doi:10.1016/j.jmb.2006.06.041 J. Mol. Biol. (2006) 361, 755–770

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Small Proteins Fold Through Transition States With Native-like Topologies

Adarsh D. Pandit[1], Abhishek Jha[2], Karl F. Freed[2] and Tobin R. Sosnick[1]*

[1]*Department of Biochemistry and Molecular Biology, and the Institute for Biophysical Dynamics, University of Chicago, 929 E. 57th St. Chicago, IL 60637, USA*

[2]*James Franck Institute and Department of Chemistry, University of Chicago, 929 E. 57th St. Chicago, IL 60637 USA*

**Corresponding author*

The folding pathway of common-type acyl phosphatase (ctAcP) is characterized using ψ-analysis, which identifies specific chain–chain contacts using bi-histidine (biHis) metal-ion binding sites. In the transition state ensemble (TSE), the majority of the protein is structured with a near-native topology, only lacking one β-strand and an α-helix. ψ-Values are zero or unity for all sites except one at the amino terminus of helix H2. This fractional ψ-value remains unchanged when three metal ions of differing coordination geometries are used, indicating this end of the helix experiences microscopic heterogeneity through fraying in the TSE. Ubiquitin, the other globular protein characterized using ψ-analysis, also exhibits a single consensus TSE structure. Hence, the TSE of both proteins have converged to a single configuration, albeit one that contains some fraying at the periphery. Models of the TSE of both proteins are created using all-atom Langevin dynamics simulations using distance constraints derived from the experimental ψ-values. For both proteins, the relative contact order of the TS models is ~80% of the native value. This shared value viewed in the context of the known correlation between contact order and folding rates, suggests that other proteins will have a similarly high fraction of the native contact order. This constraint greatly limits the range of possible configurations at the rate-limiting step.

Keywords: psi-analysis; phi-analysis; protein folding; acylphosphatase

Introduction

The characterization of the rate-limiting step is critical to defining the mechanism of any chemical reaction. Consequently, a central goal in protein folding studies is to characterize the transition state ensemble (TSE). TSEs have been described as polarized,[1–9] or an expanded version of the native state.[10,11] Even the existence of a TS for this macromolecular reaction engenders controversy.[12]

The characterization of the TSE includes the identification of the multiplicity of pathways leading up to it, and the degree of structural diversity in the ensemble itself.[13–20] Even for a singular TS, the alternative pathways leading to it can have a different folding order for the structural elements found in the TS. The existence of alternative TS structures implies that there are multiple pathways.

A meaningful discussion of whether a TSE is considered to be heterogeneous with alternative forms depends on the resolution at which two such structures differ within the ensemble. Two structures may be essentially the same, differing only by local, small-scale fluctuations such as helical or sheet fraying. This possibility is likely to be a general phenomenon in the TS because such fraying is often observed in folding intermediates[21] and native states.

This type of local or "microscopic heterogeneity" should be distinguished from a situation where

Abbreviations used: biHis, bi-histidine divalent metal ion binding site; ctAcP, common-type acyl phosphatase; GdmCl, guanidinium chloride; LRO, long-range order; RCO, relative contact distance; TSE, transition state ensemble; $\psi_o^{2\ pt}$, ψ_o-value obtained from denaturant chevrons in the absence and presence of metal ions; $\psi_o^{Leffler}$, ψ_o-value obtained from the Leffler plot; PWT, pseudo wild-type; $\Delta\Delta G_{mut}$, change in equilibrium stability as a result of mutation; $\Delta\Delta G_{eq}$, stabilization due to metal binding.

E-mail address of the corresponding author: trsosnic@uchicago.edu

members of the TSE are significantly dissimilar, for example, containing a different subset of structural elements. A further, more rigorous distinction should be made between a TSE that shares a significant number of common elements from a TSE where subpopulations are structurally disjoint (Figure 1). We believe that this latter distinction is most appropriate when discussing whether TS heterogeneity exists in protein folding.

Multiple pathways and TS heterogeneity are particularly difficult to discern with small globular proteins that fold in a kinetically two-state reaction (U↔N) because intermediates do not significantly populate.[22–25] Hence, the intermediates leading up to the TS, and the TS itself, cannot be readily studied by the usual structural methods.

Thus far, little experimental evidence supports the existence of TS heterogeneity according to our more stringent definition.[19,26] A notable exception is the dimeric GCN4 coiled coil. Multiple TSs are found with helical structures located at different regions along the length of the molecule. Its cross-linked counterpart, however, does not exhibit TS heterogeneity, nucleating only at the tethered end.[20,27] Ubiquitin (Ub) also folds through a single TS. Here, the TSE is extensive, containing four β-strands and a helix, although regions on the periphery experience some microscopic heterogeneity.

The characterization of the TSE of the GCN4 coiled coil and Ub was accomplished using a relatively new method that we developed termed ψ-analysis.[26–28] In this counterpart to mutational ϕ-analysis,[29–31] engineered bi-histidine (biHis) metal ion binding sites are introduced at specific positions on the protein surface to stabilize secondary and tertiary structures. An increase in the metal ion concentration stabilizes the interaction between the two histidine partners in a continuous fashion. As a result, this method quantitatively evaluates the metal-induced stabilization of the TS relative to the native state, as represented by the ψ-value.

The translation of a measured ψ-value to structure formation is straightforward because the positions probed are proximal in the native structure. Hence, the method is particularly well-suited to identify chain-chain contacts that define the topology and structure of TSEs. The mutational counterpart, ϕ-analysis, reports on the energetic influence of side-chain alteration and can under-report the amount of structure in the TS.[26–28]

In order to define the TS structure and determine the degree of TS heterogeneity possible in protein folding, we apply ψ-analysis to ctAcP (Figure 2(a)).[32] This 98-residue protein is composed of two helices packed against a five-stranded antiparallel sheet.[33,34] We chose this protein as it is one-third larger than Ub and more topologically complex. Therefore, it is more likely to have a TSE with structurally disjoint subpopulations.

However, we find that the TSE of ctAcP is comprised of a single, large structure containing strands β1–β4 and helix H2. The majority of the ψ-values are near zero or unity, indicating that only a minimal degree of heterogeneity is present in the TSE. The sole exception is a ψ-value of ~0.35, observed with multiple metal ions at the amino terminus of helix 2. This result is consistent with a small amount of helical fraying in the TS ensemble, and hence, with the ψ-value representing the fraction of the TSE having the metal ion binding site residues in a native-like geometry. In aggregate, however, the folding of this protein provides no evidence for

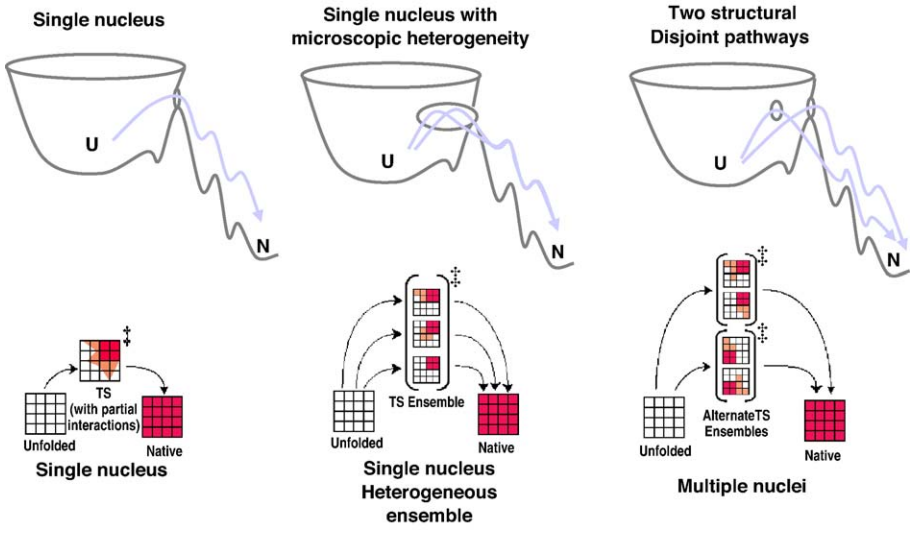

Figure 1. Classes of transition state heterogeneity. Folding may occur *via* a singular essential TS nucleus with some partially formed interactions (left). Here, some residues are completely unfolded (white boxes), or absolutely required in a given nucleus (red boxes), while others may have a fractional side-chain interaction in the TS (orange boxes). Folding may also occur through a structurally heterogeneous ensemble where some residues are critical for the folding nucleus (red boxes) but different groups of structures may exist at the TS (middle). Here the TS can exhibit "microscopic heterogeneity" (orange boxes), e.g. helical fraying, but a conserved folding nucleus. Alternatively, the nuclei can be structurally disjoint (right), each with a diverse set of necessary structures comprising distinct nuclei (red boxes).

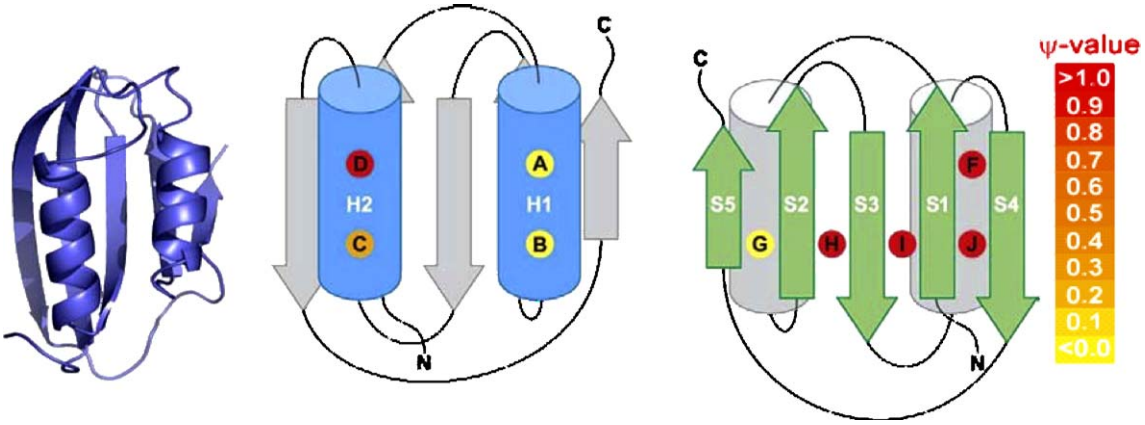

Figure 2. Schematic representation of biHis sites in ctAcP and ψ-values. Crystal structure of ctAcP (2acy.pdb)[98] and a schematized representation of ψ-analysis results at the biHis sites. The figures depict two views of the protein structure, rotated by 180°, highlighting the β-sheet network (β1–5) and α-helices (H1 and H2). Each circle represents a biHis metal binding site across two secondary structure elements. Sites are color-coded to represent resultant ψ_o-values.

structurally disjoint populations in the TSE. Implicit solvent, all-atom Langevin dynamics simulations are conducted utilizing the ψ-values as constraints on ctAcP and Ub. Both resulting TSEs have a relative contact order[35] that is ~80% of the native value, suggesting that the TSE of many proteins also will have a similar percentage of their native topology. The implications of this conclusion are discussed.

ψ-Analysis

ψ-Analysis uses engineered biHis sites to probe the fraction of the native metal ion binding energy realized in the TS. The kinetic response as a function of metal ion concentration reports on the degree to which the biHis site is present in the TSE (see Supplementary Data for detailed treatment). In a manner analogous to ϕ-analysis performed using single point mutations, the kinetic response due to metal binding can be obtained from the denaturant dependence of folding rates ("chevron analysis") at zero and high metal ion concentrations (Figure 3).

When side-chain substitution or metal binding only affects the unfolding rate, k_u, and not the free energy of the TS relative to the unfolded state, the structure probed is absent in the TSE, and the corresponding $\phi^{mutation}$ or ψ^{metal}-value is zero. Conversely, when the perturbation only affects the folding rate, k_f, the structure probed is likely to be native-like in the TSE and the associated ϕ or ψ-value is unity. When both the folding and unfolding arms shift, the ϕ or ψ-value is fractional. In both methods, the origin of a fractional value can be challenging to discern. Fractional ϕ-values may be due to either partial structure formation in the TS or the presence of multiple, distinct TS structures.[13,15,20,27,36–39] A fractional ψ-value indicates that the biHis site is either native-like in a subfraction of the TSE, or has non-native binding affinity in the entire TSE (e.g. a distorted site with less favorable binding geometry, or a flexible site that must be restricted prior to ion binding), or some combination thereof[26,28] (D. Goldenberg, private communication).

ψ-Analysis has the powerful capability of generating a large quantity of high quality kinetic data to accurately probe the degree to which a particular binding site is formed in the TSE. For each biHis variant, dozens of folding rates can be measured at increasing concentrations of metal ions. The binding of increasing concentrations of ions to the biHis site results in a near-continuous increase in the stability of TS structures that contain the binding site. Hence, the stability is perturbed yet accomplished in an isosteric and isochemical manner. The resulting series of data can be justifiably combined, which may be inappropriate in traditional mutation studies where the perturbation can arise for multiple sources including changes in backbone propensities as well as indeterminate non-local interactions.

The ψ-analysis data can be presented as the change in activation free energy relative to the change in the metal-induced stability, $\Delta\Delta G_f^{\ddagger}$ / $\Delta\Delta G_{eq}$, to produce a Leffler plot[40] (Figure 3, Supplementary Figure 3A). If the biHis site is formed in the TSE, metal binding increases its stability and folding rates increase. The associated Leffler plot has a positive slope as both $\Delta\Delta G_f^{\ddagger}$ and $\Delta\Delta G_{eq}$ increase (Supplementary Figure 3B).

The starting point in the detailed interpretation of the Leffler plot is to fit the data to a model with a single free parameter, ψ_o, which is the slope at the origin in the absence of metal:

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

Along the curve, the instantaneous slope, or ψ-value, increases with additional binding energy as the fraction of the TSE with the biHis site grows. The instantaneous slope at any point on the curve as a function of binding stability (Supplementary Figure 3A) is given by:

$$\psi = \frac{\partial\Delta\Delta G_f^{\ddagger}}{\partial\Delta\Delta G_{eq}} = \frac{\psi_o}{(1-\psi_o)e^{-\Delta\Delta G_{eq}/RT} + \psi_o} \qquad (2)$$

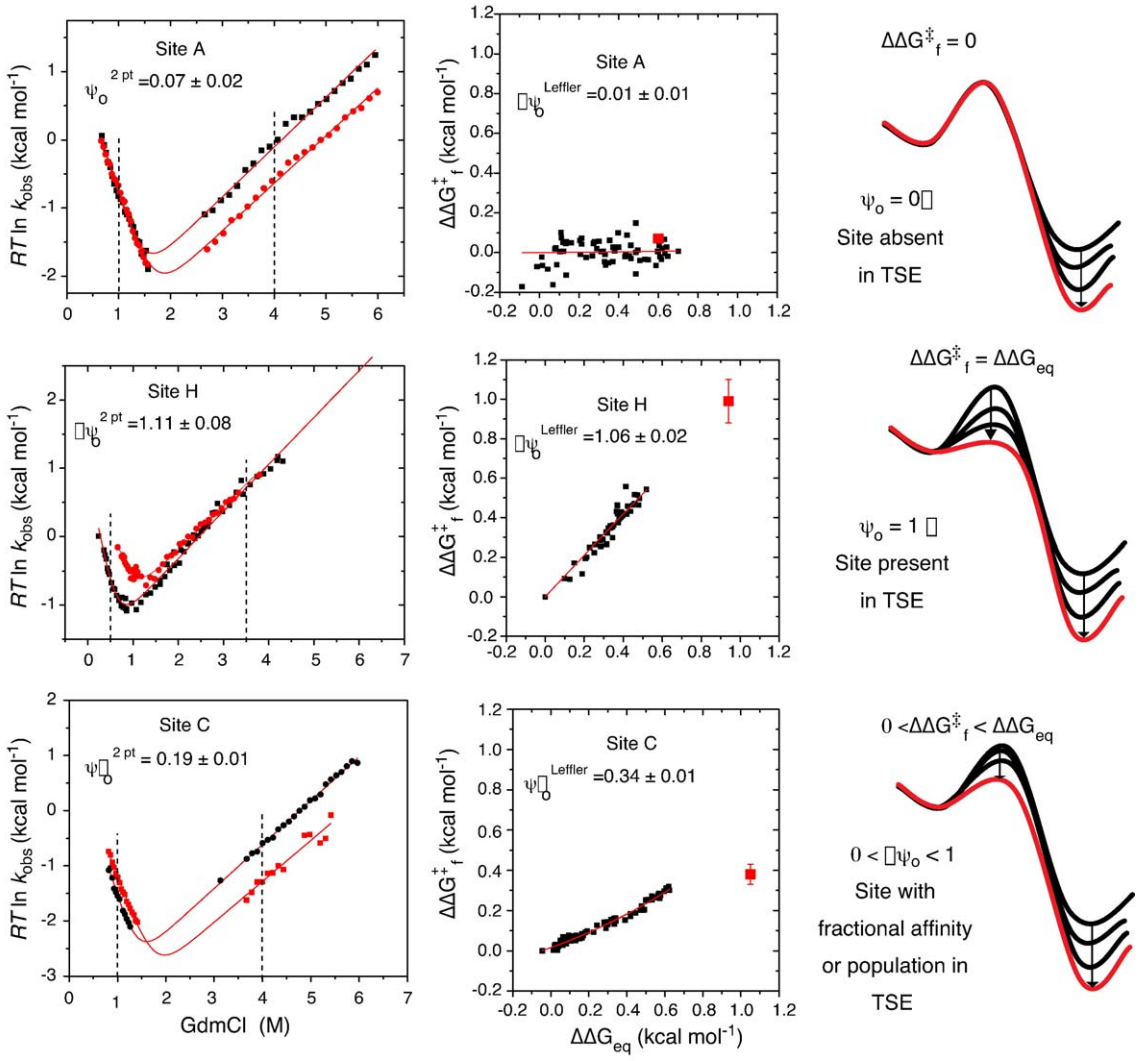

Figure 3. ψ-Analysis: metal-dependent protein folding measurements. The three classes of ψ-values are illustrated in each of the three rows. Values of zero and one indicate that the biHis site is either entirely absent or present in the TSE. A fractional value indicates that the site is either formed in a fraction of the TSE at a level given by the fraction ψ-value, a single distorted site with reduced binding affinity in the TS, or a combination thereof. ψ-Analysis can be conducted using two types of experiments. The left set of panels displays traditional denaturant dependent kinetic data (chevron analysis), measured in the absence (black) and the presence (red) of high metal ion concentration. When the site is absent in the TSE, only the unfolding arm moves and $\psi_o = 0$ (top left). Conversely, when the metal site is fully formed in the TSE, only the folding arm moves and $\psi_o = 1$ (middle left). When both arms shift, a fractional value results (bottom left). Alternatively, measurements are performed at a gradient of metal ion concentrations, and at a single folding and a single unfolding condition (dotted lines in left panels) to get $\Delta\Delta G_f^{\ddagger}$ and $\Delta\Delta G_u^{\ddagger}$. These values are used to calculate the Leffler plot (center panels). The single red data point in the center panels is the value corresponding to the shift in the denaturant chevron upon addition of saturating metal ion concentration. When the site is absent in the TSE, the data are flat with a $\psi_o^{Leffler} = 0$ (center top). If the site is fully formed in the TSE, $\psi_o^{Leffler} = 1$ (center middle). A fractional ψ-value produces a curved Leffler plot, which approaches a slope of unity at infinite stabilization (center bottom). Each of these three scenarios can be interpreted as a series of shifting free-energy diagrams (right panels). When the metal stabilization does not affect the folding rate, the metal binding stabilizes only the native state (right top). When the site is fully present, the binding energy in the TSE and native states are equal and are thus stabilized energetically by the same amount (right middle). For fractional ψ-values, the stabilization initially only affects the native state. When the metal-present pathway is stabilized enough to be dominant, the TS and native state are stabilized equally upon a further increase in stability due to metal ion binding.

The interpretation of ψ-values is clear in the two cases where the Leffler plot is linear. When ψ_o is unity, the biHis site is present with native-like binding affinity in the TS ensemble. When $\psi_o = 0$, the site is absent with unfolded-like affinity. In other cases, the Leffler plot will be curved as ligand binding continuously increases the stability of the TS ensemble.

Assuming that the metal ion binding affinity in the TS is either native-like or unfolded-like, the ψ-value obtained at any given metal concentration represents the fraction of the TS ensemble with the

biHis site formed (Supplementary Figure 3A). The other fraction of the molecules are crossing the rate-limiting barrier without the two histidine residues in a geometry capable of binding metal ions. Together, these two populations comprise the TSE.

This heterogeneous picture, rather than the scenario with a distorted site having non-native binding affinity, quantitatively described the degree of TS heterogeneity in the folding of a dimeric α-helical coiled coil,[27] a system known to have multiple nuclei.[20] This issue is examined below using different metal ions in the same biHis site to test the interpretation that fractional ψ-values quantify the fraction of the TSE that has native-like binding sites.

The complete ψ-analysis formalism takes into account the shifts in the native, unfolded and TS populations due to the binding of metal ions to each of these states (Supplementary Figure 4). Folding rates are calculated assuming two classes of TSs depending on whether the biHis site is present (k^{present}) or absent (k^{absent}). In the first class, TS_{present}, the biHis site is present in a native or near-native geometry with a dissociation constant K_{TS}^{present}. In the second class, TS_{absent}, the biHis site is essentially absent but is given a nominal effective dissociation constant K_{TS}^{absent}. According to Eyring reaction rate theory,[41] the overall reaction rate is taken to be proportional to the relative populations of the TS and U ensembles, $k_f \propto [TS]/[U]$. The net folding rate is the sum of the rates going down each of the two routes, $k_f = k^{\text{present}} + k^{\text{absent}}$, with:

$$k_f \equiv \frac{1+[M]/K_{TS}^{\text{present}}}{1+[M]/K_U} k_o^{\text{present}} + \frac{1+[M]/K_{TS}^{\text{absent}}}{1+[M]/K_U} k_o^{\text{absent}} \quad (3)$$

where $k_o^{\text{present}} \propto [TS_{\text{present}}]/[U]$, $k_o^{\text{absent}} \propto [TS_{\text{absent}}]/[U]$ are the rates through each TS class prior to the addition of metal, and [M] is the divalent metal ion concentration. In equation (3), the pre-factors of these two rates represent the increase in the population of each class's TS, relative to the increase in the population of the unfolded state, due to differential metal affinity in the TS and in the unfolded state. By examining shifts in populations and assuming that metal binding is in fast equilibrium, this treatment avoids any assumptions about possible pathways connecting the different bound and unbound states.

There are two major scenarios that one can consider. In first scenario, the TS_{present} has the biHis site present with native-like affinity ($K_{TS}^{\text{present}} = K_N$) while TS_{absent} has the site with the unfolded-like affinity ($K_{TS}^{\text{absent}} = K_U$). Hence, only TS_{present} is stabilized with respect to the unfolded state upon the addition of metal ions (Supplementary Figure 3B). The height of the kinetic barrier associated with TS_{present} decreases by the same amount as the native state's stability. As a result, the rate increases down this pathway, $k^{\text{present}} = k_o^{\text{present}}\, e^{\Delta\Delta G_{eq}/RT}$. The instantaneous slope simplifies to the fraction of the TS ensemble that has the biHis site formed at a given metal ion concentration:

$$\psi = \frac{k^{\text{present}}}{k^{\text{present}} + k_o^{\text{absent}}} \quad (4)$$

In the second scenario, curvature can also occur when the sole TS has binding affinity different from the native state (e.g. a distorted site with $K_{TS}^{\text{present}} > K_N$). It is important to note that for this and the other scenarios, the $\psi = 0$ and 1 values still imply that the biHis site is absent or 100% present in the TSE, respectively.

It has been suggested that surface mutations and metal binding in ψ-analysis may introduce complications that are difficult to unravel.[42] However, these effects can be accounted for in ψ-analysis. The altered stability due to the introduction of the two histidine residues, $-\Delta\Delta G_{\text{biHis}}$, can be corrected for by a shift in *X*-direction in the Leffler plot to produce a ψ-value appropriate for the wild-type protein (Supplementary equation S11). In addition, the ability to extrapolate to the limit of no metal binding ($\Delta\Delta G_{eq} \rightarrow 0$) implies that the ψ_o-value reports on an unperturbed and unstrained TSE. In contrast, the accurate determination of ϕ-values can require $\Delta\Delta G_{\text{mutation}} > 1$–2 kcal mol^{-1}.[43,44] This significant perturbation can result in a different TSE.[20] Furthermore, significantly destabilizing mutations can lead to non-native ground states,[45] a result that complicates the interpretation of ϕ-values.

Finally, it should be appreciated that the use of the single variant at different cation concentrations to determine a ψ-value at each site enables a perfect cancellation of the attempt frequency, or pre-factor, in reaction rate theory, $k_f = k_{\text{attempt}} e^{-\Delta G^{\ddagger}/RT}$. For the determination of the ϕ-value, however, the attempt frequency is assumed to be the same for the wild-type and mutant proteins ($\Delta\Delta G^{\ddagger}_f = RT \ln [(k_f^{\text{w-type}}/k_f^{\text{mutant}})/(k_{\text{attempt}}^{\text{w-type}}/k_{\text{attempt}}^{\text{mutant}})] \sim RT \ln [k_f^{\text{w-type}}/k_f^{\text{mutant}}]$).

Results

The TS topology of ctAcP is examined using nine separate surface biHis ion binding sites located throughout the protein (Figure 2). To eliminate any potential non-native metal ion binding, the three endogenous histidine residues are removed (H25A, H60A, H74A). BiHis sites then are engineered individually onto the surface of this pseudo-WT variant of ctAcP. Two sites are located on each of the two helices and five sites are placed across strands of the β-sheet network (Figure 2(b)). Divalent cations (Co^{2+}, Zn^{2+}, or Ni^{2+}) form complexes with biHis sites with the degree of stabilization varying among sites, presumably due in part to differences in preferred ion coordination geometry and spatial arrangement of the histidine partners (Table 1).

Initially, denaturant chevrons of all biHis variants are measured in the absence and presence of high

Table 1. Equilibrium and kinetic parameters for biHis substitutions and divalent metal ion binding[a]

Site	Mutations	ΔG_{eq} $\Delta G_f^{\ddagger}$	m^o m_f/m^o	$\Delta\Delta G_{mut}$	$\Delta\Delta G_{bind}^{Me}$ [b]	High [Metal][b] ΔG_{eq} $\Delta G^{\ddagger}_f$	m^o m_f/m^o	ϕ	$\psi_o^{2\ pt}$ [c]	$\psi_o^{Leffler}$ [d]	Metal[b] (conc)
PWT	H25A	−5.56±0.09	3.35±0.07	NA	NA	NA	NA	NA	NA	NA	NA
	H60A	2.02±0.07	0.80±0.01								
	H74A										
A, H1	K24H	−4.82±0.10	3.36±0.10	−0.89±0.04	−0.60±0.04	−5.01±0.06	3.04±0.05	0.21±0.02	0.07±0.02	0.01±0.01	Co
	A28H	1.79±0.10	0.78±.01			1.58±0.05	0.77±0.01				
B, H1	A28H	−4.68±0.12	3.64±0.13	−1.16±0.16	−0.30±0.05	−4.32±0.16	3.17±0.16	0.30±0.02	0[g]	ND	Co
	K32H	1.96±0.12	0.81±0.01			1.55±0.16	0.81±0.01				
C, H2	S56H	−5.08±0.10	3.66±0.10	−1.12±0.03	−0.67[e]	−5.39±0.21	3.08±0.13	0.83±0.01	0.19±0.01	0.34±0.01	Zn
	H60H	1.28±0.10	0.78±0.01			1.14±0.14	0.76±0.02				
C, H2	S56H	−5.08±0.10	3.66±0.10	−1.12±0.03	−0.85[e]	ND	ND	0.83±0.01	ND	0.46±0.09	Co
	H60H	1.28±0.10	0.78±0.01								
C, H2	S56H	−5.08±0.10	3.66±0.10	−1.12±0.03	−1.5[e]	ND	ND	0.83±0.01	ND	0.37±0.00	Ni
	H60H	1.28±0.10	0.78±0.01								
D, H2	R59H	−4.69±0.09	3.58±0.10	−1.38±0.04	−0.36±0.05	−4.64±0.09	3.33±0.09	0.62±0.01	1.05±0.07	1.10±0.04	Co
	E63H	1.29±0.09	0.78±0.01			1.46±0.09	0.78±0.01				
F,β1-4	E12H	−3.78±0.13	3.62±0.18	−1.64±0.06	−0.84±0.07	−4.20±0.13	2.82±0.09	0.62±0.01	1.44±0.10	ND	Zn (2)
	N79H	1.41±0.14	0.85±0.01			1.46±0.10	0.76±0.02				
G β2-5	Q40H	−5.75±0.12	3.46±0.10	−0.85±0.04	−0.59±0.05	−5.70±0.13	3.24±0.10	−0.14±0.04	−0.12±0.05	0.13±0.02	Ni
	V97H	1.93±0.12	0.71±0.01			1.72±0.12	0.74±0.01				
H,β2-3	W38H	−2.61±0.10	4.02±0.23	−3.39±0.07[f]	−0.94±0.11[f]	−2.92±0.16	3.22±0.23	0.44±0.01[f]	1.11±0.08[f]	1.06±0.02	Zn
	Q50H	0.85±0.10	0.82±0.01			1.45±0.19	0.80±0.02				
I, β1-3	Q50H	−4.35±0.11	3.27±0.12	−1.26±0.04	−1.49±0.06	−4.69±0.15	2.44±0.11	−0.03±0.02	0.53±0.03	1.05±0.03	Zn (0.5)
	D10H	1.93±0.12	0.78±0.01			2.48±0.13	0.76±0.02				
J, β1-4	D10H	−3.33±0.12	3.61±0.19	−2.08±0.06[f]	−0.28±0.06[f]	−3.39±0.13	2.30±0.17	0.47±0.02[f]	0.70±0.09[f]	ND	Zn
	N81H	1.22±0.12	0.84±0.01			1.10±0.13	0.79±0.09				

[a] To minimize extrapolation errors, values for $\Delta\Delta G_{eq}$, $\Delta\Delta G_{bind}^{Me}$, ϕ-value and $\psi_o^{2\ pt}$-value are calculated using $\Delta\Delta G_f^{\ddagger}$ and $\Delta\Delta G_u^{\ddagger}$ values determined at 1 M and 4 M GdmCl, respectively, generated from a simultaneous fit to the two relevant chevrons, with the parameter of interest being one of the fitting parameters. Units are kcal mol^{-1} (free energies) or kcal mol^{-1} M^{-1} (*m*-values). NA, Not applicable, ND, not determined.

[b] High metal concentration is 1 mM, except if noted in the last column in parenthesis, in mM.

[c] Obtained from the shift in the chevron upon the addition of high concentration of metal ions, and a two point fit to equation (1).

[d] ψ_o-Value obtained from a Leffler plot using data acquired at multiple metal ion concentrations.

[e] Obtained from the spread in the $\Delta\Delta G_{bind}^{Me}$ in the Leffler plot.

[f] Calculated using $\Delta\Delta G$ at 0.5 M and 3.5 M GdmCl, due to lowering of chevron midpoint upon biHis substitution.

[g] Ill-defined as $\Delta\Delta G_f^{\ddagger}$ is statistically zero (−0.005±0.042) while $\Delta\Delta G_{eq}$ = −0.30±0.05.

(millimolar) concentrations of divalent cations. The chevron vertices are shifted to higher denaturant concentrations as a result of metal-induced stabilization. For all variants, the slopes of the chevron arms remain largely unchanged, indicating that the surface area buried in the TS does not change as a result of either mutation or metal binding (Table 1). This invariance indicates that biHis mutations and ion binding do not induce any qualitative changes in the folding pathway.

Helical and sheet structure in the TSE

Results of ψ-analysis on ctAcP indicate that helix H1 is unstructured, H2 is partially formed, and strands β1–4 adopt the native arrangement in the TSE (Figure 2). Four sites in the two α-helices are tested: sites A and B on H1 (residues 22–32) and sites C and D on H2 (residues 55–66). From the denaturant chevrons measured in the absence and presence of metal ions, site A and site B are found to have $\psi_o = 0.02 \pm 0.03$ and 0.07 ± 0.02, respectively. Hence, H1 does not participate in the TS ensemble even after 1.6 kcal mol^{-1} of metal-induced stabilization.

For site C in H2, both the folding and unfolding arms of the chevron shift, whereas for site D on the same helix, only the folding arm shifts upon the addition of metal ions. The resulting ψ_o-values are 0.37 ± 0.01 ($NiCl_2$) and 1.11 ± 0.03, respectively. These results indicate that the carboxy-terminal end of H2 is fully formed while the other end of the helix participates at a lower level in the TSE.

The β-sheet network, with the exception of β5, adopts a native-like organization in the TS. For the biHis sites bridging β1-β3, β1-β4, and β2-β3, the ψ-values are constant and near unity over the measured range of metal ion concentrations. However, the ψ-value for the site traversing strands β2-β5 is zero, indicating that β5 does not participate in the β-sheet network in the TS.

Our observation of the formation of helix H2 and the absence of H1 in the TSE also is found in a φ-analysis study.[46] However, our ψ-value of zero between the β2-β5 contrasts with a mutational study of the human homolog muscle AcP (88% sequence identity), which concludes that the β2-β5 connection is formed in the TSE.[11] This conclusion is based on a high φ-value for F94 on β5 ($\phi^{F94L} = 0.76 \pm 0.15$), and F94's extensive interactions with the rest of the protein (F94's phenyl ring is "sandwiched" in between strand β2 and helix H1).

To investigate the apparent discrepancy, we examined two mutations, F94L and F94A. In both cases, the folding arms of the chevron are coincident with the wild-type's arm despite the large degree of destabilization, $\Delta\Delta G_{eq}^{mutation} = 1$–$3$ kcal mol^{-1} (Supplementary Figure 1). The resulting φ-values are nearly zero, which indicate that F94 is not making interactions with the rest of the protein. Furthermore, the absence of helix H1 in the TSE, as found in the present and the earlier study,[46] is inconsistent with a high φ-value for F94L because F94's phenyl ring is packed in between H1 and β2 in the native state. This packing is impossible in the TS because helix H1 is absent in the TS. Hence, a high φ-value is implausible. Finally, the mutations that make AcP susceptible to fibril formation are primarily in β5.[11,47] This result suggests that on the unfolding pathway, this strand unfolds and is available for inter-protein contacts prior to the TS.

Fractional ψ-values and TS heterogeneity

All sites have ψ-values of zero or one, except for site C located at the amino terminus of helix H2. Values of zero or unity are accepted as the absence or presence of a native-like biHis site in the entire TSE.[26,28,42,48] However, as noted by Fersht,[42] Goldenberg (private communication), Sosnick *et al.*[26,28] and subsequently by Kiefhaber *et al.*,[48] fractional ψ-values, as with fractional φ-values, can be interpreted in several ways. Fersht's earlier analysis[49] and the other three treatments of ligand binding attribute the acceleration of folding rates to the explicit binding and stabilization of the TS. Fersht's 2004 interpretation of ψ-values focuses on shifts in the unfolded state populations. In the shared view, fractional ψ-values arise either from the formation of native-like binding site in a fraction of the TSE at a level given by the fraction ψ-value, or a single site with non-native binding affinity in the TS, or a combination thereof.

Delineation between the heterogeneous and homogeneous scenarios for the lone fractional ψ-value at site C is investigated using three divalent metal ions having different coordination geometries. Ni^{2+}, Zn^{2+}, and Co^{2+} preferentially have octahedral, tetrahedral, and square planar/octahedral coordination geometries,[50] and stabilize the native state by 1.5, 0.9 and 0.7 kcal mol^{-1}, respectively. Nevertheless, the measured ψ-values, when extrapolated to the zero metal condition, are nearly the same, $\psi_o = 0.37 \pm 0.01$, 0.34 ± 0.01 and 0.46 ± 0.09, respectively (Figure 4). Given their different coordination preferences, the same fractional binding is unlikely to be maintained for a single site. If the site had a distorted geometry or was flexible, metals with different coordination geometries could stabilize the TS to different extents, relative to the stability they impart to the native state, and thus, they likely would return different ψ-values.

In contrast, in the heterogeneous scenario, the ψ-values are expected to be the same with different metal ions even if they bind the native-state with different affinities. The ψ-value should only depend on the fraction of the TSE that has native-like binding affinity, and not the magnitude of the binding affinity, which can vary from metal to metal. Although a homogeneous model wherein the three ions have the same fractional affinity in the TS cannot be ruled out, we believe that the similarity of the observed ψ-values for site C best supports the heterogeneous model.

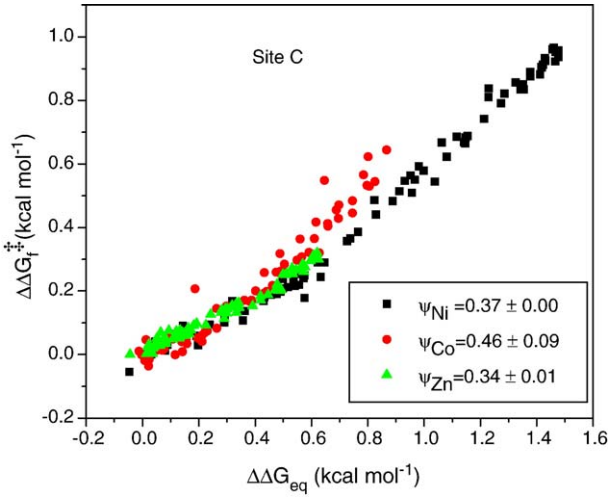

Figure 4. Multi-metal test supports intermediate ψ-value are indicative of TS heterogeneity. Leffler plot for site C using Ni^{2+}, Co^{2+}, and Zn^{2+}. The similarity of the results supports the view that the fractional ψ-value observed is due to a fraction of the TSE containing a native-like biHis binding site at the amino terminus of helix H2. The alternative interpretation of a curved Leffler plot, a distorted biHis site with lesser affinity, would have resulted in significantly different ψ-values for the three ions as they have different coordination geometries.

Furthermore, in the dimeric GCN4 coiled coil, the ψ-values quantitatively reproduce the known degree of TS heterogeneity under the assumption that sites are either formed with native-like or unfolded-like binding affinity (i.e. fully present or absent, respectively).[20,27] Additionally, the ψ-values increase in accordance to the change in flux as destabilizing mutations introduced at the other end of the coiled coil reduce the probability of nucleation at these positions located 2 1/2 turns away. This result is inconsistent with a fractional ψ-value representing a single site with non-native binding affinity. These results support the interpretation that the ψ-value can report on the degree of TS heterogeneity.

Given these results and structural considerations, we believe that fractional ψ-values generally reflect TS heterogeneity. Fractional ψ-values can readily arise from microscopic heterogeneity such as the fraying of a helix or an edge of a β-sheet, events that can occur in the native state. Such fraying events are likely to be in fast equilibrium relative to the transit time over the barrier. The equilibrium between these individual states can be established in the TSE if they interconvert right at the saddle point on the reaction pathway. Alternatively, the equilibrium can be established if the states are traversed multiple times as the molecules go back and forth across the top of the barrier. During these passages, the trajectory samples each configuration with a relative weighting equal to the ψ-value. Such repeated barrier crossings are expected for reactions in solutions having intermediate barrier heights, as envisioned by Kramers' theory.[51]

Models of ctAcP TS

Two models are created for ctAcP's TSE using the ψ-values as constraints (Figure 5). A minimalist model is created using only the $\psi = 1$ values as constraints, and a more extensive model, which includes a slightly larger β-sheet network, and the entire H2 helix to reflect the fractional ψ-value observed at the amino terminus (site C). Structured regions around the $\psi = 1$ sites are left in their native configuration while the regions outside the nucleus are fully unfolded. Native hydrogen bonds are retained where possible as high ψ-values imply near-native backbone geometries for the two histidine partners. The presence of hydrogen bonds in the TSE is further supported by the high fraction of the surface being buried in the TS ($m_f/m^o \sim 80\%$), and our previous kinetic amide isotope studies which found concomitant surface area burial and hydrogen bond formation in the TS.[52,53] After inserting the side-chain groups using the Swiss-PdbViewer V3.7†, the relative contact orders (RCO)[35] of the minimal and maximal models are 73 and 79% of the native value, respectively.

These models are further refined to reflect the mounting evidence that the polypeptide undergoes some relaxation along the folding trajectory, for example, minimizing energy through the formation of non-native hydrophobic interactions. Such relaxations have been observed in the intermediates of Rd-apocytochrome b_{562},[54] IM7,[55] and apomyoglobin.[56] In Ub, the $\phi^{L76A} = 0$ result at a core position is rationalized by the TS structure relaxing to accommodate the loss of the three methyl groups in the alanine mutant. This rearrangement is not possible in the more rigid native structure, so that the relative energetic penalty in the TS is less than anticipated for an otherwise fully buried side-chain.[28]

Given this evidence for structural relaxation and our TS models are created assuming either native-like or unfolded-like regions, the models are relaxed using an all-atom, implicit solvent Langevin dynamics (LD) simulation with the OPLS-AA force field,[57,58] using the protocol and model developed by Shen and Freed.[59,60] The unity ψ-values are reflected as distance constraints between the two C^{α}s of the residues composing the biHis site. In order to account for different, and possibly non-native binding geometries, the backbone is allowed to re-orient with the C^{α}–C^{α} distance allowed to freely vary by up to 0.5–1 Å (see Materials and Methods) with larger excursions constrained by a flat-well harmonic potential (spring constant of 100 kcal/Å^2). After the first 2 ns of the 5 ns trajectory, the minimal and maximal models have RCO values that are 75(±2)% and 80(±2)% of the native RCO value, respectively (Figure 5). From each of the two trajectories, a representative structure is selected. These two structures illustrate the range of conformations

† http://www.expasy.org/spdbv/

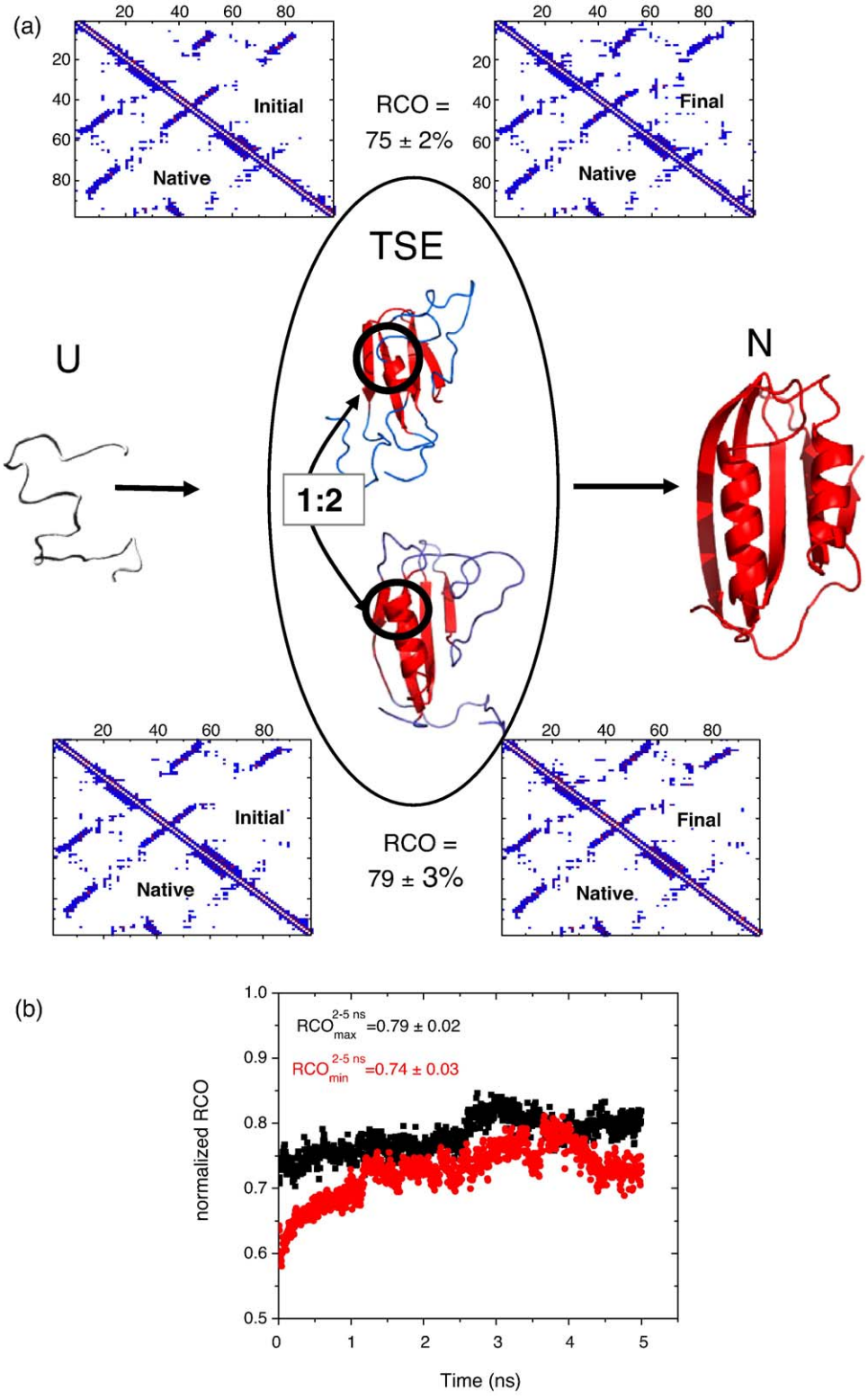

Figure 5. TS models constructed using ψ-value results and LD simulations. (a) At the TSE, two broad structural classes exist, both containing strands β1–4 and the carboxy end of H2. Due to the intermediate ψ-value for site C, a fraction of the TSE contains a fully structured H2. Models depicted are typically structures from the LD simulations using the OPLS-AA force field for the minimal model created using ψ-values ~1 (top), and maximal model created using ψ-values >0.3 (bottom) as constraints. The contact maps in the upper right triangle are for the initial (left) and the final (right) TS structures. The contact map of the native structure is shown in the lower left triangle for comparison. Blue regions are unstructured, whereas red indicates native-like structure. (b) Trajectory of RCO for the minimal and maximal models, normalized to the average value observed for a trajectory of the native state. Over the 10 ns native trajectory, the protein remained within 1.5 Å RMSD of the crystal structure.

that may be found in the TSE of ctAcP. The fractional surface burial in these models, relative to the native state and the unfolded state,[61] varies from 79–88%. This burial level is in good agreement with the high fraction of the denaturant sensitive surface buried in the TS ($m_f/m^o \sim 80\%$).

It may be considered that the models even after the LD simulations have a native-like bias as a result of the native structure serving as the starting template. To further remove the native bias and test the robustness of their RCO values, we heated the two TS models to 373 K for 5 ns. The resulting models are only slightly different compared to their 298 K counterparts (<2 Å RMSD, Supplementary Figure 5) and the relative contact order remains within the error of the original value (Supplementary Figure 6).

This modeling process followed by LD simulation is repeated for Ub using the ψ-values obtained in our previous study.[26] The minimal and maximal models, generated using the $\psi = 1$ and the $\psi > 0$ sites, have RCO values of 80(±2)% and 84(±2)%, respectively (Supplementary Figure 2). The small difference in the RCO values for the minimal and maximal models with each protein indicates that this quantity is robust to whether fractional ψ-values represent microscopic heterogeneity or the presence of a single, but distorted site.

Discussion

Single consensus TS in ctAcP

In Ub, the consensus TS structure is extensive, minimally containing strands β1–β4, and part of the major helix.[26] Having two helices, the α–β motif appears twice in ctAcP and has been considered to be critical in the folding of ribosomal protein S6.[62] This led to the possibility that the folding of ctAcP might occur through two structurally disjoint TSs, each containing one of the two helices with some fraction of the β-sheet network. However, ψ-analysis indicates that the TSE of ctAcP contains only H2 and strands β1–β4, with no evidence of the participation of H1 or strand β5 (Figure 5).

The degree to which ψ-analysis can detect a minor TS population is determined by the amount of metal-induced stabilization imparted to the structure present only in the minor TS. For example, the Leffler plot for site B in helix H1 is flat over the range $\Delta\Delta G_{eq} = 0.6$ kcal mol^{-1}, with a zero ψ_o-value (0.01±0.01). If the plot remained flat for an extra 1.4 kcal mol^{-1} of binding energy, i.e. the site still does not participate in the TSE even after being significantly stabilized, then the upper limit of its fractional population in the TSE can be decreased by tenfold to 0.1%. This reasoning can be applied to the $\psi \sim 1$ value for site D in H2, which has been destabilized by 1.4 kcal mol^{-1} because of the introduction of the biHis mutations. In spite of its destabilization, H2 still fully participates in the TSE, as indicated by the Leffler plot having a slope of unity across the entire range of stabilization ($\psi_o^{\text{Site D}} = 1.10 \pm 0.04$). After accounting for the increase in H2 stability in the pseudo-wild-type protein, and assuming that site D is present at least at the 96% level in the TSE (based on the 4% error), then H2 should be present at the 99+% level in the TSE of the pseudo-wild-type protein.

Generality

Evidence for TS heterogeneity in the folding of small proteins is limited (in the absence of proline isomerization).[15,17,26,63,64] Both the Serrano[15,65] and Baker[17] groups found no evidence for a shift in pathway upon the destabilization of elements formed in the TS of SH3 domains. Similarly, loop insertion studies did not detect pathway heterogeneity.[17,64] Other work has investigated the role of topology either through cross-linking or circular permutation. Topological changes sometimes[62,66–68] but not always[17,69] result in different TSs. The existence of distinct TSs reveals that there is structural diversity at the rate-limiting step in the different versions. However, these results do not mandate that heterogeneity exists for the same version of the protein.

In addition to the dimeric GCN4 coiled coil, an example of TS heterogeneity was observed in the folding of titan by Clarke and co-workers.[19] Heterogeneity is identified by an upturn in the unfolding arm of the chevron plot. The pattern of ϕ-values indicates that the larger of the two TSs contains an extra β strand and some additional structure formed around the periphery of the central nucleus. This type of heterogeneity seems closer to microscopic heterogeneity described earlier, rather than an example of disjoint nuclei. Generally, unless there is a strong symmetry to the protein (see Klimov & Thirumalai[70] and references therein) as observed in the dimeric coiled coil,[27] Protein L[71] and Protein G,[68] a single folding nucleus is likely to be a general phenomenon in the folding of small proteins.

TS structures and contact order

We and others proposed that the folding of small proteins is a nucleation process where the formation of a coarse version of the native chain topology is the critical step.[72–74] This proposal is bolstered by the strong correlation between log k_f, and RCO, a measure of topological complexity as discovered by Plaxco *et al.*[35] (Figure 6). This correlation and others[75–77] point to the two-state folding reaction being limited by an initial conformational search for some threshold amount of the native-like topology.

ψ-Analysis identifies residue–residue contacts, which makes it particularly well-suited to addressing the origin of the correlation between k_f and RCO, which reflect properties of the TS and the

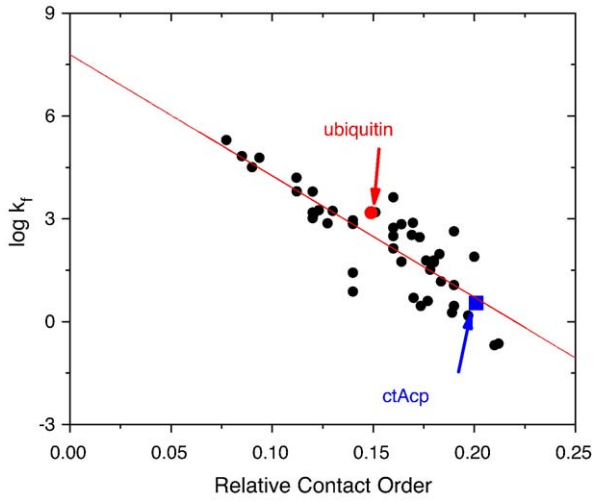

Figure 6. Correlation between relative contact order and folding rates. Each data point represents an individual two-state protein. The red line indicates the best linear fit to the data, which we believe represents proteins that form ~80% of the native topology at the TS. Data obtained from Plaxco *et al.*[35] and Maxwell *et al.*[99]

native state, respectively. The TSEs identified for both ctAcP and ubiquitin using ψ-analysis coupled with the modeling and LD simulations, have RCO values that are ~80% of the native RCO value.

The similarity between these two proteins, combined with the general RCO trend suggests an intriguing hypothesis: for the proteins that obey the trend, their TSs also should have ~80% of the native RCO. That is, in order for the correlation to be valid across a range of proteins, the TS of each protein must have a similar relationship to its native state. Given that the TSEs of ubiquitin and ctAcP have ~80% of the native RCO, the TSEs of these other proteins also should have $RCO^{TS} \sim 0.8RCO^{native}$. Effectively, ctAcP and Ub's values for the RCO^{TS} help rationalize and calibrate the connection between k_f and RCO of the native state.

Vendruscolo *et al.* came to a similar conclusion about TSs for different proteins sharing a common fraction of their native RCO.[78] Starting from the native structure, they generated a TSE using molecular dynamics constrained by ϕ-values, which are assumed to equal the fraction of native contacts formed by the residue in the TSE. However, they found that the average RCO of the TSE is only ~50% of the native value for ten proteins, rather than the ~80% observed for ubiquitin and ctAcP in our study. Potentially, the lower level observed in their simulations, which are constrained by ϕ-values, is because this parameter reflects energies and not structures, and hence, can under-report chain–chain contacts.[28]

Recently, Wallin and Chan used Go-like C^{α}-Langevin modeling for 13 proteins.[79] They observed that the average RCO of their TSEs was ~70% of the native values, in very good agreement with our work. However, as stressed here and in numerous other works,[76,77,80] the success of RCO correlation does not prove that this particular parameter best describes the physical origin of the correlation. Other parameters also can exhibit good correlation with ln k_f. These parameters include the long-range order (LRO), which is the number of contacts for sequence-distant residues separated by more than l_c residues, but within a distance of r_c, normalized with respect to L, the total number of residues,[81] and the topomer search model's Λ_D parameter, which is equal to LRO*L.[82]

Accordingly, for the minimal and maximal TS models, we calculated the fraction of the native LRO and the Λ_D values over the last 3 ns of the LD simulation at 298 K. For ctAcP and ubiquitin, $LRO^{TS}/LRO^{N} = \Lambda_D^{TS}/\Lambda_D^{N} = 0.76–0.77$ and 0.56–0.81, respectively, using a maximum $C_{i,\alpha}–C_{j,\alpha}$ contact distance of 6 Å and a minimum sequence separation of $|i–j|>12$ residues. The results are not as uniform over the LD trajectory as for RCO. The RCO, however, is calculated using the number of heavy-atom contacts, rather than the number of C^{α} contacts. When the LRO and Λ_D parameters are calculated using heavy-atom contacts, the fraction realized in the TS becomes 0.71–0.72 and 0.79–0.81, for ctAcP and ubiquitin, respectively. Presumably, the use of the heavy-atom contacts produces more uniform values by being less sensitive to small-scale fluctuations in partially folded states. Regardless, the TSE fractions of the LRO and Λ_D are similar for both proteins. Hence, the proposal that TSEs will have the same (high) fraction of native topology for proteins obeying the general RCO trend is robust to the choice of topological parameter.

Implications

The proposition that $RCO^{TS} \sim 0.8RCO^{native}$ can be used as a constraint in the delineation of TSE for ctAcP and other proteins. For example in ctAcP, the absence of an alternative TSE having helix H1 instead of H2, may be attributed to this alternative TSE having only ~50% of the native RCO. The high value for RCO^{TS} also restricts the degree to which a TS can be small and polarized.[1–9] From the ϕ-values for such proteins, we estimate that RCO^{TS} often is below $0.5RCO^{native}$, although the precise number depends on the threshold that a ϕ-value is considered to be a contact. At face value, these results would seem to be incompatible with the proposition that $RCO^{TS} \sim 0.8RCO^{native}$.

However, some caveats exist in the identification of a small, polarized TS based solely upon medium to high ϕ-values on one side of the protein. First, ϕ-analysis leads to an assignment of a small, polarized TS in ubiquitin, whereas the TS defined by just the unequivocal $\psi = 1$ sites is much more extensive.[28] As Schmid *et al.* astutely noted in their studies "the transition state of CspB folding is polarized energetically, but it does not imply that one part of the

protein is folded and the other one is unfolded. Rather, it means that the positions that have reached a native-like energetic environment in the transition state are distributed unevenly."[1] That is, energetically polarized does not necessarily mean structurally polarized.[83] Second, the low level of structure formation inferred from ϕ-analysis in these situations seems inconsistent with the high percentage of surface burial in the TS (m_f/m^o).

The third caveat is that many high ϕ-values in polarized TSs are associated with turns.[1–3,7,84] These high values may not yield a complete picture of the topology of the TS. For example, Serrano *et al.* concluded that the three SH3 homologs, SSo7D, src- and α-spectrin, fold *via* different TSs as their turns have different ϕ-values.[85] However, an alternative interpretation is that the overall TS topology is similar in all three proteins, but the turns are only folded to the degree required for the chain to turn around. Not all turns have to be native-like in order for the chain to double-back on itself. If true, the sensitivity of the ϕ-values is more a reflection of the specifics of the turn, rather than the topology of TS. For example, the distal β-hairpin in src-SH3 with high ϕ-values is a tight turn,[2] which is quite sensitive to mutation. The corresponding turn in SSo7D contains three flexible glycine residues, which are not as sensitive to mutation, and this turn has low ϕ-values,[85] Hence, ϕ-values could be different for this turn despite the two proteins having similar TS topologies.

Conclusion

The TSE of ctAcP has a single consensus nucleus with only a minor amount of structural variation, which is best described as microscopic heterogeneity. The TS is very native-like in its topology and amount of surface area burial. For both ctAcP and ubiquitin, the RCO of the TS is high, ~80% of the native value. This high value greatly diminishes the variability of possible TS structures, thereby reducing the amount of potential TS heterogeneity. Whether there are multiple pathways up to the singular TS remains an open issue.

In addition, it remains to be determined whether the high fractional RCO value is an intrinsic property of cooperatively folding proteins, for example, where on the free energy landscape the entropy–enthalpy balance shifts towards being downhill in free energy. This shift may occur upon the formation of a partially desolvated hydrophobic core, which requires the majority of the native structure to be formed. Alternatively, this high fraction could be selected for, either by the choice of fold or sequence, during evolution. A highly organized TSE would reduce the amount of partial unfolding possible on the native-side of the rate-limiting barrier. As partial unfolding and the formation of intermediates is associated with fibril formation,[86–88] a highly structured TS would reduce the possibility of unwanted fibriligenesis.

Materials and Methods

Expression and purification

BiHis double mutants were engineered sequentially using the QuikChange protocol (Stratagene) in a pseudo-WT triple-mutant background eliminating native histidine residues H25A, H60A, and H74A. ctAcP was expressed and purified as described.[89] Protein identity was confirmed by electrospray time-of-flight mass spectrometry (TOF-MS).

Folding measurements

Experiments were conducted in 50 mM Hepes (pH 7.5), 30 °C at a protein concentration of 1–10 μM. Rapid mixing was accomplished using Biologic brand SFM-4 and SFM-400 stopped-flow rapid mixing devices.[23] Fluorescence spectroscopy used $\lambda_{ex} = 280–290$ nm and $\lambda_{em} > 320$ nm.

Data analysis

The kinetic data were fit to a single exponential and analyzed according to two-state folding transition, with ΔG_{eq}, $\Delta G_f^{\ddagger}$, and $\Delta G_u^{\ddagger}$ being linearly dependent upon GdmCl concentration.[29]

$$\Delta G^o([\mathrm{GdmCl}]) = \Delta G^o_{\mathrm{H2O}} - m^o[\mathrm{GdmCl}] \quad (5a)$$

$$\Delta G_f^{\ddagger}([\mathrm{GdmCl}]) = RT\ln k_f^{\mathrm{H2O}} - m_f[\mathrm{GdmCl}] \quad (5b)$$

$$\Delta G_u^{\ddagger}([\mathrm{GdmCl}]) = RT\ln k_u^{\mathrm{H2O}} + m_u[\mathrm{GdmCl}] \quad (5c)$$

The slopes m^o, m_f and $-m_u$, represent the difference in the amount of denaturant-sensitive surface area buried between the initial and final states for the transition under consideration. The data are fit with a non-linear least-squares algorithm using the Microcal Origin software package.

The relative contact order[35] was calculated according to $\mathrm{RCO} = 1/LN_c \ \Sigma_{i,j}^{\mathrm{Nc}} \Delta L_{i,j}$ where L is the length of the proteins, $\Delta L_{i,j}$ is the sequence separation, N_c is the number of atomic contacts.

Langevin dynamics

The relaxation of the TS models uses implicit solvent LD simulation model and method as described by Shen and Freed.[59,90] The all-atom LD simulations proceed by solving the Langevin equation:

$$m_i \frac{\partial \vec{u}_i}{\partial t} = -\vec{\nabla}_i U + \zeta_i \cdot \vec{u}_i + \vec{A}_i(t),$$

where $\vec{u}_i$ is the velocity of atom i, ζ_i is its friction coefficient, m_i is its mass, $\vec{A}_i(t)$ is the random force acting on atom i, and U is the energy function. In general, the energy function is a combination of an all-atom force field to represent the solute–solute interactions and implicit solvation terms, which attempt to reproduce the influence of the solvent. These energy functions are given by:

$$U_{\mathrm{Solute}} = U_{\mathrm{bond}} + U_{\mathrm{bend}} + U_{\mathrm{torsion}} + U_{\mathrm{improper}} + U_{\mathrm{VDW}},$$
$$U_{\mathrm{Solvent}} = U_{\mathrm{charge}}(\varepsilon) + U_{\mathrm{solvation}}(\sigma)$$

such that the total potential is:

$$U = U_{\text{Solute}} + U_{\text{Solvent}}.$$

The energy function, U_{solute}, is henceforth termed the force field. We use all-atom OPLS[57,58] which has individual energy terms of the form:

$$U_{\text{bond}} = \sum_{\text{bonds}} 1/2 \cdot K_b (b - b_o)^2,$$

$$U_{\text{bend}} = \sum_{\text{bond angles}} 1/2 \cdot K_\theta (\theta - \theta_o)^2,$$

$$U_{\text{torsion}} = \sum_{\text{torsion angles}} K_\phi [1 + \cos(n\phi - \Delta)],$$

$$U_{\text{improper}} = \sum_{\text{improper torsion angles}} 1/2 \cdot K_\chi (\chi - \chi_o)^2,$$

$$U_{\text{VDW}} = \sum_{\text{non bonded pairs}} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r} \right)^{12} - \left(\frac{\sigma_{ij}}{r} \right)^{6} \right],$$

where b is the bond length, θ is the bond angle, χ is the improper torsion angle, ϕ is the torsion angle, the K are the respective force constants, and the subscript o indicates an equilibrium value. The parameter n in the torsion potential is the multiplicity, and δ is the phase. The van der Waals potential contains ε_{ij} the dispersion well depth, σ_{ij} the Lennard-Jones diameter, and r the non-bonded distance.

The first term of the solvation potential U_{Solvent} is a screened Coulomb potential of the form:

$$U_{\text{charge}}(\varepsilon) = \sum_{i<j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}},$$

where $\varepsilon(r_{ij})$ is a Ramstein-Lavery style distance-dependent dielectric "constant",[91] and the q_i are atomic partial charges. This distance-dependent dielectric constant is given by:

$$\varepsilon(r_{ij}) = D_\infty - \frac{D_\infty - D_o}{2} \left(S^2 r_{ij}^2 + 2S r_{ij} + 2 \right) e^{-S r_{ij}},$$

where D_∞ is the bulk dielectric constant of water, D_o is the limit at small distances, and S is a sigmoidal parameter. We use the parameters, $D_\infty = 78.5$, $D_o = 1$, and $S = 0.3$. The second term of the solvation model uses the Ooi-Scheraga solvent-accessible surface area (SASA) potential,[92] which is of the form:

$$U_{\text{solvation}}(\sigma) = \sum_{i=1}^{N} g_i \sigma_i$$

where σ_i is the accessible surface area of a hypersurface bisecting the first solvent shell surrounding protein atom i and g_i is an empirical (free energy) parameter dependent on atom type.

All implicit solvent simulations fix the temperature at 298 K using the random force, and the simulation is performed using a modified, speed enhanced TINKER molecular mechanics package‡.[93] Integration proceeds using the Verlet algorithm with an integration step of 2 fs. The lengths of all bonds are frozen using the RATTLE algorithm,[94] and non-bonded forces are computed using the FAST-LD routine.[95] The frictional forces and random forces are computed using the Pastor-Karplus accessible surface area model[96] with the experimental solvent viscosity of 0.89 cp. Finally, accessible surface areas, atomic friction coefficients, and solvation potentials are updated every 20 fs.

The ψ-value constraints were included in the simulations as follows. The distances between the C^α–C^α atoms of the residues composing the biHis sites that had $\psi = 1$ were allowed to move by ±0.25 Å (sites F, H), ±0.35 Å (sites D,J) or ±0.5 Å (site I) from their native distance. Larger movements were constrained using a flat-well harmonic potential with a spring constant of 100 kcal mol^{-1} Å^{-1}.

Surface area and topological metrics

The surface area was calculated using a probe radius of 1.4 Å. The surface area of the unfolded state was obtained from the average of 500 model denatured states generated using backbone geometries obtained from a coil library followed side-chain insertion using the SCWRL program[97] and "nudged" to eliminate steric overlap (hard sphere) by minor adjustment of backbone dihedral angles§.[61] The relative contact order was calculated according to $1/LN \Sigma^N \Delta L_{ij}$, where L is the length of the protein in amino acid residues, N is the number of contacts within 6 Å, and ΔL_{ij} is the number of residues separating the interacting pair of non-hydrogen atoms. The long-range order was calculated using a contact distance of 6 Å, and a minimum sequence separation distance of 12 residues.

Acknowledgements

We thank Professors S.W. Englander, N. Kallenbach, R.L. Baldwin, K. Plaxco, V. Pande, H. Chan, E. Shakhnovich and members of our group for comments and discussions, and M. Baxa for assistance in the derivations in the Supplementary Data. This work is supported by grants from the NIH (T.R.S. GM55694) and NSF (K.F.F. CHE-031226 and CHE-0416017). A.J. acknowledges the support of Burroughs Wellcome Fund Interfaces #1001774.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.06.041

References

1. Garcia-Mira, M. M., Boehringer, D. & Schmid, F. X. (2004). The folding transition state of the cold shock protein is strongly polarized. *J. Mol. Biol.* **339**, 555–569.
2. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* **5**, 714–720.
3. Gruebele, M. & Wolynes, P. G. (1998). Satisfying turns in folding transitions. *Nature Struct. Biol.* **5**, 662–665.

‡ http://www.dasher.wustl.edu/tinker

§ see http://www.unfolded.uchicago.edu

4. Guo, W., Lampoudi, S. & Shea, J. E. (2004). Temperature dependence of the free energy landscape of the src-SH3 protein domain. *Proteins: Struct. Funct. Genet.* **55**, 395–406.
5. Klimov, D. K. & Thirumalai, D. (2001). Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins: Struct. Funct. Genet.* **43**, 465–475.
6. Lindberg, M., Tangrot, J. & Oliveberg, M. (2002). Complete change of the protein folding transition state upon circular permutation. *Nature Struct. Biol.* **9**, 818–822.
7. Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016–1024.
8. Weikl, T. R. & Dill, K. A. (2003). Folding kinetics of two-state proteins: effect of circularization, permutation, and crosslinks. *J. Mol. Biol.* **332**, 953–963.
9. Yi, Q., Rajagopal, P., Klevit, R. E. & Baker, D. (2003). Structural and kinetic characterization of the simplified SH3 domain FP1. *Protein Sci.* **12**, 776–783.
10. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260–288.
11. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005–1009.
12. Ozkan, S. B., Dill, K. A. & Bahar, I. (2003). Computing the transition state populations in simple protein models. *Biopolymers*, **68**, 35–46.
13. Fersht, A. R., Itzhaki, L. S., elMasry, N. F., Matthews, J. M. & Otzen, D. E. (1994). Single versus parallel pathways of protein folding and fractional formation of structure in the transition state. *Proc. Natl Acad. Sci. USA*, **91**, 10426–10429.
14. Oliveberg, M., Tan, Y. J., Silow, M. & Fersht, A. R. (1998). The changing nature of the protein folding transition state: implications for the shape of the free-energy profile for folding. *J. Mol. Biol.* **277**, 933–943.
15. Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721–729.
16. Otzen, D. E., Kristensen, O., Proctor, M. & Oliveberg, M. (1999). Structural changes in the transition state of protein folding: alternative interpretations of curved chevron plots. *Biochemistry*, **38**, 6499–6511.
17. Grantcharova, V. P., Riddle, D. S. & Baker, D. (2000). Long-range order in the src SH3 folding transition state. *Proc. Natl Acad. Sci. USA*, **97**, 7084–7089.
18. Sanchez, I. E. & Kiefhaber, T. (2003). Hammond behavior versus ground state effects in protein folding: evidence for narrow free energy barriers and residual structure in unfolded states. *J. Mol. Biol.* **327**, 867–884.
19. Wright, C. F., Lindorff-Larsen, K., Randles, L. G. & Clarke, J. (2003). Parallel protein-unfolding pathways revealed and mapped. *Nature Struct. Biol.* **10**, 658–662.
20. Moran, L. B., Schneider, J. P., Kentsis, A., Reddy, G. A. & Sosnick, T. R. (1999). Transition state heterogeneity in GCN4 coiled coil folding studied by using multisite mutations and crosslinking. *Proc. Natl Acad. Sci. USA*, **96**, 10699–10704.
21. Krishna, M. M., Lin, Y., Mayne, L. & Englander, S. W. (2003). Intimate view of a kinetic protein folding intermediate: residue-resolved structure, interactions, stability, folding and unfolding rates, homogeneity. *J. Mol. Biol.* **334**, 501–513.
22. Jackson, S. E. (1998). How do small single-domain proteins fold? *Fold. Des.* **3**, R81–R91.
23. Krantz, B. A. & Sosnick, T. R. (2000). Distinguishing between two-state and three-state models for ubiquitin folding. *Biochemistry*, **39**, 11696–11701.
24. Krantz, B. A., Mayne, L., Rumbley, J., Englander, S. W. & Sosnick, T. R. (2002). Fast and slow intermediate accumulation and the initial barrier mechanism in protein folding. *J. Mol. Biol.* **324**, 359–371.
25. Jacob, J., Krantz, B., Dothager, R. S., Thiyagarajan, P. & Sosnick, T. R. (2004). Early collapse is not an obligate step in protein folding. *J. Mol. Biol.* **338**, 369–382.
26. Krantz, B. A., Dothager, R. S. & Sosnick, T. R. (2004). Discerning the structure and energy of multiple transition states in protein folding using psi-analysis. *J. Mol. Biol.* **337**, 463–475.
27. Krantz, B. A. & Sosnick, T. R. (2001). Engineered metal binding sites map the heterogeneous folding landscape of a coiled coil. *Nature Struct. Biol.* **8**, 1042–1047.
28. Sosnick, T. R., Dothager, R. S. & Krantz, B. A. (2004). Differences in the folding transition state of ubiquitin indicated by phi and psi analyses. *Proc. Natl Acad. Sci. USA*, **101**, 17377–17382.
29. Matthews, C. R. (1987). Effects of point mutations on the folding of globular proteins. *Methods Enzymol.* **154**, 498–511.
30. Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771–782.
31. Goldenberg, D. P. (1992). Mutational analysis of protein folding and stability. In *Protein Folding* (Creighton, T. E., ed), pp. 353–403. W. H. Freeman, New York.
32. Liguri, G., Camici, G., Manao, G., Cappugi, G., Nassi, P., Modesti, A. & Ramponi, G. (1986). A new acylphosphatase isoenzyme from human erythrocytes: purification, characterization, and primary structure. *Biochemistry*, **25**, 8089–8094.
33. Thunnissen, M. M., Taddei, N., Liguri, G., Ramponi, G. & Nordlund, P. (1997). Crystal structure of common type acylphosphatase from bovine testis. *Structure*, **5**, 69–79.
34. Saudek, V., Wormald, M. R., Williams, R. J., Boyd, J., Stefani, M. & Ramponi, G. (1989). Identification and description of beta-structure in horse muscle acylphosphatase by nuclear magnetic resonance spectroscopy. *J. Mol. Biol.* **207**, 405–415.
35. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994.
36. Kim, D. E., Yi, Q., Gladwin, S. T., Goldberg, J. M. & Baker, D. (1998). The single helix in protein L is largely disrupted at the rate-limiting step in folding. *J. Mol. Biol.* **284**, 807–815.
37. Ozkan, S. B., Bahar, I. & Dill, K. A. (2001). Transition states and the meaning of Phi-values in protein folding kinetics. *Nature Struct. Biol.* **8**, 765–769.
38. Bulaj, G. & Goldenberg, D. P. (2001). Phi-values for BPTI folding intermediates and implications for transition state analysis. *Nature Struct. Biol.* **8**, 326–330.
39. Northey, J. G., Maxwell, K. L. & Davidson, A. R. (2002). Protein folding kinetics beyond the phi value: using multiple amino acid substitutions to investigate

the structure of the SH3 domain folding transition state. *J. Mol. Biol.* **320**, 389–402.
40. Leffler, J. E. (1953). Parameters for the description of transition states. *Science*, **107**, 340–341.
41. Eyring, H. (1935). The activated complex in chemical reactions. *J. Chem. Phys.* **3**, 107–115.
42. Fersht, A. R. (2004). ϕ Value versus ψ analysis. *Proc. Natl Acad. Sci. USA*, **101**, 17327–17328.
43. Sanchez, I. E. & Kiefhaber, T. (2003). Origin of unusual phi-values in protein folding: evidence against specific nucleation sites. *J. Mol. Biol.* **334**, 1077–1085.
44. Miguel, A., de los Rios, B. K. M., Wildes, D., Sosnick, T. R., Marqusee, P. W.-S., Plaxco, K. W. & Ruczinski, I. (2006). On the precision of experimentally determined protein folding rates and f-values. *Protein Sci.* **15**, 553–563.
45. Religa, T. L., Markson, J. S., Mayor, U., Freund, S. M. & Fersht, A. R. (2005). Solution structure of a protein denatured state and folding intermediate. *Nature*, **437**, 1053–1056.
46. Taddei, N., Chiti, F., Fiaschi, T., Bucciantini, M., Capanni, C., Stefani, M. *et al.* (2000). Stabilisation of alpha-helices by site-directed mutagenesis reveals the importance of secondary structure in the transition state for acylphosphatase folding. *J. Mol. Biol.* **300**, 633–647.
47. Monti, M., Garolla di Bard, B. L., Calloni, G., Chiti, F., Amoresano, A., Ramponi, G. & Pucci, P. (2004). The regions of the sequence most exposed to the solvent within the amyloidogenic state of a protein initiate the aggregation process. *J. Mol. Biol.* **336**, 253–262.
48. Bodenreider, C. & Kiefhaber, T. (2005). Interpretation of protein folding psi values. *J. Mol. Biol.* **351**, 393–401.
49. Sancho, J., Meiering, E. M. & Fersht, A. R. (1991). Mapping transition states of protein unfolding by protein engineering of ligand-binding sites. *J. Mol. Biol.* **221**, 1007–1014.
50. Jia, Y. Q. (1991). Crystal radii and effective ionic radii of the rare earth ions. *J. Solid State Chem.* **95**, 184.
51. Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, **7**, 284–304.
52. Krantz, B. A., Moran, L. B., Kentsis, A. & Sosnick, T. R. (2000). D/H amide kinetic isotope effects reveal when hydrogen bonds form during protein folding. *Nature Struct. Biol.* **7**, 62–71.
53. Krantz, B. A., Srivastava, A. K., Nauli, S., Baker, D., Sauer, R. T. & Sosnick, T. R. (2002). Understanding protein hydrogen bond formation with kinetic H/D amide isotope effects. *Nature Struct. Biol.* **9**, 458–463.
54. Feng, H., Vu, N. D., Zhou, Z. & Bai, Y. (2004). Structural examination of Phi-value analysis in protein folding. *Biochemistry*, **43**, 14325–14331.
55. Capaldi, A. P., Kleanthous, C. & Radford, S. E. (2002). Im7 folding mechanism: misfolding on a path to the native state. *Nature Struct. Biol.* **9**, 209–216
56. Nishimura, C., Dyson, H. J. & Wright, P. E. (2006). Identification of native and non-native structure in kinetic folding intermediates of apomyoglobin. *J. Mol. Biol.* **355**, 139–156.
57. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. (2001). Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, **105**, 6474–6487.
58. Kaminski, G. A., Friesner, R. A., Tirado-Rives, J. & Jorgensen, W. L. (2000). OPLS-AA/L force field for proteins: Using accurate quantum mechanical data. *Abstr. Papers Am. Chem. Soc.* **220**, U279.
59. Shen, M. Y. & Freed, K. F. (2002). Long time dynamics of met-enkephalin: Comparison of explicit and implicit solvent models. *Biophys. J.* **82**, 1791–1808.
60. Shen, M. Y. & Freed, K. F. (2002). All-atom fast protein folding simulations: The villin headpiece. *Proteins: Struct. Funct. Genet.* **49**, 439–445.
61. Jha, A. K., Colubri, A., Freed, K. F. & Sosnick, T. R. (2005). Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl Acad. Sci. USA*, **102**, 13099–13104.
62. Lindberg, M. O., Haglund, E., Hubner, I. A., Shakhnovich, E. I. & Oliveberg, M. (2006). Identification of the minimal protein-folding nucleus through loop-entropy perturbations. *Proc. Natl Acad. Sci. USA*, **103**, 4083–4088.
63. Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010–1016.
64. Viguera, A. R. & Serrano, L. (1997). Loop length, intramolecular diffusion and protein folding. *Nature Struct. Biol.* **4**, 939–946.
65. Martinez, J. C., Viguera, A. R., Berisio, R., Wilmanns, M., Mateo, P. L., Filimonov, V. V. & Serrano, L. (1999). Thermodynamic analysis of alpha-spectrin SH3 and two of its circular permutants with different loop lengths: discerning the reasons for rapid folding in proteins. *Biochemistry*, **38**, 549–559.
66. Viguera, A. R., Serrano, L. & Wilmanns, M. (1996). Different folding transition states may result in the same native structure. *Nature Struct. Biol.* **3**, 874–880.
67. Hennecke, J., Sebbel, P. & Glockshuber, R. (1999). Random circular permutation of DsbA reveals segments that are essential for protein folding and stability. *J. Mol. Biol.* **286**, 1197–1215.
68. Grantcharova, V., Alm, E. J., Baker, D. & Horwich, A. L. (2001). Mechanisms of protein folding. *Curr. Opin. Struct. Biol.* **11**, 70–82.
69. Otzen, D. E. & Fersht, A. R. (1998). Folding of circular and permuted chymotrypsin inhibitor 2: retention of the folding nucleus. *Biochemistry*, **37**, 8139–8146.
70. Klimov, D. K. & Thirumalai, D. (2005). Symmetric connectivity of secondary structure elements enhances the diversity of folding pathways. *J. Mol. Biol.* **353**, 1171–1186.
71. Kim, D. E., Fisher, C. & Baker, D. (2000). A Breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971–984.
72. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry*, **33**, 10026–10036.
73. Sosnick, T. R., Mayne, L. & Englander, S. W. (1996). Molecular collapse: The rate-limiting step in two-state cytochrome c folding. *Proteins: Struct. Funct. Genet.* **24**, 413–426.
74. Guo, Z. Y. & Thirumalai, D. (1995). Kinetics of protein-folding: nucleation mechanism, time scales, and pathways. *Biopolymers*, **36**, 83–102.
75. Goldenberg, D. P. (1999). Finding the right fold. *Nature Struct. Biol.* **6**, 987–990.
76. Bai, Y., Zhou, H. & Zhou, Y. (2004). Critical nucleation size in the folding of small apparently two-state proteins. *Protein Sci.* **13**, 1173–1181.
77. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D. & Finkelstein, A. V. (2003). Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* **12**, 2057–2062.
78. Paci, E., Lindorff-Larsen, K., Dobson, C. M., Karplus,

M. & Vendruscolo, M. (2005). Transition state contact orders correlate with protein folding rates. *J. Mol. Biol.* **352**, 495–500.
79. Wallin, S. & Chan, H. S. (2006). Conformational entropic barriers in topology-dependent protein folding: perspectives from a simple native-centric polymer model. *J. Phys.: Condens. Matter*, **18**, S307–S328.
80. Gillespie, B. & Plaxco, K. W. (2004). Using protein folding rates to test protein folding theories. *Annu. Rev. Biochem.* **73**, 837–859.
81. Gromiha, M. M. & Selvaraj, S. (2001). Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. *J. Mol. Biol.* **310**, 27–32.
82. Makarov, D. E. & Plaxco, K. W. (2003). The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.* **12**, 17–26.
83. Merlo, C., Dill, K. A. & Weikl, T. R. (2005). Phi values in protein-folding kinetics have energetic and structural components. *Proc. Natl Acad. Sci. USA*, **102**, 10171–10175.
84. McCallister, E. L., Alm, E. & Baker, D. (2000). Critical role of beta-hairpin formation in protein G folding. *Nature Struct. Biol.* **7**, 669–673.
85. Guerois, R. & Serrano, L. (2000). The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* **304**, 967–982.
86. Jones, S., Smith, D. P. & Radford, S. E. (2003). Role of the N and C-terminal strands of beta 2-microglobulin in amyloid formation at neutral pH. *J. Mol. Biol.* **330**, 935–941.
87. Jones, S., Reader, J. S., Healy, M., Capaldi, A. P., Ashcroft, A. E., Kalverda, A. P. *et al.* (2000). Partially unfolded species populated during equilibrium denaturation of the beta-sheet protein Y74W apo-pseudoazurin. *Biochemistry*, **39**, 5672–5682.
88. Calamai, M., Chiti, F. & Dobson, C. M. (2005). Amyloid fibril formation can proceed from different conformations of a partially unfolded protein. *Biophys. J.* **89**, 4201–4210.
89. Larsen, C. N., Krantz, B. A. & Wilkinson, K. D. (1998). Substrate specificity of deubiquitinating enzymes: ubiquitin C-terminal hydrolases. *Biochemistry*, **37**, 3358–3368.
90. Zaman, M. H., Shen, M. Y., Berry, R. S., Freed, K. F. & Sosnick, T. R. (2003). Investigations into sequence and conformational dependence of backbone entropy, inter-basin dynamics and the Flory isolated-pair hypothesis for peptides. *J. Mol. Biol.* **331**, 693–711.
91. Ramstein, J. & Lavery, R. (1988). Energetic coupling between DNA bending and base pair opening. *Proc. Natl Acad. Sci. USA*, **85**, 7231–7235.
92. Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl Acad. Sci. USA*, **84**, 3086–3090.
93. Ponder, J. W. R. S., Kundrot, C., Huston, S., Dudek, M., Kong, Y., Hart, R. *et al.* (1999). *TINKER: Software Tools for Molecular Design*, 3.7 edit. Washington University, St. Louis, MO.
94. Anderson, H. C. (1983). Rattle: a velocity version of the shake algorithm for molecular dynamics calculations. *J. Comp. Phys.* **52**, 24–34.
95. Shen, M. Y. (2002). Computer simulations of proteins (Ph.D. thesis). University of Chicago.
96. Pastor, R. W. & Karplus, M. (1988). Parametrization of the friction constant for stochastic simulations of polymers. *J. Phys. Chem.* **92**, 2636–2641.
97. Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L., Jr (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **12**, 2001–2014.
98. Thunnissen, M. M., Agango, E. G., Taddei, N., Liguri, G., Cecchi, C., Pieri, A. *et al.* (1995). Crystallisation and preliminary X-ray analysis of the 'common-type' acylphosphatase. *FEBS Letters*, **364**, 243–244.
99. Maxwell, K. L., Wildes, D., Zarrine-Afsar, A., De Los Rios, M. A., Brown, A. G., Friel, C. T. *et al.* (2005). Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci.* **14**, 602–616.

Edited by C. R. Matthews

(Received 13 March 2006; received in revised form 12 June 2006; accepted 16 June 2006)
Available online 7 July 2006

Note added in proof: Similar to our results, Shakhnovich *et al.* conducted all-atom unfolding simulations of Protein G using phi-values as constraints and found that three pathways converged to common TSE having native-like topology although with the some amount of structural heterogeneity (Hubner, I.A., Shimada, J. & Shakhnovich, E.I. (2004). *J Mol Biol*, **336**, 745–761). In similar studies, they observed that three SH3 homologs also unfolded via a specific nucleus, although the TSE was less structured and more polarized than that for Protein G (Hubner, I.A., Edmonds, K.A. & Shakhnovich, E.I. (2005). *J. Mol. Biol.*, **349**, 424–434).