JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Identifying Metastable States of Folding Proteins

Abhinav Jain and Gerhard Stock*

Biomolecular Dynamics, Institute of Physics, Albert Ludwigs University, 79104 Freiburg, Germany

Supporting Information

ABSTRACT: Recent molecular dynamics simulations of biopolymers have shown that in many cases the global features of the free energy landscape can be characterized in terms of the metastable conformational states of the system. To identify these states, a conceptionally and computationally simple approach is proposed. It consists of (i) an initial preprocessing via principal component analysis to reduce the dimensionality of the data, followed by k-means clustering to generate up to 10^4 microstates, (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. By adopting two well-studied model problems, hepta-alanine and the villin headpiece protein, the potential and the performance of the approach are demonstrated.

1. INTRODUCTION

While molecular dynamics (MD) simulations account for the structure and dynamics of biomolecules in microscopic detail, they generate huge amounts of data. To extract the essential information and reduce the complex and highly correlated biomolecular motion from $3N$ atomic coordinates to a few collective degrees of freedom, dimensionality reduction methods such as principal component analysis (PCA) are commonly employed.[1−5] The resulting low-dimensional representation of the dynamics can then be used to construct the free energy landscape $\Delta G(V) = -k_B T \ln P(V)$, where P is the probability distribution of the molecular system along the principal components $V = \{V_1, V_2, \ldots\}$. Characterized by its minima (which represent the metastable conformational states of the systems) and its barriers (which connect these states), the energy landscape allows us to account for the pathways and their kinetics occurring in a biomolecular process.[6−8]

Recent simulations of peptides, proteins, and RNA have shown that in many cases the free energy landscape can be well characterized in terms of metastable conformational states.[9−12] As an example, Figure 1A shows a two-dimensional free energy landscape of hepta-alanine[13] (Ala_7) obtained from an 800 ns MD simulation with subsequent PCA of the ϕ, ψ backbone dihedral angles (see section 3). The purple circles on the contour plot readily indicate about 30 well-defined minima (or basins) of the energy surface. They correspond to metastable conformational states, which can be employed to construct a transition network of the dynamics of the system.[9−27] The network can be analyzed to reveal the relevant pathways of the considered process, or to discuss general features of the system such as the topology (i.e., a hierarchical structure) of the energy landscape and network properties such as scale-freeness. Also, in protein folding, metastable states have emerged as a new paradigm.[9,11] Augmenting the funnel picture of folding, the presence of thermally populated metastable states may result in an ensemble of (rather than one or a few) folding pathways.[19] Moreover, they can result in kinetic traps, which may considerably extend the average folding time. As an example, Figure 1B shows the free energy landscape of the villin headpiece subdomain,[28−37] obtained from a PCA of extensive folding trajectories by Pande and co-workers[33] (see section 4). Due to the high dimensionality of the energy landscape, the two-dimensional projection only vaguely indicates the multiple minima of the protein.

Although energy landscapes as in Figure 1 appear to easily provide the location of the energy minima, in general it turns out that metastable states are surprisingly difficult to identify, even for a seemingly simple system like Ala_7. To partition the conformational space into clusters of data points representing the states, one may use either geometric clustering methods such as k-means,[38] which require only data in a metric space, or kinetic clustering methods, which additionally require dynamical information on the process.[9−27] While geometrical methods are fast and easy to use, they show several well-known flaws. For example, since they usually require one to fix the number of clusters k beforehand, it easily happens that one combines two separate states into one (if k is chosen too small) or cuts one state into two (if k is chosen too large). Another problem is the appropriate definition of the border between two clusters. From a dynamical point of view, the correct border is clearly located at the top of the energy barrier between the two states. Using exclusively geometrical criteria, however, the middle between the two cluster centers appears as an obvious choice, see Figure 2A. As a consequence, conformational fluctuations in a single minimum of the energy surface may erroneously be taken as transitions to another energy minimum, see Figure 2B. The same problem may occur for systems with low energy barrier heights, say, $\Delta G_B \leq 3k_B T$.

Kinetic cluster algorithms may avoid these problems by using the dynamical information provided by the time evolution of the MD trajectory.[9−27] In a first step, the conformational space is partitioned into disjoint *microstates*, which can be obtained, e.g., by geometrical clustering (see section 2.1). Employing these microstates, we calculate the transition matrix $\{T_{mn}\}$ from the MD trajectory, where T_{mn} represents the probability that

Special Issue: Wilfred F. van Gunsteren Festschrift

Received: January 31, 2012
Published: March 26, 2012

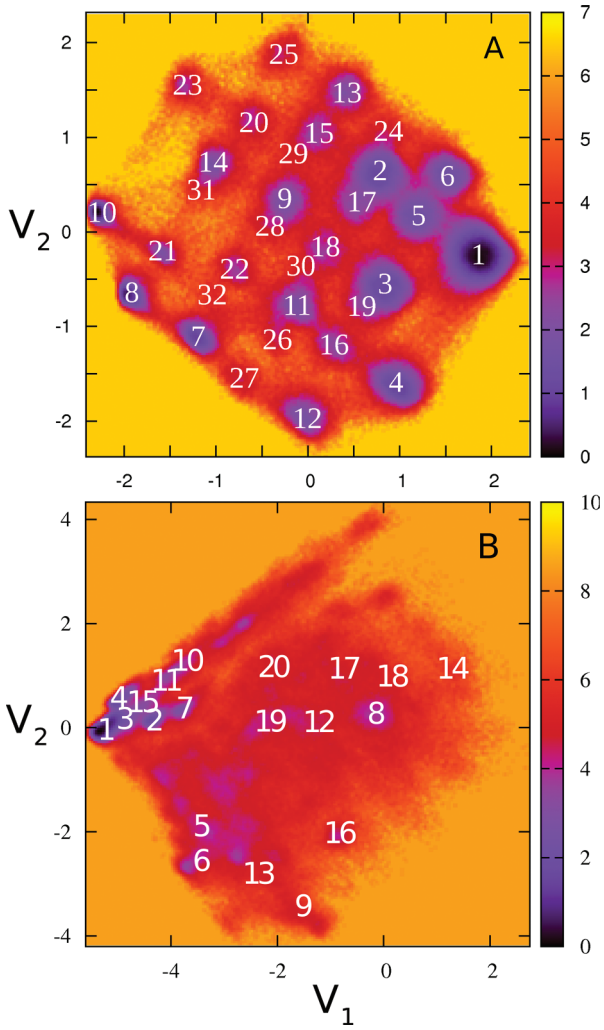

Figure 1. Free energy landscape (in units of k_BT) of (A) hepta-alanine and (B) the villin headpiece as a function of the first two principal components. The metastable conformational states of the system show up as minima of the energy surface (see Tables 1 and 2 for the labeling).

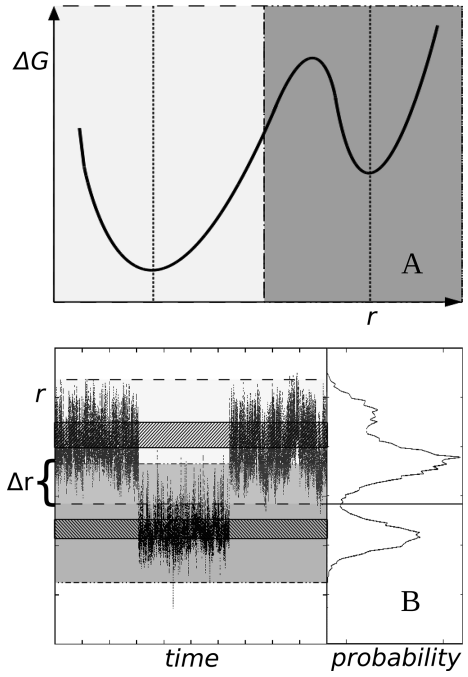

Figure 2. Common problems in the identification of metastable conformational states, illustrated for a two-state model, which is represented by a schematic free energy curve along some reaction coordinate r. (A) Although the top of the energy barrier between the two states clearly represents the correct border, geometrical clustering methods may rather choose the geometrical middle between the two cluster centers. (B) Typical time evolution of a MD trajectory along r for the two-state model and the corresponding probability distribution $P(r)$. Low barrier heights or an inaccurate definition of the separating barrier may cause *intrastate* fluctuations to be mistaken as *interstate* transitions. The overlapping region of the two states is indicated by Δr. The introduction of cluster cores (shaded areas) can correct for this.

state n changes to state m within a certain lag time τ. If the free energy landscape can be characterized by metastable conformational states (or energy basins) separated by well-defined energy barriers, there exists a time scale separation between fast intrastate motion (i.e., transitions between microstates within the energy basin) and slow interstate motion (i.e., transitions between microstates of different energy basins). By applying a suitable transformation, the transition matrix can therefore be converted into an approximately block-diagonal form, where each block corresponds to a metastable state.[14] In other words, in order to identify the metastable states of a system, we need to merge all microstates that are in the same energy basin.

While the basic idea is straightforward, the practical application of kinetic clustering to high-dimensional MD data faces several challenges. First, the choice of microstates is crucial, since they should represent the conformational space with sufficient resolution while their number still needs to be computationally manageable. The latter is particularly important in order to achieve converged estimates of the transition probabilities T_{mn} between these states. Also, the construction of metastable states from this transition matrix is an active field of research. It can be achieved, for example, by analyzing the eigenfunctions of the transition matrix[10,14] or by employing steepest-decent-type algorithms.[22,23,39,40] The choice of the method may depend to a large extent on the application in mind. For example, an important purpose of kinetic clustering is the construction of discrete Markov state models, which approximate the dynamics of the system through a memoryless jump process.[9−27] Markov models have become popular because they hold the promise of predicting the long-time dynamics of a system from relatively short trajectories. On the other hand, there has been growing interest in the interpretation of biomolecular dynamics in terms of the global features of energy landscapes. To this end, it is desirable to restrict the discussion to a relatively small number of well-defined metastable states that constitute the energy basins of the system.

In this paper, we propose an approach to identify the main metastable states of a biomolecular system, in order to achieve a global characterization of the free energy landscape. The method is simple and efficient and is therefore applicable to large biomolecular systems, which has become important with the availability of MD trajectories in the mirosecond and even millisecond range.[41] The approach detailed in section 2 consists of (i) initial PCA preprocessing to reduce the dimensionality of the data, followed by k-means clustering to generate up to 10^4 microstates, (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. As a first simple example, the method is applied to hepta-alanine, leading to a model with about 30 long-lived conformational states that obey Markovian dynamics. To demonstrate the potential of the model, we also consider the folding dynamics of the villin headpiece subdomain.[28−37]

2. THEORY AND METHODS

2.1. Generation of Microstates: dPCA and *k*-means. To provide a suitable basis for the subsequent cluster analysis, we first need to drastically reduce the large number of MD snapshots ($\sim 10^5$ to 10^7, depending on the read-out time step and the trajectory length) to a computationally manageable number of microstates ($\sim 10^3$ to 10^4). To this end, it has been proposed to employ rotamer substates of each protein residue (e.g., α-helical, polyproline II, or extended β structures) and construct a product basis of these states.[9,10,13,42] (For example, the microstate associated with the all-extended structure of a tripeptide would be labeled by $\beta\beta\beta$.) Although this procedure formally scales exponentially with the size of the system, in practice the maximum number of resulting microstates is limited by the number of MD snapshots. The approach has successfully been used to generate microstates for various peptides and small proteins.

Alternatively, microstates can be conveniently obtained from efficient geometrical clustering algorithms such as *k*-means.[38] Choosing a sufficiently large number of states ($k \sim 10^3$ to 10^4), this quite general approach can almost be employed in a black-box manner. In this case, though, it is quite helpful to first apply a PCA to the MD trajectory, in order to reduce the dimensionality of the data from $3N$ (N being the number of particles) to the order of 10. By including a suitable number of principal components (say, 5−20 for a typical protein[1]), the PCA data contain the desired amount of fluctuations of the trajectory (say, 50−90%). In the examples below, we have employed the dihedral angle PCA (dPCA),[43,44] which uses the sine/cosine-transformed φ and ψ dihedral angles of the protein backbone. 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.

2.2. Dynamical Clustering: Most Probable Path Algorithm. Given the (still relatively large number of) microstates, we next want to merge together all microstates that are in the same energy basin. The underlying assumption of a dynamical clustering approach is that the free energy landscape is characterized by metastable conformational states (also referred to as clusters or energy basins), which are separated by well-defined barriers. This results in a time scale separation of fast motion within the metastable states and slow interstate motion (i.e., rare interstate transitions). To illustrate the basic idea of the most probable path algorithm, Figure 3 shows a one-dimensional model consisting of two energy basins, *A* and *B*, which are represented by six microstates. Microstates 1−4 lie in basin *A* and microstates 5 and 6 lie in basin *B*. Given an MD trajectory that reproduces the one-dimensional free energy curve of the model, we can easily compute a transition matrix $\{T_{ij}\}$ for these microstates, where T_{ij} represents the probability of a state i to change to state j within a predefined lag time τ. T_{ij} can be employed to decide which microstates belong to which energy basin. The procedure is again based on a time scale separation argument that (for not too small barrier heights) the probability to cross the barrier is

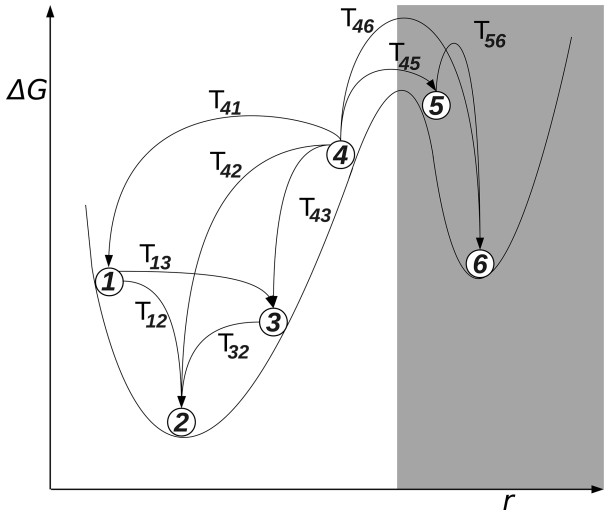

Figure 3. Illustration of the most probable path algorithm for a simple model consisting of two energy basins, *A* and *B*, and six microstates, 1−6. By following the path of the maximal transition probabilities T_{ij}, we can identify the microstates belonging to basins *A* and *B*, respectively.

significantly lower than the probability of transitions within the basins. For example, in Figure 3, microstate 1 will have a significantly higher probability to go to microstates 2, 3, and 4 (or to stay in 1) than to go to microstates 5 and 6.

Here is the basic idea of the algorithm. Starting from a given microstate (e.g., from state 4), we can calculate its most probable path by (i) evaluating the transition probabilities of this state and (ii) choosing the most probable transitions. For example, starting from state 4, we calculate T_{4j}. If, say, $\max(T_{4j}) = T_{43}$, we go to microstate 3. Then, we calculate T_{3j} and move on to the microstate with the highest transition probability, e.g., $\max(T_{3j}) = T_{32}$. This goes on until we reach a state i with $\max(T_{ij}) = T_{ii}$; i.e., it is most likely to stay in this state. In Figure 3, for example, we may find that $\max(T_{2j}) = T_{22}$. 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, thereby avoiding the problems of geometrical clustering methods discussed in Figure 2B. The method resembles previously suggested basing attraction techniques which, however, were based on the potential energy rather than on the free energy.[39,40] Similar strategies have been also applied in network theory.[22,23]

Figure 4 shows how the most probable path algorithm can be implemented in a dynamic clustering scheme. Starting with N microstates ($i = 1, ..., N$), we calculate their transition probabilities $\{T_{ij}\}$ and free energies $G_i = -k_B T \ln P_i$, with P_i being the population probability of state i. For each state i, we then evaluate the most probable path, $i \rightarrow i_1 \rightarrow ... \rightarrow i_f$ and merge all states that visited the same state of lowest free energy G_i. By choosing the state with minimal G_i (or highest population probability P_i) rather than the one with the highest metastability T_{ii}, we prefer a well sampled state over a lowly sampled state with similar metastability. This results in N_{new} "collective" states for which we again calculate all most probable paths and merge the states correspondingly. The procedure is repeated until the number of new states coincides with the state

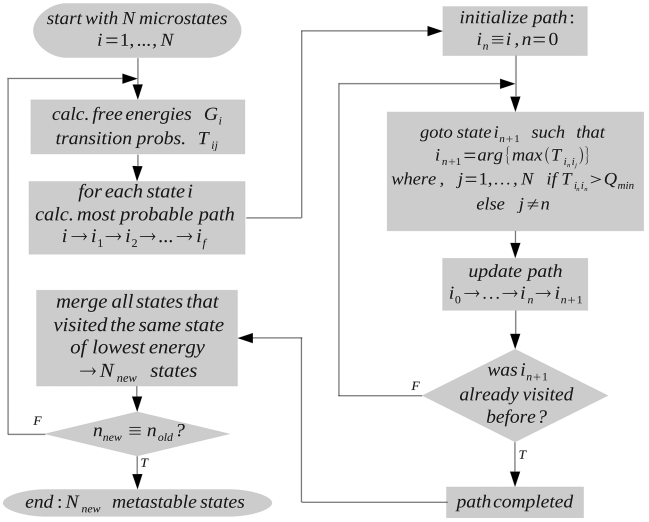

Figure 4. Scheme of the dynamic clustering method (left), using the most probable path algorithm (right).

number of the previous cycle, thus giving N_{new} metastable states.

The right-hand side of Figure 4 explains in more detail how the most probable path is found. The algorithm consists of the following steps:

1. Initialize path i_n for a given microstate i, such that initially $i_n = i$ for $n = 0$.
2. Find the next state i_{n+1}, that is, the state that has the largest transition probability to it, $T_{i_n i_{n+1}} = \max(T_{i_n i_j})$. This can be either any other state ($j \neq n$) or the same state ($j = n$). Remaining in state i_n is only allowed if its metastability $T_{i_n i_n}$ is larger than a predefined cutoff value Q_{min}. In this way, the algorithm allows us to request a minimum metastability of the resulting metastable states.
3. If the new state i_{n+1} has not been visited in the path before, we update the path (i.e., $i_0 \rightarrow i_1 \rightarrow ... \rightarrow i_n \rightarrow i_{n+1}$) and calculate the next microstate of the most probable path. Otherwise, the set of microstates visited on the most probable path is completed.

In step 2, we may face the problem of finding the same transition probability $T_{i_n j}$ for several states j. This happens in particular during the first iteration, when we have many (often insufficiently sampled) microstates. In these cases, we choose as the next state the one with the closest distance between the two cluster centers. Alternatively, we also attempted a random selection of the next state, which requires, however, that we run the algorithm several (say, 10) times. For a suitable choice of τ and Q_{min}, one typically finds good overall agreement of the metastable states obtained by these two techniques, which indicates a stable partitioning.

2.3. Boundary Corrections: Cluster Cores. If the energy barrier between two metastable states is on the order of $k_B T$, the underlying assumption of a time scale separation between fast intrastate and slow interstate motion becomes questionable. As a consequence, intrastate conformational fluctuations may reach over the interstate borders defined by the clustering, see Figure 1C. Because intrastate fluctuations by mistake are interpreted as interstate transitions, this effect can lead to an artificial short lifetime of a metastable state. To avoid this problem, one may restrict the counting of transitions to events that remain at least some minimum time in the new state. Alternatively, we can employ the concept of milestones[45] or cluster cores.[17] A cluster core is defined as the region around a cluster center that contains a certain percentage of its population (shaded areas in Figure 1C). The idea is to require that a transition from a cluster to another cluster must reach the core region of the other cluster. Otherwise, it is not counted as a transition. The quality of this approximation was recently analyzed in ref 27. Coring schemes were also employed in refs 18 and 24.

To implement this idea, we start with the MD data given in a suitable representation such as the time series of its first few principal components. By discretizing the principal components into bins, we first calculate for each cluster i one-dimensional probability distributions along these coordinates. Next, the center bin of each cluster is determined as the minimum of the free-energy along each curve, and one-dimensional cluster cores are defined as the portion of bins around the cluster center that contribute to the first, say, 50% of the population of the cluster. This yields the set of points lying within a given one-dimensional core. The actual core of the cluster i in all considered dimensions is then given by the common subset of all points of the individual one-dimensional cores. Hence, the multidimensional cluster core is in general smaller than the chosen size (e.g., 50% of the population) of the one-dimensional cores. Finally, we convert the original MD time series to a time series of cluster cores $i(t)$. This is done by checking if all coordinates of a given MD snapshot are within the limits of the above-defined cluster cores. In between two cluster cores, we set $i(t)$ to the last visited core. From this time series, the transition matrix of the cored clusters is readily calculated.

To explain the limitations and virtues of the proposed coring procedure, we reconsider Figure 1C, which illustrates why intrastate fluctuations may be interpreted as interstate transitions. Obviously, the problem is caused by the overlapping region (indicated by Δr) of the two uncored clusters. Hence, the cores of the two clusters need to be chosen small enough that the core regions are just short of touching the overlapping region. If this is done in a similar way for both states, the cored population probabilities of the metastable states are typically similar to their original populations (see Tables 1 and 2 below), since on average the gain of population due to an outgoing trajectory is approximately compensated by the loss due to an incoming trajectory. The procedure may become questionable, however, in the case of entropic or diffusive states. Diffusive clusters contain many weakly populated microstates and therefore give rise to a spatially extended and flat minimum of the free energy. As a consequence, the overlapping region becomes large and the cores necessarily small, such that the cored populations of a diffusive state may be significantly smaller than its original population. On the other hand, we note that the coring may correct for improper assignments of microstates close to a barrier, which are often insufficiently sampled and therefore difficult to assign to a metastable state by the most probable path algorithm. The explanation above shows that the precise assignment of these barrier microstates does not really matter during the dynamic clustering if coring is used afterward.

3. A SIMPLE EXAMPLE: HEPTA-ALANINE

Polyalanines represent a popular test-bed to study the potential and performance of dimensionality reduction approaches and

Table 1. Characterization of the 10 Most Populated Metastable States of Ala_7, as Obtained from the Most Probable Path Algorithm for τ = 10 ps and Q_{min} = 0.7[a]

#	structure	n	P[%]	$\sum P$ [%]	P_c [%]	Q [%]	Q_c [%]
1	$\beta\beta\beta\beta\beta$	228	24.18	24.18	23.65	93.4	95.2
2	$\beta\beta\alpha\beta\beta$	79	7.58	31.76	7.61	89.8	92.2
3	$\beta\beta\beta\alpha\beta$	61	7.15	38.90	7.45	87.9	93.0
4	$\beta\alpha\beta\beta\beta$	63	6.94	45.84	6.92	89.3	92.0
5	$\beta\beta\beta\beta\alpha$	61	5.37	51.21	5.30	88.2	91.1
6	$\alpha\beta\beta\beta\beta$	53	5.10	56.32	5.12	90.2	93.0
7	$\beta\alpha\alpha\alpha\alpha$	34	3.80	60.12	3.80	92.6	95.3
8	$\beta\alpha\alpha\alpha\beta$	37	3.83	63.95	3.77	88.7	92.6
9	$\beta\beta\alpha\alpha\beta$	37	3.39	67.33	3.48	86.8	91.0
10	$\beta\alpha\alpha\beta\beta$	31	3.17	70.51	3.33	85.8	90.8

[a]Shown are secondary structure, number of microstates n, population P, cumulative population $\sum P$, and metastability Q of the states. Subscript c refers to the respective quantities after coring. The five-letter code for the structure refers to the conformation of the inner five residues, which is described by either α-helical (α) or β/polyproline II (β) structure.

Table 2. Structure, Number of Microstates n, Population P, Cumulative Population $\sum P$, and Metastability Q of the 12 Most Populated Metastable States of HP-35[a]

#	structure	n	P [%]	$\sum P$[%]	P_c [%]	Q [%]	Q_c [%]
1	FFFFF	2264	29.23	29.23	28.86	99.3	99.7
2	$I_1I_1P_1FP_1$	663	8.34	37.56	4.49	97.8	98.8
3	FFP_2FF	386	4.91	42.48	4.23	97.9	99.3
4	NFFFF	406	4.81	47.29	4.87	99.7	99.8
5	FFFFN	230	3.00	50.29	4.31	96.2	99.4
6	$FFFUP_2$	178	2.15	52.44	2.03	98.6	99.8
7	$P_1I_1P_2FF$	117	1.66	54.10	1.50	97.7	98.7
8	P_2I_2FFN	123	1.64	55.74	1.77	99.9	99.9
9	P_1PFFP_1	97	1.35	57.10	2.15	97.3	99.3
10	FRUFF	103	1.14	58.24	1.03	99.6	99.9
11	I_2FFFF	85	1.10	59.34	0.65	97.3	99.1
12	UFFFN	72	0.99	60.33	0.99	97.8	99.3

[a]Subscript c refers to the respective quantities after coring. Following ref 34, the conformation of the five secondary structure parts (helix-1, turn-1, helix-2, turn-2, and helix-3, see Figure 9) is described by a five-letter code using the labels U (unfolded), R (random), I_n (intermediate), P_n (partially folded), N (near native), and F (folded).

Markov state models.[10,16,44] For example, Altis et al.[13] performed a comprehensive MD study of the free energy landscape of hepta-alanine (Ala_7). On the basis of an 800 ns MD simulation in explicit water, they showed that a five-dimensional dPCA energy landscape provides a suitable and accurate representation of the conformational dynamics of Ala_7 (cf. Figure 1A). To identify the metastable conformational states of the system, k-means clustering was employed, which resulted in k = 23 states. This number is somewhat smaller than 2^5 = 32, which is the expected number of states when the conformation of the inner five residues is described by either the α-helical (α) or β/polyproline II (β) structure.[46] Using these rotamer states to characterize the 23 k-means states, it was indeed found that several of the lowly populated k-means states were given as a mixture of rotamer states. Moreover, it was shown that the life times of the k-means states were on average surprisingly short (~ 10 ps), thus hampering a Markovian modeling.

It is interesting to study if these findings really are features of the dynamics of hepta-alanine or rather are caused by problems associated with the definition of the metastable states. To this end, we applied the above-described clustering approach to the five-dimensional dPCA trajectory of ref 13. First, we performed a k-means clustering on the data to generate a suitably fine grid of microstates. We used k = 1000, which is much larger than the anticipated number of metastable states. Next, we partitioned these microstates into metastable states by using the most probable path algorithm. As explained above, the number of metastable states generally depends on the lag time used in the calculation of the transition matrix as well as on the required minimum metastability Q_{min} of the states. There are two limiting cases of Q_{min} which hold for all lag times: If no metastability is required (Q_{min} = 0), all 1000 microstates qualify as metastable states; in the opposite limit of Q_{min} = 1, all microstates are merged into a single metastable state.

For intermediate Q_{min}, the number of metastable states depends on the lag time τ. Choosing τ between 1 and 100 ps, Figure 5 shows the results for the number of metastable state n

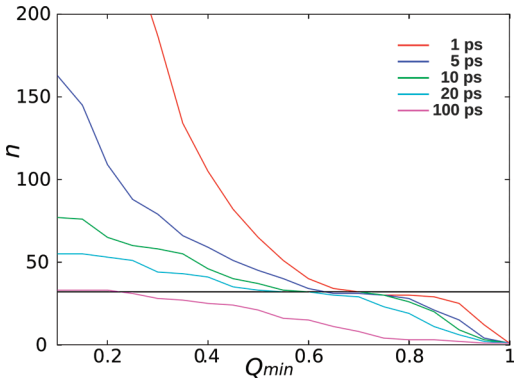

Figure 5. Number of metastable states $n(Q_{min})$ of Ala_7, as obtained from the most probable path algorithm using various lag times as indicated. For all considered lag times, $n(Q_{min})$ exhibits a plateau at n = 32 (dotted line).

as a function of the required minimum metastability Q_{min}. (That is, the algorithm was run multiple times, where Q_{min} was increased from 0 to 1 in steps of 0.05.) For all lag times, we find an initial decay of $n(Q_{min})$ which levels off at $n \approx 32$, before it ultimately decays to zero. That is, for a quite large range of parameters, the most probable path algorithm gives exactly the number of metastable states that is expected for this simple molecule. We note that our previous k-means clustering[13] resulted in 23 metastable states and led to different state combinations as we obtain here, thus clearly reflecting the limits of the geometric clustering method.

As may be expected, Figure 5 reveals a certain trade-off between Q_{min} and τ. For example, to obtain $n \approx 32$, only a small metastability ($Q_{min} \approx 0.1$) is required for τ = 100 ps, while for τ = 1 ps we need to request a large metastability ($Q_{min} \approx 0.7$). In practice, one wants to choose the lag time (i) short enough that the dynamics of interest is resolved in time but (ii) long enough that one can assume memoryless transitions between the states, in order to construct a Markov model. These conditions can be tested, e.g., by analyzing the autocorrelation function of the time series or testing the Chapman Kolmogorov condition (see below). The required minimum metastability, on the other hand, is dictated by the energy barriers between the metastable states. That is, Q_{min} should be as large as possible to get states

that are truly metastable but still small enough that the metastable states are well separated by the barriers. In other words, by adjusting Q_{min}, we can choose the energy resolution of the resulting state representation of the energy landscape.

This virtue of the most probable path algorithm can be used to construct a dendrogram shown in Figure 6, which

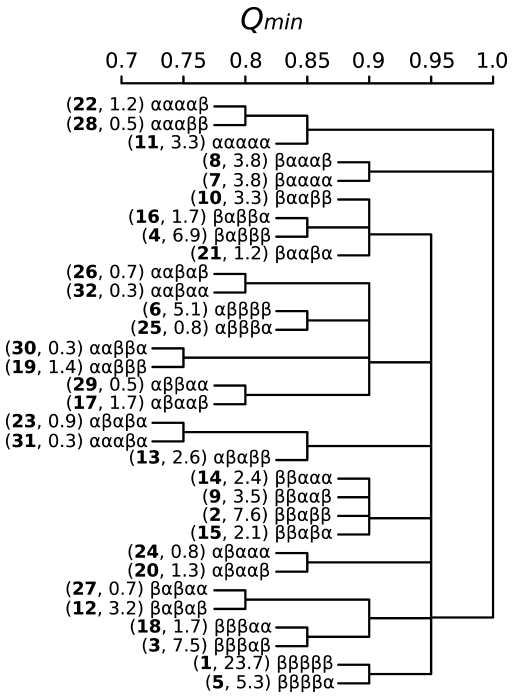

Figure 6. Dendrogram showing how the metastable states of Ala_7 merge into energy basins with increasing minimum metastability Q_{min}. Numbers in parentheses indicate the number (in bold face) and the thermal population (in %) of the states. The five-letter code of the secondary structure is as in Table 1.

demonstrates how various metastable states merge into basins with increasing minimum metastability Q_{min}. For example, while we get 32 states for $0.6 \leq Q_{min} \leq 0.7$, we find for $Q_{min} = 0.8$ that in six cases two states were merged together, resulting in 26 remaining states. Note that the highly populated states are usually not affected by this coarse-graining; that is, we obtain the same main states in a large range of Q_{min}. Hence, the dendrogram nicely illustrates the topology and the hierarchical structure of the free energy landscape of Ala_7. Since Q_{min} determines the minimum barrier heights of the energy landscape, the dendrogram also reveals how structurally similar conformations (as indicated by the five-letter code) are dynamically linked. For example, the barrier between the *ααααβ* state and the *αααββ* state is lower than the barrier between *ααααβ* and the all-helical state *ααααα*. In other words, it is more likely to observe the conformational transition *ααααβ* → *αααββ* than the transition *ααααα* → *ααααβ*. Similarly, we see that an $\alpha \rightarrow \beta$ transition is energetically easier in the case of *ααααβ* → *αααββ* than for *ααααβ* → *βαααβ*.

Table 1 characterizes the 10 most populated metastable states of Ala_7, which contain ~70% of the population of the in total 32 states. In agreement with experimental results,[47] we find that the β/polyproline II structure clearly dominates the α-helical structure, with the all-extended state *βββββ* being the most populated state. Interestingly, we notice that the number of microstates n of each metastable state is approximately proportional to the state population. This indicates that the k-means microstates are fine-grained enough ($k = 1000$) that they contain a similar amount of conformational space. More generally speaking, it reveals that the most probable path algorithm generates reasonably populated states (see also S1 in the Supporting Information, which shows the distribution of all metastable states in terms of their number of microstates). All states exhibit high metastability ($Q \approx 0.9$), which is somewhat increased by the coring (see below). The state population changes only a little with coring, which indicates that the coring procedure works well in the case of Ala_7.

We are now in a position to study the dynamical properties of the metastable states of Ala_7. To this end, it is instructive to consider the time-dependent population probability $P_i(t)$ of a specific metastable state i, given that the system started in state i at time $t = 0$. $P_i(t)$ can be readily calculated from the MD data by simply computing the distribution of times that the system stays in this state (i.e., the lifetime distribution, or eqivalently, the escape-time distribution). Choosing states $i = 1$, 3, and 5 as representative examples, Figure 7A shows the time evolution of

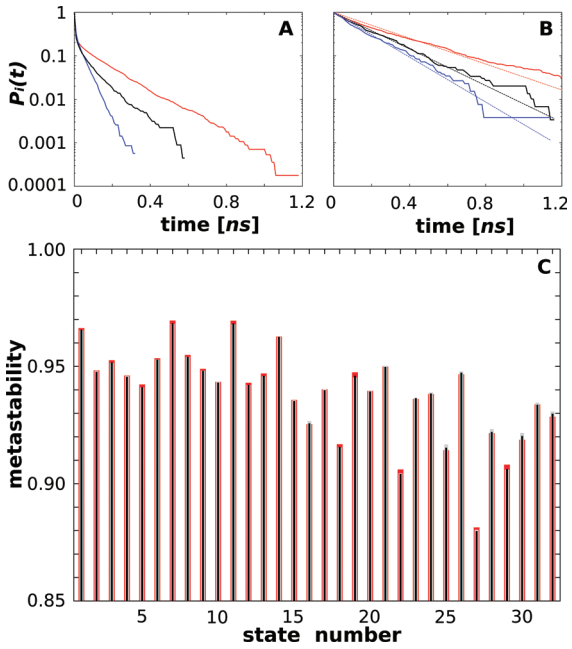

Figure 7. Markovianity of the 32-state model of Ala_7 as revealed by the time evolution of the population probability $P_i(t)$ of the metastable states $i = 1$ (red), 3 (black), and 5 (blue), obtained (A) without and (B) with the introduction of cluster cores. The dashed lines in panel B represent the corresponding results of the Markov model constructed from the cored states. (C) Verification of the Chapman–Kolmogorov equation for the metastable states i of Ala_7 after coring. Compared are the metastabilities $T_{ii}(\tau = 10$ ps; gray), T_{ii}^{2} (5 ps; red), and T_{ii}^{10} (1 ps; black).

the population of these states. In all cases, $P_i(t)$ is seen to perform a biexponential decay: a rapid initial decay on the time scale of the lag time (10 ps) and a slower relaxation with decay times of 180 ps (state 1), 120 ps (state 3), and 50 ps (state 5). These findings are in line with the results of Altis et al.[13] and suggest that the dynamics of the above introduced metastable states is not of Markovian nature.

A closer analysis of the data reveals, however, that the rapid initial decay is caused by boundary errors, as indicated in Figure 2B. As discussed in section 2.3, this problem can be avoided,

when one introduces cluster cores and requires that a transition from a state to another state must reach the core region of the other state.[17] Otherwise, the event is not counted as a transition. Doing so, Figure 7B shows that the resulting time-dependent populations indeed lack the initial picosecond decay and exhibit a single-exponential relaxation. As a consequence, the mean lifetimes of 300 ps (state 1), 210 ps (state 3), and 170 ps (state 5) are significantly longer than the decay times found for the data without coring.

To test the Markovianity of the metastable states after coring, we consider the Chapman−Kolmogorov equation:

$$P(n\tau) = P(0)\ T(n\tau) = P(0)\ T^n(\tau)$$

where $P(t) = (P_1(t), ..., P_{32}(t))$ represents the state vector at time t and $T(\tau)$ is again the (row-normalized) transition matrix obtained for lag time τ. Choosing τ = 1, 5, and 10 ps, Figure 7 reveals that all 32 metastable states of Ala_7 exhibit almost perfect Markovian behavior already on a time scale of 1 ps. By calculating the time-dependent population probability $P_i(t)$ from this Markov model, in fact, Figure 7B shows excellent agreement between the results obtained from the model and the results directly obtained from the MD data. Performing the Chapman−Kolmogorov test for the uncored states, on the other hand, we found significant deviations from Markovianity (data not shown). We note in passing, that the Markovianity property was found to deteriorate for a number of metastable states that is smaller or larger than 32.

4. ENERGY LANDSCAPE OF VILLIN HEADPIECE PROTEIN

We now apply the above introduced methodology to the description of the free energy landscape of a fast folding variant of the villin headpiece subdomain (HP-35), which has been studied in numerous experimental and computational works.[28−37] As in previous work,[34] we use the extensive (about 350 μs in total) simulations of HP-35 in explicit solvent which were carried out by Pande and co-workers on the Folding@home distributed computing platform.[33] Despite its small size (35 aa), HP-35 shows many of the typical properties of larger proteins such as a compact core and tertiary contacts. To obtain a first impression, Figure 1B shows the free energy landscape of HP-35 along the first two principal components. Compared to the energy surface of Ala_7 with numerous well resolved metastable states (Figure 1A), the energy landscape of the small protein is diffusive and quite structureless. By performing a "PCA by parts" of the various secondary structure elements (three helices and two turns) of HP-35, however, it has been shown that the energy landscape of the parts is well structured and reveals multiple minima.[34] The apparent lack of structure in Figure 1B therefore simply reflects the fact that the dimensionality of the energy landscape of HP-35 is high and cannot be resolved in a two-dimensional projection of the data.

As in the case of Ala_7, we first performed a dPCA of the data[48] and chose for the subsequent analysis the principal components, which exhibited non-Gaussian distributions.[34] In this 40-dimensional space, we employed k-means to generate 8000 microstates of HP-35. We then applied the most probable path algorithm in order to generate sets of metastable states as a function of the minimum metastability Q_{min}. Using various lag times from 0.25 to 5 ns, Figure 8 reveals that the number n of metastable states monotonically decays with Q_{min}. This behavior is different from the case of Ala_7, which showed a plateau at a certain number of states (Figure 5). As the choice

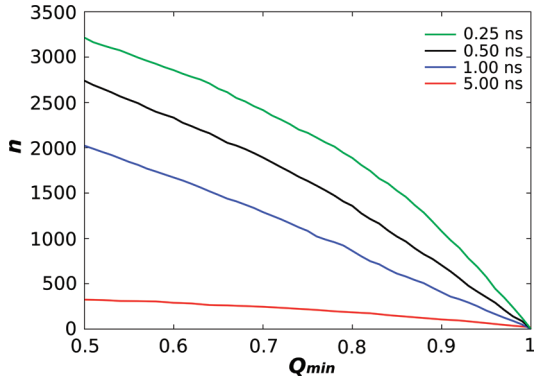

Figure 8. Number of metastable states $n(Q_{min})$ of HP-35, shown for various lag times.

of Q_{min} defines the minimum barrier height that a metastable state is separated from its neighboring states, a plateau indicates a separation of time scales in the dynamics of Ala_7. Interestingly, this clear separation of energy or time scales does not exist in the case of HP-35. Rather, we find that there are numerous (at least partly) overlapping time scales of the system, which are a consequence of the high dimensional dynamics of the system.

While the lag time can again be chosen by checking the Markovianity of the dynamics (see Figure 10 below), various

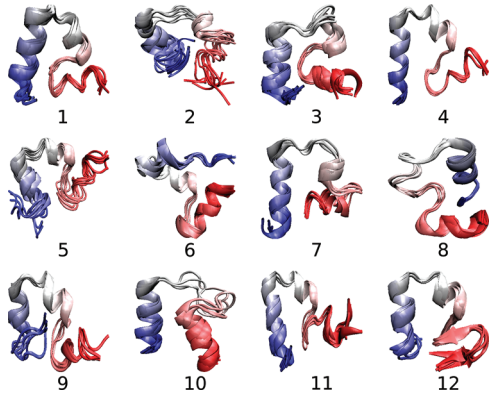

Figure 9. Molecular structure of the 12 highest populated metastable states of HP-35. The protein consists of three helices, helix-1 (red), helix-2 (gray), and helix-3 (blue), which are connected by two turns, turn-1 and turn-2, respectively. To indicate the heterogeneity of the states, 10 representative MD snapshots are shown for each cluster.

options for Q_{min} and the resulting number n of metastable states are possible. In practice, one wants to (i) obtain a reasonable (i.e., not too large) number of metastable states and (ii) request that the underlying microstates of a metastable state should have similar molecular structures (which favor a large n). As a compromise between these opposing objectives, we chose τ = 0.5 ns and Q_{min} = 0.95. This results in ∼300 metastable states, where the highest populated 12 states carry about 60% of the total population. Table 2 comprises various properties of these states, and Figure 9 shows their molecular structure. Again, we find that the number of microstates n of each metastable state is approximately proportional to the state population (see S1 for the distribution of all metastable states). To indicate the heterogeneity of the states, 10 representative MD snapshots are shown in Figure 9 for each cluster. With the

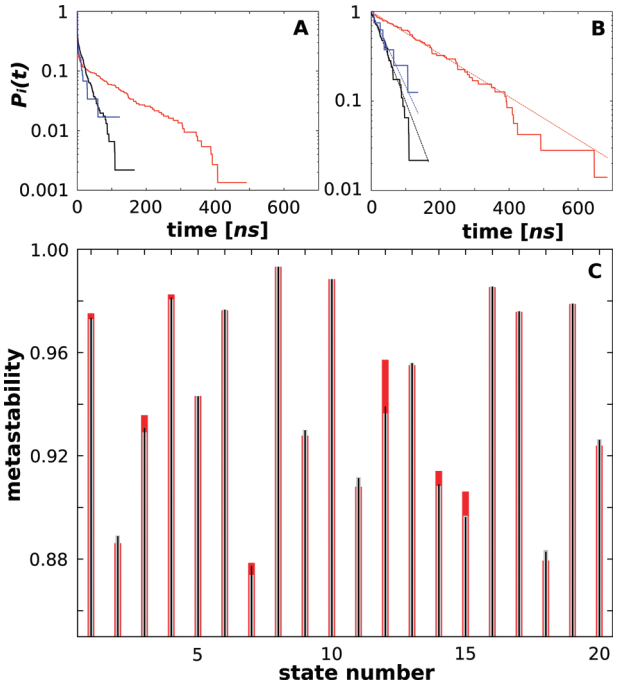

Figure 10. Markovianity of the 300-state model of HP-35 as revealed by the time evolution of the population probability $P_i(t)$ of the metastable states i = 1 (red), 2 (black), and 12 (blue), obtained (A) without and (B) with the introduction of cluster cores. The dashed lines in panel B represent the corresponding results of the Markov model constructed from the cored states. (C) Verification of the Chapman−Kolmogorov equation for the metastable states after coring. Compared are the metastabilities T_{ii}^{10} (τ = 0.5 ns; red), T_{ii}^{5} (1 ns; gray), and T_{ii} (5 ns; black).

exception of state 2 (see below), we find that the structure of the various microstates of each cluster are indeed quite similar.

Let us discuss a few important states of the system. State 1 is the native state of HP-35 which contains ∼30% of the population. An analysis of the transition matrix shows that it is well connected to numerous other metastable states. The native state has a well-defined structure and a hydrophobic core formed by three phenyl-alanine side chains, which add to the stability of the folded main chain. State 2 contains the starting configurations of all trajectories and holds ∼8% of the population. A closer study of this state and its vicinity reveals that—besides a low-energy minimum—it also contains a relatively large and flat area of the free energy surface. As the trajectories initially start out in this flat part,[48] they may explore various structures before they reach the local minimum of state 2. As a consequence, this state is structurally quite heterogeneous, see Figure 9. Regarding the boundary corrections, the flatness of the energy landscape in state 2 indicates that the population of the state is significantly reduced after coring, see Table 2. This is because an incoming trajectory only counts for state 2, once it has reached the low-energy minimum. Another interesting finding that is supported by recent experiments concerns the existence of kinetic traps in the folding process of HP-35.[36,37] In this data set, states 8 and 16 are highly metastable (but misfolded) states in which the system gets trapped (i.e., it remained in it until the trajectory was terminated). Finally, we note that state 12 exhibits an extended β-strand structure instead of helix 1.

To study the dynamical properties of the metastable states, we again consider their time-dependent population $P_i(t)$, see Figure 10. Similar to the case of Ala_7, we find (A) biexponential decays of $P_i(t)$ if no boundary corrections are employed and (B) single-exponential decays after coring. Coring was performed in the subspace of the first 13 principal components. Performing the Chapman−Kolmogorov test for the cored states, panel C shows that the first 20 metastable states exhibit Markovian behavior already at a time scale of 0.5 ns. Regarding the boundary corrections, it is interesting to note from Table 2 that the state population before and after coring are quite similar (with the exception of state 2).

5. CONCLUSIONS

Recently, various methods have been proposed that aim to identify metastable conformational states of an MD trajectory.[12−27] Although in principle the problem may be approached in a rigorous and straightforward manner via the transition operator formalism,[14] the large number of recent publications on the topic indicates that, in practice, the successful dynamical clustering of high-dimensional MD data often turns out to be a complex matter. It depends significantly on the nature and the quality of the data (e.g., on the intrinsic dimensionality, the underlying noise and the (typically insufficient) sampling), on various technical issues such as the choice of coordinates, microstates, and the lag time, and on the ultimate goal to be achieved (e.g., metastable states vs a Markov state model[25]). In this work, we have added a conceptionally and computationally simple new approach to identify metastable states, which relies on the suitable combination of various techniques.

- First, we achieve a preprocessing of the data by performing a dPCA of the trajectory. This greatly reduces the dimensionality of the data (typically from 10^4 to 10^1) without a loss of required information (a PCA becomes an exact representation of the data if a sufficient number of components are included). We stress that the use of internal coordinates such as dihedral angles is essential for describing large amplitude motion like folding,[44] i.e., a standard Cartesian PCA is not appropriate. The preprocessing greatly facilitates the subsequent generation of microstates via standard routines such as k-means. Moreover, it is essential for the coring procedure, which is hardly feasible in high dimensionality.
- To merge all microstates that belong to the same metastable state, we have introduced the most probable path algorithm. The method is similar in spirit to previous work but differs in important aspects (such as the treatment of microstates with equal probability). The algorithm depends on two parameters, the lag time τ used in the calculation of the transition matrix and the minimum metastability Q_{min} that defines the minimum barrier height that a metastable state is separated from its neighboring states. τ can be chosen according to the shortest relevant time scale or to facilitate a Markovian treatment of the dynamics. The choice of Q_{min} may reflect a separation of time scales of the dynamics and/or warrant the partitioning of the data into structurally homogeneous metastable states. In this way, the approach provides a natural way to decide how many metastable states there are. Moreover, the variation of Q_{min} facilitates the construction of dendrograms that

reveal the topology and the hierarchical structure of the free energy landscape.

- To correct for boundary errors of interstate transitions (Figure 2B), we have suggested a simple implementation to introduce cluster cores of each metastable state and required that a transition from a state to another state must reach the core region of the other state. For the systems considered, this conceptionally simple modification indeed greatly facilitates a Markovian description of the transitions. The proposed scheme extends previous works to the treatment of multidimensional data.

To demonstrate the applicability and the potential of the approach, we considered two well-studied model problems, hepta-alanine (Ala_7) and the villin headpiece (HP-35). For the small peptide Ala_7, the number of metastable states $n(Q_{min})$ exhibited a well-defined plateau at $n = 32$ (Figure 5), thus indicating a time scale separation of the conformational dynamics. The resulting 32-state model of Ala_7 was shown to nicely obey Markovian dynamics (Figure 7) if coring was applied. The analysis of the folding dynamics of the small protein HP-35, on the other hand, turned out to be a far more challenging problem. It involved 13 principal components, 8000 microstates, and 300 metastable states (compared to 5/1000/32 for Ala_7). As a consequence of the high dynamical dimensionality of the system, a two-dimensional representation of the free energy landscape as in Figure 1B could not resolve the underlying metastable states of the system. Moreover, the absence of a plateau in $n(Q_{min})$ in Figure 8 indicates that a clear separation of energy or time scales does not exist in the case of HP-35. Nonetheless, the above-described dynamical clustering approach allowed us to construct a network of 300 metastable states that are structurally homogeneous and exhibit Markovian behavior after coring (Figure 10). The detailed analysis of the resulting protein folding energy landscape and conformational network is beyond the scope of this paper and represents the topic of ongoing work.

■ ASSOCIATED CONTENT

Supporting Information

Size distribution of metastable states in terms of number of k-means microstates as obtained for hepta-alanine and villin headpiece. This material 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

This paper is dedicated to Wilfred van Gunsteren in honor of his 65th birthday. Wilfred's exceptionally creative and sharp mind has inspired the MD community for several decades. Furthermore, we thank Vijay Pande for providing the trajectories of HP-35 used in this study and Francesco Rao, Frank Noé, Laura Riccardi, and Rainer Hegger for instructive and helpful discussions.

■ REFERENCES

(1) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins* **1993**, *17*, 412−425.
(2) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9885−9890.
(3) Nguyen, P. H. *Proteins* **2006**, *65*, 898.
(4) Lange, O. F.; Grubmüller, H. *Proteins* **2006**, *70*, 1294−1312.
(5) Hegger, R.; Altis, A.; Nguyen, P. H.; Stock, G. *Phys. Rev. Lett.* **2007**, *98*, 028102.
(6) Onuchic, J. N.; Schulten, Z. L.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545−600.
(7) Dill, K. A.; Chan, H. S. *Nat. Struct. Biol.* **1997**, *4*, 10−19.
(8) Wales, D. J. *Energy Landscapes*; Cambridge University Press: Cambridge, U. K., 2003.
(9) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299−306.
(10) Noe, F.; Horenko, I.; Schütte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.
(11) Bowman, G. R.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 10890−10895.
(12) Riccardi, L.; Nguyen, P. H.; Stock, G. *J. Phys. Chem. B* **2009**, *113*, 16660−16668.
(13) Altis, A.; Otten, M.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.* **2008**, *128*, 245102.
(14) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J. Comp. Phys* **1999**, *151*, 146−168.
(15) Swope, W.; Pitera, J.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571−6581.
(16) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(17) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057−69.
(18) Krivov, S.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701−8714.
(19) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011−19016.
(20) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J. Chem. Phys.* **2009**, *131*, 124101.
(21) Keller, B.; Daura, X.; van Gunsteren, W. F. *J. Chem. Phys.* **2010**, *132*, 074110.
(22) Carmi, S.; Krapivsky, P. L.; ben Avraham, D. *Phys. Rev. E* **2008**, *78*, 066111.
(23) Rao, F. *J. Phys. Chem. Lett.* **2010**, *1*, 1580−1583.
(24) Rao, F.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 9152−9157.
(25) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noe, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(26) Keller, B.; Hünenberger, P.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2011**, *7*, 1032−1044.
(27) Schütte, C.; Noe, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134*, 204105.
(28) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740−4.
(29) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. *Nature (London)* **2002**, *420*, 102.
(30) Kubelka, J.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2003**, *329*, 625−30.
(31) Fernández, A.; Yi Shen, M.; Colubri, A.; Sosnick, T. R.; Berry, R. S.; Freed, K. F. *Biochemistry* **2003**, *42*, 664−71.
(32) Lei, H.; Wu, C.; Liu, H.; Duan, Y. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4925−30.
(33) Ensign, D. L.; Kasson, P. M.; Pande, V. S. *J. Mol. Biol.* **2007**, *374*, 806−16.
(34) Jain, A.; Hegger, R.; Stock, G. *J. Phys. Chem. Lett.* **2010**, *1*, 2769−2773.
(35) Rajan, A.; Freddolino, P. L.; Schulten, K. *PLoS One* **2010**, *5*, e9890.
(36) Reiner, A.; Henklein, P.; Kiefhaber, T. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 4955−4960.
(37) Beauchamp, K. A.; Ensign, D. L.; Das, R.; Pande, V. S. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 12734−12739.
(38) Hartigan, J. A.; Wong, M. A. *Appl. Stat.* **1979**, *28*, 100−108.
(39) Stillinger, F. H.; Weber, T. A. *Science* **1984**, *225*, 983−989.
(40) Wales, D. J.; Scheraga, H. A. *Science* **1999**, *285*, 1368−1372.

(41) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517.
(42) Cossio, P.; Laio, A.; Pietrucci, F. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10421−10425.
(43) Mu, Y.; Nguyen, P. H.; Stock, G. *Proteins* **2005**, *58*, 45.
(44) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. *J. Chem. Phys.* **2007**, *126*, 244111.
(45) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880−10889.
(46) The dihedral angles of both end groups were found to be virtually uncorrelated to the rest of the system.
(47) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. *J. Am. Chem. Soc.* **2007**, *129*, 1179−1189.
(48) Ensign et al.[33] generated 410 molecular dynamics (MD) trajectories at 373 K, starting from nine unfolded conformations, as well as 120 trajectories starting from the experimental crystal structure. They reported that two of the unfolded starting structures (4 and 7) folded much faster than the others. Hence, we concentrate our analysis on the trajectories starting from structure 4, which amount to about 35 μs in total.