
A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling

STEFAN WALLIN AND HUE SUN CHAN

Department of Biochemistry and Department of Medical Genetics & Microbiology, University of Toronto, Toronto, Ontario M5S 1A8, Canada

(RECEIVED December 16, 2004; FINAL REVISION February 23, 2005; ACCEPTED February 23, 2005)

Abstract

Recently, a series of closely related theoretical constructs termed the “topomer search model” (TSM) has been proposed for the folding mechanism of small, single-domain proteins. A basic assumption of the proposed scenarios is that the rate-limiting step in folding is an essentially unbiased, diffusive search for a conformational state called the native topomer defined by an overall native-like topological pattern. Successes in correlating TSM-predicted folding rates with that of real proteins have been interpreted as experimental support for the model. To better delineate the physics entailed, key TSM concepts are examined here using extensive Langevin dynamics simulations of continuum C_{α} chain models. The theoretical native topomers of four experimentally well-studied two-state proteins are characterized. Consistent with the TSM perspective, we found that the sizes of the native topomers increase with experimental folding rate. However, a careful determination of the corresponding probabilities that the native topomers are populated during a random search fails to reproduce the previously predicted folding rates. Instead, our results indicate that an unbiased TSM search for the native topomer amounts to a Levinthal-like process that would take an impossibly long average time to complete. Furthermore, intraprotein contacts in all four native topomers considered exhibit no apparent correlation with the experimental ϕ -values determined from the folding kinetics of these proteins. Thus, the present findings suggest that certain basic, generic yet essential energetic features in protein folding are not accounted for by TSM scenarios to date.

Keywords: protein folding; topomer search model; topomer-sampling model; Levinthal search; explicit-chain modeling

Many single-domain proteins fold via apparent two-state processes with rates that vary widely, from μsec^{-1} to sec^{-1} (Jackson 1998; Baker 2000). In 1998, an insightful discovery was made: These folding rates were found to be predictable to approximately within one to two orders of magnitude using a very simple topology-based parameter

called the relative contact order, CO (Plaxco et al. 1998). This parameter quantifies the average sequence separation between native contacts divided by protein length. Proteins with low CO values (typically α -helical proteins) were found to fold faster than proteins with high CO values (typically α/β - and β -proteins). Since this seminal finding, other topology-based parameters have been found that exhibit similar correlation with folding rates. These include long-range order (Selvaraj and Gromiha 2001), total contact distance (Zhou and Zhou 2002), “cliquishness” (Micheletti 2003), local secondary structure content (Gong et al. 2003), and the topomer-derived parameter considered in this work (Makarov and Plaxco 2003).

Reprint requests to: Hue Sun Chan, Department of Biochemistry, University of Toronto, 1 King's College Circle, Toronto, Ontario M5S 1A8, Canada; e-mail: chan@arrhenius.med.utoronto.ca; fax: 1-416-978-8548.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.041317705>.

Several theoretical constructs that did not directly consider explicit chain representations have had notable successes in reproducing this empirically observed rate–topology relationship (Alm and Baker 1999; Galzitskaya and Finkelstein 1999; Muñoz and Eaton 1999; Weikl and Dill 2003). For more realistic explicit-chain models that involve direct simulations of folding kinetics, however, it has proven less straightforward (Koga and Takada 2001; Faisca and Ball 2002; Cieplak and Hoang 2003); nonetheless, significant recent advances have been made (Jewett et al. 2003; Kaya and Chan 2003c; Chavez et al. 2004; Ejtehadi et al. 2004).

Among the theoretical efforts aiming to elucidate the remarkable contact order-dependent folding rates, a series of closely related models under the name of “topomer search model” or “topomer-sampling model” (TSM) has been proposed in the past five years as a possible mechanistic basis for the empirically observed rate–topology relationship (Debe and Goddard 1999; Debe et al. 1999; Makarov and Metiu 2002; Makarov et al. 2002; Makarov and Plaxco 2003; Gillespie and Plaxco 2004). These topomer constructs share the idea that the folding of small single-domain proteins is a two-step process: The first step is an essentially unbiased, diffusive search for an overall native-like topological state termed the “native topomer state.” In the second step, after the chain has formed the correct gross topology, the native contacts rapidly zipper to form the completed native structure. Since this second step is assumed to occur very rapidly, the rate-limiting step in folding is the essentially unbiased search process in the first step, and folding rates are therefore determined by the “size,” or extent in conformational space, of the native topomer state.

The topomer folding picture may be schematically illustrated by the highly simplified energy landscape caricatures in Figure 1. Here the energy landscape “outside” of the native topomer (the part represented by a steep funnel in Fig. 1) has been drawn flat to highlight the unbiased nature of the hypothetical TSM process (Debe and Goddard 1999). This corresponds to the case in

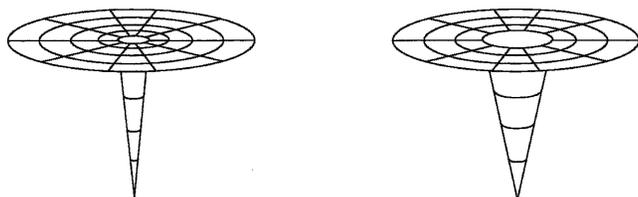


Figure 1. Simple schematic illustration of the energy landscape in the TSM, for a slow-folding (*left*) and a fast-folding (*right*) protein. The horizontal dimensions represent protein conformational variation or conformational entropy; the vertical dimension provides the free energy of every given protein conformation—with appropriate averaging of solvent degrees of freedom (Dill and Chan 1997).

Makarov et al. (2002) for which “the barrier to folding is purely entropic,” although the possibility of a general energy favoring contact formation has also been considered as part of a TSM process (Makarov et al. 2002, p. 3538; see Discussion below). Thus, in a proposed unbiased topomer-search scenario, the first step of the folding process is a search on this flat region of the energy landscape. Then, as soon as the native topomer is found, folding proceeds quickly and downhill to the native state until all native contacts are formed. As a consequence of this folding picture, the folding rate of a protein is controlled by a quantity P_{top} , the probability that the native topomer state is populated during a random search. Slow- and fast-folding proteins thus have “tight” and “loose” native topomers, respectively, corresponding to the two landscapes in Figure 1 with smaller and larger “golf holes.”

To our knowledge, the original version of TSM was first proposed by Debe et al. (1999). Using a discrete chain growth method to generate large numbers of random conformations, the total number of distinct topomer states (of which the native topomer is one) was estimated for proteins with chain lengths up to $N = 100$ amino acid residues. For instance, they found the number of topomers to be $\sim 3 \times 10^7$ for $N = 100$. It was then argued that if the topomer state space was sampled on a nanosecond timescale, native topomers could be found on typical folding times, even in the absence of a mechanism that simplifies and expedites the search. In a subsequent paper, Debe and Goddard (1999) set out to test the model further by estimating the quantities P_{top} for a set of 18 β - and α/β -proteins using a similar chain growth method (all- α proteins were excluded from this data set). The folding rates that resulted from their calculations showed a good correlation with experimental data (correlation coefficient 0.78).

In all TSM scenarios, the proposed folding process is quickly completed once the native topomer is located. It follows that if TSM is to provide a rationalization of apparent two-state folding mechanisms, the precise definition of the native topomer state is of critical importance. In the original TSM study by Debe and colleagues, “topomers are tubes of topologically equivalent conformations” (Debe and Goddard 1999): Two structures (conformations) were classified as topomeric (i.e., belonging to the same topomer state) based on a test procedure in which equal-strength harmonic forces were applied between corresponding C_{α} atoms of two optimally superimposed structures. If a conjugate gradient minimization procedure could completely relax the spring forces, that is, if small backbone moves of one structure could bring it onto the other without getting trapped in a local minimum, the two structures were classified as belonging to the same topomer (Debe et al. 1999).

In the subsequent development of the TSM approach by Makarov and colleagues, however, the native topomer

state was defined by specific collections of native contacts, as the set of conformations in which all sequence-distant native contacts are formed. In other words, TSM is the conformational ensemble in which all sequence-distant residues (as defined by the formalism) that are in contact in the native state are in close spatial proximity. Hence, the number of sequence-distant native contacts, Λ_D , plays a central role in this more recent TSM formulation, as will be detailed below. This approach is different in many respects from the original TSM of Debe and Goddard (1999) and Debe et al. (1999). Certain features introduced in the later formulation of Makarov and colleagues are seen as instrumental in extending the TSM's ability to predict folding rates to include also purely α -helical proteins, as opposed to only β - and α/β -proteins (Makarov and Metiu 2002; Makarov et al. 2002; Makarov and Plaxco 2003).

The kinetic process of topomer sampling was envisioned by Debe and colleagues to be mainly among relatively compact conformations with low (favorable) solvation energies, as was depicted by the kinetic path on their "Rose Bowl" (a stadium in Pasadena, California) landscape (see Debe et al. 1999, Fig. 5). This amounts to postulating a folding mechanism that involves a fast, "burst-phase" collapse to an ensemble of compact conformations that are partitioned into different topomers. In that case, the flat areas in Figure 1 represent only such compact conformations but not the open conformations that presumably have higher solvation energies. The bulk of the search time for the native topomer is then spent in different compact conformations. But such a folding mechanism does not appear to correspond to that of many real, apparent two-state proteins—typified by chymotrypsin inhibitor 2 (Jackson and Fersht 1991)—that exhibit no significantly populated compact folding intermediates. In contrast, the topomer search approach of Makarov and colleagues seeks the probability of locating the native topomer among all possible conformations, most of which are presumably open (Makarov et al. 2002), and is therefore conceptually more in line with the principles of cooperative protein folding (Chan et al. 2004; Gillespie and Plaxco 2004).

For this reason, our present analysis focuses primarily on the more recent TSM formulation of Makarov and colleagues with the above-described contact-based definition of the native topomer (Makarov and Metiu 2002; Makarov et al. 2002; Makarov and Plaxco 2003). In their approach, TSM prediction of folding rates is reduced to solving the following problem: What is the probability $P(n_D)$ that n_D sequence-distant residue pairs are brought into proximity by a random search process? In this notation, the native topomer probability P_{top} is identified with $P(\Lambda_D)$. The parameter Λ_D (which is denoted Q_D in Makarov and Plaxco 2003) is closely

related to the long-range order parameter LRO proposed previously by Selvaraj and Gromiha (2001), namely, $\Lambda_D/N = LRO$ provided that the same definition of a contact is used. Based on results from inert Gaussian chains (without consideration of polymer excluded volume), Makarov and Plaxco (2003) introduced the approximate formula $P(\Lambda_D) = \gamma a^{\Lambda_D}$ where a and γ are constants. This expression leads to a simple topology-based rate-prediction formula that was tested on a set of 24 two-state proteins (Makarov and Plaxco 2003), a diverse set that did not exclude all- α proteins as had been done previously (Debe and Goddard 1999). The rate-prediction formula showed excellent correlation with experimental folding rate data with a correlation coefficient ~ 0.9 (Makarov and Plaxco 2003), even though folding rates predicted using Λ_D can sometimes be drastically different from those predicted using CO (Jones and Wittung-Stafshede 2003).

Because of this empirical success and the potential physical insights it may offer, it is imperative to take an in-depth look at the TSM perspective as several key aspects of it remain to be better elucidated. As it stands, TSM is not a self-contained, explicit-chain model. TSM investigations thus far have not used direct simulations of folding kinetics. Instead, kinetic behaviors and folding rates were deduced from presumed thermodynamics–kinetics relationships. In this basic respect, TSM is akin to other non-explicit-chain modeling of CO -dependent folding that have also enjoyed high degrees of success (Alm and Baker 1999; Muñoz and Eaton 1999). However, the relationship between the thermodynamics and kinetics of chain molecules can be subtle, and explicit-chain results do not always agree with expectations from non-explicit-chain considerations (Karanicolas and Brooks 2003; Chan et al. 2004). Therefore, ultimately, TSM or any other non-explicit-chain model of protein folding has to be evaluated by ascertaining whether and how the model assumptions can emerge from polymer physics. With this in mind, this article explores the feasibility of the proposed TSM folding mechanism by performing extensive Langevin molecular dynamics simulations using an explicit C_α chain model. We focus on four apparent two-state proteins with diverse CO values. These chain models allow us to assess the logic of the physical picture afforded by TSM, and to perform structural comparisons with experiments.

In the analysis below, we first derive a clear structural characterization of the native topomers for the four proteins. Then, by using standard histogram reweighting techniques, we estimate their $P(\Lambda_D)$ values. We find them to be much smaller than that stipulated by the TSM approximation $P(\Lambda_D) = \gamma a^{\Lambda_D}$. In fact, our calculations show that finding the correct native topomer state

in a random, unbiased TSM search is so unlikely that it is comparable to the hypothetical Levinthal search process. Finally, to connect TSM with common understanding of folding kinetics, we compare experimental ϕ -values with TSM-predicted ϕ -values calculated by taking the ensemble of conformations constituting a native topomer state to be the TSM-prescribed folding transition state.

Results

Proteins

We focus on the four single-domain proteins in Table 1 that cover a variety of chain lengths, secondary structure contents, and folding rates. They are part of a larger set of roughly 30 single-domain proteins that have been shown to fold via simple, apparent two-state thermodynamics and kinetics. A list of these proteins along with experimental folding data can be found in Ivankov et al. (2003).

Preliminaries: Definitions and assumptions

Following Makarov and Plaxco (2003), we use a spatial proximity cutoff parameter r_c and a sequence separation cutoff parameter l_c to determine the set of sequence-distant native contacts: An amino acid pair ij is considered part of the native topomer contact set if the C_α 's of amino acids i and j are separated by a spatial distance $r_{ij}^n < r_c$ and their sequence separation $|i - j| > l_c$. Thus, the number of contacts Λ_D in this set depends on the choice of l_c and r_c . In Table 2, we report the Λ_D values for our four proteins, for several different choices of l_c and r_c within the limits $4 \leq l_c \leq 12$, and $6 \text{ \AA} \leq r_c \leq 8 \text{ \AA}$. Table 2 shows that Λ_D depends rather strongly on l_c and r_c , and this dependence is particularly strong for 1lmb. For this protein, Λ_D varies almost by a factor 10 between the different values of l_c and r_c considered here.

We then define the native topomer state of a protein as the set of conformations in which every Λ_D amino acid pair has formed a "contact" or is otherwise in close

Table 1. Data for the four proteins studied

Protein	pdb	N	R_g^n (Å)	k_f (sec ⁻¹)	T_{exp} (°C)
Acylphosphatase	laps	98	12.6	2.3×10^{-1}	28
Chymotrypsin inhibitor 2	2ci2	64	10.8	4.8×10^1	25
Spliceosomal protein U1A	lurn	96	12.2	3.2×10^2	25
λ -repressor	1lmb	80	11.5	1.1×10^4	25

Protein databank id, pdb; number of amino acids, N ; radius of gyration, R_g^n ; and experimental folding rate in water, k_f , determined at temperature T_{exp} (Jackson and Fersht 1991; Silow and Oliveberg 1997; van Nuland et al. 1998; Myers and Oas 1999).

Table 2. Number of sequence-distant native contacts Λ_D

	$r_c = 8 \text{ \AA}$			$r_c = 7 \text{ \AA}$			$r_c = 6 \text{ \AA}$		
	$l_c = 4$	8	12	$l_c = 4$	8	12	$l_c = 4$	8	12
laps	236	216	205	167	155	147	92	85	79
2ci2	108	93	88	80	69	67	47	41	40
lurn	180	156	140	123	108	98	70	59	53
1lmb	85	58	49	45	25	19	24	12	9

Here we focus on the cutoff values l_c and r_c within the limits suggested by Makarov and Plaxco (2003).

spatial proximity. To simplify the terminology, in the following we will use "contact" to indicate simply that two C_α atoms are close in space. For this purpose, two definitions of contact are used. Unless stated otherwise, we use the contact criterion $r_{ij} < 1.2r_{ij}^n$, where r_{ij} is the C_α - C_α distance between i and j . This criterion has been used in previous studies of native-centric C_α models (Clementi et al. 2000; Kaya and Chan 2003a). In certain considerations, we find it useful to use a slightly more permissive contact criterion, $r_{ij} < r_{ij}^n + 3.0 \text{ \AA}$, which entails a broader definition of a native topomer. It should be noted that both of these contact criteria are independent of the cutoff parameter r_c .

Chain simulation of the native topomer search process

To assess the TSM folding picture by an explicit-chain approach, it is necessary to (1) construct an overall conformational ensemble that is essentially unbiased, (2) determine the size of the conformational space covered by the native topomer as a sub-ensemble of the overall conformational ensemble, and (3) estimate by explicit construction the probability that this conformational sub-ensemble can be located by a diffusive search, as envisioned by TSM. We adopt a simple C_α chain model for this purpose, with the following potential energy function:

$$E_0 = k_{\text{rep}} \sum_{i < j - 3} \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} + k_{\text{bon}} \sum_{\text{bonds}} (b_i - b_i^n)^2 + k_{\text{ben}} \sum_{\text{angles}} (\theta_i - \theta_i^n)^2 + k_{\text{tor}} \sum_{\text{torsion}} [1 - \cos 3(\phi_i - \phi_i^n)], \quad (1)$$

where b_i , θ_i , ϕ_i , and r_{ij} are the virtual bond lengths, bond angles, torsion angles, and C_α - C_α distances, respectively, and b_i^n , θ_i^n , ϕ_i^n , and r_{ij}^n are the corresponding native values. This energy function is designed to capture the basic polymer statistics of a generic self-avoiding polypeptide chain. The use of the θ_i^n and ϕ_i^n values in the last

two terms introduces a small bias toward the native state. These terms are included here for computational efficiency. Their presence does not alter our main conclusion, based on the results from Equation 1, that the probability of locating the native topomer by an unbiased search is vanishingly small (see below). This is because the native-centric nature of these terms in Equation 1 implies that the relative size of the native topomer sub-ensemble would be even smaller in an entirely unbiased overall conformational ensemble. The first term of the above potential function accounts for polymer excluded volume. Its summation is over all amino acid pairs ij , and the parameters σ_{ij} are chosen in the following manner: $\sigma_{ij} = 0.85r_{ij}^n$ for pairs ij in contact in the native state ($r_{ij}^n < r_c$); otherwise, $\sigma_{ij} = 4.0 \text{ \AA}$. To speed up the calculations, the repulsive energy term is evaluated using a cutoff procedure with cutoff radius $2\sigma_{ij}$. This choice of the σ_{ij} parameters ensures that the native conformation would not be made inaccessible by unphysical repulsive forces.

A motivation for adopting a simplified C_α model in this study rather than using chain representations that account for more structural details is computational efficiency: A simplified chain model allows for extensive conformational sampling and reliable probability estimations that are essential in addressing issues pertinent to the TSM. In this regard, it is noteworthy that the TSM itself—like several other topology-based folding scenarios—is also an inherently structurally low-resolution formulation that does not refer directly to sequence-specific interactions (Makarov and Plaxco 2003). Hence, a C_α chain model enjoys the advantage of making direct connections to the underlying mathematical approximations of the TSM. Continuum (off-lattice) C_α models have been used extensively in the literature for studying protein folding, mostly in combination with $G\ddot{o}$ -type energy functions (Clementi et al. 2000) but also for structure prediction (Nanias et al. 2003). In particular, our model is similar to the one introduced in Clementi et al. (2000) and further studied in Kaya and Chan (2003a). Following Kaya and Chan (2003a), we choose the strengths of energy terms in Equation 1 to be $k_{\text{rep}} = \epsilon$, $k_{\text{bon}} = 100 \epsilon$, $k_{\text{ben}} = 20 \epsilon$, and $k_{\text{tor}} = 0.5 \epsilon$, respectively, where the energy unit ϵ is set to 1, and use Langevin dynamics to calculate the thermodynamic behavior. The temperature is kept constant at $T = 1.0$, and snapshots are taken every 100 time steps. Simulation details are otherwise the same as in Kaya and Chan (2003a).

The native topomers

The energy function in Equation 1 describes a disordered and “floppy” protein chain that samples many different conformations. To obtain a subset of these conformations that corresponds to the native topomer

state of each of the four proteins studied here, we impose harmonic constraints representing the sequence-distant topomer contact pairs on the C_α chain. This is achieved by using the energy function

$$E_{\text{top}} = E_0 + k_{\text{top}} \sum_{\text{constr.}} \left(r_{ij} - r_{ij}^n \right)^2, \quad (2)$$

where the summation shown is over the Λ_D sequence-distant native contacts as defined by the cutoff parameters l_c and r_c (see Table 1). The conformational sampling procedure here is similar to the determination of ensembles of partially disordered protein conformational states that uses experimental data as constraints (Choy and Forman-Kay 2001; Vendruscolo et al. 2001), the main difference being that the constraints in this study are supplied by the TSM hypothesis instead of experimental measurements. In Equation 2, the topomer constraint strength is taken to be $k_{\text{top}} = 10 \epsilon$, making these harmonic forces one order of magnitude weaker than the strength k_{bon} of the virtual bonds between sequentially adjacent C_α atoms. We perform Langevin dynamics simulations of the energy function E_{top} for each of our four proteins and different choices of the parameters l_c and r_c . Each simulation is started from the native structure. The system is first equilibrated for 10^8 simulation time steps, after which sampling is performed for 10^9 time steps.

In general, a smaller l_c and a larger r_c will produce a larger number of topomer constraints, resulting in a topomer ensemble that more closely resembles the native pdb structure. This trend can be seen from Table 3, which shows thermodynamic averages of the radius of gyration R_g , the root-mean-square deviation (rmsd) from the corresponding native structure, and the fraction of native contacts Q for four different parameter choices of l_c and r_c . Smaller l_c and larger r_c naturally lead to smaller $\langle \text{rmsd} \rangle$ and larger $\langle Q \rangle$ values, consistent with more native-like ensembles. We note also that $\langle R_g \rangle$ remains fairly close to the corresponding native values (see Table 1) across the various choices of l_c and r_c .

The conformational ensembles of native topomers obtained here are a direct logical consequence of the TSM assumptions of Makarov and Plaxco (2003). An advantage of our explicit-chain approach is that it can supply relatively detailed conformational information that was not available in their TSM formulation that used an analytical approximation of conformations of chains without excluded volume. To provide a graphic structural illustration of the ensemble of topomer conformations obtained from our simulations, we display in Figure 2 a selected set of these conformations optimally superimposed on the corresponding native structure. These conformations are the first 25 centroids from a

Table 3. Native topomer thermodynamic averages

	$l_c = 4, r_c = 8 \text{ \AA}$			$l_c = 12, r_c = 8 \text{ \AA}$			$l_c = 4, r_c = 6 \text{ \AA}$			$l_c = 12, r_c = 6 \text{ \AA}$		
	$\langle Q \rangle$	$\langle \text{rmsd} \rangle$	$\langle R_g \rangle$	$\langle Q \rangle$	$\langle \text{rmsd} \rangle$	$\langle R_g \rangle$	$\langle Q \rangle$	$\langle \text{rmsd} \rangle$	$\langle R_g \rangle$	$\langle Q \rangle$	$\langle \text{rmsd} \rangle$	$\langle R_g \rangle$
laps	0.99	0.97	12.5	0.96	1.43	12.5	0.97	3.09	12.6	0.86	5.37	12.6
2ci2	0.95	2.19	10.4	0.90	2.68	10.3	0.92	4.74	10.8	0.79	5.21	11.0
lurn	0.96	2.45	11.9	0.86	3.06	12.0	0.90	6.73	12.9	0.70	7.41	13.4
lmb	0.87	3.07	10.8	0.58	4.44	11.1	0.83	9.42	12.9	0.34	11.0	14.3

Thermodynamic averages (denoted by $\langle \dots \rangle$) of fraction of native contacts (Q), root-mean-square deviation (rmsd, in \AA), and radius of gyration (R_g in \AA), are obtained from simulations of the energy function E_{top} (see equation 2) at temperature $T = 1.0$. Q is calculated using the contact definition $r_{ij} < 1.2r_{ij}^0$.

simple clustering scheme on the complete topomer ensembles (see figure caption). It is interesting to note that, qualitatively, these drawings appear to be consistent with the TSM perspective (Makarov et al. 2002; Makarov and Plaxco 2003) that a slow-folding protein such as laps has a “tighter” transition state ensemble than, for example, the much faster folding protein lmb, which has a very “loose” native topomer state (Fig. 2).

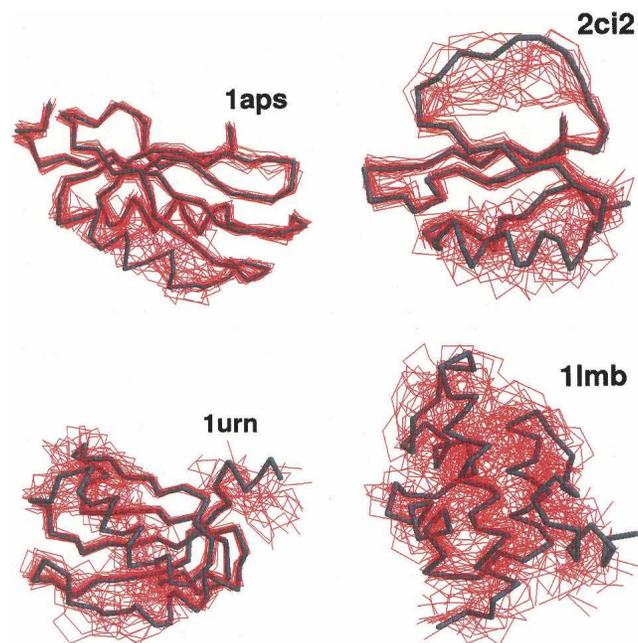


Figure 2. Illustrations of the native topomers for the proteins laps, lci2, lurn, and lmb. Shown are the top 25 conformations (red) selected by a simple clustering procedure on the full $l_c = 12, r_c = 8 \text{ \AA}$ topomer ensembles. Each of the 25 conformations is optimally superimposed on the corresponding native structure (dark trace). In the clustering procedure, the highest ranking conformation is the one with the largest number of conformational neighbors, where two conformations are considered to be neighbors if their rmsd is less than a certain cutoff (1.3–4.0 \AA). The next highest ranking conformation is the conformation with most neighbors in the reduced ensemble, where the highest-ranking conformation and its neighbors have been excluded, and so on.

A similar picture is offered by a more recent non-explicit-chain analysis of the correlation between folding rates and presumed ensemble sizes of “transition states” (Bai et al. 2004) derived from the total contact distance measure (Zhou and Zhou 2002). Here, the trend exhibited in Figure 2 is further demonstrated by Figures 3 and 4, which show the normalized probability distributions $P(\text{rmsd})$ and $P(Q)$ for $l_c = 12$ and $r_c = 8 \text{ \AA}$. The order of these distributions in terms of their closeness to the native structure follows the order of the experimental folding rates for these proteins (see Table 1), as expected from the TSM picture. The same rmsd and Q order holds for practically all the other choices of l_c and r_c within the limits $4 \leq l_c \leq 12$ and $6 \text{ \AA} \leq r_c \leq 8 \text{ \AA}$ studied here.

Native topomer ensemble sizes and folding rate predictions

We turn now to a more quantitative assessment of the TSM. In the previous section we saw that the native topomer ensembles of the four proteins display various degrees of “floppiness”; the lmb topomer ensemble is, in a sense, larger than that of laps (see Fig. 2). The TSM perspective hypothesizes that it is these differences in ensemble sizes that underlie the different folding rates observed for different proteins. More precisely, TSM stipulates that the folding rate is essentially proportional to $\Lambda_D P(\Lambda_D)$, the probability of populating the native topomer during a random, diffusive search, since TSM considers this to be the rate-limiting step in the folding of small, single-domain proteins (Makarov and Plaxco 2003).

As mentioned briefly above, Makarov and Plaxco (2003) made the approximation, based on considerations of Gaussian inert chains, that $P(\Lambda_D) = \gamma a^{\Lambda_D}$, where γ and a are constants applicable to all proteins. This means that bringing each additional sequence-distant native pair together contributes on average a constant multiplicative factor a (< 1) to the overall probability $P(\Lambda_D)$ of the native topomer state among

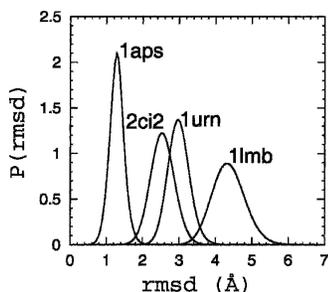


Figure 3. Probability distributions $P(\text{rmsd})$, where rmsd is from the corresponding native pdb structure, for 1aps, 2ci2, 1urn, and 1lmb, as obtained using $l_c = 12$ and $r_c = 8 \text{ \AA}$ in Equation 1, and sampling at $T = 1.0$.

the full ensemble of all accessible conformations. (The first three contact pairs are assumed to contribute with somewhat different factors, which are taken into account by the overall multiplicative factor $\gamma \neq 1$.) The topomer search rate was then postulated to be given approximately by the formula $k_f = \kappa \Lambda_D P(\Lambda_D)$, where k_f is the folding rate, and $\kappa \Lambda_D$ is the frequency of attempted contact formation. This formula was fitted to experimental folding rate data for a set of 24 two-state proteins yielding an excellent correlation coefficient of ≈ 0.88 (Makarov and Plaxco 2003). It should be noted that the dominant, exponential dependence on Λ_D in this Makarov-Plaxco expression is in $P(\Lambda_D)$. Thus, this rate correlation formula resembles closely the $\ln k_f \propto LRO$ hypothesis proposed previously by Selvaraj and Gromiha (2001) because $LRO = \Lambda_D/N$.

The success of this empirical two-parameter fit implies that the structural quantity Λ_D or LRO —like the original CO parameter (Plaxco et al. 1998)—is capturing significant aspects of the physics of apparent two-state protein folding. Obviously, the more important issue, then, is whether the physical picture offered by the TSM perspective is supported by this remarkable folding rate correlation as well. Does the empirical success of the TSM $k_f(\Lambda_D)$ formula necessarily imply that the proposed process of topomer search is correct or even physically plausible? To provide a logical answer to this question, an essential first step is to determine the mathematical validity of the $P(\Lambda_D) = \gamma a^{\Lambda_D}$ formula; more specifically, whether this formula, indeed, provides the native topomer probability it purports to describe. Obviously, unless the answer to this basic question is affirmative, the empirical success of the TSM $k_f(\Lambda_D)$ formula can only be compared with empirical rate correlation with LRO and other proposed parameters but cannot logically be translated into support for the broader topomer-search discourse.

We therefore first focus on the TSM formula $P(\Lambda_D) = \gamma a^{\Lambda_D}$. By fitting experimental data, the two

constants γ and a have been determined to be $a = 0.86$ and $\gamma = 4 \times 10^{-5}$ for $l_c = 12$ and $r_c = 6 \text{ \AA}$ (Makarov and Plaxco 2003). Using these fitted values of Makarov and Plaxco in conjunction with the Λ_D values for $l_c = 12$ and $r_c = 6 \text{ \AA}$ in Table 2, the γa^{Λ_D} values for our four proteins are found to be ranging from 10^{-5} to 10^{-10} (Table 4). Thus, according to this Makarov-Plaxco formula, populating the native topomer state during a random conformational search is about 3×10^4 times more likely for 1lmb than for 1aps, for example; and this predicted probability difference is approximately equal to the folding rate difference of these two proteins.

Now, is this predicted probability difference valid? Does the Makarov-Plaxco $P(\Lambda_D)$ formula adequately describe the relative size of the native topomer sub-ensemble relative to the full conformational ensemble? To tackle this question, we compute the probability $P(\Lambda_D)$ for our four proteins in explicit-chain conformational ensembles defined by the energy function E_0 in Equation 1. To err on the side of affording a higher probability to the native topomer, in this analysis we use the more permissive contact criterion $r_{ij} < r_{ij}^n + 3.0 \text{ \AA}$ to determine if amino acid residues (C_α positions) are in contact. The most direct way to ascertain $P(\Lambda_D)$ would be to perform Langevin dynamics simulations with E_0 and count the fraction of conformations that fulfill the native topomer criterion, that is, having the Λ_D amino acid pairs close in space. But such a brute-force approach is not computationally feasible because of the smallness of $P(\Lambda_D)$, making the native topomer very unlikely to be visited by unbiased conformational sampling. Thus, to enhance sampling of the native topomer, we introduce the auxiliary energy function

$$E_\lambda = E_0 + \lambda E_{\text{bias}}, \quad (3)$$

where E_{bias} is a potential with a bias toward the native topomer and λ is a bias strength parameter. Two-

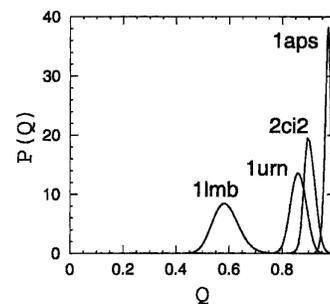


Figure 4. Probability distributions $P(Q)$ for 1aps, 2ci2, 1urn, and 1lmb, as obtained using $l_c = 12$ and $r_c = 8 \text{ \AA}$ in Equation 1, and sampling at $T = 1.0$.

Table 4. *Estimated probabilities* $P(\Lambda_D)$

	$r_c = 8 \text{ \AA}$		$r_c = 6 \text{ \AA}$		γa^{Λ_D}
	$l_c = 4$	$l_c = 12$	$l_c = 4$	$l_c = 12$	
1aps	2×10^{-86}	8×10^{-85}	8×10^{-82}	3×10^{-75}	3×10^{-10}
2ci2	2×10^{-47}	7×10^{-45}	2×10^{-38}	1×10^{-36}	1×10^{-7}
1urn	8×10^{-73}	4×10^{-68}	8×10^{-60}	1×10^{-47}	1×10^{-8}
11mb	7×10^{-60}	3×10^{-48}	2×10^{-37}	1×10^{-18}	1×10^{-5}

$P(\Lambda_D)$ is the probability that the Λ_D sequence-distant native contacts are formed during a search process described by the energy function E_0 (see Equation 1). Two amino acid residues i and j are assumed to contact each other if $r_{ij} < r_{ij}^n + 3.0 \text{ \AA}$. For comparison, we have included the corresponding values from the Makarov-Plaxco $P(\Lambda_D) = \gamma a^{\Lambda_D}$ formula with $a = 0.86$ and $\gamma = 4 \times 10^{-5}$.

dimensional histograms $P_\lambda(n_D, E_{\text{bias}})$ are recorded for a series of λ , where n_D denotes the number of sequence-distant native contacts formed. A standard histogram reweighting technique (Ferrenberg and Swendsen 1989) is then applied to obtain an improved estimate of $P_0(n_D, E_{\text{bias}})$, and hence of $P_0(n_D = \Lambda_D) = P(\Lambda_D)$. Further details of this reweighting procedure, including the precise form of E_{bias} , are given in Materials and Methods. It should be noted that by design E_{bias} is not a function of all C_α positions, but only of those that are part of the native topomer constraint set, such that the added bias is toward the somewhat disordered native topomer state (whose sampling we aim to enhance) rather than toward the fully native pdb structure.

Table 4 shows the results of our $P(\Lambda_D)$ calculations for four different choices of l_c and r_c . For the four proteins studied, our explicit-chain estimates of $P(\Lambda_D)$ lie in the range $\sim 10^{-86}$ to 10^{-18} , in contrast to the much larger values produced by the formula $P(\Lambda_D) = \gamma a^{\Lambda_D}$ using the particular a and γ parameters of Makarov and Plaxco. In fact, some of the computed explicit-chain $P(\Lambda_D)$ probabilities in Table 4 are as small as what would be expected for a Levinthal search. For instance, assuming that each Ramachandran torsion angle can occupy three discrete values, the total number of conformations for a 64-amino-acid protein (such as 2ci2) would be roughly $(3 \times 3)^{64} \approx 10^{61}$. (The essential exponential increase with chain length N of the number $\sim \mu^N$ of all possible conformations—most of which are noncompact—applies to chains with excluded volume as well. The effect of excluded volume in three dimensions leads to an appreciable but not large decrease in μ . For instance, for chains configured on simple cubic lattices, μ decreases from 6 to ≈ 4.68 ; e.g., see Barber and Ninham 1970; Chan and Dill 1991). From a polymer physics perspective, the results in Table 4 are not surprising. This is because the native topomer ensembles, especially those with relatively large Λ_D 's (e.g., 1aps),

are constituted of conformations quite closely resembling that of the native pdb structure itself (cf. Figs. 3 and 4). A probable cause of the gross overestimation of native topomer probabilities by the Makarov-Plaxco γa^{Λ_D} formula is its neglect of excluded volume. Excluded volume places enormous restrictions on the conformational freedom of compact chains (Chan and Dill 1990), but excluded volume effects are not taken into account in the Makarov-Plaxco TSM formulation. We note that the relative values of the $P(\Lambda_D)$ probabilities computed here mostly follow the same order as that provided by the $P(\Lambda_D) = \gamma a^{\Lambda_D}$ expression, suggesting that this formula might apply approximately for a different set of a and γ . For example, fitting this formula to our explicit-chain $P(\Lambda_D)$ results for $l_c = 12$ and $r_c = 8 \text{ \AA}$ yields the parameter values of $a = 0.60$ and $\gamma = 4.5 \times 10^{-33}$; but in that case the exceedingly small value of γ would not conform to the TSM picture of Makarov and Plaxco (2003), who have related a $\gamma \sim 10^{-5}$ value to “the extra entropy associated with the first few ordering events” (Makarov and Plaxco 2003, p. 21). More fundamentally, however, the very small values of the present explicit-chain $P(\Lambda_D)$ values strongly support the argument against the TSM speculation that the “entropic cost of finding the native topomer may be reasonable even in the absence of native-like interactions that may favor this set of conformations” (Makarov and Plaxco 2003, p. 22). Quite to the contrary, the vanishingly small native topomer probabilities we obtained imply that an unbiased, diffusive search for the native topomer among all accessible conformations, like the hypothetical Levinthal search, is highly unlikely to succeed and is therefore not a viable mechanism for apparent two-state folding.

ϕ -values and rate-limiting formation of the native topomer

In the TSM picture, the rate-limiting step of folding is the search for the native topomer. Thus, TSM predicts how the conformational ensemble of any given protein should look at the point when the rate-limiting step is achieved during the folding process. In principle, these structural predictions of TSM can be independently verified or falsified by experiments, irrespective of the mathematical/physical question raised above regarding whether the TSM formula provides the correct probability for the hypothetical native topomer state.

The native topomer state has been identified with the folding transition state. According to Makarov et al. (2002), the TSM model “assumes that the folding rate is controlled by the rate of forming all the native contacts observed in the folded protein” (note that these “contacts” refer to native topomer contacts that require

a contacting pair to be separated by at least a certain number of residues along the chain), and that the state created when all contacts that define the native topomer is formed is a “transition state” of the model (Makarov et al. 2002, pp. 3536, 3539). Therefore, it is instructive to compare the structural characteristics of the native topomer state with the folding transition state properties inferred by experimental techniques.

A common approach to characterize the transition state ensemble for two-state proteins is through ϕ -value analysis (Fersht et al. 1992). In these experiments, collection of single amino acid mutations is performed and their effects on the folding rate and stability are measured. The goal of ϕ -value analysis is to map out the degree of “interaction” of the mutated amino acids in the transition state. The final result for an amino acid is expressed as a ϕ -value, which states whether the amino acid is “ordered” ($\phi \approx 1$) or “disordered” ($\phi \approx 0$) in the transition state ensemble. Notwithstanding the recent controversy about the interpretation of experimental ϕ -values (Sanchez and Kiefhaber 2003; Fersht 2004; Hubner et al. 2004), ϕ -value analysis is a widely adopted approach. Thus, it is worthwhile to delineate the logical relationship between experimental ϕ -values and the corresponding quantities implied by TSM, even though ϕ -values were not computed by Makarov et al. themselves.

It is relatively straightforward to use the conformational ensembles from our simulations of the hypothetical native topomers to calculate TSM-predicted ϕ -values, which can then be compared with available experimental ϕ -value data for the four proteins studied here. Similar procedures have been applied before by other researchers to compare experimental ϕ -values and theoretical ϕ -values predicted by non-explicit-chain (Alm and Baker 1999; Galzitskaya and Finkelstein 1999; Muñoz and Eaton 1999) as well as explicit-chain (Clementi et al. 2000, 2003; Ejtehadi et al. 2004) models.

Before turning to more detailed calculations, several salient conclusions about TSM-predicted ϕ -values can be drawn by considering the general properties of the native topomer state (cf. Fig. 2). Any given amino acid residue making mostly sequence-distant native contacts in a protein's pdb structure will, because of the topomer constraints, have many of its spatially neighboring amino acid residues in the native structure also in proximity in the native topomer state. Consequently, TSM would predict a high ϕ -value for the given amino acid residue. In fact, if all of this amino acid residue's native pdb contacts were sequence-distant, ϕ would be predicted by TSM to be ≈ 1 . On the other hand, an amino acid residue making mostly local contacts, as might be the case for residues in α -helical regions, is likely to have a lower TSM-predicted ϕ -value because by definition the amino acid residues it contacted in the native pdb structure

are not required to be in spatial proximity in the native topomer state.

To calculate the TSM-predicted ϕ -values, we use the following definition of ϕ -value for an amino acid residue i :

$$\phi_{\text{calc}}^i = \frac{\langle N_i \rangle}{N_i^n} \quad (4)$$

where $\langle N_i \rangle$ is the average number of native contacts made by amino acid residue i in the native topomer conformational ensemble, and N_i^n is the total number of native pdb contacts for this amino acid residue. This definition is identical to that introduced by Vendruscolo et al. (2001) to construct transition-state ensembles from experimental data. In the present calculation, a contact is considered formed if $r_{ij} < 1.2r_{ij}^n$. Figure 5 compares ϕ_{calc}^i and ϕ -values obtained from experiments (Itzhaki et al. 1995; Burton et al. 1997; Chiti et al. 1999; Ternström et al. 1999). As expected, the TSM-predicted ϕ_{calc}^i -values display a tendency to be lower in α -helical regions than in β -sheet regions. It is quite clear from Figure 5 that the TSM-predicted and experimental ϕ -values do not follow the same trend along the chain sequences. The correlation coefficient between ϕ_{calc}^i and ϕ_{exp}^i is < 0.1 in each of the four cases, although there are isolated individual calculated ϕ -values that match the corresponding experimental values well. These results indicate that, at least for the four proteins studied, the TSM predictions about the structural characteristics of the rate-limiting transition-state conformational ensemble are very different from that obtained from common experimental ϕ -value analysis.

TSM “ ψ -values”

A more informative way to characterize the folding transition state than the ϕ -values for individual amino acid residues is to consider the participation of pairs of contacting amino acid residues in the transition state. We apply this to the conformational ensembles of the four native topomer states, and determine the probability ψ_{calc}^{ij} for each amino acid residue pair ij in the native contact set to be in contact in these topomer ensembles (Fig. 6). As before, a contact in these ensembles is defined by $r_{ij} < 1.2r_{ij}^n$. These TSM-predicted quantities are analogous to the ψ -values in experimental analysis, wherein bi-histidine metal binding sites are introduced for selected pairs of amino acid residue positions, and the stability of folding pathways with the metal binding pair formed relative to other pathways is regulated by metal ion concentration. The resulting ψ -value of a native contact ij , a number between 0 and 1, then represents the ratio of protein molecules that have the given

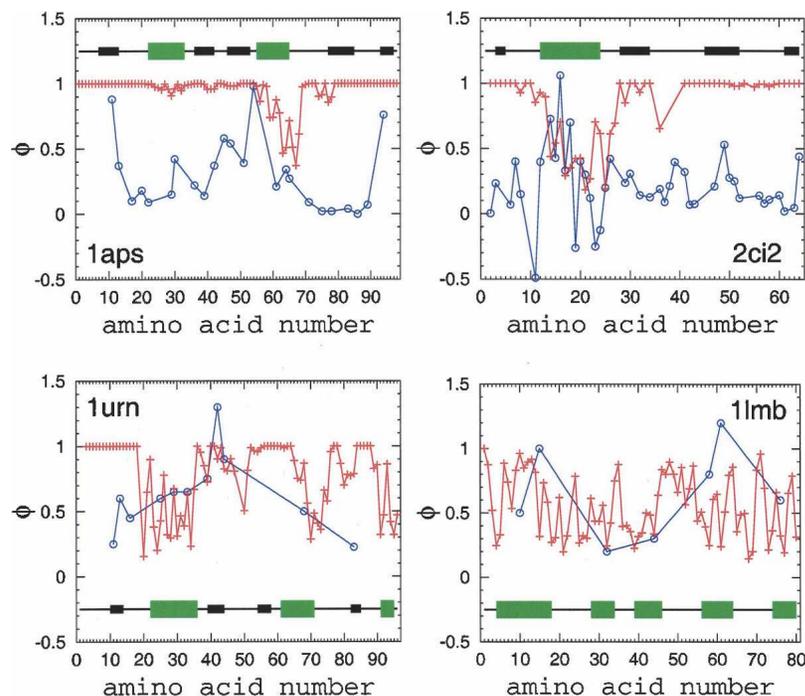


Figure 5. Comparison between TSM-predicted (+) and experimental (o) ϕ -values. The lines between data points serve merely as a guide for the eye. The secondary structure of the proteins along the sequences is indicated by large green and small black rectangular shapes for α -helical and β -sheet regions, respectively.

metal binding site in the transition state (Sosnick and Krantz 2001).

Figure 6 displays our computed ψ_{calc}^{ij} values for the $l_c = 12$, $r_c = 8$ Å native topomer ensembles. By TSM definition, the probability of being in contact is close to 1 for pairs ij that are more than l_c steps from the diagonal $i = j$. It is thus a general TSM prediction that $\psi_{\text{TSM}}^{ij} \approx 1$ for amino acid pairs ij that are distant in sequence. Not surprisingly, in line with the calculated ϕ -values, the TSM-predicted ψ -values in Figure 6 indicate that the α -helical regions are relatively more disordered in the native topomer states. ψ -value analysis is a relatively novel method, and experimental data are quite limited thus far. None of our four proteins has been studied experimentally with this technique. Nonetheless, TSM predictions in Figure 6 should be useful for future assessment when pertinent experimental data become available.

TSM and native-centric explicit-chain folding models

In the above, we have shown by explicit-chain sampling that the population of the hypothetical native topomer state as a fraction of all accessible conformations is dramatically lower than that originally stipulated by Makarov and Plaxco (2003), and that the structural properties of the native topomer states of the four

proteins considered in this study do not appear to resemble that of the corresponding folding transition states inferred from conventional experimental ϕ -value analysis. To further elucidate the relationship between the TSM picture and explicit-chain dynamics, we now consider a class of native-centric models that allow for direct kinetic simulations of folding rates (Clementi et al. 2000; Koga and Takada 2001; Kaya and Chan 2003a).

One of the main justifications for the use of native-centric models to study the folding process has been the empirically observed relationship between native topology and folding rate, emphasizing the important energetic information embodied in the native topology of natural proteins. This principle is, naturally, also central to the TSM. We therefore find it worthwhile to compare and to put into context these two types of approaches. In particular, we choose to compare the TSM with the Langevin dynamics version (Kaya and Chan 2003a) of the continuum C_α model of Clementi et al. (2000), because a recent variation of this explicit-chain modeling construct has been shown to provide a good correlation between model-predicted and real folding rates of 16 apparent two-state proteins (Chavez et al. 2004).

Conformational search in these native-centric models is directed, not unbiased. The main energy term driving chain collapse and folding in this native-centric model

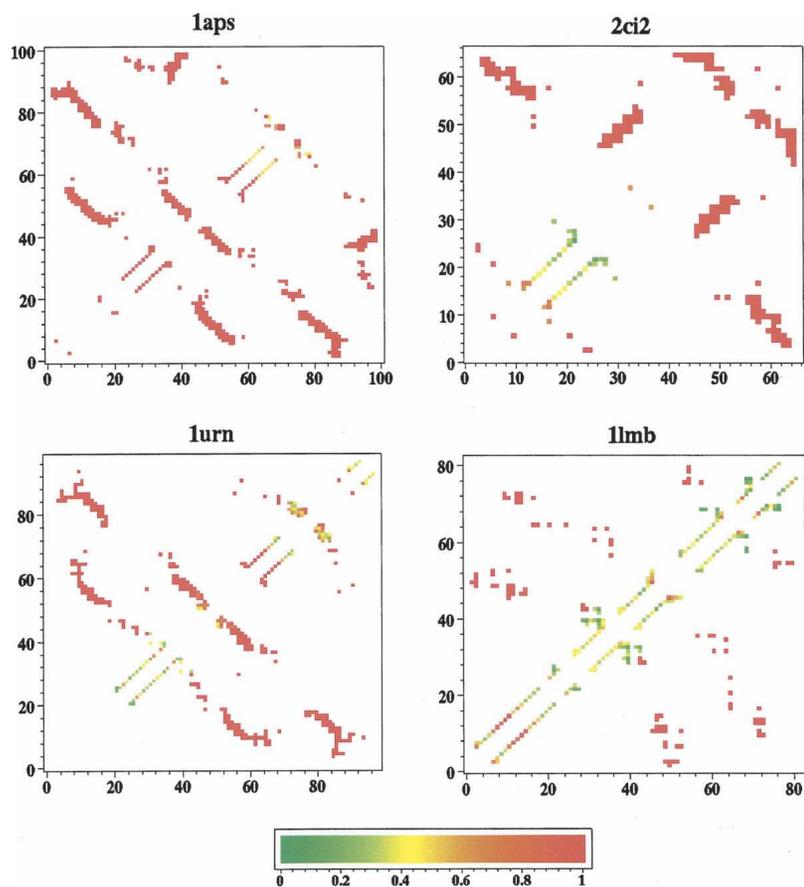


Figure 6. Probabilities of native contacts being formed (ψ_{calc}^{ij} -values) in the $l_c = 12$, $r_c = 8 \text{ \AA}$ topomer ensembles. The color scale goes from $\psi = 0$ (green) to $\psi = 1$ (red). The triangular regions above and below the main diagonal provide identical information.

consists of Lennard-Jones-like interactions between every pair of amino acid residues ij that satisfies $|i - j| > 3$ and forms a contact in the native pdb structure (Clementi et al. 2000). We consider two different definitions for two residues to be a native contact pair in this model: (1) if their C_α - C_α separation in the native pdb structure is less than a certain cutoff distance r_c ; (2) if any two heavy (non-hydrogen) atoms, one from each residue, are within 4.5 \AA ; this is identical to the definition used recently by Chavez et al. (2004). Definition (1) is sensitive to the choice of r_c . It turns out that the thermodynamic behaviors of three of our four proteins do not appear to be two-state-like for $r_c = 6 \text{ \AA}$ and $r_c = 7 \text{ \AA}$ (data not shown). Here we focus only on the better-behaved $r_c = 8 \text{ \AA}$ case. Computational runs of 10^9 simulation time steps are executed for each of the four proteins for the two different native contact sets we just described, using the potential energy function and Langevin dynamics simulation protocol detailed elsewhere (Equations 1, 2; see related discussion in Kaya and Chan 2003a). In the present investigation, the

simulations are performed at these model proteins' respective folding temperature T_f , the temperature at which the given chain model's specific heat capacity is at its maximum.

Figure 7 shows our simulated free-energy profiles in the fraction of native contacts Q . The location of the free-energy barrier along these profiles (the single peak between the unfolded and folded minima at low and high Q , respectively) varies slightly between the four proteins and the two different contact sets. The chain structures in these peak regions are commonly taken as the transition state conformations for these and similar native-centric explicit-chain models (Clementi et al. 2000; Nymeyer et al. 2000; Chavez et al. 2004). However, it has been pointed out that not all structural details of transition-state conformations can be captured by such simple procedures, because the factors determining whether a conformation belongs to the folding transition state can be complex (Du et al. 1998; Hubner et al. 2004). In other words, some true transition-state conformations of the model may reside

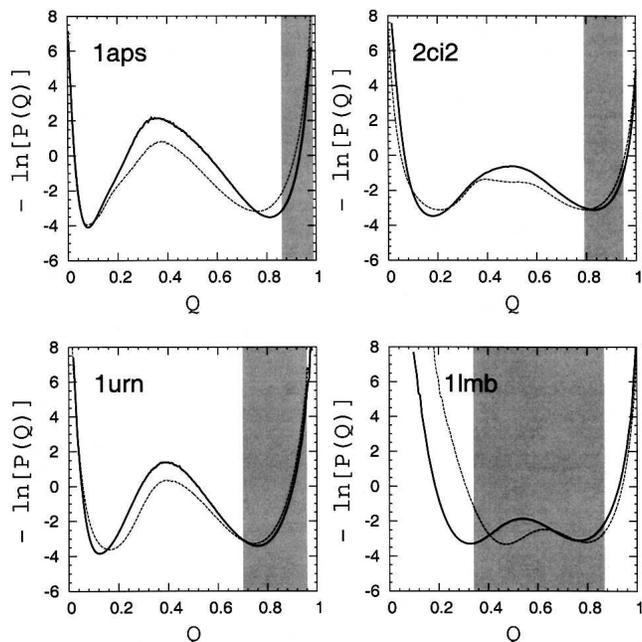


Figure 7. Free-energy profiles for the four proteins are given by the negative logarithm of the conformational distribution $P(Q)$ in Q , obtained in this study by simulations of native-centric models (Kaya and Chan 2003a) at $T \approx T_f$. Results from two native contact sets are shown: (1) contact pairs are defined by native C_α - C_α distance $r_{ij}^n < 8 \text{ \AA}$ (dashed curves); and (2) contact pairs are defined by the shortest spatial separation between non-hydrogen atoms in the two amino acid residues as described by Chavez et al. (2004) and in the text (solid curves). For comparison with the hypothetical TSM picture, the range of variation of $\langle Q \rangle$ of the corresponding native topomers across different (l_c, r_c) criteria (Table 3) is indicated by the shaded areas.

outside the Q -based free-energy peak region. Nonetheless, inasmuch as kinetic transitions between different Q -values are not too discontinuous, conformations populating the free-energy barrier region are indicative of general features of the conformations associated with the rate-limiting step in folding (Kaya and Chan 2002). With this in mind, in the following discussion we refer to the conformations in the free-energy barrier regions in Figure 7 as the putative transition states of the native-centric explicit-chain models.

Figure 7 shows that these putative transition-state conformations are found roughly around $Q \approx 0.4$ – 0.6 , and that they are quite dissimilar to the corresponding native topomer ensembles, which have very different $\langle Q \rangle$ -values (Fig. 7, shaded areas). For 1aps, 2ci2, and 1urn, in terms of Q , the native topomers are far more native-like than the explicit-chain putative transition states. In these cases, even if some of the true transition-state conformations do reside outside the Q -based free-energy peak region, it is highly unlikely that they would share many similarities with the native topomer ensembles that have Q -values essentially equal to (2ci2

and 1urn) or higher than (1aps) the Q -based native free-energy minima of the explicit-chain models (although the possibility cannot be ruled out entirely without an exhaustive determination of transition-state conformations in the explicit-chain models). For 1lmb, the situation is more complicated because of the strong dependence of $\langle Q \rangle$ on the cutoff parameters l_c and r_c . Taken together, these observations suggest strongly that the native topomer states do not generally correspond to the folding transition state in these models. Instead, they are more representative of conformations that are visited after the rate-limiting step in the folding of these explicit-chain models.

Finally, we compare the TSM rate-prediction formula $k_f \propto \Lambda_{\text{DA}}^{a_b}$ and the folding rate results from the native-centric models. Following Chavez et al. (2004), we determine folding rates in the native-centric model at T_f using the direct Langevin dynamics simulation method of Kaya and Chan (2003a). Folding and unfolding rates are identical at the transition midpoint. Thus, the folding rate k_f^{nc} of the native-centric models at T_f is defined as the reciprocal of the mean first passage time (MFPT) from the beginning of an unfolded “phase” to the initiation of a folded phase, or vice versa, during the Langevin dynamics simulation [i.e., $k_f^{\text{nc}} = (\text{MFPT})^{-1}$]. The unfolded and folded phases correspond to $Q \leq Q_u$ and $Q \geq Q_f$, respectively, where Q_u and Q_f are the unfolded (low Q) and folded (high Q) free-energy minima. The MFPT was determined by averaging over 40 (1aps) to 3225 (1lmb) folding/unfolding events recorded during 10^9 simulation time steps. Figure 8 shows that the correlation between $\Lambda_{\text{DA}}^{a_b}$ and k_f^{nc} is good, which is perhaps not surprising given that both sets of results have been shown previously to correlate strongly with real folding rate data (Makarov and Plaxco 2003; Chavez et al. 2004). We note, however, that both the topomer search rate-prediction formula and

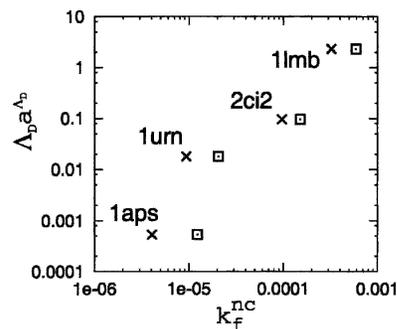


Figure 8. Comparison between the TSM quantity $\Lambda_{\text{DA}}^{a_b}$ ($a = 0.86$) (Makarov and Plaxco 2003) of the four proteins and the corresponding folding rate k_f^{nc} computed in this study from direct kinetic simulations of the native-centric Langevin dynamics models at each model’s T_f . Results for k_f^{nc} denoted by \times and \square are, respectively, for the native contact sets (1) and (2) specified in the caption for Fig. 7.

the native-centric model predict 1urn to fold roughly an order of magnitude slower than 2ci2, but in fact the opposite is true (see Table 1).

Despite the limitations of native-centric models with essentially pairwise additive interactions (which include the models used here) in their ability to capture experimental features of kinetic cooperativity when native stability is significantly higher than that at the transition midpoint (Kaya and Chan 2003a; Chan et al. 2004), comparing them with TSM predictions offers valuable insights. The predicted folding rates of the two approaches correlate well, but the native topomer states most likely do not correspond to the explicit-chain transition states. Moreover, as an unbiased search for the native topomer state is not viable (see above), it is not clear whether there is any definitive way to construct viable explicit-chain models that are consistent with the TSM. In any event, the results in Figures 7 and 8 indicate clearly that a good rate correlation by itself is far from sufficient to pin down the underlying folding mechanism.

For the native-centric explicit-chain model considered here, every point in the conformational space has an energetic bias toward the native state. Frustration in the model arises solely from chain connectivity and excluded-volume effects. As Q increases, the combined effects of these incremental energetic favorabilities and the reduced conformational freedom as the chain acquires an increasing number of favorable contacts result in a free-energy slope (with respect to Q) that becomes first zero and then negative at intermediate values of Q . In contrast, TSM predicts conformational characteristics that are much more native-like at the rate-limiting step of folding. In this connection, it has been pointed out that, as a consequence of experimentally observed protein folding cooperativity, “an excess of 90% of the native structure is required for the free energy of a typical single-domain protein to drop below zero” (Gillespie and Plaxco 2004, p. 855). However, the configurational position of the transition state is determined by the gradient of free energy with respect to the reaction coordinate(s), not by the value of free energy itself. As an illustration of this basic principle, we note that a chain model’s stability and cooperativity can be arbitrarily enhanced by adding an ad hoc favorable energy to the unique native structure as a whole without affecting the folding kinetics (Kaya and Chan 2003b). Therefore, the above observation of Gillespie and Plaxco does not necessarily imply that transition states of real, cooperative proteins have to be very close to the native state. In fact, such a proposal would appear to be inconsistent with many chevron predictions of intermediate solvent exposure of folding transition states (Plaxco et al. 1998).

Discussion

The point of contention of this article is not the remarkable correlation between folding rates of apparent two-state proteins and parameters derived from native topology. The empirical success of these parameters—these include the original CO , LRO , and the $\Lambda_{\text{D}}^{\text{Ad}}$ of TSM, and several others—suggests strongly that they are capturing important aspects of protein physics, which we should try to better understand. The subject matter of our investigation is whether the folding mechanism offered by the TSM formulation of Makarov et al. (2002) and Makarov and Plaxco (2003) is physically viable. In other words, is it mathematically and physically possible for the TSM $\Lambda_{\text{D}}^{\text{Ad}}$ formula to really describe the topomer search process postulated by the TSM picture? Generally speaking, multiple models based on very different underlying mechanisms can give similar predictions (cf. Fig. 8), thus a good correlation with real folding rate is not sufficient for validating a proposed mechanism. To ensure that the predictions of a model formulation do, indeed, follow logically from the stated assumptions, issues of mathematical validity and internal consistency should be examined as well.

In obtaining ensembles of structures consistent with the Makarov-Plaxco definition of a native topomer, we found it natural to apply harmonic forces between the Λ_{D} sequence-distant contacts required to be in proximity by the TSM. It is possible that using a different strength k_{top} of the harmonic forces can have some effect on the results. However, in view of the rather strong dependence of Λ_{D} on the parameters l_c and r_c , the dependence on k_{top} is likely not critical. It has also been suggested by Makarov and Plaxco (2003) that “proximity” in their topomer definition does not necessarily have to mean close in space but refers to “any orientation in which elements can collide to form the native contacts more rapidly than the rate-limiting step in folding” (Makarov and Plaxco 2003, p. 22). But it is unclear how such a description should be implemented in an explicit-chain context. Moreover, under such a definition, there is no longer a close conceptual link between the topomer search picture of folding and the parameter Λ_{D} defined by the simple criterion $r_{ij}^{\text{N}} < r_c$.

Our explicit-chain computation demonstrates that the Makarov-Plaxco formulation has grossly overestimated the probability of finding the hypothetical native topomer among all accessible conformations, by tens of order of magnitudes in some cases (Table 4). This means that the TSM $\Lambda_{\text{D}}^{\text{Ad}}$ formula does not actually describe the postulated topomer search process in the TSM picture. The likely reason for this failure is that excluded volume effects on chain conformations were neglected. Unlike the original topomer formulation of

Debe et al. that took excluded volume into account (Debe and Goddard 1999; Debe et al. 1999), the conformational statistics of Makarov et al. (2002) were based on Gaussian inert (phantom) chains. It is not difficult to see how a neglect of excluded volume can lead to such large errors: If there were no excluded volume, all C_α positions of a phantom protein chain can be confined within a sphere of diameter $\sim 3.8 \text{ \AA}$, the virtual bond length between two sequential C_α 's. Since all inter- C_α distances in such conformations are less than the cutoff r_c , these unphysical conformations are included in the native topomer. By the same token, unphysical topomer conformations can be constructed for spheres with different diameters $< r_c$. These and other similarly unphysical "super-compact" conformations satisfying a given (l_c, r_c) criterion are all accepted as part of the native topomer state in the Makarov-Plaxco formulation, leading to a hugely inflated conformational count.

Another way to illustrate this problem of gross conformational overcount is the following. For real chains with excluded volume, when a sufficiently large set of consistent pairwise contact constraints is imposed, the number of viable conformations would reduce to unity or very close to unity. After all, this polymer principle is what allows compact protein native structures to be determined by NMR. (Presumably, hypothetical native topomer ensembles can be explored using common softwares such as XPLOR for NMR structure determination by imposing TSM interresidue distance restraints. Indeed, Figure 2 is reminiscent of NMR structural displays. It should be noted, however, that in NMR structure determination, instead of simple harmonic constraints, biharmonic, square-well, and other restraining potentials are usually adopted for distance geometry; see, e.g., Brünger et al. 1986; Brünger 1992.) But this basic condition is not satisfied by the Makarov-Plaxco formulation. Take 2ci2 as an example. For $r_c = 8 \text{ \AA}$, there are a total of 124 native contacts ($r_{ij}^n < 8 \text{ \AA}$, $|i - j| > 3$; the corresponding numbers for 1aps, 1urn, and 1lmb are 263, 213, and 141, respectively). This number is only slightly larger than the 108 native contacts that define the $l_c = 4, r_c = 8 \text{ \AA}$ 2ci2 native topomer (Table 2). Direct explicit-chain simulations of C_α models have indicated that the set of 124 native contacts is sufficient to essentially specify the 2ci2 native conformation uniquely (up to a mirror symmetry). If we substitute $\Lambda_D = 124$ into the Makarov-Plaxco formula $P(\Lambda_D) = \gamma a^{\Lambda_D}$ with $a = 0.86$, we obtain a probability of $P(124) = 3.0 \times 10^{-13}$. Since the unfolded states of small, two-state proteins are expected to consist of mostly open conformations (Kohn et al. 2004), the number of all accessible 2ci2 conformations may be very roughly estimated as $\sim 10^{61}$ by assuming three rotational states per dihedral angle (see above).

Even a much more conservative consideration based on three rotational states per amino acid residue yields an estimate of $3^{64} \approx 3 \times 10^{30}$. Therefore, if the Makarov-Plaxco $P(\Lambda_D)$ expression is adequate for a realistic chain, the probability factor it predicts for $\Lambda_D = 124$ should be sufficiently small to reduce the huge number of all accessible 2ci2 conformations to order unity. But the predicted $P(124) \sim 10^{-13}$ factor is 17 to 48 orders of magnitude too large to do so, underscoring the formula's gross overestimation of the conformational freedom of compact chains with excluded volume.

Polymer excluded volume is notoriously difficult to treat by analytical methods, especially for compact conformations (Barber and Ninham 1970; Freed 1987; Chan and Dill 1991). Moreover, the impact of excluded volume effects is much more severe for compact conformations than for open conformations (Chan and Dill 1989; Camacho and Thirumalai 1993). In this context, explicit-chain modeling is indispensable in many applications. In the absence of the mathematical discipline and the deductive logic it affords, errors incurred by analytical approximations and semiquantitative arguments are often difficult to delineate.

Based on their $P(\Lambda_D)$ estimate, Makarov and Plaxco (2003) have stated that "this is precisely how the topomer search model solves Levinthal's paradox; whereas the number of conformations in the native topomer is small relative to the total number of conformations available in the unfolded ensemble, it is enormously larger than unity" (Makarov and Plaxco 2003, p. 22). The simple considerations above and our explicit-chain sampling data have made it quite clear that this assertion is not tenable. Although the number of conformations in a native topomer is larger than unity, in general it is very far from sufficiently large to allow an unbiased search of the native topomer among all accessible conformations to be a viable solution to Levinthal's paradox.

It has long been recognized that protein folding cannot be achieved by unbiased conformational search (Baldwin 1994). In light of the present findings, it is even more clear that the neglect of energetic contributions to the rate-limiting step of folding in the bulk of the TSM argument is not appropriate. The proponents of TSM have acknowledged that this is "obviously a potentially significant omission" (Makarov and Plaxco 2003, p. 22). Indeed, in order for protein folding to be achievable in physiological time scales, conformational sampling has to have elements of directed search (Leopold et al. 1992; Zwanzig et al. 1992; Bryngelson et al. 1995; Thirumalai and Woodson 1996; Dill and Chan 1997; Dobson et al. 1998), although the high degrees of thermodynamic and kinetic cooperativity of many natural proteins imply that their underlying

energy landscapes may be somewhat different from those derived from certain common protein chain models (Kaya and Chan 2003b; Chan et al. 2004; Scalley-Kim and Baker 2004).

The main thrust of the TSM discourse is on the estimation of the (incorrectly) presumed affordable entropic cost of unbiased search for the native topomer. Nonetheless, the possibility of incorporating a general native-favoring energy term in the topomer search process has also been explored (Makarov et al. 2002). A desire to capture the high degree of cooperativity in protein folding was one of the motivations for the TSM proposal (Gillespie and Plaxco 2004). An explicit-chain construct that has apparently been inspired by the TSM perspective is the lattice model, with excluded volume, of Jewett et al. (2003). Instead of the usual linear relationship between Q and energy E in common $G\bar{o}$ models, Jewett et al. (2003) introduced a nonlinear Q - E relationship, controlled by a parameter s , to enhance the cooperativity of their models and yet (when $s \neq \infty$) allow for a native-centric directed conformational search. Relative to the corresponding common $G\bar{o}$ models, the Jewett et al. construct results in an improvement in correlation between model folding rate and CO , and a slight increase in the folding rate diversity. But these improvements are considerably less significant than another lattice model that uses a different cooperative interaction scheme involving local-nonlocal coupling (see Chan et al. 2004, Fig. 7). It is unclear whether folding rate correlation with CO can be meaningfully improved further within the class of model interaction schemes of Jewett et al. (2003) by increasing their cooperative parameter s beyond the highest value of $s = 3$ that they used because $s \rightarrow \infty$ entails a Levinthal search. Such an unphysical process would most likely also abolish the dependence of folding rates on native topology. While these models are rudimentary, the following conceptual implication of this comparison is clear: Cooperativity per se is insufficient for protein-like correlation between folding rate and native topology. The specific energetic interactions that give rise to the cooperativity is the determining factor (Kaya and Chan 2003c). The TSM folding scenario of unbiased search for the native topomer has not been verified by any explicit-chain model with a proper account of excluded volume. In fact, the analysis in this paper demonstrates that such a search cannot succeed. Inasmuch as the lattice model of Jewett et al. (2003) is an explicit-chain implementation of a more relaxed version of the TSM idea (i.e., without insisting on an unbiased search), the lattice model results mean that even such a relaxed TSM-inspired mechanism, when evaluated in an explicit-chain context, is quite far from capturing the trend of topology dependence in real protein folding.

Besides the unrealistic unbiased conformational search process it advocated, the hypothetical TSM folding process may be regarded as a particular form of a general nucleation mechanism (Fersht et al. 1992; Abkevich et al. 1994; Fersht 1995; Thirumalai and Guo 1995), with TSM offering a more definitive, and therefore falsifiable, prediction on the structural characteristics of folding transition states (cf. Fig. 2). It follows that not every protein model that produces a high degree of cooperativity can be regarded as an implementation of the TSM idea. Leaving aside the question of how the TSM-predicted sets of rate-limiting structures are located by conformational search, we have attempted a simple experimental evaluation of these structural predictions. The results so far are not positive.

An aspect of the TSM perspective worth further exploring is the idea relating to the size of transition state conformational ensemble as part of the entropic component of the free-energy barrier to folding (Bai et al. 2004). However, as is clear from the discussion above, energetic or enthalpic contributions to the rate-limiting step in folding cannot be neglected. Favorable intrachain interactions must exist in chains in the unfolded ensemble to direct the conformational search. Experimental observations of protein folding cooperativity using low-resolution techniques should serve as an important modeling constraint (Chan et al. 2004). But cooperativity as such does not eliminate all conformations with free energies intermediate between that of the folded and unfolded states. Sparsely populated or "hidden" intermediate conformations exist (Bai et al. 1995; Ollerenshaw et al. 2004; Kaya and Chan 2005). An elucidation of the protein folding process must involve deciphering the stability and kinetic connectivities of these conformations, because this information holds the key to how conformational search is directed during folding. Further advance would require concerted efforts in theory and experiment. In this endeavor, self-contained, explicit-chain modeling that embodies an adequate account of polymer excluded volume is an essential tool.

Materials and methods

Estimating the probabilities $P(\Lambda_D)$

To estimate the thermodynamic quantities $P(\Lambda_D)$ in the explicit-chain model defined by the energy function E_0 in Equation 1, we use the following computational strategy: An auxiliary energy function $E_\lambda = E_0 + \lambda E_{\text{bias}}$ is used to calculate two-dimensional probability histograms $P_\lambda(n_D, E_{\text{bias}})$, where n_D is the number of sequence-distant native contacts, for a series of K discrete bias strengths $\lambda_1, \dots, \lambda_K$. The maximum bias strength, λ_K , is chosen so that there is a significant population of the native topomer state at that λ , namely, $P_{\lambda_K}(\Lambda_D) > 10^{-4}$.

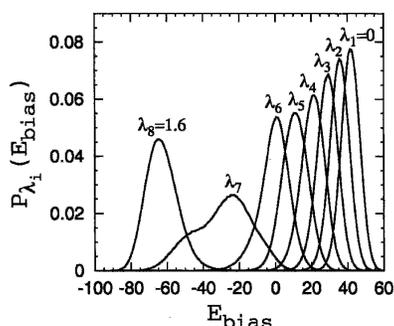


Figure 9. Marginal distributions $P_{\lambda_i}(E_{\text{bias}})$ for eight linearly separated λ_i values (from $\lambda_1 = 0$ to $\lambda_8 = 1.6$), as obtained from a simulation of the energy function $E_\lambda = E_0 + \lambda E_{\text{bias}}$ for 2ci2 and $l_c = 12$ and $r_c = 8 \text{ \AA}$ in which λ is a dynamical parameter (see text).

These histograms can then be combined through standard histogram reweighting technique (Ferrenberg and Swendsen 1989) to give improved estimates of $P_\lambda(n_D, E_{\text{bias}})$, for any λ . Hence, by using $\lambda = 0$ in this reweighting procedure we obtain $P_0(n_D, E_{\text{bias}})$, from which the desired probability $P_0(\Lambda_D) = P(\Lambda_D)$ can easily be calculated.

The above sampling strategy is straightforward. In principle, if infinite sampling were possible, the result should be independent of the specific choice of the bias potential. In practice, however, care must be taken so that the native topomer state is sampled in a representative way. We therefore construct E_{bias} not as a function of all the $N C_\alpha$ positions of the C_α chain, but only of the subset of the C_α positions that are involved in the Λ_D sequence-distant native contacts as defined by TSM. This bias function is composed of a “torsion” and a contact term:

$$E_{\text{bias}} = \sum_k \left[1 - \cos(\phi'_k - \phi_k^n) \right] - \sum_{kl} e^{-r_{kl} - r_{kl}^n)^2 / 2}, \quad (5)$$

where ϕ'_k are the torsion angles formed by four consecutive C_α positions in the TSM-defined subset, and ϕ_k^n the corresponding values in the native pdb structure. The first summation is over all such torsion angles. The second summation is over all native contacts kl ($r_{kl} < r_c$) for which both k and l are part of the TSM-defined subset of C_α positions.

We obtain the two-dimensional probability distributions $P_{\lambda_i}(n_D, E_{\text{bias}})$ by a dynamical-parameter method closely related to simulated tempering (Lyubartsev et al. 1992; Marinari and Parisi 1992; Irbäck and Potthast 1995). In simulated tempering, the configurational space of the system is expanded to include the temperature, which becomes a dynamical parameter. Here we instead let the bias parameter λ become a dynamical parameter. We use eight different linearly separated values ($K = 8$) from $\lambda_1 = 0$ to λ_K . The maximum bias used, λ_K , ranges from 1.15 (for 1aps and $l_c = 4$ and $r_c = 8 \text{ \AA}$) to 8.0 (for 1lmb and $l_c = 12$ and $r_c = 6 \text{ \AA}$). Conformational sampling is performed using Langevin dynamics, exactly as before for the constrained chain simulations, and changes between adjacent λ_i values are performed as elementary Monte Carlo moves. Frequent visits to $\lambda_1 = 0$ effectively decorrelate the generated sequence of conformations and improve sampling. For each l_c and r_c in Table 4, a simulation with a sampling time of 10^9 steps is performed to obtain the eight $P_{\lambda_i}(n_D, E_{\text{bias}})$

histograms, which are then combined to give the corresponding $P(\Lambda_D)$ value. As an example, we display in Figure 9 the marginal distributions $P_{\lambda_i}(E_{\text{bias}})$ obtained from one of these simulations, showing that sufficient overlaps between adjacent E_{bias} distributions are obtained despite the relatively few λ_i values used. Each simulation is initiated from a disordered conformation. We note that the dynamical parameter method used here works well together with the histogram reweighting technique since it provides the free energies at each chosen λ_i value (the so-called g-parameters), which are needed to achieve a self-consistent solution to the multiple-histogram reweighting equations (Ferrenberg and Swendsen 1989).

Acknowledgments

Illuminating discussions with Kevin Plaxco and Eugene Shakhnovich are gratefully acknowledged. We thank Julie Forman-Kay for a helpful discussion on distance restraints in NMR structure determination, and an anonymous referee for instructive comments. This work was partly supported by the Swedish Research Council (S.W.), the Protein Engineering Network of Centres of Excellence (Canada), and a Canadian Institutes of Health Research grant (MOP-15323) to H.S.C., who holds a Canada Research Chair in Biochemistry.

References

- Abkevich, V.I., Gutin, A.M., and Shakhnovich, E.I. 1994. Specific nucleus as the transition state for protein folding: Evidence from the lattice model. *Biochemistry* **33**: 10026–10036.
- Alm, E. and Baker, D. 1999. Prediction of protein folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci.* **96**: 11305–11310.
- Bai, Y., Sosnick, T., Mayne, L., and Englander, S.W. 1995. Protein folding intermediates: Native-state hydrogen exchange. *Science* **269**: 192–197.
- Bai, Y.W., Zhou, H.Y., and Zhou, Y.Q. 2004. Critical nucleation size in the folding of small apparently two-state proteins. *Protein Sci.* **13**: 1173–1181.
- Baker, D. 2000. A surprising simplicity to protein folding. *Nature* **405**: 39–42.
- Baldwin, R.L. 1994. Folding intermediates in protein folding. *BioEssays* **16**: 207–210.
- Barber, M.N. and Ninham, B.W. 1970. *Random and restricted walks: Theory and applications*. Gordon and Breach, New York, NY.
- Brünger, A.T. 1992. *XPLOR manual, Version 3.1*. Yale University, New Haven, CT.
- Brünger, A.T., Clore, G.M., Gronenborn, A.M., and Karplus, M. 1986. Three-dimensional structures of proteins determined by molecular dynamics with interproton distance restraints: Application to crambin. *Proc. Natl. Acad. Sci.* **83**: 3801–3805.
- Bryngelson, J.D., Onuchic, J.N., Socci, N.D., and Wolynes, P.G. 1995. Funnels, pathways and the energy landscape of protein folding: A synthesis. *Proteins Struct. Funct. Genet.* **21**: 167–195.
- Burton, R.E., Huang, G.S., Daugherty, M.A., Calderone, T.L., and Oas, T.G. 1997. The energy landscape of a fast-folding protein mapped by Ala → Gly substitutions. *Nat. Struct. Biol.* **4**: 305–310.
- Camacho, C.J. and Thirumalai, D. 1993. Minimum energy compact structures of random sequences of heteropolymers. *Phys. Rev. Lett.* **71**: 2505–2508.
- Chan, H.S. and Dill, K.A. 1989. Compact polymers. *Macromolecule* **22**: 4559–4573.
- . 1990. The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.* **92**: 3118–3135.
- . 1991. Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.* **20**: 447–490.
- Chan, H.S., Shimizu, S., and Kaya, H. 2004. Cooperativity principles in protein folding. *Methods Enzymol.* **380**: 350–379.
- Chavez, L.L., Onuchic, J.N., and Clementi, C. 2004. Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.* **126**: 8426–8432.

- Chiti, F., Taddei, N., White, P.M., Bucciantini, M., Magherini, F., Stefani, M., and Dobson, C.M. 1999. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* **6**: 1005–1009.
- Choy, W.Y. and Forman-Kay, J.D. 2001. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* **308**: 1011–1032.
- Cieplak, M. and Hoang, T.X. 2003. Universality classes in folding times of proteins. *Biophys. J.* **84**: 475–488.
- Clementi, C., Nymeyer, H., and Onuchic, J.N. 2000. Topological and energetic factors: What determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? *J. Mol. Biol.* **298**: 937–953.
- Clementi, C., Garcia, A.E., and Onuchic, J.N. 2003. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: All-atom study of Protein L. *J. Mol. Biol.* **326**: 933–954.
- Debe, D.A. and Goddard, W.A. 1999. First-principles prediction of protein folding rates. *J. Mol. Biol.* **294**: 619–625.
- Debe, D.A., Carlson, M.J., and Goddard III, W.A. 1999. The topomer-sampling model of protein folding. *Proc. Natl. Acad. Sci.* **96**: 2596–2601.
- Dill, K.A. and Chan, H.S. 1997. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **4**: 10–19.
- Dobson, C.M., Šali, A., and Karplus, M. 1998. Protein folding: A perspective from theory and experiment. *Angew. Chem. Int. Ed. Engl.* **37**: 868–893.
- Du, R., Pande, V.S., Grosberg, A.Y., Tanaka, T., and Shakhnovich, E.I. 1998. On the transition coordinate for protein folding. *J. Chem. Phys.* **108**: 334–350.
- Ejtehad, M.R., Avall, S.P., and Plotkin, S.S. 2004. Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proc. Natl. Acad. Sci.* **101**: 15088–15093.
- Faisca, P.F.N. and Ball, R.C. 2002. Topological complexity, contact order, and protein folding rates. *J. Chem. Phys.* **117**: 8587–8591.
- Ferrenberg, A.M. and Swendsen, R.H. 1989. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **63**: 1195–1198.
- Fersht, A.R. 1995. Optimization of rates of protein folding: The nucleation–condensation mechanism and its implications. *Proc. Natl. Acad. Sci.* **92**: 10869–10873.
- . 2004. Relationship of Leffler (Bronsted) α values and protein folding ϕ values to position of transition-state structures on reaction coordinates. *Proc. Natl. Acad. Sci.* **101**: 14338–14342.
- Fersht, A.R., Matouschek, A., and 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.
- Freed, K.F. 1987. *Renormalization group theory of macromolecules*. Wiley, New York, NY.
- Galzitskaya, O.V. and Finkelstein, A.V. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci.* **96**: 11299–11304.
- Gillespie, B. and Plaxco, K.W. 2004. Using protein folding rates to test protein folding theories. *Annu. Rev. Biochem.* **73**: 837–859.
- Gong, H., Isom, D.G., Srinivasan, R., and Rose, G.D. 2003. Local secondary structure content predicts folding rates for simple, two-state proteins. *J. Mol. Biol.* **327**: 1149–1154.
- Hubner, I.A., Shimada, J., and Shakhnovich, E.I. 2004. Commitment and nucleation in the protein G transition state. *J. Mol. Biol.* **336**: 745–761.
- Irbäck, A. and Potthast, F. 1995. Studies of an off-lattice model for protein folding: Sequence dependence and improved sampling at finite temperature. *J. Chem. Phys.* **103**: 10298–10305.
- Itzhaki, L.S., Otzen, D.E., and 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.
- Ivankov, D.N., Garbuzynskiy, S.O., Alm, E., Plaxco, K.W., Baker, D., and Finkelstein, A.V. 2003. Contact order revisited: Influence of protein size on the folding rate. *Protein Sci.* **12**: 2057–2062.
- Jackson, S.E. 1998. How do small single-domain proteins fold? *Fold. Des.* **3**: R81–R91.
- Jackson, S.E. and Fersht, A.R. 1991. Folding of chymotrypsin inhibitor 2. I. Evidence for a two-state transition. *Biochemistry* **30**: 10428–10435.
- Jewett, A.I., Pande, V.S., and Plaxco, K.W. 2003. Cooperativity, smooth energy landscapes and the origins of topology-dependent protein folding rates. *J. Mol. Biol.* **326**: 247–253.
- Jones, K. and Wittung-Stafshede, P. 2003. The largest protein observed to fold by two-state kinetic mechanism does not obey contact-order correlation. *J. Am. Chem. Soc.* **125**: 9606–9607.
- Karanicolas, J. and Brooks, C.L. 2003. The importance of explicit chain representation in protein folding models: An examination of Ising-like models. *Proteins Struct. Funct. Genet.* **53**: 740–747.
- Kaya, H. and Chan, H.S. 2002. Towards a consistent modeling of protein thermodynamic and kinetic cooperativity: How applicable is the transition state picture to folding and unfolding? *J. Mol. Biol.* **315**: 899–909.
- . 2003a. Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: How adequate is native-centric topological modeling? *J. Mol. Biol.* **326**: 911–931. (Corrigendum **337**: 1069–1070.)
- . 2003b. Simple two-state protein folding kinetics requires near-Levinthal thermodynamic cooperativity. *Proteins Struct. Funct. Genet.* **52**: 510–523.
- . 2003c. Contact order dependent protein folding rates: Kinetic consequences of a cooperative interplay between favorable nonlocal interactions and local conformational preferences. *Proteins Struct. Funct. Genet.* **52**: 524–533.
- . 2005. Explicit-chain model of native-state hydrogen exchange: Implications for event ordering and cooperativity in protein folding. *Proteins Struct. Funct. Bioinform.* **58**: 31–44.
- Koga, N. and Takada, S. 2001. Roles of native topology and chain-length scaling in protein folding: A simulation study with a Gō-like model. *J. Mol. Biol.* **313**: 171–180.
- Kohn, J.E., Millett, I.S., Jacob, J., Zagrovic, B., Dillon, T.M., Cingel, N., Dothager, R.S., Seifert, S., and Thiyagarajan, P., Sosnick, T.R., et al. 2004. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci.* **101**: 12491–12496.
- Leopold, P.E., Montal, M., and Onuchic, J.N. 1992. Protein folding funnels: A kinetic approach to the sequence–structure relationship. *Proc. Natl. Acad. Sci.* **89**: 8721–8725.
- Lyubartsev, A.P., Martsinovski, A.A., Shevkunov, S.V., and Vorontsov-Velyaminov, P.V. 1992. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.* **96**: 1776–1783.
- Makarov, D.E. and Metiu, H. 2002. A model for the kinetics of protein folding: Kinetic Monte Carlo simulations and analytical results. *J. Chem. Phys.* **116**: 5205–5216.
- Makarov, D.E. and Plaxco, K.W. 2003. The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.* **12**: 17–26.
- Makarov, D.E., Keller, C.A., Plaxco, K.W., and Metiu, H. 2002. How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc. Natl. Acad. Sci.* **99**: 3535–3539.
- Marinari, E. and Parisi, G. 1992. Simulated tempering: A new Monte Carlo scheme. *Europhys. Lett.* **19**: 451–458.
- Micheletti, C. 2003. Prediction of folding rates and transition-state placement from native-state geometry. *Proteins Struct. Funct. Genet.* **51**: 74–84.
- Muñoz, V. and Eaton, W.A. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci.* **96**: 11311–11316.
- Myers, J.K. and Oas, T.G. 1999. Contribution of a buried hydrogen bond to λ repressor folding kinetics. *Biochemistry* **38**: 6761–6768.
- Nanias, M., Chinchio, M., Pillardy, J., Ripoll, D.R., and Scheraga, H.A. 2003. Packing helices in proteins by global optimization of a potential energy function. *Proc. Natl. Acad. Sci.* **100**: 1706–1710.
- Nymeyer, H., Socci, N.D., and Onuchic, J.N. 2000. Landscape approaches for determining the ensemble of folding transition states: Success and failure hinges on the degree of frustration. *Proc. Natl. Acad. Sci.* **97**: 634–639.
- Ollershaw, J.E., Kaya, H., Chan, H.S., and Kay, L.E. 2004. Sparsely populated folding intermediates of the Fyn SH3 domain: Matching native-centric essential dynamics and experiment. *Proc. Natl. Acad. Sci.* **101**: 14748–14753.
- Plaxco, K.W., Simons, K.T., and Baker, D. 1998. Contact order, transition state placement and the refolding rates of single-domain proteins. *J. Mol. Biol.* **277**: 985–994.
- Sanchez, I.E. and Kiefhaber, T. 2003. Origin of unusual ϕ -values in protein folding: Evidence against specific nucleation sites. *J. Mol. Biol.* **334**: 1077–1085.
- Scalley-Kim, M. and Baker, D. 2004. Characterization of the folding energy landscapes of computer generated proteins suggests high folding free energy barriers and cooperativity may be consequences of natural selection. *J. Mol. Biol.* **338**: 573–583.
- Selvaraj, S. and Gromiha, M.M. 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.
- Silow, M. and Oliveberg, M. 1997. High-energy channeling in protein folding. *Biochemistry* **36**: 7633–7637.
- Sosnick, T.R. and Krantz, B.A. 2001. Engineered metal binding sites map the heterogeneous folding landscape of a coiled coil. *Nat. Struct. Biol.* **8**: 1042–1047.
- Ternström, T., Mayor, U., Akke, M., and Oliveberg, M. 1999. From snapshot to movie: ϕ analysis of protein folding transition states taken one step further. *Proc. Natl. Acad. Sci.* **96**: 14854–14859.
- Thirumalai, D. and Guo, Z. 1995. Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labeling experiments. *Biopolymers* **35**: 137–140.
- Thirumalai, D. and Woodson, S.A. 1996. Kinetics of folding of proteins and RNA. *Acc. Chem. Res.* **29**: 433–439.
- van Nuland, N.A., Chiti, F., Taddei, N., Raugei, G., Ramponi, G., and Dobson, C.M. 1998. Slow folding of muscle acyphosphatase in the absence of intermediates. *J. Mol. Biol.* **283**: 883–891.
- Vendruscolo, M., Paci, E., Dobson, C.M., and Karplus, M. 2001. Three key residues form a critical contact network in a protein folding transition state. *Nature* **409**: 641–645.
- Weikl, T.R. and Dill, K.A. 2003. Folding rates and low-entropy-loss routes of two-state proteins. *J. Mol. Biol.* **329**: 585–598.
- Zhou, H. and Zhou, Y. 2002. Folding rate prediction using total contact distance. *Biophys. J.* **82**: 458–463.
- Zwanzig, R., Szabo, A., and Bagchi, B. 1992. Levinthal's paradox. *Proc. Natl. Acad. Sci.* **89**: 20–22.