Published in final edited form as:
J Phys Chem B. 2013 October 24; 117(42): . doi:10.1021/jp404331d.

Characterization of Folding Mechanisms of Trp-cage and WW-domain by Network Analysis of Simulations with a Hybrid-resolution Model

Wei Han[†,‡] and **Klaus Schulten**[†,‡,*]

[†]Beckman Institute, University of Illinois at Urbana-Champaign, USA

[‡]Center for Biophysics and Computational Biology, University of Illinois at Urbana-Champaign, USA

Abstract

In this study, we apply a hybrid-resolution model, namely PACE, to characterize the free energy surfaces (FESs) of trp-cage and a WW domain variant along with the respective folding mechanisms. Unbiased, independent simulations with PACE are found to achieve together multiple folding and unfolding events for both proteins, allowing us to perform network analysis of the FESs to identify folding pathways. PACE reproduces for both proteins expected complexity hidden in the folding FESs, in particular, meta-stable non-native intermediates. Pathway analysis shows that some of these intermediates are, actually, on-pathway folding intermediates and that intermediates kinetically closest to the native states can be either critical on-pathway or off-pathway intermediates, depending on the protein. Apart from general insights into folding, specific folding mechanisms of the proteins are resolved. We find that trp-cage folds via a dominant pathway in which hydrophobic collapse occurs before the N-terminal helix forms; full incorporation of Trp6 into the hydrophobic core takes place as the last step of folding, which, however, may not be the rate-limiting step. For the WW domain variant studied we observe two main folding pathways with opposite orders of formation of the two hairpins involved in the structure; for either pathway, formation of hairpin 1 is more likely to be the rate-limiting step. Altogether, our results suggest that PACE combined with network analysis is a computationally efficient and valuable tool for the study of protein folding.

Keywords

multiscale simulation; free energy landscape; Markov state model; transition path theory

Introduction

How proteins assume their three-dimensional (3D) structures remains an intriguing and fundamental question.[1,2,3,4,5] To address this question, it is necessary to acquire thorough

[*]Corresponding author: kschulte@ks.uiuc.edu; phone: +1-217-244-1604.

Supporting Information: Figure S1, evolution of key structural properties during the folding of trp-cage; Figure S2, distributions of p_{fold} for hairpins 1 and 2 during the folding of WW-domain; Figure S3, comparison between p_{fold} values obtained through the transition matrix and obtained directly from trajectories; Figure S4, scheme of HB potentials in PACE; Table S1, comparison of the vdW radii in the original PACE and those from statistical analysis; Tables S2, re-optimized parameters for the vdW interactions of side chain atoms; Table S3, comparison of evaporation free energies calculated with the original PACE and the new PACE and the experimental values; Table S4, comparison of RMSDs of six proteins simulated with the original PACE and the new PACE and with all-atom models; introduction of PACE and discussion about how PACE has been refined for the present study. This material is available free of charge via http://pubs.acs.org.

knowledge of free energy surfaces (FES) governing protein folding and of the associated folding mechanisms.[6,7,8,9] Computational characterization of folding mechanisms and their underlying FES is a difficult task. One of the major reasons for the difficulty is that due to high dimensionality of the folding space, it is complicated to determine and represent the FES and subsequently extract folding kinetics from it. One common approach to characterize the FES governing folding is to project it onto a small number of progress variables, such as root mean square deviation from the native structure (RMSD), radius of gyration (R_g) of proteins and the first few principal components of atomic coordinates of proteins. However, such an approach may fail to unmask hidden complexity of the FES.[10,11] In the past decade, new methods based on network analysis have been proposed to characterize the FES of folding through unprojected representations of the FES.[12,13,14,15,16,17,18,19,20] These methods have been applied to folding trajectories from molecular dynamics (MD) simulations, revealing the complexity of the folding FES for small peptides and mini-proteins.[21,22,23,24,25,26,27,28]

The network analysis of the folding FES usually requires unbiased folding simulations of proteins. Ideally, the simulations need to include full atomistic detail and need to sample multiple folding and unfolding events. In practice, obtaining such simulations is difficult owing to the long timescale for folding events to arise and due to the requirement of powerful computational resources needed for all-atom molecular dynamics (MD) simulations. A recent special-purpose computer makes it feasible to simulate atomistic systems over millisecond duration,[29,30] having lead already to numerous new insights of folding.[31,32] However these computers are not generally available and, therefore, it is desirable to explore the FES of folding through a computational approach of wide availability. One option for such approach is to replace the atomistic representation by a coarse-grained (CG) model, which groups multiple atomic sites into one site so that the computational effort needed for folding simulations becomes significantly reduced.[33,34,35] Folding simulations with CG models have been assisted often by Go-like potentials, which accelerate folding tremendously, but only at the price of adding undesirable biases that distort folding. Several CG models that do not rely on Go-like potentials have been applied to fold unstructured proteins to their native structures,[36,37,38,39,40,41] but none of them have been used to perform also a network analysis to examine the folding FES. Exceptions include a model for prediction of protein structures, namely SimFold, developed by Fujitsuka et al,[42] which has been implemented recently in a CG simulation platform, namely CafeMol,[43] and a united-atom model with transferable atomistic potentials by Hubner et al.[41] The former model has been applied to characterize the folding landscapes and network dynamics of protein G and SH3 domain;[44] the latter model has been applied to the network analysis of the FES of a helical protein, namely ENH, to identify folding intermediates.[45]

In the present study, we performed folding simulations with a hybrid-resolution model, namely PACE.[46,47,48] In PACE, proteins are described in united-atom (UA) representation and embedded in a CG solvent environment modeled with the MARTINI force field.[49] PACE has been used already to successfully fold proteins with up to 73 amino acids into their native structures.[50,51] Here we investigate whether PACE can furnish also satifactory descriptions of the folding FES and, thereby, of the folding mechanism for, as depicted in Fig 1, the helical protein trp-cage[52] and the β-sheet protein hPin WW,[53,54] the latter modified through loop-enhancing mutations. For this purpose, we carried out with PACE multiple folding simulations of the two proteins. Folding mechanisms of both proteins had been intensively investigated previously in experimental[55,56,57,58] and computational studies.[25,29,59,60,61] In the present study we construct the folding networks of the two proteins from folding simulations with PACE. Folding pathways were analyzed with a mathematic framework based on transition path theory (TPT)[25,62] to elucidate the folding mechanisms. The results demonstrate that simulations with PACE, indeed, capture several key features of the folding mechanisms of the two proteins that had been seen earlier only in atomistic simulations.

Methods

The folding simulations carried out employed a CG description, yet results rival those based on an all-atom description. The results from folding simulations are formulated usually in terms of a Markov state model that identifies main folding pathways and their kinetics.[16,17] In the following we describe first the CG model, PACE,[47,51] and its simulation through the NAMD MD program.[63] We then specify the trajectory analysis method employed which started with conformational clustering and continued with the identification of macro-states corresponding to basins of attraction; lastly, the folding pathway analysis in the framework of the macro-states is outlined.

PACE model for proteins

All simulations were performed with a hybrid-resolution model, namely PACE. The potential functions of PACE and their parametrization are described in detail in two prior reports[47,51] as well as in Supplementary Information (SI). In short, the protein model of PACE includes all heavy atoms as well as hydrogen atoms on amide groups so that packing of side chains and hydrogen bonding interactions are modeled realistically. The interactions of atomic sites in proteins are parameterized by using different data as references, such as experimental thermodynamic data, potential of mean force (PMF) profiles of polar and charged interactions from atomistic simulations and statistical backbone potentials from a PDB coil library.[64] The protein model is embedded in an environment modeled with the MARTINI coarse-grained (CG) force field.[49] Cross-resolution parameters were optimized through a thermodynamic-based approach, *i.e.,* through reproducing experimental thermodynamic quantities, which is in line with the parameterization scheme of MARTINI. Comprising such different ways of parametrization, PACE has been used successfully in folding both small α-helical and β-sheet proteins, not relying on information of their native structures.[50,51] To improve the description of packing of side chains, the version of PACE used here includes a further refine-ment of parameters for excluded volume of side chain atoms, based on previous statistical analysis on PDB structures.[65] With the improved parameters, native structures of six proteins with various structural motifs were well maintained in 100 ns simulations, with average root mean square deviation (RMSD) to PDB structures of 2.5 ± 0.3Å, comparable to 1.8 ± 0.4 Å arising in all-atom simulations of the same protein set. The RMSD values of the present PACE simulations are better than 3.1 ± 0.3 Å, the value obtained with the original PACE.[47] The detail of the optimization is discussed in SI. PACE is available at http://www.ks.uiuc.edu/~whan/PACE/PACEvdw/ and can be employed readily with the simulation program NAMD.[63]

Simulation setup

NAMD 2.9,[63] with a minor modification for using PACE parameters, was employed for all folding simulations. Switching functions were applied to Coulomb potentials from 0.0 to 1.2 nm, and to LJ potentials from 0.9 to 1.2 nm. Temperature was maintained by Langevin dynamics with a damping coefficient of 0.1 ps^{-1}; pressure was maintained by a Nosé-Hoover Langevin piston barostat[63] with a period of 200 fs and a decay rate of 100 fs. PDB structures were solvated in cubic boxes of CG water particles with 1.5 nm clearance from the solutes, neutralized with Na^+ or Cl^- ions and buffered at 0.15M ionic concentration. The resulting systems contained ~1300 CG particles for trp-cage and ~2500 CG particles for the hPin WW variant. To study folding, each system was subjected, after optimization, to 10-ns heating at 700 K. The denatured structures, randomly selected from the heating simulations, were employed as starting points for independent folding simulations lasting 1 μs and considerably longer in some cases as specified in Results. Simulated structures were stored every 0.1 ns.

Conformational clustering

We clustered conformations in two steps. In a first step, conformations were collected from one of every five frames in the trajectories. An iterative clustering algorithm proposed by Daura et al.[66] was used to group the collected conformations pairwise according to RMSD of selected atoms. In each iteration, a cluster with the largest number of neighbors that are within a given cutoff of RMSD to the center of the cluster is removed from a pool of conformations. The remaining pool was used for the next iteration. The algorithm ended when the pool became empty. In a second step, each conformation from the trajectories was mapped to the clusters obtained. A conformation was thought to belong to a cluster if its RMSD to the center of the cluster is the smallest among the RMSDs to the centers of all the clusters.

Partitioning conformational states

We considered each conformational cluster as a conformational state, counting the number of transitions (n_{ij}) between states i and j from the trajectories. An undirected graph was built with edge capacities $c_{ij} = (n_{ij} + n_{ji})/2$ to represent the kinetic network among conformational states at equilibrium.[10,67] As the conformational clustering described above usually generates large numbers of conformational states, the resulting kinetic network is normally too complicated for computational analysis. Thus, it is helpful to group multiple conformational states into one macro-state to reduce the original network into a smaller one, involving only macro-states.[17] A desirable grouping of conformational states should yield meta-stable states as macro-states.[16,17] Transitions among the conformational states within the same meta-stable states are much more frequent than those across meta-stable states. In the present study, we grouped conformational states with the PCCA+ algorithm[17,68,69] that optimizes the partition of states by minimizing the ratio of inter-macrostate transitions to intra-macrostate ones.

In short, the PCCA+ algorithm starts with a transition probability matrix $\{T_{ij}\}$, constructed through $T_{ij} = c_{ij}/\Sigma_k c_{ik}$, namely, the transition probability from state i to state j at a lag time Δt, which is 0.1 ns in our case. In order to partition n conformational states into m macro-states, the transition probability matrix is diagonalized and the right eigenvectors corresponding to the 2_{nd}–$(m + 1)_{\mathrm{th}}$ largest eigenvalues are used to construct an m-dimensional vector space. In this m-dimensional space, the i_{th} conformational state has coordinate $(\phi_{2,i}, \ldots, \phi_{m+1,i})$, where $\phi_{k,i}$ is the i_{th} component of the k_{th} right eigenvector. The m most distant conformational states in this space are selected as vertices of a simplex, each representing the center of a macro-state. The other conformational states are then, according to their convex distances to the vertices, assigned to the closest vertex and, thereby, to the corresponding macro-state. The PCCA+ analysis was performed with the module *mm–pcca* implemented in the program package EMMA (http://simtk.org/home/emma).[70] The direct transitions between two macro-states I and J are

$$c_{IJ} = \sum_{i \in I} \sum_{j \in J} c_{ij}. \quad (1)$$

Folding pathway analysis

We followed Noé et al. in analyzing folding pathways based on transition networks employing Transition Path Theory (TPT).[25,62] An essential part of the analysis of folding pathways involves the calculation of the committor probability $p_{\text{fold},\,i}$ of folding,71,72 defined as the probability that a protein, when being in state *i*, hits folded (*N*) states before reaching unfolded states (*U*). The p_{fold} for all states can be obtained by solving the system of linear equations,

$$p_{\text{fold},i}=\sum_j T_{ij}\, p_{\text{fold},j}, \quad (2)$$

for the conditions $p_{\text{fold},\,i} = 0$ for $i \in U$ and $p_{\text{fold},\,i} = 1$ for $i \in N$. For each edge between states *i* and *j*, its contribution to the *U*–*N* transitions can be evaluated through the effective flux f_{ij}. For molecules at equilibrium, this flux is

$$f_{ij}=p_{\text{eq},i}(1-p_{\text{fold},i})\, T_{ij}\, p_{\text{fold},j}, \quad (3)$$

where $p_{\text{eq},\,i}$ is the probability of state *i* at equilibrium. However,the effective flux f_{ij} still accounts for recrossing events that do not contribute to the folding transitions.[25] The actual net folding flux f_{ij}^{+} is calculated by removing the part for recrossings in f_{ij}, namely by

$$f_{ij}^{+}=\max\{0, f_{ij}-f_{ji}\}. \quad (4)$$

The network $\{f_{ij}^{+}\}$ of folding fluxes can be decomposed into individual pathways.[25] The way to decompose the network is not unique. Here we followed the approach proposed by Noé et al.[25]. In this approach the strongest pathway was chosen each time. The folding flux (*f*) of the chosen pathway is the minimum f_{ij}^{+} along the chosen pathway. The chosen pathway was removed from the network of folding fluxes by subtracting, along the chosen pathway, its folding flux *f* from the network $\{f_{ij}^{+}\}$. The decomposition yields a set of folding pathways and their fluxes. The probability of pathway *i* is then calculated as

$$p_{\text{path},i}=\frac{f_i}{\sum_i f_i}, \quad (5)$$

where f_i is the folding flux of the pathway *i*. Similarly, the probability of state *i* being present on the folding pathways is estimated as

$$p_{\text{onpath},i}=\frac{\sum_j f_j P(i,j)}{\sum_j f_j}, \quad (6)$$

where $P(i, j) = 1$ if state *i* is on pathway *j* and $P(i, j) = 0$ otherwise.

Results

For the purpose of the present study 84 *μ*s of PACE folding simulations were carried out altogether and analyzed; given that PACE simulations, due to PACE's simplified solvent model, accelerate folding fivefold to tenfold,[51] the sampling achieved in our study compares effectively to roughly a millisecond of all-atom molecular dynamics. In the following, we will describe first the folding of trp-cage and then folding of a hPin WW variant. For each protein, we will first present its free energy surface (FES) characterized by means of conformational clustering and subsequent PCCA+ analysis, and discuss then associated folding pathways characterized by analyzing a network of folding fluxes based on the FES. Finally, we summarize for each of the two proteins the main folding mechanisms.

Folding simulations of trp-cage

An essential part in the network analysis of protein folding is to construct transition matrices based on unbiased simulations in which multiple folding and unfolding events are observed. We examined the occurrence of multiple folding and unfolding events in thirty 1-μs folding simulations of trp-cage that we have carried out, each simulation starting from a different unfolded conformation (see Methods). In each simulation, trp-cage folding was simulated at 320 K with PACE.

Folding and unfolding events of the trp-cage simulations were monitored by evaluating the root mean square deviation (RMSD) from the available NMR structures (PDB ID: 1L2Y),[52] based on C_α-atoms of residue 3–19 as well as the side chain of Trp6. Since the Trp6 side chain is the central part of the hydrophobic core of trp-cage, inclusion of Trp6 in the RMSD calculation is necessary for discriminating the native structures from near-native ones having misfolded hydrophobic cores. Such near-native conformations have been observed in previous folding simulations of trp-cage.[60] We considered a folding event to occur if the RMSD drops, for the first time, below 1.5 Å, down from 5.5 Å, the value above which the protein is thought to be unfolded. Similarly, an unfolding event happens once the RMSD rises above 5.5 Å. The analysis revealed that there were 25 folding and 20 unfolding events sampled in the 30 simulations. Given the stated high number of events, these simulations are considered to offer sufficient statistics to construct transition matrices for analysis of trp-cage folding.

Free energy surface of trp-cage

To characterize the free energy surface (FES) of trp-cage (see Fig 1b), we first clustered ~ 2.8×10^5 conformations from the 30 simulations, using the RMSD-based clustering method as described in Methods. The RMSD of C_α-atoms of residues 3–19 and atoms of the Trp6 side chain was used to measure distances between conformations. Conformations closer than a cutoff of 2 Å were grouped in one state. The clustering analysis revealed 2631 conformational states for trp-cage. As these conformational states were obtained from multiple simulations, it is important to know whether the sampling of each conformational state converges. We compared, therefore, for each conformational state its observed probability p_{obs} to its expected equilibrium value p_{eq}. The latter can be evaluated based on the transition probability matrix $\{T_{ij}\}$, calculating its left eigenvector that has an eigenvalue of unity.[16,17] The relative deviation between p_{obs} and p_{eq} turns out to be ~2.5% on average for all the conformational states, suggesting that the sampling is close to convergence.

FESs containing a large number of conformation states, like the one here with 2631 conformational states, in general are complex and, thus, difficult to analyze. In order to achieve a less complex and clearer picture of the FES, we partitioned the 2631 conformational states into 250 meta-stable macro-states, using the PCCA+ algorithm[17] as described in Methods. For the analysis of folding pathways as discussed below, we also identified as a native macro-state the conformational states that have less than 2 Å average RMSDs to the PDB structure. The population of the native macro-state is ~25.6%, close to the experimental value (30–35%) at 320 K.[52]

The population analysis revealed that the native macro-state is the most populated macro-state in the FES. Nevertheless, other non-native macro-states, as listed in Table 1, have a comparable population, with a free energy within several k_BT of that of the native macro-state. These non-native macro-states are composed of a wide spectrum of conformations, ranging from the one that is native-like (macro-state 1 with RMSD of ~3.4 Å) to the ones that are significantly different from the native structure (macro-state 6 with RMSDs of ~6 Å) (Table 1). Obviously, the FES of trp-cage is more complicated than expected for a folding funnel with only a single dominant basin.

Folding mechanism of trp-cage

Considering the complexity of FES as revealed above, it is natural to wonder what are the folding pathways of trp-cage governed by such a FES. We have carried out an analysis of folding pathways based on the committor probability p_{fold},[71,72] as described in Methods. To obtain p_{fold}, both the native states and unfolded states need to be defined. For trp-cage, we defined "native states" as those included in the native macro-state and "unfolded states" as conformations with RMSD> 6 Å, i.e., a value high enough to guarantee that most of native structural elements are lost. On the basis of this definition, we performed the folding pathway analysis on the 250 macro-states obtained as described in Methods.

The outcome of the analysis described is a network of folding fluxes involving transitions between macro-states that contribute to actual transitions from unfolded states to folded states. The folding flux network can be decomposed into contributions from individual pathways, revealing 550 distinct pathways, the large number indicating a heterogeneous folding mechanism. We depict in Fig 2 the top seven pathways that account for 30% of the overall folding flux. These pathways go through seven intermediate macro-states, six of which (labeled in Fig 2) are among the top non-native macro-states discovered in the FES analysis (Table 1). The importance of each intermediate macro-state in folding of trp-cage can be estimated by calculating the probability of folding flux (p_{onpath}) flowing through this macro-state (Eq 7). Macro-states 1–3, as listed in Table 1, exhibit high p_{onpath} values, namely 98%, 38% and 24%, respectively. As shown in Fig 2, macro-states 1 and 3 act as kinetic routers at which upstream folding pathways converges. In particular, almost all the folding pathways converge at macro-state 1 before reaching the native states, indicating that macro-state 1 could be a critical on-pathway intermediate for trp-cage folding.

Series of pictures of the progressive structural change during the course of folding can resolve the folding mechanism. As suggested in previous studies,[25,27] such pictures may be derived from the folding network, by evaluating various structural properties of on-pathway intermediates and projecting them onto progress variables, such as the committor probability p_{fold}. We analyzed properties of key parts of trp-cage, including the hydrophobic core of four prolines (Pro12 and Pro17–19) and Trp6, the N-terminal helix and the loop region (see Fig 1a). All on-pathway intermediates with $p_{\mathrm{onpath}} > 0.01$ were covered by the analysis. The projections, as shown in Figs 3a–e, suggest that despite the heterogeneous folding pathways, the protein follows common mechanisms to fold. For example, the RMSDs of the loop region ($\mathrm{RMSD}_{\mathrm{loop}}$) are about 1.5 Å for most of the intermediates with $p_{\mathrm{fold}} > 0.15$, indicating the early formation of the native loop conformation during the course of folding (see Fig 3c). Similarly, the calculation of radii of gyration of the hydrophobic side chains (R_{g}) suggests that the protein collapses early, normally before $p_{\mathrm{fold}} \sim 0.5$, into a compact hydrophobic core (Fig 3e). However, as shown by the RMSD calculation (Fig 3b), no intermediate actually has correctly folded cores ($\mathrm{RMSD}_{\mathrm{core}} > 3.5$ Å). Instead, the side chain of Trp6 is more exposed (SASA> 90 Å^2) than seen in the native states (SASA= 45 ± 5 Å) (Fig 3d). Compared to the loop and the core, the N-terminal helix forms relatively late, only after $p_{\mathrm{fold}} \sim 0.5$ (Fig 3a). Macro-state 1, as discussed earlier, involves the last folding steps ($p_{\mathrm{fold}} = 0.97$) of almost all the folding pathways. Macro-state 1 already has almost well folded structures ($\mathrm{RMSD}_{\mathrm{helix}} = 1.2 \pm 0.7$ Å and $\mathrm{RMSD}_{\mathrm{loop}} = 1.3 \pm 0.4$ Å) except for the misfolded core with partially exposed Trp6 (SASA= 89 ± 9 Å ^{2}). Thus, in the last folding steps the protein rearranges the side chain of Trp6 to fully bury it and assume the correct core conformation.

Besides the involvement of the above major folding mechanisms, there is only about 8.9% of a chance for trp-cage to form the N-terminal helix before hydrophobic collapse takes place (Fig 3f). However, the protein may adopt intermediate conformations containing two separated hydrophobic clusters, one formed by Pro17–19 and the other by Trp6-Pro12. Interestingly, such conformations have been proposed in other computational studies as on-pathway intermediates.[59,73]

We characterized the folding mechanism of trp-cage using a small number (250) of macro-states to represent the FES of folding. However, pathway analysis relying on a FES represented by too few macro-states may result in misleading folding mechanisms. To ensure that the 250 macro-states investigated here are enough to reveal the folding mechanism of trp-cage, we also performed analysis of folding pathways using 500 macro-states to partition the FES. The resulting folding mechanism, as shown in Fig S1 (SI), is the same as that discussed above.

Folding simulations of hPin WW variant

To further explore the potential of PACE for protein folding, we performed simulations and a subsequent folding analysis for a β-sheet protein, hPin WW-domain, which has three β-strands forming two hairpins. WW-domains are known to fold much slower than helical proteins like trp-cage. To facilitate the folding of hPin WW for the present study, we replaced SRSSGR in the 4:6 loop of hairpin 1 (Fig 1b) by NPATGK, a segment engineered to promote formation of a 4:6 loop in a β-hairpin protein GB1p;[74] we also mutated residues 24–26 in the loop of hairpin 2 into a GTT sequence, which had been shown in both experimental and computational studies[75] to enhance folding of the WW domain by promoting formation of the loop of hairpin 2. Moreover, the simulations were performed at an elevated temperature (T = 350 K) for enhanced sampling.

We first performed 48 1-μs simulations starting each simulation with a distinct unfolded structure. To ensure unbiased sampling of folding events, 8 and 4 of the 48 simulations were further extended by 0.8 and 2.4 μs, respectively, so that unfolding events could also be observed once the protein folded. Following a previous folding study,[76] we thought that the protein is folded if its backbone RMSD compared to the native structure (PDB: 1PIN) of three β-strands, namely regions 6–11, 16–21 and 25–28 (S1–S3 in Fig 1b), is below 3 Å. The protein was considered unfolded when RMSDs of both hairpins 1 (S1–S2) and 2 (S2–S3) are larger than 5.5 Å. With these criteria, we found 12 folding and 10 unfolding events in the 48 simulations. The folded population is ~6.8%.

Folding mechanism of hPin WW variant

The conformational clustering was based on RMSD values of backbone and C_β-atoms, discarding three flexible residues at each terminus, and using a cutoff of 3 Å in pairwise RMSD calculations. We clustered ~5.4 $\times 10^5$ conformations into 4110 conformational states. Conformational sampling of these conformational states is close to converged, as reflected by a small average relative deviation (~1.3%) between p_{obs} and p_{eq}. The 4110 states were partitioned into 500 macro-states using the PCCA+ algorithm. All the native conformational states were included by a native macro-state. As in case of trp-cage, the FES of hPin WW exhibits multiple non-native macro-states. Table 2 lists the most populated ones.

With the definitions of folded and unfolded states the same as those in the analysis of folding events, we constructed the network of folding fluxes for the hPin WW variant. The analysis of the network of folding fluxes yielded 903 distinct folding pathways, the six most populated pathways accounting for 25% of the overall folding flux (Fig 4). Six of the top non-native macro-states, as listed in Table 2, are among the intermediates on these major pathways. In particular, about 33% and 47% folding flux goes through macro-state 5 and 9, respectively, indicating that these two macro-states are critical on-pathway intermediates.

Having obtained the folding pathways, one can address an important question regarding the folding mechanism of the WW domain: which hairpin is the first to form (Fig 1b)? We performed structural analysis of each intermediate on the folding pathways, monitoring the formation of hairpins 1 and 2 through calculation of average RMSDs of the two hairpins for the intermediate. Hairpins 1 and 2 are thought to form if $\text{RMSD}_{\text{S1–S2}}$ and $\text{RMSD}_{\text{S2–S3}}$, respectively, are less than 2.8 Å. The structural analysis of the on-pathway intermediates (Fig 4) revealed that both hairpins could be the first to form during the course of folding via two folding mechanisms. We estimated the probability of each folding mechanism according to

$$P(\mathrm{E})=\frac{\sum_i f_i \delta(\mathrm{E}, i)}{\sum_i f_i}, \quad (7)$$

where f_i is the folding flux of pathway i, E is the condition of whether hairpin 1 or 2 is first to form, and $\delta(\mathrm{E}, i)$ is 1 if E is true for pathway i and otherwise is 0. The calculation revealed that hairpin 2 has about 80% of a chance to be the first to form while hairpin 1 has about 20% of a chance.

We further examined whether formation of either hairpin or both hairpins is likely to be the rate-limiting step of the folding of the hPin WW variant. The rate-limiting steps of folding reactions are characterized through transition states on the FES. According to Du et al.,[71] transition states for folding reactions are thought to be the conformational states which have committor probabilities p_{fold} close to 0.5. We calculated p_{fold} values of the earliest on-pathway intermediates in which only one of the hairpins is observed. The resulting p_{fold} values were then binned for both mechanisms as presented in Fig 5. For the first mechanism, where hairpin 2 is the first to form, the distribution is mainly below p_{fold} values of 0.5, peaking at 0.1–0.4. For the second mechanism, where hairpin 1 is the first to form, the distribution is mainly at p_{fold} values above 0.5, peaking at 0.7–0.8. According to the stated definition of transition states, the p_{fold} distribution of the first mechanism indicates that the formation of hairpin 2 is mostly completed before the transition states are reached. Thus, the rate-limiting steps could be any of the remaining conformational transitions including the formation of hairpin 1 with an already completed hairpin 2. The p_{fold} distribution of the second mechanism indicates that hairpin 1 appears late during the course of folding, mainly after the transition states have been passed by. Thus, the rate-limiting steps that have already taken place may involve the formation of hairpin 1. Taken together, for either mechanism formation of hairpin 1 is more likely to be the rate-limiting step of the folding rather than formation of hairpin 2.

We further examined the robustness of the stated conclusion. First, the conclusion does not change when we employed a 1000-macrostate partition for the pathway analysis. (Fig S2 in SI). Second, we examined whether the reaction coordinates (p_{fold}) were properly estimated by comparing the p_{fold} values obtained with the transition matrix as discussed above with those derived directly from the simulation trajectories. In the latter approach, we followed Rao et al.[77] to calculate p_{fold} of a macro-state as a ratio between the number of frames of the macro-state from which the trajectories lead to folded states before reaching unfolded states and the total number of frames of the same macro-state. The p_{fold} values by the two approaches agree well with each other (Fig S3 in SI), with an average deviation of 0.17 for all the important intermediates ($p_{\text{onpath}} > 0.01$).

Mean first passage time of folding

To explore the folding kinetics of the two proteins investigated, trp-cage and the hPin WW variant, the mean first passage time (MFPT) of folding was determined. Two ways of obtaining folding events were applied to the calculation of the MFPT. One is to extract the folding events observed in the MD trajectories as discussed earlier, and the other is to perform Monte Carlo (MC) simulations with the transition probability matrix for conformational states obtained from the simulations, as was done in previous studies.[10,21] For the latter case, 10^4 MC trajectories were performed, each starting from unfolded states and ending at folded states. The cumulative distribution of first passage times (FPTs) for the WW domain matches well ones arising in case of a single exponential decay (Fig 6b), indicating simple two-state folding kinetics. However, the same distribution for trp-cage (Fig 6a) suggests that the folding process does not follow single exponential kinetics. Interestingly, non-single exponential folding kinetics has also been observed in recent bias exchange metadynamics simulations of trp-cage using atomistic models.[60] The cumulative distributions of FPT from the MD and MC trajectories, as shown in Fig 6, agree with each other well for both proteins, validating that the transition matrices derived here encapsulate the proper folding kinetics.

The MFPT of trp-cage was calculated to be 180 ± 30 ns through MD and 127 ± 7 ns through MC, about 2 times shorter than the MFPT of ~400 ns seen in a previous folding study with PACE,[51] owing to the higher simulation temperature here (320 K vs 300 K), and about 10 times shorter than the relaxation time of (~4.1 μs) measured at 296 K in T-jump experiments.[78] The MFPT of the hPin WW variant was calculated to be 900 ± 200 ns through MD and 1080 ± 10 ns through MC, about 5 times shorter than that (~4.3 μs at 353 K) of a closely related WW-domain protein measured in mutagenesis and T-jump experiments.[75] Our results agree with the previous study, supporting the previously observed general trend of PACE to accelerate the folding kinetics of small proteins by 5–10 fold.[51]

Mean first passage time for transitions from non-native to native macro-states

We further examined whether a non-native macro-state that is kinetically close to the native state also contributes significantly to the folding processes. We evaluated, through MC simulations, the MFPT of the transition from each major non-native macro-state to the native state. The resulting MFPTs measure kinetic distances between the macro-states and the native one. As shown in Tables 1 and 2, the macro-states close to the native state are not always important for the folding. For example, macro-state 1 of trp-cage and macro-state 2 of WW-domain are kinetically the closest to the native states, exhibiting MFPTs of ~10 – 20 ns, which is much shorter than the folding time. However, the former macro-state has the highest contribution ($p_{\mathrm{onpath}} = 0.98$) to the folding pathways while the contribution (p_{onpath}) by the latter is negligible ($p_{\mathrm{onpath}} = 0.06$).

Discussion

In the present study, we addressed the following question: can a hybrid-resolution force field like PACE[47,51] be used to capture folding mechanisms of small proteins? To this end, two known fast-folding proteins have been investigated through a combination of multiple folding simulations and a subsequent network analysis of free energy surfaces (FES).[10,25,27] One protein studied is a helical protein, namely trp-cage[52], and the other is a three-stranded β-sheet protein, namely one related to the hPin WW domain.[53,54] One benefit of PACE is that folding is accelerated by about one order of magnitude in PACE simulations, as demonstrated in previous folding simulations of various small proteins.[47] Therefore, with PACE multiple folding and unfolding events have been seen in μs MD stimulations. These simulations allowed us to construct transition matrices and to perform folding pathway analysis based on transition path theory (TPT).[25,62]

The FESs of trp-cage and of a hPin WW variant (Tables 1 and 2) revealed that neither protein has a funnel-like FES with only a single dominant macro-state containing the native state. Although for either protein the native states are the most-populated states among all states, several non-native macro-states on the FESs are clearly discernible and, thereby, act as meta-stable non-native folding intermediates. Such a feature has also been observed in previous computational studies, including a transition disconnectivity graph analysis of implicit-solvent simulations of GB1p hairpin and a conformational space network analysis of implicit-solvent simulations of a three-stranded β-sheet protein, β3s.[10,12,21]

An immediate question regarding the major non-native intermediates on FESs is whether they are on-pathway or off-pathway during the course of folding. The same studies on GB1p and β3s have sought to address this question by analyzing a simplified kinetics network, obtained by grouping states by their kinetic connection.[10,21] The conclusion from these studies was that all the major non-native intermediates are off the folding pathways and, instead, the folding crosses an entropic basin that essentially comprises a large collection of states, each of low probability. On the other hand, a combined analysis based on the Markov State Model (MSM)[18] and TPT demonstrated that a millisecond folding protein, namely NTL9, has several non-native macro-states present on the folding pathways.[27] The present study demonstrates that some, but not all, major non-native intermediates are on the folding pathway (Figs 2 and 4). Several non-native intermediates, such as macro-states 1–3 of trp-cage and macro-states 5 and 9 of hPin WW, have high probabilities (24–98%) of being on the folding pathways.

Macro-state 1 of trp-cage and macro-state 2 of hPin WW are two special non-native intermediates. They are both separated from the native macro-state by relatively low kinetic barriers. Macro-state 1 of trp-cage has a very high probability (98%) of being on-pathway and arises at the very late stage of folding ($p_{\mathrm{fold}} \sim 0.98$), acting as a router intervening between the native macro-state and other macro-states (Fig 2). Similar macro-states have also been observed in the folding of NTL9.[27] However, macro-state 2 of the hPin WW variant is mostly off-pathway ($p_{\mathrm{onpath}} \sim 6\%$) despite the low barrier between the macro-state and the native one. In fact, the macro-state is kinetically separated from other parts of the FES. The native macro-state as a router intervenes between other parts of the FES and the macro-state. Thus, the intermediates that are both geometrically and kinetically close to the native states do not always play critical roles in folding.

Besides general insights into the folding mechanisms, our simulations with PACE also provided specific pictures of how the two proteins fold. We identified two folding mechanisms for trp-cage. In the major mechanism, the loop region (residues 9–15, Fig 1a) folds into the native structure at the early stage of folding, followed by collapse of the hydrophobic side chains into misfolded cores. The misfolded cores have their Trp6 partly exposed. The N-terminal helix forms after the hydrophobic collapse. Such a folding mechanism is reminiscent of a nucleation-condensation route.[79] We also observed a minor mechanism, though ones that arises only with relative low probability (~8–9%), in which the helix forms before any collapse occurs. The minor mechanism is reminiscent of a diffusion-collision route.[80] The two mechanisms share the same last step of folding. In this step, the protein buries the side chain of Trp6 to assume correct conformations in the core via macro-state 1, which already adopts near-native conformations, but with Trp6 still exposed (Fig 2). A similar intermediate has also been seen in previous computational studies[60] (cluster 2). Our findings agree well with a scenario of multiple folding mechanisms for trp-cage. Such a scenario has been proposed by previous atomistic simulations using transition path sampling[59] and, more recently, by means of bias exchange metadynamics.[60]

The on-pathway intermediates introduced here are consistent with an experimental study which suggests through florescence correlation spectroscopy that the unfolded ensemble of trp-cage could include molten-globular intermediates with Trp6 partly exposed.[55] Based on the observation of molten-globular intermediates, this study also suggests that full incorporation of Trp6 into a pre-formed hydrophobic core is the rate-limiting step of folding. Our observation, however, is the opposite. We did not observe full incorporation of Trp6 in any folding pathway until the last step of folding. As this step takes place very late ($p_{\mathrm{fold}} \sim 0.97$), the results here imply that the final step of folding, in which the partly exposed Trp6 is fully incorporated into the core, may not be rate-limiting.

One question regarding folding mechanisms of WW domains is which of its two hairpins fold first.[58] Our folding analysis of the loop variant of hPin WW showed that there could be two folding mechanisms that have different orders of formation of hairpins. In the first mechanism, hairpin 1 forms before hairpin 2. This mechanism accounts for ~20% of folding transitions. The second mechanism, which accounts for ~80% of folding transitions, has a reverse order of formation of the hairpins. Both mechanisms have been observed in previous computational studies of WW domains and the probability of each mechanism was also estimated. For example, Noé et al. showed a 30:70 probability ratio for the two mechanisms for hPin WW through MSM analysis of a large number of short (~200 ns) simulations in explicit solvent.[25] For FiP35 (a fast-folding variant of hPin WW), Beccara et al. showed a 70:30 ratio of the two mechanisms and a 60:40 ratio at an elevated temperature through implicit-solvent simulations biased by Go-like potentials.[61] A 20:80 ratio was obtained by Krivov[81] through re-optimization of reaction coordinates of a ms-long atomistic simulation of FiP35 by Shaw et al.[29] As the above studies and ours investigated different variants of WW domains, using different force fields, at different temperatures and with different sampling techniques and analysis methods, it is not surprising that the observed probability ratios vary among these studies by 1–2k_{B}T. However, our study and those of others do suggest that coexistence of the two mechanisms are general in the folding of WW domains.

One assumption made in some studies of folding mechanisms of WW domains is that the rate-limiting step of folding is formation of one of the hairpins while the other is unstructured.[82] By this assumption, the formations of hairpins 1 and 2 would be the rate-limiting steps in the first and second folding mechanisms, respectively. However, the present study suggests rather that formation of hairpin 1 is the rate-limiting step of the folding of our hPin WW variant for both mechanisms. In particular, although hairpin 2 is the first to form in the second mechanism, the committor analysis (Fig 5) showed that formation of hairpin 2 is mostly completed before transition states could be reached ($p_{\mathrm{fold}} < 0.5$) and, therefore, is not the rate-limiting step. We attribute early formation of hairpin 2 to the GTT mutation that we introduced into the loop of hairpin 2. Such a mutation has also been introduced to FiP35 to speed up folding 3–4 fold.[75] Interestingly, a long time MD simulation revealed that the same mutant, namely GTT, exhibits considerable amount of β-sheet and native loop structures for hairpin 2 in the unfolded ensemble.[75] In fact, in our case the conformations with only hairpin 2 formed represent meta-stable non-native macro-states that are early intermediates on the folding pathways (such as macro-states 1 and 6 as shown in Fig 4). A similar intermediate has also been observed by Krivov[81] (state I2) through analysis of the folding simulation of FiP35. Therefore, we suspect that the assumption about the rate-limiting step as mentioned above may become invalid for some WW variants like the one studied here and perhaps the GTT mutant. Our hypothesis awaits experimental tests, probably through ϕ-value analysis as was done for other WW domains.[57,83]

Conclusion

In summary, we have demonstrated through folding analysis of trp-cage and a hPin WW variant that PACE is not only useful for folding simulations, but also for the characterization of free energy surfaces and folding pathways. Simulations with PACE are fast, normally running at ~1 μs per day for the proteins studied here utilizing a 16-core machine. PACE simulations also accelerate folding fivefold to tenfold.[51] As a result, multiple folding and unfolding events could be sampled for subsequent network analysis of folding kinetics, which in turn provides insight into the folding mechanism. Therefore, we suggest that PACE combined with network analysis is a valuable tool to tackle problems of protein folding.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work is supported by grant(s) from the National Institutes of Health (9P41GM104601, P41-RR005969, R01-GM067887) and from the National Science Foundation (MCB02-34938). Computer time was provided by Texas Advanced Computing Center through grant MCA93S028 allocated by the Extreme Science and Engineering Discovery Environment (XSEDE) program funded by the National Science Foundation.

References

1. Wolynes PG, Onuchic JN, Thirumalai D. Navigating the Folding Routes. Science. 1995; 267:17.
2. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem. 1997; 48:545–600. [PubMed: 9348663]
3. Kennedy D, Norman C. What don’t we know? Science. 2005; 309:75–75. [PubMed: 15994521]
4. Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. Annu Rev Biophys. 2008; 37:289. [PubMed: 18573083]
5. Onuchic JN, Wolynes PG. Theory of protein folding. Curr Opin Struct Biol. 2004; 14:70–75. [PubMed: 15102452]
6. 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]
7. Plotkin SS, Onuchic JN. Understanding protein folding with energy landscape theory part I: basic concepts. Q Rev Biophy. 2002; 35:111–167.
8. Plotkin SS, Onuchic JN. Understanding protein folding with energy landscape theory Part II: Quantitative aspects. Q Rev Biophy. 2002; 35:205–286.
9. Frauenfelder H, Sligar S, Wolynes P. The energy landscapes and motions of proteins. Science. 1991; 254:1598–1603. [PubMed: 1749933]
10. Krivov SV, Karplus M. Hidden complexity of free energy surfaces for peptide (protein) folding. Proc Natl Acad Sci USA. 2004; 101:14766–14770. [PubMed: 15466711]
11. Allen LR, Krivov SV, Paci E. Analysis of the free-energy surface of proteins from reversible folding simulations. PLoS Comput Biol. 2009; 5:e1000428. [PubMed: 19593364]
12. Rao F, Caflisch A. The protein folding network. J Mol Biol. 2004; 342:299–306. [PubMed: 15313625]
13. Krivov SV, Karplus M. One-dimensional free-energy profiles of complex systems: Progress variables that preserve the barriers. J Phys Chem B. 2006; 110:12689–12698. [PubMed: 16800603]
14. Caflisch A. Network and graph analyses of folding free energy surfaces. Curr Opin Struct Biol. 2006; 16:71–78. [PubMed: 16413772]
15. Gfeller D, De Los Rios P, Caflisch A, Rao F. Complex network analysis of free-energy landscapes. Proc Natl Acad Sci USA. 2007; 104:1817–1822. [PubMed: 17267610]

16. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. J Chem Phys. 2007; 126:155101–155101. [PubMed: 17461665]

17. Noé F, Horenko I, Schütte C, Smith JC. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. J Chem Phys. 2007; 126:155102. [PubMed: 17461666]

18. Noé F, Fischer S. Transition networks for modeling the kinetics of conformational change in macromolecules. Curr Opin Struct Biol. 2008; 18:154–162. [PubMed: 18378442]

19. Buchete NV, Hummer G. Coarse master equations for peptide folding dynamics. J Phys Chem B. 2008; 112:6057–6069. [PubMed: 18232681]

20. Berezhkovskii A, Hummer G, Szabo A. Reactive flux and folding pathways in network models of coarse-grained protein dynamics. J Chem Phys. 2009; 130:205102. [PubMed: 19485483]

21. Krivov SV, Muff S, Caflisch A, Karplus M. One-dimensional barrier-preserving free-energy projections of a β-sheet miniprotein: New insights into the folding process. J Phys Chem B. 2008; 112:8701–8714. [PubMed: 18590307]

22. Muff S, Caflisch A. Kinetic analysis of molecular dynamics simulations reveals changes in the denatured state and switch of folding pathways upon single-point mutation of a β-sheet miniprotein. Proteins. 2008; 70:1185–1195. [PubMed: 17847092]

23. Bowman GR, Pande VS. Protein folded states are kinetic hubs. Proc Natl Acad Sci USA. 2010; 107:10890–10895. [PubMed: 20534497]

24. Jayachandran G, Vishal V, Pande VS. Using massively parallel simulation and Marko-vian models to study protein folding: Examining the dynamics of the villin headpiece. J Chem Phys. 2006; 124:164902–164902. [PubMed: 16674165]

25. Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proc Natl Acad Sci USA. 2009; 106:19011–19016. [PubMed: 19887634]

26. Wales DJ. Energy landscapes: some new horizons. Curr Opin Struct Biol. 2010; 20:3–10. [PubMed: 20096562]

27. Voelz VA, Bowman GR, Beauchamp K, Pande VS. Molecular simulation of ab initio protein folding for a millisecond folder NTL9 (1–39). J Am Chem Soc. 2010; 132:1526–1528. [PubMed: 20070076]

28. Bowman GR, Voelz VA, Pande VS. Atomistic folding simulations of the five-helix bundle protein 6–85. J Am Chem Soc. 2010; 133:664–667. [PubMed: 21174461]

29. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W. Atomic-level characterization of the structural dynamics of proteins. Science. 2010; 330:341–346. [PubMed: 20947758]

30. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How Fast-Folding Proteins Fold. Science. 2011; 334:517–520. [PubMed: 22034434]

31. Beauchamp KA, McGibbon R, Lin YS, Pande VS. Simple few-state models reveal hidden complexity in protein folding. Proc Natl Acad Sci USA. 2012; 109:17807–17813. [PubMed: 22778442]

32. Piana S, Lindorff-Larsen K, Shaw DE. Atomic-level description of ubiquitin folding. Proc Natl Acad Sci USA. 2013

33. Tew GN, Liu D, Chen B, Doerksen RJ, Kaplan J, Carroll PJ, Klein ML, DeGrado WF. De novo design of biomimetic antimicrobial polymers. Proc Natl Acad Sci USA. 2002; 99:5110–5114. [PubMed: 11959961]

34. Nielsen SO, Lopez CF, Srinivas G, Klein ML. Coarse grain models and the computer simulation of soft materials. J Phys Condens Mat. 2004; 16:R481.

35. Klein ML, Shinoda W. Large-scale molecular dynamics simulations of self-assembling systems. Science. 2008; 321:798–800. [PubMed: 18687954]

36. Chebaro Y, Dong X, Laghaei R, Derreumaux P, Mousseau N. Replica exchange molecular dynamics simulations of coarse-grained proteins in implicit solvent. J Phys Chem B. 2008; 113:267–274. [PubMed: 19067549]

37. Liwo A, Oldziej S, Pincus M, Wawak R, Rackovsky S, Scheraga H. A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J Comput Chem. 1997; 18:849–873.

38. Takada S, Luthey-Schulten Z, Wolynes PG. Folding dynamics with nonadditive forces: A simulation study of a designed helical protein and a random heteropolymer. J Chem Phys. 1999; 110:11616.

39. Ding F, Tsao D, Nie H, Dokholyan NV. Ab initio folding of proteins with all-atom discrete molecular dynamics. Structure. 2008; 16:1010–1018. [PubMed: 18611374]

40. Hills RD, Lu L, Voth GA. Multiscale coarse-graining of the protein energy landscape. PLoS Comput Biol. 2010; 6:e1000827. [PubMed: 20585614]

41. Hubner IA, Deeds EJ, Shakhnovich EI. High-resolution protein folding with a transferable potential. Proc Natl Acad Sci USA. 2005; 102:18914–18919. [PubMed: 16365306]

42. Fujitsuka Y, Chikenji G, Takada S. SimFold energy function for de novo protein structure prediction: consensus with Rosetta. Proteins. 2006; 62:381–398. [PubMed: 16294329]

43. Kenzaki H, Koga N, Hori N, Kanada R, Li W, Okazaki K-i, Yao XQ, Takada S. CafeMol: a coarse-grained biomolecular simulator for simulating proteins at work. J Chem Theory Comput. 2011; 7:1979–1989.

44. Hori N, Chikenji G, Berry RS, Takada S. Folding energy landscape and network dynamics of small globular proteins. Proc Natl Acad Sci USA. 2009; 106:73–78. [PubMed: 19114654]

45. Zhang Y, Hubner IA, Arakaki AK, Shakhnovich E, Skolnick J. On the origin and highly likely completeness of single-domain protein structures. Proc Natl Acad Sci USA. 2006; 103:2605–2610. [PubMed: 16478803]

46. Han W, Wan CK, Wu YD. Toward a coarse-grained protein model coupled with a coarse-grained solvent model: Solvation free energies of amino acid side chains. J Chem Theory Comput. 2008; 4:1891–1901.

47. Han W, Wan CK, Jiang F, Wu YD. Pace force field for protein simulations. 1. full parameterization of version 1 and verification. J Chem Theory Comput. 2010; 6:3373–3389.

48. Wan CK, Han W, Wu YD. Parameterization of PACE force field for membrane environment and simulation of helical peptides and helix–helix association. J Chem Theory Comput. 2011; 8:300–313.

49. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH. The MARTINI force field: coarse grained model for biomolecular simulations. J Phys Chem B. 2007; 111:7812–7824. [PubMed: 17569554]

50. Han W, Wan CK, Wu YD. PACE force field for protein simulations. 2. Folding simulations of peptides. J Chem Theory Comput. 2010; 6:3390–3402.

51. Han W, Schulten K. Further Optimization of a Hybrid United-Atom and Coarse-Grained Force Field for Folding Simulations: Improved Backbone Hydration and Interactions between Charged Side Chains. J Chem Theory Comput. 2012; 8:4413–4424. [PubMed: 23204949]

52. Neidigh JW, Fesinmeyer RM, Andersen NH. Designing a 20-residue protein. Nat Struct Mol Biol. 2002; 9:425–430.

53. Jäger M, Zhang Y, Bieschke J, Nguyen H, Dendle M, Bowman ME, Noel JP, Gruebele M, Kelly JW. Structure–function–folding relationship in a WW domain. Proc Natl Acad Sci USA. 2006; 103:10648–10653. [PubMed: 16807295]

54. Liu F, Du D, Fuller AA, Davoren JE, Wipf P, Kelly JW, Gruebele M. An experimental survey of the transition between two-state and downhill protein folding scenarios. Proc Natl Acad Sci USA. 2008; 105:2369–2374. [PubMed: 18268349]

55. Neuweiler H, Doose S, Sauer M. A microscopic view of miniprotein folding: Enhanced folding efficiency through formation of an intermediate. Proc Natl Acad Sci USA. 2005; 102:16650–16655. [PubMed: 16269542]

56. Mok KH, Kuhn LT, Goez M, Day IJ, Lin JC, Andersen NH, Hore P. A pre-existing hydrophobic collapse in the unfolded state of an ultrafast folding protein. Nature. 2007; 447:106–109. [PubMed: 17429353]

57. Jäger M, Nguyen H, Crane JC, Kelly JW, Gruebele M. The folding mechanism of a β-sheet: The WW domain. J Mol Biol. 2001; 311:373–393. [PubMed: 11478867]

58. Deechongkit S, Nguyen H, Jager M, Powers ET, Gruebele M, Kelly JW. *β*-Sheet folding mechanisms from perturbation energetics. Curr Opin Struct Biol. 2006; 16:94–101. [PubMed: 16442278]

59. Juraszek J, Bolhuis P. Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. Proc Natl Acad Sci USA. 2006; 103:15859–15864. [PubMed: 17035504]

60. Marinelli F, Pietrucci F, Laio A, Piana S. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. PLoS Comput Biol. 2009; 5:e1000452. [PubMed: 19662155]

61. a Beccara S, Škrbić T, Covino R, Faccioli P. Dominant folding pathways of a WW domain. Proc Natl Acad Sci USA. 2012; 109:2330–2335. [PubMed: 22308345]

62. Metzner P, Schütte C, Vanden-Eijnden E. Transition path theory for Markov jump processes. Multiscale Model Simul. 2009; 7:1192–1219.

63. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. J Comput Chem. 2005; 26:1781–1802. [PubMed: 16222654]

64. Jiang F, Han W, Wu YD. Influence of Side Chain Conformations on Local Conformational Features of Amino Acids and Implication for Force Field Development. J Phys Chem B. 2010; 114:5840–5850. [PubMed: 20392111]

65. Li AJ, Nussinov R. A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. Proteins. 1998; 32:111–127. [PubMed: 9672047]

66. Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, Mark AE. Peptide folding: when simulation meets experiment. Angew Chem Intl Ed. 1999; 38:236–240.

67. Krivov SV, Karplus M. Free energy disconnectivity graphs: Application to peptide models. J Chem Phys. 2002; 117:10894.

68. Deuflhard P, Weber M. Robust Perron cluster analysis in conformation dynamics. Linear Algebra Appl. 2005; 398:161–184.

69. Prinz JH, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera JD, Schütte C, Noé F. Markov models of molecular kinetics: Generation and validation. J Chem Phys. 2011; 134:174105. [PubMed: 21548671]

70. Senne M, Trendelkamp-Schroer B, Mey AS, Schutte C, Noe F. EMMA: A Software Package for Markov Model Building and Analysis. J Chem Theory Comput. 2012; 8:2223–2238.

71. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES. On the transition coordinate for protein folding. J Chem Phys. 1998; 108:334.

72. Bolhuis PG, Chandler D, Dellago C, Geissler PL. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. Annu Rev Phys Chem. 2002; 53:291–318. [PubMed: 11972010]

73. Zhou R. Trp-cage: Folding free energy landscape in explicit water. Proc Natl Acad Sci USA. 2003; 100:13280–13285. [PubMed: 14581616]

74. Fesinmeyer RM, Hudson FM, Andersen NH. Enhanced hairpin stability through loop design: the case of the protein G B1 domain hairpin. J Am Chem Soc. 2004; 126:7238–7243. [PubMed: 15186161]

75. Piana S, Sarkar K, Lindorff-Larsen K, Guo M, Gruebele M, Shaw DE. Computational design and experimental testing of the fastest-folding *β*-sheet protein. J Mol Biol. 2011; 405:43–48. [PubMed: 20974152]

76. Ensign DL, Pande VS. The Fip35 WW domain folds with structural and mechanistic heterogeneity in molecular dynamics simulations. Biophys J. 2009; 96:L53–L55. [PubMed: 19383445]

77. Rao F, Settanni G, Guarnera E, Caflisch A. Estimation of protein folding probability from equilibrium simulations. J Chem Phys. 2005; 122:184901–184901. [PubMed: 15918759]

78. Qiu L, Pabit SA, Roitberg AE, Hagen SJ. Smaller and faster: The 20-residue Trp-cage protein folds in 4 *μ*s. J Am Chem Soc. 2002; 124:12952–12953. [PubMed: 12405814]

79. Fersht AR. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. Proc Natl Acad Sci USA. 1995; 92:10869–10873. [PubMed: 7479900]

80. Karplus M, Weaver DL. Protein-folding dynamics. Nature. 1976; 260:404–406. [PubMed: 1256583]

81. Krivov SV. The free energy landscape analysis of protein (FIP35) folding dynamics. J Phys Chem B. 2011; 115:12315–12324. [PubMed: 21902225]

82. Weikl TR. Transition States in Protein Folding Kinetics: Modeling Φ-Values of Small β-Sheet Proteins. Biophys J. 2008; 94:929–937. [PubMed: 17905840]

83. Petrovich M, Jonsson AL, Ferguson N, Daggett V, Fersht AR. F-analysis at the experimental limits: mechanism of β-hairpin formation. J Mol Biol. 2006; 360:865–881. [PubMed: 16784750]

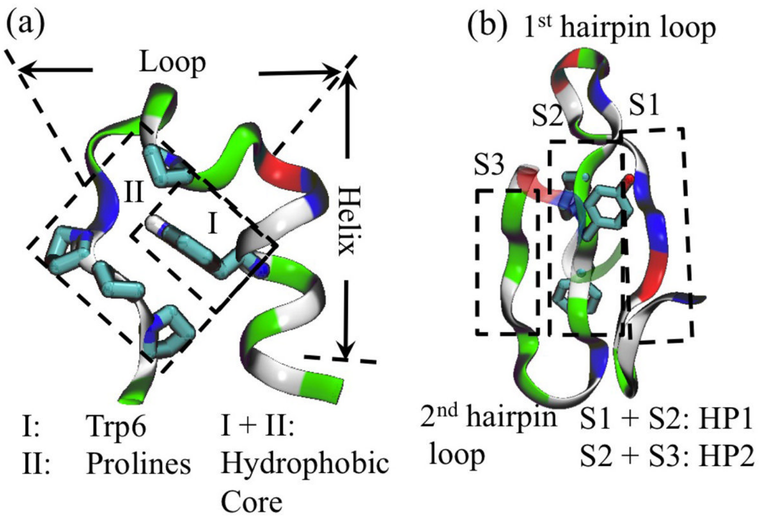

Figure 1.

Native structures of trp-cage (a) (PDB ID: 1L2Y) and hPin WW (b) (PDB ID: 1PIN). For both proteins, backbones are shown in ribbon representation. Hydrophobic, polar, basic and acidic amino acids are shown in gray, green, blue and red, respectively. Side chains of Pro12, Pro17–19 and Trp6 in (a) and those of Y17, Y18 and F19 (b) are shown, respectively, in stick representation (color: light blue is carbon, red is oxygen, dark blue is nitrogen, white is hydrogen).

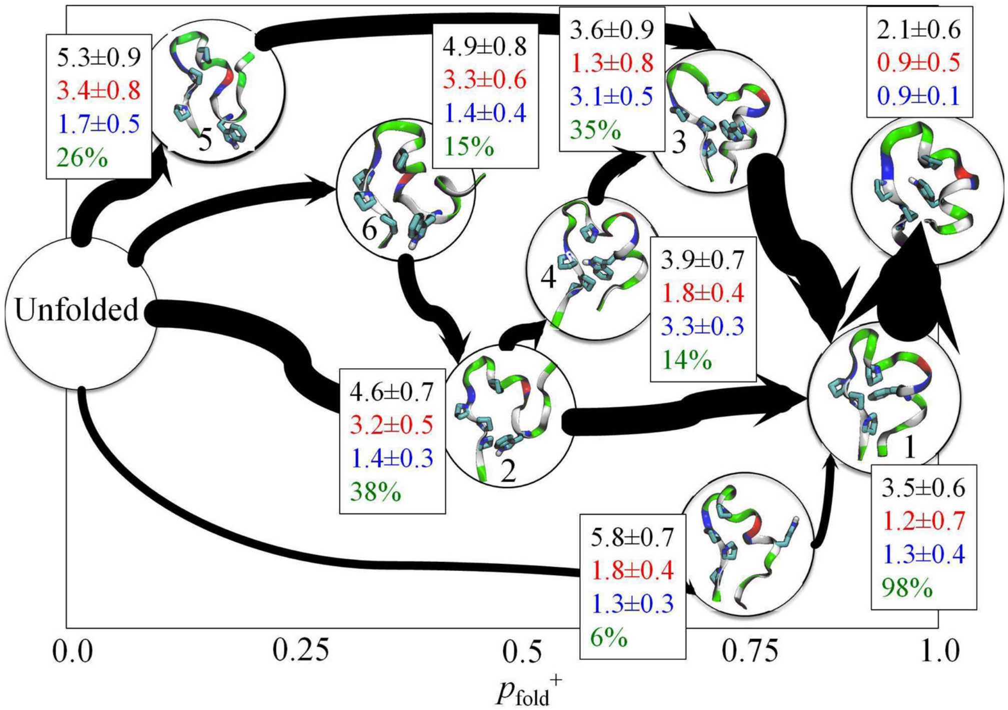

Figure 2.

Major folding pathways of trp-cage accounting of 30% folding flux. Shown are the representative structures of on-pathway intermediates. They are plotted against the committor probability p_{fold}. Next to each representative structure are stated for each intermediate the average RMSD to the native structure of the hydrophobic core (black), the N-terminal helix (red) and the loop (blue); the on-pathway probability (p_{onpath}) of the same intermediate is shown in green. The thickness of the curved arrows is proportional to the folding flux. Representative structures are shown as in Fig 1.

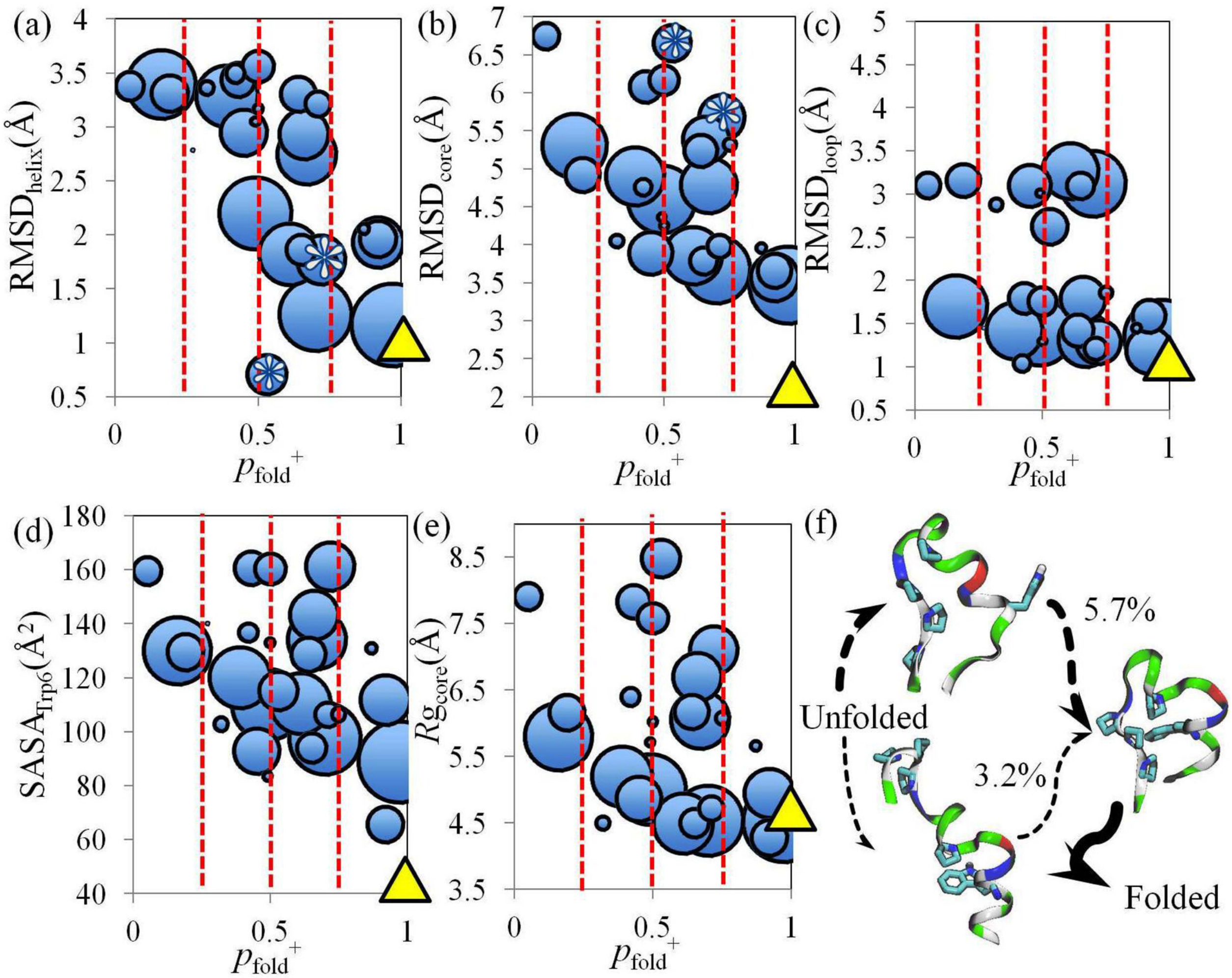

Figure 3.

Evolution of key structural properties during the course of the folding of trp-cage. Shown are average RMSD to the native structure of the N-terminal helix, the hydrophobic core and the loop (a)–(c), the solvent accessible surface area of the side chain of Trp6 (d) and the radii of gyration (R_g) (e). The quantities in (a)–(e) are plotted against p_{fold} for the intermediates with $p_{onpath} > 0.01$. The radii of spheres in (a)–(e) are proportional to ln p_{onpath}. The triangles indicates the same properties of the native states. (f) Folding pathways in which the N-terminal helix forms first. The two helical intermediates are indicated by an asterisk in (a) and (b). The dashed arrows indicate that there could be other intervening intermediates. Representative structures in (f) are shown as in Fig 1.

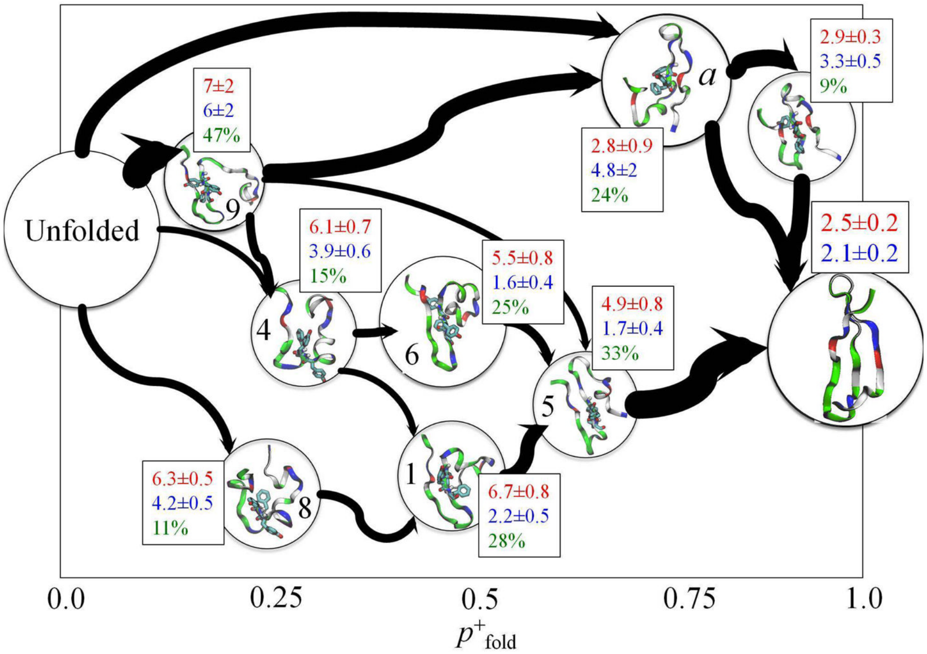

Figure 4.

Major folding pathways of the hPin WW variant accounting for 25% folding flux. Shown are the representative structures of on-pathway intermediates. They are plotted against the committor probability p_{fold}. Next to each representative structure are given the average RMSD of hairpin 1 (red), and hairpin 2 (blue) (relative to the corresponding part of the native structure) and the on-pathway probability (p_{onpath}) of the same intermediate (green). The thickness of the curved arrows is proportional to the folding flux. Representative structures are shown as in Fig 1.

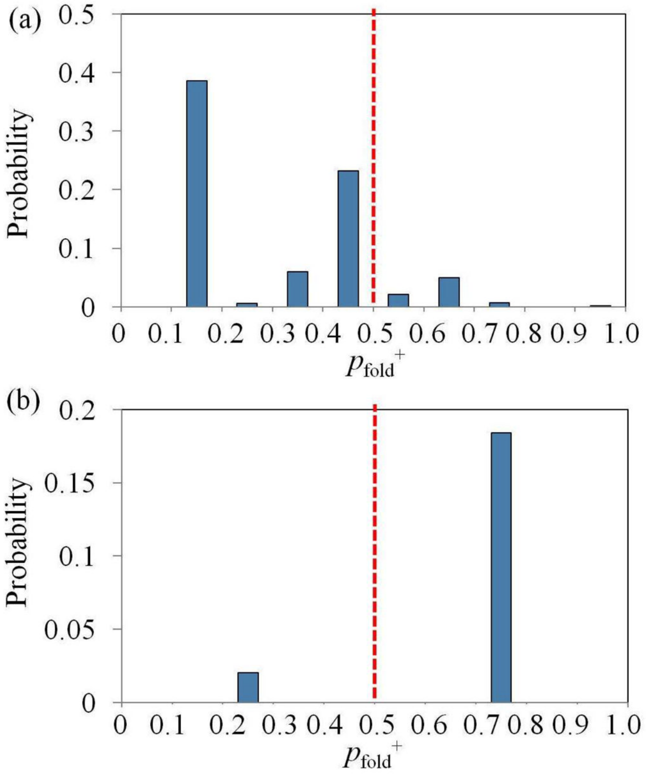

Figure 5.

Distributions of p_{fold} for the folding of the hPin WW variant earliest on-pathway intermediates in which one of the two hairpins forms. (a) Distribution for the pathways in which hairpin 2 is the first to form. (b) Distribution for the pathways in which hairpin 1 is the first to form.

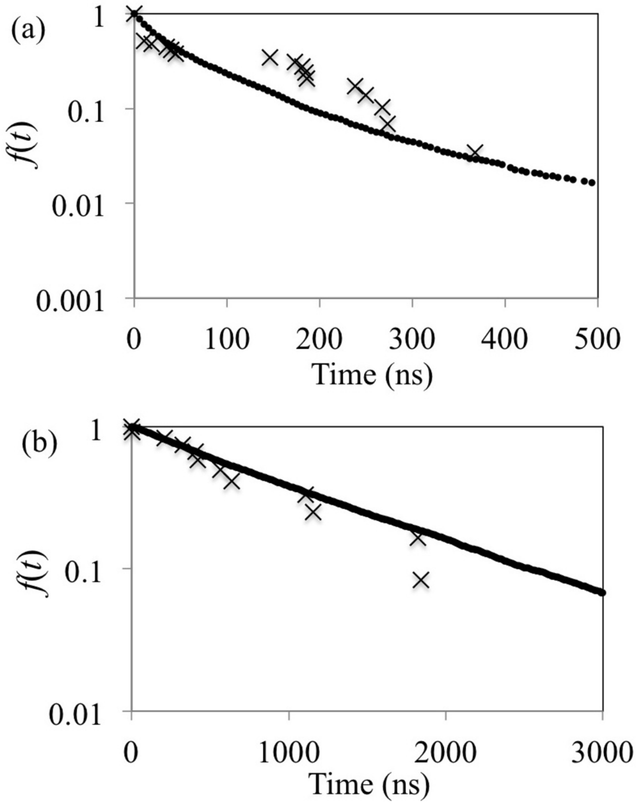

Figure 6.

Cumulative distribution (f) of the first passage times of folding of trp-cage (a) and of folding of the hPin WW variant (b), with $f(t)=\int_t^{\infty} p(\tau)d\tau$, where $p(\tau)$ is the probability distribution of the first passage time. Crosses indicate the distributions obtained from MD trajectories; dotted lines indicate the distributions obtained from MC simulations based on transition probability matrices.

Table 1

Comparison of observed probabilities (p_{obs}), average root mean square distances (RMSD) to the native structures, probabilities (p_{onpath}) of being on the folding pathways and mean first passage time (MFPT) of transitions to the native states for the most populated non-native macro-states of trp-cage.

Macro-state ID	p_{obs}	RMSD (Å)	p_{onpath}	MFPT (ns)
1	0.167	3.4±0.6	0.98	9.3±0.4
2	0.070	5.4±0.9	0.38	94±2
3	0.040	4.4±0.6	0.35	48±1
4	0.025	4.4±0.6	0.14	55±1
5	0.022	5.4±0.8	0.26	111±2
6	0.021	6.0±0.7	0.15	120±2

Table 2

Comparison of observed probabilities (p_{obs}), average root mean square distances (RMSD) to the native structures, probabilities (p_{onpath}) of being on the folding pathways and mean first passage time (MFPT) of transitions to the native states for the most populated non-native macro-states of the WW-domain variant.

Macro-state ID	p_{obs}	RMSD (Å)	p_{onpath}	MFPT (ns)
1	0.032	6.6±0.7	0.28	571±7
2	0.031	3.5±0.1	0.06	20±1
3	0.031	3.7±0.2	0.11	36±2
4	0.030	6.7±0.6	0.15	722±7
5	0.029	4.8±0.9	0.33	377±6
6	0.025	5.4±0.7	0.25	524±7
7	0.025	4.4±0.3	0.02	334±6
8	0.023	6.8±0.5	0.11	732±7
9	0.018	9±2	0.47	744±7