THE JOURNAL OF PHYSICAL CHEMISTRY B

ARTICLE

pubs.acs.org/JPCB

Delineation of Folding Pathways of a β-Sheet Miniprotein

Wenwei Zheng,[†] Bo Qi,[‡] Mary A. Rohrdanz,[†] Amedeo Caflisch,[¶] Aaron R. Dinner,[‡] and Cecilia Clementi*[,†]

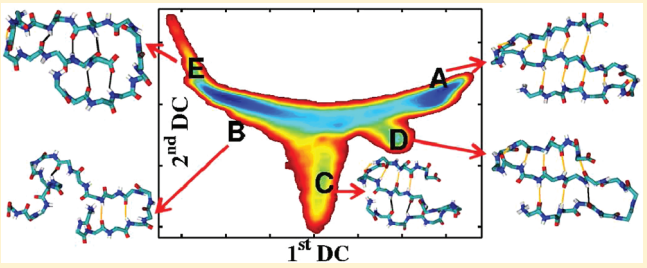
[†]Department of Chemistry, Rice University, Houston, Texas 77005, United States

[‡]Department of Chemistry, University of Chicago, Chicago, Illinois 60637, United States

[¶]Department of Biochemistry, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

Supporting Information

ABSTRACT: Several methods have been developed in the past few years for the analysis of molecular dynamics simulations of biological (macro)molecules whose complexity is difficult to capture by simple projections of the free-energy surface onto one or two geometric variables. The locally scaled diffusion map (LSDMap) method is a nonlinear dimensionality reduction technique for describing the dynamics of complex systems in terms of a few collective coordinates. Here, we compare LSDMap to two previously developed approaches for the characterization of the configurational landscape associated with the folding dynamics of a three-stranded antiparallel β-sheet peptide, termed Beta3s. The analysis is aided by an improved procedure for extracting pathways from the equilibrium transition network, which enables calculation of pathway-specific cut-based free energy profiles. We find that the results from LSDMap are consistent with analysis based on transition networks and allow a coherent interpretation of metastable states and folding pathways in terms of different time scales of transitions between minima on the free energy projections.



1. INTRODUCTION

Molecular dynamics (MD) simulations are routinely used to collect information on the motion of high dimensional systems, in many different areas of research. Proteins serve as paradigms for analyzing complex dynamics because of the many degrees of freedom involved and the important role of entropy, for example, in the protein folding mechanism and free energy surface.[1] The ability to interpret the results of peptide or protein simulations often requires the definition of collective coordinates with which to describe the dynamics and understand the nature of transition ensembles. Several research groups have proposed different methods to accomplish this task. For example, recent approaches to analyze MD data include the definition of isocommitor surfaces,[2−4] Bayesian analysis methods,[5] nonlinear dimensionality reduction,[6−9] genetic neural network algorithms,[10,11] and likelihood maximization;[12] these approaches take advantage of the large number of rare events that can be harvested from methods like the string method,[13−15] transition path sampling,[16−19] metadynamics,[20] and milestoning[21,22] once means for distinguishing the stable states of a reaction are known.

In the current work, we apply the recently developed Locally Scaled Diffusion Map (LSDMap) technique to the characterization of the dynamics of a designed 20-residue, three-stranded antiparallel β-sheet miniprotein (termed Beta3s). LSDMap has already been successfully applied to other systems: isomerization of alanine dipeptide, folding of a coarse-grained model of the SH3 protein domain,[23] and polymer reversal inside a nanopore.[24] Beta3s provides an ideal additional test for the LSDMap method, as it has been extensively studied with different approaches,[25−30] and these studies have shown that, although Beta3s consists of a single β-sheet, its folding dynamics is far from simple, as it involves a well-defined native state and several misfolded metastable states.

In previous work, the conformational space of Beta3s has been sampled by implicit solvent[31] MD simulations at 330 K for a total of 20 μs during which about 100 folding and unfolding events were observed.[28] In these simulations, Beta3s folds reversibly, without any bias, irrespective of the starting conformation. The interpretation of Beta3s simulations is significantly more challenging than the test cases previously used in the LSDMap approach. In particular, as a transferable force-field is used, there is no a priori information on the nature of the folded state, nor on the possibility of populating misfolded conformations. In contrast, previous applications of LSDMap have been limited to simplified protein models where empirical reaction coordinates are more easily used to interpret the results.[23,24]

The free energy surface of Beta3s has been previously characterized by two different methods for determining metastable states: the minimum-cut based free energy profile method[27,32] and kinetic grouping analysis.[28] Both methods require (geometric) clustering of the MD snapshots into nodes of a network whose links are the transitions observed during the equilibrium MD sampling.

Received: August 10, 2011
Revised: September 22, 2011
Published: September 23, 2011

The essential idea of these two methods is to group the clusters into free energy minima, according not to their standard structural features, but rather to the equilibrium dynamics. In other words, the MD trajectory is used to determine the populations of the states, which provide the relative free energies and the rates of transition between the states, which yield the free energy barriers. Notably, both methods yield the same free energy basins of Beta3s whose most populated state is the designed antiparallel β-sheet structure (population of about 35% at 330 K). Interestingly, the denatured state of Beta3s presents several misfolded traps stabilized by enthalpy (with a cumulative population of about 20%), as well as a basin with fluctuating helical conformations and a heterogeneous entropic state populated at about 10% and 35%, respectively.

Additionally, in the work of Qi, et al.[11] the same long simulations obtained by Muff, et al.[28] were used to construct networks of transition ensembles. Committor probabilities were then determined to the native state and misfolded states, and these were used as input to a genetic neural network (GNN) algorithm to extract physically meaningful collective coordinates. They found the sum of the distances of eight key hydrogen bonds along the backbone to be the best geometric coordinate for the overall folding/unfolding reaction and identified three distinct folding pathways and the distinct coordinates that characterized them.

In the present work, we relate the diffusion coordinates and time scales that emerge from LSDMap analysis to the metastable states identified in Krivov et al.[27] and the reaction coordinates of Qi et al.[11] We also introduce an improved procedure for defining the pathways. This, together with the LSDMap analysis, provides insights into the dynamics unique to each pathway.

2. METHODS

2.1. LSDMap Analysis. As mentioned above, we apply the LSDMap[23] formalism to the same Beta3s MD data used by Krivov et al.[27] and Qi et al.[11] As both the locally scaled diffusion map (LSDMap) methodology and Beta3s MD simulation have been detailed in previous publications, we give only a brief overview here, and refer the interested reader to the original publications.

As detailed in ref 27, MD simulations of the Beta3s system were performed with the program CHARMM.[33] All of the polar hydrogen atoms and the heavy atoms are included in the simulation; solvation effects are incorporated through the solvent accessible surface area implicit solvent model.[31] The use of an implicit solvent model is supported by the fact that the same model used here (with the same surface-tension like parameters) has been used previously to collect statistically significant sampling of different processes that are computationally prohibitive with explicit solvent. These processes include the reversible folding of a simplified-sequence version of protein G,[34] the mechanical (un)folding of a helical peptide,[35,36] and the early steps of aggregation of the Alzheimer's amyloid-β peptide, which have revealed the amyloidogenic "hot-spots".[37] Moreover, using the same implicit solvent model, MD simulation of amyloid-forming peptides in the presence of small-molecule inhibitors of aggregation have shed light on the mechanism of inhibition.[38,39]

In the current work, configurations were sampled every 20 ps of MD simulation, for a total of 10^6 snapshots. We use a subset of the original data, by collecting configurations every 100 ps, for a total of 200 000 configurations that serve as input to the LSDMap calculation.

LSDMap is a recently developed[23] nonlinear dimensionality reduction technique for characterizing dynamics at different time scales in terms of a few collective coordinates. In general, the application of dimensionality reduction techniques is based on the assumption that the high dimensional data set under consideration lies on a manifold of much lower dimensionality than the full configuration space. LSDMap treats the whole ensemble of configurations sampled from an MD simulation as a noisy data set that resides on an low-dimensional underlying "manifold". This "manifold" is not a mathematically defined manifold with constant dimension, but it is locally heterogeneous, presenting a different intrinsic dimensionality in different regions of the configurational landscape.

These local differences are taken into account in the LSDMap formalism and are used to locally "renormalize" the landscape in the construction of global eigenfunctions of the Fokker−Planck (FP) operator. These eigenfunctions have been shown to represent good reaction coordinates for describing the diffusive dynamics of systems sampled by MD simulations. The mathematical details on diffusion maps can be found in the original paper by Coifman and Lafon,[40] and applications to model nonmolecular Fokker−Planck systems in ref 41. The full details on the motivation and construction of the locally scaled version of diffusion map (that is, LSDMap) are presented in ref 23. For completeness, a brief overview of the LSDMap procedure is provided below.

Macromolecular systems in the high-friction regime obey the FP equation

$$\frac{\partial p}{\partial t} = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i}\left(\frac{1}{\beta}\frac{\partial}{\partial x_i} + \frac{\partial E}{\partial x_i}\right)p = -\mathbf{H}_{\mathrm{FP}}p \tag{1}$$

where $p = p(x,t)$ is the probability density, N is the number of degrees of freedom, $\beta = 1/(k_{\mathrm{B}}T)$, k_{B} is Boltzmann's constant, T is the temperature, $E = E(x)$ is the potential energy function, and $\mathbf{H}_{\mathrm{FP}}$ is the FP operator. The FP equation can be cast as an eigenvalue problem, and in systems for which there is a separation of time scales between m slow collective modes and the remaining faster ones, the solution can be written as

$$p(x,t) \simeq \phi_0(x) + \sum_{i=1}^{m} c_i\phi_i(x)e^{-\lambda_i t} \tag{2}$$

where ϕ_i and λ_i are the eigenfunctions and associated eigenvalues of $\mathbf{H}_{\mathrm{FP}}$, respectively, and the coefficients c_i are determined by the initial distribution $p(x,t = 0)$. The zeroth eigenfunction ϕ_0 corresponds to the Boltzmann distribution; the first eigenfunction ϕ_1 to the collective motion with the slowest time scale; the second eigenfunction ϕ_2 to the second slowest collective motion, etc. When normalized by the Boltzmann distribution, these eigenfunctions ϕ_i/ϕ_0 possess the qualities of good reaction coordinates in the sense that the dynamics on the longer time scales can be described using only these m collective degrees of freedom.[42,43] The functions $\psi_i = \phi_i/\phi_0$ are eigenfunctions of the backward FP operator. These coordinates are the "diffusion coordinates" (DCs); the "diffusion map" is the nonlinear mapping from the molecular configuration space to the diffusion coordinate space.

The LSDMap is based on the kernel

$$K_{ij} = \exp\left(-\frac{||x_i - x_j||^2}{2\varepsilon_i\varepsilon_j}\right) \tag{3}$$

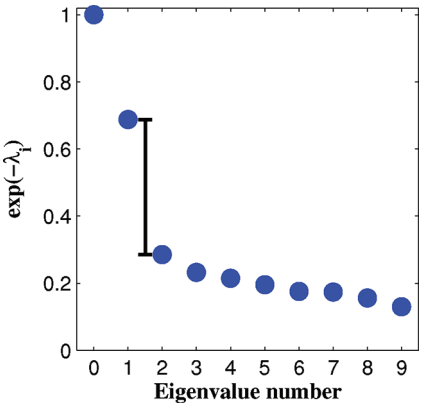


Figure 1. The exponential of the negative of the FP eigenvalues λ_i as a function of eigenvalue number. The presence of a spectral gap, denoted by the vertical black bar, indicates that the essential features of the dynamics can be captured by the first DC.

where $\|x_i - x_j\|$ is the root-mean-square deviation (rmsd) between the two configurations x_i and x_j, and ε_i is the local scale for x_i. This local scale represents the radius in configuration space around x_i within which the underlying manifold can be approximated by a hyperplane tangent to the manifold, that is, is approximately linear. The procedure to estimate the local scale ε_i around every point x_i in the data set is detailed in ref 23. The kernel K_{ij} is related to the "ease" with which x_i can diffuse into x_j. A normalized version of this kernel (see ref 23) represents the Markov matrix for the data set of molecular configurations, and the diagonalization of such a matrix yields a set of vectors that serve as diffusion coordinates.

2.2. Construction of Pathways. The network of states emerging from the dynamics of Beta3s is detailed in Krivov et al.[27] and Qi et al.[11] Nine major non-native states were previously identified (see Figure 7 of ref 27). The different pathways are constructed by considering the connections between the native state and all the non-native states; only a small number of transitions occur between the non-native states. Consequently, the transitions between different non-native states and the native state define the possible folding and unfolding pathways.

In earlier work,[11] the dynamics of structures were characterized by their commitment probabilities for folding from state i to the native state ($p_{\mathrm{fold},i}$) and for unfolding from the native state to state i ($p_{\mathrm{unfold},i}$). Ideally, if a structure is specifically associated with pathway i, it would satisfy the condition $p_{\mathrm{fold},i} + p_{\mathrm{unfold},i} = 1$. However, because $p_{\mathrm{fold},i}$ and $p_{\mathrm{unfold},i}$ are calculated separately based on the statistics of the network, their sum can exceed one. Qi et al.[11] grouped structures together when $p_{\mathrm{fold},i} + p_{\mathrm{unfold},i} \in (0.8,1.2)$. Here, we employ a more restrictive scheme based on a series of state-to-state transitions. Specifically, we identify trajectory segments that go from the native state to non-native state i and remove those segments that also visit the other major non-native states. By this procedure, three main pathways are detected with sufficient statistics needed for a detailed analysis; these are the three pathways that are characterized in the following.

We identify 17 folding events and 18 unfolding events that follow pathway 1. There are 48533 structures that are clustered based on rmsd into 2523 nodes with 1569 pairwise links within nodes and 7579 pairwise links between nodes. We varied the commitment time to maximize the number of structures with

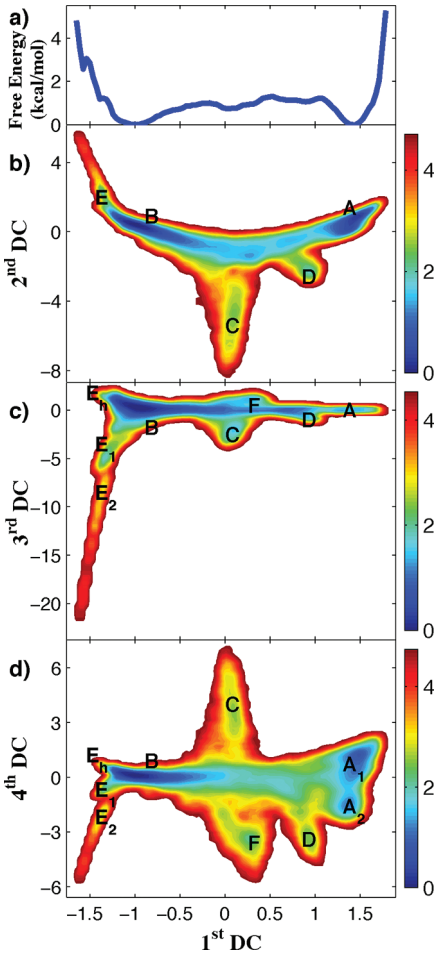


Figure 2. Free energy (in units of kcal/mol) as a function of the diffusion coordinates. Free energy as a function of the first DC (a), free energy as a function of the first DC and second DC (b), first DC and third DC (c), and first DC and fourth DC (d). The metastable states associated with the free energy minima are labeled with letters A−F.

$p_{\mathrm{fold},i} + p_{\mathrm{unfold},i} \in (0.8,1.2)$; this yields a commitment time of 11.3 ns, compared with 10 ns in previous work.[11] Using this commitment time, 45131 of the 48533 structures (nearly 93%) satisfy the commitment probability sum condition even though it was not explicitly used in construction of the pathway, which validates the procedure. By the same token, we identify 11 folding events and 10 unfolding events that follow pathway 2; there are 20658 structures that are clustered into 1966 nodes with 1118 pairwise links within nodes and 4728 pairwise links between nodes. Pathway 3 contains 6 folding events and 7 unfolding events; there are 9826 structures that are clustered into 1043 nodes with 607 pairwise links within nodes and 2478 pairwise links between nodes. 80% of the structures in the trajectories of both pathways 2 and 3 satisfy the commitment probability sum criterion with commitment times of 13.6 and 12.5 ns, respectively.

3. RESULTS

Figure 1 shows the first ten FP eigenvalues obtained by LSDMap applied to the MD data of Beta3s. As mentioned above, the zeroth eigenvalue $\lambda_0 = 0$ corresponds to the zeroth eigenfunction ϕ_0, that is the Boltzmann distribution. The first eigenvalue λ_1 corresponds to

Table 1. Pearson Correlation of Hydrogen Bonds with the First DC

hydrogen bond	Pearson correlation with the first DC
$H_{10}-O_3$	0.77
H_3-O_{10}	0.84
H_8-O_5	0.59
H_5-O_8	0.75
$H_{18}-O_{11}$	0.71
$H_{11}-O_{18}$	0.77
$H_{16}-O_{13}$	0.56
$H_{13}-O_{16}$	0.69

the collective motion with the slowest time scale, the second eigenvalue λ_2 corresponds to the collective motion with the second slowest time scale, and so on. The large gap denoted by the black vertical bar in Figure 1 shows that there is a separation of time scales, and one collective slow motion dominates the dynamics of Beta3s on long time scales. This slow collective motion corresponds to the folding and unfolding of the peptide. A detailed discussion on the relation between the different collective motions at different time scales and the diffusion coordinates is presented in section 3.1.

Figure 2a shows the free energy as a function of the first DC. The one-dimensional folding and unfolding free energy barrier of about 1.5 kcal/mol predicted by the first DC is comparable to the free energy barrier as a function of the best overall reaction coordinate presented in Figure 7a of the work of Qi, et al.[11] where a genetic neural network (GNN) analysis was used. Indeed, we find a high correlation (Pearson correlation coefficient ~0.89) between the first DC and the best reaction coordinate describing the overall folding reaction in ref 11, that is, the sum of the hydrogen bond distances for the eight hydrogen bonds between atoms H_3-O_{10}, H_5-O_8, $H_{11}-O_{18}$, $H_{13}-O_{16}$, $H_{10}-O_3$, H_8-O_5, $H_{18}-O_{11}$, and $H_{16}-O_{13}$. Throughout our analysis, we use the same numbering scheme for the hydrogen bonds as in ref 11, that is also shown in Figure 8. For example the hydrogen bond distance for H_3-O_{10} corresponds to the distance between the H atom on the third nitrogen atom from the N-terminus and the tenth O atom from the N-terminus. As shown in Table 1 and consistent with the results presented in ref 11, the correlation of the first DC with the O−H distance of each hydrogen bond individually is lower than that with their sum, suggesting that the formation of this set of eight hydrogen bonds is a better variable for describing the overall folding than any single hydrogen bond by itself.

Figure 2b shows the free energy as a function of the first DC and the second DC. The two deepest minima correspond to the folded (A) and unfolded (B) states. The second DC represents a motion deviating from the main folding path, and defines several local free energy minima corresponding to misfolded states. The local minima are also evident, from a different "angle", in the plot of the free energy as a function of the first DC and third DC (Figure 2c), and first DC and fourth DC (Figure 2d). State E, which appears as one misfolded minimum in the free energy as a function of the first DC and the second DC, splits into three substates (E_1, E_2, and E_h) when the additional coordinates third DC and fourth DC are used. Additionally, the fourth DC also describes a motion internal to the folded state, that appears as a split minimum in the folded basin of the free energy in Figure 2d.

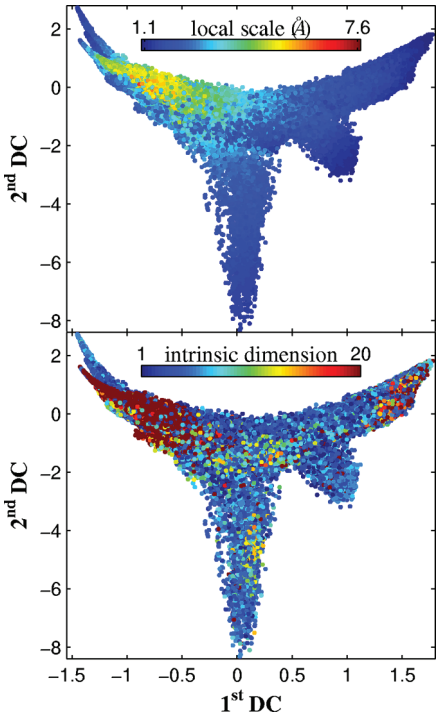


Figure 3. Local geometric indicators associated with each molecular configuration are plotted as a function of the first DC and second DC. Each dot corresponds to one of the configurations in the data set, and different colors indicate different values for the local scale ε (top panel), and number of local intrinsic dimensions (bottom panel).

As discussed in section 2, the local geometric properties of high-dimensional MD data sets are expected to vary from region to region in the configuration space. The LSDMap approach quantifies this heterogeneity by providing a local length scale and local dimensionality around each point in the data set. Figure 3 illustrates the variability of the local scales and the number of local intrinsic dimensions on the Beta3s landscape as determined by LSDMap, plotted as a function of first DC and second DC. For each configuration of the sample, the local scale is shown on the top panel and the local intrinsic dimension on the bottom panel. By comparing this figure with the free energy landscape as a function of the first DC and second DC reported in Figure 2b, we observe that, roughly speaking, configurations near the free energy minima have a larger number of intrinsic dimensions than those close to free energy barriers, as expected.[23,24] In addition, the local scale is much larger in the unfolded state than in any other states. This large difference in local scale suggests that configurations in the entropically stabilized region of the unfolded state are separated by a larger rmsd than configurations within the native basin or the enthalpic traps.

3.1. Characterization of States. The states corresponding to the different free energy minima are labeled with letters from A−F in Figure 2 and correspond to different metastable states: folded (A), unfolded (B), and several partially misfolded states (C−F). The average structural features of each of these states is investigated by considering the probability to form hydrogen bonds in each of these minima. The results are reported in terms of "hydrogen bond maps" in Figure 4 and Figure 5. A hydrogen bond is considered formed if the OH distance is smaller than 2.4 Å and the NHO angle is larger than 2.44 rad. The folding pathways that emerged in the analysis discussed above

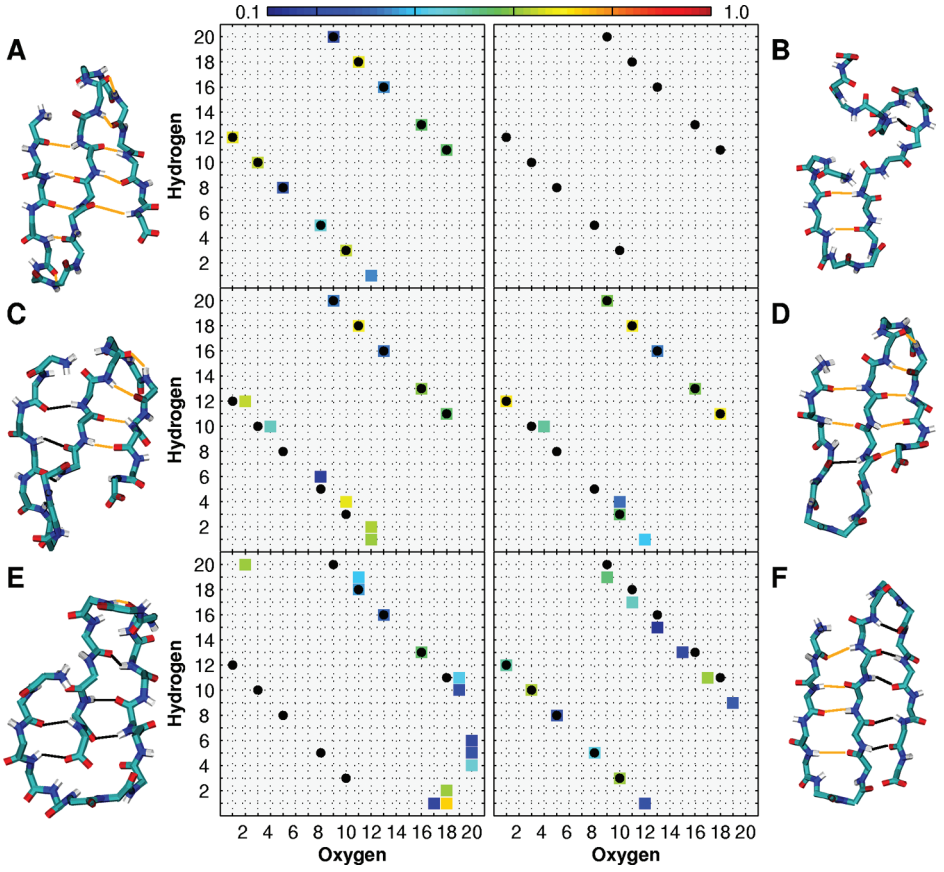


Figure 4. Representative structure and probability to form hydrogen bonds in the main different metastable states identified by the first four DCs. The letters A−F correspond to the states identified in Figure 2. Native and non-native hydrogen bonds formed in the different structures are shown in orange and black, respectively. The black dots in the hydrogen bond maps indicate the residue pairs that are identified by Qi, et al.[11] as forming native hydrogen bonds.

(see section 2.2 and ref 11) are then interpreted in terms of transitions between these minima. The states identified by LSDMap are also compared with the states determined by the minimum-cut based free-energy profile method for the same system in ref 27.

3.1.1. Folded State. The hydrogen bond map presented in Figure 4 for state A shows that all ten native hydrogen bonds identified in the paper by Qi et al.[11] are formed in state A, although with different probabilities. One additional hydrogen bond, H_1-O_{12}, can be formed with probability ∼0.4 within this basin. This extra hydrogen bond is at the beginning of the N-terminal antiparallel β-sheet, and it is consistent with the overall native structure of Beta3s.

While the native basin appears as a single minimum in the free energy as a function of the first three DCs, the fourth DC splits the native state into two substates, indicating fluctuations inside the folded state on a fast time scale (compared with the much slower overall folding). The two substates present a slightly different hydrogen bond pattern: the hydrogen bond H_9-O_{20} is formed with probability less than 0.1 in state A_1, while it is formed with probability ∼0.5 in state A_2; additionally, the hydrogen bond H_6-O_4 can form with probability ∼0.1 in state A_2. The probability of formation of all the other hydrogen bonds is essentially indistinguishable from the map reported in Figure 4, both in A_1 and A_2 (see Figure S1 in Supporting Information).

3.1.2. Unfolded State. The free-energy minimum labeled B in Figure 2 corresponds to the unfolded state, with no hydrogen bonds formed with probability higher than 0.1. The configurations in state B are quite different from each other and, although different sets of hydrogen bonds can be transiently formed, on average they do not present a persistent hydrogen bond pattern. The unfolded state can be understood as the gathering of many low-populated partially misfolded states. The reason why they are gathered together by the LSDMap analysis is that these mostly unfolded structures can diffuse easily into each other despite of the large rmsd between them. Additionally, it is clear

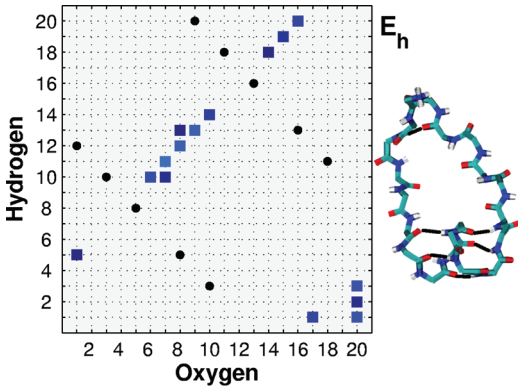


Figure 5. Representative structure and probability to form hydrogen bonds in the helical state E_h. The color scheme is the same as Figure 4.

dx.doi.org/10.1021/jp2076935 |J. Phys. Chem. B 2011, 115, 13065–13074

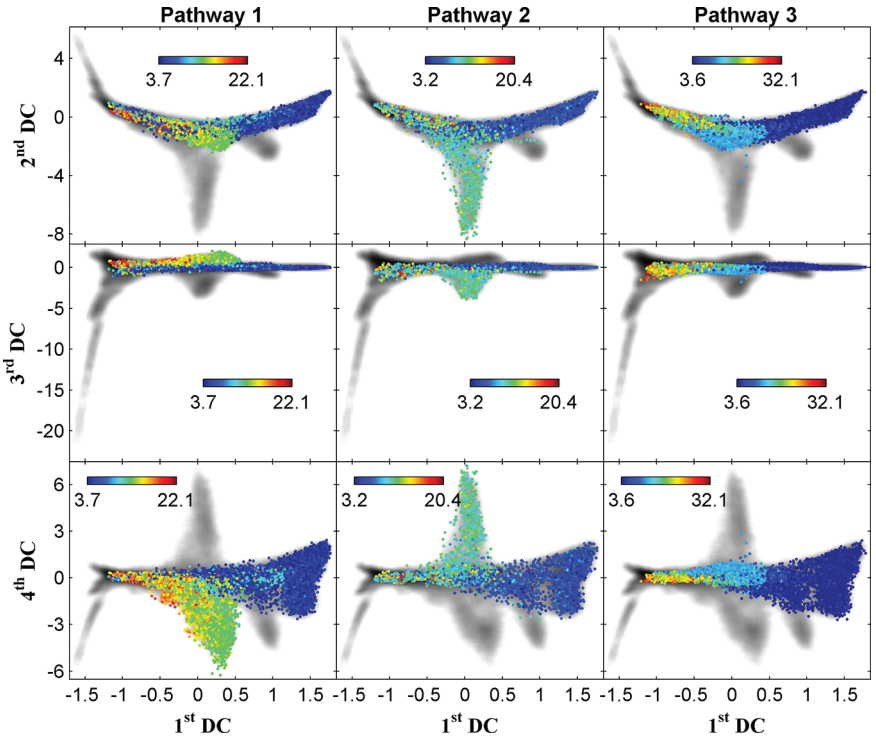


Figure 6. Projection of the configurations corresponding to the three main folding pathways onto the free energy landscape as a function of first DC and second DC (first row), first DC and third DC (second row), and first DC and fourth DC (third row). The projected points are colored according to the value of the best physical variable identified by the one-descriptor NN models (see text for details). The free energy projections are shown in grayscale. The figures in the first column correspond to pathway 1, and the projections are colored by the sum of the hydrogen bond distances between atoms $H_{13}-O_{16}$ and $H_{16}-O_{13}$; in the second column to pathway 2, colored by the sum of the distances between the geometric centers of the side chains of residue 4 and 9; and in the third column to pathway 3, colored by the sum of the hydrogen bond distances between atoms H_3-O_{10} and $H_{10}-O_3$.

Table 2. Top One-Descriptor NN Models for Major Folding and Unfolding Pathways[a]

pathway	$p_{\mathrm{unfold},i}$ as target	rms error	$p_{\mathrm{unfold},i}$ and $p_{\mathrm{fold},i}$ as target	rms error
1	d_{HB} of 13−16	0.1908	d_{HB} of 3−10, 13−16	0.2778
	distance of $H_{13}-O_{16}$	0.1935	distance of $H_{13}-O_{16}$	0.2888
	q_{23}	0.1987	d_{HB} of 13−16	0.2898
2	d_{SC} of 4−9	0.2034	d_{SC} of 4−9	0.2746
	$E_{\mathrm{SC}}^{\mathrm{VDW}}$ of 4−9	0.2093	distance of $H_{10}-O_3$	0.2766
	$E_{\mathrm{SC}}^{\mathrm{Elec}}$ of 4−9	0.2100	d_{SC} of 4−9, 12−17, 5−8, 13−16	0.2787

pathway	$p_{\mathrm{fold},i}$ as target	rms error	$p_{\mathrm{unfold},i}$ and $p_{\mathrm{fold},i}$ as target	rms error
3	d_{HB} of 3−10	0.1968	CA_3-CA_4-CA_5-CA_6	0.2924
	CA_3-CA_4-CA_5-CA_6	0.2016	CB_4-CA_4-CA_5-CB_5	0.3018
	$E_{\mathrm{HB}}^{\mathrm{Elec}}$ of 3−10	0.2017	CA_2-CA_3-CA_4-CA_5	0.3239

[a] d_{HB} and d_{SC} denote the sum of distances between hydrogen bonding backbone O and H atoms and between the geometric centers of side chains, respectively. $E_{\mathrm{HB}}^{\mathrm{Elec}}$, $E_{\mathrm{SC}}^{\mathrm{VDW}}$, and $E_{\mathrm{SC}}^{\mathrm{Elec}}$ are energy terms. HB and SC subscripts denote hydrogen bond and side chain interactions, respectively; elec and VDW superscripts denote the electrostatic and van der Waals parts of the energy function. q_{23} indicates the fraction of contacts within the C-terminal hairpin; contacts were defined as in ref 25.

Table 3. Definitions of the Native and Non-native States for $p_{\mathrm{fold}}^{\mathrm{MSM}}$ Calculations[a]

pathway	native states	non-native states
1	d_{HB} of 13−16 < 4.5 Å	d_{HB} of 13−16 > 12.5 Å
2	d_{SC} of 4−9 < 4.75 Å	d_{SC} of 4−9 > 10.75 Å
3	(a) d_{HB} of 3−10 < 4.25 Å	(a) d_{HB} of 3−10 > 12.05 Å
	(b) d_{HB} of 3−10 < 4.25 Å and $\lvert CA_3\text{-}CA_4\text{-}CA_5\text{-}CA_6\rvert$ > 157°	(b) d_{HB} of 3−10 > 12.05 Å and $CA_3\text{-}CA_4\text{-}CA_5\text{-}CA_6 \in (-33°, 59°)$

[a] We define the native and non-native states according to the distributions of the reaction coordinates (Figure S6 in Supplementary Material).

from Figure 2 that both the native state and most of the misfolded states that appear as separate free energy minima are directly connected to the entropically stabilized region of the unfolded state; this is not surprising as we expect the misfolded structures to at least partially unfold before proceeding to the folded state, consistent with the results of ref 28.

3.1.3. Misfolded States. The main six misfolded states (labeled as C, D, E_1, E_2, E_h, and F in Figure 2 and Figure 4) can be defined according to the free energy as a function of the first several DCs. The probability of forming different hydrogen bond patterns provides information on the average structure of the peptide in these states that can be compared to the results of previous studies.[11,27]

Figure 4 shows that in state C, the C-terminal part of the antiparallel β-sheet is completely formed, while a mismatch of the native hydrogen bond pattern is observed in the N-terminal part. This state corresponds to the state named "Ns-or" (N terminal strand out of register) in ref 27.

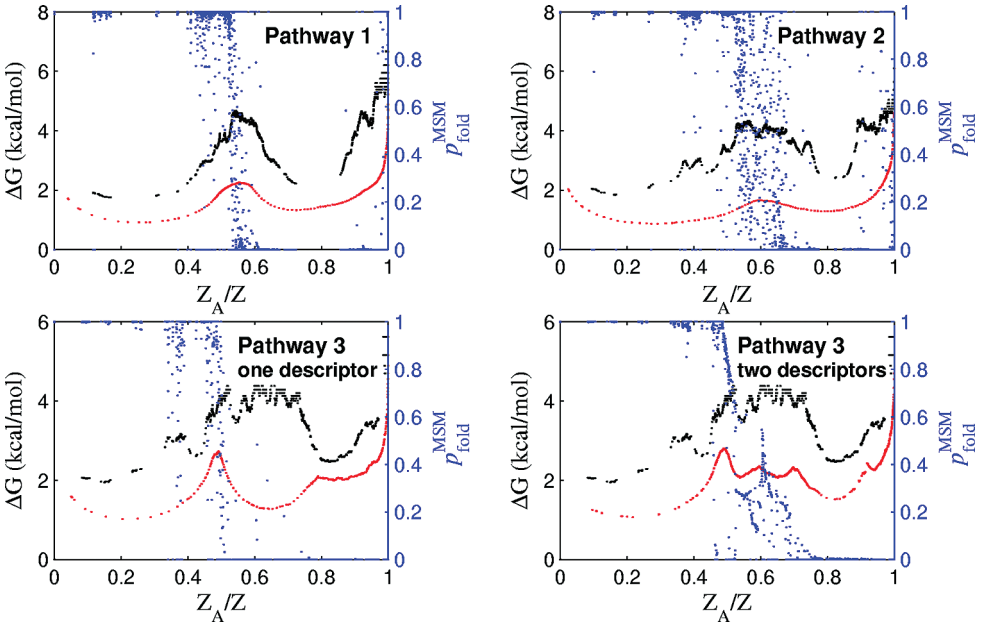


Figure 7. Cut-based free energy profile for networks clustered according to the selected coordinates (red) or all-atom rmsd (black), and the folding probability corresponding to a Markov state model (blue) as a function of the relative partition function Z_A/Z for specific pathways. Pathways are as marked.

The average hydrogen bond pattern in state D looks very similar to what observed in state C, although the N-terminal hairpin in state D appears to be a little closer to the native state structure (or, equivalently, state C is more misfolded than state D), consistent with their relative position along the first DC. In particular, configurations both in state C and in state D have high probability of forming the pair of non-native hydrogen bonds $H_{10}-O_4$ and H_4-O_{10}. In addition, non-native hydrogen bonds $H_{12}-O_2$, H_6-O_8, and H_2-O_{12} can be formed in state C, while the native hydrogen bonds $H_{12}-O_1$, H_3-O_{10}, and H_1-O_{12} can be formed in state D. It is worth noting that although state C and D only differ by a few hydrogen bonds in the N-terminal hairpin, the free energy projections in Figure 2 suggests that state C and D are not directly connected: the configurations in state C need to unfold to some extent before they can proceed to state D (and vice versa).

The average hydrogen bond pattern of state F shows that the N-terminal β-hairpin is correctly formed while the C-terminal one is out of register. This state corresponds to the basin named as "Cs-or" (C-terminal strand out of register) in ref 27.

The free energy as a function of the first few DCs (Figure 2) suggests that while states C, D, and F are accessible from partially folded states along the folding process, states E_1, E_2, and E_h are directly connected only to the unfolded state. Indeed, these states have a completely non-native topology and need to unfold, at least partially, to proceed toward the correctly folded state. The average hydrogen bond map of states E_1 and E_2 presents the C-terminal β-hairpin partially out of register, with a few native hydrogen bonds formed with probability ~0.2, and the formation of a non-native parallel β-hairpin connecting the N and C termini. These states correspond to the basins named "Ch-curl1" and "Ch-curl2" in ref 27.

The probability of forming different hydrogen bonds is very similar in states E_1 and E_2, and the average over the two ensembles is presented in Figure 4. The main difference between these two states is that hydrogen bonds $H_{19}-O_{10}$, $H_{12}-O_{17}$ and $H_{13}-O_{17}$ can form with probability ~0.2−0.3 in state E_1 but are not present in state E_2, while $H_{14}-O_{12}$ appear with a probability ~0.2 in E_2 but not in E_1 (Figure S2 in Supporting Information).

As shown in Figure 5, state E_h in Figure 2 corresponds to helix-like configurations, that is, to the state named "helix" in the analysis reported in ref 27. This state appears projected onto the same position of the first DC as misfolded state E_1 and E_2, along the left boundary of the entropic state B defined by the LSDMap analysis, and emerges as an independent state when at least the first three DCs are used. The fact that the helical state is included in state E in Figure 2b, and it is very close to the entropic basin along the first DC indicates that the formation and unfolding of this misfolded helix from the entropic basin, and the interconnection of the helix into the Ch-curl state, is very rapid with respect to the formation/unfolding of other misfolded states. However, according to the distance between the helical state and the native state along the first DC, it is clear that the E_h state, together with the Ch-curl state, are the states kinetically most distant from the native state. In other words, although the helix can be rapidly form from the entropic basin, it takes longer to fold from the helical state than from the other misfolded states. This result is consistent with the findings by the minimum-cut based free energy profiles of Krivov et al.[27]

3.2. Heterogeneity of Folding Pathways. As shown in the free energy projections onto the first four DCs Figure 2, different misfolded states are involved in the folding and unfolding process with the longest time scale along the first DC. Therefore, when extracting physical details of the dynamics, multiple pathways should be taken into account, despite the similar time scale of the overall folding transitions in each pathway. Previous work[11] on Beta3s analyzed three folding pathways for this protein, involving different misfolded configurations. As discussed in section 2.2, here we group states into these pathways according to more restrictive criteria than previously used, to eliminate trajectory segments that visit more than one major misfolded state. This new procedure allows us to calculate free energy profiles along the relative partition function Z_A/Z and selected descriptors using only the states in a single pathway, in contrast to Figure 13

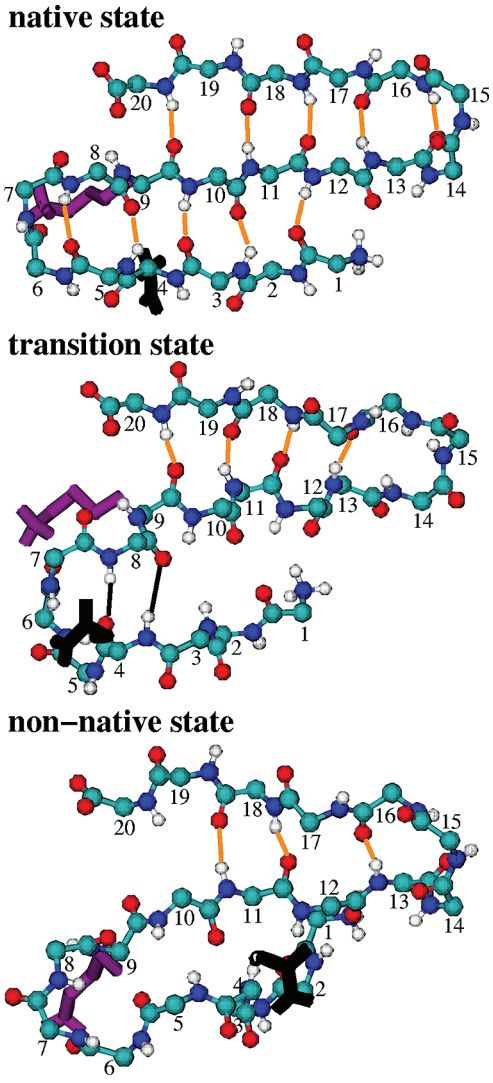


Figure 8. Representative structures in pathway 2 illustrating the rearrangement of the side chains of residues 4 and 9, shown in black and purple, respectively. Native and non-native hydrogen bonds formed in the different structures are shown in orange and black lines, respectively.

of ref 11, in which free energy profiles using all states were calculated for pathway-specific variables. We are also able to obtain pathway-specific dynamic information, for example, position-dependent diffusion constants (see Figure S3 in Supporting Information). Such information is important as it can help to evaluate the quality of a selected coordinate: simple Brownian motion with uniform diffusion constant is expected for a "good" coordinate.[44]

We begin by interpreting the pathways in terms of the population of the metastable states identified by LSDMap. Figure 6 presents the projection of the configurations visited in the different pathways on the free energy landscape as a function of the first few DCs. It is clear that different sets of metastable states are visited in different pathways. In particular, pathway 1 visits the misfolded state F (corresponding to Cs-or in ref 27), which equilibrates with the unfolded basin on a time scale much faster than the overall folding process (as state F clearly emerges as a distinct state only when projected on the fourth DC, and it is not visible at all when only the first two DCs are used). This interpretation of pathway 1 is consistent with the description presented in Figure 11 of ref 11: the main non-native state visited during this pathway involves the partial misfolding of the C-terminal hairpin, as also evident from the comparison of Figure 6 and Figure 4. On the contrary, the main non-native state visited in pathway 2 involves the partial misfolding of the N-terminal hairpin (Ns-or), again consistent with the results of Qi et al.[11] Pathway 3 appear to visit a misfolded state included in the entropic basin, which suggests that the misfolded configuration in pathway 3 is more rapidly unfolded, compared with that of pathway 1 and 2. This is in consistence with the misfolded configurations from the three pathways in Figure 11 of ref 11. The distinct population of different regions of the free energy landscape as a function of the first few DCs further validates the improved pathway construction procedure introduced in section 2.2.

To relate the DCs to physically intuitive variables, we apply the NN part of the GNN procedure to these pathways. Details of the method and the choice of descriptors is the same as in ref 11 except that we exhaustively enumerate the one- and two-descriptor models rather than searching them with a genetic algorithm. In constructing the database of input structures for the NN procedure, it is important to have a roughly uniform distribution of commitment probability values, so that no particular values dominate the fit. We selected 1200 and 1920 structures with roughly uniform distributions of $p_{\mathrm{unfold},i}$ for pathways 1 and 2. For pathway 3, there were more limited statistics, and we could not obtain a sufficient number of structures with intermediate $p_{\mathrm{unfold},i}$ values. However, as the commitment probabilities are approximate (so $p_{\mathrm{unfold},i} \neq 1 - p_{\mathrm{fold},i}$), we were able to construct a database of 500 structures with roughly uniform distribution of $p_{\mathrm{fold},i}$ values for pathway 3. Given these databases, two sets of NN calculations are performed for each: one with only a single commitment probability as the target ($p_{\mathrm{fold},i}$ or $p_{\mathrm{unfold},i}$, depending on the pathway); the other with simultaneous prediction of $p_{\mathrm{fold},i}$ and $p_{\mathrm{unfold},i}$. The best one-descriptor models obtained are listed in Table 2. The mean square errors per target are comparable and the descriptors chosen are consistent in both cases. Consistent with the descriptions of the pathways above, individual pathways are best characterized by coordinates that track the formation of specific hairpins. We consider each pathway individually below.

To evaluate the different descriptors, we compare cut-based free energy profiles[27] for networks clustered according to either the selected coordinates or all-atom rmsd. For pathway 1, we group the 48533 structures into 186 bins (nodes) with width 0.1 Å in d_{HB} of 13−16 (the sum of the hydrogen bond distances between atoms $H_{16}-O_{13}$ and $H_{13}-O_{16}$). There are 144 pairwise links within nodes and 9269 pairwise links between nodes. Good correspondence is found between the two cut-based free energy profiles (Figure 7), which suggests that the simple geometric variable captures the relevant dynamics. We also calculated the folding probability corresponding to a Markov state model $p_{\mathrm{fold}}^{\mathrm{MSM}}$ using the definitions of folded and unfolded states in Table 3. The small spread in $p_{\mathrm{fold}}^{\mathrm{MSM}}$ in the transition region further supports the dynamic relevance of the selected descriptors (see Figure 7). The projection of the free energy onto d_{HB} of atoms 13−16 itself (Figure S3 in Supporting Information) has a broad barrier between 6 and 12 Å, which corresponds to breaking of backbone hydrogen bonds (Figure S4 in Supporting Information).

Analogous calculations are performed for pathway 2. The most highly ranked descriptor is d_{SC} of 4−9 (the sum of the distances

Table 4. Top Two-Descriptor NN Models[a]

pathway	reaction coordinates		rms error
1	CB_{15}-CA_{15}-CA_{16}-CB_{16}	N_{15}-CA_{15}-C_{15}-N_{16}	0.1750
	O_{14}-C_{14}-C_{15}-O_{15}	N_{15}-CA_{15}-C_{15}-N_{16}	0.1781
	cos angle b/w C═O of 15−14	cos angle b/w C═O of 16−15	0.1783
	cos angle b/w C═O of 16−15	N_{14}-CA_{14}-C_{14}-N_{15}	0.1785
	N_{13}-N_{15}-N_{17}	distance of H_{13}−O_{16}	0.1789
	distance of H_{13}−O_{16}	ASA of 14	0.1794
	distance of H_{13}−O_{16}	sum of CA_5-CA_7-CA_9 and CA_{12}-CA_{14}-CA_{16}	0.1794
2	interaction b/w residues 1−5 and 8−12	d_{SC} of 4−9	0.1779
	interaction b/w residues 1−5 and 8−12	E_{SC}^{VDW} of 4−9	0.1785
	interaction b/w residues 1−5 and 8−12	E_{SC} of 4−9	0.1796
	distance of H_{12}−O_1	E_{SC}^{VDW} of 4−9	0.1801
	distance of H_{12}−O_1	d_{SC} of 4−9	0.1802
	d_{SC} of 4−9	CA_5-CA_7-CA_9	0.1804
3	$ASA_{side\ chain}$ of 2	E_{HB} of 3−10	0.1773
	$ASA_{side\ chain}$ of 2	E_{HB}^{Elec} of 3−10	0.1787
	ASA of 1	E_{HB}^{Elec} of 3−10	0.1788
	distance of CB_5-CB_8	CA_2-CA_4-CA_6	0.1803
	E^{Elec} of H_3−O_{10}	CA_3-CA_4-CA_5-CA_6	0.1806
	$ASA_{sidechain}$ of 2	d_{HB} of 3−10	0.1809

[a] ASA denotes accessible surface area.

between the geometric centers of the side chains of residue 4 and 9). This descriptor was selected in ref 11, but less consistently. As suggested by the representative structures in Figure 8, unfolding is necessary for the side chains to move from opposite sides of the plane (as in the non-native state) defined by the β-sheet to the same side (as in the native state). This finding is consistent with the free energy (Figure 2b) as a function of the first two DCs that shows the configurations in state C need to unfold partially before they can fold to the native state. We group the 20658 structures of pathway 2 into 171 bins (nodes) with width 0.1 Å in d_{SC} of atoms 4−9. There are 102 pairwise links within nodes and 6189 pairwise links between nodes. Again, there is good correspondence with the rmsd based profile (Figure 7).

The best descriptor for pathway 3 is obtained by using d_{HB} of atoms 3−10. While there is a barrier along the cut-based free energy profile obtained from a network clustered according to d_{HB} of 3−10 (Figure 7), considerably better correspondence with the rmsd-based network is obtained when we also group structures according to the descriptor giving the second best prediction, the pseudodihedral angle between the CA atoms of residue 3, 4, 5, and 6 (Figure 7). The free energy projection on these two variables (Figure S3 in Supporting Information) indicates that there are multiple minima that are well-separated but overlap in each of the individual coordinates, consistent with the need for two variables.

Given the results for pathway 3, we exhaustively enumerated all possible two-descriptor models for the three pathways (Table 4). These do yield noticeably better prediction (rms error <0.18). However, because the descriptors are often selected to complement each other to maximize their joint information content, they can be considerably more challenging to interpret. Interestingly, the accessible surface area of the side chain of residue 2 is selected multiple times for pathway 3. This residue is a tryptophan, so it is likely that this descriptor reports on misfolding. Other combinations of descriptors result in particularly well-defined minima when the free energy is projected onto them (e.g., E^{Elec} of H_3−O_{10} and CA_3-CA_4-CA_5-CA_6 in Figure S5 in Supporting Information, the two blue-black regions on the left are the native basin, and the blue region on the right is the non-native basin). Overall, the two-descriptor analysis shows that very good prediction of commitment probability values can be obtained, and that the complexity revealed in the LSDMap analysis is reflected in the two-descriptor free energy projections.

To better connect the NN results with the LSDMap analysis, we show how the descriptors of the best one-descriptor models vary with the first few DCs (Figure 6) by coloring structures projected onto the DCs according to their NN descriptor values. The fact that structures of similar color fall together supports the choice of the selected descriptors to represent the different folding transitions. At the same time, the adjacency of different parts of the color scale is consistent with the improved prediction obtained with two descriptors and the need for multiple DCs.

4. CONCLUSION

We have applied the recently proposed LSDMap approach to the characterization of the folding process of Beta3s, a 20-residue, three stranded, antiparallel β-sheet miniprotein. Despite its small size Beta3s is a complex system as previous analysis of equilibrium all-atom MD simulations of reversible folding close to the melting temperature have revealed a very heterogeneous unfolded state ensemble[27,28] and multiple folding pathways.[11,25,28,30] The LSDMap approach has been previously tested and applied to extract collective coordinates to describe more simple dynamics, such as the isomerization of alanine dipeptide, the folding of a coarse-grained model of the SH3 protein domain,[23] and a coarse-grained polymer reversal inside a nanopore.[24] The identification of folding reaction coordinates is more challenging in the case of the Beta3s protein model used here: the peptide dynamics is simulated with an all-atom transferable potential and its associated free energy landscape is expected to be significantly more frustrated than in minimalist protein models.

The LSDMap analysis provides a set of diffusion coordinates (DCs) that allow the definition of a low-dimensional free energy landscape on multiple time scales. The global diffusion coordinates are generated in a multiscale fashion that takes into account the local heterogeneity and noise intrinsic to the MD data set. This approach can be seen as a "renormalization" of the configuration space according to the local geometry.

The first few DCs correspond to different collective motions on different time scales. The first DC describes the overall transition between the folded and unfolded state, the slowest motion of the system, and has a high correlation (~0.89) with the best empirical reaction coordinate for the folding process that has been identified in previous work.[11] While there is only a single well separated time scale, there are multiple pathways with similar barriers that contribute to this longest time scale. To obtain a complete understanding of these pathways, a genetic neural network algorithm has been used to extract physically intuitive variables for the first three most populated pathways. All the states and pathways identified here are in good agreement

with those analyzed previously,[11,27] and moreover, in the picture of a set of global orthogonal coordinates from the LSDMap analysis. In summary, the results presented show that LSDMap appears to correctly extract the local and global diffusion dynamics of Beta3s without any a priori knowledge of the system.

In this work we have applied the LSDMap approach to an equilibrium sampling of the Beta3s obtained with an implicit solvent model, as statistically meaningful data are computationally inaccessible with an explicit solvent model. However, in principle the application of LSDMap to a system sampled with an explicit solvent model is essentially the same as what presented here. It would be interesting to see if application of the LSDMap analysis to folding of Beta3s in explicit solvent yields DCs involving solvent degrees of freedom once such simulations become feasible.

We believe that the approach presented here can be used in general to characterize complex diffusion processes involving several (meta)stable states and multiple pathways.

■ ASSOCIATED CONTENT

Supporting Information. Additional figures. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
*E-mail: cecilia@rice.edu.

■ ACKNOWLEDGMENT

This work was supported by NSF (CDI-type I grant 0835824 and CAREER award CHE-0349303 to C.C. and CAREER award MCB-0547854 to A.R.D.), and the Welch Foundation (C-1570 to C.C.). The original MD simulations of Beta3s were carried out on the Schrödinger compute cluster at the University of Zurich. A.C. acknowledges the University and Canton of Zurich for the computational resources. Simulations and other computations were performed on the following shared resources at Rice University: the Cyberinfrastructure for Computational Research funded by NSF under grant CNS-0821727; the Shared University Grid at Rice University funded by NSF under grant EIA-0216467 and in partnership between Rice University, Sun Microsystems, and Sigma Solutions, Inc.; and a 2010 IBM Shared University Research (SUR) Award on IBM's Power7 high performance cluster (BlueBioU) to Rice University as part of IBM's Smarter Planet Initiatives in Life Science/Healthcare and in collaboration with the Texas Medical Center partners, with additional contributions from IBM, CISCO, Qlogic and Adaptive Computing.

■ REFERENCES

(1) Karplus, M. *J. Phys. Chem. B* **2000**, *104*, 11–27.
(2) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334.
(3) E, W.; Vanden-Eijnden, E. *J. Stat. Phys.* **2006**, *123*, 503–523.
(4) E, W.; Ren, W.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2005**, *413*, 242–247.
(5) Best, R.; Hummer, G. *Proc. Nat. Acad. Sci.U.S.A.* **2005**, *102*, 6732–37.
(6) Nguyen, P. H. *Proteins* **2006**, *65*, 898–913.
(7) Das, P.; Moll, M.; Stamati, H.; Kavraki, L. E.; Clementi, C. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 9885.
(8) Plaku, E.; Stamati, H.; Clementi, C.; Kavraki, L. E. *Proteins* **2007**, *67*, 897–907.
(9) Stamati, H.; Clementi, C.; Kavraki, L. E. *Proteins: Struct., Funct., Bioinf.* **2009**, *78*, 223.
(10) Ma, A.; Dinner, A. *J. Phys. Chem. B* **2005**, *109*, 6769–6779.
(11) Qi, B.; Muff, S.; Caflisch, A.; Dinner, A. R. *J. Phys. Chem. B* **2010**, *114*, 6979–6989.
(12) Peters, B.; Trout, B. L. *J. Chem. Phys.* **2006**, *125*, 054108.
(13) E, W.; Ren, W.; Vanden-Eijnden, E. *Phys. Rev. B* **2002**, *66*, 052301.
(14) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. *J. Chem. Phys.* **2006**, *125*, 024106.
(15) Maragliano, L.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2007**, *446*, 182–190.
(16) Dellago, C.; Bolhuis, P. G.; Csajka, F. S.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964.
(17) Dellago, C.; Bolhuis, P.; Geissler, P. L. *Adv. Chem. Phys.* **2002**, *123*, 1–78.
(18) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877.
(19) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.
(20) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.
(21) Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880.
(22) Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. *J. Chem. Phys.* **2008**, *129*, 174102.
(23) Rohrdanz, M. A.; Zheng, W.; Maggioni, M.; Clementi, C. *J. Chem. Phys.* **2011**, *134*, 124116.
(24) Zheng, W.; Rohrdanz, M. A.; Maggioni, M.; Clementi, C. *J. Chem. Phys.* **2011**, *134*, 144109.
(25) Ferrara, P.; Caflisch, A. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10780.
(26) Rao, F.; Caflisch, A. *J. Chem. Phys.* **2003**, *119*, 4035–4042.
(27) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701–8714.
(28) Muff, S.; Caflisch, A. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1185–1195.
(29) Muff, S.; Caflisch, A. *J. Phys. Chem. B* **2009**, *113*, 3218–3226.
(30) Muff, S.; Caflisch, A. *J. Chem. Phys.* **2009**, *130*, 125104.
(31) Ferrara, P.; Apostolakis, J.; Caflisch, A. *Proteins: Struct., Funct., Bioinf.* **2002**, *46*, 24–33.
(32) Krivov, S. V.; Karplus, M. *J. Phys. Chem. B* **2006**, *110*, 12689–12698.
(33) Brooks, B. R.; et al. *J. Comput. Chem.* **2009**, *30*, 1545–1614.
(34) Guarnera, E.; Pellarin, R.; Caflisch, A. *Biophys. J.* **2009**, *97*, 1737–1746.
(35) Ihalainen, J. A.; Paoli, B.; Muff, S.; Backus, E. H. G.; Bredenbeck, J.; Woolley, G. A.; Caflisch, A.; Hamm, P. *Proc. Nat. Acad. Sci.USA* **2008**, *105*, 9588.
(36) Paoli, B.; Pellarin, R.; Caflisch, A. *J. Phys. Chem. B* **2010**, *114*, 2023–2027.
(37) Cecchini, M.; Curcio, R.; Pappalardo, M.; Melki, R.; Caflisch, A. *J. Mol. Biol.* **2006**, *357*, 1306–1321.
(38) Convertino, M.; Pellarin, R.; Catto, M.; Carotti, A.; Caflisch, A. *Protein Sci.* **2009**, *18*, 792–800.
(39) Scherzer-Attali, R.; Pellarin, R.; Convertino, M.; Frydman-Marom, A.; Egoz-Matia, N.; Peled, S.; Levy-Sakin, M.; Shalev, D. E.; Caflisch, A.; Gazit, E.; Segal, D. *PloS One* **2010**, *5*, e11101.
(40) Coifman, R. R.; Lafon, S. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 5–30.
(41) Coifman, R. R.; Kevrekidis, I. G.; Lafon, S.; Maggioni, M.; Nadler, B. *Multiscale Model. Simul.* **2008**, *7*, 842.
(42) Jones, P. W.; Maggioni, M.; Schul, R. *Proc. Nat. Acad. Sci. U.S.A.* **2008**, *105*, 1803–1808.
(43) The functions $\psi_i = \phi_i/\phi_0$ are eigenfunctions of the backward FP operator.
(44) Krivov, S. V. *PLoS Comput. Biol.* **2010**, *6*, No. e1000921.