THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

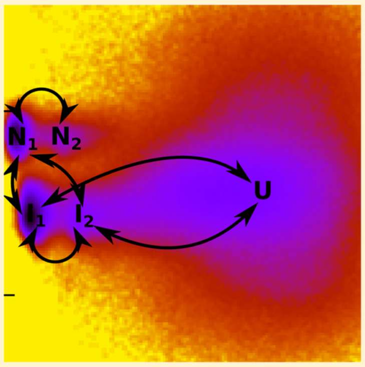
Hierarchical Folding Free Energy Landscape of HP35 Revealed by Most Probable Path Clustering

Abhinav Jain and Gerhard Stock*

Biomolecular Dynamics, Institute of Physics and Freiburg Institute for Advanced Studies (FRIAS), Albert Ludwigs University, 79104 Freiburg, Germany

Supporting Information

ABSTRACT: Adopting extensive molecular dynamics simulations of villin headpiece protein (HP35) by Shaw and co-workers, a detailed theoretical analysis of the folding of HP35 is presented. The approach is based on the recently proposed *most probable path* algorithm which identifies the metastable states of the system, combined with dynamical coring of these states in order to obtain a consistent Markov state model. The method facilitates the construction of a dendrogram associated with the folding free-energy landscape of HP35, which reveals a hierarchical funnel structure and shows that the native state is rather a kinetic trap than a network hub. The energy landscape of HP35 consists of the entropic unfolded basin *U*, where the prestructuring of the protein takes place, the intermediate basin *I*, which is connected to *U* via the rate-limiting *U* → *I* transition state reflecting the formation of helix-1, and the native basin *N*, containing a state close to the NMR structure and a native-like state that exhibits enhanced fluctuations of helix-3. The model is in line with recent experimental observations that the intermediate and native states differ mostly in their dynamics (locked vs unlocked states). Employing dihedral angle principal component analysis, subdiffusive motion on a multidimensional free-energy surface is found.



INTRODUCTION

Molecular dynamics (MD) simulations represent a versatile approach to describe the structure, dynamics, and function of biomolecules in microscopic detail.[1] Due to the development of supercomputing devices and improved force fields, even the "in silico" folding of small proteins has become possible recently.[2−4] Assuming that these simulations provide a statistically meaningful ($\gtrsim$ 100 events) and sufficiently accurate (compared to experiment) description of protein folding, the challenge is now to analyze the resulting huge amount of data and construct a simple model that explains the essential physics of the process. Dimensionality reduction methods such as principal component analysis (PCA) attempt to reduce the description of the highly correlated molecular motion of 3N atomic coordinates to some collective degrees of freedom.[5−8] The resulting low-dimensional representation of the dynamics can then be used to construct the free-energy landscape of the process. Alternatively, one may partition the continuous MD trajectory in discrete metastable states and construct a Markov state model, which approximates the dynamics of the system by a memoryless jump process.[9−14]

Owing to the dynamical complexity of even the smallest proteins, various theoretical approaches may arrive at quite different pictures for the same folding trajectory. For example, there is an ongoing debate on whether a single or multiple pathways dominate the folding process,[11,15] whether the free-energy landscape is described by a folding funnel or rather by a native hub model,[9,13,16,17] whether the motion on the energy landscape should be described by a diffusive or by a subdiffusive process,[18−20] and the required dimensionality of the energy landscape.[21−24] Using different reaction coordinates or state partitionings, different methods may even predict conflicting folding mechanisms (e.g., the presence of intermediates vs an almost downhill scenario).[15,25,26]

We have recently introduced a mixed PCA/state-based approach to characterize biomolecular energy landscapes.[23,27] It consists of (i) a careful selection of microstates using PCA preprocessing and *k*-means clustering, (ii) the *most probable path* algorithm to identify the metastable states of the system, and (iii) boundary corrections of these states via the introduction of cluster cores in order to obtain the correct dynamics. In an attempt to clarify the above-posed questions, in this work we apply the methodology to a reference example, the folding of villin headpiece protein.[28−41] To this end, we adopt ≈ 300 μs long trajectories of the fast folding variant HP-35 NleNle (henceforth simply referred to as HP35) recently published by D. E. Shaw Research,[35] which show more than 100 folding and unfolding events. The simulations reproduce at least qualitatively some of the main experimental findings for the system,[37−39] including the melting temperature (370 K in MD vs 361 K in experiment), the folding enthalpy (21 vs 25 kcal/mol), as well as the folding time (3 vs 1 μs). Moreover, the relative folding times for various mutations of the villin

Special Issue: James L. Skinner Festschrift

Received: October 21, 2013
Revised: January 9, 2014
Published: January 9, 2014

dx.doi.org/10.1021/jp410398a | J. Phys. Chem. B 2014, 118, 7750−7760

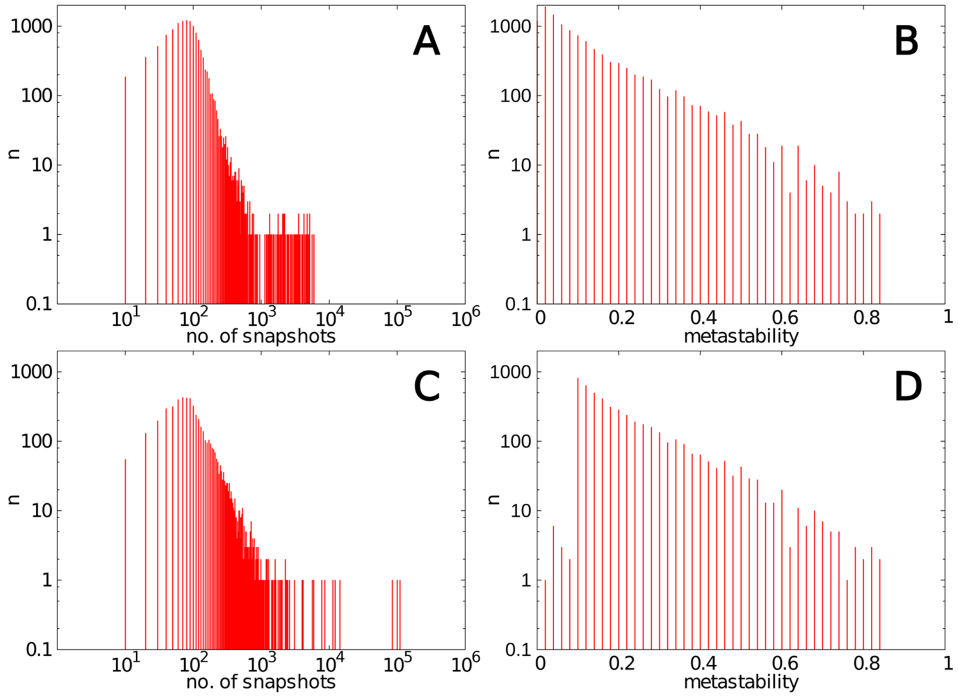


Figure 1. (a) Distribution of the MD data among the microstates shown as number of microstates that contain a certain number of data points. (b) Distribution of the metastability T_{ii} of these microstates. (c,d) As above but for the distribution of states obtained by the most probable path algorithm for $Q_{\min} = 0.1$.

headpiece are nicely reproduced. Performing a detailed PCA/state-based analysis of these trajectories, we identify the metastable states of HP35, characterize the folding mechanism and pathways, and clarify the applicability and validity of the above-mentioned theoretical concepts of protein folding.

■ METHODS

Piana et al.[35] performed equilibrium MD simulations at various temperatures for wild-type HP35 and various mutants, using the Amber ff99SB*-ILDN force field[42−44] and TIP3P explicit water.[45] In this work, we adopted a ≈ 300 μs long trajectory of the fast folding variant HP35 NleNle at 380 K, which shows 140 folding and unfolding events. The data contains $\approx 1.5 \times 10^6$ snapshots with a time step of 200 ps. As a consistency check, we also considered a similar trajectory at 360 K. HP35 consists of helix-1 (residues 4−10), turn-1 (residues 11−14), helix-2 (residues 15−19), turn-2 (residues 20−22), and helix-3 (residues 23−32).

To reduce the dimensionality of the trajectory, we employed the dihedral angle principal component analysis (dPCA), which uses the sine/cosine-transformed ϕ and ψ dihedral angles of the protein backbone (for details see ref 7). This avoids possible artifacts due to the mixing of overall rotation and internal motion, which may occur when a PCA using Cartesian coordinates is employed to study large amplitude processes such as folding.[46] As the terminal ends exhibit largely uncorrelated fluctuations, we excluded the dihedral angles of the first and last two residues of HP35 from the analysis, yielding in total $31 \times 2 \times 2 = 124$ variables. Figure S1 in the Supporting Information shows the cumulative fluctuations covered by the first n principal components as well as their one-dimensional distributions. Following ref 23, we included for the further analysis all principal components with a multipeaked distribution, which typically also reflect the longest time scales of the system. We also performed a dPCA using only the data of the folded and the unfolded basin, respectively, see Figures S2 and S3.

From the dPCA, the time-dependent mean squared displacement MSD(t) of the trajectory was calculated via

$$\mathrm{MSD}(t) = \frac{1}{T - t} \sum_{\tau=0}^{T-t} \sum_{k=1}^{K} [V_k(t + \tau) - V_k(\tau)]^2 \tag{1}$$

where T denotes the total number of MD frames and K is the total number of included principal components V_k. Alternatively, we have calculated the MSD using the Cartesian coordinates $\boldsymbol{x}_i$ of all M backbone atoms via

$$\mathrm{MSD}(t) = \frac{1}{T - t} \sum_{\tau=0}^{T-t} \sum_{i=1}^{M} |\boldsymbol{x}_i(t + \tau) - \boldsymbol{x}_i(\tau)|^2 \tag{2}$$

where we performed a pairwise rotational and translational fit between structures at times $t + \tau$ and τ in order to subtract the overall motion.

To reduce the large number of MD snapshots ($\sim 10^7$) to a computationally manageable number of microstates ($\sim 10^4$), we used the k-means geometric clustering algorithm.[47] Adopting the 11-dimensional conformational space obtained from the dPCA described above, we thus discretized the conformational space into $k = 12\,000$ microstates (for details see ref 23). The microstates were employed to calculate one-dimensional barrier-preserving free-energy profiles, using the native structure as reference state and P_{fold} as progress variable.[48]

Employing the most probable path algorithm,[27] we next merge all microstates that belong to the same metastable state, that is, we assume a time scale separation of fast motion within the metastable states and slow interstate motion (i.e., rare interstate transitions). To this end, we calculate the transition matrix $\{T_{ij}\}$ for the microstates, where T_{ij} represents the probability of a state i to change to state j within a predefined

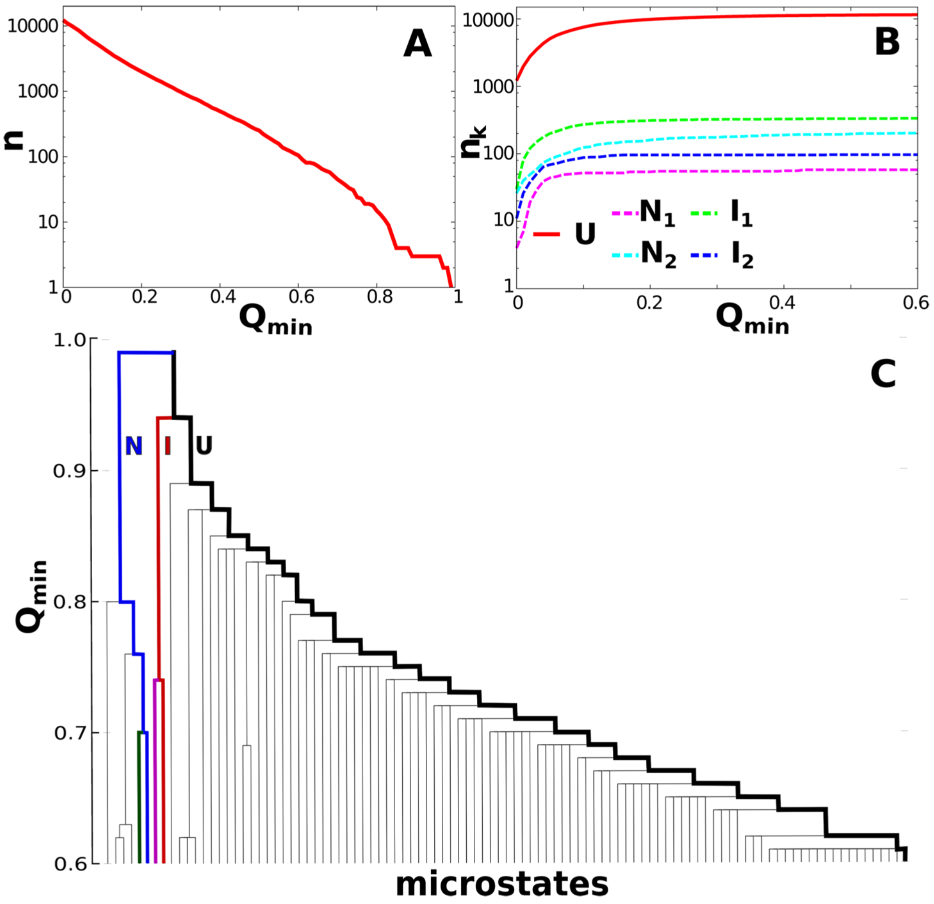


Figure 2. Partitioning of state space obtained by the most probable path algorithm. (a) Evolution of the number of metastable states n as a function of the minimum metastability Q_{min}. (b) Number of microstates that end up in the "final" metastable states U, I_1, I_2, N_1, and N_2. (c) Dendrogram showing the hierarchical structure of the energy landscape. Thick lines indicate the main metastable states which cover 98% of the population. Thin lines correspond to states with a population of $\lesssim$0.3%.

lag time (τ = 2 ns) and the self-transition probability T_{ii} represents the metastability of state i. Starting from a given microstate, the basic idea of the most probable path algorithm is to calculate the transition probabilities of this state and to choose the most probable transitions for all states with $T_{ii} \leq Q_{\mathrm{min}}$, where Q_{min} is a chosen minimum value of the metastability. This goes on until we reach a state i with $\max(T_{ij}) \leq T_{ii}$ (i.e., it is most likely to stay in this state). As *intra*-basin transitions are much more likely than *inter*-basin transitions (i.e., barrier crossing), the scheme collects all microstates of a basin, thereby defining the basin in terms of the included microstates. Moreover, the approach by construction places the boundaries between the metastable states in the middle of the separating barriers. A recently published comparison[49] of alternative methods to identify metastable states found the overall performance of the most probable path algorithm similar to more established methods based on the eigenvectors of the transition matrix.[14] Unlike the latter approaches, though, the most probable path algorithm does not require the diagonalization of the transition matrix and moreover affords an illustrative analysis and interpretation of the clustering process (see Figure 2). A python implementation of the most probable path algorithm can be downloaded from www.moldyn.uni-freiburg.de.

RESULTS

Identification of Metastable States. As described above, we first employed dPCA and k-means geometric clustering in order to generate a suitable number of *microstates*, which represent the conformational space of HP35 sampled by the MD trajectory in a fine-grained state space. Figure 1a shows how the 1.5 × 10^6 MD snapshots are divided up into the 12 000 microstates. As is expected from a density-based algorithm, k-means populate the majority of the states with roughly the same number (≈100) of MD points. Only some hundred microstates are significantly higher or lower populated, that is, almost all microstates are sufficiently well sampled to represent a statistical meaningful entity. Figure 1b displays the distribution of the metastability T_{ii} of the microstates (i.e., the probability that state i does not change to another state within the chosen lag time of τ = 2 ns).[50] The exponential shape of the distribution indicates that the partitioning of microstates is statistically unbiased, because a quantity, for which only a fixed mean ($\langle T_{ii}\rangle$ = const) is required, is most likely exponentially distributed according to maximum entropy considerations.[51]

Using these microstates, we next wish to construct *metastable states* (sometimes also called "macrostates"), which—at least in principle—can be observed in experiment. To this end, we first need to define a criterion to decide if a microstate will be assigned to some metastable state or not. As explained above, in the most probable path algorithm, we only merge a microstate if its metastability T_{ii} is smaller than a chosen minimum value $Q_{\min}$. In this way, $Q_{\min}$ defines the minimum barrier height by which the resulting metastable states are separated from their neighboring states. By gradually increasing $Q_{\min}$ from 0 to 1, the number of metastable states $n(Q_{\min})$ decreases from 12 000 (all microstates remain) to 1 (all microstates are merged into a single state). Figure 2a reveals a roughly exponential decay of $n(Q_{\min})$, which reflects the likewise exponential decay of the distribution of the metastability in Figure 1b.

In spite of the continuous overall decrease of $n(Q_{\min})$, a clear division of the microstates into a few metastable conformational states emerges from early on. As a first indicator, we consider the distribution of the density for $Q_{\min} = 0.1$. Figure 1c reveals that already at this point most of the data (about 60%) are included in only a few metastable states (shown as isolated "bars" on the right-hand side). As detailed below, they consist of the unfolded state U, two intermediate states I_1 and I_2, as well as two native states N_1 and N_2. The remaining data are distributed among ≈4500 small states, which are fairly unstable (see Figure 1d). The immediate partitioning of the state space already at $Q_{\min} \lesssim 0.1$ is most directly observed in Figure 2b, which shows the number of microstates that end up in the "final" metastable states U, I_1, I_2, N_1, and N_2. For $Q_{\min} \gtrsim 0.1$, the remaining small states get assigned to the five main metastable states. This phase of the most probable path algorithm is important also, because the residual states are often located in the transition state regions and may therefore considerably affect the overall transition rate.

The partitioning of state space is further illustrated by the dendrogram in Figure 2c, which shows the process from a top-down perspective (i.e., for decreasing minimum metastability $Q_{\min}$). Already close to $Q_{\min} = 1$, we find an overall division of the energy landscape in native (N), intermediate (I), and unfolded (U) basins, respectively. The unfolded basin is actually a single metastable state, in the sense that it never splits up into a few similarly populated substates. Rather, the dendrogram shows that U accumulates small states for increasing $Q_{\min}$. Containing the vast majority of microstates, U is of entropic nature. The intermediate states are separated from the unfolded basin early on ($Q_{\min} \leq 0.95$) and split up in I_1 and I_2 at $Q_{\min} \leq 0.73$. In contrast to U, they contain only a few highly populated microstates and are thus of enthalpic nature. A similar situation is found for the two native states N_1 and N_2. For $Q_{\min} \lesssim 0.7$, no further relevant subdivision of the states occurs, because the five metastable states are already formed for $Q_{\min} \lesssim 0.1$. The dendrogram thus reveals a surprisingly simple but nevertheless hierarchical structure of the folding free-energy landscape of HP35.

The properties of the metastable states are collected in Table 1. Due to the relatively high temperature (380 K) of the simulation, the unfolded state U is the most populated state (76%) and contains most (11 900) of the microstates. Its high metastability (0.99) stems from the fact that it connects to the intermediate states only via a small transition state region. The intermediate states I_1 and I_2 provide two separate pathways to connect the unfolded and the folded states. They are fairly localized in space (together they contain 170 microstates and 14% population) and exhibit relatively low metastability (0.8 and 0.74). Both intermediate states lead to the N_1 state, which is localized (61 microstates and 7.5% population) and shows the smallest RMSD to the experimental structures. The second native state, N_2, is of relatively low population and stability and connects only to N_1.

Table 1. Population P_i, Number of Microstates n_i, Metastability T_{ii}, Lifetime τ_i, and Total Number of Visits $\sum_j C_{ji}$ of the Five Metastable States of HP35. Superscript "c" Indicates the Respective Quantity after Dynamic Coring

state	P_i [%]	n_i	T_{ii}	P_i^c [%]	T_{ii}^c	τ_i^c [μs]	$\Sigma_j C_{ji}$
U	76.34	11 907	0.9887	73.16	0.9992	2.50	107
I_1	8.12	98	0.7972	9.34	0.9511	0.04	942
I_2	5.98	71	0.7391	7.59	0.9359	0.03	891
N_1	7.52	61	0.9465	8.23	0.9825	0.11	219
N_2	2.04	182	0.8178	1.68	0.9553	0.04	115

Structural Distribution of Metastable States. Comprising a vast variety of conformations, the unfolded basin U contains by far the majority (97%) of all microstates. Here, the "prestructuring" of the protein takes place, starting from largely random structures to an almost folded system with preformed secondary structures. The temporal order of this prestructuring was found to depend to some extent on the force field used in the simulations.[34] For example, the Amber force field ff99SB*-ILDN[42−44] used here preferably folds first helix-3, then helix-2, and helix-1 last, whereas the CHARMM27 force field[52,53] typically folds helix-2 last. The conformational substates of U can be resolved by a "PCA by parts",[54,55] in which a separate PCA for each secondary-structure element is performed. The analysis reveals that 1, 7, 27, 41, and 24% of the MD frames in U have 0, 1, 2, 3, and 4 secondary-structure parts formed in native-like conformations, respectively. See also Figure S4, which shows the distribution of the number of helical residues in the three basins. This energetic bias of the prestructuring process toward the native state combined with the loss of configurational entropy when the number of accessible states decreases is the essence of the "folding funnel" paradigm.[16,17] It is interesting to note that the barrier heights between dynamically adjacent substates in U are $\lesssim k_BT$. This is indicated by the rapid initial increase of the number of microstates in U (Figure 2b) as well as by the structureless appearance of the dPCA distributions of the unfolded state (Figure S3). Due to the many possible destabilizing interactions in an unfolded protein, the substates of U are not metastable and therefore hard to observe in experiment.

Next, we wish to describe the transition state region between the unfolded and intermediate states. To this end, we first determined the "hopping times" t_i of all $U \leftrightarrow I$ transitions from the trajectory in state space. By considering a time window of 40 ns (i.e., 200 MD frames) right before and right after t_i, we then calculated the (ϕ,ψ) distributions of all residues before and after the transition (averaged over all events). The most significant conformational changes occur for residues 9−13, where Val9 and Phe10 are at the C-terminus of helix-1 and Gly11, Met12 and Thr13 are part of turn-1. Figure 3 shows the associated (ϕ,ψ) distributions for the $U \to I$ folding transition. The transition consists of (i) a compacting of helix-1 residues from a mixed extended/helical to a helix-only conformation, (ii) a change from a mixed left/right to a left-only helical conformation of Gly11, and (iii) of a stretching of turn-1 via a change from a mixed extended/helical to a extended-only conformation. Starting in U with already-folded helices 2 and 3 as well as a partially folded helix-1, the $U \to I$ transitions essentially folds helix-1 (except for residue Asp3, see below). As shown in Figure S5, the conformational change of the $I \to U$ unfolding transition is quite similar. However, the temporal

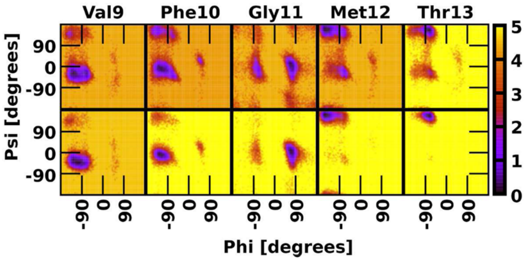


Figure 3. Characterization of the transition state region of HP35 as reflected in the distributions of the backbone (ϕ,ψ) dihedral angles of residues 9–13, which were calculated (top) right before and (bottom) right after the $U \rightarrow I$ transition.

order of the transitions changes (i.e., the first residues to unfold are the last to fold).

With respect to basins N and I, one finds that all conformations in these basins are relatively close to the experimentally found native structure of HP35.[36] This is in line with the experimental observation that the intermediate and native states are structurally quite similar and differ mostly in their dynamics (locked vs unlocked states).[41] To identify a possible structural origin of this dynamics, we considered the Ramachandran (ϕ,ψ) plots of all residues of the protein. It turns out that residue Asp3 plays a key role in the discrimination of intermediate and native states. This is demonstrated by Figure 4, which shows the (ϕ,ψ) distribution of Asp3 for (a) the unfolded state U (which is identical to the results for the complete trajectory), (b) the native states N_1 and N_2, and (c) and (d) of the intermediate states I_1 and I_2, respectively. We find that the native states sample the α_L-helical region, the intermediate states populate the α_R-helical (I_1) and the 3_{10}-helical/extended (I_2) conformations, and the unfolded state covers all of them. Due to this change of residue Asp3 from various right-handed conformations to a stable α_L-helical conformation, the $I \rightarrow N$ transition affects a compacting of hydrophobic core of the protein ("locked state"), because two of the three hydrophobic residues of HP35 (Phe7 and Phe11) are located in helix-1. On the other hand, when changing from $N \rightarrow I$, the intermediate states show significant fluctuations of helix-1 ("unlocked state").

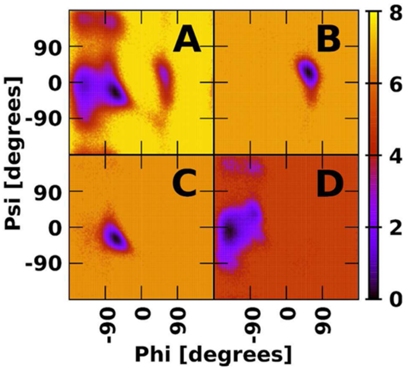


Figure 4. (ϕ,ψ) Distribution of residue Asp3 of HP35 in the (a) unfolded state U, (b) native states N_1 and N_2, (c) intermediate state I_1, and (d) intermediate state I_2.

The two native states N_1 and N_2 differ in the structure of residues 29–33 (i.e., the C-terminus of helix-3). The Ramachandran (ϕ,ψ) plots of these residues displayed in Figure S6 clearly show that in N_1 these residues stay in a helical structure, although—due to the lack of a stabilizing hydrogen bond—in N_2 also extended conformations are populated. This results in a flexible C-terminus of helix-3, that is, N_2 represents an unlocked state as well.

Dynamic Coring and Markov Modeling. Once the metastable states are defined, we can describe the folding process along a given MD trajectory by assigning a state k to each MD frame. Showing the time evolution of the state variable $k(t)$ for a short section of the total trajectory, Figure 5a readily reveals that the folding from the unfolded basin U via the intermediate states I_1/I_2 to the native state N_1 occurs on a microsecond time scale. Somewhat unexpectedly, however, we find that most transitions are accompanied by frequent recrossings (i.e., the metastable states appear to be fairly unstable). A closer analysis reveals that the recrossings are mainly due to insufficient partitioning of the conformational space in the barrier regions.[27] For example, the grid of microstates might not be chosen fine-grained or high-dimensional enough to clearly locate the top of each barrier. As a consequence, intrastate conformational fluctuations that reach over the interstate borders may be mistaken as interstate transitions, which leads to artificial short lifetimes of the metastable states. As an illustrative example, Figure 5b shows the time-dependent population probability $P_{N_1}(t)$ of the native state N_1 (red line), given that the system started in this state at time $t = 0$. $P_{N_1}(t)$ is seen to perform a rapid initial decay on the time scale of the lag time (2 ns) followed by some slower relaxation. The fast decay is a clear consequence of the above-described spurious recrossings, which defy a simple Markov modeling of the data.

To avoid this problem, one may employ the concept of milestones[56] or cluster cores.[10] The idea is to require that a valid transition from a state (or cluster) to another must reach the core region of the other cluster.[48,57,58] However, in a high-dimensional system, the geometric definition of the core of a metastable state may become difficult, in particular, if the state is of entropic nature and exhibits several subminima. Alternatively, we can invoke a dynamical criterion and restrict the counting of transitions to events that remain at least some minimum time $t_{\min}$ in the new state. If this condition is not met, the trajectory points are reassigned to the last visited state. Obviously, the coring time $t_{\min}$ needs to be much shorter than any dynamical time scale of interest.

To illustrate the effect of this "dynamic coring", Figure 5b shows the population probability $P_{N1}(t)$ of the native state N_1 for $t_{\min} = 1$ and 2 ns. A large part of the spurious initial decay is found to disappear already for a time window of 1 ns. Using a 2 ns window, we obtain a single exponential decay, which does not change for $t_{\min} = 3$ ns (data not shown). Performing this procedure for all five metastable states, we determined time windows of 1 ns for the states N_1, I_1, and I_2, 8 ns for N_2, and 60 ns for the entropic state U. Using these defined cored states, Figure 5c,d show that all five metastable states decay exponentially with lifetimes collected in Table 1. We note that in general the time windows of connected metastable states are not independent of each other and may therefore need to be determined in a self-consistent manner.

To verify the Markovianity of the metastable states after coring, we first tested the detailed balance condition $P_i^{\mathrm{eq}}T_{ij} = P_j^{\mathrm{eq}}T_{ji}$, where P_i^{eq} denotes the equilibrium population probability of state $i = N_1, N_2, I_1, I_2, U$, and T_{ij} is the transition probability

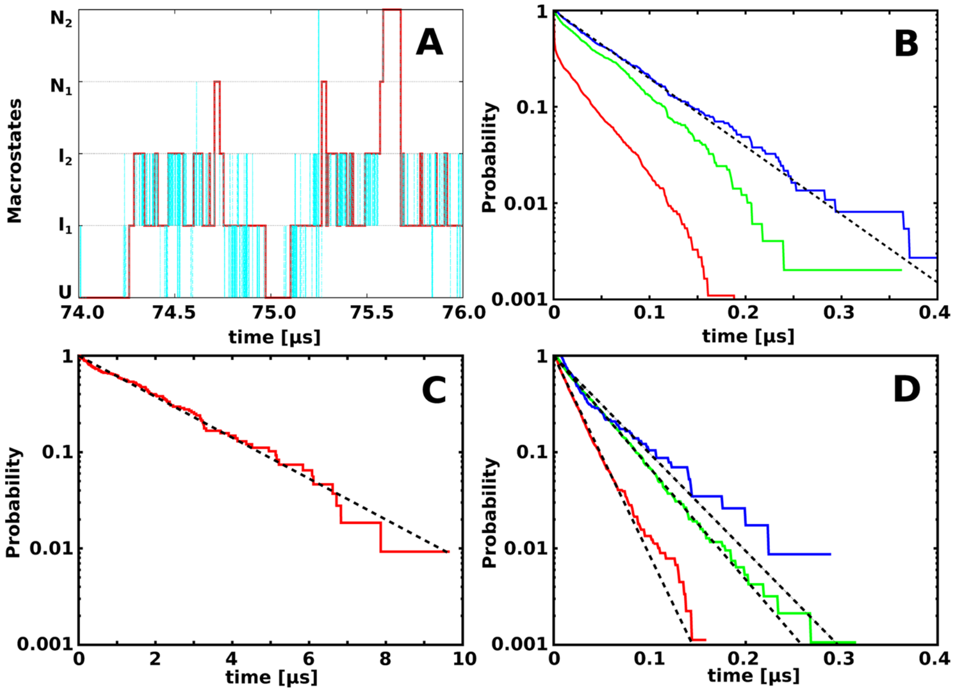


Figure 5. (a) Time evolution of a representative piece of the state trajectory $k(t)$ with (red) and without (blue) dynamic coring. (b) Effect of the coring on the population probability $P_{N_1}(t)$ of the native state N_1 for coring time $t_{\min}$ = 0 (red, no coring), 1 ns (green), and 2 ns (blue). Population probability $P_k(t)$ of (c) the cored state U as well as of (d) I_1 (green), I_2 (red), and N_2 (blue), compared to the corresponding results from the Markov model (dashed lines).

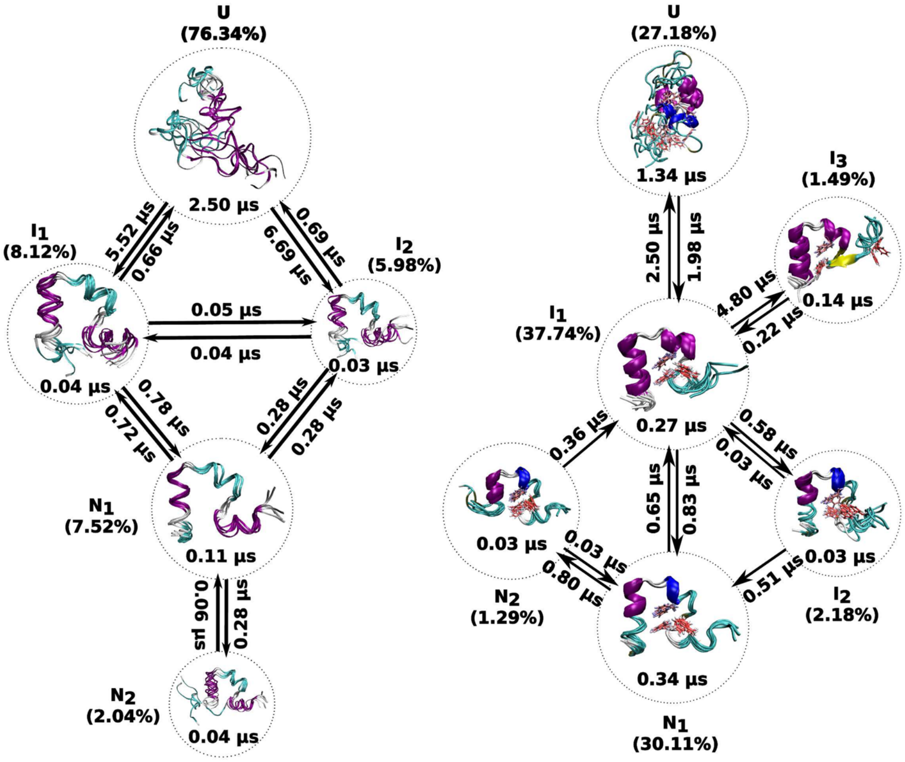


Figure 6. Markov state model of the reversible folding and unfolding of HP35 obtained at 380 K (left) and 360 K (right). Shown are the metastable states including their populations (in %), lifetimes (in μs), and representative structures, as well as the transition times (in μs). Links with less than 10 transitions or a rate lower than 1/10 μs are neglected.

from state i to state j obtained from a row-normalized transition count matrix. Subsequently, we considered the Chapman–Kolmogorov equation $P(n\tau) = P(0)T(n\tau) = P(0)T^n(\tau)$, where $P(t)$ represents the state vector at time t and $T(\tau)$ is the transition matrix obtained for lag time τ. Table S1 and Figure 5 show that both conditions are satisfied with high precision.

The resulting Markov state model illustrates the energy landscape and the pathways of the folding and unfolding of HP35. Figure 6a displays a network representation of the transition matrix, where the nodes are the five metastable states, and the edges are the transitions among them. Starting from the unfolded state U, we obtain the folding pathways and folding times: (1) $U \rightarrow I_1 \rightarrow N_1$ (5.7 μs), (2) $U \rightarrow I_1 \rightarrow I_2 \rightarrow N_1$ (5.3 μs), (3) $U \rightarrow I_2 \rightarrow N_1$ (6.9 μs), and (4) $U \rightarrow I_2 \rightarrow I_1 \rightarrow N_1$ (7.4 μs). Although path 2 represents the fastest individual pathway, the average folding time $\langle \tau_F \rangle = (\sum_i 1/\tau_i)^{-1}$ amounts to 3.5 μs, that is, the collective flux rather than the fastest pathway determines the folding speed. Similarly, the average time for complete unfolding from N_1 to U is estimated $\langle \tau^{-1}_{N_1 \rightarrow U} \rangle$ = 0.88 μs.

To obtain an impression of the overall significance and consistency of the above-analyzed MD data, we repeated the analysis for a HP35 trajectory at 360 K (with similar length and simulation conditions[35]). Figures S7–S11 and Table S2 show that one obtains a quite similar structural and dynamical characterization of the metastable states as for 380 K. The main differences of the simulations at 380 and 360 K are depicted in Figure 6, which compares the resulting Markov state models for the two cases. Due to the lower temperature, the run at 360 K preferably populates the native (31%) and the intermediate (42%) states rather than the unfolded state (27%). Again, we obtain a weakly populated native-like state N_2 with a flexible helix-3 that can only be reached from the main native state N_1. Interestingly, though, we find that the intermediate states differ for the two simulations. At 360 K, we find one main intermediate state (I_1) containing 39% of the population and two weakly populated states (I_2 and I_3), which differ mainly in the structure of the N-terminal residues of HP35. Similar to that found in a previous folding trajectory of HP35,[27] I_3 exhibits an extended β-strand structure instead of helix-1. The overall transition times between the basins U, I, and N are 3 and 1.4 μs ($U \rightarrow I$) as well as 0.4 and 0.8 μs ($I \rightarrow N$) for 380 and 360 K, respectively. They differ by a factor of 2, which reflects the different thermal populations of the basins. Put together, the Markov state models at 380 and 360 K are found to provide a largely consistent picture of the folding process. Although the separation of the native states in a main state (N_1) and a native-like state (N_2) seems to prevail, the details of the intermediate states are found to somewhat vary with temperature.

Free-Energy Landscape. It is instructive to also consider *continuous* representations of the free-energy landscape of HP35. With the choice of a suitable reaction coordinate, the separation of the three basins U, I, and N can be already achieved by a one-dimensional free-energy curve. As representative examples, Figure 7 shows the free energy as a function of (a) the root-mean-square deviation (RMSD) with respect to the experimental crystal structure and (b) a barrier-preserving coordinate,[48] constructed from the partition function of the microstates (see Methods). Alternatively, one may focus on a particular transition (e.g., $U \rightarrow I_1$) and define the reaction coordinate r as the difference of the coordinate vectors of their geometric centers (e.g., $r_{U,I_1} = |\vec{R}_U - \vec{R}_{I_1}|$). Figure 7d reveals that the resulting free-energy curve $\Delta G(r_{U,I_1})$ exhibits a barrier between U and I_1 of about 2.5 k_BT, which is in agreement with the result obtained by the partition function-based coordinate in Figure 7b. A very similar curve is found for the transition between U and I_2 (data not shown). The highest barriers ($\approx$ 5 k_BT) are found between the intermediate states I_1/I_2 and the native state N_1 (see $\Delta G(r_{I_1,N_1})$ in Figure 7c). The partial unfolding of the C-terminus of helix-3 during the $N_1 \rightarrow N_2$ transition is also energetically expensive ($\approx$ 4.75 k_BT), whereas the reverse transition only shows a barrier of $\approx$2 k_BT (see $\Delta G(r_{N_1,N_2})$ in Figure 7e). Using the transition state expression $k = k_0 \exp(-\Delta G/k_BT)$ for the rate of a transition with barrier height ΔG, we estimate the prefactors for the $U \leftrightarrow I$ and the $I \leftrightarrow N$ reactions as $1/k_0^{UI}$ = 0.25 μs and $1/k_0^{IN}$ = 3 ns, respectively, which are in good agreement with previous work.[35,38]

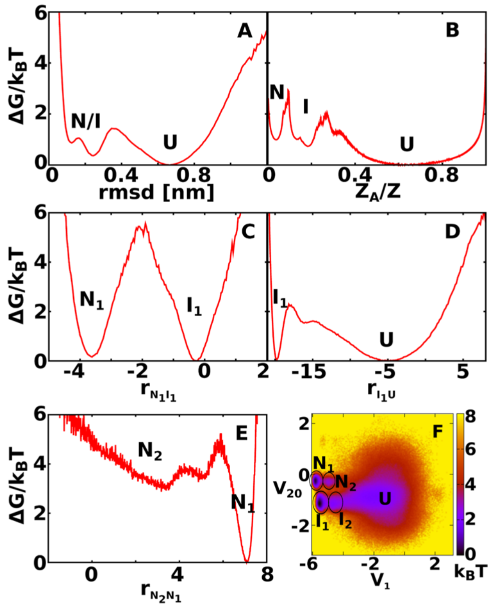


Figure 7. Free-energy landscape of HP35. Shown are one-dimensional free-energy curves constructed from (a) the RMSD, (b) a partition function-based coordinate, and (c,d,e) the difference vectors of the geometric centers of a pair of metastable states. (f) Two-dimensional representation of energy landscape, constructed from two components of the dPCA.

It should be kept in mind, however, that low-dimensional projections of the high-dimensional energy surface may be misleading, because they generally cannot reproduce the correct connectivity and the barriers of the metastable states.[21,23] For example, the one-dimensional "optimal reaction coordinate" introduced in ref 59 to analyze the identical 380 K trajectory of HP35 results in a two-state picture of the free-energy landscape. That is, the method cannot discriminate the structurally similar intermediate states I_1 and I_2 and native states N_1 and N_2, although these states are separated by significant barriers (cf. Figure 7). The fact that different optimizations of one-dimensional reaction coordinates may result in largely different pictures of the free-energy landscape was also discussed in ref 26. To resolve all five metastable states rather than two or three basins, one needs to invoke (at least) a two-dimensional representation of the energy landscape. Figure 7f shows such a landscape as constructed from two components of the dPCA (see Methods), which clearly discriminates all metastable states of HP35. Interestingly, we find that we need to include 10–20 principal components to cover a substantial amount (say, $\gtrsim$ 50%) of the system's fluctuations (see Figure S1). Upon performing separate dPCAs for the folded and the unfolded parts of the trajectory (see Figures S2 and S3), respectively, one learns that this relatively high dimensionality of HP35 is a consequence of the many different possible molecular motions of the protein in the unfolded state. This finding seems to be in variance with previous studies that indicated a relatively small dimensionality of biopolymers.[22,24]

Following the construction of a continuous free-energy landscape, we may characterize the molecular motion on this landscape by considering the time-dependent mean squared displacement, MSD(t), of the trajectory (see Methods). In the case of diffusive motion, one finds MSD(t) = $2Dt$ (with D being the diffusion constant), while one obtains MSD(t) = $2Dt^{\alpha}$ with $0 \leq \alpha \leq 1$ for subdiffusive motion.[18−20] We stress that the notion of diffusion or subdiffusion is not a general characterization of the dynamics but naturally depends on the choice and the dimensionality of the reaction coordinate. This ambiguity is similar to the finding that different physical observables (which may be described by different reaction coordinates) may also reflect different aspects of the dynamics.[19] As the dPCA represents a simple and systematic approach to represent the free-energy landscape, in the following we calculate the MSD along individual components of the dPCA as well as in the full-dimensional space. As a representative example, Figure 8a shows the MSD of HP35 along the first and the second principal component. In both cases, we find subdiffusive behavior with $\alpha \approx 0.5$ which, however, lasts for different time scales. Reflecting the main folding-unfolding transition, the first component undergoes subdiffusion up to $t \approx 1$ μs (i.e., the time scale of this transition). The MSD of the second principal component only rises up to $t \approx 100$ ns, a time scale corresponding to intrabasin motion. The higher components show quite similar behavior (Figure.S12). Due to the ordering of the principal components by variance, however, the associated diffusion constant D decreases for higher components.

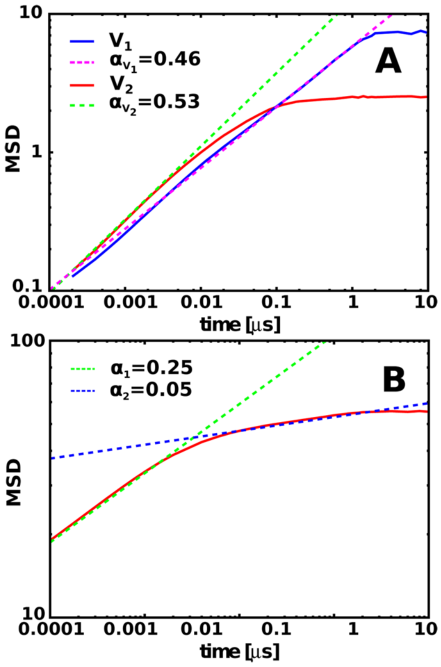


Figure 8. Time-dependent mean squared displacement (MSD) of the MD trajectory of HP35, obtained (A) for the first two principal components and (B) in the full-dimensional dPCA space. Dashed lines are fits to t^{α}.

Given as the sum over the MSD of each component (cf. eq 1), the total MSD exhibits a biphasic behavior with $\alpha_1 = 0.25$ for $t \lesssim 20$ ns and $\alpha_2 = 0.05$ for longer times (Figure 8). That is, in the full-dimensional space the exploration of conformational space appears to be more subdiffusive, because the MSD of the numerous higher components rise only at short times compared to the MSD of the first component. Calculating the MSD in the full-dimensional *Cartesian* space of all backbone atom coordinates (cf. Equation 2), we obtain a very similar biphasic behavior with $\alpha_1 = 0.43$ and $\alpha_2 = 0.04$ (data not shown). In line with the interpretation of the helicity of U in ref 35, one may assign the fast component to contact formation in U and the slow component to the unfolding transition. Indeed, when we calculate the autocorrelation function of the number of alpha helical residues (Figure S13), we also find two time scales (30 and 700 ns) which qualitatively match the results of Piana et al.[35] Calculating the autocorrelation function from the five-state Markov model, on the other hand, only the slow time scale is recovered, while the fast time scale is by construction averaged out in the model.

■ DISCUSSION AND CONCLUSION

Adopting recently published[35] MD simulations of HP35 protein, we have analyzed in detail the structure and dynamics of the folding process and characterized the underlying free-energy landscape in state space as well as in coordinate space. The quality of the computational study rests on the following features: (i) The considered simulations of Piana et al.[35] use a state-of-the-art force field (Amber ff99SB*-ILDN[42−44]) and are long enough (≈ 300 μs) to warrant a sufficient statistical sampling of the considered process. (ii) We first employed a dPCA preprocessing of the trajectory, as it is computationally advantageous to reduce the dimensionality and the noise of the MD data. To construct microstates in the resulting 11-dimensional space, we used k-means clustering with a sufficiently large number of clusters ($k \approx 10^4$), in order to warrant a balance between the resolution of conformational space and the statistical sampling of the microstates. (iii) To identify the metastable states of the system, we employed the most probable path algorithm,[27] which allows for a systematic and physically intuitive partitioning of state space. Subsequently, boundary errors of these metastable states are corrected by dynamic coring, thus yielding a consistent Markov state model.

On the basis of this analysis of the folding trajectory of HP35, in the following we wish to discuss first, the folding mechanism, and second, the applicability and validity of several prominent theoretical concepts of protein folding.

Folding Mechanism and Pathways. According to the above analysis of the MD data of Piana et al.,[35] the folding of HP35 proceeds in the following steps: (i) Prestructuring of the protein in the unfolded basin U, where helix-3 and helix-2 fold first, while helix-1 is only partially folded. As the barriers between dynamically adjacent substates in U are $\lesssim k_BT$, these substates are not metastable and therefore difficult to observe. (ii) During the main $U \rightarrow I$ folding transition, helix-1 is folded except for residue Asp3. This results in a native-like structure of HP35 which, however, exhibits significant fluctuations of helix-1 (unlocked state). (iii) The $I \rightarrow N_1$ transition corresponds to a change of residue Asp3 from various right-handed conformations to a stable α_L-helical conformation (locked state), which significantly compacts the hydrophobic core of the protein. A second native-like state N_2 exists, which can only be reached from N_1 and exhibits a flexible C-terminus of helix-3 (unlocked state).

Let us discuss these findings in the context of the ongoing discussion on the conformational heterogeneity and the existence of multiple folding pathways of HP35. Although the prestructuring of the protein in U certainly occurs via numerous pathways,[34,55] the substates of U are not metastable and therefore hardly observable in experiment. Apart from the unfolded basin, on the other hand, we have identified several native-like states, which are metastable and therefore could also be observed in principle. While the separation of the native states in a main state (N_1) and a native-like state (N_2) seems to prevail in our analysis, the structural details of the intermediate states were found to vary somewhat with temperature. Considering the Markov state model with two intermediate states I_1 and I_2 (Figure 6), we have found four folding pathways which collectively contribute to the average folding time. As the barrier between I_1 and I_2 is relatively small, however, in a more coarse-grained picture they may be merged into a single state I, thus giving a four-state model with a single folding pathway $U \rightarrow I \rightarrow N_1$. The precise definition of the intermediate states may depend on the temperature, the requested metastability of these states, and most likely also on the force field employed. Nonetheless, the analysis indicates that there are either a single or a few intermediates but not many.

On the experimental side, there is likewise clear evidence for some intermediate state(s) in the folding of HP35.[40,41,60] Performing triplet−triplet energy transfer experiments on various labeled variants of the villin headpiece, Reiner et al.[41] recently proposed a four-state model, which comprises a locked and an unlocked native state that differ mainly in the flexibility of helix-3, an intermediate state I that has still a native-like structure, and the unfolded state U. This scenario is remarkably similar to the four-state model discussed above. While our native-like state N_2 shows also an increased flexibility of helix-3, it is not *on-route*, as it can only be reached from N_1. The intermediate state I, on the other hand, exhibits an increased flexibility of helix-1.

Testing Theoretical Concepts of Protein Folding. *Hierarchical Energy Landscape.* As demonstrated by the dendrogram in Figure 2, the most probable path algorithm allows for a systematic and physically intuitive partitioning of the conformational space as a function of the minimum metastability $Q_{\min}$, which defines the minimum barrier height that separates a metastable state from its neighboring states. We have shown that already close to $Q_{\min} = 1$ the energy landscape consists of three basins U, I, and N containing the unfolded, intermediate, and native states, respectively. For $Q_{\min} \approx 0.7$, the latter two split up in two substates each. No further relevant subdivision of states occurs, because these five metastable states are already formed for $Q_{\min} \lesssim 0.1$. Hence the dendrogram nicely reveals a simple hierarchical structure of the folding free-energy landscape of HP35. This is in variance to the more "democratic" structure found for the energy landscape of small peptides with numerous similarly populated metastable states.[23]

Folding Funnel. Containing the vast majority of microstates, U is of entropic nature, whereas I and N contain only a few highly populated microstates and are thus of enthalpic nature. To proceed from the entropic unfolded state U to the enthalpic intermediate state I, the resulting entropic penalty needs to be compensated by a lowering of the potential energy. This is achieved during the "prestructuring" of the protein in U, from largely random structures to an almost folded system with preformed secondary structures. The energetic bias ΔE along the prestructuring process combined with the decrease of the configurational entropy ΔS when the number of accessible states decreases is the essence of the "folding funnel" diagram.[16,17] Usually plotted as the function $\Delta E(\Delta S)$, this diagram is not to be confused with the free-energy surfaces along some reaction coordinates as in Figure 7, which do not necessarily look like a funnel.[61]

Markov State Model and Network Theory. After dynamic coring, the metastable states are of Markovian nature, and their kinetics matches quantitatively the original MD data (Figure 5). The resulting Markov state model can be drawn as a network (Figure 6), where the main native state N_1 is rather a kinetic trap[62] than a network hub,[9,13] at least at temperatures close to the melting point. Concepts of complex network theory[63] such as hub, "small world", or "scale-free" naturally apply more to networks constructed on a microstate level.[9,13] Recalling that the main metastable states already emerge at very low metastability $Q_{\min}$, however, (see Figure 2), in the present study of HP35 it is not obvious to what extent these concepts are helpful to understand the folding process.

Dimensionality of Reaction Coordinates. By introducing various kinds of reaction coordinates, we have also considered several continuous representations of the free-energy landscape of HP35 (Figure 7). It has been found that a suitable one-dimensional reaction coordinate may reveal the main basins of the energy landscape and therefore provides a qualitative understanding of the folding process. However, none of the considered one-dimensional free-energy curves yielded all metastable states of HP35 or reproduced quantitatively the barriers between these states. As a more systematic approach to describe the conformational dynamics of biomolecules in reduced dimensionality, we have performed a dPCA of the HP35 trajectory, where 10−20 principal components need to be included in order to cover a substantial amount of the system's fluctuations. This relatively high dimensionality of the reduced model is caused by the many different possible molecular motions of the protein in the unfolded state.

Diffusion versus Subdiffusion. We have calculated the time-dependent mean squared displacement, MSD(t) $\propto t^{\alpha}$, along individual principal components as well as in the full-dimensional dPCA space. The first few components exhibit subdiffusive behavior with $\alpha \approx 0.5$ (Figure 8). The total MSD exhibits a biphasic behavior with $\alpha = 0.25$ for $t \lesssim 20$ ns and $\alpha = 0.05$ for longer times. This is because the MSD along the numerous higher components rises for shorter times than the MSD along the first component. Hence, the motion of HP35 is clearly subdiffusive, at least for the present trajectory and the choice of coordinates.

To summarize, we have shown that the most probable path method represents a powerful approach to study the folding dynamics of proteins. In straightforward extension of this work, the method is currently applied to investigate the functional dynamics of proteins as well as the aggregation of peptides.[64]

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Tables of detailed balance, figures of details of the principal component analyses, mean square displacements, helicity, and the structural characterization of the transition path region as well as states N_1 and N_2. This information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: stock@physik.uni-freiburg.de.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank D. E. Shaw Research for sharing their protein folding trajectories and Laura Riccardi, Florian Sittel, Rainer Hegger, Peter Hamm, and Francesco Rao for numerous instructive and helpful discussions.

REFERENCES

(1) van Gunsteren, W. F.; et al. Biomolecular modelling: goals, problems, perspectives. *Angew. Chem., Int. Ed.* **2007**, *45*, 4064−4092.

(2) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. Challenges in protein-folding simulations. *Nat. Phys.* **2010**, *6*, 751−758.

(3) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517−520.

(4) Bowman, G. R.; Voelz, V. A.; Pande, V. S. Taming the complexity of protein folding. *Curr. Opin. Struct. Biol.* **2011**, *21*, 4−11.

(5) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. Essential dynamics of proteins. *Proteins* **1993**, *17*, 412−425.

(6) Lange, O. F.; Grubmüller, H. Generalized Correlation for Biomolecular Dynamics. *Proteins* **2006**, *62*, 1053−1061.

(7) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.* **2007**, *126*, 244111.

(8) Rohrdanz, M. A.; Zheng, W.; Clementi, C. Discovering Mountain Passes via Torchlight: Methods for the Definition of Reaction Coordinates and Pathways in Complex Macromolecular Reactions. *Annu. Rev. Phys. Chem.* **2013**, *64*, 295−316.

(9) Rao, F.; Caflisch, A. The protein folding network. *J. Mol. Biol.* **2004**, *342*, 299−306.

(10) Buchete, N.-V.; Hummer, G. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.

(11) Noe, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. Constructing the Full Ensemble of Folding Pathways from Short Off-Equilibrium Simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011−19016.

(12) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. Progress and challenges in the automated construction of Markov state models for full protein systems. *J. Chem. Phys.* **2009**, *131*, 124101.

(13) Bowman, G. R.; Pande, V. S. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 10890−10895.

(14) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noe, F. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(15) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341−346.

(16) Onuchic, J. N.; Schulten, Z. L.; Wolynes, P. G. Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545−600.

(17) Dill, K. A.; Chan, H. S. From Levinthal to Pathways to Funnels: The "New View" of Protein Folding Kinetics. *Nat. Struct. Biol.* **1997**, *4*, 10−19.

(18) Neusius, T.; Daidone, I.; Sokolov, I. M.; Smith, J. C. Subdiffusion in Peptides Originates from the Fractal-Like Structure of Configuration Space. *Phys. Rev. Lett.* **2008**, *100*, 188103.

(19) Krivov, S. Is Protein Folding Sub-Diffusive? *PLoS Comput. Biol.* **2010**, *6*, e1000921.

(20) Milanesi, L.; Waltho, J. P.; Hunter, C. A.; Shaw, D. J.; Beddard, G. S.; Reid, G. D.; Dev, S.; Volk, M. Measurement of energy landscape roughness of folded and unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 19563−19568.

(21) Krivov, S. V.; Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766−14770.

(22) Hegger, R.; Altis, A.; Nguyen, P. H.; Stock, G. How Complex is the Dynamics of Peptide Folding? *Phys. Rev. Lett.* **2007**, *98*, 028102.

(23) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. Construction of the free energy landscape of biomolecules via dihedral angle principal component analysis. *J. Chem. Phys.* **2008**, *128*, 245102.

(24) Piana, S.; Laio, A. Advillin Folding Takes Place on a Hypersurface of Small Dimensionality. *Phys. Rev. Lett.* **2008**, *101*, 208101.

(25) Liu, F.; Du, D.; Fuller, A. A.; Davoren, J. E.; Wipf, P.; Kelly, J. W.; Gruebele, M. An experimental survey of the transition between two-state and downhill protein folding scenarios. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 2369−2374.

(26) Krivov, S. V. The Free Energy Landscape Analysis of Protein (FIP35) Folding Dynamics. *J. Phys. Chem. B* **2011**, *115*, 12315−12324.

(27) Jain, A.; Stock, G. Identifying metastable states of folding proteins. *J. Chem. Theory Comput.* **2012**, *8*, 3810−3819.

(28) Duan, Y.; Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, *282*, 740−744.

(29) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **2002**, *420*, 102.

(30) Fernández, A.; Shen, M. Y.; Colubri, A.; Sosnick, T. R.; Berry, R. S.; Freed, K. F. Large-scale context in protein folding: villin headpiece. *Biochem.* **2003**, *42*, 664−671.

(31) Lei, H.; Wu, C.; Liu, H.; Duan, Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4925−4930.

(32) Ensign, D. L.; Kasson, P. M.; Pande, V. S. Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J. Mol. Biol.* **2007**, *374*, 806−816.

(33) Rajan, A.; Freddolino, P. L.; Schulten, K. Going beyond clustering in MD trajectory analysis: an application to villin headpiece folding. *PLoS One* **2010**, *5*, e9890.

(34) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophys. J.* **2011**, *100*, L47−L49.

(35) Piana, S.; Lindorff-Larsen, K.; Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17845−17850.

(36) McKnight, C. J.; Matsudaira, P. T.; Kim, P. S. NMR structure of the 35-residue villin headpiece subdomain. *Nat. Struct. Biol.* **200**, *4*, 180−184.

(37) Kubelka, J.; Eaton, W. A.; Hofrichter, J. Experimental tests of villin subdomain folding simulations. *J. Mol. Biol.* **2003**, *329*, 625−630.

(38) Kubelka, J.; Hofrichter, J.; Eaton, W. A. The protein folding "speed limit". *Curr. Opin. Struct. Biol.* **2004**, *14*, 76−88.

(39) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. Sub-microsecond protein folding. *J. Mol. Biol.* **2006**, *359*, 546−553.

(40) Kubelka, J.; Henry, E. R.; Cellmer, T.; Hofrichter, J.; Eaton, W. A. Chemical, physical, and theoretical kinetics of an ultrafast folding protein. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18655−18662.

(41) Reiner, A.; Henklein, P.; Kiefhaber, T. An unlocking/relocking barrier in conformational fluctuations of villin headpiece subdomain. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 4955−4960.

(42) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712−725.

(43) Best, R. B.; Hummer, G. Optimized Molecular Dynamics Force Fields Applied to the Helix-Coil Transition of Polypeptides. *J. Phys. Chem. B* **2009**, *113*, 9004−9015.

(44) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950–1958.

(45) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926.

(46) Mu, Y.; Nguyen, P. H.; Stock, G. Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins* **2005**, *58*, 45–52.

(47) Hartigan, J. A.; Wong, M. A. A K-means clustering algorithm. *Appl. Stat.* **1979**, *28*, 100–108.

(48) Krivov, S.; 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.

(49) Bowman, G. R.; Meng, L.; Huang, X. Quantitative comparison of alternative methods for coarse-graining biological networks. *J. Chem. Phys.* **2013**, *139*, 121905.

(50) Varying the lag time around the chosen value of 2 ns (e.g., 1 or 5 ns) does not affect the results.

(51) Dill, K.; Bromberg, S. *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*, 2nd ed.; Garland Science: New York, 2010.

(52) MacKerell, A. D., Jr.; et al. All-atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586.

(53) MacKerell, A. D.; Feig, J. M.; Brooks, C. L. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

(54) Jain, A.; Hegger, R.; Stock, G. Hidden complexity of protein energy landscape revealed by principal component analysis by parts. *J. Phys. Chem. Lett.* **2010**, *1*, 2769–2773.

(55) Stock, G.; Jain, A.; Riccardi, L.; Nguyen, P. H. In *Protein and Peptide Folding, Misfolding and Non-Folding*; Schweitzer-Stenner, R., Ed.; Wiley: New York, 2012; p 57.

(56) Faradjian, A. K.; Elber, R. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **2004**, *120*, 10880–10889.

(57) Rao, F.; Karplus, M. Protein dynamics investigated by inherent structure analysis. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 9152–9157.

(58) Schütte, C.; Noe, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. Markov state models based on milestoning. *J. Chem. Phys.* **2011**, *134*, 204105.

(59) Banushkina, P. V.; Krivov, S. V. High-Resolution Free-Energy Landscape Analysis of alpha-Helical Protein Folding: HP35 and Its Double Mutant. *J. Chem. Theory Comput.* **2013**, *9*, 5257–5266.

(60) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. Simple few-state models reveal hidden complexity in protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 17807–17813.

(61) Karplus, M. Behind the folding funnel diagram. *Nat. Chem. Biol.* **2011**, *7*, 401–404.

(62) Dickson, A.; Brooks, C. L. Native States of Fast-Folding Proteins Are Kinetic Traps. *J. Am. Chem. Soc.* **2013**, *135*, 4729–4734.

(63) Helms, V. *Principles Of Computational Cell Biology*; Wiley-VCH: Weinheim, 2008.

(64) Riccardi, L.; Nguyen, P. H.; Stock, G. Construction of the Free Energy Landscape of Peptide Aggregation from Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2012**, *8*, 1471–1479.