J|A|C|S
ARTICLES
Published on Web 06/17/2004

Quantifying the Roughness on the Free Energy Landscape: Entropic Bottlenecks and Protein Folding Rates

Leslie L. Chavez,[†] José N. Onuchic,[†] and Cecilia Clementi*,[‡]

Contribution from the Center for Theoretical Biological Physics and Department of Physics, University of California at San Diego, La Jolla, California 92093, Department of Chemistry and W. M. Keck Center for Computational and Structural Biology, Rice University, 6100 Main Street, Houston, Texas 77005, and Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030

Received January 27, 2004; E-mail: cecilia@rice.edu

Abstract: The prediction of protein folding rates and mechanisms is currently of great interest in the protein folding community. A close comparison between theory and experiment in this area is promising to advance our understanding of the physical−chemical principles governing the folding process. The delicate interplay of entropic and energetic/enthalpic factors in the protein free energy regulates the details of this complex reaction. In this article, we propose the use of topological descriptors to quantify the amount of heterogeneity in the configurational entropy contribution to the free energy. We apply the procedure to a set of 16 two-state folding proteins. The results offer a clean and simple theoretical explanation for the experimentally measured folding rates and mechanisms, in terms of the intrinsic entropic roughness along the populated folding routes on the protein free energy landscape.

I. Introduction

Experimental evidence and its comparison with theoretical models have shown proteins to be robust folders, capable of folding in many environments and despite many mutations to the amino acid sequence.[1−4] Results on small (single domain) proteins suggest that evolution has selected amino acid sequences with low enough energetic frustration in the free energy landscape that sensitivity to a particular mutation appears to be an exception, not the rule.[5−7] This has two main observable effects: these proteins fold quickly—on the scale of microseconds to seconds in typical laboratory conditions—and the structural details of the folding mechanism are predominantly due to what is usually referred to as *topological frustration*.[8−17] *Topological frustration* (though nonstandard terminology) effectively evokes the ruggedness of the folding landscape that arises as chain connectivity interplays with the energy bias to reach the native state. More rigorously, this ruggedness results from the heterogeneous loss of conformation entropy associated with the formation of partially folded structures throughout the free energy landscape. For proteins where the heterogeneity of the conformation entropy is much larger than the energetic heterogeneity, the main folding route(s) in the free energy landscape are strongly constrained and shaped by the protein topology.[18,19] This implies that for proteins with a large *topological frustration* the main features of the folding routes can be traced back from the geometrical information contained in the native state. This backtracking functionality of the native state has spawned the birth and growth of a number of theoretical models designed to recover the folding mechanism of single domain proteins by exploiting the information contained in the native structure.[16,17,20,21] In this context, we have proposed the use of an energetically unfrustrated Hamiltonian,[22] in both minimalist[11−13] and all-atom protein representations,[14]

[†] University of California at San Diego.
[‡] Rice University and Baylor College of Medicine.

(1) Itzhaki, L. S.; Otzen, D. E.; Fersht, A. R. *J. Mol. Biol.* **1995**, *254*, 260−288.
(2) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 167−195.
(3) Kuhlman, B.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10383−10388.
(4) Bryngelson, J. D.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524−7528.
(5) Brockwell, D. J.; Smith, D. A.; Radford, S. E. *Curr. Opinion Struct. Biol.* **2000**, *10*, 16−25.
(6) Grantcharova, V.; Alm, E. J.; Baker, D.; Horwich, A. L. *Curr. Opinion Struct. Biol.* **2001**, *11*, 70−82.
(7) Gunasekaran, K.; Eyles, S. J.; Hagler, A. T.; Gierasch, L. M. *Curr. Opinion Struct. Biol.* **2001**, *11*, 83−93.
(8) Alm, E.; Morozov, A. V.; Kortemme, T.; Baker, D. *J. Mol. Biol.* **2002**, *322*, 463−476.
(9) Dinner, A. R.; Karplus, M. *Nat. Struct. Biol.* **2001**, *8*, 21−22.
(10) Nymeyer, H.; Socci, N. D.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 634−639.
(11) Clementi, C.; Nymeyer, H.; Onuchic, J. N. *J. Mol. Biol.* **2000**, *298*, 937−953.
(12) Clementi, C.; Jennings, P. A.; Onuchic, J. N. *J. Mol. Biol.* **2001**, *311*, 879−890.
(13) Clementi, C.; Jennings, P. A.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5871−5876.
(14) Clementi, C.; García, A. E.; Onuchic, J. N. *J. Mol. Biol.* **2003**, *326*, 933−954.
(15) Klimov, D.; Thirumalai, D. *J. Mol. Biol.* **2002**, *317*, 721−737.
(16) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. *J. Chem. Phys.* **1999**, *111*, 10375−10380.
(17) Plaxco, K. W.; Simmons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985−994.
(18) Plotkin, S. S.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 6509−6514.
(19) Plotkin, S. S.; Onuchic, J. N. *J. Chem. Phys.* **2002**, *116*, 5263−5283.
(20) Koga, N.; Shoji, T. *J. Mol. Biol.* **2001**, *313*, 171−180.
(21) Gromiha, M. M.; Selvaraj, S. *J. Mol. Biol.* **2001**, *310*, 27−32.
(22) Taketomi, H.; Ueda, Y.; Go, N. *Int. J. Pept. Protein Res.* **1975**, *7*, 445−459.

10.1021/ja049510+ CCC: $27.50

to sample the relevant structures populated in the transition-state ensemble and/or intermediate state during the folding process.

Although a general understanding of protein folding is emerging from studies of relatively small proteins, the existence of a large variety of folding scenarios is becoming increasingly clear: the measured folding rates (k_f, in units of s^{-1}) for proteins of about 100 residues have been found to span almost 6 orders of magnitude (from microseconds for simple helical proteins[23,24] to seconds for more complex topologies[25,26]); the large variation in the degree of "structural polarization" detected at the folding transition state for different proteins reflects a large variation in the regions of the free energy landscape populated during folding. The transition-state structures emerging from experiments range from the formation of a very localized folding nucleus (suggesting a more "pathway-like" folding process)[27,28] to the population of a large ensemble of different partially folded structures (suggesting a more "funnel-like" folding landscape). Moreover, as the latest theoretical and experimental findings are generally confirming that proteins with a similar native fold share a similar folding mechanism,[5,7] they are also bringing to light remarkable exceptions.[5,29−33] These evidences call for a more quantitative understanding of the specific factors shaping the protein landscape.

The importance of entropic factors on the free energy landscape of a large number of small proteins suggests that a deeper understanding of protein conformation entropy may prove fundamental toward a more quantitative understanding of the folding mechanisms.[34] In this article, we take a first step toward a quantitative characterization of the roughness on the folding free energy landscape due to the heterogeneity in the conformation entropy.

To practically perform this analysis, we built a database of energetically unfrustrated single domain proteins. This database represents the extrapolation of the minimal (energetic) frustration principle to the limit of completely unfrustrated protein-like chains. The study of the folding landscape on this computer-generated, energetically unfrustrated protein world allows us to concentrate on the features determined solely by the backbone topology (i.e., configuration entropy). Toward this goal, we define and use several theoretical probes to assess the degree of structural heterogeneity at different stages of the folding process for all the proteins in our database and examine the results synoptically with the available experimental data. The amount of entropic roughness emerging from this analysis on different proteins provides a clean explanation for the different folding scenarios experimentally detected: bottlenecks in the configuration entropy are identified along the folding routes of slow folding proteins, whereas the smooth entropy landscapes associated with the fastest folders leave more room for energetic perturbations (sequence dependence) to shape the minimal free energy pathways to the native state.

This study offers a solid starting point toward a quantitative understanding of the delicate entropy/energy balance shaping the folding free energy landscape and offers an essential step for connecting theory and experiment in protein folding. An extension of the analysis to longer proteins, exhibiting a more complex kinetics, is already in progress.

II. A Representative Database for an Unfrustrated Protein World

The analysis is performed on a database of completely unfrustrated protein models. Proteins in the database are selected to span more than 6 orders of magnitude in their experimentally measured folding rates and to have different overall folding topology and secondary structure composition. Sixteen two-state, single domain proteins (ranging from 36 to 115 residues in length) are considered. Table 1 summarizes the structural information and folding rates of the selected proteins.

The database is built by associating a C-α representation to each protein and dressing it with a Gō-like potential. Simulation procedures and more technical details are provided in the Supporting Information (section A). The selection of a Gō-like potential is motivated by our goal of separating the different sources of "frustration" in the free energy that arise from either the conformation entropy or interaction energy terms. By a priori removing any energetic heterogeneities from the protein Hamiltonian, we can concentrate on the effect of conformational entropic heterogeneity on the folding landscape. Elsewhere we have considered the effect of increasing energetic frustration on folding, for a fixed protein structure,[35] and observed that the induced rate enhancement/reduction is limited to less than 1 order of magnitude even up to a reasonably large amount of frustration. Energetic heterogeneity thus cannot be used to explain the much larger variation of folding rates experimentally observed for single-domain, two-state folding proteins.[17,36]

For each of the selected proteins, kinetics and thermodynamic quantities are extracted from simulations. The amount of "frustration" (roughness) on the most relevant regions of the free energy landscape is then quantified by properly defined indicators (see Section III.B).

III. Results and Discussion

A. Simulation Rates versus Experimental Rates. Figure 1 compares the folding rates obtained from simulations with the corresponding experimental data. The rates obtained with the C-α unfrustrated protein models correlate remarkably well ($r \simeq 0.9$) with the experimental rates.

The selection of experimental data on folding rates has been the subject of debate in the protein folding community.[9,37−39]

(23) Wittung-Stafshede, P.; Lee, J. C.; Winkler, J. R.; Gray, H. B. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 6587−6590.
(24) Chang, I. J.; Lee, J. C.; Winkler, J. R.; Gray, H. B. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 3838−3840.
(25) Chiti, F.; Taddei, N.; van Nuland, N. A. J.; Magherini, F.; Stefani, M.; Ramponi, G. M.; Dobson, C. M. *J. Mol. Biol.* **1998**, *283*, 883−891.
(26) Roumestand, C.; Boyer, M.; Guignard, L.; Barthe, P.; Royer, C. A. *J. Mol. Biol.* **2001**, *312*, 247−259.
(27) Martinez, J. C.; Pisabarro, M. T.; Serrano, L. *Nat. Struct. Biol.* **1998**, *5*, 721−729.
(28) Grantcharova, V. P.; Riddle, D. S.; Santiago, J. V.; Baker, D. *Nat. Struct. Biol.* **1998**, *5*, 714−720.
(29) Burns, L. L.; Dalessio, P. M.; Ropson, I. J. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 107−118.
(30) Dalessio, P. M.; Ropson, I. J. *Biochemistry* **2000**, *39*, 860−871.
(31) Ferguson, N.; Capaldi, A. P.; James, R.; Kleanthous, C.; Radford, S. E. *J. Mol. Biol.* **1999**, *286*, 1597−1608.
(32) Kim, D.; Fisher, C.; Baker, D. *J. Mol. Biol.* **2000**, *298*, 971−984.
(33) McCallister, E. L.; Alm, E.; Baker, D. *Nat. Struct. Biol.* **2000**, *7*, 669−673.
(34) Pappu, R. V.; Srinivasan, R.; Rose, G. D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 12565−12570.
(35) Clementi, C.; Plotkin, S. S. *Protein Sci.*, in press.
(36) Plaxco, K. W.; Simons, K. T.; Ruczinski, I.; Baker, D. *Biochemistry* **2000**, *39*, 11177−11183.
(37) Lindberg, M. O.; Tangrot, J.; Otzen, D. E.; Dolgikh, D. A.; Finkelstein, A.; Oliveberg, M. *J. Mol. Biol.* **2001**, *314*, 891−900.
(38) Ivankov, D. N.; Finkelstein, A. V. *Biochemistry* **2001**, *40*, 9957−9961.
(39) Finkelstein, A. V.; Badretdinov, A. *Folding Des.* **1997**, *2*, 115−121.

Table 1. Summary of Most Relevant Structural and Folding Features for All Two-State Folding Proteins Used in Our Study

	protein (or protein domain)	L	structural information	k_f (s^{-1})	ref
HP36	headpiece subdomain of chicken villin	36	all helical; smallest naturally occurring, independently folding protein domain	$\sim 10^5$	63
λ_{6-85}	monomeric version of N-terminal domain of λ-repressor protein	80	five-helix bundle	10^4-10^5	64
Psbd	peripheral subunit-binding domain of pyruvate dehydrogenase multi-enzyme complex	43	very small three-helix bundle	$\sim 10^4$	65
N−L9	N-terminal domain of ribosomal protein L9	56	three-stranded antiparallel β-sheet sandwiched between two helices	$\sim 10^3$	60
CspB	cold-shock protein from *Bacillus subtilis*	67	small β-barrel	10^2-10^3	66−68
PtG	IgG-binding domain of protein G	56	four-stranded β-sheet spanned by an α-helix (similar to PtL)	10^2-10^3	33, 69
CI2	chymotrypsin inhibitor 2	64	six-stranded β-sheet packed against an α-helix	$\sim 10^2$	70−72
PtL	IgG-binding domain of protein L	61	four-stranded β-sheet spanned by an α-helix (similar to PtG)	$\sim 10^2$	59
Im9	colicin-binding bacterial immunity protein 9	86	four-helix bundle	$\sim 10^2$	73
SH3	sarcoma homology 3 domain	57	two antiparallel β-sheets orthogonally packed	$10-10^2$	61, 74, 75
TI−I27	immunoglobulin-like domain from human muscle titin protein	89	two antiparallel β-sheets packed against each other (Greek key topology)	1−10	62, 76
HPr	histidine containing photocarrier protein	85	three α-helices packed against a four-stranded antiparallel β-sheet	1−10	77
MerP	mercury binding protein	72	antiparallel four-stranded β-sheet, with two helices packed on one side	~ 1	49
TWIg	immunoglobulin-like domain from *Caenorhabditis elegans* twitchin protein	93	two antiparallel β-sheets packed against each other (Greek key topology)	~ 1	62
AcP	human enzyme muscle acylphosphatase	98	two antiparallel α-helices packed against a five-stranded β-sheet	$\sim 10^{-2}$	25, 78
P13	oncogene product of MTCP1 gene, involved in T-cell leukemia	115	filled β-barrel	$10^{-2}-10^{-1}$	26, 79

Clearly, folding rates depend on external parameters such as temperature and denaturant concentration. It is paramount to have a robust criterion in the measurements of rates in both simulation and experiments to have a meaningful and quantitative comparison. We have addressed this issue in depth (the details are described in the Supporting Information section C.1): there is absolutely no arbitrariness nor any adjustable parameters in the selection of data used here.

An important point emerging from the rate analysis is on the definition of coherent units for the measured physical quantities in the comparison between theoretical and experimental data, as detailed in the Supporting Information section C.

B. Definition of Structural Probes. Several empirical parameters have been proposed to summarize key characteristics of a protein topology and as such are able to correctly order protein folding rates of single domain proteins.[16,17,20,21,36,40−42] The most famous of these parameters is the *contact order*, originally introduced by Plaxco et al.[17,36] All proposed topological parameters similarly condense the information of a protein native state into a single number, which is undoubtedly very convenient for comparing theoretical results with folding rates and supposedly related experimental quantities (such as β-values or *m* values. See, for example, ref 40). The realization that the kinetics of such a complex reaction as protein folding could be essentially summarized by simple a priori considerations of the native state geometry profoundly affected the field and jump-started a new generation of theoretical models.[43] However, a single parameter cannot fully explain similarities and differences in different proteins' folding mechanisms. Recently, proteins have been found for which folding kinetics seem to escape the predictive power of existing topological parameters (see, for example, refs 26, 44, and 45). We propose here a deeper analysis of the determinants of folding kinetics by monitoring the evolution of appropriate structural probes along the folding landscape.

B.1. Topological Parameters along the Folding Landscape. To fully explore the connection between the folding process and topology, it is necessary to look beyond the information contained in the native state and examine the progression of topological descriptors from the unfolded to the folded ensemble.

We use the reaction coordinate Q (see Supporting Information section B) to chart the protein's progress through configuration space, starting from those states accessible to a floppy-chain molecule to those that define its native form. The ensemble of states defined by each value of Q is dissected by means of several functions, properly defined to capture the progression of entropic/structural information as the folding proceeds.

The first function we consider is the *route measure*, $R(Q)$, that mirrors the route entropy (similar in concept to mixing entropy; see, for example, ref 46). We then introduce two partner functions that we call *effective loop length*, $L_{\text{eff}}(Q)$, and *partial contact order*, $\text{pCO}(Q)$. The effectiveness of these functions in capturing the intrinsic roughness of a folding landscape is illustrated in the following sections by detailing their behavior on four proteins from our model set: AcP, SH3, PtG, and Psbd. These proteins are selected because they span the whole range of folding rates and are well documented in both theory and experiment. The corresponding results for the remaining proteins are fully consistent with what is detailed here.

(40) Micheletti, C. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 74−84.
(41) Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. *Protein Sci.* **2003**, *12*, 2057−2062.
(42) Galzitskaya, O. V.; Garbuzynskiy, S. O.; Ivankov, D. N.; Finkelstein, A. V. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 162−166.
(43) Baker, D. *Nature* **2000**, *405*, 39−42.
(44) Jones, K.; Wittung-Stafshede, P. *J. Am. Chem. Soc.* **2003**, *125*, 9606−9607.
(45) Tang, K. S.; Fersht, A. R.; Itzhaki, L. S. *Structure* **2003**, *11*, 67−73.
(46) Plotkin, S. S.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 6509−6514.

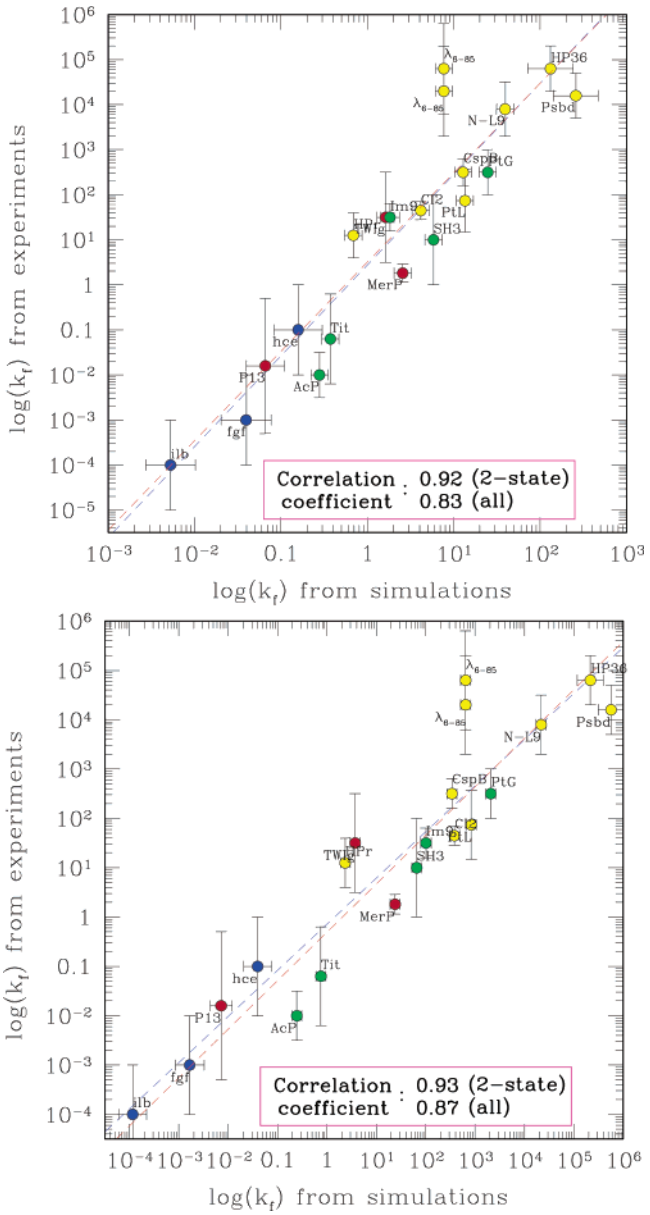

Figure 1. Folding rates from simulations correlate remarkably well with experimentally determined folding rates. The "raw" rates from simulations are shown in (a), while the rates corrected to have the same physical units (see text for details) are shown in (b). In both (a) and (b), yellow dots represent rates experimentally measured at the protein melting point (T_f), in absence of denaturant; green dots represent proteins for which enough experimental data are available to extrapolate the rate to T_f in pure water by using a Brønsted plot, as described in the Supporting Information section C. Proteins marked with red have insufficient data to permit a similar extrapolation; experimental rates for these proteins are selected among the available data as measured in the condition closest to zero stability in pure water. Blue dots correspond to folding rates of three three-state folding proteins (Interleukin-1β IL1b, fibroblast growth factor FGF, and hisactiphilin Hce) from a different study (C. Clementi, unpublished results), reported here to show for comparison.

B.2. Route Measure. The route measure outlines the free energy landscape by showing the breadth of configuration space sampled by the protein as it folds, at different values of Q (i.e., at different stages of the folding reaction). This idea is quantified by measuring the fraction of configurations that are actually accessible among all the possible ones with the same degree of nativeness, from the unfolded to the folded extremes on the free energy landscape. A more direct measure of the fraction of conformational space accessible at different stages of the folding process could be obtained by monitoring the total configurational entropy as a function of the reaction coordinate. However, the definition of an entropy function on the folding landscape would require multiple approximations (see, for example, refs 19 and 46−48). For this reason, we monitor here the evolution of the route measure along the folding reaction: The route measure is cleanly defined, yet directly related to the route entropy, an important component of the total configurational entropy of the protein chain.[19,46]

The route measure function is defined by the following equation:

$$R(Q) = \sum_{i=1}^{M} \frac{\langle(\langle Q_i\rangle_Q - Q)^2\rangle_Q}{MQ(1-Q)} \qquad (1)$$

where Q_i, $i = 1, ..., M$ are the native contacts. In any given configurations $Q_i = 1$ if contact i is formed, $Q_i = 0$ if it is open. The ensemble averages $\langle\cdot\rangle_Q$ are evaluated over all structures with the same fraction $Q = 1/M\sum_{i=1}^{M} Q_i$ of native contacts formed. For each contact i, the frequency of occurrence is determined across the ensemble of configurations at a value Q, yielding the probability of contact formation $0 \leq \langle Q_i\rangle_Q \leq 1$. The distribution of these probabilities is normalized by the maximum number of routes possible. To clarify the meaning of route measure, consider a protein in the process of folding, at a stage such that a fraction Q of total M native contacts are made. There are two extreme values in the route measure:

$R(Q) = 1$. At this value, there is only one populated structure: the ensemble Q contains only one configuration. That means that in any folding event, the same QM native contacts are always formed, $\langle Q_i(Q)\rangle = 1$, while the remaining $(1 - Q)M$ are never made, and $\langle Q_i(Q)\rangle = 0$. In this case a particular folding pathway emerges, while the rest of the landscape is totally inaccessible. In other words, the local folding free energy landscape is extremely rough, with a very narrow accessible path surrounded by much higher peaks.

$R(Q) = 0$. Every native contact is formed with the same probability as all the other native contacts ($\langle Q_i(Q)\rangle = Q$ for all contacts). The folding does not proceed through a well-defined pathway, rather, the local free energy landscape is completely flat: any configuration consistent with the selected Q value is equally accessible.

Both in the unfolded ($Q \approx 0$) and folded ($Q \approx 1$) state $R(Q) = 0$, by definition. Because of the large number of partially folded configurations, at intermediate folding stages $R(Q)$ is not expected to take one of the extreme values. Nevertheless, the evolution of $R(Q)$ from ($Q \approx 0$) to ($Q \approx 1$) reflects the local landscape geography. Where the route measure is lower (more routes), the contact formation is largely unordered: all the available configuration space is represented by the partial configurations making the free energy landscape. Where the route measure is higher (fewer routes), the order of contact formation does matter; if $R(Q)$ increases in proceeding from lower to higher Q values, that means that some configurations that are accessible at a certain stage of folding are very unlikely to advance toward more structured states. On the contrary, a

(47) Plotkin, S. S.; Onuchic, J. N. *Q. Rev. Biophys.* **2002**, *35*, 111−167.
(48) Plotkin, S. S.; Onuchic, J. N. *Q. Rev. Biophys.* **2002**, *35*, 205−286.

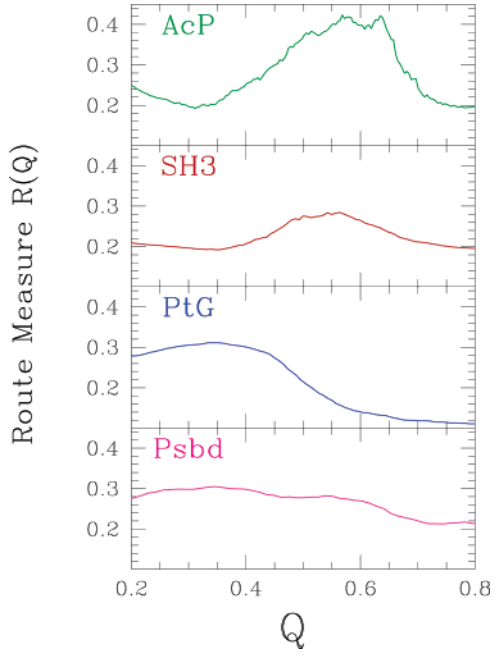

Figure 2. Route measure $R(Q)$ calculated for four of our simulated proteins (from top to bottom: AcP, SH3, PtG, and Psbd). The greater the route measure, the fewer pathways are available for the protein to progress from the unfolded to the folded state. Being more routed may either help or hurt the folding rate; PtG and Psbd are more routed early on, limiting the search through configuration space later, whereas AcP and SH3 encounter a routing barrier in the transition region ($Q \simeq 0.5$), showing fewer correct paths leading to the native state.

decreasing $R(Q)$ can be interpreted as an increase on the local smoothness of the landscape. The behavior of $R(Q)$ on the selected representative proteins demonstrates the usefulness and meaning of this function. The results are shown in Figure 2 and commented in the following, from the slowest to the fastest protein.

AcP has mild routing early in the folding process so the protein may sample a large volume of configuration space. A large bottleneck region slows the folding of this protein around the transition-state barrier ($Q \simeq 0.4-0.5$). Diffuse structure flickers between conformations before the transition-state region, then the formation must become more ordered; a specific structure is required for the protein to fold. If native contacts form out of the preferred route, it may be overall easier (faster) to partially unfold than evolve in the folding. This may be seen as backtracking from an "entropic trap". Unfolded AcP may sample a large region of the configuration space available to it before it manages to pass through the narrow bottleneck that leads to the native state. These effects cause AcP to be one of the slowest two-state proteins.

SH3 is a slower folding protein[49] and shows little routing initially (similar to AcP), and then a small bump appears in the route measure curve. This indicates that at $Q \simeq 0.5$ some portion of the protein is more likely to be structured relative to the rest of the protein. These contacts may need to be in place for SH3 to continue folding. After this short region is crossed, the rest of the curve is quite uniform and low. Overall, the probability that any contact is made is close to Q. This protein never has its configuration space strongly reduced as it searches for its native state. The overall shape of the $R(Q)$ curve is similar to AcP, although the entropic bottleneck is smaller for SH3.

PtG is a relatively fast folding protein and has residual structure in the denatured state.[50,51] Our results corroborate this idea, showing an unfolded free energy minimum at $Q \approx 0.25$; almost a quarter of PtG's structure remains structured in the denatured state (data not shown). Figure 2 shows that $R(Q)$ is highest at lower Q ($0.2 \leq Q \leq 0.4$), meaning specific contacts have a high frequency of occurrence. PtG is more routed early in the folding process, and this limits the search through configuration space for the remaining unstructured portion. The structured portion of PtG found at this lower Q region is correctly formed and guides the protein into the native fold. The rest of the folding process is not routed, approaching zero as the protein reaches the native state.

Psbd is a very fast folding protein and shows almost constant route measure throughout folding (see Figure 2). The folding process is moderately routed from the early stages, with a very slow decreasing of the route measure from unfolded to folded states.

The route measure curves of these four proteins reveal the width of the accessible energy landscape. Changes in the width of the landscape may either help or hinder the folding rate. If the landscape narrows at the early stages of the folding process, due to residual structure in the denatured state ensemble, this can steer the protein through the landscape more quickly. If a bottleneck suddenly arises on a later stage of the landscape, specific contacts need to come together before the process can progress, and the process is slowed. It has been shown in lattice simulations and analytical theory[18] that routing a protein by making already favored contacts more likely to occur increases the folding speed. Recently, this has been proposed as the possible mechanism speeding the folding of circular permutants of S6.[52] Routing this protein by making a circular permutant where contacting residues that were energetically favored but far apart in native sequence are now both energetically and entropically favored (the incision and reconnection places these residues close together) increases the folding rate. It has also been shown that folding rates for a large set of two-state proteins correlate very well with the variance of Φ-values (both simulated and experimentally determined), confirming that the degree of structural polarization at the transition state strongly influences the folding kinetics (S. S. Plotkin, personal communication).

B.3. Effective Loop Length and Partial Contact Order. The loop entropy of contact formation is an important component of protein topology, as shown by the success of the contact order as a predictor of folding rates. The contact order is the average sequence separation between residues computed over all residue pairs that form native contacts. The effective size of the loops formed by contacting pairs vary throughout the folding process, however, because more local, inner loops may already be formed, decreasing the loss of loop entropy for larger loops (see Figure 3). The ordering and heterogeneity of loop formation is an important component of topology and is captured by the effective loop length, $L_{\text{eff}}(Q)$, as defined by the following equation:

(49) Aronsson, G.; Brorsson, A. C.; Sahlman, L.; Jonsson, B. H. *FEBS Lett.* **1997**, *411*, 359−364.
(50) Kuszewski, J.; Clore, G. M.; Gronenborn, A. M. *Protein Sci.* **1994**, *3*, 1945−1952.
(51) Park, S. H.; Oneil, K. T.; Roder, H. *Biochemistry* **1997**, *36*, 14277−14283.
(52) Lindberg, M. O.; Tangrot, J.; Oliveberg, M. *Nat. Struct. Biol.* **2002**, *9*, 818−822.

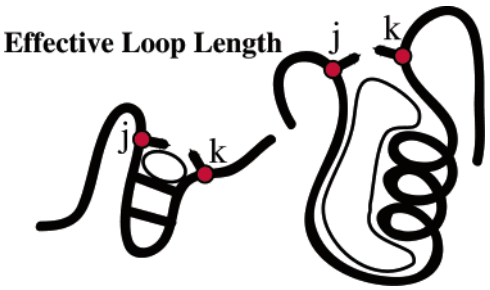

Figure 3. Traditional loop length considered in Plaxco's contact order (the thick line) is larger than the effective loop (thin line). The effective loop size is reduced by the formation of inner contacts, allowing a representation of "inner-contact cooperativity".

$$L_{\mathrm{eff}}(Q) = \frac{\sum_{i=1}^{M}(L_i - \sum_{j\in i} L_j \langle Q_j \rangle_Q)}{\sum_{k=1}^{M} L_k} \qquad (2)$$

where L_i is the loop defined as the chain segment between the pair of residues identifying contact i. The sum $\sum_{j\in i}$ is computed over all nonintersecting inner loops j inside the loop L_i (see Figure 3); the fact that each inner loop L_j is formed with probability $\langle Q_j \rangle$ is also taken into account. $L_{\mathrm{eff}}(Q)$ is normalized to unity in the unfolded state: the denominator coincides with the "standard" absolute contact order (multiplied by total number of contacts M) that is also the value taken by the numerator at $Q = 0$, when no contacts are formed. $L_{\mathrm{eff}}(Q)$ yields a smaller value as folding occurs ($Q > 0$) because all formed inner loops are subtracted.

The partial contact order, pCO(Q), considers the reduction of the loop entropy in $L_{\mathrm{eff}}(Q)$ and additionally includes the probability that a given contact is formed:

$$\mathrm{pCO}(Q) = \frac{1}{M}\sum_{i=1}^{M} \frac{L_{\mathrm{eff}}(Q)\langle Q_i(Q)\rangle_Q}{L_{\mathrm{eff}}(Q{=}1)} \qquad (3)$$

This function describes the evolution of an average effective contact order along the folding landscape. The synoptic analysis of partial contact order and effective loop length illustrates the heterogeneity and order in loop formation (Figure 4). Let's consider our four selected representative proteins.

PtG and Psbd: $L_{\mathrm{eff}}(Q)$ smoothly decreases, and pCO(Q) smoothly increases. For both PtG and Psbd proteins, $L_{\mathrm{eff}}(Q)$ decreases with a smaller slope than the slower folding proteins SH3 and AcP. This indicates that the inner loops are being formed early in the folding of PtG and Psbd, smoothing out the energy landscape for subsequent contacts. As the folding proceeds from the unfolded to the transition state ($Q \simeq 0.5$), pCO(Q) increases much more slowly for PtG and Psbd than for SH3 and AcP. Overall, the loss of configuration entropy is homogeneous; all contacts become in essence local, because the effective loop length is sufficiently reduced by the formation of inner contacts. This homogeneity in entropy loss may be thought of as an inner loop cooperativity. A rapid increase of pCO(Q) distinguishes the post-transition-state region of PtG, while pCO(Q) remains extremely smooth over the entire folding process.

SH3 and AcP: the folding process involves a more heterogeneous (or more "frustrated") formation of contacts. The steeper slope of $L_{\mathrm{eff}}(Q)$ in SH3 and AcP compared to PtG and Psbd indicates that longer loops form earlier in the folding process. The benefit of forming inner loops first is not observed in these proteins. Moreover, AcP (the slowest folding proteins among the selected four) has a "dip" in pCO(Q), revealing that some contacts that were formed early on are less likely to be made in the region of negative slope, thus confirming that a large entropic bottleneck slows down the folding process of this protein. A more detailed analysis of AcP shows that several nonlocal contacts have a 10% higher probability of being formed at $Q \simeq 0.3$ than at $Q \simeq 0.4$ (data not shown). These contacts may disrupt the folding procession if they are made too early and be required to unform. Experimental and simulated data show AcP to have three residues (Y11, P54, and F94) with relatively high Φ-values: these residues are indeed involved in long-range contacts. Our measurements support the importance of long-range contacts in the transition-state ensemble of AcP as has been shown in experiment.[53]

Conclusions

During the last five years several experimental, computational, and theoretical results have shown that *topological frustration* is an important determinant of the folding mechanism in two-state proteins.[6,8,11−14,16,17,36,40,43,54−57]

In this article we have quantified this assertion by introducing topological probes extracting information on the intrinsic (entropic) roughness of the free energy folding landscape. We have studied the behavior of these probes on a model database of two-state proteins. The picture emerging from this analysis is fully consistent with the folding rates and mechanisms experimentally obtained for these proteins. We have compared folding rates from simulations and experiments in an absolute sense, by establishing a solid criterion to report the data in the same physical units (see Supporting Information section C). When the data are in the same units and the same conditions, the agreement between simulated and experimentally measured folding rates is quite remarkable.

The route measure, $R(Q)$, of the fastest of these proteins typically shows a landscape that is more funneled early on and unordered beyond the transition state. This indicates a very smooth folding landscape, in which local contacts may easily form first because they have a smaller entropy cost. Longer range contacts may have longer-range or stronger energetic attraction to help funnel the energy landscape of the protein. These longer range contacts may additionally be aided by the inner loop cooperativity measured through the effective loop length, making the overall loss of loop entropy more homogeneous.

Proteins in which the loss of loop entropy is shown to be more heterogeneous (as demonstrated by a steeper negative slope of the effective loop length) are generally slower folding. To pack the protein core correctly may require a specific order of

(53) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. *Nature* **2001**, *409*, 641−645.
(54) Micheletti, C.; Banavar, J.; Maritan, A.; Seno, F. *Phys. Rev. Lett.* **1999**, *82*, 3372−3375.
(55) Alm, E.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11305−11310.
(56) Munoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311−11316.
(57) Klimov, D. K.; Thirumalai, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 7254−7259.

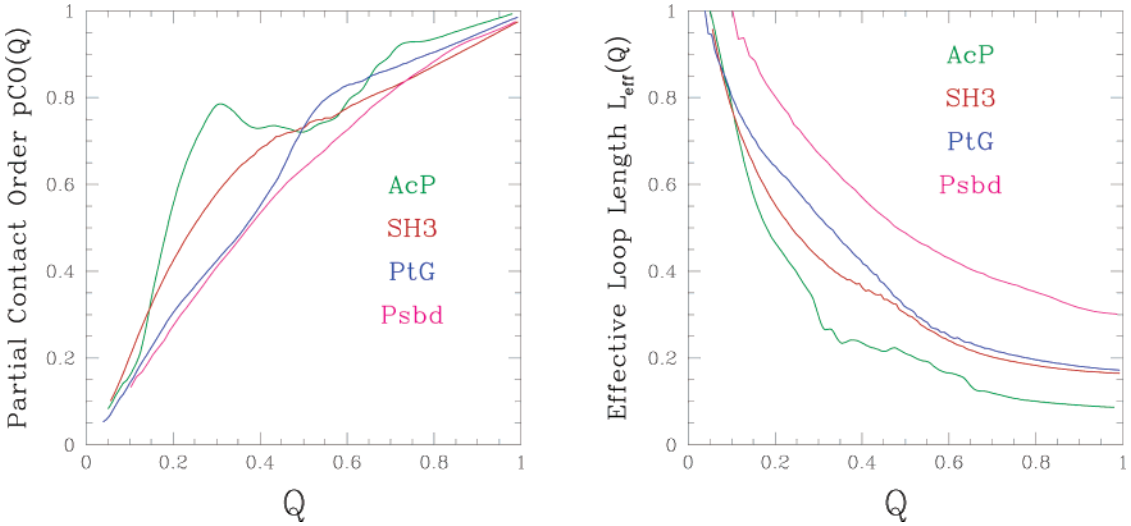

Figure 4. Effective loop length (right) shows the heterogeneity of loop formation during the folding process. The steeper negative slopes of the slower folding AcP (green) and SH3 (red) show that more long-range contacts are being formed early. The slope of the partial contact order (left) portrays the loop entropy cooperativity of the folding process. PtG (blue) and Psbd (purple) have smoother, more gently sloped curves prior to the transition state than AcP and SH3, indicating that longer-range contacts are made only after more local contacts in their interior have formed in faster-folding proteins. Additionally, AcP's curve dips at $Q = 0.5$, showing that some contacts form around $Q = 0.3$ and then unfold, slowing the folding process.

contact formation, perhaps involving more nonlocal contacts; see, for example, ref 58 In such a case, the protein could search through many configurations in which local contacts come together only to drift apart, until finally the correct contacts are made. When the ensemble of contacts leading to the native state appears to have a more heterogeneous loss of loop entropy, the details of the amino acid sequence may have less influence in defining the transition-state ensemble than in proteins where local contacts come together first, such as in PtG and PtL.[35] Overall, our results suggest that while entropic heterogeneity may be the definitive sculptor of the free energy landscape for slow-folding proteins, it is not the case for fast-folding proteins, where important features may be added to the free energy landscape by energetic (native and non-native) heterogeneity.

Acknowledgment. C.C. acknowledges funds from the Welch foundation (Norman Hackerman Welch Young Investigator award), NSF (CAREER Award CHE-0349303), and generous start-up funds provided by Rice University. L.L.C. is supported by MARC Predoctoral Fellowship of the NIH-NIGMS. Most of this work was performed during a visit by L.L.C. to the Chemistry Department of Rice University, that is hereby gratefully acknowledged for the kind hospitality. The work in San Diego was funded by the NSF-sponsored Center for Theoretical Biological Physics (Grants PHY-0216576 and 0225630) and additional support from NSF (Grant MCB-0084797). We acknowledge the W. M. Keck Foundation for computer support (through the W. M. Keck Laboratory for Integrated Biology at UCSD). C.C. is indebted to Giovanni Fossati for valuable suggestions, continuous encouragement, and technical assistance on computer related issues, and to Jim Kinsey for critical reading of the manuscript. Insightful and enjoyable discussions with Steve Plotkin and Peter Wolynes have been precious to the development of this work. Members of Clementi's group are warmly acknowledged for stimulating discussions.

Supporting Information Available: Details for simulation procedure and choice of the reaction coordinate, discussion on the criterion for selecting experimental folding rates from multiple data at different conditions, and details on the theoretical determination of folding rates from simulations for absolute comparison of theoretical and experimental data. This material is available free of charge via the Internet at http://pubs.acs.org.

JA049510+

(58) Garcia, A. E.; Onuchic, N. J. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 13898−13903.
(59) Scalley, M. L.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *94*, 10636−10640.
(60) Kuhlman, B.; Luisi, D. L.; Evans, P. A.; Raleigh, D. P. *J. Mol. Biol.* **1998**, *284*, 1661−1670.
(61) Cobos, E. S.; Filimonov, V. V.; Vega, M. C.; Mateo, P. L.; Serrano, L.; Martinez, J. C. *J. Mol. Biol.* **1998**, *328*, 221−233.
(62) Clarke, J.; Cota, E.; Fowler, S. B.; Hamill, S. J. *Structure with Folding and Design* **1999**, *7*, 1145−1153.
(63) Wang, M.; Tang, Y.; Sato, S.; Vugmeyster, L.; McKnight, C. J.; Raleigh, D. P. *J. Am. Chem. Soc.* **2003**, *125*, 6032−6033.
(64) Yang, W. Y.; Gruebele, M. *Nature* **2003**, *423*, 193−197.
(65) Spector, S.; Raleigh, D. P. *J. Mol. Biol.* **1999**, *293*, 763−768.
(66) Perl, D.; Jacob, M.; Stupak, M.; Antalik, M.; Schmid, F. X. *Biophys. Chem.* **2002**, *96*, 173−190.
(67) Schindler, T.; Schmid, F. *Biochemistry* **1996**, *35*, 16833−16842.
(68) Jacob, M.; Geeves, M.; Holtermann, G.; Schmid, F. X. *Nat. Struct. Biol.* **1999**, *6*, 923−926.
(69) Alexander, P.; Orban, J.; Bryan, P. *Biochemistry* **1992**, *31*, 7243−7248.
(70) Jackson, S. E.; Fersht, A. R. *Biochemistry* **1991**, *30*, 10436−10443.
(71) Oliveberg, M.; Tan, Y.-J.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 8926−8929.
(72) Tan, Y.; Oliveberg, M.; Fersht, A. *J. Mol. Biol.* **1996**, *264*, 377−389.
(73) Friel, C. T.; Capaldi, A. P.; Radford, S. E. *J. Mol. Biol.* **2003**, *326*, 293−305.
(74) Riddle, D. S.; Grantcharova, V. P.; Santiago, J. V.; Alm, E.; Ruczinski, I.; Baker, D. *Nat. Struct. Biol.* **1999**, *6*, 1016−1024.
(75) Martinez, J. C.; Serrano, L. *Nat. Struct. Biol.* **1999**, *6*, 1010−1016.
(76) Fowler, S. B.; Clarke, J. *Structure with Folding and Design* **2001**, *9*, 355−366.
(77) Van Nuland, N. A. J.; Meijberg, W.; Forge, V.; Scheek, R. M.; Robillard, G. T. M.; Dobson, C. M. *Biochemistry* **1998**, *37*, 622−637.
(78) Chiti, F.; Taddei, N.; White, P. M.; Bucciantini, M.; Magherini, F.; Stefani, M. M.; Dobson, C. M. *Nat. Struct. Biol.* **1999**, *6*, 1005−1009.
(79) Kitahara, R.; Royer, C.; Yamada, H.; Boyer, M.; Soldana, J.-L.; Akasaka, K.; Roumestand, C. *J. Mol. Biol.* **2002**, *320*, 609−628.