model fit to the naphthyl quantum yield changes.

absorption at 1632 cm^{-1}. In proteins that fold downhill or over marginal barriers, there is an additional effect of the helix lengths themselves changing with temperature that in turn decreases the alignment of the dipoles in a temperature-dependent fashion. All of these factors combine to produce a net effect on ε that is explained quantitatively here. The temperature dependence changes the ε by ~14% in going from 273 – 373 K providing the first direct estimate of this quantity decoupling it from changes in helix length.

8.3.4 NALA QY

The QY of NALA attached to the protein (at the lowest temperature) is higher than the free dye suggesting transient interactions with the structure. It is referred to as transient because the C-terminus following helix 2 is unstructured in the NMR structure. Since NALA is tagged to the C-terminus, the most probable region for such an interaction is the structured region of helix 2. Specifically, there is a large hydrophobic patch with the sequence Ala-Phe-Leu-Ala corresponding to the last turn

of the helix 2. Interaction with this region would shield it from the surrounding polar environment thus stimulating the fluorescence. Melting of the hydrophobic patch due to the progressive unraveling of the helix 2 would weaken this interaction and thus the QY should approach that of the free dye. However, the QY at the highest temperature (0.066) is significantly smaller than the free dye (0.085), possibly due to quenching in the unfolded state. Furthermore, there temperature dependent quantum yield data shows a minor pre- and post-transition slope. This intrinsic temperature dependence is the result of non-radiative transitions from the 1st singlet to the ground state. The probability of such a transition increases with temperature mainly due to increased collision with solvent molecules. The observed slopes thus need not have a structural origin though the higher temperature slopes are more than that expected from intrinsic temperature dependence alone (see below).

The effects discussed above were modeled as:

$$QY^{F}(m,n,T) = QY_{o}(T) \quad \text{when helix 2 is structured}$$

and

$$QY^{U}(m,n,T) = \frac{QY_{o}(T)}{1 + r_{solv}} \quad \text{otherwise}$$

where

$$QY_{o}(T) = QY + a \times T$$

$QY_{o}(T)$ accounts for the intrinsic temperature dependent quantum yield for all the species. The slope of this dependence (a) was calculated from experiments on free dye and was fixed to -1.6955 x 10^{-4} K^{-1}. r_{solv} is the product between the rate of quenching and the intrinsic life time of the fluorophore. Therefore, the model required

two parameters: QY and r_{solv}. The changes in QY were calculated from expressions similar to equations 8.5 and 8.6. The model reproduced the observed changes in QY very well (Figure 8.4B) resulting in QY = 0.18 (± 0.01) and r_{solv} = 0.86 (± 0.02). The fact that $r_{solv} >> 0$ indicates that there is significant perturbation of the quantum yield in the unfolded state/species.

8.3.5 End-to-end Distance Changes

The main conclusion from modeling the temperature effects on far- and near-UV CD spectra was that the helices unravel gradually from the ends. This provided a simple yet physical way to model the changes in end-to-end distance. The <r> was assumed to be constant and equal to r_F when the protein is completely folded, i.e. when the residues 5 to 39 are structured. As the helices melt, the end-to-end distance was assumed to increase linearly and in proportion to the number of unwound residues (n_U) from either ends,

$$r(m,n) = r_F + r_n n_U$$

Species that have both helices unstructured were considered to be unfolded. The unfolded state of PDD becomes more compact at higher temperatures as evident from the steep negative post-transition slopes at pH 7.0 and the general behavior of pH 3.0 data. In view of this observation, the end-to-end distance of the unfolded state(s) was represented as

$$r(m,n) = r_U + r_T T$$

where r_T is the temperature dependence and r_U is the reference distance. Modeling the end-to-end changes therefore required 4 parameters: r_F, r_n, r_U and r_T. The changes in end-to-end distance were then calculated using an expression similar to equation 8.5.

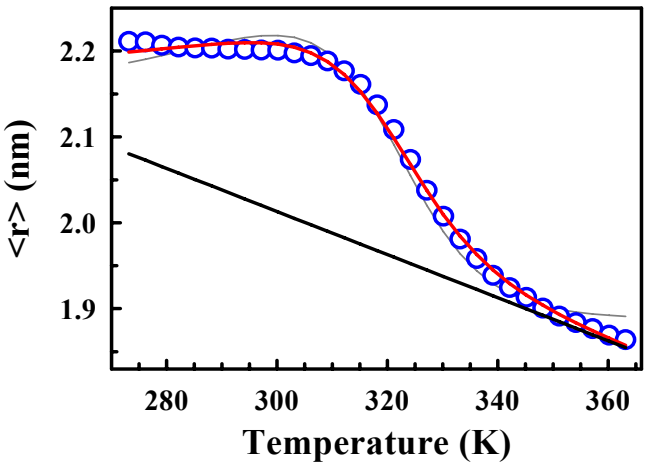


Figure 8.5 End-to-end distance changes modeled assuming a temperature dependent (red curve) and independent (light gray line) unfolded state. The unfolded state baseline is represented in black.

Figure 8.5 plots the fit (red curve) with the final parameters: r_F = 2.18 (± 0.01) nm, r_U = 2.71 (± 0.15) nm, r_n = 0.020 (± 0.006) nm and r_T = -0.0024 (± 0.0004) nm K^{-1}. The agreement with between the data and fit is very good. The end-to-end distance of the folded state predicted by the model is similar to the lowest temperature point and that calculated from the NMR structure[137] (~2.6 nm). The slope r_T is also similar to that calculated from the high temperature points of pH 3.0 data alone. The decrease in average end-to-end distance of the unfolded states with temperature (negative r_T) compensates for the increase upon unraveling (positive r_n), thus producing an apparent baseline at lower temperatures. This effect is illustrated in Figure 8.5, where the gray curve was computed without assuming any temperature dependence on the unfolded states. Such a calculation produces an increase in the end-to-end distances at lower temperatures together with a flat post-transition baseline. These two observations further validate the need for a temperature dependent phenomenological slope on the unfolded state.

However, the discretization of end-to-end distances and a narrow dynamic range (~4 Å) effectively result in a small magnitude of r_n. This suggests that

distributions of distances have to be employed for statistical systems like proteins as earlier discussed. But, the limited number of species employed by this model, the lack of corresponding structural information on the unfolded segments and the non-availability of alternate models that directly characterize the temperature dependent distance distributions preclude such an analysis.

8.4 Analysis of IR T-jump Kinetics

The 2-dimensional probabilities generated from the structure-based statistical mechanical model were projected onto a single reaction co-ordinate – the # of structured residues (see Figure 8.2D for example). This enabled performing diffusive kinetic calculations on a simple one-dimensional surface as opposed to a more complex 2D treatment. The details of the computation are discussed in Chapter 5 with the only difference being the use of just two parameters $E_{a,res}$ and k_0 for a complete description, as the free energy surface is already known. The shape of the temperature versus relaxation rate plot was reproduced very well by this calculation (fit not shown). However, the maximum of the amplitude was over-estimated by ~5-10 K for various approximations of the signal that included linear, sigmoidal and step functions. This suggests that the projection of the two-dimensional probabilities onto this specific reaction co-ordinate fails to reproduce the average changes in the IR signal.

To obtain a reasonable fit for both the amplitude and the rates, the data was analyzed by the DM model described in Chapter 5. The signal was approximated as a step function changing from 0 to 1 at a nativeness value of 0.65 (Figure 8.6C black

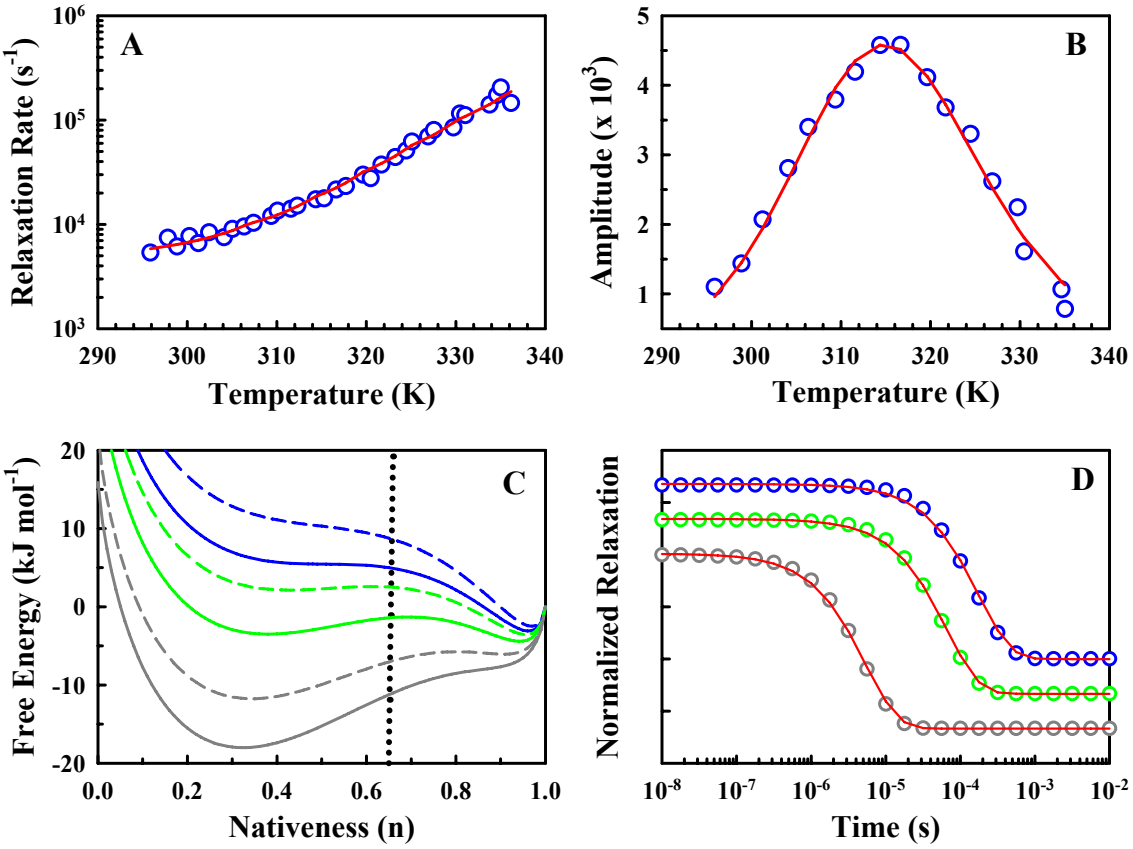


Figure 8.6 A & B) Fits (red curve) to the kinetic relaxation rates and amplitude. C) Calculated free energy profiles before (dashed lines) and after (continuous lines) a T-jump of 10 K for final temperatures of ~296 K (blue), ~312 K (green) and ~336 K (dark gray), respectively. The assumed signal is represented by the dotted black line. D) The relaxation traces at the same temperatures in panel C together with single-exponential fits.

dotted line). Previously, the stabilization energy per residue had to be fixed to specific values for a precise reproduction of the experimental apparent T_ms from thermodynamic measurements (Chapter 5). The availability of amplitude information in the case of PDD provides a more rigorous constraint on the thermodynamics thus enabling ΔH^0_{res} to be used as a floating parameter. More importantly, this treatment gives an independent estimate of the apparent kinetic midpoint temperature. The fitted temperature dependent relaxation rate and amplitude using a ΔC_p^{res} = 10 J K^{-1}

mol^{-1}, $\Delta S_{res}^{n=0}$ = 16.5 J mol^{-1} K^{-1} per residue (at 385 K) and a $k_{\Delta C_p}$ = 4.3 is shown in panels 8.6A and 8.6B. The striking agreement indicates that model clearly reproduces the overall behavior of the system with the following final parameters: $k_{\Delta H}$ = 1.83, ΔH_{res}^{0} = 5.27 kJ mol^{-1} at the reference temperature of 316 K (maximum of the amplitude), k_0 = $10^{12.64}$ and $E_{a,res}$ = 1.35 kJ mol^{-1}. The T_m, defined as the temperature at which the probability weighted nativeness is equal to (n_U + n_F)/2 is calculated to be ~312 K. Interestingly, this value is ~9-10 K lower than that estimated from thermodynamic analysis of DSC profiles. The large discrepancy between the estimated apparent T_ms from kinetic and thermodynamics is an evidence to the non-two-state nature of the transition. The differences in T_ms were also apparent in the set of 9 fast-folding proteins analyzed in Chapter 5. However, the lack of amplitude information for these proteins prevented a more detailed analysis as in principle the difference between kinetic and thermodynamic T_ms can be used as a scale to test the validity of two-state hypothesis. The $E_{a,res}$ of 1.35 kJ mol^{-1} is higher than the value of ~1 kJ mol^{-1} estimated from the analysis presented in Chapter 5, suggestive of a significantly rough free energy surface in PDD.

Figure 8.6C plots the generated free energy profiles before (dashed lines) and after (continuous lines) a 10 K jump to the final experimental temperatures of ~296 K (blue), ~312 K (green) and ~336 K (gray). The profiles are downhill (zero or negative barriers) at both higher and lower temperatures with a folding barrier height of ~2.2 kJ mol^{-1} at the kinetic T_m (ΔC_p^{res} = 10 J mol^{-1} K^{-1}). This value is in agreement with the estimates from DSC analysis and is insensitive to the typical $k_{\Delta C_p}$ values of 2

- 4.5. However, the barrier height at the T_m does increase successively from ~1.3 kJ mol^{-1} for a ΔC_p^{res} = 0 J mol^{-1} K^{-1} to ~6.3 kJ mol^{-1} for a ΔC_p^{res} of 30 J mol^{-1} K^{-1}. The barrier heights at 298 K also increase from being downhill to ~4.2 kJ mol^{-1} for ΔC_p^{res} in the range of 0 - 30 J mol^{-1} K^{-1}. Figure 8.6D plots the simulated kinetic relaxations for the three temperatures and the corresponding single-exponential fits (red curve). It is apparent that single-exponential functions are sufficient to describe the kinetics even at temperatures in which the protein folds downhill, corroborating the earlier theoretical and experimental studies (see Introduction).

8.5 ΔC_p, Barrier Height and D_{eff} of PDD

8.5.1 Apparent T_m

It is not surprising for proteins that fold downhill or over small barriers to report on different apparent T_ms and barrier heights when monitored by different techniques[16,47]. Probe dependent relaxation kinetics has been reported for mutants of lambda repressor that folds over marginal barrier[58]. Moreover, the shape of the temperature dependent relaxation rates for villin variants from fluorescence and IR are drastically different (see Figure 5.7). These observations suggest that the computed barrier height will ultimately depend on the ability to extract precise probability densities from the structural features perceived by the spectroscopic probe. This is further compounded by the assumption that a single reaction-coordinate is sufficient to describe the kinetics and thermodynamics of protein folding. In such cases the barrier heights will also depend on the particular reaction

co-ordinates employed and the various approximations that go with modeling experimental data.

The above limitation in calculating precise barrier heights and T_ms is true in the case of PDD. Analysis of calorimetric data using three different reaction co-ordinates produces thermodynamic T_ms in the range of 321-326 K while a characterization of IR kinetics reveals an apparent kinetic T_m of ~312 K. The significantly lower T_m reported by the IR kinetic analysis is consistent with the nature of this spectroscopic probe as it monitors local structural features, i.e. the changes in vibrational frequency arising out of H-bonding in an alpha helix that spans just 5 residues. This is in contrast to DSC that senses the total changes in heat capacity of the entire system. The spread of T_ms and the trends are already evident in model-free first derivative analysis of raw experimental data that predicts apparent T_ms in the range of 316-325 K with the estimates from FTIR data being the lowest. Furthermore, the definition of a T_m for proteins that fold downhill/marginal barriers is not as straightforward as characterizing a two-state system. This is because of the significant contributions to the dynamics and thermodynamics from sub-ensembles with varying degrees of structure in contrast to two-state systems whose properties are governed by just two ensembles. Given these considerations, an apparent T_m of ~320 K which is an average estimate from different probes and computational treatments seems to be appropriate for PDD.

8.5.2 Heat Capacity Change and Barrier Height

What is the barrier height of PDD at this temperature? The answer to this question relies on the estimates of heat capacity change upon unfolding arising out of

solvation (ΔC_p^{res}). For larger two-state like proteins, this value is estimated to be in the range of 50-58 J mol^{-1} K^{-1}. But, experimental observations do suggest a much smaller value in the range of 0 - 20 J mol^{-1} K^{-1} (Chapter 7). Results from computational calculations and evolutionary arguments are also consistent with this observation:

a) The statistical mechanical model that directly incorporates solvation effects as arising due to changes in accessible surface area upon unfolding predicts a ΔC_p^{res} of just 10 J K^{-1} mol^{-1}.

b) The DM model of protein folding produces significantly worse fits to the DSC profiles for values of ΔC_p^{res} > 20 J mol^{-1} K^{-1}.

c) Protein domains from thermophilic and hyper-thermophilic organisms have to maintain a sufficient thermodynamic stability at high growth temperatures to be functional. Reduction in ΔC_p^{res} is known to be one of the more common mechanisms to achieve higher stability as this broadens temperature stability curve thus maximizing the range of temperatures at which the protein can remain 'folded'[154-157]. The double perturbation analysis of BBL from *Escherichia coli* (a mesophile) results in a significant cold-denaturation at high denaturant concentrations with an average ΔC_p^{res} of 30 – 35 J K^{-1} mol^{-1}. PDD, from the thermophilic *Bacillus stearothermophilus* is then bound to have ΔC_p^{res} smaller than this estimate.

The above arguments propose that ΔC_p^{res} for PDD probably lies within the range of 0 – 20 J mol^{-1} K^{-1}. This would then translate to a predicted barrier height between 0.15

– 4 kJ mol^{-1} at the apparent T_m of 320 K. At 298 K, the barrier height should span from zero (downhill) to ~1.4 kJ mol^{-1}. It is important to note that the estimated barrier heights are not significant at 298 K as the thermal energy (RT) is ~2.5 kJ mol^{-1} at these temperatures. Therefore, PDD folds downhill at room temperatures while crossing a marginal barrier of at most 4 kJ mol^{-1} around 320 K.

8.5.3 Effective Diffusion Coefficient

Using the calculated range of barrier heights it is then possible to estimate the limits for the effective diffusion coefficients (D_{eff}) assuming a simple transition state like expression for the temperature dependence of rates (equation 2.24). This renders a D_{eff} of ~1/(147 μs) – 1/(84 μs) at 298 K and a value between 1/(66 μs) – 1/(15 μs) at 320 K. The D_{eff} at ~336 K – the typical temperatures at which the T-jump data of fast-folding proteins are reported - is ~1/(6 μs) that is within range of numbers predicted in Chapter 5 (Table 5.1). Moreover, the upper estimates for the D_{eff} are of similar magnitude to those calculated before. The significantly smaller lower limits and the relatively larger activation energy (~1.35 kJ mol^{-1} per residue) for D_{eff} indicate that the free energy surface of PDD gets progressively rougher with decreasing temperatures, consistent with theory.

Intriguingly, the mesophilic homolog BBL folds ~7-8 times faster in the same range of temperatures (data not shown). The slower folding observed in PDD at the T_m is due to the larger barrier height compared to BBL that folds globally downhill. But at 298 K, the various models predict downhill folding profiles for PDD. This results in a smaller diffusion coefficient for PDD compared to BBL at 298 K (1/16 μs)). What is the reason for this discrepancy? Comparing the electrostatic potential

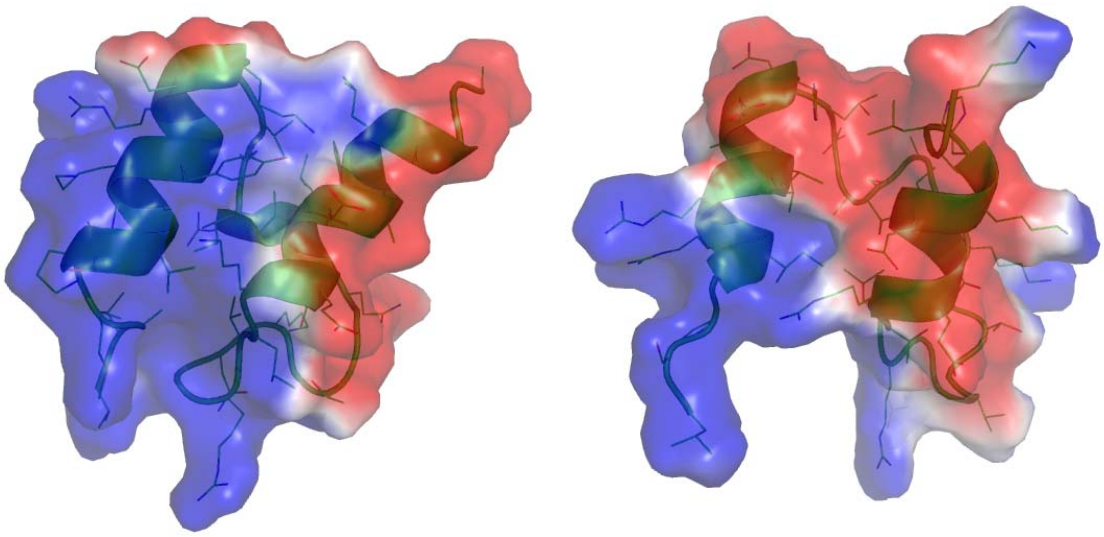


Figure 8.7 Electrostatic potential maps of PDD (left) and BBL (right) calculated using APBS[158] and plotted with PyMol (http://www.pymol.org).

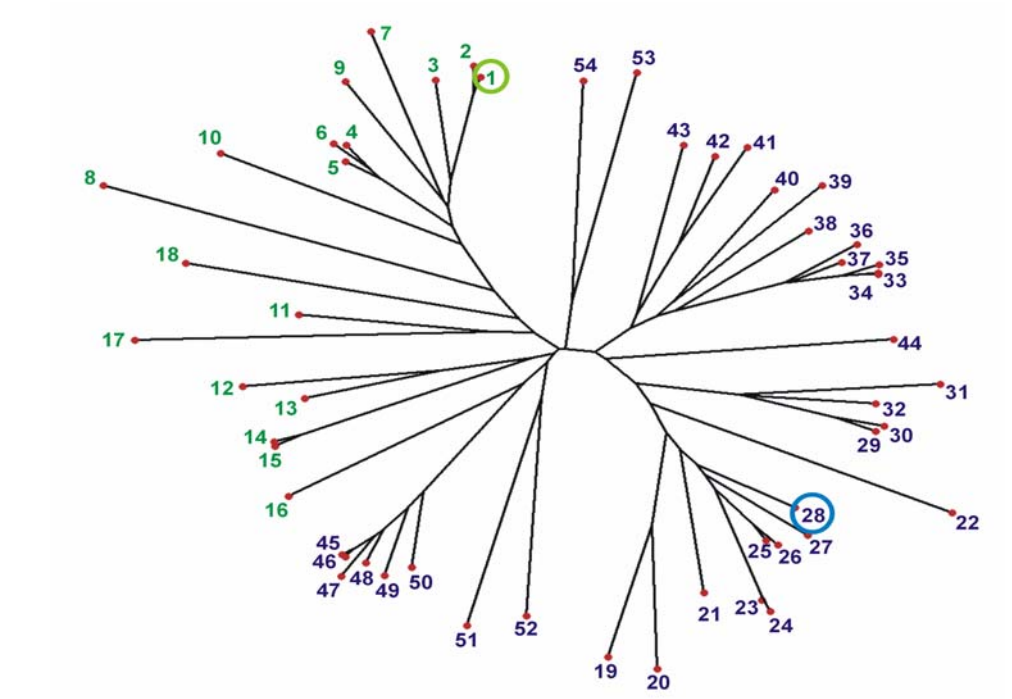
energy surfaces for these proteins (Figure 8.7), it is clear that the charges in PDD are unfavorably placed with positive charges on one face of the protein and negative charges on the other. In other words, the system is highly 'frustrated' with a propensity to form a number of non-native interactions with oppositely charged segments farther along the sequence. In fact, relieving the electrostatic repulsion by mutations has been shown to increase the stability of this protein by Raleigh and co-workers[142], though they do not report the kinetic effects. In addition to the electrostatic repulsions, the hydrophobicity of PDD is higher than that of BBL due to the presence of tyrosine and phenyl-alanine that could in principle slow down the rate due to the stickiness of these residues. Though these two observations stand out as possible reasons, there could be other subtle factors at work. This is because of the fact that this domain is not evolutionarily selected for folding or functionality at low temperatures as the optimal growth temperature for *Bacillus stearothermophilus* is ~328 K. Even more interestingly, the relaxation times of both these domains are very

similar ~10 μs at the respective growth temperature of the source organisms, perhaps suggestive of the necessary link between dynamics and function. It would be interesting to see if this observation holds true for other mesophilic-thermophilic pairs.

8.6 Phylogenetic Analysis

These results together with the earlier analysis of BBL indicate that the functional homologs are both downhill folders. To test whether this result is representative of the behavior of the two protein families (2-oxoglutarate and pyruvate dehydrogenase) a simple phylogenetic analysis was carried out. The sequence homologs of BBL and PDD were obtained by querying them against the database of non-redundant protein sequences provided by NCBI, i.e. BLAST (http://130.14.29.110/BLAST/). The resultant dataset of 16 and 38 sequences each were grouped together. For simplicity, only the sequence boundaries defined by the structures of BBL and PDD were considered for further analysis. A multiple alignment of the 54 sequences was performed using CLUSTALX and the distance scores were plotted as an unrooted phylogenetic tree using Phylodraw (Version 0.8, Graphics Application Lab, Pusan National University).

These two enzymes perform biological functions at the core of the glucose metabolism, which are essential for all known organisms. Therefore, these functions must have arisen by very early divergent evolution, and any common trait between the families has withstood an evolutionary process of billions of years. The unrooted phylogenetic tree of all the known sequence members for the two families supports this view (Figure 8.8). Sequences of each family cluster together in the tree. The



1 *Escherichia coli*
2 - *Salmonella enterica*
3 - *Yersinia pestis*
4 - *Vibrio vulnifucus*
5 - *Vibrio parahaemolyticus*
6 - *Vibrio cholerae*
7 - *Haemophilus influenzae*
8 - *Candidatus blochmannia*
9 - *Pasteurella multocida*
10 - *Haemophilus ducreyi*
11 - *Shewanella oneidensis*
12 - *Pseudomonas aeruginosa*
13 - *Pseudomonas putida*
14 - *Streptomyces coelicolor*
15 - *Streptomyces avermitilius*
16 - *Clostridium magnum*
17 - *Pirellula sp.*
18 - *Coxiella burnetti*
19 - *Lactobacillus reuteri*
20 - *Lactobacillus plantarum*
21 - *Enterococcus faecalis*
22 - *Lactobacillus lactis*
23 - *Listeria monocytogenes*
24 - *Listeria innocua*
25 - *Bacillus cereus*
26 - *Bacillus anthracis*
27 - *Bacillus subtilis*
28 - *Bacillus stearothermophilus*
29 - *Staphylococcus aureus*
30 - *Staphylococcus epidermidis*
31 - *Oceanobacillus iheyensis*
32 - *Bacillus halodurans*
33 - *Salmonella enterica*
34 - *Escherichia coli K12*
35 - *Yersinia pestis*
36 - *Haemophilus influenzae*
37 - Pasteurella multocida
38 - Shewanella oneidensis
39 - Coxiella burnetti
40 - Nitrosomonas europaea
41 - Neisseria meningitidis
42 - Chromobacterium violaceum
43 - Bradyrhizobium japonicum
44 - Thermoplasma volcanium
45 - Rattus norvegicus
46 - Mus musculus
47 - Sus scrofa
48 - Homo sapiens
49 - Xenopus laevis
50 - Danio rerio
51 - Brucella melitensis
52 - Streptococcus mutans
53 - Arabidopsis thaliana
54 - Gallus gallus

Figure 8.8 Unrooted tree depicting the sequence space covered in studying the proteins BBL and PDD (shown within green and blue circles). Sequences of the BBL and PDD family are shown in dark green and dark blue, with the corresponding organisms listed below.

phylogenetic distance between members of the same family from very distant organisms (e.g. chordates and archaebacteria in the pyruvate dehydrogenase family) is shorter than the phylogenetic distance between homolog sequences of very close

organisms (e.g. BBL from *Escherichia coli* and PDD from *Bacillus stearothermophilus*). The unrooted tree reveals that BBL and PDD are representative members of the two families from a phylogenetic standpoint. In fact, the sequence homology between these two proteins (i.e. 0.3876) is somewhat lower than the average sequence homology between members of the two families (0.4427). A parsimonious analysis (i.e. two proteins are evolutionary connected by the minimal number of sequence changes) shows that all the proteins from a given family are closer to the representative member of the family than to the representative of the other – 0.5200 and 0.4708, for 2-oxoglutarate and pyruvate dehydrogenase families, respectively. A similar result was obtained upon analyzing 157 sequences that included even the distant homologs of these domains (using PSI-BLAST; data not shown). In other words, BBL and PDD are evolutionary connected only through the primordial ancestor. The implication is that the downhill folding character is conserved in these two protein families. This result supports the molecular rheostat hypothesis because the evolutionary conservation of downhill folding suggests that it is essential for the biological function of these proteins.

8.7 Conclusions

A comprehensive analysis of the equilibrium and kinetic signals indicates that PDD folds downhill at 298 K while crossing a maximum folding barrier of 4 kJ mol^{-1} at ~320 K. This renders a D_{eff} of ~1/(116 ± 32 μs) at 298 K and ~1/(41 ± 26 μs) at 320 K. The ability to accurately reproduce the signals without employing arbitrary baselines provides a direct access to the mechanism of unfolding - the gradual unraveling of the helices followed by the melting of the hydrophobic core.

Evolutionary arguments based on sequence alignment indicate that folding over marginal/negligible barriers should be conserved among the various species. Given the strategic location of PSBDs in the E2 subunit, the conservation of downhill folding behavior suggests that it has an important role in the functioning of pyruvate dehydrogenase and 2-oxoglutarate dehydrogenase multi-enzyme complexes.

9. Perspectives

The energy landscape theory provides an intuitive base to approach the problem of short time scales involved in protein folding while offering a number of experimentally testable predictions. Recently, many of these predictions including downhill folding, small folding barriers and the principle of minimal frustration have been shown to hold good for natural proteins. The work presented here is a step further in this direction highlighting the diffusive nature of the folding process and the resultant complex experimental observations.

However, protein folding has been over-simplified by the widespread use of the chemical two-state model aided by arbitrary baselines and assumption of large barriers. Moving a step away from a chemical treatment to just a one-dimensional free energy surface analyses is shown here to explain a number of apparent paradoxes in protein folding suggesting that physical models are more appropriate in dealing with proteins. In other words, an unbiased analysis of the shapes of experimental signals (for example, the temperature versus relaxation rate plot) is more informative than the ability to individually reproduce the data points. Due to want of techniques that give *a priori* estimates of barrier heights or the pre-exponential, DSC experiments and multi-probe characterization should be a must as they provide 'model-free' tests to statistical nature of the transition. Nevertheless, some proteins are definitely more two-state like than others; but given the number of examples presented here the same can be said of downhill folders as well.

The prevalence of downhill folding also raises important questions. What factors contribute to the plastic nature of these domains and what are the functional

consequences? This could be approached in the future by employing a reverse engineering approach – mutate proteins iteratively to make a two-state folder out of a downhill folding protein. The functionality of the protein can then be tested. Moreover, the sequence of steps involved in this process would provide valuable information on the relation between hydrophobic forces, electrostatics and packing requirements. Such protein engineering experiments though common in the field, have not been tested against the changes in barrier heights upon mutation (a two-state system is always assumed). With the availability of models that could in principle differentiate the various folding scenarios and measure precise effective diffusion coefficients, this now offers an interesting avenue of research to extricate the elusive dynamic and energetic contributions to folding.

In this aspect, the hydrophobic effect is seen to be a dominant force that drives the folding of a protein to a compact structure[159]. But the ability of the variable barrier model to successfully reproduce the DSC thermograms of both downhill and two-state-like proteins without invoking the idea of solvation indicates that the interplay of molecular forces is more subtle than previously thought. In fact, the significance of equilibrium fluctuations – that forms the basis of heat capacity and hence the variable barrier model - in dictating of function of a protein is well-known. Therefore cataloging of proteins based on the barrier heights and the asymmetry factor should be seen as a step forward in connecting mechanistic aspects of folding and the function of a protein. All of these together with the recent 'backbone-based theory' of protein folding[160] indicate that a lot still needs to be done in elucidating the physico-

chemical forces that determine the ensemble of structures that populate at a given denaturational stress.

Given the current expertise to probe nanosecond processes by T-jump experiments and to monitor single diffusing molecules, the recent advances in molecular dynamics simulations that afford exhaustive sampling, and the development of a number of statistical mechanical models to explain experimental data, it should be possible in the near future to develop a 'unified theory' of protein folding.

Bibliography

(1) Anfinsen, C. B. *Science* **1973**, *181*, 223-230.

(2) Moult, J. *Phil. Trans. Royal Soc. London B* **2006**, *361*, 453-458.

(3) Levinthal, C. *J. Chim. Phys.* **1968**, *65*, 44.

(4) Bryngelson, J. D.; Wolynes, P. G. *J. Phys. Chem.* **1989**, *93*, 6902-6915.

(5) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 167-195.

(6) Onuchic, J. N.; LutheySchulten, Z.; Wolynes, P. G. *Ann. Rev. Phys. Chem.* **1997**, *48*, 545-600.

(7) Watters, A. L.; Deka, P.; Corrent, C.; Callender, D.; Varani, G.; Sosnick, T.; Baker, D. *Cell* **2007**, *128*, 613-624.

(8) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. *J. Chem. Phys.* **1996**, *104*, 5860-5868.

(9) Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11311-11316.

(10) Doshi, U.; Muñoz, V. *Chem. Phys.* **2004**, *307*, 129-136.

(11) Muñoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196-199.

(12) Cho, S. S.; Levy, Y.; Wolynes, P. G. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 586-591.

(13) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334-350.

(14) Muñoz, V. *Curr. Opin. Struct. Biol.* **2001**, *11*, 212-216.

(15) Zwanzig, R. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 9801-9804.

(16) Muñoz, V. *Int. J. Quantum Chem.* **2002**, *90*, 1522-1528.

(17) Thompson, P. A.; Muñoz, V.; Jas, G. S.; Henry, E. R.; Eaton, W. A.; Hofrichter, J. *J. Phys. Chem. B* **2000**, *104*, 378-389.

(18) Jones, C. M.; Henry, E. R.; Hu, Y.; Chan, C. K.; Luck, S. D.; Bhuyan, A.; Roder, H.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 11860-11864.

(19) Bieri, O.; Wirz, J.; Hellrung, B.; Schutkowski, M.; Drewello, M.; Kiefhaber, T. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 9597-9601.

(20) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 7220-7225.

(21) Buscaglia, M.; Schuler, B.; Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2003**, *332*, 9-12.

(22) Krieger, F.; Fierz, B.; Bieri, O.; Drewello, M.; Kiefhaber, T. *J. Mol. Biol.* **2003**, *332*, 265-274.

(23) Sadqi, M.; Lapidus, L. J.; Muñoz, V. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 12117-12122.

(24) Pollack, L.; Tate, M. W.; Darnton, N. C.; Knight, J. B.; Gruner, S. M.; Eaton, W. A.; Austin, R. H. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 10115-10117.

(25) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76-88.

(26) Li, M. S.; Klimov, D. K.; Thirumalai, D. *Polymer* **2004**, *45*, 573-579.

(27) Schuler, B.; Lipman, E. A.; Eaton, W. A. *Nature* **2002**, *419*, 743-747.

(28) Muñoz, V. *Ann. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 395-412.

(29) Ikai, A.; Tanford, C. *J. Mol. Biol.* **1973**, *73*, 145-163.

(30) Jackson, S. E.; Fersht, A. R. *Biochemistry* **1991**, *30*, 10428-10435.

(31) Jackson, S. E. *Folding Des.* **1998**, *3*, R81-R91.

(32) Fersht, A. R.; Matouschek, A.; Serrano, L. *J. Mol. Biol.* **1992**, *224*, 771-782.

(33) Hagen, S. J. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 1-4.

(34) Merchant, K. A.; Best, R. B.; Louis, J. M.; Gopich, I. V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1528-1533.

(35) Zimm, B. H.; Bragg, J. K. *Journal of Chemical Physics* **1959**, *31*, 526-535.

(36) Doshi, U. R.; Muñoz, V. *J. Phys. Chem. B* **2004**, *108*, 8497-8506.

(37) Venyaminov, S. Y.; Hedstrom, J. F.; Prendergast, F. G. *Proteins: Struct., Funct., Genet.* **2001**, *45*, 81-89.

(38) Maity, H.; Maity, M.; Krishna, M. M. G.; Mayne, L.; Englander, S. W. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 4741-4746.

(39) Bahar, I.; Wallqvist, A.; Covell, D. G.; Jernigan, R. L. *Biochemistry* **1998**, *37*, 1067-1075.

(40) Chattopadhyay, K.; Saffarian, S.; Elson, E. L.; Frieden, C. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 14171-14176.

(41) Li, H.; Frieden, C. *Biochemistry* **2005**, *44*, 2369-2377.

(42) Li, H.; Frieden, C. *Biochemistry* **2007**, *46*, 4337-4347.

(43) Lakshmikanth, G. S.; Sridevi, K.; Krishnamoorthy, G.; Udgaonkar, J. B. *Nat. Struct. Mol. Biol.* **2001**, *8*, 799-804.

(44) Ding, K. Y.; Louis, J. M.; Gronenborn, A. M. *J. Mol. Biol.* **2004**, *335*, 1299-1307.

(45) Werner, J. H.; Joggerst, R.; Dyer, R. B.; Goodwin, P. M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 11130-11135.

(46) Klimov, D. K.; Thirumalai, D. *J. Comp. Chem.* **2002**, *23*, 161-165.

(47) Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Muñoz, V. *Science* **2002**, *298*, 2191-2195.

(48) Sadqi, M.; Fushman, D.; Muñoz, V. *Nature* **2006**, *442*, 317-321.

(49) Muñoz, V.; Sanchez-Ruiz, J. M. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 17646-17651.

(50) Zuo, G. H.; Wang, J.; Wang, W. *Proteins: Struct., Funct., Bioinf.* **2006**, *63*, 165-173.

(51) Knott, M.; Chan, H. S. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 373-391.

(52) Oliva, F. Y.; Muñoz, V. *J. Am. Chem. Soc.* **2004**, *126*, 8596-8597.

(53) Naganathan, A. N.; Doshi, U.; Fung, A.; Sadqi, M.; Muñoz, V. *Biochemistry* **2006**, *45*, 8466-8475.

(54) Ma, H. R.; Gruebele, M. *J. Comp. Chem.* **2006**, *27*, 125-134.

(55) Yang, W. Y.; Gruebele, M. *Nature* **2003**, *423*, 193-197.

(56) Yang, W. Y.; Gruebele, M. *Biophys. J.* **2004**, *87*, 596-608.

(57) Yang, W. Y.; Gruebele, M. *J. Am. Chem. Soc.* **2004**, *126*, 7758-7759.

(58) Ma, H. R.; Gruebele, M. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 2283-2287.

(59) Liu, F.; Gruebele, M. *J. Mol. Biol.* **2007**, *370*, 574-584.

(60) Hagen, S. J. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 205-217.

(61) Religa, T. L.; Johnson, C. M.; Vu, D. M.; Brewer, S. H.; Dyer, R. B.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 9272-9277.

(62) Akmal, A.; Muñoz, V. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 142-152.

(63) Thirumalai, D. *J. Phys. I* **1995**, *5*, 1457-1467.

(64) Naganathan, A. N.; Muñoz, V. *J. Am. Chem. Soc.* **2005**, *127*, 480-481.

(65) Naganathan, A. N.; Sanchez-Ruiz, J. M.; Muñoz, V. *J. Am. Chem. Soc.* **2005**, *127*, 17970-17971.

(66) Naganathan, A. N.; Doshi, U.; Muñoz, V. *J. Am. Chem. Soc.* **2007**, *129*, 5673-5682.

(67) Ferguson, N.; Schartau, P. J.; Sharpe, T. D.; Sato, S.; Fersht, A. R. *J. Mol. Biol.* **2004**, *344*, 295-301.

(68) Ferguson, N.; Sharpe, T. D.; Schartau, P. J.; Sato, S.; Allen, M. D.; Johnson, C. M.; Rutherford, T. J.; Fersht, A. R. *J. Mol. Biol.* **2005**, *353*, 427-446.

(69) Naganathan, A. N.; Perez-Jimenez, R.; Sanchez-Ruiz, J. M.; Muñoz, V. *Biochemistry* **2005**, *44*, 7435-7449.

(70) Makhatadze, G. I.; Medvedkin, V. N.; Privalov, P. L. *Biopolymers* **1990**, *30*, 1001-1010.

(71) Kholodenko, V.; Freire, E. *Anal. Biochem.* **1999**, *270*, 336-338.

(72) Georgescu, R. E.; Garcia-Mira, M. M.; Tasayco, M. L.; Sanchez-Ruiz, J. M. *Eur. J. Biochem.* **2001**, *268*, 1477-1485.

(73) Freire, E.; Biltonen, R. L. *Biopolymers* **1978**, *17*, 463-479.

(74) Jelesarov, I.; Bosshard, H. R. *J. Mol. Recog.* **1999**, *12*, 3-18.

(75) Frank, H. S.; Evans, M. W. *J. Chem. Phys.* **1945**, *13*, 507.

(76) Silverstein, K. A.; Haymet, A. D. J.; Dill, K. A. *J. Am. Chem. Soc.* **1998**, *118*, 5163-5168.

(77) Myers, J. K.; Pace, C. N.; Scholtz, J. M. *Protein Sci.* **1995**, *4*, 2138-2148.

(78) Aune, K. C.; Tanford, C. *Biochemistry* **1969**, *8*, 4586-4590.

(79) O'Brien, E. P.; Dima, R. I.; Brooks, B.; Thirumalai, D. *J. Am. Chem. Soc.* **2007**, *129*, 7346-7353.

(80) Schellman, J. A. *Biopolymers* **1994**, *34*, 1015-1026.

(81) Knapp, J. A.; Pace, C. N. *Biochemistry* **1974**, *13*, 1289-1294.

(82) Oliveberg, M.; Tan, Y. J.; Fersht, A. R. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 8926-8929.

(83) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985-994.

(84) Makarov, D. E.; Plaxco, K. W. *Prot. Sci.* **2003**, *12*, 17-26.

(85) Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. *Prot. Sci.* **2003**, *12*, 2057-2062.

(86) Ivankov, D. N.; Finkelstein, A. V. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 8942-8944.

(87) Finkelstein, A. V.; Badretdinov, A. Y. *Folding Des.* **1997**, *2*, 115-121.

(88) Wolynes, P. G. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 6170-6175.

(89) Koga, N.; Takada, S. *J. Mol. Biol.* **2001**, *313*, 171-180.

(90) Kouza, M.; Li, M. S.; O'Brien, E. P.; Hu, C. K.; Thirumalai, D. *J. Phys. Chem. A* **2006**, *110*, 671-676.

(91) Spolar, R. S.; Record, M. T. *Science* **1994**, *263*, 777-784.

(92) Taylor, J. W.; Greenfield, N. J.; Wu, B.; Privalov, P. L. **1999**.

(93) Maynard, A. J.; Sharman, G. J.; Searle, M. S. *J. Am. Chem. Soc.* **1998**, *120*, 1996-2007.

(94) Cooper, A. *Biophysical Chemistry* **2000**, *85*, 25-39.

(95) Cooper, A. *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 2740-2741.

(96) Cooper, A. *Prog. Biophys. Mol. Biol.* **1984**, *44*, 181-214.

(97) Robertson, A. D.; Murphy, K. P. *Chem. Rev.* **1997**, *97*, 1251-1267.

(98) Zhou, Y. Q.; Hall, C. K.; Karplus, M. *Protein Sci.* **1999**, *8*, 1064-1074.

(99) Gomez, J.; Hilser, V. J.; Xie, D.; Freire, E. *Proteins: Struct., Funct., Genet.* **1995**, *22*, 404-412.

(100) Knott, M.; Chan, H. S. *Chem. Phys.* **2004**, *307*.

(101) Kaya, H.; Chan, H. S. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 637-661.

(102) Chan, H. S. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 543-571.

(103) Ferguson, N.; Sharpe, T. D.; Johnson, C. M.; Fersht, A. R. *J. Mol. Biol.* **2006**, *356*, 1237-1247.

(104) Garcia-Mira, M. M.; Boehringer, D.; Schmid, F. X. *J. Mol. Biol.* **2004**, *339*, 555-569.

(105) Ferguson, N.; Johnson, C. M.; Macias, M.; Oschkinat, H.; Fersht, A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 13002-13007.

(106) Chiti, F.; Taddei, N.; White, P. M.; Bucciantini, M.; Magherini, F.; Stefani, M.; Dobson, C. M. *Nat. Struct. Biol.* **1999**, *6*, 1005-1009.

(107) Teilum, K.; Thormann, T.; Caterer, N. R.; Poulsen, H. I.; Jensen, P. H.; Knudsen, J.; Kragelund, B. B.; Poulsen, F. M. *Proteins: Struct., Funct., Bioinf.* **2005**, *59*, 80-90.

(108) Gianni, S.; Guydosh, N. R.; Khan, F.; Caldas, T. D.; Mayor, U.; White, G. W. N.; DeMarco, M. L.; Daggett, V.; Fersht, A. R. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 13286-13291.

(109) Nguyen, H.; Jager, M.; Moretto, A.; Gruebele, M.; Kelly, J. W. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 3948-3953.

(110) Jager, M.; Nguyen, H.; Crane, J. C.; Kelly, J. W.; Gruebele, M. *J. Mol. Biol.* **2001**, *311*, 373-393.

(111) Kubelka, J.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2003**, *329*, 625-630.

(112) Brewer, S. H.; Vu, D. M.; Tang, Y. F.; Li, Y.; Franzen, S.; Raleigh, D. P.; Dyer, R. B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16662-16667.

(113) Wang, T.; Zhu, Y. J.; Gai, F. *J. Phys. Chem. B* **2004**, *108*, 3694-3697.

(114) Mayor, U.; Johnson, C. M.; Daggett, V.; Fersht, A. R. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 13518-13522.

(115) Vu, D. M.; Myers, J. K.; Oas, T. G.; Dyer, R. B. *Biochemistry* **2004**, *43*, 3582-3589.

(116) Zhu, Y.; Alonso, D. O. V.; Maki, K.; Huang, C. Y.; Lahr, S. J.; Daggett, V.; Roder, H.; DeGrado, W. F.; Gai, F. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 15486-15491.

(117) Yang, W. Y.; Gruebele, M. *Biochemistry* **2004**, *43*, 13018-13025.

(118) Kim, J.; Keyes, T. *J. Phys. Chem.* **2007**, *111*, 2647-2657.

(119) Lapidus, L. J.; Steinbach, P. J.; Eaton, W. A.; Szabo, A.; Hofrichter, J. *J. Phys. Chem. B* **2002**, *106*, 11628-11640.

(120) Sanchez, I. E.; Kiefhaber, T. *J. Mol. Biol.* **2003**, *327*, 867-884.

(121) Hagen, S. J.; Qiu, L. L.; Pabit, S. A. *J. Phys.: Condens. Matter* **2005**, *17*, S1503-S1514.

(122) Takada, S. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 85-98.

(123) Portman, J. J.; Takada, S.; Wolynes, P. G. *Phys. Rev. Letters* **1998**, *81*, 5237-5240.

(124) Cecconi, F.; Guardiani, C.; Livi., R. *Biophys. J.* **2006**, *91*, 694-704.

(125) Religa, T. L.; Markson, J. S.; Mayor, U.; Freund, S. M. V.; Fersht, A. R. *Nature* **2005**, *437*, 1053-1056.

(126) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2006**, *359*, 546-553.

(127) Nguyen, H.; Jager, M.; Kelly, J. W.; Gruebele, M. *J. Phys. Chem. B* **2005**, *109*, 15182-15186.

(128) Sabelko, J.; Ervin, J.; Gruebele, M. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 6031-6036.

(129) Robien, M. A.; Clore, G. M.; Omichinski, J. G.; Perham, R. N.; Appella, E.; Sakaguchi, K.; Gronenborn, A. M. *Biochemistry* **1992**, *31*, 3463-3471.

(130) Desai, U. R.; Osterhout, J. J.; Klibanov, A. M. *J. Am. Chem. Soc.* **1994**, *116*, 9420-9422.

(131) Griebenow, K.; Klibanov, A. M. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 10969-10976.

(132) Spera, S.; Bax, A. *J. Am. Chem. Soc.* **1991**, *113*, 5490-5492.

(133) Muñoz, V.; Serrano, L. *J. Mol. Biol.* **1995**, *245*, 275-296.

(134) Chen, Y. H.; Yang, J. T.; Chau, K. H. *Biochemistry* **1974**, *13*, 3350-3359.

(135) Scholtz, J. M.; Barrick, D.; York, E. J.; Stewart, J. M.; Baldwin, R. L. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 185-189.

(136) Felitsky, D. J.; Record, M. T. *Biochemistry* **2003**, *42*, 2202-2217.

(137) Kalia, Y. N.; Brocklehurst, S. M.; Hipps, D. S.; Appella, E.; Sakaguchi, K.; Perham, R. N. *J. Mol. Biol.* **1993**, *230*.

(138) Perham, R. N. *Annu. Rev. Biochem.* **2000**, *69*, 961-1004.

(139) Spector, S.; Kuhlman, B.; Fairman, R.; Wong, E.; Boice, J. A.; Raleigh, D. P. *J. Mol. Biol.* **1998**, *276*.

(140) Spector, S.; Raleigh, D. P. *J. Mol. Biol.* **1999**, *293*, 763-768.

(141) Spector, S.; Young, P.; Raleigh, D. P. *Biochemistry* **1999**, *38*, 4128-4136.

(142) Spector, S.; Wang, M.; Carp, S. A.; Robblee, J.; Hendsch, Z. S.; Fairman, R.; Tidor, B.; Raleigh, D. P. *Biochemistry* **2000**, *39*, 872-879.

(143) Chakrabartty, A.; Kortemme, T.; Padmanabhan, S.; Baldwin, R. L. *Biochemistry* **1993**, *32*, 5560-5565.

(144) Williams, S.; Causgrove, T. P.; Gilmanshin, R.; Fang, K. S.; Callender, R. H.; Woodruff, W. H.; Dyer, R. B. *Biochemistry* **1996**, *35*, 691-697.

(145) Manas, E. S.; Getahun, Z.; Wright, W. W.; DeGrado, W. F.; Vanderkooi, J. M. *J. Am. Chem. Soc.* **2000**, *122*, 9883-9890.

(146) Yoder, G.; Pancoska, P.; Keiderling, T. A. *Biochemistry* **1997**, *36*, 15123-15133.

(147) Murphy, K. P.; Privalov, P. L.; Gill, S. J. *Science* **1990**, *247*.

(148) Athawale, M. V.; Goel, G.; Ghosh, T.; Truskett, T. M.; Garde, S. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 733-738.

(149) Zagrovic, B.; Snow, C. D.; Khaliq, S.; Shirts, M. R.; Pande, V. S. *J. Mol. Biol.* **2002**, *323*, 153-164.

(150) Fitzkee, N. C.; Rose, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12497-12502.

(151) Makhatadze, G. I.; Privalov, P. L. *J. Mol. Biol.* **1990**, *213*, 375-384.

(152) Graff, D. K.; PastranaRios, B.; Venyaminov, S. Y.; Prendergast, F. G. *J. Am. Chem. Soc.* **1997**, *119*, 11282-11294.

(153) Chirgadze, Y. N.; Fedorov, O. V.; Trushina, N. P. *Biopolymers* **1975**, *14*, 679-694.

(154) McCrary, B. S.; Edmondson, S. P.; Shriver, J. W. *J. Mol. Biol.* **1996**, *264*, 784-805.

(155) Motono, C.; Oshima, T.; Yamagishi, A. *Protein Engineering* **2001**, *14*, 961-966.

(156) Kumar, S.; Nussinov, R. *Biophys. Chem.* **2004**, *111*, 235-246.

(157) Razvi, A.; Scholtz, J. M. *Prot. Sci.* **2006**, *15*, 1569-1578.

(158) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037-10041.

(159) Dill, K. A. *Biochemistry* **1990**, *29*, 7133-7155.

(160) Rose, G. D.; Fleming, P. J.; Banavar, J. R.; Maritan, A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 16623-16633.