Sampling the multiple folding mechanisms of Trp-cage in explicit solvent

J. Juraszek and P. G. Bolhuis*

Van 't Hoff Institute for Molecular Sciences, Universiteit van Amsterdam, Nieuwe Achtergracht 166, 1018 WV, Amsterdam, The Netherlands

We investigate the kinetic pathways of folding and unfolding of the designed miniprotein Trp-cage in explicit solvent. Straightforward molecular dynamics and replica exchange methods both have severe convergence problems, whereas transition path sampling allows us to sample unbiased dynamical pathways between folded and unfolded states and leads to deeper understanding of the mechanisms of (un)folding. In contrast to previous predictions employing an implicit solvent, we find that Trp-cage folds primarily (80% of the paths) via a pathway forming the tertiary contacts and the salt bridge, before helix formation. The remaining 20% of the paths occur in the opposite order, by first forming the helix. The transition states of the rate-limiting steps are solvated native-like structures. Water expulsion is found to be the last step upon folding for each route. Committor analysis suggests that the dynamics of the solvent is not part of the reaction coordinate. Nevertheless, during the transition, specific water molecules are strongly bound and can play a structural role in the folding.

protein folding | reaction coordinate | transition path sampling | replica exchange | transition state ensemble

Elucidating the mechanism by which proteins fold into their native state remains a central issue in molecular biology. For single domain two-state folding proteins, several decades of experimental, theoretical, and simulation studies have revealed two major qualitative folding mechanisms. In the diffusion–collision mechanism (1), proteins first form secondary structure elements followed by a diffusive search toward the tertiary native state structure. In the nucleation–condensation mechanism (2), a nucleus of crucial tertiary contacts is made, around which the native structure condensates. In recent years, these two mechanisms were combined in a unified view (3).

By bridging the gap between experiments and computer simulation, the discovery of small and fast folding proteins has contributed much to the understanding of generic folding mechanisms. The fastest of those is the designed 20-residue miniprotein Trp-cage (NLYIQ WLKDG GPSSG RPPPS) (4), which folds in 4 μs to a native state with an α-helix, a salt bridge, and a polyproline II helix shielding the central tryptophan from solvent. Laser temperature-jump spectroscopy experiments by Qiu *et al.* (5) indicated two-state folding. Subsequently, fluorescent correlation spectroscopy by Neuweiler *et al.* (6) revealed that the protein (un)folds in a more complicated manner via an intermediate molten globule-like state, characterized by exposure of the tryptophan to the solvent. It remains unclear at what stage of folding the helix is being formed. Recent UV-resonance Raman spectroscopy measurements show some evidence of a helical structure in the denatured state of Trp-cage, and thus suggest that an early formation of the helix is possible (7). Many molecular dynamics (MD) simulations were performed to investigate thermodynamic stability of the protein and elucidate possible folding pathways by using, for instance, all-atom models with implicit solvent (8–12) as well as explicit solvent (13) or simplified models such as Go models (14). Replica exchange MD (REMD) simulations in explicit solvent (13) confirmed the two-state nature of Trp-cage. An intermediate state structure, containing two hydrophobic cores, was proposed as the reason of Trp-cage being such a fast folder. Folding events of

Trp-cage have been observed in all-atom implicit solvent MD simulations (8, 9, 12) and in a coarse-grained model (15). Misfolded states identified in implicit solvent calculations suggested that Trp-cage is a less reliable and efficient folder than was previously assumed (12). Few simulation studies exist that aim to understand the dynamics of the mechanism, and employ either implicit force field (FF) (8) or a simplified model (14). Recent work suggests that the solvent does play a crucial role in protein folding, one that current implicit solvent models are not able to capture (16). Therefore, it is important to study the dynamical trajectories of folding in which the solvent molecules are treated explicitly. Here, we show that the Trp-cage (un)folding path ensemble consists of two main routes. In contrast to the predictions in ref. 14, we find that 20% of the paths first form the helix, whereas 80% first form the tertiary contacts. It appears that Trp-cage folds via both diffusion-collision and nucleation-condensation (1–3). Existence of alternative pathways was also found in larger proteins, e.g., lysozyme (17). In addition, we predict the nature of the transition states (TS) and investigate the influence of solvent motion on the reaction coordinate during the folding, something that can only be studied with the use of an accurate FF for protein and solvent.

It is still a computational challenge to perform all-atom MD simulations of folding of any protein in explicit solvent, including Trp-cage, because even the microsecond folding time scale is only barely reachable for such simulations. The required long time scales are due to the presence of relatively high free energy (FE) barriers between the native and unfolded states. Protein folding can thus be seen as a rare event. Many computational methods exist that aim to overcome such FE barriers. A direct way of accessing kinetics is to initiate many simultaneous simulations, of which a small percentage succeed in crossing the barrier (8). Using for instance Markovian state models, one is then able to study complex reaction networks (18). Other approaches try to overcome the barriers by increasing the temperature [high-temperature MD (19), temperature accelerated dynamics (20), replica exchange/parallel tempering (21)]. In these methods, accurate information about the rare event at room temperature is lost, because transitions only occur at high temperatures. Therefore, many other methods employ biasing potentials to enhance conformational sampling at room temperature [e.g., umbrella sampling (21), metadynamics (22), hyperdynamics (23), and flooding (24)]. Although most of these methods are well suited for computing thermodynamic properties, they do not preserve the dynamical pathways due to the bias they introduce, and hence cannot be used to obtain accurate kinetics and mechanisms in complex systems. Moreover, because the biasing potential must be applied as a function of an order parameter, the important reaction coordinates must be known *a priori* to enable proper

---

BIOPHYSICS

sampling. However, a poor choice of reaction coordinate in these methods leads to bad statistics, a wrong mechanism, and overestimation of the rate constants. To address this "reaction coordinate" problem, Chandler and coworkers (25–27) developed the transition path sampling (TPS) methods, a suite of techniques that enable the collection of an ensemble of transition paths between an initial and final state, without prior knowledge of the TS region. The TPS algorithm samples short trajectories (of several nanoseconds) connecting folded and unfolded states of the protein at room temperature.[†] In contrast, a straightforward molecular dynamics trajectory would take many microseconds to achieve the same. The TPS method also allows evaluation of the mechanism, TS ensemble, and rate constant. Recently, a more efficient TPS-based algorithm was introduced to evaluate the rate (28). Path sampling was successfully applied to investigate the kinetics of the folding of the GB1 β-hairpin (29, 30). The path sampling method requires order parameters to describe and distinguish the initial and final stable states. We find these order parameters by calculating the FE landscapes as a function of several variables via REMD.

Preparation of the Native State

The Trp-cage system was prepared from the NMR structure from the Protein Data Bank (entry 1L2Y). The 304-atom protein was solvated with 2,797 water molecules in a rhombic dodecahedral periodic box with a 5-Å diameter. One water molecule was replaced with a Cl^- counterion to neutralize the system. All MD simulations in this work were performed by using the Gromacs molecular simulation package (31), in most cases with the OPLSAA FF (32) and SPC water model (31). Imposing bond constraints, by LINCS (33) for protein atoms and by SETTLE (34) for water, allowed for a time step of 2 fs. Long-range electrostatics was treated with Fast Particle-Mesh Ewald (35, 36) with a grid spacing of 1.2 Å. The van der Waals interactions, truncated and shifted, decayed smoothly to zero between 6 and 9 Å. After energy minimization and a short protein position restraint run, the sample was equilibrated for 1 ns at a constant pressure of 1 bar and a temperature of 300 K using Berendsen coupling. From then on, all simulations were performed in the NVT ensemble, with a box size of 5.02 Å, corresponding to ambient pressure at 300 K. The Nosé–Hoover thermostat (37, 38) ensured a constant temperature with a time constant of 0.2 ps. The system was equilibrated for 10 ns at 300 K. To investigate the effect of the FF on the stability of the native state, we extended this run for another 10 ns for both the GROMOS96 43a1 (39) and OPLSAA FFs (32). The G43a1 FF trajectory did not preserve the native structure of Trp-cage, and within a few nanoseconds we observed a spontaneous rearrangement of the protein structure, including partial solvation of Trp-6 (see Fig. 1). We concluded that the global minimum for the G43a1 FF does not coincide with the NMR structure (4). In contrast, although, for the OPLSAA FF, minor fluctuations occurred around the 3_{10}-helix region, caused by a single water molecule entering the Trp-pocket, the protein was stable and stayed within 1.5 Å C_α rmsd of the NMR structures over the entire trajectory of 10 ns. Therefore, we decided to use the OPLSAA FF in our simulations.

Order Parameters

In our simulations we monitor the following order parameters: the protein radius of gyration using the α-carbons (rg), the rg including the side-chain atoms (rg_{sch}), the fraction of native contacts (ρ), the root mean square deviation (rmsd) from the native α-carbons structure ($rmsd$), the rmsd of the α-helical residues (2–8) from an ideal helix ($rmsd_{hx}$), the solvent accessible surface (sas) of the whole protein (31), the salt-bridge distance (sb) defined as the minimum distance between donors and acceptors in the hydrogen bond
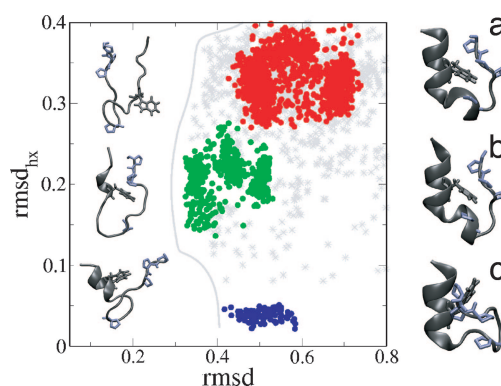

[†]Other path-sampling techniques exist (e.g., ref. 51) but are often based on unrealistic dynamics.

Fig. 1. Trp-cage conformations occurring in the MD and REMD simulations. (*Right*) PDB NMR structure (*a*) compared with typical simulated native state structures at room temperature for the OPLSAA (*b*) and G43a1 (*c*) FFs. The protein is plotted in cartoon representation. Pro-12, -17, -18, and -19 and Trp-6 are plotted in stick representation. Images were made with VMD (50). (*Left*) Three groups of clusters found in the unfolded initiated REMD ensemble. The top cluster (red, 45%) is a set of several loop structures, with Trp-6 fully exposed to the solvent. The middle cluster (green, 15%) consist of the loop structures with Trp-6 positioned in between prolines. The bottom cluster (blue, 3%) denotes fully helical structures. Structures plotted as gray stars belong to clusters that did not reach the minimum abundance of 2%.

between Arg-16 and Asp-9, and the number of water molecules within 4 Å around Trp-6 (nw_{trp}). We use these order parameters to construct FE diagrams and subsequently extract stable state definitions for TPS.

REMD

The advantage of REMD with respect to biasing rare event methods is that no *a priori* knowledge of the reaction coordinates is required. However, a known problem with the method is that it is difficult to decide whether the simulation at a given temperature has already reached the canonical distribution (40). Reaching the Boltzmann distribution in a local part of conformational space does not guarantee that the sampling over the entire configuration space is satisfactory. One way of assessing the sampling quality is to initiate two independent simulations with significantly different protein structures, and test whether the two simulations converge to the same equilibrium distribution and corresponding FE surfaces (40).

We performed two independent 64-replica REMD simulations of Trp-cage in explicit solvent (13). The first REMD simulation series (REMD-fol) was started from the final configuration of the equilibration run of the native state. The unfolded REMD series (REMD-unf) was initiated from an extended structure, selected from a 4-ns high-temperature (500 K) MD trajectory in which the protein unfolded. The temperature distribution, ranging from 272 K to 555 K, was optimized to yield a constant acceptance probability of $\approx$20–30% for exchanging adjacent replicas (41, 42). We allowed for exchange every 0.5 ps (250 MD steps), and saved protein structures with the same frequency. The REMD-fol simulation was performed for 13 ns per replica (a total integrated time of 0.8 μs). The REMD-unf series was run 3 times longer (36 ns, an integrated time of 2.3 μs) because we did not observe convergence with the folded simulation and, more importantly, the native state was not reached. In the data analysis, we excluded the first 5 ns to allow for relaxation toward equilibrium.

The FE landscape of the REMD-fol simulation plotted in the (rg, ρ) plane roughly agrees with the results by Zhou (13) (see Fig. 2). In contrast, the REMD simulation initialized from the unfolded state did not converge to the same FE surfaces. The lack of convergence is clearly visible in the (rg, ρ) representation of Fig. 2.
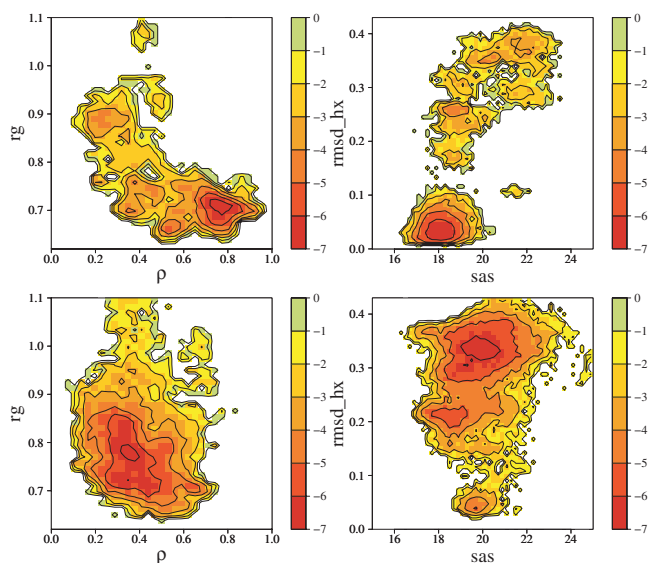
Fig. 2. FE contour maps at 300 K for folded-initiated REMD (*Upper*) and unfolded-initiated REMD (*Lower*). The contours are separated by 1 kT. (*Left*) Radius of gyration (*rg*) versus the ratio of native contacts (ρ). (*Right*) Helix root mean square deviation ($rmsd_{hx}$) versus solvent accessible surface (*sas*).

Minima in the FE plots do not always correspond to a single stable state. Depending on the order parameters, overlap can occur between states, and hide a (meta)stable state inside another one. To identify the stable states, we performed cluster analysis of our REMD ensembles. We used the full-linkage clustering algorithm (31) with a cutoff of 3 Å and with the distance between two structures defined as their relative rmsd. Application of this procedure to the unfolded initiated REMD ensemble resulted in ≈150 clusters. The first and largest cluster (45%) contains a variety of different, strongly twisted hairpin like structures and bent loop structures, all having Trp-6 fully exposed to the solvent. The second cluster (15%) comprises U-shaped structures, with the tryptophan oriented correctly and packed in the center of the protein and one turn of the helix formed. The remainder of the clusters was lumped together and divided into two groups: those clusters containing a helix and those without proper helix. We performed this procedure because the helical cluster could not correctly be recognized by the clustering algorithm due to large fluctuations of the coil part. The first group (3% of the REMD ensemble) consists of fully helical structures with the polyproline detached from the rest of the hydrophobic part, whereas the rest (37%) are molten globular structures with no distinct features.

These three different groups are depicted in Fig. 1 and are also visible as FE minima in the $rmsd_{hx}$ vs. *sas* plot in Fig. 2, suggesting two routes for the folding process. In the first mechanism, loop formation is followed by packing of the tryptophan between the proline residues. The second route would go via helix formation and subsequent packing of Trp-6 in the core. Both pathways seem possible according to the REMD-unf data, and the FE barriers from the unfolded state toward both intermediates don not appear to be very high. However, we did not find any completely folded protein structure within 30 ns per replica. Such slow convergence of REMD is not unprecedented. For instance, Rhee and Pande (40) found no proper convergence within 50 ns for BBA5, a protein with a size comparable to Trp-cage. Results from our TPS simulations (see below) indicate two reasons for the convergence issues in the REMD-unf runs. First, a substantial free-energy barrier separates the native state from the intermediates. Second, this barrier is very diffusive, and crossing it can take up to 2–3 ns. Even if a particular structure would have reached the TS (not observed

Table 1. Order parameters (OP) defining the upper (max) and lower (min) boundaries of the stables state

OP	A_{min}	A_{max}	B_{min}	B_{max}
rmsd, nm	0	0.25	0.45	0.8
$rmsd_{hx}$, nm	0	0.05	0	1
sas, nm^2	17	18.5	0	30
ρ	0.75	0.90	0.20	0.50
nw_{trp}	0	7	12	25

in our REMD-unf simulation), the shorter residence time in the low-temperature replicas would prevent the folding into the native state.

Because of the above reasons, REMD cannot be used for accurate estimating FEs. Nevertheless, the REMD simulations allowed for the definition of the initial and final state for TPS and simultaneously yielded insight in some important on-pathway intermediate structures for the (un)folding process.

Transition Path Sampling

We use the stochastic, flexible path length version of the shooting algorithm. The Andersen thermostat ensured the stochastic nature of the MD trajectories (see *Supporting Text* and Fig. 5, which are published as supporting information on the PNAS web site, for details). The TPS algorithm relies on a proper definition of the stable states. This requires order parameters that are able to distinguish between the stable states and at the same time are representative for these states (27, 29, 30), i.e., when the system is in a stable state, the order parameter is frequently within the definition (at least every few picoseconds). Although the path length depends on the definition of the stable states, the path ensemble itself is rather insensitive to this, as long as the definition of a stable state is within the basin of attraction of that state (28).

We obtain insight in these order parameters from straightforward MD trajectories and REMD simulations. The stable states were defined by the order parameters listed in Table 1. Although the folded state (A) is defined rather rigorously, we do not use several order parameters to define the final state (B). The $rmsd_{hx}$ is excluded because we did not *a priori* know at what stage of unfolding the helix dissolves. Leaving out the *sas* results in including molten globular structures with low *sas* into the final state. The final region B includes both intermediate states found in the REMD simulation to avoid unnecessary long pathways in the TPS simulations.

We initialized TPS with a high-temperature (500 K) unfolding trajectory connecting the folded and unfolded state. Subsequent equilibration of the path ensemble at room temperature (300 K) in four different simulation runs resulted in four entirely independent trajectories, which we used as input for 20 independent TPS simulation runs. The number of trial shootings was 3,200, and the acceptance ratio was ≈44%. This production part of the TPS procedure required 8.5-μs simulation time, and resulted in an ensemble of 3,200 pathways, on average 3.1 ns long (representing an aggregate simulation time of 10 μs), of which ≈100 are independent uncorrelated trajectories. Thus, for Trp-cage, TPS is around two orders of magnitude more efficient than a straightforward MD simulation (apart from the trivial parallelizability of TPS). The distribution of path lengths is quite broad, as expected (27) (see *Supporting Text* and Fig. 5).

Fig. 3 plots a few selected paths in different representations. The paths broadly follow two qualitatively different folding routes,[‡] which are also schematically depicted. Starting from the interme-

---

[‡]Note that as the trajectories are completely time reversible the A and B labels are arbitrary. The final path ensemble is equally valid for the unfolding as well as for the folding process [only between the predefined states, of course. Other possible folding processes (e.g., to misfolded states) are most unlikely to be picked up].
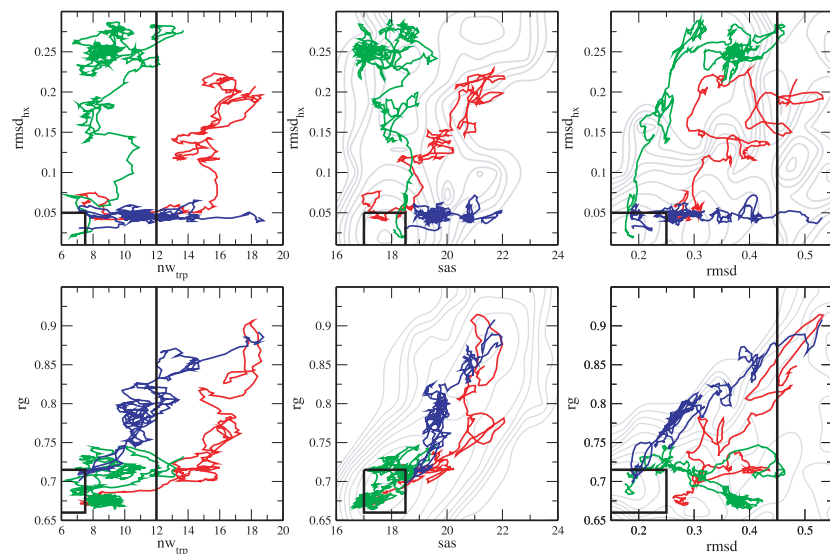
BIOPHYSICS

Fig. 3. Results of the TPS simulations. (*Left*) Three trajectories (of 3,200) in different representations. Blue lines indicate a representative of the *NP_dI* pathway. Green lines indicate the *NL* pathway. Red lines indicate an intermediate ''switching'' pathway. Underlying REMD FE contours are smoothed and plotted in gray. (*Right*) the two main folding scenarios schematically depicted in the *rmsd–rmsd_hx* plane. Protein backbone structures were rendered in cartoon, hydrophobic residues in licorice and waters within 4 Å of the tryptophan in space-filling representations. At the first stage of folding (*U*), key hydrophobic contacts are formed: either Pro-12 or Pro-17/18/19 collapse on the helical residues. The first scenario gives rise to the formation of two hydrophobic clusters. The α-helix appears quickly in the larger of the two (*I*) and eventually the smaller hydrophobic cluster approaches Trp-6 (*P_d*). When this is accomplished, the protein finds its native state (*N*) relatively easily. In the second scenario, a loop structure (*L*) with correct tertiary contacts precedes the formation of the α-helix. All scenarios show water expulsion step in the last stage of folding.

diate states, both pathways reach the native state within a few nanoseconds. The barrier between the initial and final states is substantial, as we observe no spontaneous transitions in a straightforward MD simulation at room temperature, nor in the REMD-unf simulations. Below we describe both (un)folding routes, first starting from the native state.

All of the trajectories begin their unfolding by changing the position of Pro-12 with respect to Trp-6. This process results in a partial exposure of the tryptophan to the solvent and initiates the solvation process. If the water starts to penetrate the hydrophobic region between Trp-6 and Pro-12, resulting in a thread of waters going through the protein, then the protein will end up in the loop state *L*. On the other hand, if hydrogen bonds form between residues Gly-10–Gly-11 and Ser-13–Ser-14, thus impeding solvation of this region, solvation between Trp-6 and Pro-18 occurs instead and results in the detachment of the polyproline helix. Before this happens, the protein can stay in the ''proline detached'' (*P_d*) state for up to 1 ns. The first group of pathways (*NL*) preserves the compact form of the protein upon the desolvation of the α-helix and Trp-6. When Trp-6 is solvated, the remaining steps to complete unfolding are relatively small. The second group of pathways (*NP_dI*) is characterized by the breaking of the hydrophobic core, resulting in the formation of two hydrophobic clusters separated by a layer of water molecules. We do not see a significant decrease in helicity in this group of pathways. A characteristic of the *L* and *I* state is the presence of the salt bridge. Our TPS ensemble shows that the breaking of the salt bridge toward complete unfolding *U* can occur spontaneously with a probability of 9.2% for the *NI* pathways and 28% for the *NL* pathways. These high percentages show that the salt bridge is, in fact, not so stable. Indeed, we find that in both REMD and MD, it also forms and breaks spontaneously. This finding indicates that the *NL* and *NI* transitions contain the rate-limiting steps.

The folding pathways occur in the reversed order. Starting from the U state the protein undergoes either a fast initial collapse or shows helix formation. When the protein forms the loop structure *L*, it is most probably that Trp-6 will point to the solvent. According
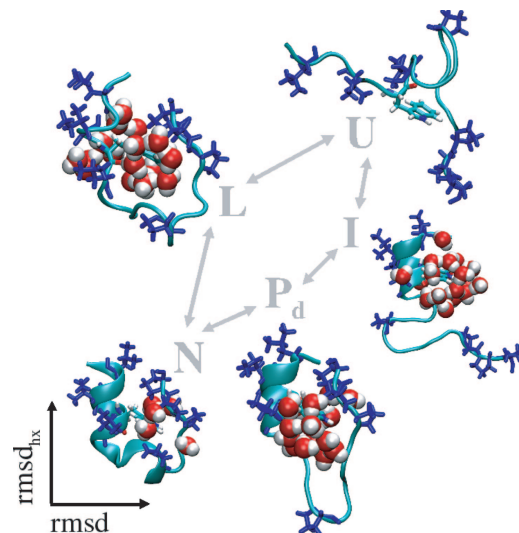
to our REMD-unf simulations, this structure resembles a β-hairpin and is stabilized by tertiary contacts and hydrogen bonds. The barrier to incorporate tryptophan in the protein center seems of the order of 1 kT. If this happens, the protein follows the *LN* route. On the other hand, when at the beginning of the folding process the helix is formed, the protein will follow the *IN* route. There is a small probability (4% in the REMD-unf ensemble) of misfolding to structures with a salt-bridge on the opposite side of the protein.

We observe ''switching'' events in the TPS simulations, i.e., changes between the *NL* to *NP_dI* pathways or vice versa. These switches tend to take place via an intermediate *NP_dL* pathway (see Fig. 3), followed by quick equilibration to the other main route. The conditional probability of going from *NP_dI* to *NL* is ≈4 times higher (0.029) than the other way around (0.007). These probabilities agree with the occurrence of four times more *LN* as *IN* pathways in the ensemble. This finding suggests that the path ensemble is equilibrated and the influence of the initial starting pathway has vanished. Another indication of convergence of the path ensemble follows from the fact that the least-changed pathway (see supporting information) covers a wide region in the *rmsd–rmsd_hx* projection.

We stress that the TPS procedure selects the *L* and *I* structures from a much broader set of allowed configurations within the final state (see Table 1). More importantly, these two states were also found in the REMD-unf simulations (Fig. 1) as helical and loop structures. This finding shows that the *L* and *I* states are the only relevant states for the folding starting from the molten-globule and U-shaped clusters with the solvent exposed Trp-6. Application of TPS allowed the crossing of the highest (un)folding barrier and connected the native state and the unfolded ensemble. In contrast, this barrier was not crossed within the 64 × 36 ns REMD simulation time.

Our results agree with experiments, because existence of two pathways can explain both signs of helical content in the early stages of folding (7) and the structurally restricted intermediate (6) (*L* state). Comparing our study to earlier simulations, we recognize the intermediate *I* state on the *NP_dI* pathway as the intermediate found

by Zhou (13), but we find that the proposed folding scenario was not complete. Identifying the *NL* pathway may have been difficult in the REMD ensemble because of the order parameter issues (see Fig. 3). In our study, we show this is very straightforward when examining dynamical trajectories.

Water Dynamics

Comparison of the two routes reveals significant differences in the water dynamics. Examination of the residence times of water molecules in the vicinity of the surface of the protein along the trajectories in the ensemble shows that 90% of all water-protein contacts exist <50 ps. Water molecules in close proximity to the Trp-6 NH group exhibit a longer residence time, exceeding 100 ps in 25% of the cases. For the *NL* route, the carbonyl oxygen of Trp-6 can strongly bind water molecules, with 25% of the residence times longer than 100 ps. These trapped waters are hydrogen bonded to the backbone oxygens of both the tryptophan and the glycine, thus connecting two secondary structure elements: the α-helix and the 3_{10}-helix. This double-hydrogen bonding increases the residence time.

In the case of the NP_dI route, water cannot bridge tryptophan and the glycine after the hydrophobic collapse, because the 3_{10}-helical part is separated from the α-helix, leading to only 2% of the residence times above 100 ps. These findings show that the explicit presence of solvent molecules clearly influences the water dynamics, and hence the folding kinetics, something that cannot easily be captured by an implicit solvent.

TS Calculation

To gain insight in the nature of the TS ensemble, we calculated the commitment probabilities (a.k.a., committor, or p_{fold}) for trajectories between the native and loop states (*NL* transition) and for the polyproline detachment (P_dI transition). We identify configurations with the same committor for both folding and unfolding as TS (43). A few of such TS structures are shown in the Fig. 4. These structures are in fact very similar to the native state, but have a rather wide range (50–80%) of native contacts. For the (*NI*) path, previously estimated TS structures based on a Go-model only have 50–60% of the native contacts (14).

Furthermore, in the *NL* TS structure (Fig. 4b) the helix is mainly dissolved ($rmsd_{hx} \approx 1.5$ Å). The tryptophan is centered in the middle of the structure surrounded by a shell of water molecules, separating it from other hydrophobic amino acids. The TS structures of the P_dI pathway (Fig. 4b) reveal a polyproline helix perpendicular to the surface of the tryptophan aromatic ring, separated by a single layer of water molecules. In this case, Trp-6 is also fully solvated but without a chain of hydrogen bonded water molecules underneath it. Fig. 4a summarizes the identified stable and TS in a scatter plot. Clearly, the $rmsd_{hx}/rmsd$ order parameters distinguish between the different states of the protein, whereas the ρ/rg_{sch} order parameters show a substantial overlap and separate only the native state. Moreover, the location of the TS is hidden in a stable state, making the ρ/rg_{sch} order parameters unfit to act as reaction coordinates.

The gray curve demarcates the projected accessible region of the REMD-unf simulation. However, even REMD structures coinciding with the TS in this projection still had a relative RMSD of >2 Å with respect to the true TS structure. Assuming that the REMD boundary line approximately delimits the highest barriers, we conclude that the last step to folding, the water expulsion transition, is indeed the rate-limiting step in both Trp-cage folding routes. In an early simulation study of barnase in explicit solvent, Caflisch and Karplus (44) showed that water entrance is a key step in unfolding. A water-expulsion transition occurring during the packing of the hydrophobic core was also observed in other computational studies (44–46). Interestingly, it has not been observed in folding simulations BBA5, a small protein forming a hydrophobic core (16).
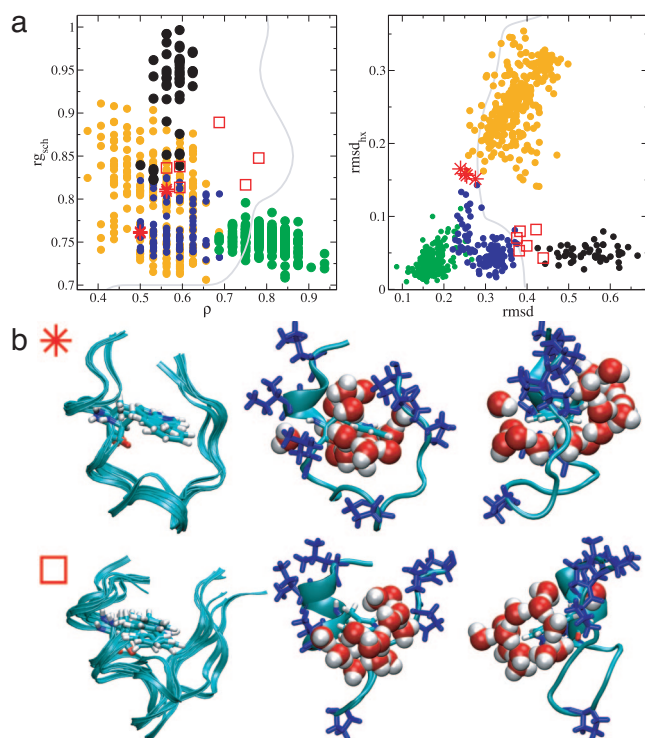

Fig. 4. Results of the TS analysis. (*a*) Stable states and TS obtained by committor calculation, plotted as rg_{sch} versus ρ (*Left*) and $rmsd$ versus $rmsd_{hx}$ (*Right*). TS for the *NL* and NP_d transitions are plotted as stars and squares, respectively. The scatter points are taken from the corresponding TPS trajectories. Native states (*N*) are plotted in green, loop structures (*L*) in orange, close to native structures with Pro-12 detached (P_d) in blue and the *I*-state in black. The region to the left of the gray line on the $rmsd_{hx}/rmsd$ (to the right on the ρ/rg_{sch}) plot is the part of the configuration space not reachable by our REMD-unf simulation. (*b*) (*Left*) Superimposed TS structures for four different *NL* (star) and six for P_dI paths (square). (*Center* and *Right*) One of the TS structures for both routes and its side view plotted in cartoon representation. Hydrophobic residues are plotted in licorice representation (blue), water molecules within 4 Å of Trp-6 are plotted in space-filling representation.

Reaction Coordinate

Rhee and Pande (47) examined the influence of the instantaneous water structure on the committor to estimate the contribution of the water dynamics to the reaction coordinate. We perform a similar calculation for Trp-cage by freezing the protein coordinates and randomizing the water structure. If the committor for these randomized structures deviates systemically from 0.5, the solvent dynamics is an important ingredient of the reaction coordinate. We randomized the solvent by performing MD at 400 K with a frozen protein, followed by a 1-ns equilibration at 300 K. Using the Nosé–Hoover thermostat, we observed temperature equilibration after 40 ps. The committors for three TS taken from both IP_dN and *NL* pathways are 0.5 ± 0.1. We can therefore conclude that the committor is independent on the dynamics of the water. Instead, the instantaneous water configuration has a structural role, as discussed above. This conclusion seems to contradict a recent study by Ma and Dinner (48) on alanine dipeptide in which they found that solvent dynamics is subtly coupled to the isomerization of the peptide. This discrepancy might be due to the difference in length and time scales, as the dynamics of the Trp-cage backbone chain is orders of magnitude slower than that of the solvent, and its size larger than correlation lengths in water.

Interestingly, when performed with a 100-ps equilibrated water configuration, our committor calculation was biased toward the unfolded state ($p_B = 0.7$). The water molecules that penetrated the

hydrophobic core at high temperature apparently are only fully equilibrated at 300 K after 1 ns. This finding highlights the slow relaxation of water next to proteins and agrees with the observation that waters can be bound to the protein for a long time. These results also explain why we see no coupling to the solvent motion due to a hydrophobic drying transition (49): our system is too small to exhibit a reasonable vapor–liquid interface nucleation barrier. It would be interesting to test the water dynamics involvement in the reaction coordinate on a protein with a much larger hydrophobic core.

Conclusion

In this paper, we present all-atom TPS simulations of Trp-cage folding in explicit solvent. After fast initial collapse, we find that two pathways are possible, one in which the helix forms first, the second where the loop, the tertiary contact between Trp-6 and the polyproline part form before the helix. The second pathway is four times more likely than the first. This finding is contradictory to implicit solvent results, where the paths through the I state were more dominant. The two different classes of pathways are reminiscent of the two prevalent generic protein folding mechanisms, the diffusion–collision mechanism (1) and the nucleation–condensation mechanism (2). Our results indicate that both mechanisms can act simultaneously in one protein (3, 17). Another finding of this work is the existence of a long-lived partly solvated intermediate on the NI path, the P_d state, characterized by a detached proline. To our knowledge, this intermediate was not recognized before. Committor analysis indicated that the water dynamics is not a part of the reaction coordinate for the folding of this small protein. However, the observation that some water molecules are extremely long-lived during the folding might contribute to the improvement of implicit solvent models. The TPS methodology also allows for a rate constant calculation, but this is beyond the scope of this paper.

1. Karplus D, Weaver DL (1976) *Nature* 260:404–406.
2. Abkevich VI, Gutin AM, Shakhnovich EI (1994) *Biochemistry* 33:10026–10036.
3. Gianni S, Guydosh NR, Khan F, Caldas TD, Mayor U, White GWN, DeMarco ML, Daggett V, Fersht AR (2003) *Proc Natl Acad Sci USA* 100:13286–13291.
4. Neidigh JW, Fesinmeyer RM, Andersen HH (2002) *Nat Struct Biol* 9:425–430.
5. Qiu L, Pabit SA, Roitberg AE, Hagen SJ (2002) *J Am Chem Soc* 124:12952–12953.
6. Neuweiler H, Doose S, Sauer M (2005) *Proc Natl Acad Sci USA* 102:16650–16655.
7. Ahmed Z, Beta IA, Mikhonin AV, Asher SA (2005) *J Am Chem Soc* 127:10943–10950.
8. Snow CD, Zagrovic B, Pande VS (2002) *J Am Chem Soc* 124:14548–14549.
9. Simmerling C, Strockbine B, Roitberg AE (2002) *J Am Chem Soc* 124:11258–11259.
10. Pitera JW, Swope W (2003) *Proc Natl Acad Sci USA* 100:7587–7592.
11. Chowdhury S, Lee MC, Duan Y (2004) *J Phys Chem B* 108:13855–13865.
12. Ota M, Ikeguchi M, Kidera A (2004) *Proc Natl Acad Sci USA* 101:17658–17663.
13. Zhou R (2003) *Proc Natl Acad Sci USA* 100:13280–13285.
14. Linhananta A, Boer J, MacKay I (2005) *J Chem Phys* 122:114901.
15. Ding F, Buldyrev SV, Dokholyan V (2005) *Biophys J* 88:147–155.
16. Rhee YM, Sorin EJ, Jayachandran G, Lindahl E, Pande VS (2004) *Proc Natl Acad Sci USA* 101:6456–6461.
17. Dinner AR, Sali A, Smith LJ, Dobson CM, Karplus M (2000) *Trends Biochem Sci* 25:331–339.
18. Jayachandran G, Vishal V, Pande VS (2006) *J Chem Phys* 124:164902.
19. Beck DAC, Dagget V (2004) *Methods* 34:112–120.
20. Sørensen MR, Voter AF (2000) *J Chem Phys* 112:9599.
21. Frenkel D, Smit B (2002) *Understanding Molecular Simulation* (Academic, San Diego), 2nd Ed.
22. Laio A, Parrinello M (2002) *Proc Natl Acad Sci USA* 99:12562–12566.
23. Voter AF, Sørensen MR (1999) *Mat Res Soc Symp Proc* 538:427.
24. Grubmüller H (1995) *Phys Rev E* 52:2893–2906.
25. Dellago C, Bolhuis PG, Csajka FS, Chandler D (1998) *J Chem Phys* 108:1964–1977.
26. Bolhuis PG, Chandler D, Dellago C, Geissler PL (2002) *Annu Rev Phys Chem* 53:291–318.
27. Dellago C, Bolhuis PG, Geissler PL (2002) *Adv Chem Phys* 123:1–78.
28. van Erp T, Moroni D, Bolhuis PG (2003) *J Chem Phys* 118:7762–7774.
29. Bolhuis PG (2003) *Proc Natl Acad Sci USA* 100:12129–12134.
30. Bolhuis PG (2005) *Biophys J* 88:50–61.
31. Lindahl E, Hess B, van der Spoel D (2001) *J Mol Mod* 7:306–317.
32. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) *J Phys Chem* 105:6474–6487.
33. Hess B, Bekker B, Berendsen HJC, Fraaije, JGEM (1997) *J Comp Chem* 18:1463–1472.
34. Miyamoto S, Kollman PA (1997) *J Comp Chem* 13:952–962.
35. Darden T, York D, Pedersen LG (1993) *J Chem Phys* 98:10089–10092.
36. Essman U, Perera L, Berkowitz ML, Daren T, Lee H, Pedersen LG (1995) *J Chem Phys* 103:8577–8592.
37. Nosé S (1984) *Mol Phys* 52:255–268.
38. Hoover WG (1985) *Phys Rev A* 31:1695–1697.
39. Scott WRP, Hünenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, vanGunsteren WF (1999) *J Phys Chem* 103:3596–3607.
40. Rhee YM, Pande VS (2003) *Biophys J* 84:775–786.
41. Sugita Y, Okamoto Y (1999) *Chem Phys Lett* 314:141–145.
42. Rathore N, Chopra M, dePablo JJ (2005) *J Chem Phys* 122:024111.
43. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES (1998) *J Chem Phys* 108:334–350.
44. Caflisch A, Karplus M (1995) *J Mol Biol* 252:672–708.
45. Guo W, Lampoudi S, Shea JE (2003) *Biophys J* 85:61–69.
46. Cheung MS, Garcia AE, Onuchic JN (2002) *Proc Natl Acad Sci USA* 99:685–690.
47. Rhee YM, Pande VS (2006) *Chem Phys* 323:66–77.
48. Ma A, Dinner AR (2005) *J Phys Chem B* 109:6769–6779.
49. ten Wolde PR, Chandler D (2002) *Proc Natl Acad Sci USA* 99:6539–6543.
50. Humphrey W, Dalke A, Schulten K (1996) *J Mol Graphics* 14:33–38.
51. Olender R, Elber R (1996) *J Chem Phys* 105:9299–9315.