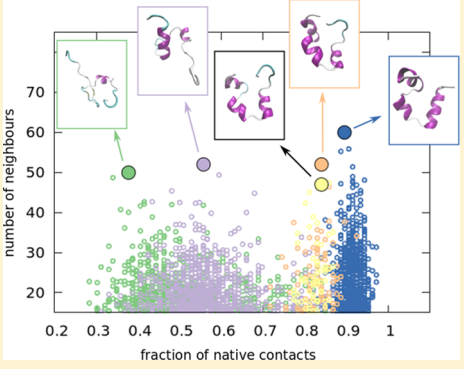
Explicit Characterization of the Free-Energy Landscape of a Protein in the Space of All Its C_α Carbons

Giulia Sormani,[†] Alex Rodriguez,[‡] and Alessandro Laio*[,†]

[†]SISSA, Via Bonomea 265, Trieste 34136, Italy
[‡]ICTP, Str. Costiera 11, Trieste 34151, Italy

ABSTRACT: By using an approach that allows computing the free energy in high-dimensional spaces together with a clustering technique capable of identifying kinetic attractors stabilized by conformational disorder, we analyze a molecular dynamics trajectory of the Villin headpiece from Lindorff-Larsen, K.; et al. How fast-folding proteins fold. *Science* **2011**, *334*, 517−520. We compute its free-energy landscape in the space of all its C_α carbons. This landscape has the shape of a 12-dimensional funnel with the free energy decreasing monotonically as a function of the native contacts. There are no significant folding barriers. The funnel can be partitioned in five regions, three mainly folded and two unfolded, which behave as Markov states. The slowest relaxation time among these states corresponds to the folding transition. The second slowest time is only twice smaller and corresponds to a transition within the unfolded state. This indicates that the unfolded part of the funnel has a nontrivial shape, which induces a sizable kinetic barrier between disordered states.



INTRODUCTION

Protein folding is possibly the most biologically relevant and studied conformational transition in biomolecules. The development of molecular dynamics simulations played an essential role in understanding this process. Indeed, protein folding is a rapid and complex process and thus very hard to study through experimental techniques, which either have a good spatial resolution or a good temporal resolution, but rarely both. The first simulation of a folding event in an explicit solvent was obtained using the idle processing time of thousands of personal computers.[2] More recently, a computer was specifically designed to perform molecular dynamics of biomolecules on the millisecond timescale.[3] By using this computer, it was possible simulating the folding process of 12 small proteins.[1] Other works from the same group followed, increasing the availability of all-atom folding trajectories.[4−7] All this wealth of data can nowadays be used to derive more and more reliable models of the folding process.

The possibly most famous and inspiring paradigm in the field is the folding funnel.[8,9] According to this theory, evolution has shaped the energy landscape of proteins[10,11] in such a way that it resembles a funnel: the native state is at its bottom, whereas the non-native local minima are small ripples on the walls of the funnel. Going down along the funnel, the number of possible states must decrease and the number of native contacts must increase. This landscape allows the polypeptide chain finding the folded structure through a large number of pathways, "solving" Anfinsen's paradox.[12] Experimental data on fast-folding proteins support the funnel paradigm pointing to the so-called downhill folding scenario,[13] in which the entropic cost for forming native contacts is much smaller than the energy gain, to the point that the barrier created by the configuration entropy disappears.[10]

The most direct procedure to visualize the folding landscape is projecting it on a one-dimensional reaction coordinate, for example, the fraction of native contacts.[1,4] However, the projection on a single variable can bring to a description that is thermodynamically meaningful but which does not capture the complexity of the kinetics. In particular, the folding barrier obtained by this procedure unavoidably depends on the variable which is chosen for the projection. A more rigorous procedure for describing kinetics is offered by Markov state modeling (MSM).[14] The core idea of this method is to describe the dynamics as a Markov process between a few metastable states. The most common procedure to obtain these states involves first grouping the conformations in a high number of microstates (e.g., using *k*-means clustering[15] or the Ward algorithm[16]). The microstates are then grouped in Markov states, for example, using most probable path (MPP) algorithm[17] or Perron cluster cluster analysis (PCCA[18]/PCCA +).[19] The picture emerging from these studies is extremely rich. For example, in refs 20 and 21, the native states act as kinetic hubs. Indeed, according to these analyses, the native state is easily reached from every unfolded state, whereas there is a low flux between the unfolded states. Moreover, a

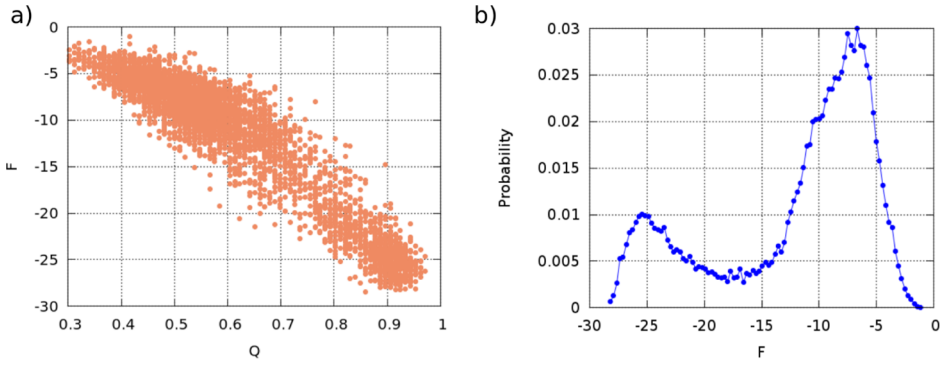


Figure 1. (a) Free energy (F) vs fraction of native contacts (Q) for each trajectory frame. Q is evaluated comparing the contact matrix of a structure with the contact matrix of the crystallographic structure (PDB_id 2F4K); two heavy atoms form a contact if their distance is less than 4.5 Å. (b) Probability distribution of the free energy. At an intermediate free energy ($F \approx -18$), corresponding to $Q \approx 0.75$, there are fewer states than at high and low free energies.

systematic analysis performed by MSM on 14 protein simulations (from refs 1, 3, 22) revealed significant deviations from the two-state behavior.[23] More recently, the folding landscape of the double mutant of the Villin headpiece was studied by an approach combining clustering in high dimensions and an advanced dimensional reduction technique based on the analysis of kinetic properties.[17,24]

Here, we characterize explicitly the free-energy landscape of a double mutant of the Villin headpiece[1,13] as a function of the coordinates of all the C_α carbons, demonstrating that it is indeed funnel-shaped, namely, that the free energy of each configuration is a monotonic function of the number of native contacts. Moreover, by exploiting a clustering technique capable of finding also kinetic attractors stabilized by conformational disorder, we find that the funnel can be partitioned in five subregions, which behave as Markov states. The core of our approach is a technique that allows estimating the free energy in spaces of high dimensionality,[25] in particular, the space defined by the positions of all the C_α carbons. The effective dimension of this space, estimated by the approach in ref 26, is approximately 12 for the Villin trajectory. We will show that the manifold in which the data are embedded is curved and topologically complex, which implies that it is not possible to obtain an explicit expression of these 12 coordinates. However, by using our approach, one can compute the free energy as an implicit function of these coordinates. Our results are fully consistent with the funnel theory but interpreting the free energy (calculated as a function of the position of all C_α-s) as an efficacious conformational energy. Indeed, the free energy of each configuration decreases monotonically with the fraction of native contacts. The depth of the native minimum is $\simeq 15 k_B T$, and all the barriers on the funnel are of a few $k_B T$. We then analyze the folding kinetics on the funnel, by using an extension of Density Peak clustering developed specifically for this work, capable to locate within the same framework the enthalpic and entropic traps. In concrete, our algorithm is able to detect directly the Markov states without defining any collective variable and working directly on the coordinates of the C_α carbons, without defining the microstates, and without optimizing the properties of the states using information on the dynamics. We find five relevant states, neatly mapping different regions of the funnel: three with a high fraction of native contacts and two unfolded. Our model predicts four relevant relaxation times. The slowest is associated with the folding−unfolding transition, and the second one, only 2 times smaller, is associated with an internal relaxation in the unfolded state.

RESULTS

We performed the analysis using two different sets of coordinates and metrics. The two metrics are the Euclidean distance of the backbone Ψ dihedral angles (imposing periodic boundary conditions with a box size of 2π) and the root-mean-square deviation (rmsd) of the backbone atoms. As we get analogous results using the two different metrics, we choose to present a detailed description of the results using the Euclidean distance between the Ψ dihedral angles in the following sections of the paper, summarizing in the Supporting Information the ones obtained using the rmsd.

We use 122 μs of the trajectory of a double mutant of the Villin headpiece from ref 1 (shortly "Villin"). We take one frame every 4 ns. We measure the Ψ dihedral angles, discarding the first and the last as their fluctuations are not related to the structure of the protein. The simulation is thus summarized by a set of 30 500 points in a space of 32 coordinates.

By using the approach in ref 26, we estimate an intrinsic dimension (ID) in this space of approximately 12, indicating that, on average, from every configuration the system can move in 12 linearly independent directions. A similar value is obtained by using the rmsd metric (see Supporting Information).

Free-Energy Landscape. For each point of the dataset (i.e., frame of the trajectory), we evaluated the free energy and its uncertainty using the PAk estimator, described in ref 25. This method allows estimating the free energy in a space where the ID is lower than the number of coordinates (ID = 12 in our case), without the need of specifying explicitly the collective variables that define this reduced space. In this approach, the free energy is evaluated using an extension of the k-nearest neighbor density estimator[27] where the optimal k, which becomes position-dependent and is denoted as $\hat{k}$, is chosen by finding the largest neighborhood in which the free energy can be considered constant within a confidence threshold. We observe a strong anticorrelation between the free energy (F) and the fraction of native contacts (Q): the folded state is the free-energy minimum (see Figure 1, panel (a)). The free energy is a monotonic function of Q. The free-energy landscape is therefore a funnel in 12 dimensions, with the global minimum corresponding to the crystallographic structure and a wide area corresponding to the unfolded

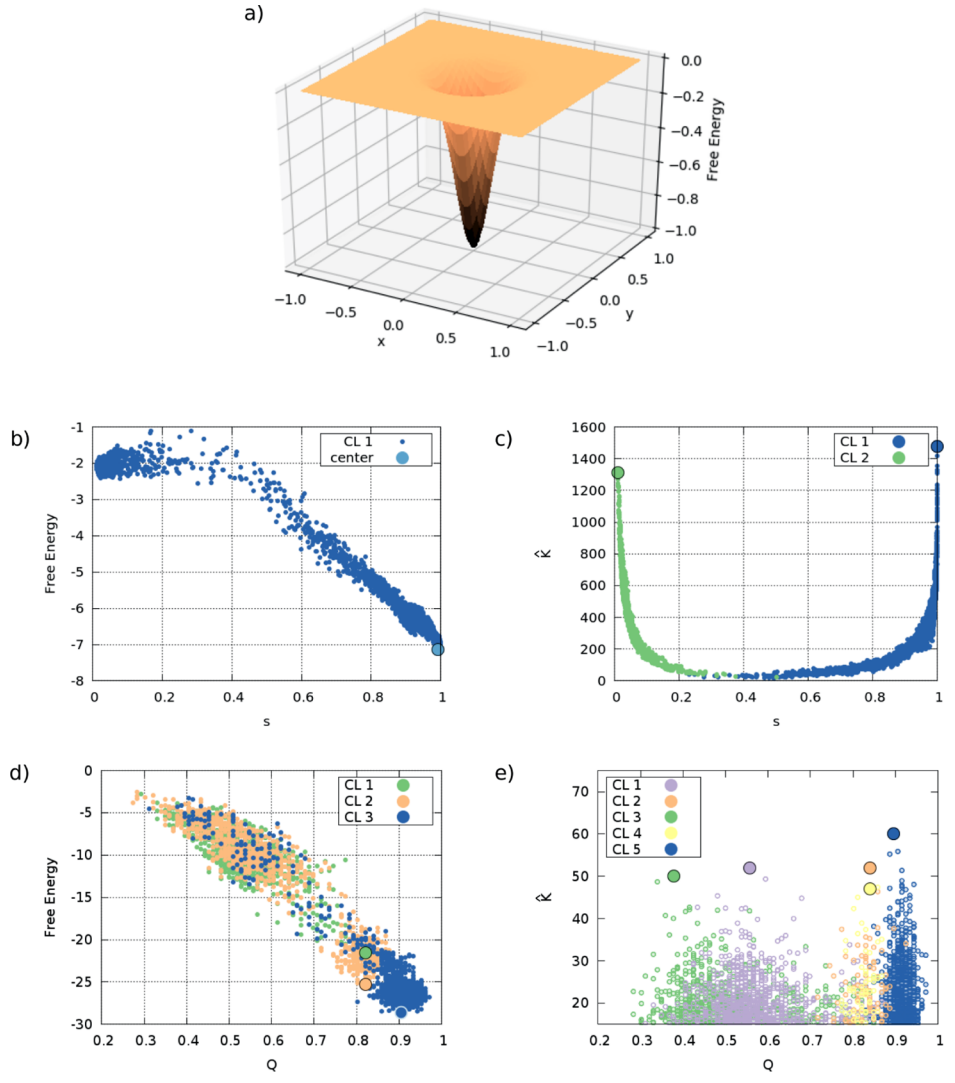


Figure 2. Comparison between DP clustering and $\hat{k}$-peak clustering. (a) Funnel-shaped two-dimensional free-energy distribution. (b,c) Results of the comparison for the toy model in panel (a): (b) free energy of each point vs an order parameter $s = \frac{1-(d/0.3)^3}{1-(d/0.3)^6}$, where d is the distance from the origin; the cluster analysis, performed with DP clustering,[28] finds only one cluster; (c) optimal value of the nearest neighbors of each point ($\hat{k}$) vs s. The cluster analysis, performed with $\hat{k}$-peak clustering, finds two clusters. (d,e) Results of the comparison for the Villin trajectory: (d) fraction of native contacts Q vs the free energy F for the frames belonging to the three most populated clusters found with DP clustering; (e) fraction of native contacts Q vs the optimal number of neighbors ($\hat{k}$) for the five most populated clusters found with $\hat{k}$-peak clustering. In panels (b–e), the cluster centers are shown as points with bigger radius.

region. Moreover, we see in panel (b) that there are few states with a free energy of $\simeq -17$ (corresponding to the transition region): the funnel has a bottleneck with a lower number of available states in the "intermediate" region.

Kinetic Attractors on the Funnel. We then attempted analyzing the free-energy landscape using the approach described in ref 28. This method is an extension of Density Peak clustering,[29] in which each free-energy minimum corresponds to a cluster, and the connections among clusters are obtained measuring the height of the free-energy barriers between the minima.

This procedure, the results of which are presented in detail in the Supporting Information, identifies the folded state, but it is hindered by some serious pitfalls. The key problem is that there are no free-energy minima corresponding to the unfolded state: the position of the cluster centers is shown with blue dots in the F versus Q plane in Figure S1 (panel a). There are no centers with $Q < 0.6$. Indeed, the unfolded state is composed by configurations that are significantly different from each other, with very little or no secondary structure. Clearly, an algorithm attempting to find free-energy minima in the space of the C_α positions is not an appropriate tool for studying such a system.

$\hat{k}$-Peak Clustering. In order to address this problem, we developed a procedure which allows performing clustering on systems in which some of the metastable states are stabilized by conformational disorder. To find these states, we consider, for each frame i, the number of neighbors $\hat{k}_i$ for which the free energy can be considered constant within a given level of confidence. The idea is that there are two situations in which $\hat{k}_i$ assumes high values. The first one is in the free-energy minima, where the high density of points leads to a high value of $\hat{k}_i$. The second one is in the flat regions of the free-energy landscape, where the low variation of the density of points leads also to

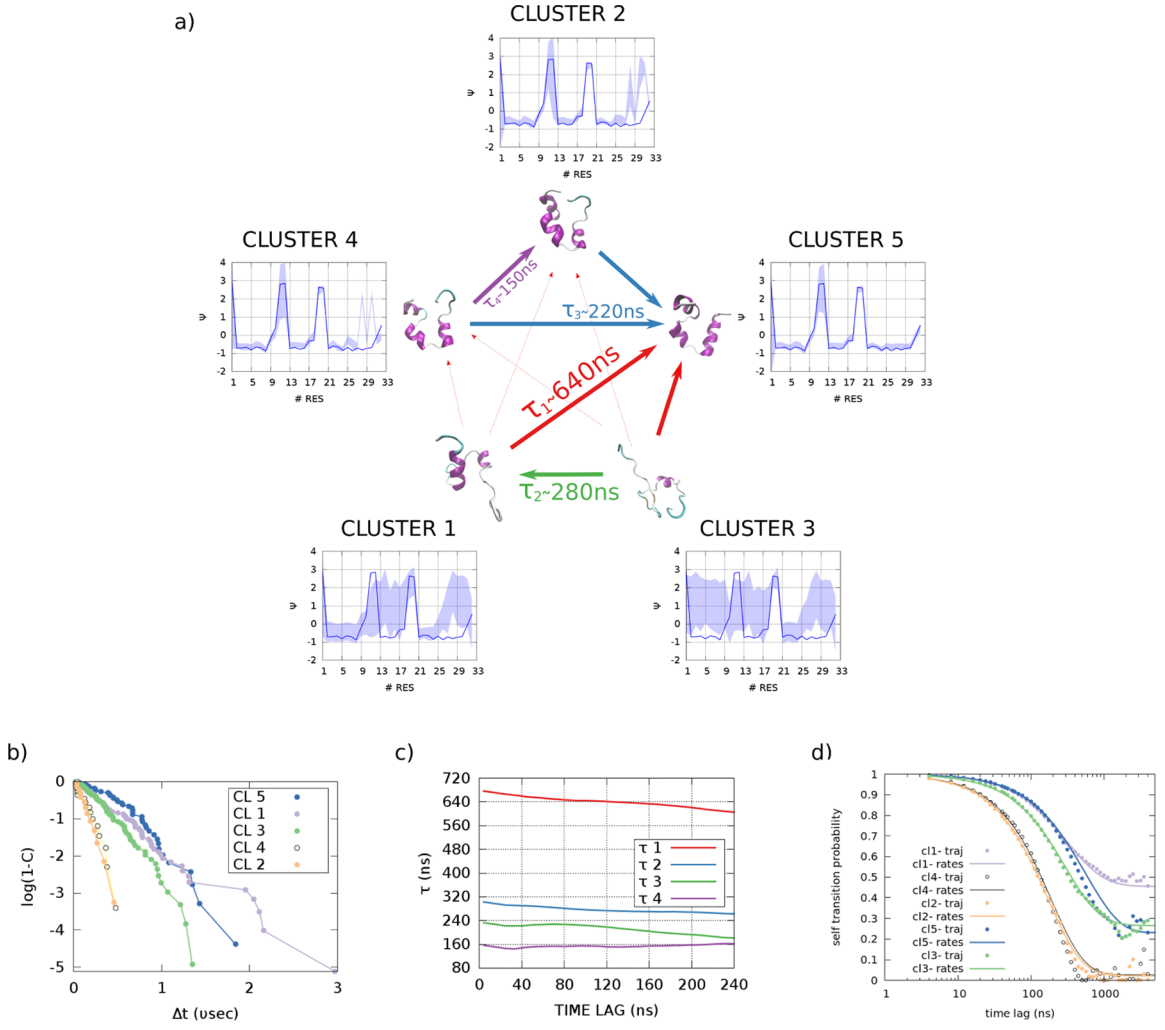


Figure 3. (a) Diagrams representing the dihedral angle values and their variance for the core set structures of each cluster. Next to the diagrams, the structures of the centers of the clusters are shown. The arrows link the clusters involved in the relevant transitions; for each transition, a color code is assigned, and the relaxation times (τ_1, τ_2, τ_3, τ_4) are written over the arrows. (b) Logarithm of the negative cumulative distribution [i.e., log(1 − cumulative)] of the permanence times (Δt) in each of the five clusters. (c) Relaxation times obtained from the transition matrix, as a function of the time lag. (d) Self-transition probabilities as a function of the time lag for the five clusters. Dots represent probabilities obtained directly from the trajectory and lines represent probabilities obtained from the rescaling of the matrix $\Pi(dt = 120$ ns).

high values of $\hat{k}_i$. Therefore, we propose that in order to characterize the kinetics of a system in which at least one state is stabilized by conformational disorder, it is convenient to look for the peaks of $\hat{k}_i$. The approach for finding the clusters is identical to the one described in ref 28, with the optimal number of neighbors $\hat{k}$ playing the role of the free energy in the original implementation. The centers of the clusters therefore are the local maxima of $\hat{k}$. The algorithm is as follows.

- Estimation of $\hat{k}_i$ and of $r_{\hat{k}_i}$ for each frame, using the approach in ref 25. $r_{\hat{k}_i}$ is the radius of the neighborhood in which the density is approximately constant.
- Estimation of the uncertainty σ_i of $\hat{k}_i$ as the standard deviation of $\hat{k}$ among the points which are inside the constant density neighborhood of point i.
- Search of the peaks of $g_i = \hat{k}_i - \sigma_i$. The local maximum of g_i (defined putative center) is a cluster center if two conditions are satisfied: (1) $\delta_i > r_{\hat{k}_i}$, where δ_i is the distance from the nearest point with higher g and (2) the point i does not belong to the constant density neighborhood of any other point with higher g.
- Assignation of all the points that are not centers to the same cluster as the nearest point with higher g_i. This assignation is performed in the order of decreasing g_i.
- Search of the saddle points, which are the points with highest g among the border points. A point i belonging to cluster A is at the border between cluster A and B if (1) its distance to the closest point j belonging to B is less than $r_{\hat{k}_i}$ and (2) i is the closest point to j belonging to A.
- Merging of the clusters which are not meaningful. In particular, cluster A is merged with cluster B if: $\hat{k}_{AB} - \hat{k}_A < Z(\sigma_{\hat{k}_A} + \sigma_{\hat{k}_{AB}})$, where $\hat{k}_{AB}$ is the optimal number of neighbors of the saddle point between cluster A and B, $\hat{k}_A$ is the optimal number of neighbors of the center of cluster A, and $\sigma_{\hat{k}_{AB}}$ and $\sigma_{\hat{k}_A}$ are the corresponding uncertainties. Z is a free parameter of our approach.

We fixed the value of the merging parameter to $Z = 0.2$. If Z is increased, the description becomes less detailed; if Z is increased, it becomes more detailed. We verified that the description does not change significantly if Z is lowered to 0.1

or increased to 0.3, indicating that our results are robust with respect to the choice of this parameter. Even if Z is set to zero, the most populated clusters are the same as the ones in Z = 0.2, but there are several additional clusters with very small populations, or which are explored only once during the dynamics, and are therefore likely to be numerical artifacts. We choose to present the results, fixing Z = 0.2, as this value allows a detailed description of the system, maintaining a high level of statistical significance: the relevant states are visited a significant number of times (>14), and each state has a significant population.

In order to test the new algorithm, we devise a toy model in which the free-energy distribution has a funnel shape (Figure 2, panel a). We generated ≃15 000 points in the x, y plane ($x, y \subset [-1, 1]$), from the density distribution given by the sum of a narrow Gaussian centered at the origin and a uniform distribution. We then applied both the density peak algorithm[28] and the $\hat{k}$-peak algorithm, giving as input the coordinates. The results are compared in Figure 2. In panel (b), we show, for each point, the dependence of its free energy on an order parameter s, defined as a function of the distance from the origin which is close to 1, if the distance is small, and close to zero, if the distance is large. Clearly, the free energy has a single minimum; thus, the density peak algorithm finds a single cluster. On the other hand, the optimal number of neighbors ($\hat{k}$), as a function of s (panel d), has two peaks, one at the center of the Gaussian ($s \approx 1$) and the other at the maximum distance from this center ($s \approx 0$). Two clusters are thus found by the $\hat{k}$-peak algorithm: the point assignations to these two clusters are shown with two different colors in panel (d). These results show the ability of the algorithm to localize within the same framework a metastable state stabilized by the (free) energy and a metastable state stabilized by conformational disorder, namely a flat area of the (free) energy landscape.

Encouraged by the results obtained with the toy model, we applied the $\hat{k}$-peak algorithm to study the Villin trajectory. In Figure 2 (panel e), the points of the five biggest clusters are shown in the $\hat{k}$ versus Q plane. These clusters alone contain 93% of all the trajectory frames. In this representation, we see the presence of several peaks of $\hat{k}$ as a function of Q, both for low and high values of Q. The crystallographic state is easily identified in cluster 5, which contains the frames with the highest Q. There are other two peaks with a high value of Q, corresponding to clusters 2 and 4. These two clusters specifically select a region with $0.75 < Q < 0.85$. The unfolded region is mainly represented by clusters 1 and 3. These two clusters almost do not contain any structure with $Q > 0.75$. There is a significant overlap in the Q value between the two unfolded clusters, but this is not surprising: Q is a good reaction coordinate for describing the folding process, not the dynamics within the unfolded state. Also, the value of Q of clusters 2 and 4 overlaps with the value in cluster 5: indeed, as we will see, these two clusters correspond to the defective folded states, with only a few non-native contacts.

In panel (a) of Figure 3 we present the average values of the dihedral angles (Ψ) and their variance for the core set structures of each cluster. The structure i is assumed to belong to the cluster core if its value of $\hat{k}$ is sufficiently high ($\hat{k}_i \geq 25$) and if the following or the previous configuration satisfying the first condition belongs to the same cluster. The first condition selects the frames which are within the lower part of the basin defining the cluster. The second condition discards isolated configurations classified as core states. Once the core set of the clusters are determined, the remaining frames are assigned to the cluster of the previous visited core state. At the end of this procedure, we discard all the clusters that have less than five visits. This is done as a minimum number of visits is necessary to have a sufficient statistics in order to describe the dynamics.

The blue thick line in the figure represents the value of the dihedral angles for the crystallographic structure. Looking at the crystallographic dihedral angles, we see the presence of helices when their value is $\Psi \approx -0.8$. This happens in three different regions: from residue 2 to 8, from 13 to 16, and from 21 to 30. The presence of folded clusters (5, 4, 2) and of unfolded clusters (1, 3), already seen in Figure 2 (panel e), is confirmed in this representation. The folded region is characterized by structures very similar to each other as the dihedral variance is small. Cluster 5 corresponds to the crystallographic Villin, with the formation of three helices. The other two folded clusters (2 and 4) are characterized by structures that are almost totally folded, except for the final part of the C-terminal helix. This kind of structure has already been seen as a possible intermediate state between the folded and the unfolded ones, both experimentally,[30] in a computer simulation of triplet-triplet energy transfer experiments,[4] and in an MSM built on the same trajectory.[23] The unfolded clusters are characterized by a high value of the dihedral variance, but their core sets contain structures which are different from each other. Cluster 1 is characterized by structures in which the N-terminal helix (1 < res < 8) and the first part of the C-terminal helix (21 < res < 24) are formed, whereas the rest of the protein is basically unfolded. Cluster 3 mainly contains totally unfolded frames. As we will see, this distinction has an impact on the relaxation kinetics within the unfolded state.

Kinetics. Using the $\hat{k}$-peak clustering algorithm, we have partitioned the entire conformation space into five clusters which, as we will see, allow describing satisfactorily also the dynamics. In panel (b) of Figure 3, we show the negative cumulative distribution of the permanence times (Δt) in each of the five clusters in a semilogarithmic scale. These curves are well fitted by straight lines. This means that the probability distribution $P(\Delta t)$ is approximately exponential, and the process of moving from one cluster to another is a Poisson process.

We then built an MSM directly on the five states. The kinetics is assumed to be a memoryless jump process between the five clusters, and it is summarized using a 5 × 5 transition probability matrix (Π). A matrix element Π_{ij} represents the conditional probability that the system is in state j at time t + dt (dt is the lag time), given that it was in state i at time t. The matrix Π is thus dependent on the parameter dt. However, if the conformation space partition has been done in a correct way, there should be an interval of time lag dt, in which all the dynamics predicted by Π are invariant. Analyzing the spectrum of Π, we get the relaxation times of the system (from the eigenvalues) and the connections between the clusters (from the eigenvectors). In panel (c) of Figure 3, we show that there is a wide range of dt for which the relaxation times τ_i are almost constant. This proves that our model is approximately Markovian. Specifically, there are four relaxation times related to transitions between different clusters. None of them is low enough compared to the others:

- $\tau_1 \approx 640$ ns. This value represents the main relaxation time of the system. Indeed, the corresponding transition

is the general folding/unfolding transition, between clusters (1, 3) and clusters (5, 4, 2). The eigenvector is shown in Figure S4 of the Supporting Information.

- $\tau_2 \approx 280$ ns. The second largest relaxation time is internal in the unfolded state, between cluster 3 (containing totally unfolded structures) and cluster 1 (containing unfolded structures but with the N-terminal helix formed and C-terminal helix partially formed).
- $\tau_3 \approx 220$ ns. The corresponding transition is another internal transition in the folded state, from cluster 5 (crystallographic state) and clusters 2 and 4 (containing folded structures but with the C-terminal helix partially unformed).
- $\tau_4 \approx 150$ ns. The corresponding transition is between clusters 2 and 4.

In panel (a), the arrows represent the transitions. The relaxation times are indicated above the arrows. The longest relaxation time we found ($\tau_1 = 640$ ns) is of the same order of magnitude as that of the one from ref 23 (three-state MSM on the same trajectory, $\tau = 400$ ns).

In order to evaluate the folding time and compare it with the analysis from ref 1, we gathered our five clusters into two states: the folded one (clusters 2, 4, and 5) and the unfolded one (clusters 1 and 3). We then evaluated the folding time (t_f) as the average time spent in the unfolded state and the unfolding time (t_u) as the average time spent in the folded state. We obtained $t_f = 2.26$ μs and $t_u = 0.915$ μs, in good agreement with the ones obtained in ref 1. This folding time is however bigger than the experimental one, estimated to be $\sim$1 μs.[13,31]

We then performed an extra Markovianity test on the five-state model. We compared the transition probabilities between states (Π_{ij}) as a function of the time lag (dt), evaluated in two different ways:

1 Directly counting the number of transitions from the trajectory (method 1)
2 Scaling the transition matrix evaluated at a fixed time lag ($\Pi(dt = 120\ ns) = \Pi(120)$, method 2)

Indeed, if the model is Markovian, the Chapman–Kolmogorov equation should hold: $\Pi(dt) = (\Pi(1))^{dt} = \Pi(120)^{dt/120}$. The rescale of $\Pi(120)$ is performed from its eigenvalues (λ) and left and right eigenvectors (Ψ^{left}, Ψ^{right}): $\Pi_{ij}(t) = \sum_\alpha \lambda_\alpha^{t/120}\ \Psi_i^{left}\ \Psi_j^{right}$. We choose to scale the matrix $\Pi(120)$, as in this range the relaxation times of the system are independent of the time lag. In panel (d) of Figure 3, we compare the self-transition probabilities for the five clusters, evaluated from method 1 (shown with dots) and from method 2 (shown with lines). For low time lags, there is a perfect correspondence for all clusters, and the correspondence holds until $\simeq$200 ns in the worst case (cluster 4) and until $\simeq$300/400 ns for the other clusters. This shows that the Markovianity is respected for a wide range of time lags. The same test has been performed on the three-state MSM from ref 23, where a similar high-quality agreement is observed only at long timescales (>100 ns).

In Figure S3 from the Supporting Information, we instead compare the self-transition probabilities among different clusters, obtained with methods (1) and (2). The correspondence is good for a low lag time, but it is however lost at a large time lag. This is due to the fact that there are fewer transitions among different clusters than self-transitions; the statistics is poorer, and so the counts from the trajectory can deviate from the theoretical curve. Indeed, we see that the best correspondence is for transition 1 → 3, which has good statistics as it is among clusters that are highly populated and similar to each other (both unfolded). In summary, this test is a strong proof of Markovianity and of the precision of our five-state model.

We finally evaluated the heights of the free-energy barriers between each couple of clusters. The free energy-barrier between cluster A and cluster B is given by $\Delta F_{A-B} = F_{AB} - F_A$, where F_A is the free energy of the center of the cluster A and F_{AB} is the free energy of the saddle point between cluster A and cluster B. In agreement with the experimental picture of a downhill free-energy landscape,[31] all the barriers from unfolded to folded clusters are really low (around 1 KT). This is not surprising: indeed, the folding process in this system is not a rare event because of the presence of a barrier but rather because of the structure of the free-energy landscape, which resembles the one of the toy model in Figure 2. In this model, there are no barriers, and yet finding the (only) free-energy minimum is a rare event, as it requires diffusing through a large region where the free energy is approximately flat. On the other hand, the barriers between the folded clusters and unfolded ones are large: the highest unfolding barrier, corresponding to the depth of the funnel, is of $\sim$15 KT, between cluster 3 and cluster 5. Finally, the barriers between the folded state and the two defective folded states (clusters 4 and 5) are of the order of 4 KT.

DISCUSSION

In this work, we described a procedure which gives a detailed description of the free-energy folding landscape and of the kinetics on this landscape, avoiding the definition of any collective variable and the use of information from the dynamics for deriving the model. Our procedure consists of two main steps. The first one is the free-energy calculation for each frame of the trajectory using the method described in ref 25; the second one is the analysis of the free-energy landscape using the $\hat{k}$-peak algorithm, described in this work. The salient feature of our technique is the capability of identifying both the flat regions of the free-energy landscape, corresponding to the unfolded states, and the minima of the free energy corresponding to the native or near-native states. The main difference with the other procedures for building an MSM is that the relevant kinetic states are here identified simply by analyzing the structure of the free-energy landscape, *without using kinetic information to optimize the partition or for choosing the number of states.*

Applying our algorithm to the MD simulation of Villin from Anton, we observe that the free-energy landscape as a function of the 32 dihedral coordinates of the protein is actually funnel-shaped. This sheds a new light on the works from Wolynes and Onuchic:[8,9] our method allows an explicit calculation of the efficacious energy function which describes the folding process. Like in ref 9, this function is defined as a function of the coordinates of the C_α carbons. We find that, as predicted in these works, the free energy is a monotonic function of the fraction of native contacts Q. However, at variance with what was observed in the model in ref 9, the number of states is not a monotonic function of Q: indeed, the scatter plot of the value of Q versus the value of the free energy F (Figure 1) indicates that a bottleneck is present at the intermediate values of Q and F ($F \approx -20$ and $Q \approx 0.75$ in the figure). Moreover, our work allows characterizing explicitly the shape of the funnel: even if

the coordinate space is 32-dimensional, the presence of correlations makes the manifold on which the funnel is defined 12-dimensional. In order to investigate the structure of this manifold, we performed an analysis of the trajectory using ISOMAP, an approach which allows recovering an explicit representation of a manifold when it is topologically equivalent to a hyperplane. We find that the spectrum of the ISOMAP covariance matrix does not show any visible gap (see Supporting Information). Therefore, a meaningful dimensional reduction cannot be performed on this system using this technique. This suggests that the manifold on which the data are lying is not isomorphic to a hyperplane.

Studying the kinetics, we obtain five main states. Three of these states are folded: one of them corresponds to the native state and two of them to near-native states in which the C-terminal helix is partially unraveled. The remaining two states are unfolded: one contains totally unfolded conformations and the other contains unfolded conformations but with the tendency of having parts of the N-terminal and C-terminal helices folded. The permanence times in these two states are long, meaning that these two states are separated by a well-defined kinetic barrier. In the trajectory we analyzed, we observe 117 direct transitions between the two states, without visiting the native state in between. This implies that the description that we present is not consistent with the kinetic hub scenario.[20] To the best of our knowledge, the existence of two well-defined kinetic attractors in the unfolded state of Villin has never been reported before.

The predicted folding time is similar to the one obtained in ref 1, which is however longer than the experimental one (from refs 13 and 31). The folding barriers between the unfolded states and folded ones are really low (<KT), in agreement with the experimental results of a downhill folding landscape.

The Markovianity of our model is assessed by various tests (panel b–d of Figure 3 and Figure S3 from the Supporting Information). We remark that in our approach Markovianity is not imposed iteratively but is only verified a posteriori.

In order to evaluate the reliability of our results, we applied our procedure to a second trajectory of the same protein, obtained with a different force field (from ref 4). The whole analysis is presented in the Supporting Information In this second simulation, performed at the same temperature, the relative time spent by the system in the folded state is ≃70%, much more than that in the simulation performed with the other force field. Despite this difference, there is an important consistency between the two analyses: in both cases, the main relaxation time corresponds to the folding–unfolding transition and the second one corresponds to a transition internal in the unfolded state. The presence of two kinetic attractors in the unfolded state is observed with two different force fields. The same trajectory was analyzed by Sittel and Stock.[24] They first performed a dimensional reduction with principal component analysis, followed by a density-based clustering and a final step of dynamic clustering using MPP. After this procedure, they obtained 12 metastable states, described according to the secondary structure propensity of each residue. Some of the states they found are similar to ours. A precise comparison on the description of the kinetics is not possible, as the relevant relaxation times of their model and the states involved in the main transitions are not indicated.

In conclusion, thanks to the high quality of the description and to the simplicity of the method, we believe our algorithm will become a popular tool for the study of the structure of (free)-energy landscapes, in particular when these landscapes include metastable states stabilized by conformational disorder.

■ ASSOCIATED CONTENT

● Supporting Information

The Supporting Information is available free of charge a https://pubs.acs.org/doi/10.1021/acs.jctc.9b00800.

Description of the clusters obtained with density peak clustering; results obtained using the $X-Y-Z$ positions of the atoms as coordinates and rmsd as distance; transition probabilities between different clusters as a function of time lag; spectrum of the transition matrix Π; Analysis of an MD trajectory generated with the Amber ff99SD*-ILDN force field (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: laio@sissa.it.

ORCID

Alessandro Laio: 0000-0001-9164-7907

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517–520.

(2) 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.

(3) Shaw, D. E.; et al. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341–346.

(4) 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.

(5) Kruse, A. C.; et al. Structure and dynamics of the M3 muscarinic acetylcholine receptor. *Nature* **2012**, *482*, 552.

(6) 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.

(7) Nygaard, R.; et al. The dynamic process of β2-adrenergic receptor activation. *Cell* **2013**, *152*, 532–542.

(8) Onuchic, J. N.; Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.

(9) Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein folding funnels: a kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 8721–8725.

(10) Wolynes, P. G. Recent successes of the energy landscape theory of protein folding and function. *Q. Rev. Biophys.* **2005**, *38*, 405–410.

(11) Vendruscolo, M.; Dobson, C. M. Towards complete descriptions of the free–energy landscapes of proteins. *Philos. Trans. R. Soc., A* **2004**, *363*, 433–452.

(12) Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **1973**, *181*, 223–230.

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

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

(15) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A k-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1979**, *28*, 100–108.

(16) Ward, J. H., Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

(17) Jain, A.; Stock, G. Identifying metastable states of folding proteins. *J. Chem. Theory Comput.* **2012**, *8*, 3810−3819.
(18) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.* **1999**, *151*, 146−168.
(19) Deuflhard, P.; Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra Appl.* **2005**, *398*, 161−184.
(20) Bowman, G. R.; Pande, V. S. Protein folded states are kinetic hubs. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 10890−10895.
(21) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. Markov state model reveals folding and functional dynamics in ultra-long MD trajectories. *J. Am. Chem. Soc.* **2011**, *133*, 18413−18419.
(22) Beauchamp, K. A.; Ensign, D. L.; Das, R.; Pande, V. S. Quantitative comparison of villin headpiece subdomain simulations and triplet−triplet energy transfer experiments. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 12734−12739.
(23) 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.
(24) Sittel, F.; Stock, G. Robust density-based clustering to identify metastable conformational states of proteins. *J. Chem. Theory Comput.* **2016**, *12*, 2426−2435.
(25) Rodriguez, A.; d'Errico, M.; Facco, E.; Laio, A. Computing the Free Energy without Collective Variables. *J. Chem. Theory Comput.* **2018**, *14*, 1206−1215.
(26) Facco, E.; dErrico, M.; Rodriguez, A.; Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci. Rep.* **2017**, *7*, 12140.
(27) Mack, Y. P.; Rosenblatt, M. Multivariate k-nearest neighbor density estimates. *J. Multivar. Anal.* **1979**, *9*, 1−15.
(28) d'Errico, M.; Facco, E.; Laio, A.; Rodriguez, A. Automatic topography of high-dimensional data sets by non-parametric Density Peak clustering. arXiv preprint arXiv:1802.10549, **2018**,
(29) Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492−1496.
(30) 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.
(31) Beauchamp, K. A.; Ensign, D. L.; Das, R.; Pande, V. S. Quantitative comparison of villin headpiece subdomain simulations and triplet−triplet energy transfer experiments. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 12734−12739.