Published in final edited form as:
J Phys Chem B. 2013 October 24; 117(42): . doi:10.1021/jp401962k.

How Kinetics within the Unfolded State Affects Protein Folding: an Analysis Based on Markov State Models and an Ultra-Long MD Trajectory

Nan-jie Deng, **Wei Dai**, and **Ronald M. Levy**
BioMaPS Institute for Quantitative Biology and Department of Chemistry and Chemical Biology, Rutgers, the State University of New Jersey, Piscataway, NJ 08854, Telephone: 732-445-3947

Ronald M. Levy: ronlevy@lutece.rutgers.edu

Abstract

Understanding how kinetics in the unfolded state affects protein folding is a fundamentally important yet less well-understood issue. Here we employ three different models to analyze the unfolded landscape and folding kinetics of the miniprotein Trp-cage. The first is a 208 μs explicit solvent molecular dynamics (MD) simulation from D. E. Shaw Research containing tens of folding events. The second is a Markov state model (MSM-MD) constructed from the same ultra-long MD simulation; MSM-MD can be used to generate thousands of folding events. The third is a Markov state model built from temperature replica exchange MD simulations in implicit solvent (MSM-REMD). All the models exhibit multiple folding pathways, and there is a good correspondence between the folding pathways from direct MD and those computed from the MSMs. The unfolded populations interconvert rapidly between extended and collapsed conformations on time scales ≤ 40 ns, compared with the folding time of ≈ 5 μs. The folding rates are independent of where the folding is initiated from within the unfolded ensemble. About 90 % of the unfolded states are sampled within the first 40 μs of the ultra-long MD trajectory, which on average explores ~27 % of the unfolded state ensemble between consecutive folding events. We clustered the folding pathways according to structural similarity into “tubes”, and kinetically partitioned the unfolded state into populations that fold along different tubes. From our analysis of the simulations and a simple kinetic model, we find that when the mixing within the unfolded state is comparable to or faster than folding, the folding waiting times for all the folding tubes are similar and the folding kinetics is essentially single exponential despite the presence of heterogeneous folding paths with non-uniform barriers. When the mixing is much slower than folding, different unfolded populations fold independently leading to non-exponential kinetics. A kinetic partition of the Trp-cage unfolded state is constructed which reveals that different unfolded populations have almost the same probability to fold along any of the multiple folding paths. We are investigating whether the results for the kinetics in the unfolded state of the twenty-residue Trp-cage is representative of larger single domain proteins.

Keywords

Protein Folding; Unfolded State Kinetics

Correspondence to: Ronald M. Levy, ronlevy@lutece.rutgers.edu.

Supporting Information Available The Supporting Information contains figures that illustrate the results of the kinetics of the unfolded state and its effects on protein folding. This information is available free of charge via the Internet at http://pubs.acs.org.

Introduction

Although much progress has been made on the protein folding problem, unresolved questions still exist concerning some of the fundamental aspects of how proteins fold[1–15]. For example, how does the energy landscape of the unfolded state affect folding[3,5,16,17]? Does residual structure within the unfolded ensemble influence folding rates? Why do some proteins which theory and simulation suggest have multiple folding pathways exhibit two-state, single exponential kinetics? Molecular dynamics simulations (MD) in atomic detail provide the spatial and temporal resolution required to investigate the mechanisms of protein folding in aqueous solutions. However, the time scale covered by MD is usually too short for direct unbiased folding simulations. In recent years, the D. E. Shaw lab has developed a special-purpose computer that greatly accelerates MD simulations of biomolecules, and the gap between direct simulation and biological time scales is now beginning to be closed[18–21]. Using their ANTON technology on twelve structurally diverse fast-folding proteins, Shaw and coworkers were able to fold eleven of them to experimental structures and observe numerous reversible folding transitions in simulations ranging from microseconds to milliseconds [20].

Other methods which do not require special purpose hardware are being developed to overcome the time scale limitation of direct MD and more efficiently sample the rare events associated with biomolecular transitions [22–26]. In this area, Markov state models (MSM) constructed from atomistic simulations have been particularly successful in sampling the rare events associated with protein folding and protein conformational transitions [27–38]. In this approach, the protein conformational space is discretized into a network of coarse grained substates. Transitions on the network are modeled by a master equation; the kinetics on the network is Markovian. The network approach provides an efficient way to extract mechanistic insights from a large amount of MD trajectory data without losing the accuracy of the underlying atomistic simulations. The folding pathways and their fluxes can be obtained by applying transition path theory (TPT) on the network [30,31,39], yielding a statistical description of how a protein acquires the specific native conformation starting from an extremely large number of possibilities. Noe et al. studied the folding of PinWW domain by constructing a Markov state network model from many relatively short MD simulations.[31] They found many non-overlapping pathways passing through intermediate regions to reach the native state. Based on their Markov state modeling of small single domain proteins, Pande and coworkers proposed that protein native states act as kinetic hubs connected to unfolded structures by stochastic jumps through metastable states [34,36]. In this kinetic hub model, the unfolded state ensemble is divided into collections of states that fold along different folding paths; to get from states which fold along one path to those which fold along another involves transiting through the folded state [34].

To overcome the sampling limitations of constant temperature MD in constructing a network model, over the past several years our group has developed an approach that takes advantage of replica exchange molecular dynamics (REMD) in accelerating barrier crossing, and extracts kinetic information from REMD by assuming that a network of transitions can be reconstructed by applying structural similarity criteria together with reweighting techniques [28,40,41]. By exploiting transition path theory together with stochastic simulations, the kinetic network can be interrogated and information concerning the temperature dependent folding pathways can be obtained. Application of this approach to the miniprotein Trp-cage indicated that below the folding temperature, the folding flux is dominated by a small number of localized pathways [40]. Above the folding temperature, the folding pathway ensemble becomes much more diverse.

The effect of the unfolded state heterogeneity on folding was the focus of an insightful study by Ellison and Cavagnero[16]. One important finding from their simple kinetic model is that for proteins with heterogeneous folding pathways, deviations from single exponential are observed only when unfolded conformations exchange at rates slower than folding. This result may provide a simple explanation for the apparent two-state, single exponential kinetics shown by some proteins, even though these proteins may fold through multiple diverse pathways.

In the present study, we employ stochastic simulations, transition path theory and Markov state models constructed from atomistic simulations to investigate how the kinetics within the unfolded state ensemble affects folding. The Trp-cage miniprotein (Fig. 1) has served as a model system for studying folding in numerous experimental [42–45] and theoretical studies [46–52]. Here we investigate the kinetics in the unfolded state and its effects on folding using the following models: (1) a 208 μs explicit solvent MD simulation from the Shaw lab[20] that contains several folding events; (2) a Markov state model constructed from the same ultra-long MD trajectory (MSM-MD); and (3) a kinetic network model built from REMD simulations using an implicit solvent effective potential over a wide temperature range (MSM-REMD). Direct comparison between the ultra-long MD and MSM-MD trajectories serves to test the validity of the Markov model. Pande and coworkers have reported the first such comparison using two 100 μs folding trajectories of FIP35 WW domain [53]. They found that the MSM has a hub-like topology and the analysis yielded more insights into the diversity of folding pathways and dynamics between two alternative native structures. Here our emphasis is on sampling within the unfolded state and its effects on folding pathways. Because stochastic simulations on a discretized network are extremely efficient, we use it in the present study to extensively explore the kinetics within the unfolded ensemble. We have developed techniques to map the reactive stochastic trajectories onto the folding pathways computed using TPT [40]. By combining stochastic simulations with TPT pathway analysis we can evaluate the folding rate along each pathway and the probability of folding along any pathway from any place within the unfolded state ensemble. By analyzing the Trp-cage kinetics in the light of a simple kinetic model calculation we determine a general relationship between the folding kinetics and the rate of mixing in the unfolded states. We discuss our network model analysis in relation to the study of Ellison and Cavagnero [16] and other folding models [5,34,54]. The main result of the present study is that proteins with heterogeneous pathways will fold with single exponential kinetics, as long as the rate of mixing within the unfolded state is comparable to or faster than folding. While the mixing within the unfolded state modulates the apparent waiting times for folding along individual paths, the overall folding rate depends only on the total folding flux and the equilibrium unfolded state population.

Methods

Analysis of the Ultra Long MD Trajectory

A MD simulation of Trp-cage was performed by Shaw and coworkers on the Anton computer for 208 μs using a modified charmm22 all-atom force field in the TIP3P explicit solvent [20]. The MD trajectory contains ~10^6 snapshots saved every 200 ps. During the course of the simulation the Trp-cage fluctuates between the low and high rmsd regions, via transiently occupied intermediate region (Fig. S1a). The distribution of rmsd is bimodal, containing a sharp peak at rmsd = 1.3 Å, which is separated from a broad peak at rmsd = 6 Å by a weakly populated intermediate region (Fig. S1b). Based on the rmsd distribution, we define three macrostates as follows: folded (rmsd ≤ 2.2 Å), intermediate (2.2 Å < rmsd ≤ 5 Å), and unfolded states (rmsd > 5 Å); the populations carried by these macrostates are: 17.5 % folded, 15.1 % intermediate, and 67.4 % unfolded. Note that the definition of the three macrostates is somewhat different from that used in the study by Shaw et al.[20], where the

folded and unfolded states are defined based on the fraction of native contacts, Q. As a result, there are some quantitative differences between the kinetic properties calculated in this work and those found by Shaw and coworkers. In particular, using the RMSD-based cutoff scheme, the trajectory contains a number of rapid folding transitions that are not considered as folding events in the study of Shaw and coworkers which used Q-based definition of the macrostates. This does not affect the interpretation of the main results except for the reported value of the folding transit time as discussed in the following section.

Construction of MSM-MD

We constructed a Markov state model based on the 208 μs MD simulation to analyze the Trp-cage folding kinetics. The MSM-MD consists of a collection of conformational microstates and the transition probability matrix describing the memoryless jumps among these microstates. A set of 25000 microstates is generated by geometrically clustering the ~10^6 MD snapshots according to their mutual rmsd using the k-means clustering method. The average rmsd between a structure and its cluster center is 2.45 Å. The transition matrix $T_{ij}(\tau)$ is estimated by projecting the MD trajectory onto the network nodes and counting the number of transitions from node i to node j within lag time τ, i.e. $T_{ij}(\tau) \equiv P(j,\tau|i,0) = C_{i\to j}(\tau) / \sum_k C_{i\to k}(\tau)$. To choose a lag time for which the transitions on the network are Markovian, we used a criteria based on the mean first passage time (MFPT) of folding: when the transitions are Markovian, the folding MFPT computed using $T_{ij}(\tau)$ should not depend on the choice of lag time τ: see Fig. S2a. Here folding MFPT is obtained from the inverse of the folding rate $k_f = \frac{J}{\sum_i P_{eq}(i)[1-P_{fold}(i)]}$[31], where $P_{fold}(i)$ is the commitor probability of folding; J is the folding flux computed using TPT[31], $J = \frac{\sum_{i\in U, j\notin U} T_{ij}(\tau) P_{eq}(i) P_{fold}(j)}{\tau}$. The calculated folding MFPT at different lag times are shown in Fig. S2a. At τ ≥ 5 ns, the folding time for the 25000-node MSM begins to level off to a plateau value close to the MFPT observed in the ultra-long MD simulation, suggesting that the model behaves Markovian for lag times t ≥ 5 ns. We also tested a coarser model with 6000 states: although the MFPT shows similar curvature, the plateau value of MFPT is considerably smaller than the folding time observed from the MD trajectory.

To further verify the Markovian property of the network, we used another criteria based on the implied timescales t_i, calculated from the eigenvalues $\lambda_i(\tau)$ of T'(τ), $t_i = -\tau / \ln[\lambda_i(\tau)]$.[27] Here T′(τ) is identical to the transition matrix T(τ) except that all the rates leaving the folded state are set to zero, which corresponds to the absorbing boundary condition for folding. The use of T'(t) allows its slowest eigenvalue to reproduce the MFPT of folding computed from TPT. As seen from Fig. S2b, at τ ≈ 5ns, the implied timescales begin to level off. The implied timescale computed for the slowest decaying eigenmode of the T'(τ) at τ = 5ns is 5.5 μs, which is in excellent agreement with the MFPT obtained using the flux from TPT calculation.

The choice of lag time τ not only determines the Markovian behavior of the MSM, but also strongly affects the kinetic resolution of the network model. At small lag time, $T_{ij}(\tau)$ is equivalent to the rate matrix in continuous-time Markov model via $k_{ij} = \lim_{\tau\to 0} \frac{T_{ij}(\tau)}{\tau}$, which gives the highest possible kinetic resolution on the network. At large τ, many of the unfolded states are connected to the folded state in one jump. At τ = 1 ns, 5 ns and 20 ns, the one-jump folding pathways were found to account for 2.7 %, 10 % and 28 % of the total folding flux, respectively. In contrast, in the 208 μs MD simulation, one-jump folding event

was observed only once out of the 31 folding transitions, which corresponds to 3.2 % folding flux. Therefore, our results show that, while the lag time needs to be long enough to satisfy Markov property for memoryless transition, τ should also be small enough to allow sufficient kinetic resolution for studying folding mechanism. Likewise, the radius of the clustered nodes need to be large enough to have adequate statistics; but too much coarse graining could lead to non-Markovian behavior by grouping structures separated by significant barriers. The MSM-MD used in the present study is based on 25000-state clustering and a lag time of 5 ns, which we found give a good tradeoff between satisfying Markov property and providing adequate kinetic resolution and statistics.

Construction of MSM-REMD from Replica Exchange Simulation

We also constructed a kinetic network model for Trp-cage (MSM-REMD) by clustering the 150,000 snapshots, obtained from REMD simulations at temperatures from 363 K to 566 K, into a set of 20,000 conformational microstates. The details about the REMD simulations are described in the reference 34 [40]. The clustering is performed based on the Cα-RMSD between pair of the snapshots, using a cutoff radius of 1.1 Å. All the neighboring conformations found within the cutoff RMSD from a selected central node are merged to create a composite node. The resulting clustered nodes generally consist of contributions from many REMD snapshots observed at several temperatures.

The rates for the memoryless transitions on the network were parameterized using a scheme involving many short MD simulations. The rate constant k_{ij} for the transition from state j to state i is

$$k_{ij} = C_{ij}\left(\frac{P_{i,eq}}{P_{j,eq}}\right)^{\frac{1}{2}} \quad (1)$$

Here the prefactor $C_{ij} = C_{ji}$, which satisfies the detailed balance $k_{ij}P_{j,eq} = k_{ji}P_{i,eq}$. By definition, the rate k_{ij} can be expressed in erms of the branching probability $P_{j\to i}$ and the mean lifetime at node j, $<T_j>$:

$$k_{ij} = \frac{P_{j\to i}}{<T_j>} \quad (2)$$

Eq. 2 suggests a way to parameterize k_{ij} based on the lifetime observed from many short MD trajectories. The branching probability $P_{j\to i}$ can be approximately expressed in terms of the RMSD distance between node i and j, Δr_{ij}. From running many short MD trajectories, we found that the average probabilities of jumping to a neighboring node at Δr can be fitted with

$$P_{j\to i}(\Delta r_{ij}) \propto \frac{\Delta r_{ij}^{-6}}{<\Delta r_{ij}^{-6}>_j} \quad (3)$$

(Fig. S3). Additionally, $P_{j\to i}$ decreases approximately with the number of neighbors of node j, i.e. $P_{j\to i}(\Delta r_{ij}) \propto \frac{1}{N_j^{nb}}$. From Eq. 2 and Eq. 3, the rate constant k_{ij} is expressed as

$$k_{ij} \approx \frac{\Delta r_{ij}^{-6}}{<\Delta r_{ij}^{-6}>_j N_j^{nb} T_{j,MD}}\left(\frac{P_{i,eq}}{P_{j,eq}}\right)^{\frac{1}{2}} \quad (4)$$

Here the prefactor is identified with C_{ij} in Eq. 1, i.e. $C_{ij} \approx \frac{\Delta r_{ij}^{-6}}{<\Delta r_{ij}^{-6}>_j N_j^{nb} T_{j,MD}}$. Since $C_{ij} = C_{ji}$ (needed for maintain detailed balance), we symmetrize C_{ij} and write $C_{ij} \approx \frac{1}{2}\left(\frac{\Delta r_{ij}^{-6}}{<\Delta r_{ij}^{-6}>_j N_j^{nb} T_{j,MD}} + \frac{\Delta r_{ij}^{-6}}{<\Delta r_{ij}^{-6}>_i N_i^{nb} T_{i,MD}}\right)$. Taken these considerations together, we obtain

$$k_{ij} \approx \frac{1}{2}\Delta r_{ij}^{-6}\left(\frac{1}{<\Delta r_{ij}^{-6}>_j N_j^{nb} T_{j,MD}} + \frac{1}{<\Delta r_{ij}^{-6}>_i N_i^{nb} T_{i,MD}}\right)\left(\frac{P_{i,eq}}{P_{j,eq}}\right)^{\frac{1}{2}} \quad (5)$$

To test how well the rates parameterized using Eq. 5 describe the kinetics, we compare the distributions of state lifetimes obtained from many short MD simulations and those from stochastic simulations on the MSM-REMD. The results show that the two distributions of the state lifetimes agree well with each other (Fig. S4).

The procedures of the decomposition of the folding flux into folding pathways, the clustering of folding pathways into folding tubes, and the mapping of stochastic simulation trajectories onto folding tubes were described in a previous paper[41].

Results

Below we first present the results of sampling the Trp-cage unfolded state by MD and the MSM built from MD (MSM-MD), including the time scales of structural reorganization in the unfolded state, the kinetic partitioning of the unfolded state into populations that fold along different paths and the folding rates associated with different folding paths. We then analyze the distribution of the folding passage times, transit times, and the nature of the heterogeneity in the folding pathways. Finally, we discuss the results for the MSM constructed from REMD sampling (MSM-REMD) to investigate how the folding kinetics is influenced as a function of temperature and for comparison with the ultra-long MD trajectory from the Shaw group.

Sampling of the Unfolded States by MD and MSM-MD

We first examine to what extent the sampling of the unfolded states has converged in the 208 μs MD trajectory, by estimating the fraction of the unfolded conformational space sampled by the trajectory as a function of simulation time. To this end we cluster the trajectory; the clustering scheme we employed is described in the Methods. We calculated the fraction of the unfolded state clusters visited by the trajectory as a function of simulation time and found that about 90 % of the unfolded states are sampled within the first 40 μs, which is one-fifth of the total simulation time (Fig. S5, Supporting Information). The trajectory spends the remaining 80 % of the simulation time mostly revisiting the structures seen earlier. This result is reasonably robust with respect to the variation in the granularity of the clustering (see Fig. S5). It is therefore an indication that the ultra-long MD simulation exhibits good convergence in the sampling of the unfolded state ensemble.[55]

We characterized the structural reorganization in the unfolded state to address (1) how heterogeneous are the structures explored between two adjacent unfolding/folding events? (2) What is the time scale for chain extension and collapse for unfolded Trp-cage before it folds? We choose the radius of gyration (R_g) as the order parameter to characterize the structural reorganization in the unfolded region. Fig. 2 shows the distribution of R_g and its fluctuations in the MD trajectory. The folded structure has an R_g of ~7 Å; for the unfolded

state R_g spans broadly the range from 6.5 Å to 15 Å, which correspond respectively to compact unfolded conformations and fully extended chains, two examples of which are shown in Fig. S6. As seen from Fig. 2b, the MD trajectory visits both extended conformations ($R_g \geq 14$ Å) and compact unfolded structures ($R_g \leq 8$ Å) many times before it folds. We computed the distribution $\Gamma(\tau)$ of relaxation times τ for the radius of gyration in the unfolded state (see Fig. S7) and found a dominant relaxation mode at $\tau = 6$ ns along with a much weaker mode centered at $\tau = 38$ ns. The relaxation times for the fluctuation between the extended and collapsed forms of Trp-cage are much shorter than the average residence time of (~5 μs) in the unfolded state between adjacent folding events. We also examined the time scale of collective motions in the unfolded state by computing the autocorrelation functions for the principal components. The relaxation time along the slowest principal component is found to be ≤ 40 ns, i.e. similar to the time scale of fluctuations in R_g (Fig. S8).

In order to gain further insight into the kinetic properties of the unfolded state ensemble, we analyze the fraction of the total conformational space of the unfolded ensemble visited by the MD trajectory, between two adjacent unfolding/folding transitions. We compute this quantity by analyzing unfolded intervals between each consecutive unfolding/folding event. Here, an unfolded interval starts from the time when the trajectory enters the unfolded region and ends when the trajectory enters the folded state. Fig. 3 shows the fraction of unfolded conformations visited during each of the unfolded intervals before the trajectory folds. In 45 % of the folding events, the trajectory visits > 30 % of the unfolded states before it folds. On average, a trajectory typically explores about 27 % of the unfolded conformational space between consecutive unfolding/folding events. Fig. 3 also shows that the fraction of unfolded states visited is strongly correlated with the folding passage time. This correlation is an indication of substantial mixing within the unfolded state ensemble, as discussed below.

Using the ultra-long MD trajectory we constructed a 25,000-state kinetic network model (MSM-MD): see Methods. We performed a kinetic Monte Carlo (KMC) simulation for 64 milliseconds, which contains ~10,000 folding and unfolding events. The radius of gyration time series are nearly indistinguishable from those observed in the direct molecular dynamics simulation MD (Fig. 2b; see also Table 1).

We computed the folding pathways and their fluxes using transition path theory (TPT) [31,39] to analyze the MSM-MD network model; ~5000 pathways were generated. To obtain mechanistic insights, the pathways were clustered into a much smaller number of folding tubes (~100), each containing between 10–100 structurally similar pathways. The grouping of folding pathways into tubes is based on structural similarity between the structures along two pathways [40,41]; the average RMSD distance between two pathways in different tubes is at least 4 Å.

Kinetic Partition of the Unfolded State by Folding Tubes

We now discuss the results on the kinetic partitioning of the unfolded state by folding tubes and the characterization of the unfolded populations and folding rates associated with different folding tubes. By projecting the stochastic MSM-MD trajectories onto the different folding tubes, we determined three important kinetic quantities associated with each folding tube: $J(\alpha)$, the flux through tube α; $k(\alpha)$, the folding rate corresponding to the tube; and $P(\alpha)$, the fraction of the unfolded population that folds through tube α. The flux $J(\alpha)$ is defined by the number of folding events through tube α per unit time. The tube rate constant $k(\alpha)$ is obtained from the inverse of the mean first passage time for the folding events through tube α. The population $P(\alpha)$ of the unfolded state which fold through tube α is

calculated using $P(\alpha)=\sum_{i\in U} t(i|\alpha)/T_{\text{total}}$, where t(i|α) is the residence time that trajectories which fold through tube α spend on unfolded node i, and T_{total} is the total simulation time. The set {P(α)} corresponds to a kinetic partition of the unfolded state ensemble into populations which fold along each tube; the partition has the property that $\sum_{\alpha} P(\alpha)=P_{\text{Unfold}}^{\text{total}}$. Additionally, for the hub folding model, as different unfolded populations folding independently along different paths, $\sum_{i} P(i|\alpha)P(i|\beta)$ is expected to be small, although this is not observed for the kinetic network model of Trp-cage folding constructed from the ultra-long MD trajectory (see below).

The values for P(α), J(α) and k(α) calculated for the top 16 folding tubes are shown in Fig. 4. Although the fluxes vary by more than three fold along the different folding tubes, they all have very similar folding rates i.e. k(α) ≈ constant. Consequently, the tube fluxes are proportional to the corresponding populations, i.e. P(α) ∝ J(α). We discuss how these results are a direct consequence of the significant mixing within the unfolded state before folding.

We have also computed the overlap between the distributions of the unfolded state populations which fold along different tubes. This is another indication of the extent of mixing within the unfolded state between folding events. We define the conditional probability $P(i|\alpha)=\frac{t(i|\alpha)}{\sum_{i\in U} \mathrm{t(i|\alpha)}}$. It corresponds to the fraction of the time the system spends on unfolded node i given that it folds along tube α. It is obtained by normalizing t(i|α) with the total time trajectories which fold through tube α spend in the entire unfolded region. The distribution P(i|α) over all the unfolded nodes describes the extent to which the unfolded states are explored before folding through tube α. In the case of extensive mixing between unfolded state populations, P(i|α) should be only weakly dependent on i. In contrast, for a kinetic hublike scenario, in which the exchanges between unfolded states are severely limited, each folding tube's P(i|α) distribution is confined to a local area of the unfolded ensemble.

To examine the extent to which the P(i|α) distributions overlap, we define a quantitative measure of the overlap between the two normalized distributions P(i|α) and P(i|β) in discretized space $\Omega(\alpha,\beta)=\frac{\sum_{i\in U} \mathrm{P(i|\alpha)} \times \mathrm{P(i|\beta)}]}{\sqrt{\sum_{i\in U} P(i|\alpha)^2}\sqrt{\sum_{i\in U} P(i|\beta)^2}}$. For the case of rapid mixing, Ω(α,β) will be ≈ 1. In the opposite regime, if the two folding tubes α and β are connected with very different regions of the unfolded ensemble, then Ω(α,β) will be ≈ 0. We found that all the matrix elements of Ω(α,β) for the top 16 folding tubes are greater than 0.95, which implies extensive mixing prior to folding.

Another unresolved question in protein folding concerns the role of residual structures in the unfolded states in modulating folding kinetics [14]. For example, UV-Raman measurements found significant α-helical content of Trp-cage under denaturing condition [43]. It has been speculated that residual secondary structure may help accelerate Trp-cage folding. To probe the role of preexisting residual structure in folding, we performed a large number of stochastic simulations initiated from unfolded conformations with and without the residual secondary structure. In the MD trajectory, about 7 % of the unfolded conformations contain an intact N-terminal α-helix (residues 2-9). This value is consistent with the UV-resonance Raman study [43]. We initiated 8000 folding simulations from (1) unfolded conformations

with intact N-terminal α-helix and (2) unfolded states with a disordered N-terminal segment. The folding starting from the conformations with α-helix is only slightly faster than that starting from those conformations without the secondary structure.

We also examine in a similar fashion the influence of nonnative compactness in the unfolded region. Folding simulations starting from collapsed unfolded conformations (Rg < 7.0 Å) and from extended conformations (Rg > 15 Å) result in virtually identical MFPT. Therefore, neither the preexisting N-terminal α-helix nor nonnative compactness was found to significantly influence the Trp-cage folding rate.

Comparison of the Folding Kinetics and Pathways from MD and MSM-MD

There are 31 folding events in the 208 μs MD trajectory. The distribution of the first passage times of the folding events can be approximately fit to a single exponential (Fig. S9). The mean first passage time (MFPT) of the 31 folding transitions is found to be ≈ 5.5 μs, in good agreement with the experimental folding time of 4 μs at room temperature [56]. Another quantity describing the folding kinetics is the transit time, which is the time for a folding trajectory to traverse the intermediate region. The average transit time of all the folding transitions is 23.6 ns; the range is between 1.8 ns and 267 ns. The observation that the average transit time is ~ 200 times smaller than the mean folding passage time of ~ 5 μs indicates that the Trp-cage folding is highly cooperative.

Examination of the folding transitions sampled by the ultra-long MD trajectory revealed heterogeneous structural pathways leading to the folded state. Here we discuss two representative paths (Fig. 5). In pathway A, the polypeptide chain first undergoes a hydrophobic collapse, forming a compact molten globule containing multiple non-native H-bonds; later on, the non-native interactions are loosened, which is followed by the formation of the N-terminal α-helix and native hydrophobic core. In pathway B, the folding starts from more extended conformation with pre-formed α-helix in the unfolded state; the hydrophobic core and the 3_{10}-helix then form in concert to complete the folding process. The two pathways have very different transit times: in pathway A, the trajectory has to loosen the non-native contacts and gradually replace them with native hydrophobic core; these localized structural rearrangements take place in a relatively long transit time of 44 ns. By contrast, the folding along pathway B is much simpler because the starting unfolded structure contains fewer non-native interactions; the associated transit time in this pathway is only 3 ns. We also found that the pathway B is a more dominant pathway, i.e. there are more folding transitions in which the α-helix forms before the hydrophobic collapse.

By carrying out stochastic simulations on the kinetic network generated from the MD simulation, a large number of folding transitions are obtained. The folding passage time distribution exhibits single exponential decay (Fig. S10), with a folding time close to the average of the 31 transitions observed in the ultra-long MD simulation.

We have compared the folding tubes constructed using transition path theory applied to the MSM-MD kinetic network with the folding transitions observed in the MD trajectory. Using a rmsd = 3.0 Å as the cutoff distance between a TPT folding pathway and an MD folding pathway, we found that 29 out of 31 MD folding transitions can be assigned to TPT folding tubes. Fig. 6 shows the fluxes of the top 12 TPT folding tubes compared with the number of the MD folding transitions assigned to each folding tube. There is a general correspondence between the folding transitions observed in the ultra-long MD simulation and the flux through folding tubes generated from the kinetic network (Fig. 6). The MSM folding tube with the largest flux is also the one that contains the largest number of MD folding transitions among all the folding tubes. The folding mechanism in this tube is the same as in the MD pathway B discussed above, which features an early formation of the α-helix (Fig.

5). The analysis of the TPT folding tubes shows that about 45 % of the flux is carried by pathways in which the α-helix forms early. In the remaining pathways the hydrophobic compaction either occurs early or forms in concert with the α-helix.

It should be noted that in general, MSM predicts more folding pathways than that contained in the raw MD trajectory. For example, two such pathways that are predicted by MSM-MD but not observed in the original MD data are shown in Fig. S11. The reason for the richer folding pathways in MSM can be qualitatively understood by considering the schematic transition diagram shown in Fig. S12, where a MD trajectory contains transitions U1→I→U2 and separately N→I→N. There is no direct folding transition in this MD transition diagram. The corresponding MSM, however, would predict folding pathways U1→I→N and U2→I→N.

MSM Constructed from REMD Simulations

The MSM-MD model we have analyzed in the previous sections was based on MD simulations performed well above the Trp-cage folding temperature with just 17 % native population [20]. What is the kinetic picture of Trp-cage folding below the folding temperature? To address the temperature dependence of folding kinetics, here we study a Markov network model of Trp-cage built from temperature replica exchange (REMD) simulations with implicit solvation over a wide temperature range.[40,41] We call this Markov network model MSM-REMD.

Using the REMD data obtained over a wide temperature range we determined the Trp-cage melting behavior (Fig. 7); the folding temperature T_f was found to be ≈ 468 K. The high melting temperature compared to the experimental T_f is typical of the results found with implicit solvent models and is partially attributable to the overly attractive intramolecular interactions in the OPLS-AA force field with AGBNP implicit solvent model [57] used in the REMD simulations.

To investigate Trp-cage kinetics below and above the folding temperature, we performed TPT pathway calculations and stochastic simulations using the MSM-REMD model at T = 465 K and 539 K, at which 54 % and 11 % of the populations are folded, respectively. We found that both the rate of mixing within the unfolded state and the diversity of the folding pathways vary strongly with temperature. The folding pathway ensemble becomes more diverse at the higher temperature. At T = 465 K, the top folding tube carries 61 % of the total flux and the top three folding tubes account for ≥ 90 % of the total flux. In contrast at T = 539 K, the top folding tube carries just 30 % of the flux and it takes nine folding tubes to accumulate 90 % of the total flux (Fig. 8, top row). Below the folding temperature we observe very slow folding through one of the folding tubes (Fig. 8). The folding through this tube is ~ 60 times slower than the fastest folding tube.

Next, we examine the temperature dependence for the mixing within the unfolded states by computing the conditional probabilities P(i|α) for the different folding tubes α, and the overlaps of P(i|α) with P(i|β) among the folding tubes at both T = 465 K and T = 539 K. Table 2 shows the overlap factor Ω(α,β) between different pairs of unfolded population distributions P(i|α) and P(i|β) associated with the top folding tubes. At T = 465 K, overlaps between the P(i|α) of the slow folding tube (No. 3) and that of the rest of the tubes are zero, indicating that the unfolded populations associated with the slow tube and those with the other tubes fold independently. At the high temperature T = 539 K, the overlaps between the P(α) of slow tube (No. 6) and those of the other tubes increased significantly (Table 2). This trend reflects more extensive mixing within the unfolded ensemble above the folding temperature. How does this enhanced mixing in the unfolded state affect the folding kinetics at higher temperature? For this, we compare the tube populations P(α), fluxes J(α) and

folding rates k(α) for the different folding tubes at the two temperatures (Fig. 8). It can be seen that the difference in the folding rates between the slowest and fastest folding tubes decreases significantly as the temperature is increased (Fig. 8, bottom row). At the lower temperature T = 465 K the ratio of the slowest folding rate to the fastest folding rate is $k_{slow}/k_{fast} \approx 0.015$. At the higher temperature T = 539 K this ratio becomes $k_{slow}/k_{fast} \approx 0.2$.

Another observation from Fig. 8 is that at the higher temperature T = 539 K, there is a clear correlation between J(α) and P(α), i.e. $J(\alpha) \propto P(\alpha)$. The plot of J(α) and P(α) at this temperature shows that the correlation coefficient R-squared ≈ 0.7 (Fig. S13). Such correlation between J(α) and P(α) is not observed at the lower temperature T = 465 K.

We have identified the conformational species that folds through the slow folding tube at the lower temperature. The average structure of the slow folding population adopts a hairpin-like conformation stabilized by between 5 and 7 nonnative hydrogen bonds. It also contains a nonnative hydrophobic core featuring Trp6-Arg16 stacking. It is found that the same compact conformation is also sampled by the ultra-long MD trajectory in explicit solvent, but in explicit solvent, these conformations are not metastable. In contrast, their lifetime is ~ 500 ns with the AGBNP implicit solvent model.

The results using the MSM-REMD trajectory at the lower temperature reflects the increased ruggedness in the free energy landscape of the unfolded ensemble at the lower temperatures with the AGBNP implicit solvent model, and a more hub-like partitioning of the unfolded state ensemble, in which slow folding populations and fast folding populations folding independently. At the higher temperature however the MSM-REMD results are more qualitatively similar to those observed using the MSM-MD model (compare Fig. 4 and the right half of Fig. 8).

Discussion

In this study we have focused on kinetics within the unfolded state ensemble and its influence on folding, which is less well understood compared with other aspects of protein folding. We begin with the following observations: First, the sampling of the unfolded states in the ultra-long MD simulation shows good convergence (Fig. S5). Second, the kinetic properties observed in the direct MD simulation are well reproduced by the Markov state model constructed from the MD simulation: The folding passage times, transit times, unfolded state dynamics and folding pathways obtained from the 208 μsec MD simulation and 64 millisecond stochastic MSM-MD simulation are in good agreement (Table 1 and Fig. 6).

The analysis of the MD and MSM-MD data suggests that the unfolded population of Trp-cage mixes well before folding. The relaxation time of the autocorrelation function for the radius of gyration and the principal components are ≤ 40 ns, which are much faster than the folding time ~ 5 μs. The experimentally determined time scales for large scale motions in unfolded proteins have been reported in several studies [58–61]. Using laser-temperature jump Sadqi et al. found that the hydrophobic collapse of the acid-denatured 40-residue BBL occurs on a ~ 60 ns time scale [58]. Using single-molecule spectroscopy Schuler and coworkers found that the chain reconfiguration time for the unfolded, 70-residue cold shock protein (Csp) was approximately 100 ns [59,61]. The orders of magnitude for the relaxation times for the radius of gyration calculated in the present study for the 20-residue Trp-cage are consistent with those measured for the somewhat larger polypeptides BBL and Csp.

Additional evidence of significant mixing within the unfolded state ensemble comes from the fact that the folding rates are independent of where the folding is initiated from within the unfolded basin and the extensive overlaps among the unfolded state populations which

fold along different pathways. The strong correlation between folding passage times and the fraction of the unfolded nodes visited before each MD folding transitions also reflects the absence of major internal barriers in the unfolded basin (Fig. 3).

To further analyze how the kinetics within the unfolded state affects folding, we have studied a simple 5-state model (Fig. S14a), in which two unfolded nodes 1 and 2 have very different microscopic escape rates k_{13} and k_{24}, with $k_{13}/k_{24} = 10$. We examine how the rates of the fast folding tube α and slow folding tube β are affected by changes in the U-state interconverting rate k_{12}. The simulation shows that the tube folding rates k_α and k_β strongly depend n the rate of transition within U-state (Fig. S14b). When the transition rate within U-state k_{12} is small relative to the microscopic escape rates k_{13} and k_{24}, the unfolded populations on nodes 1 and 2 fold independently with very different k_α and k_β respectively governed by the intrinsic escape rates k_{13} and k_{24}, producing bi-exponential folding time distributions (Fig. S15). As k_{12} increases, the difference between the k_α and k_β decreases monotonically. When k_{12} is comparable to or faster than k_{13} and k_{24}, the two tube folding rates k_α and k_β converge to the overall folding rate k_{tot} (Fig. S14b and Fig. S15).

On the basis of the simple model results and using the concept of P_α introduced earlier, we can write expressions for the folding rates k(α) when mixing within the unfolded free energy basin is much slower or much faster than folding (see Table 3). The tube folding rate k_α has the simple, general expression $k_\alpha = \frac{J_\alpha}{P_\alpha}$, where the tube flux is obtained from transition path theory $J_\alpha = \sum_{i \in U, j \notin U} k_{ij} p_i^{ep} p_j^{fold}$[30,31]. In the fast exchange limit the folding rate along a folding tube becomes $k_\alpha \approx \frac{J_{total}}{P^{eq}_{total,U}} \equiv k_{tot}$, which is the same for all the folding tubes, independent of the intrinsic rates (k_{12} and k_{24}). In the limit of slow exchange within the unfolded basin, the result is $k_\alpha = \frac{J_\alpha}{P^{eq}_{\alpha,U}}$, where $P^{eq}_{\alpha,U}$ is the unfolded population locally associated with tube α. In this regime k_α depends on the escape rates from the local population $P^{eq}_{\alpha,U}$. While rates along individual folding tubes are modulated by the rate of U-state mixing, the k_{tot}, which is the simple average of folding events per unit time (also the same as the inverse of the mean first passage time), is constant and can be written as the weighted average of k_α: $k_{tot} = \frac{J_{total}}{P_{total,U}} = \frac{J_\alpha + J_\beta}{P_{total,U}} = k_\alpha \frac{P_\alpha}{P_{total,U}} + k_\beta \frac{P_\beta}{P_{total,U}}$.

We now apply these insights to interpret the results of Trp-cage folding obtained from the MSM-MD model. As shown in Fig. 4, for the results from the stochastic simulations on the MSM-MD kinetic network, $P_\alpha \propto J_\alpha$ and $k_\alpha \approx$ constant. Comparing with Table 3, we can see that such behavior is consistent with the scenario of significant U-state mixing.

The result that under the fast exchange condition the folding kinetics is single-exponential was first pointed out by Ellison and Cavagnero in an insightful study on the role of unfolded state kinetics [16]. The authors studied different types of folding energy landscapes using simple kinetic models and concluded that under the condition of fast exchange in the unfolded basin, it is not possible to determine the microscopic rate constants for different parallel folding routes by a simple experiment in bulk solution. They also observed that the folding flux along a given route is controlled by the intrinsic escape rate along that route. These results agree well with our analysis of the Trp-cage folding kinetics and the simple model. As we show in the Table 3 here, the flux for a folding route is determined by the product of intrinsic rate and the equilibrium population of the unfolded region from which the folding route originates.

The results for P_α, J_α and k_α calculated using the MSM-REMD model (Fig. 8) reveal the temperature dependence of the unfolded states landscape: at low temperature, the landscape contains a deep basin whose population folds through a slow folding tube only, and does not exchange with other regions of the unfolded ensemble; at higher temperature, there is considerable mixing between the slow folding population and the rest of the unfolded basin, which is reflected in the overlap factor $\Omega(\alpha,\beta)$ (Table 2). The relationships between J_α and P_α at the different temperatures provide additional evidence for the greater mixing within the unfolded basin at higher temperature. We have shown that a strong correlation between J_α and P_α is a signature for significant mixing in the unfolded state relative to folding (Table 3), here we look at J_α and P_α obtained from the MSM-REMD model. At T = 465 K, there is little correlation between J_α and P_α; however at T = 539 K, a stronger correlation between the two quantities emerges ($R^2 \approx 0.7$, Fig. S13). This suggests that at the higher temperature, the unfolded state landscape becomes substantially smoother and this allows for more rapid exchange between the different folding tube populations. The MSM-REMD result at the higher temperature is qualitatively similar to the results we obtained at ambient temperature using the MSM-MD model based on the ultra-long MD trajectory.

We now examine our results in the light of the insightful paper by Bicout and Szabo,[5] who studied different folding landscapes by modeling the protein dynamics in conformational space as diffusion under a spherically symmetric potential. They showed that the folding kinetics on both a golf-course landscape (Levinthal) and funnel landscape[2] is single exponential, which arises from the entropic barrier to folding. They also showed that to get such two-state behavior a folding trajectory on these landscapes needs not explore most of the unfolded states before folding.[5] Our results for the Trp-cage folding are consistent with theirs: for the single exponential, two-state folding behavior of Trp-cage, a trajectory typically explores ~ 27 % of the unfolded space before it folds (Fig. 3).

Finally, we discuss our results from the perspective of the kinetic hub model of folding introduced by Pande.[34,53,54,62] In this model, the folded state F acts as a hub, so that most paths which connect pairs of unfolded states U1 and U2 pass through F. Hub like behavior also appears to imply that the unfolded state partitions into subspaces which largely fold along different pathways.[14] However, as we have reported in this paper, we find no evidence of a kinetic partitioning of the U state space into regions which mostly fold along different pathways. Dickson and Brooks[63] introduced a hub score to quantify the hub-like character of a network; the hub score for (U1,U2) corresponds to the fraction of trajectories starting at U1 which pass through native state F before reaching U2. We have calculated the distribution of hub scores for the MSM-MD network constructed from the Shaw trajectory and obtained an average hub score of 0.88. Such a high hub score is not inconsistent with the observation of single exponential folding kinetics of Trp-cage and the rapid mixing within its unfolded state. It is simply a manifestation that on a funnel landscape, because of the energetic bias towards the native state, two sufficiently separated unfolded states will be connected by pathways which include folding events. It is not clear therefore how the hub score can be used to distinguish a rugged landscape from a smooth folding funnel.

Conclusions

An important problem in protein folding is to understand the relationship between the structural heterogeneity and kinetics within the unfolded free energy basin and the folding kinetics. We have investigated the unfolded state kinetics and folding pathways of the miniprotein Trp-cage using (1) a 208 μsec MD trajectory in explicit solvent; (2) Markov state model simulations based on the ultra-long MD trajectory; and (3) a Markov state model constructed from replica exchange molecular dynamics simulations in implicit solvent over a wide temperature range. Using stochastic simulations and transition path theory we have

explored the kinetics of the unfolded state ensemble and studied its impact on the kinetics of folding. By comparing the folding behavior observed in the fully atomistic Trp-cage simulations with the kinetics in a simple 5-state folding model, we have obtained a relationship between the rate of mixing in the unfolded state and the folding kinetics along individual pathways (tubes). Here the main result is that the conformational mixing in the unfolded state modulates the apparent protein folding rates by affecting the waiting times for folding along different routes. When this mixing is comparable to or faster than folding, the folding rates associated with different folding routes converge to the same value which is independent of the intrinsic rates along any given route; despite the presence of multiple folding routes with non-uniform barriers, the folding kinetics is essentially single exponential. In the slow exchange limit, the folding rate of along folding route is controlled by the intrinsic rates along the route; In this case the different unfolded populations fold independently and the overall folding kinetics can deviate from single exponential.

We have presented results showing that, based on atomistic Trp-cage models in explicit and implicit solvent the Trp-cage unfolded state ensemble does not contain long-lived metastable states; there exists significant mixing in the unfolded state. These include the time scale for chain extension and compaction within the unfolded state, the approximately uniform folding rates among different folding tubes, the extensive overlaps among the unfolded populations associated with the different folding tubes, and the strong correlation between the flux along folding tubes and the unfolded state populations associated with the corresponding tubes. Because of the significant internal mixing of the unfolded state, the probability to fold along any of the multiple folding paths is almost the same regardless of where in the unfolded state the folding is initiated

Analysis of the Markov state model constructed from the temperature replica exchange data provides an opportunity to probe the temperature dependence of the unfolded states kinetics. By studying the results below and above the midpoint of the folding transition, we found that in implicit solvent at low temperature the unfolded state landscape contains a slow folding basin; internally the exchange between the slow folding population and other regions of the unfolded state basin is much slower than folding. Above the folding temperature, the unfolded state landscape becomes less rugged allowing more rapid mixing and considerable overlap among the unfolded populations associated with the different folding tubes.

Our study reinforces and extends the simple kinetic model of Ellison and Cavagnero[16] in providing a physical basis for the apparent two-state, single exponential kinetics exhibited by many proteins with heterogeneous folding pathways. The current work makes use of Markov state kinetic network models built from atomic simulations, stochastic simulations on the network and transition path theory to analyze how kinetics within the unfolded state affects folding rates. For the models we have studied the unfolded state of Trp-cage is well mixed, and the rate of exchange within the unfolded state ensemble is comparable to or faster than the folding rate. We emphasize that Trp-cage is a small system and its kinetics may not be representative of the folding of larger and more complex proteins. It would be interesting to apply the computational tools and the concepts of P_α and the overlap matrix introduced in the present study to investigate the folding mechanisms of proteins with different native topology and more complex unfolded state kinetics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work has been supported by a grant from the National Institute of Health (GM30580). Some of the calculations were performed using the XSEDE allocation TG-MCB100145. We thank Dr. David Shaw and Dr. Piana-Agostinetti for reading the manuscript and for making the long MD trajectory of Trp-cage available for analysis. Dr. Dmitrii Makarov read the manuscript and made very helpful comments. Dr. Emilio Gallicchio also made helpful suggestions. Dr. Weihua Zheng performed the REMD simulations of Trp-cage. Dr. Junchao Xia helped with the figures.

This manuscript has been prepared for the special issue of the Journal of Physical Chemistry in honor of the 60th birthday of Peter Wolynes. My (RL) interactions with Peter go back to the days of Prince House II at Harvard thirty-five years ago. His passion for science was clear from the first time I spoke with him. And so too was his brilliance and strong opinions. It is always exciting and energizing talking with Peter Wolynes. Happy Birthday!

References

1. Bryngelson JD, Wolynes PG. Intermediates and barrier crossing in a random energy model (with applications to protein folding). The Journal of Physical Chemistry. 1989; 93:6902–6915.
2. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins. 1995; 21:167–195. [PubMed: 7784423]
3. Wang J, Onuchic J, Wolynes P. Statistics of Kinetic Pathways on Biased Rough Energy Landscapes with Applications to Protein Folding. Physical Review Letters. 1996; 76:4861–4864. [PubMed: 10061399]
4. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annual Review of Physical Chemistry. 1997; 48:545–600.
5. Bicout DJ, Szabo A. Entropic barriers, transition states, funnels, and exponential protein folding kinetics: A simple model. Protein Science. 2000; 9:452–465. [PubMed: 10752607]
6. Shea JE, Brooks CL III. FROM FOLDING THEORIES TO FOLDING PROTEINS□: A Review and Assessment of Simulation Studies of Protein Folding and Unfolding. Annual Review of Physical Chemistry. 2001; 52:499–535.
7. Onuchic JN, Wolynes PG. Theory of protein folding. Current Opinion in Structural Biology. 2004; 14:70–75. [PubMed: 15102452]
8. Wolynes PG. Energy landscapes and solved protein-folding problems. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2005; 363:453–467.
9. Kubelka J, Hofrichter J, Eaton WA. The protein folding ‘speed limit’. Current Opinion in Structural Biology. 2004; 14:76–88. [PubMed: 15102453]
10. Shakhnovich E. Protein Folding Thermodynamics and Dynamics: Where Physics, Chemistry, and Biology Meet. Chemical Reviews. 2006; 106:1559–1588. [PubMed: 16683745]
11. Dill KA, Ozkan SB, Shell MS, Weikl TR. The Protein Folding Problem. Annual Review of Biophysics. 2008; 37:289–316.
12. Thirumalai D, O’Brien EP, Morrison G, Hyeon C. Theoretical Perspectives on Protein Folding. Annual Review of Biophysics. 2010; 39:159–183.
13. Karplus M. Behind the folding funnel diagram. Nature Chemical Biology. 2011; 7:401–404.
14. Sosnick TR, Barrick D. The folding of single domain proteins—have we reached a consensus? Current Opinion in Structural Biology. 2011; 21:12–24. [PubMed: 21144739]
15. Zheng W, Schafer NP, Wolynes PG. Frustration in the energy landscapes of multidomain protein misfolding. Proceedings of the National Academy of Sciences. 2013; 110:1680–1685.
16. Ellison PA, Cavagnero S. Role of unfolded state heterogeneity and en-route ruggedness in protein folding kinetics. Protein Science. 2006; 15:564–582. [PubMed: 16501227]
17. Gin BC, Garrahan JP, Geissler PL. The Limited Role of Nonnative Contacts in the Folding Pathways of a Lattice Protein. Journal of Molecular Biology. 2009; 392:1303–1314. [PubMed: 19576901]
18. Shaw, DE.; Bowers, KJ.; Chow, E.; Eastwood, MP.; Ierardi, DJ.; Klepeis, JL.; Kuskin, JS.; Larson, RH.; Lindorff-Larsen, K.; Maragakis, P., et al. Millisecond-scale molecular dynamics simulations on Anton. ACM Press; 2009.

19. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, et al. Atomic-Level Characterization of the Structural Dynamics of Proteins. Science. 2010; 330:341–346. [PubMed: 20947758]
20. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How Fast-Folding Proteins Fold. Science. 2011; 334:517–520. [PubMed: 22034434]
21. Piana S, Lindorff-Larsen K, Shaw DE. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? Biophysical Journal. 2011; 100:L47–L49. [PubMed: 21539772]
22. Dellago C, Bolhuis PG, Csajka FS, Chandler D. Transition path sampling and the calculation of rate constants. The Journal of Chemical Physics. 1998; 108:1964.
23. Faradjian AK, Elber R. Computing time scales from reaction coordinates by milestoning. The Journal of Chemical Physics. 2004; 120:10880. [PubMed: 15268118]
24. Laio A. Escaping free-energy minima. Proceedings of the National Academy of Sciences. 2002; 99:12562–12566.
25. Beccara SA, Skrbic T, Covino R, Faccioli P. Dominant folding pathways of a WW domain. Proceedings of the National Academy of Sciences. 2012; 109:2330–2335.
26. Zheng W, Qi B, Rohrdanz MA, Caflisch A, Dinner AR, Clementi C. Delineationof Folding Pathways of a β-Sheet Miniprotein. The Journal of Physical Chemistry B. 2011; 115:13065–13074. [PubMed: 21942785]
27. Swope WC, Pitera JW, Suits F. Describing Protein Folding Kinetics by Molecular Dynamics Simulations. 1. Theory†. The Journal of Physical Chemistry B. 2004; 108:6571–6581.
28. Andrec M, Felts A, Gallicchio E, Levy RM. Chemical Theory and Computation Special Feature: Protein folding pathways from replica exchange simulations and a kinetic network model. Proceedings of the National Academy of Sciences. 2005; 102:6801–6806.
29. Chodera JD, Swope WC, Pitera JW, Dill KA. Long-Time Protein Folding Dynamics from Short-Time Molecular Dynamics Simulations. Multiscale Modeling & Simulation. 2006; 5:1214–1226.
30. Berezhkovskii A, Hummer G, Szabo A. Reactive flux and folding pathways in network models of coarse-grained protein dynamics. The Journal of Chemical Physics. 2009; 130:205102. [PubMed: 19485483]
31. Noe F, Schutte C, Vanden-Eijnden E, Reich L, Weikl TR. From the Cover: Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proceedings of the National Academy of Sciences. 2009; 106:19011–19016.
32. Bowman GR, Beauchamp KA, Boxer G, Pande VS. Progress and challenges in the automated construction of Markov state models for full protein systems. The Journal of Chemical Physics. 2009; 131:124101. [PubMed: 19791846]
33. Pande VS, Beauchamp K, Bowman GR. Everything you wanted to know about Markov State Models but were afraid to ask. Methods. 2010; 52:99–105. [PubMed: 20570730]
34. Bowman GR, Pande VS. Protein folded states are kinetic hubs. Proceedings of the National Academy of Sciences. 2010; 107:10890–10895.
35. Marinelli F, Pietrucci F, Laio A, Piana S. A Kinetic Model of Trp-Cage Folding from Multiple Biased Molecular Dynamics Simulations. PLoS Computational Biology. 2009; 5:e1000452. [PubMed: 19662155]
36. Voelz VA, Bowman GR, Beauchamp K, Pande VS. Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1-39). Journal of the American Chemical Society. 2010; 132:1526–1528. [PubMed: 20070076]
37. Prinz JH, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera JD, Schütte C, Noé F. Markov models of molecular kinetics: Generation validation. The Journal of Chemical Physics. 2011; 134:174105. [PubMed: 21548671]
38. Prinz JH, Keller B, Noé F. Probing molecular kinetics with Markov models: metastable states, transition pathways and spectroscopic observables. Physical Chemistry Chemical Physics. 2011; 13:16912–16927. [PubMed: 21858310]
39. Metzner P, Schütte C, Vanden-Eijnden E. Transition Path Theory for Markov Jump Processes. Multiscale Modeling & Simulation. 2009; 7:1192–1219.
40. Zheng W, Gallicchio E, Deng N, Andrec M, Levy RM. Kinetic Network Study of the Diversity and Temperature Dependence of Trp-Cage Folding Pathways: Combining Transition Path Theory

with Stochastic Simulations. The Journal of Physical Chemistry B. 2011; 115:1512–1523. [PubMed: 21254767]

41. Deng N, Zheng W, Gallicchio E, Levy RM. Insights into the Dynamics of HIV-1 Protease: A Kinetic Network Model Constructed from Atomistic Simulations. J Am Chem Soc. 2011; 133:9387–9394. [PubMed: 21561098]
42. Neidigh JW, Fesinmeyer RM, Andersen NH. Designing a 20-residue protein. Nature Structural Biology. 2002; 9:425–430.
43. Ahmed Z, Beta IA, Mikhonin AV, Asher SA. UV-Resonance Raman Thermal Unfolding Study of Trp-Cage Shows That It Is Not a Simple Two-State Miniprotein. Journal of the American Chemical Society. 2005; 127:10943–10950. [PubMed: 16076200]
44. Neuweiler H. A microscopic view of miniprotein folding: Enhanced folding efficiency through formation of an intermediate. Proceedings of the National Academy of Sciences. 2005; 102:16650–16655.
45. Mok KH, Kuhn LT, Goez M, Day IJ, Lin JC, Andersen NH, Hore PJ. A pre-existing hydrophobic collapse in the unfolded state of an ultrafast folding protein. Nature. 2007; 447:106–109. [PubMed: 17429353]
46. Simmerling C, Strockbine B, Roitberg AE. All-Atom Structure Prediction and Folding Simulations of a Stable Protein. Journal of the American Chemical Society. 2002; 124:11258–11259. [PubMed: 12236726]
47. Zagrovic B, Snow CD, Shirts MR, Pande VS. Simulation of Folding of a Small Alpha-helical Protein in Atomistic Detail using Worldwide-distributed Computing. Journal of Molecular Biology. 2002; 323:927–937. [PubMed: 12417204]
48. Chowdhury S, Lee MC, Xiong G, Duan Y. Ab initio Folding Simulation of the Trp-cage Mini-protein Approaches NMR Resolution. Journal of Molecular Biology. 2003; 327:711–717. [PubMed: 12634063]
49. Pitera JW. Understanding folding and design: Replica-exchange simulations of “Trp-cage” miniproteins. Proceedings of the National Academy of Sciences. 2003; 100:7587–7592.
50. Zhou R. Trp-cage: Folding free energy landscape in explicit water. Proceedings of the National Academy of Sciences. 2003; 100:13280–13285.
51. Paschek D, Hempel S, Garcia AE. Computing the stability diagram of the Trp-cage miniprotein. Proceedings of the National Academy of Sciences. 2008; 105:17754–17759.
52. Juraszek J, Bolhuis PG. Rate Constant and Reaction Coordinate of Trp-Cage Folding in Explicit Water. Biophysical Journal. 2008; 95:4246–4257. [PubMed: 18676648]
53. Lane TJ, Bowman GR, Beauchamp K, Voelz VA, Pande VS. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. Journal of the American Chemical Society. 2011; 133:18413–18419. [PubMed: 21988563]
54. Bowman GR, Voelz VA, Pande VS. Taming the complexity of protein folding. Current Opinion in Structural Biology. 2011; 21:4–11. [PubMed: 21081274]
55. Du R, Grosberg A, Tanaka T. Random Walks in the Space of Conformations of Toy Proteins. Physical Review Letters. 2000; 84:1828–1831. [PubMed: 11017636]
56. Qiu L, Pabit SA, Roitberg AE, Hagen SJ. Smaller and Faster: The 20-Residue Trp-Cage Protein Folds in 4 μs. Journal of the American Chemical Society. 2002; 124:12952–12953. [PubMed: 12405814]
57. Gallicchio E, Paris K, Levy RM. The AGBNP2 Implicit Solvation Model. Journal of Chemical Theory and Computation. 2009; 5:2544–2564. [PubMed: 20419084]
58. Sadqi M, Lapdius L, Munoz V. How fast is protein hydrophobic collapse? Proceedings of the National Academy of Sciences. 2003; 100:12117–12122.
59. Nettels D, Gopich IV, Hoffmann A, Schuler B. Ultrafast dynamics of protein collapse from single-molecule photon statistics. Proceedings of the National Academy of Sciences. 2007; 104:2655–2660.
60. Neuweiler H, Johnson CM, Fersht AR. Direct observation of ultrafast folding and denatured state dynamics in single protein molecules. Proceedings of the National Academy of Sciences. 2009; 106:18569–18574.

61. Soranno A, Buchli B, Nettels D, Cheng RR, Muller-Spath S, Pfeil SH, Hoffmann A, Lipman EA, Makarov DE, Schuler B. Quantifying internal friction in unfolded intrinsically disordered proteins with single-molecule spectroscopy. Proceedings of the National Academy of Sciences. 2012; 109:17800–17806.

62. Bowman GR, Voelz VA, Pande VS. Atomistic Folding Simulations of the Five-Helix Bundle Protein λ6–85. Journal of the American Chemical Society. 2011; 133:664–667. [PubMed: 21174461]

63. Dickson A, Brooks CL. Quantifying Hub-like Behavior in Protein Folding Networks. Journal of Chemical Theory and Computation. 2012; 8:3044–3052. [PubMed: 24027492]

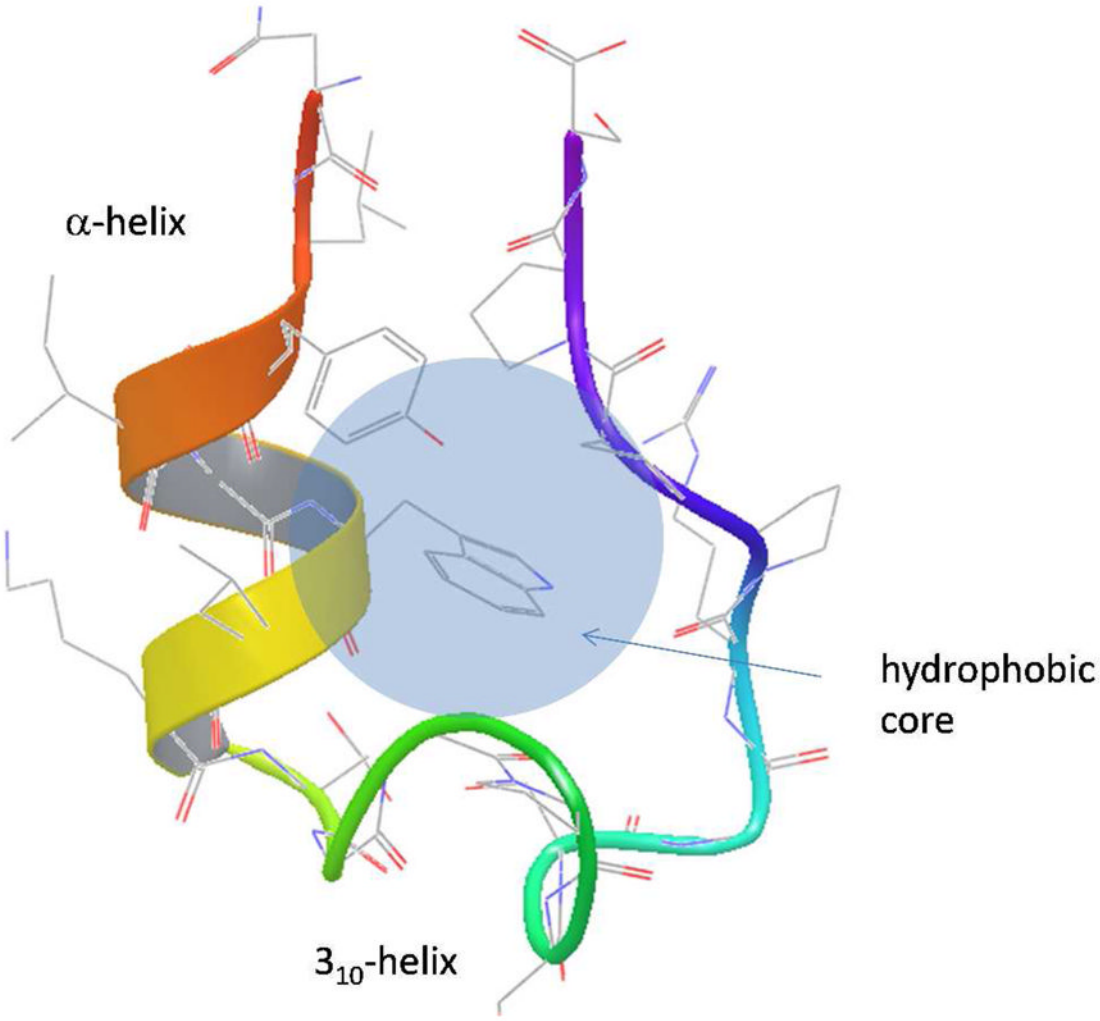

Figure 1.
NMR structure of Trp-cage miniprotein.

Fig. 2a

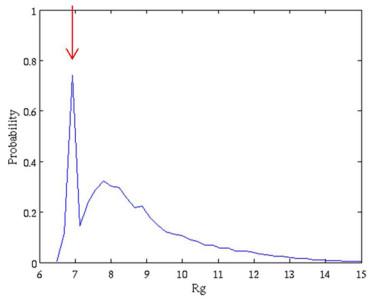

Fig. 2b

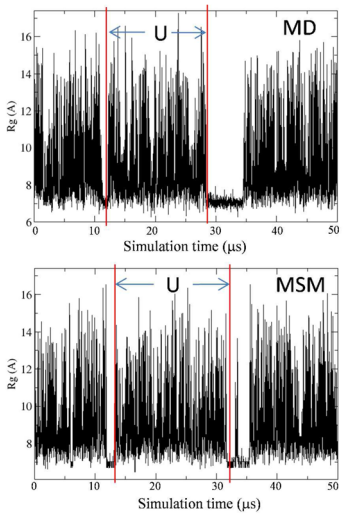

Figure 2.

(a) The distribution of radius of gyration from the 208 μs MD trajectory [20]. The R_g corresponding to the native structure is indicated by the red arrow. (b) A 50 μs portion of the time series of R_g obtained from the MD trajectory and from the stochastic simulation trajectory on the MSM-MD. The letter U indicates a time span in the unfolded state.

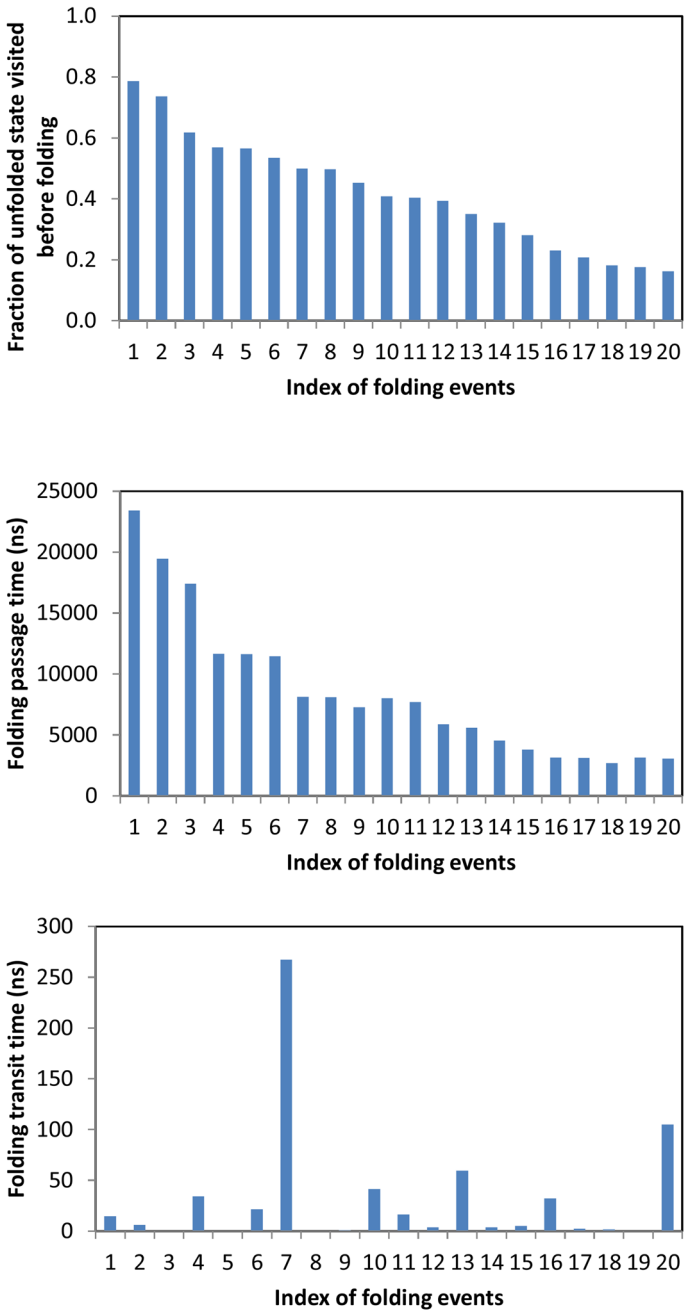

Figure 3.

For each unfolding/folding transition observed in the ultra-long MD trajectory, from top to bottom: the fraction of unfolded state space sampled before folding ordered from largest to smallest; the corresponding passage time for folding; the folding transit time. The x-axis is the index of each folding transition, in descending order of fraction of visited unfolded states.

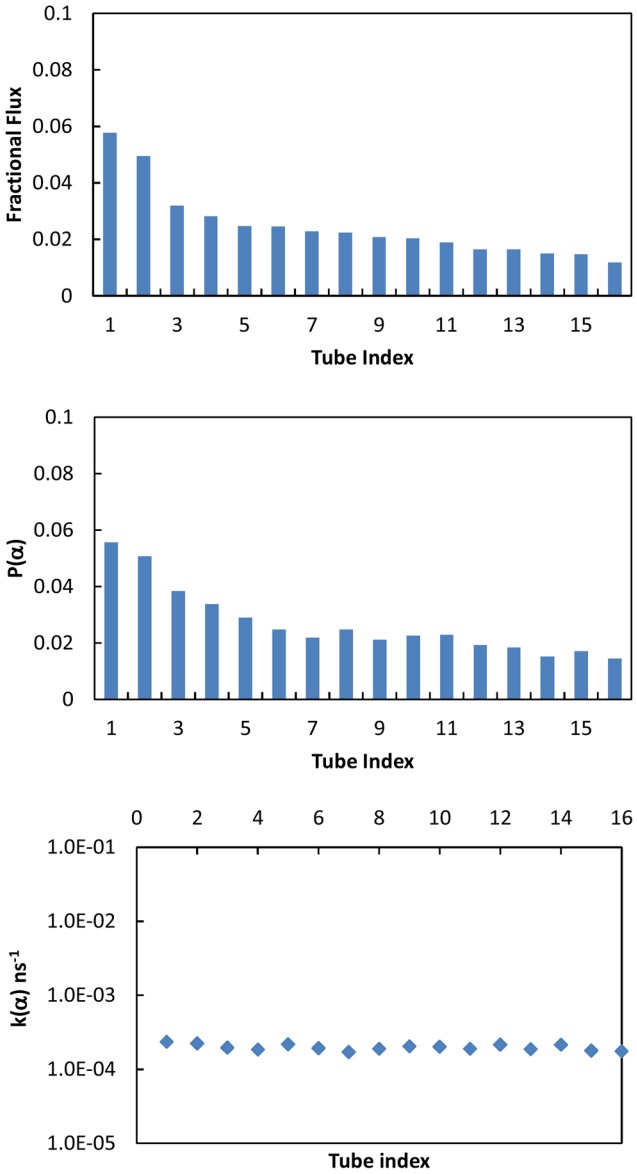

Figure 4.
The tube fluxes J(α), tube population P(α) and tube folding rates k(α) for the top 16 folding tubes obtained using MSM-MD.

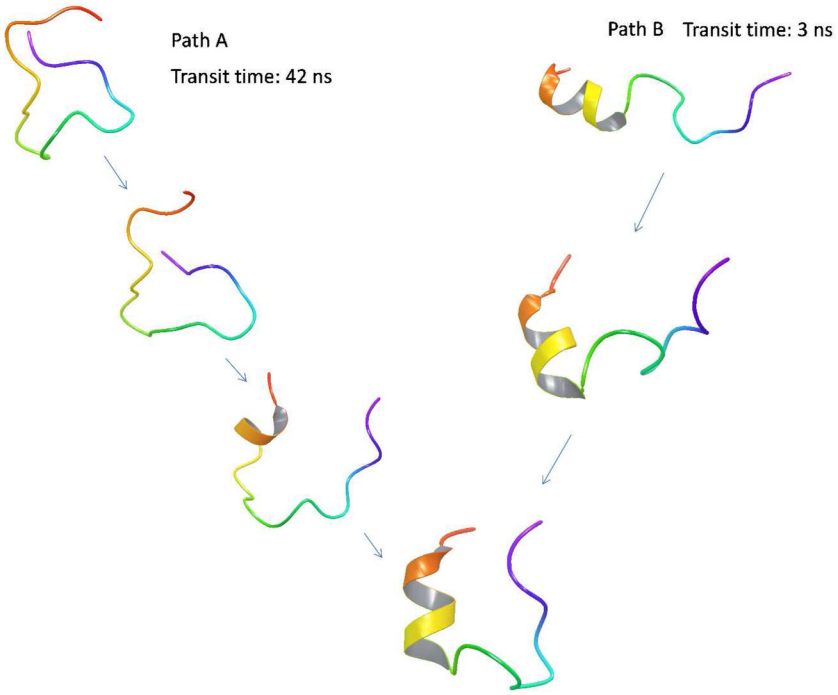

Figure 5.
Two representative folding transitions extracted from the 208 μs MD trajectory.

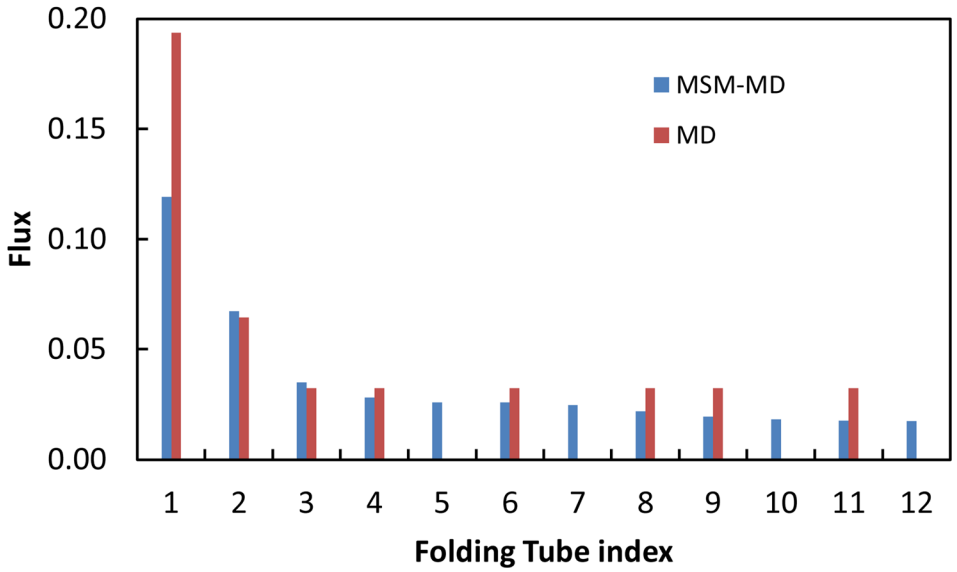

Figure 6.
The blue histogram shows the distribution of fluxes along the top 12 folding tubes generated from the MSM-MD kinetic network using TPT. The red histogram represents the fluxes of MD folding transitions projected onto the folding tubes.

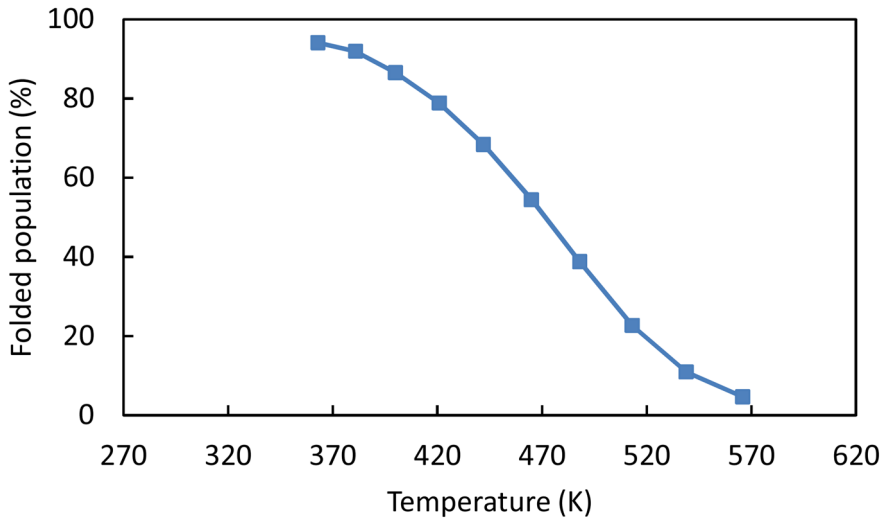

Figure 7.
The melting curve of Trp-cage obtained from REMD simulation.

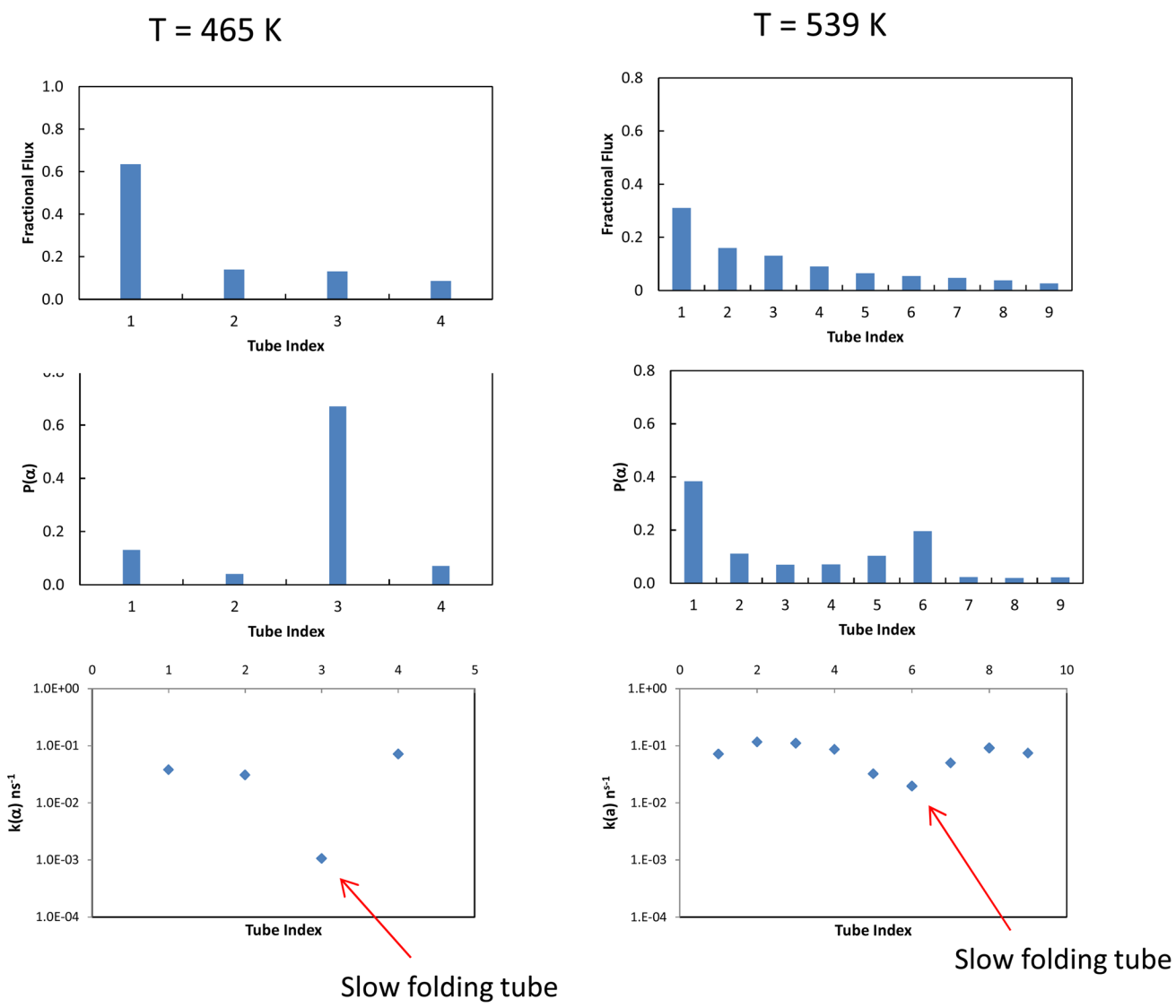

Figure 8.
Results of J(α), P(α) and k(α) calculated using the MSM-REMD model at T = 465 K and T = 539 K.

Table 1

Time scales of Trp-Cage folding and unfolded state kinetics from the ultra-long MD and MSM-MD model.

	Mean folding time	Conformational relaxation time in the unfolded states[3]	Time to sample 90% of the unfolded states	Transit time of folding
MD (208 μs)[1]	5.5 μs	Mode 1: 6 ns; Mode 2: ~ 38 ns	~ 40 μs	23.6 ns
MSM-MD[2]	5.3 μs	Mode 1: 7 ns; Mode 2: ~ 30 ns	~ 44 μs	30 ns

[1] The MD trajectory contains 31 folding events.

[2] The Markov state model contains 25000 microstates. The time scales are obtained by running kinetic Monte-Carlo simulation which generated 10000 folding events.

[3] Estimated from the time correlation functions of R_g.

Table 2

MSM-REMD results: The overlap factor matrix Ω(α,β) involving the top folding tubes that account for ≥ 95 % of the total folding flux. The slow folding tubes are indicated by red.

(a) T = 465 K				
Folding tube	**1**	**2**	**3**	**4**
1	1.00	0.24	0.00	0.33
2	0.24	1.00	0.00	0.62
3	0.00	0.00	1.00	0.00
4	0.33	0.62	0.00	1.00

(b) T = 539 K									
Folding tube	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	1.00	0.65	0.87	0.73	0.61	0.79	0.26	0.82	0.78
2	0.65	1.00	0.37	0.85	0.25	0.29	0.35	0.28	0.89
3	0.87	0.37	1.00	0.56	0.71	0.97	0.15	0.95	0.50
4	0.73	0.85	0.56	1.00	0.46	0.55	0.38	0.37	0.80
5	0.61	0.25	0.71	0.46	1.00	0.75	0.11	0.61	0.40
6	0.79	0.29	0.97	0.55	0.75	1.00	0.09	0.88	0.38
7	0.26	0.35	0.15	0.38	0.11	0.09	1.00	0.02	0.31
8	0.82	0.28	0.95	0.37	0.61	0.88	0.02	1.00	0.47
9	0.78	0.89	0.50	0.80	0.40	0.38	0.31	0.47	1.00

Table 3

Results of k_α, P_α, and J_α for folding tube α, determined from studying a simple folding model (Fig. S13) (neglecting the small intermediate state population).

	k_α	P_α	J_α
General	$\frac{J_\alpha}{P_\alpha}$	$\frac{\sum_{i\in U} t(i\|\alpha)}{T_{\text{total}}}$	$\sum_{i\in U, j\notin U, j\in\alpha} k_{ij} p_i^{\text{eq}} p_j^{\text{fold}}$
Fast U-State mixing (funneled folding landscape)	$\frac{J_{\text{total}}}{P^{\text{eq}}_{\text{total},U}}$	$\frac{J_\alpha}{J_{\text{total}}} P^{\text{eq}}_{\text{total},U}$	
Slow U-state mixing (hub folding landscape)	k_{13}	p_1^{eq}	$k_{13} p_1^{\text{eq}}$
	k_{24}	p_2^{eq}	$k_{24} p_2^{\text{eq}}$