Multiple Protein Folding Nuclei and the Transition State Ensemble in Two-State Proteins

D.K. Klimov[1,2] **and D. Thirumalai**[1,2]*

[1]*Department of Chemistry and Biochemistry, Center for Biomolecular Structure and Organization, University of Maryland, College Park, Maryland*

[2]*Institute for Physical Science and Technology, University of Maryland, College Park, Maryland*

ABSTRACT **Using exhaustive simulations of lattice models with side-chains, we show that optimized two-state folders reach the native state by a nucleation-collapse mechanism with multiple folding nuclei (MFN). For both the full model and the Go version, there are certain contacts that on an average participate in the critical nuclei with higher probability than the others. The high- (⩾0.5) probability contacts are largely determined by the structure of the native state. Comparison of the results for the full sequence and the Go model shows that non-native interactions compromise the degree of cooperativity and stability of the native state. From an extremely detailed analysis of the folding kinetics, we find that non-native interactions are present in the folding nuclei. The folding times decrease if the non-native interactions in the folding nuclei are made neutral or repulsive. Using cluster analysis and making no prior assumption about reaction coordinate, we show that both full and Go models have three distinct transition states that give a structural description for the MFN. In the transition states, on an average, about two-thirds of the sequence is structured, whereas the rest is disordered, reminiscent of the polarized transition state in the SH3 domain. Our studies show that Go models cannot describe the transition state characteristics of two-state folders at the molecular level. As a byproduct of our investigations, we establish that our method of computing the transition state ensemble is numerically equivalent to the technique based on the stochastic separatrix, which also does not require a priori knowledge of the folding reaction coordinate. Proteins 2001;43:465–475.**

© 2001 Wiley-Liss, Inc.

Key words: lattice protein models; protein folding kinetics and thermodynamics; nucleation-collapse mechanism; cooperativity; non-native interactions

INTRODUCTION

The study of folding mechanisms of proteins continues to be of abiding interest[1–4] because of its importance in understanding more complex biological phenomena, such as interaction between biomolecules and assembly of complex structures. Advances in experimental methodologies[4] and computational models[1–3] have provided an outline of how single-domain two-state proteins fold. Such proteins (e.g., CI2, SH3, and FKBP12) are assumed to fold by a nucleation-collapse (NC)[3,5–8] (also referred to as nucleation condensation[9]) mechanism. According to this model, the folded state is reached rapidly, once a critical number of tertiary contacts is formed.[5–8,10,11] In the NC model, the rate-determining step is the search for one of the critical nuclei. The set of distinct structures corresponding to the critical nuclei represents the transition state (TS) ensemble. General arguments and specific computations suggest that typically there are multiple folding nuclei (MFN),[3,7,11] consistent with the notion that the TS ensemble is typically heterogeneous.[12] Most experiments can be interpreted within the MFN model of the NC mechanism.

Although there is general agreement that NC mechanism explains the folding routes followed by two-state proteins, the precise relationship between the multiplicity of the folding nuclei[7,11] and the diversity of the TS ensemble[12] has not been fully clarified. Moreover, with the exception of one recent study, the role of non-native interactions in the TS has not been addressed.[13] This is particularly important because a number of recent studies ignore non-native interactions completely (i.e., adopt a Go model[14]), in order to speed up folding.[15] The extent to which this drastic simplification affects the results is unknown. In this article, we use lattice model with side-chains to explore both issues. Using numerically rigorous computations, we show that (1) for optimized two-state folders, there are multiple folding nuclei. Analysis of the corresponding structures over 100 folding trajectories shows that they can be clustered into a small number of TS. The structures corresponding to these define the TS ensemble; and (2) the Go model does give a qualitatively correct description of the folding thermodynamics and kinetics, as long as the sequence is minimally frustrated. However, the Go model fails to capture fully the nature of the TS ensemble at the molecular level.

Grant sponsor: National Science Foundation; Grant number: CHE99-75150.

*Correspondence to: D.K. Klimov, Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742. E-mail: klimov@glue.umd.edu

Received 20 October 2000; Accepted 26 January 2001

MODEL AND SIMULATION METHODS

Lattice Model With Side-Chains

A brief description is provided in this section, as the details of the model are given elsewhere.[16] A polypeptide chain is modeled by a sequence of $N = 15$ backbone beads, representing the C_α-carbons of a protein backbone. Side-chain beads, which mimic amino acid residues, are attached to each backbone bead. In all, there are $2N$ beads in the model, which occupy the vertices of cubic lattice. The conformation of a protein is specified by $2N$ vectors $\mathbf{r}_{b,i}$, $\mathbf{r}_{s,i}$, $i = 1, 2, \ldots, N = 15$, where $\mathbf{r}_{b,i}$ and $\mathbf{r}_{s,i}$ are the positions of backbone and side-chain beads, respectively. Interactions of two types are included in the energy function: (1) self-avoidance condition imposes the restriction that a given lattice site can be occupied only once (either by a backbone or by a side-chain bead), and (2) The contact interactions between the side chain beads i and j B_{ij} ($|i - j| \geq 1$) are given by pairwise statistical potentials computed by Kolinski et al.[17] We consider short-range contact interactions by assuming that side-chains form a contact only when they are nearest neighbors on a lattice. The energy of a conformation is

$$E = \sum_{i<j} B_{ij}\delta(|\mathbf{r}_{s,i} - \mathbf{r}_{s,j}| - a) \qquad (1)$$

where δ is the Kronecker delta function and a is the lattice spacing. The contact energies B_{ij} are expressed in units of RT, where T is the absolute temperature.

Computation of Thermodynamic Quantities

Thermodynamic functions describing folding and collapse transitions were computed using the multiple histogram technique[16,18] and the λ-expansion method.[19] According to the λ-expansion method[19] the energy function is divided into contributions resulting from native (E_{N}) and non-native (E_{NN}) interactions. The non-native energies are scaled by a factor λ ($|\lambda| \leq 1$), which permits continuous exploration of thermodynamic properties at any strength of non-native interactions without performing additional simulations. In particular, $\lambda = 0$ corresponds to the Go model, in which all non-native interactions are suppressed, while at $\lambda = 1$ the full scale of non-native interactions is restored. We apply these methods to compute the thermal averages of the overlap function χ[16]

$$\chi = 1 - \frac{1}{2N^2 - 3N + 1}\left[\sum_{i<j}\delta(r_{ij}^{ss} - r_{ij}^{ss,\mathrm{N}}) + \sum_{i<j+1}\delta(r_{ij}^{bb} - r_{ij}^{bb,\mathrm{N}}) + \sum_{i \neq j}\delta(r_{ij}^{bs} - r_{ij}^{bs,\mathrm{N}})\right] \qquad (2)$$

where r_{ij}^{bb}, r_{ij}^{ss}, r_{ij}^{bs} refer to the distances between backbone beads, side beads, and between backbone and side beads, respectively, and superscript N refers to the native state. The factor $2N^2 - 3N + 1$ ensures that in the native conformation $\chi = 0$. The thermodynamic average $\langle\chi(\lambda)\rangle$ is calculated according to

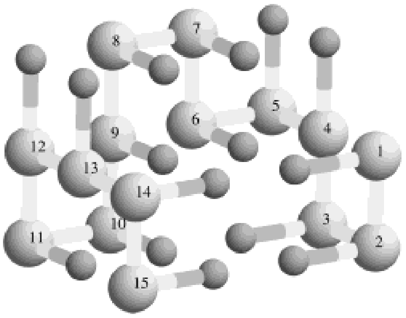

Fig. 1. Native structure for sequences A and AGO. Side-chains and C_α-carbons are shown in grey and light grey, respectively. The sequence is WVVEKWHYYVVANNAV, where a one-letter code for amino acids is used. The native conformation is compact with well-defined hydrophobic core and contains 20 contacts between side-chains. In all, cubic lattice topology permits formation of 56 nearest-neighbor contacts between side-chains, out of which only 20 are native. The native contacts (using the format $i - j(q)$, where i and j are the residues involved and q is the native contact index) are 1–2 (1), 2–3 (2), 4–5 (3), 6–7 (4), 7–8 (5), 8–9 (6), 9–10 (7), 10–11 (8), 12–13 (9), 14–15 (10), 3–6 (11), 4–7 (12), 6–9 (13), 1–6 (14), 8–13 (15), 9–14 (16), 10–15 (17), 3–10 (18), 1–14 (19), 2–15 (20). The contacts are arranged in the ascending order of the distance along the sequence between residues $|i - j|$. Generated using RasMol2.6.[20]

$$\langle\chi(\lambda)\rangle = Z^{-1} \sum_{E_{\mathrm{N}},E_{\mathrm{NN}},\chi} \chi e^{-(E_{\mathrm{N}} + \lambda E_{\mathrm{NN}})/T} \frac{\sum_{r=1}^{R} h(E_{\mathrm{N}}, E_{\mathrm{NN}}, \chi)(T_r)}{\sum_{r=1}^{R} n_r e^{f_r - (E_{\mathrm{N}}/T_r)}} \qquad (3)$$

where Z is the partition function; $h(E_{\mathrm{N}}, E_{\mathrm{NN}}, \chi)$ is the histogram collected for the variables E_{N}, E_{NN}, χ; R is the number of such histograms; T_r is the temperature of simulations, at which the rth histogram is collected; n_r is the number of states in the rth histogram; and f_r is the scaled free energy, which is calculated self-consistently from the following equation:

$$e^{-f_r} = \sum_{E_{\mathrm{N}},\chi} e^{-(E_{\mathrm{N}}/T_r)} \frac{\sum_{m=1}^{R} h(E_{\mathrm{N}}, \chi)(T_m)}{\sum_{m=1}^{R} n_m e^{f_m - (E_{\mathrm{N}}/T_m)}} \qquad (4)$$

Further details concerning the implementation of the multiple histogram method for lattice models can be found elsewhere.[16]

Sequence Selection

The native state of sequence A, which is a two-state folder,[16] is displayed in Figure 1. To illustrate the importance of non-native interactions, we also consider the Go version of sequence A, labeled AGO, which was obtained by setting $\lambda = 0$. The native states of AGO and A are identical (Fig. 1). In accord with the λ-expansion method[19] the histograms in eq. 3 were collected for AGO, and the original sequence A was recovered when $\lambda = 1$. This approach is computationally very efficient, because while the actual equilibrium simulations are done only for the AGO sequence, the spectrum of other sequences differing by the strength of non-native interactions (including A) may be studied by simply changing λ and recomputing any thermodynamic average using eq. 3. Thermodynamic functions for A computed by λ expansion and by multiple histogram method based on the direct simulations of A[16] are in excellent agreement.

In this article, we focus on two sequences, A and AGO, and compare their folding behavior to illustrate the role of non-native interactions in determining the folding mechanism of optimized two-state folders. Folding thermodynamics and kinetics have been obtained using standard Monte Carlo dynamics algorithm with the set of moves described earlier.[16] First, we discuss the thermodynamics of folding for A and AGO and consider the effect of non-native interactions. Then, we describe the folding kinetics, including the transition states for both sequences and evaluate the role of non-native contacts and the diversity of the TS ensemble.

RESULTS

Thermodynamics of Folding: Role of Non-native Interactions

Characteristic Temperatures

The ground state for sequences A and AGO is displayed in Figure 1. The collapse temperature T_θ for AGO, inferred from the location of the maximum in the specific heat C_v, is 0.29 (in reduced temperature units). We have also determined T_θ from the temperature dependence of the radius of gyration $\langle R_g \rangle$ by identifying the location of maximum in the derivative $d\langle R_g \rangle/dT$. This gives the same value of 0.29. Thus, as has been repeatedly shown in our earlier simulations,[21,22] equating T_θ with collapse transition is indeed valid. The temperature, at which the steepest decrease in energy is observed, corresponds to that at which a dramatic compaction of a chain takes place (Fig. 2). Because we are dealing with finite systems, such transitions are not very sharp. In fact, the maximum change in $\langle R_g \rangle$ from the unfolded random coil state to compact conformations occurs over a temperature interval $\Delta T_\theta \approx 0.2$. The smaller ΔT_θ becomes the sharper the collapse transition would be.

The folding transition temperature T_F is defined as the temperature, at which the fluctuations in the overlap function $\Delta\chi$ reach maximum.[16] For AGO, folding to the native state takes place at $T_F = 0.28$. The acquisition of the native state is monitored by the overlap function $\langle \chi(T) \rangle$ in Figure 2. The value of the foldability index $\sigma = (T_\theta - T_F)/T_\theta$[23] for this sequence is 0.03, which implies that AGO is expected to be a fast-folding two-state sequence, in which the processes of specific collapse and folding are nearly synchronous. Sequence A collapses at the temperature $T_\theta = 0.27$ and folds to the native state at $T_F = 0.26$ (Fig. 2). The value of σ was found to be 0.04. Thus, the data imply that both sequences should have similar folding properties. These are two-state sequences, in which folding and collapse occur almost simultaneously, as reflected in very small values of σ. It is interesting that non-native interactions do not appreciably affect the values of T_θ and T_F, which determine the "phases" of proteins.

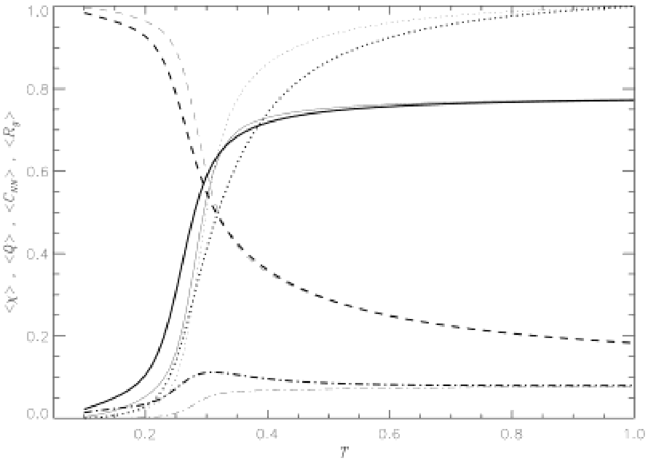

Fig. 2. Thermodynamic functions describing collapse and folding of sequences A (thick curves) and AGO (thin curves). The dotted curves indicate the normalized radii of gyration $\langle R_g \rangle$. The thermal averages of $\langle \chi \rangle$ and the fraction of native contacts $\langle Q \rangle$ are given by solid and dashed curves. The fractions of non-native contacts $\langle C_{\mathrm{NN}} \rangle$ are shown by dash dot curves. We compute the folding temperature T_F to be 0.26 and 0.28 for A and AGO, respectively (see text for details). The corresponding estimates for the collapse temperatures T_c are 0.27 and 0.29. $\langle R_g \rangle$ and $\langle \chi \rangle$ (or equally $\langle Q \rangle$) imply that collapse and folding are synchronous. $\langle C_{\mathrm{NN}} \rangle$ for A suggests a weak transient accumulation of non-native interactions at $\approx T_F$.

Stability

The effect of non-native interactions becomes more evident if we consider the stability and cooperativity of folding for A and AGO. The stability of the native state with respect to the unfolded state may be estimated by computing the free energy profile

$$F(\chi_{cg}) = -T \ln Z(\chi_{cg})$$

(Fig. 3), in which χ_{cg} is the coarse-grained values of χ (see legend to Fig. 3), and $Z(\chi_{cg})$ is the restricted partition function. The free energy difference between folded and unfolded states ΔF_{U-N} calculated at $T = 0.24 < T_F$ (see below) is -3.5 for AGO and only -1.5 for A. As expected, this shows that complete elimination of non-native interactions (as in the Go model) increases the stability of A by a factor of 2.3.

Cooperativity

The changes in the stability are coupled with the enhancement in folding cooperativity for AGO as compared with A. The cooperativity of folding transition may be inferred from the cooperativity index[16]:

$$\Omega_c = \frac{T^2_{\max}}{\Delta T}\frac{d\langle \chi \rangle}{dT} \tag{5}$$

where $T_{\max} \approx T_F$ is the temperature at which $d\langle \chi \rangle/dT$ is maximum and ΔT is the full transition width at half-

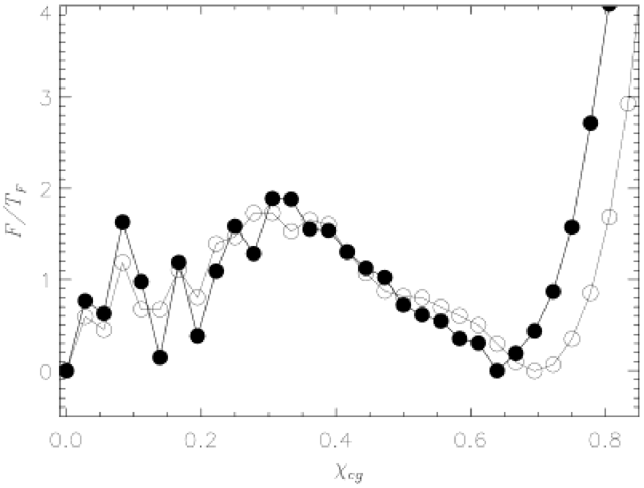

Fig. 3. Free energy profiles $F(\chi_{cg})/T_F$ for A (full circles) and AGO (open circles) at $T = T_F$ (which is 0.26 for A and 0.28 for AGO). The free energy barrier $\Delta F^{\ddagger}/kT_F$ is 1.9 and 1.7, respectively. The profiles show that folding for both sequences is two-state, because only two thermodynamic states are populated under these conditions: the unfolded with $\chi \geqslant 0.6$ and native with $\chi \approx 0$. Note that elimination of non-native interactions in AGO makes the free energy landscape considerably smoother and reduces the folding barrier. The overlap χ_{cg} is coarse-grained in such a way that each 11 discrete original values of χ are merged into one value.

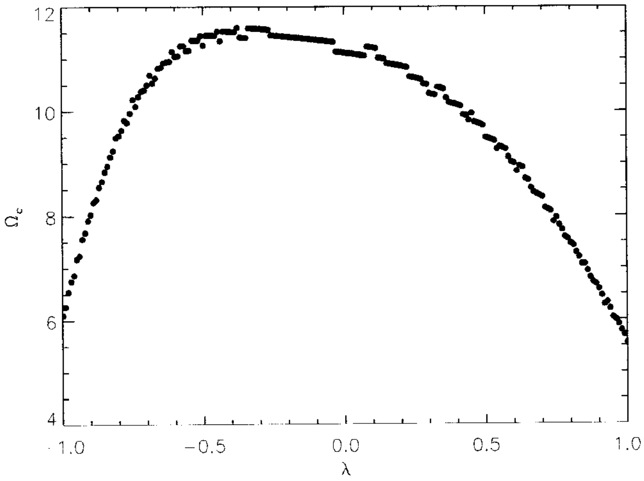

Fig. 4. Cooperativity index Ω_c as a function of the parameter λ, which controls the strength of non-native interactions. Generally, reducing the amplitude of λ results in the rise of folding cooperativity. Because the average non-native contact energy is non-zero, the maximum Ω_c is reached at $\lambda \neq 0$. This plot shows that the role of non-native interactions is confined to compromising the cooperativity of folding. Small random variations in Ω_c are due to finite accuracy in computing parameters in eq. 5.

maximum of $d\langle\chi\rangle/dT$. It is clear from eq. 5 that for a first-order transition in an infinite system $\Omega_c \rightarrow \infty$. For small systems Ω_c would be finite. Experimental data suggest that for two-state proteins $5 < \Omega_c < 100$.[16] For sequence AGO, we find $\Omega_c = 11.4$, which given the very small size of the model protein, indicates a highly cooperative two-state transition. By contrast, the cooperativity of folding of A is $\Omega_c = 5.3$, which is a factor of 2.2 lower, compared with AGO.

The cooperative transition is further evident in Figure 3, which shows that A and AGO populate only two thermodynamic states at $T \approx T_F$—unfolded state with $\chi \geqslant 0.6$ and folded with $\chi \approx 0.0$; i.e., folding is indeed two-state. Furthermore, the plot reveals that complete suppression of non-native interactions makes the free energy profile smoother and leads to a modest decrease in the free energy barrier. Although sequence A has several native-like conformations ($\chi \sim 0.1$–0.2) with low free energy, these states are completely destabilized in AGO, that further contributes to the higher stability of this sequence. Thus, cooperativity qualitatively follows the changes observed in the stability. These changes in folding characteristics are not evident, when only the differences in the characteristic temperatures are considered.

Variation of Ω_c With λ

The Go version of any sequence is an extreme limit, in which all non-native interactions are completely ignored. The power of the λ-expansion method[19] lies in the possibility of systematically examining the effect of non-native interactions as their strength is varied. Here we examine the variations in Ω_c with respect to λ (Fig. 4). In this case, λ is allowed to vary from -1 to 1; i.e., we consider not only cases in which non-native interactions are in full strength or completely suppressed, but cases in which all non-native energies reverse their signs. Figure 4 demonstrates that reducing the strength of non-native interactions leads to a gradual increase in cooperativity; Ω_c increases as λ is decreased reaching 11.6 at $\lambda = -0.38$. The maximum in cooperativity is reached at $\lambda < 0$, not at $\lambda = 0$ (the Go limit). The reason for this is that the average energy for non-native contacts is not strictly zero. For sequence A, the average energy of native contacts is -0.725, while the average non-native contact energy is -0.06. It is remarkable that although the native interactions are on an average stronger by about a factor of 10, the non-native contacts still have a profound effect on the folding thermodynamics. Thus, setting non-native interactions to zero significantly enhances cooperativity of folding, but the absolute maximum in Ω_c is attained when non-native contacts (on average) become unfavorable (not merely neutral). For sequence A, this occurs at $\lambda = -0.38$. Obviously, the precise location of maximum in Ω_c depends on the particular sequence and the model, but the generic feature of the plot shown in Figure 4 remains unchanged; that is, reducing the strength of non-native interactions or even their complete elimination results in a considerable gain in folding stability and cooperativity. The analysis also suggests that there is a range of λ ($-0.6 \leqslant \lambda \leqslant 0.1$), in which Ω_c shows little variation ($\Omega_c \geqslant 11$). Thus, many mutations that can be represented by these values of λ do not affect cooperativity and hence stability. This is consistent with previous theoretical results showing that single-domain proteins can tolerate a surprisingly large number of mutations without altering the native structures.[25]

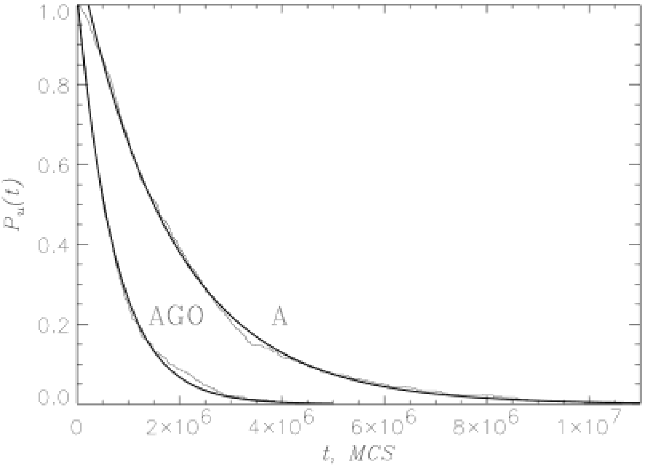

Fig. 5. Time dependence of the fraction of unfolded molecules $P_u(t)$ for A and AGO computed at $\langle\chi(T_s)\rangle = 0.25$. Data are averaged over 600 individual trajectories. $P_u(t)$ decays exponentially with the time scales $\tau_F = 2.07 \pm 0.07 \times 10^6$ MCS (A) and $0.76 \pm 0.01 \times 10^6$ MCS (AGO). Complete elimination of non-native interactions speeds up folding by a factor of 3. Exponential fits are shown by thick smooth curves.

Folding Kinetics

Folding to the native state was initiated by a temperature jump. Initial high-temperature conformations are obtained by equilibrating the sequence at high $T_h \approx 10\,T_F$. The temperature was then quenched to $T_s < T_F$ and the kinetics of folding to the native state was monitored. The folding conditions were chosen in such a way that the thermal average $\langle\chi(T_s)\rangle = 0.25$. This gives $T_s = 0.94T_F$ or $0.95T_F$ for A and AGO, respectively. To calculate the folding time τ_F we computed the distribution of first passage times τ_{1i} from hundreds (≥600) of individual folding trajectories. The first passage time τ_{1i} is the first instance a given trajectory labeled i reaches the native conformation starting from initially unfolded state. From the distribution of τ_{1i} the fraction of unfolded molecules $P_u(t)$, which have not reached the native state at time t, may be readily calculated. The profiles of $P_u(t)$ can then be accurately fit with exponentials that provides a reliable method for computing τ_F using $\tau_F = \int_0^\infty P_u(t)dt$

For both sequences A and AGO $P_u(t)$ decays exponentially (Fig. 5) over a wide temperature range. The folding time at $T_s = 0.94T_F$ for A is $\tau_F = 2.07 \pm 0.07 \times 10^6$ MCS and folding is two-state, i.e., $P_u(t) \sim e^{-t/\tau_F}$ (Fig. 5). The Go version of A folds about three times faster ($\tau_F = 0.76 \pm 0.01 \times 10^6$ MCS) and also displays two-state kinetics. Thus, we conclude that non-native interactions in A retard the folding rates by approximately a factor of 3. Similar results for the ratio of folding times are obtained at other temperatures as well.

Multiple Folding Nuclei Model

To probe the nature of the folding nuclei and how their characteristics depend on non-native interactions we employed the approach used in our previous studies.[11] We describe only the major steps in our method. The conformations from the folding trajectories generated at the quench temperature $T_s(\langle\chi(T_s)\rangle = 0.25)$ were recorded over regular intervals. Each trajectory begins with a high-temperature initial conformation and ends at the first passage time, when the native state is reached. The stored conformations were then clustered using the cluster algorithm[11] that allows us to reduce the structural "noise" and describe the sampled conformations until the first passage time is reached. We focused our analysis on the propagation of stable native as well as non-native contacts. The stability of contacts is defined with certain tolerance to permit occasional short-lived disruptions.

We assume that the nucleation contacts are the minimal set of native contacts that (1) occur relatively close to the first passage time (i.e., at $\delta\tau_{1i}$, where $\delta \leqslant 1$), and (2) remain stable until τ_{1i}.[7,11] We also note that the choice of δ must satisfy the condition $\max[(1 - \delta)\tau_{1i}] < \tau_u$, where τ_u is the unfolding time. This implies that unfolding may not occur after crossing TS. Operationally, we search for the nucleation contacts (or TS) within the last 10–20% of the folding trajectory; i.e., $\delta = \delta_{\mathrm{TS}}$ is about 0.8–0.9. We have also varied δ_{TS} to ensure that the conclusions are not sensitive to its precise value.

The time evolution of individual trajectories shows that for both sequences the rate determining step involves a search for a set of contacts (critical nucleus). After the nucleus is formed, the native conformation is reached rapidly. For A and AGO, we find that this occurs for values of $\delta \gtrsim 0.8$ (see discussion below). For each folding trajectory we identify the minimal set of native contacts that are involved in the folding nucleus. This allows us to unambiguously obtain the set of residues that are critical for forming the native conformation. As shown in previous works,[11,26] no contact is found with unit probability in the folding nucleus.

The propagation of minimal set of stable native contacts for A (Fig. 6, top) and AGO (Fig. 6, bottom) shows striking heterogeneity in their formation. The local native contact $q = 4$ formed between residues 6 and 7 becomes stable very early in the folding process. As a rule, local (i.e., $|i - j| \leq 3$, where i and j are residue indexes) native contacts are formed much faster than their nonlocal counterparts. However, most native contacts remain unstable until late in the folding process, suggesting that transition states in our model occur closer to the native state. This observation is demonstrated more clearly in Figure 7. Figure 7 shows derivatives $dP/d\delta$ as a function of δ, where P is the probability of forming a native contact which is part of a minimal set of stable native contacts at a given δ averaged over all 20 native contacts. It is seen that $dP/d\delta$ remains negligibly small until late in the folding stage, but after about $\delta = 0.85$ it experiences an explosive growth. The value of δ at which $dP/d\delta$ begins the steep rise is identified with the folding TS region, and the minimal set of stable native contacts, formed at this δ, is identified with the folding nucleus. Thus, for A, we obtain that TS is crossed at $\delta_{\mathrm{TS}} \simeq 0.90$, while for AGO $\delta_{\mathrm{TS}} \simeq 0.85$. Qualitatively, the propagation of native interactions in A and AGO are very

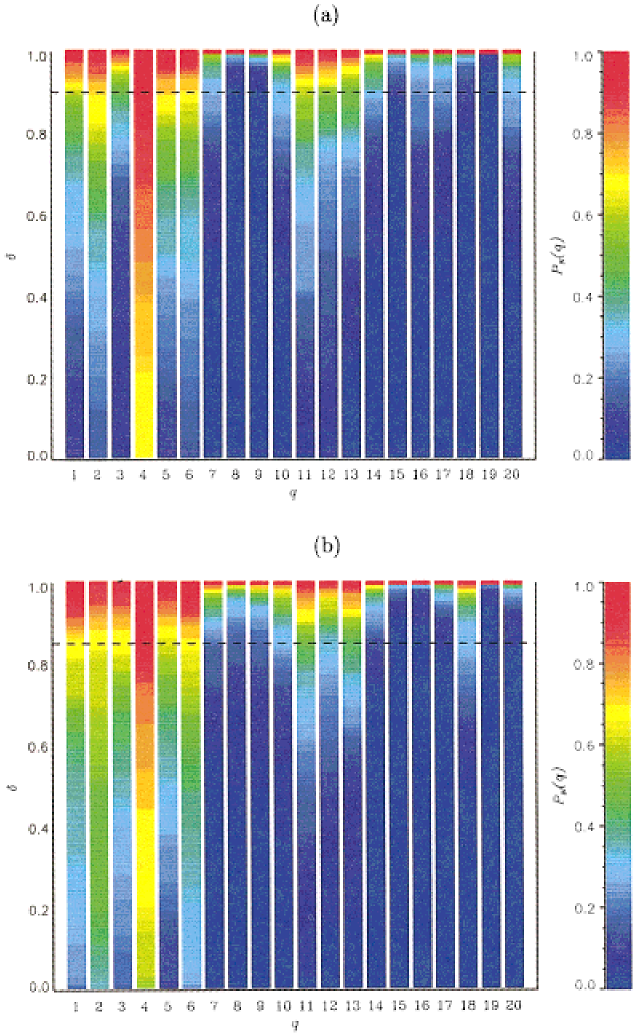

Fig. 6. Kinetic probabilities $P_N(q)$ as a function of δ averaged over 100 trajectories for sequences A (**a**) and AGO (**b**). $P_N(q)$ gives the probability that a native contact labeled q belongs to a minimal set of stable native contacts at a given δ. The color codes to $P_N(q)$ are given on the right side. The plots show a remarkable heterogeneity in the formation of native contacts. Most q become stable only in the vicinity of the native state; i.e., the TS occurs late in the folding process. The native contact indexes q are arranged in the ascending order of $|i - j|$, the distance between interacting residues i and j along the sequence. For the list of native contacts, see legend to Fig. 1.

similar, which is natural since native contacts being the driving force of folding are identical in both sequences. The differences are associated with the participation of non-native contacts in the folding process. The formation of native (mostly local) interactions in A is somewhat delayed as compared with AGO (cf. both panels in Fig. 6).

To get additional insights into formation of contacts in the TS we plot the profiles of native contacts probabilities $P_N(q)$ (where q, the native contact index, runs from 1 to 20 (see legend to Fig. 1) at several δ (including the TS value δ_{TS}) for both sequences (Fig. 8a,b). The profiles are obtained by averaging over 100 trajectories. It is clear that TS contains a mixture of local and nonlocal contacts (contact indexes 1 to 6, 10 to 14, and 16, 17, and 20 for A; for the list of native contacts, see caption to Fig. 1), among which local contacts clearly dominate. The TS for AGO features similar set of native interactions. More importantly, Figure 8c plots the kinetic probability of forming stable (over some interval) non-native contacts $P_{NN}(c)$, indicating that the TS for A contains several non-native interactions that disintegrate only close to the native state. These are the non-native contacts 8–11 (9), 9–12 (10), 11–14 (12), 6–11 (18), 1–10 (27), and 3–14 (35), using the format $i - j(c)$, where i and j are the residues involved, and c is the non-native contact index. Some of the listed non-native contacts are nonlocal and exceptionally stable (Fig. 8c).

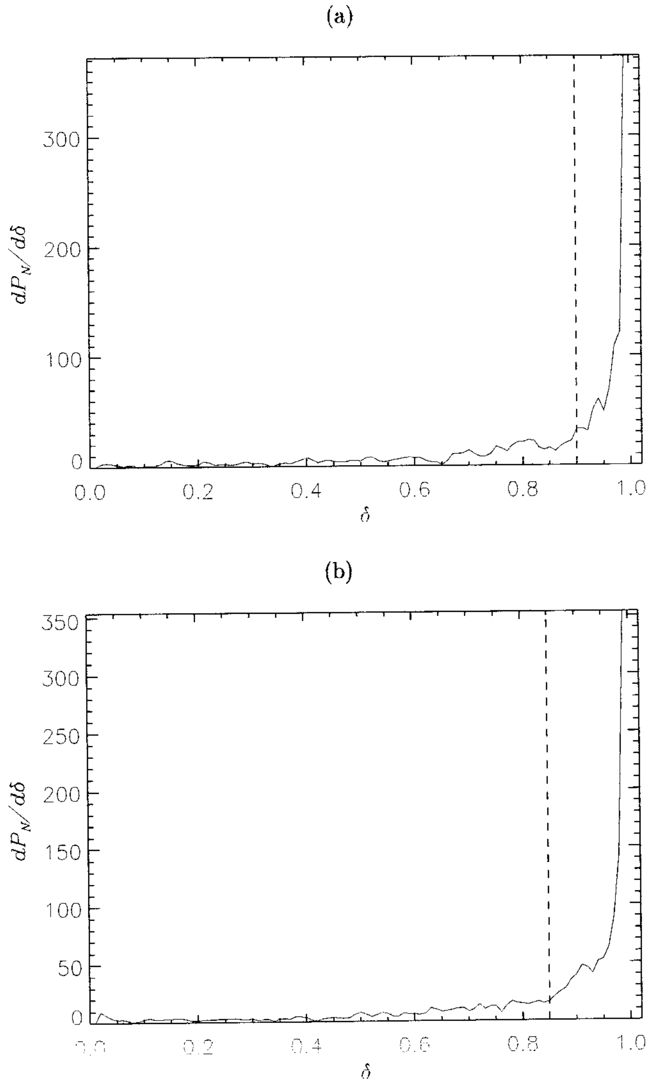

Fig. 7. Derivatives of kinetic probabilities P_N averaged over all native contacts with respect to δ as a function of δ for A (**a**) and AGO (**b**). Dashed vertical lines mark the values of δ_{TS} at which an explosive growth in $dP_N/d\delta$ is observed. We identify these regions with crossing the folding transition state (TS).

Structural Heterogeneity of Folding Nuclei

Experiments based on ϕ-value analysis[4] are interpreted in terms of the extent to which a particular residue is

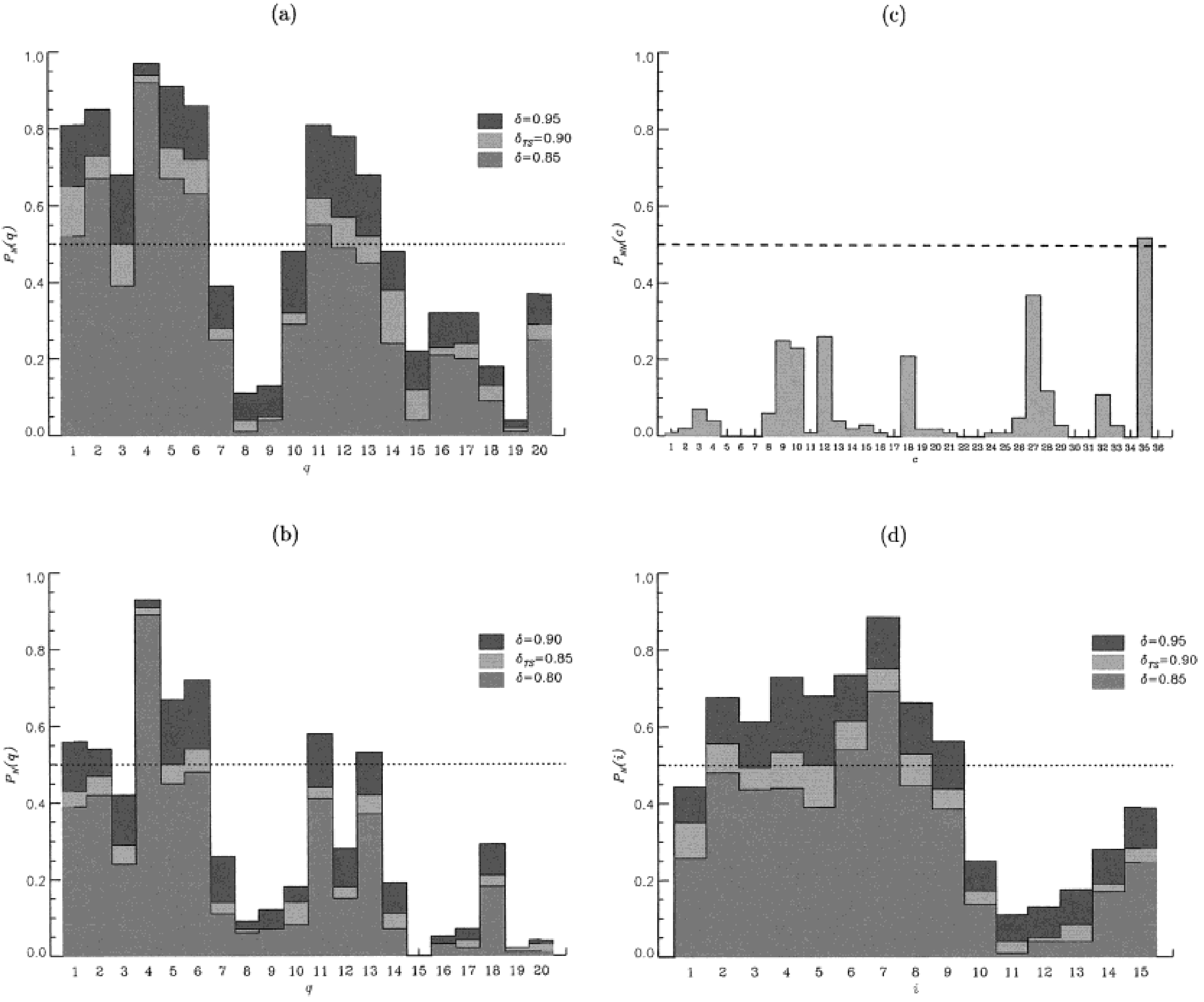

Fig. 8. Profiles of kinetic probabilities $P_N(q)$ for native contacts at three values of δ, where δ_{TS} (= 0.90 for A and 0.85 for AGO) is associated with crossing transition state (TS). **a:** Sequence A. **b:** Sequence AGO. Note that $P_N(q)$ are qualitatively similar at all δ proximal to δ_{TS}. Thus, the precise numeric value for δ_{TS} is not that important. The plots show that TS contain a mixture of local and non-local contacts. **c:** Profile of kinetic probabilities $P_{NN}(c)$ for non-native contacts for A at $\delta_{TS} = 0.90$; *c*, the non-native contact indexes, are arranged in ascending order of $|i - j|$ distance. In all, there 36 topologically possible non-native contacts. Surprisingly, several non-native contacts participate in folding TS with high probabilities and are kinetically important. **d:** Kinetic probabilities $P_N(i)$ for the residues i ($i = 1, \ldots, 15$) to form nucleation native contacts at three values of δ, including the TS value δ_{TS}, for A. $P_N(i)$, which estimates the degree to which a particular residue i is in the stable native conformation in TS, shows that the part of a sequence (i = 2–9) is more structured in the TS than the rest. Note that none of the residues completely forms native interactions in all folding trajectories.

structured in the TS. Figure 8d plots the probabilities for the residues to form nucleation native contacts in the TS $P_N(i)$ $(i = 1, \ldots, 15)$ for three values of δ, including δ_{TS} (sequence A). If there any residue were fully structured in the TS in all folding trajectories, the corresponding probability would be unity. We find that none of the nucleation residues has all its native contacts fully established in the TS; i.e., $P_N(i) < 1$ for all i (Fig. 8d). Because no residue is fully structured in all trajectories, no specific set of residues constitutes a folding nucleus. By contrast, from the results in Figure (8d) we conclude that residues form the native structure to a varying degree in the TS. For example, Figure 8d shows that residues 2–9 are more structured than 10–15; i.e., on average, the TS structures are "polarized."[27] This result clearly depends on the structure of the native state.

The analysis, summarized in Figures 6–8, rigorously shows that both sequences reach the native conformation by a nucleation-collapse mechanism with MFN. When averaged over hundreds of trajectories, we find that certain contacts (for structural and energetic reasons) appear in the folding nuclei with higher probabilities than the others. The high-probability ($\geqslant$0.5) contacts are determined largely by the structural restrictions of the native state. It is important to note that not a single contact participates with a unit probability in the TS, even when

$\delta = 0.95$. Consequently, no residue is fully structured in the TS. These observations are in agreement with earlier studies[11,13] and further affirm the MFN model for NC mechanism.

Identifying δ_{TS} With the TS Region

The reaction coordinate, a useful concept in rate theories, is difficult to define for multidimensional systems. Nevertheless, it is often found that the motions in slow degrees of freedom are confined to low dimensions. Because of the difficulty in identifying the nature of such low-dimensional attractors physical arguments are often used to obtain appropriate reaction coordinates. For lattice models of proteins, the fraction of native contacts, Q (a collective coordinate analogous to χ), has been suggested as a reaction coordinate.[28] For sequences A and AGO, we find that on an average certain contacts occur with high probability in the folding nuclei (Fig. 8a,b). If all the contacts were to participate with significant probabilities, a diffused delocalized nuclei would be an appropriate description of the TS structures. In this case, Q would be a natural choice for the reaction coordinate.[28]

In the generic case of small MFN, containing a mixture of local and nonlocal contacts (Fig. 8a,b), Q can also be a reasonable (but approximate) coordinate for the folding reaction $\mathbf{U} \rightleftharpoons \mathbf{N}$. In the absence of transparent physical arguments the choice for reaction coordinate becomes ambiguous.[29] In this work, δ is used as a surrogate progress variable to locate the TS region. There are at least two other ways of defining the TS structures without assuming any reaction coordinate. First, following Klosek et al.,[30] the TS region can be taken to be the locus of structures (stochastic separatrix) that have equal probability to lead either to **U** or **N** states. This idea was numerically implemented by Du et al.[31] to describe folding kinetics in lattice models without side-chains. Second, the locus of structures in the dividing surface separating **U** and **N**, across which the flux from **U** to **N** is maximum, can also be identified with the TS ensemble. This is the basic idea underlying the variational transition state theory, in which a dividing hypersurface is sought that minimizes the mean first passage time. Huo and Straub have used the maximum flux method to obtain the temperature-dependent optimal reaction pathway in the structural transitions in dipeptide.[32] If the basins of attraction for **U** and **N** are not symmetric with respect to the dividing surface, the two definitions of the TS ensemble may not be equivalent. In particular, the optimum paths (including passage over a barrier) in the maximum flux method[32] are independent of friction, while the stochastic separatrix method does depend on the friction along the reaction coordinate.[30]

The method for locating TS region employed in this study is tested against the stochastic separatrix method.[30,31] A reasonable definition of the folding transition state for symmetric free energy profiles at $T \approx T_F$ demands that, upon crossing TS, a protein subsequently folds to the native state or unfolds with approximately equal probability.[31] Therefore, if one initiates folding simulations with TS conformations, the fraction of molecules that finds the native state before unfolding (i.e., "forward" trajectories) shall be $p_{\text{fold}} \approx 0.5$. This requirement forms the computational basis for obtaining the TS structures.[31] The advantage of this method is that it does not make assumptions about the nature of the folding reaction coordinate, just as in our technique. Unlike our method, the obvious computational inefficiency of the p_{fold} technique severely limits its practical implementation.

Our goal is to calculate p_{fold} for the TS structures obtained at $\delta = \delta_{TS}$. Because the simulations for A (and AGO) are performed at $T \approx T_F$ and the free energy wells for these systems are expected to be roughly symmetric, p_{fold} should be approximately 0.5, if δ_{TS} indeed corresponds to the TS region. In all, we recorded 1769 TS conformations for A and subsequently selected 176 (every tenth) conformations for further analysis. Each conformation serves as a starting point for $M = 100$ individual folding trajectories generated at $T = T_s$, which either find the native state ($\chi = 0$) or lead to sequence unfolding ($\chi > \chi_u = 0.60$). The value of $\chi_u = 0.60$ is chosen such that it approximately corresponds to the boundary of the unfolded state in the free energy profile shown in Figure 3. Thus, for each TS conformation, we calculated the coefficient $p_{\text{fold}} = M_F/M$, where M_F is the number of trajectories, in which the folded state is found first. We then averaged p_{fold} over all sampled TS conformations. We found that for A p_{fold} at $\delta_{TS} = 0.9$ is 0.56. (Note that this results depends on the precise definition of the unfolded state so that changing χ_u to higher values would correspondingly increase p_{fold}. However, this change is very insignificant until $\chi_u < 0.7$.) Thus, the numerical equivalence of the methods based on clustering technique and the stochastic separatrix for obtaining TS provides solid evidence that the conformations located at the value of $\delta_{TS} = 0.90$ indeed correspond to the TS region. Similar results were obtained for AGO as well.

Transition State Ensemble

To analyze the TS conformations directly, we applied the clustering technique to the ensemble of structures found upon crossing the TS region at $\delta = \delta_{TS}$. We found three major TS clusters for both A and AGO. For the sake of brevity, Figure 9 shows TS structures for A alone. The most nativelike TS cluster (TS1) contains about $Q_{TS} = 18$ native contacts (90% of native interactions), has very low overlap $\chi_{TS} \simeq 0.2$ and potential energy, and is as compact as the average native state ($R_{g,\text{TS}}/R_{g,\text{N}} \simeq 1$). TS1 occurs in about 30% (A) (20% for AGO) of folding trajectories. Thus, it is not the most probable TS cluster. The cluster, which appears most frequently, is TS2 (found in 40% or 50% of folding trajectories for A and AGO, respectively). This TS has smaller average native content as measured by Q_{TS} (= 15 or 16) and χ_{TS}, which is about 0.5. More importantly, TS2 corresponds to the minimum free energy path, for which the folding barrier $\Delta F^{\ddagger}/T_s$ is the smallest with respect to other TS. This explains why TS2 appears most frequently in folding pathways and the folding time calculated for this pathway is minimal. TS3 appears less frequently than TS2 and has rather similar native content as TS2.

We conclude that the folding TS ensemble is heterogeneous in both sequences, consisting of three major TS clusters. This observation is consistent with the MFN model and reflects heterogeneous participation of contacts in the critical nuclei. As a result, we expect that optimized two-state sequences would be robust with respect to the mutations which affect even the residues participating with high probability in the TS. It is also worth emphasizing that because of the restriction that the critical nuclei should rapidly lead to the formation of native state, we do not expect that TS would exhibit extreme diversity.[7] This is clearly reflected in the TS ensemble analysis, in which we find that there are only three distinct structurally unrelated transition states[33] (Fig. 9). Generally, the TS structures are compact and nativelike, as measured by the number of formed native contacts, although the fraction of native interactions may vary. The local contacts are dominant in TS. Our TS analysis establishes a direct link between the free energy landscape and TS clusters. The most frequently occurring TS cluster (TS2) corresponds to the minimal free energy path, along which the folding barrier is the smallest. Consequently, the molecules that take the TS2 pathway reach the native state faster than along any other folding pathway (TS1 or TS3).

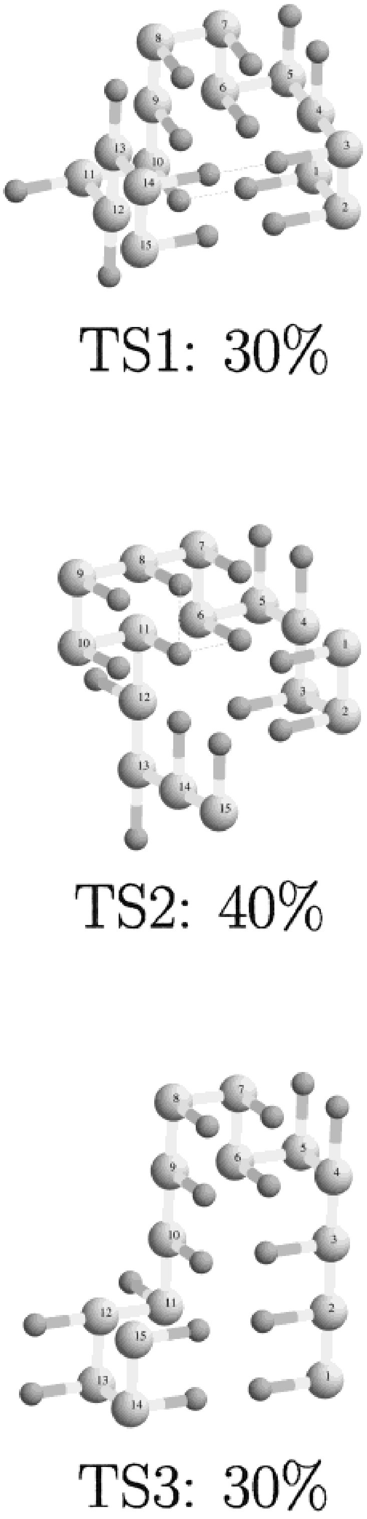

Fig. 9. Representative conformations for the transition state (TS) clusters obtained for sequence A. The conformation from the TS1 contains 18 native contacts and two non-native contacts (1–10 (27) and 3–14 (35), shown by dashed lines). The structures in TS1 are more compact and native-like than those in TS2 and TS3 clusters. TS2 conformation contains 13 native and 4 non-native contacts, two of which (8–11 (9) and 6–11 (18), shown by dashed lines) appear most frequently in the TS2. TS3 cluster is characterized by very low content of non-native contacts, therefore a typical conformation from this cluster contains only native contacts (16 in all). The percentage of folding trajectories that follow through these TS is indicated.

Role of Non-native Interactions in the TS Ensemble

The differences between TS for A and AGO lie in the degree to which non-native contacts participate in the TS1 and TS2 for A. Both TS clusters contain two non-native contacts. Li et al.[13] suggested that non-native contacts, which are present in the folding nucleus with high probability, are important kinetically, but not thermodynamically, because they apparently guide folding into the native basin of attraction. Disruption of these non-native interactions decreases the folding rates without compromising the native state stability.[13]

To assess the role of non-native contacts in the TS1 and TS2 in A, we created a mutant sequence AM, in which non-native contacts $c = 9, 18, 27$, and 35 (Fig. 9) are turned off (the contact interaction energies are set to zero). We calculated the folding time and found that $\tau_F = 1.44 \pm 0.04 \times 10^6$ MCS, which implies that elimination of these four non-native contacts speeds up folding by about 30%. Remarkably, the control mutant sequence AMR, in which four other randomly selected non-native contacts were turned off, folds within $2.21 \pm 0.01 \times 10^6$ MCS; i.e., no significant change with respect to A in folding rates is observed. Thus, we conclude that non-native contacts present in TS1 and TS2 are the "remnants" of non-native interactions that transiently stabilize intermediate (non-native) folding conformations. From the kinetic perspective, these contacts are destined to be rapidly disrupted because the native nucleation contacts are already formed at δ_{TS}. In line with this, calculation of the free energy profile (similar to those in Figure 3) for AM shows significant destabilization of the states with $\chi \sim 0.1$–0.25.

To assess the role of non-native interactions further, we ran simulations for three sequences that were generated by altering λ. As indicated above, Figure 4 shows that thermodynamics of folding remains, to a large extent, unaffected, even if λ varies in a broad interval ($-0.6 \lesssim \lambda \lesssim 0.1$). In this interval of λ, the stability of the sequences is essentially unchanged. This permits examination of the role non-native interactions play by decoupling stability and kinetics. To determine the impact of λ variation on folding kinetics, we calculated the folding times τ_F at $\lambda = -0.61$, -0.38, and 0.14. The respective results are $0.88 \pm 0.02 \times 10^6$, $0.79 \pm 0.03 \times 10^6$, $0.70 \pm 0.01 \times 10^6$ MCS. For the Go model ($\lambda = 0$) $\tau_F = 0.76 \pm 0.01 \times 10^6$ MCS. Furthermore, in all cases, the folding kinetics remains two-state. The computations show that two-state folders that have comparable stabilities fold at nearly the same rates at temperatures close to T_F. This simple exercise also suggests that sequence mutations described by $-0.6 \lesssim \lambda \lesssim 0.1$ are, in fact, neutral as these produce negligible effect on folding thermodynamic and kinetic characteristics. We conclude that the overall effect of non-native interactions on folding is disruptive, because their complete or partial elimination speeds up folding and enhances cooperativity. However, our calculations also indicate a strong heterogeneity in the participation of non-native contacts in folding. Most individual non-native contacts have very little impact on folding, whereas few of them (as those in the TS1 and TS2) do slow down folding considerably. In this sense, non-native contacts in TS1 and TS2 may be considered kinetically important.

CONCLUSIONS

We have considered the folding thermodynamics and kinetics of models of two-state proteins, using lattice representation with side-chains that we introduced a few years ago.[16] The advantage of such models is that issues, such as the nature of the folding nuclei and the associated TS ensemble, can be directly and precisely addressed, provided exhaustive simulations are undertaken. Using an optimized sequence and its Go version (in which all the non-native interactions are set to zero), we arrived at the following conclusions:

1. We showed rigorously, analyzing hundreds of folding trajectories, that folding occurs by a nucleation-collapse mechanism, in which there are multiple folding nuclei. This implies that the transition state is, in general, heterogeneous.[12] To quantify the degree of heterogeneity we determined the structures of the distinct transition states, which correspond to the average conformations in the critical nuclei. For both sequences, which have identical native states, there are only three structurally unrelated transition states. TS structures occur late in the folding process. This study confirms our previous suggestion[7,11] that the number of distinct folding nuclei in these protein-like models is small but is greater than unity.
2. A direct comparison between sequence A and the Go version (sequence AGO) affords critical examination of the role of non-native interactions in the thermodynamics and kinetics of two-state folders. Although the characteristic temperatures are unaffected by elimination of non-native interactions, we find that the stability and the degree of cooperativity are greatly enhanced. However, the qualitative features of the folding kinetics for both systems are unchanged. From a structural point of view, it appears that Go model (by definition) cannot account for non-native interactions in the transition state structures. This is the major qualitative kinetic difference between the two models.
3. Li et al.[13] recently used a similar model to study the role of non-native interactions in folding kinetics. They showed that the most probable TS (a folding nucleus) has a small fraction of non-native interactions. Our study is in agreement with this conclusion. For their sequence it appears that softening the non-native interactions in the TS leads to a decrease in the folding rates. This prompted them to conclude that optimizing certain non-native interactions may be necessary for enhancing the refolding rates. By contrast, we find that the folding rate increases upon eliminating or destabilizing the non-native interactions. Nevertheless, both studies establish the kinetic importance of non-native contacts. We can surmise that the role of the non-native interactions in refolding may be complex and will depend on the precise sequence and structure of the native state.
4. We find that elimination of non-native interactions invariably leads to faster folding. Shea et al.[19] showed that by systematically attenuating non-native interactions T_F can be made to approach T_θ. Since the foldability index $\sigma = (T_\theta - T_F)/T_\theta$ is good indicator of the folding rates, our conclusions are in accord with Shea et al. The free energy profiles $F(\chi)$ (Fig. 3) also become less rugged as λ decreases, suggesting that for optimized sequences (in which energetic and topological frustration is minimized) a suitable one-dimensional reaction coordinate may be found to describe qualitatively the refolding kinetics.[15] Identification of low dimensional reaction coordinate for non-optimized sequences may be difficult, if not impossible.
5. A byproduct of this work is the demonstration that the clustering technique that was used in our previous works is a very efficient tool for locating the TS structures in protein folding simulations. Our method allows one to obtain the TS structures directly from the folding trajectories without using any underlying assumptions about the nature of the reaction coordinate. The computational efficiency of this approach will enable a structural description of TS provided an adequate number of folding trajectories is generated.

ACKNOWLEDGMENTS

The authors are grateful to John E. Straub and Robert Zwanzig for discussions on finding TS in multidimensional systems.

REFERENCES

1. Dill KA, Chan HS. From Levinthal to pathways to funnels. Nature Struct Biol 1997;4:10–19.
2. Onuchic JN, Luthey-Schulten ZA, Wolynes PG. Theory of protein folding: an energy landscape perspective. Annu Rev Phys Chem 1997;48:545–600.
3. Thirumalai D, Klimov DK. Deciphering the time scales and mechanisms of protein folding using minimal off-lattice models. Curr Opin Struct Biol 1999;9:197–207.
4. Fersht A. Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. New York: WH Freeman; 1999.
5. Wolynes PG. Folding funnels and energy landscapes of larger proteins within the capillarity approximation. Proc Natl Acad Sci USA 1997;94:6170–6175.
6. Guo Z, Thirumalai D. Kinetics of protein folding: nucleation mechanism, time scales, and pathways. Biopolymers 1995;36:83–103.
7. Guo Z, Thirumalai D. The nucleation-collapse mechanism in protein folding: evidence for the non-uniqueness of the folding nucleus. Folding Design 1997;2:377–391.
8. Bryngelson JD, Wolynes PG. A simple statistical field theory of heteropolymer collapse with application to protein folding. Biopolymers 1990;30:177–188.
9. Fersht AR. Nucleation mechanism of protein folding. Curr Opin Struct Biol 1997;7:10–14.
10. Shakhnovich E, Abkevich VI, Ptitsyn O. Conserved residues and the mechanisms of protein folding. Nature 1996;379:96–98.
11. Klimov DK, Thirumalai D. Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. J Mol Biol 1998;282:471–492.
12. Onuchic JN, Socci ND, Luthey-Schulten Z, Wolynes PG. Protein folding funnels: the nature of the transition state ensemble. Folding Design 1996;1:441–450.
13. Li L, Mirny LA, Shakhnovich EI. Kinetics, thermodynamics and evolution of non-native interactions in protein folding nucleus. Nature Struct Biol 2000;7:336–342.
14. Go N. Theoretical studies of protein folding. Annu Rev Biophys Bioeng 1983;12:183–210.
15. Clementi C, Jennings PA, Onuchic JN. How native-state topology affects the folding of dihydrofolate reductase and interleukin-1. Proc Natl Acad Sci USA 2000;97:5871–5876.
16. Klimov DK, Thirumalai D. Cooperativity in protein folding: from lattice models with side chains to real proteins. Folding Design 1998;3:127–139.
17. Kolinski A, Godzik A, Skolnick J. A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: application to designed helical proteins. J Chem Phys 1993;98:7420–7433.
18. Ferrenberg AM, Swendsen RH. Optimized Monte Carlo data analysis. Phys Rev Lett 1989;63:1195–1198.
19. Shea JE, Nochomovitz YD, Guo ZY, Brooks CL III. Exploring the space of protein folding hamiltonians: the balance of forces in a minimalist β-barrel model. J Chem Phys 1998;96:12512–12517.
20. Sayle R, Milner-White EJ. RasMol: biomolecular graphics for all. Trends Biochem Sci 1995;20:374.
21. Klimov DK, Thirumalai D. Linking rates of folding in lattice models of proteins with underlying thermodynamic characteristics. J Chem Phys 1998;109:4119–4125.
22. Thirumalai D, Klimov DK, Betancourt MR. Exploring the folding mechanisms of proteins using lattice models. In: Grassberger P, Barkema GT, Nadler W, editors. Monte Carlo approach to biopolymers and protein folding. Singapore: World Scientific; 1998. p 19–28.
23. Camacho CJ, Thirumalai D. Kinetics and thermodynamics of folding in model proteins. Proc Natl Acad Sci USA 1993;90:6369–6372.
24. Klimov DK, Thirumalai D. A criterion that determines the foldability of proteins. Phys Rev Lett 1996;76:4070–4073.
25. Bussemaker HJ, Thirumalai D, Bhattacharjee JK. Thermodynamic stability of folded proteins against mutations. Phys Rev Lett 1997;79:3530–3533.
26. Thirumalai D, Klimov DK. Fishing for folding nuclei in lattice models and proteins. Folding Design 1998;3:R112–R118.
27. Grancharova VP, Riddle DS, Santiago JV, Baker D. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. Nature Struct Biol 1998;5:714–720.
28. Socci ND, Onuchic JN, Wolynes PG. Diffusive dynamics of the reaction coordinate for protein folding funnels. J Chem Phys 1996;104:5860–5868.
29. Zwanzig R. Two-state models of protein folding kinetics. Proc Natl Acad Sci USA 1997;94:148–150.
30. Klosek MM, Matkowsky BJ, Schuss Z. The Kramers problem in the turnover regime: the role of the stochastic separatrix. Ber Bunsenges Phys Chem 1991;95:331–337.
31. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich EI. On the transition coordinate for protein folding. J Chem Phys 1998;108:334–350.
32. Huo S, Straub JE. The MaxFlux algorithm for calculating variationally optimized reaction paths for conformational transitions in many body systems at finite temperature. J Chem Phys 1997;107:5000–5006.
33. Pande VS, Grosberg AY, Tanaka T, Rokhsar DS. Pathways for protein folding: is a new view needed? Curr Opin Struct Biol 1998;8:68–79.