Systematic characterization of protein folding pathways using diffusion maps: Application to Trp-cage miniprotein

Sang Beom Kim, Carmeline J. Dsilva, Ioannis G. Kevrekidis, and Pablo G. Debenedetti

Citation: The Journal of Chemical Physics **142**, 085101 (2015); doi: 10.1063/1.4913322
View online: http://dx.doi.org/10.1063/1.4913322
View Table of Contents: http://scitation.aip.org/content/aip/journal/jcp/142/8?ver=pdfcov
Published by the AIP Publishing

Systematic characterization of protein folding pathways using diffusion maps: Application to Trp-cage miniprotein

Sang Beom Kim,[1] Carmeline J. Dsilva,[1] Ioannis G. Kevrekidis,[1,2] and Pablo G. Debenedetti[1,a)]

[1]*Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, USA*
[2]*Program in Applied and Computational Mathematics, Princeton University, Princeton, New Jersey 08544, USA*

(Received 20 December 2014; accepted 9 February 2015; published online 26 February 2015)

Understanding the mechanisms by which proteins fold from disordered amino-acid chains to spatially ordered structures remains an area of active inquiry. Molecular simulations can provide atomistic details of the folding dynamics which complement experimental findings. Conventional order parameters, such as root-mean-square deviation and radius of gyration, provide structural information but fail to capture the underlying dynamics of the protein folding process. It is therefore advantageous to adopt a method that can systematically analyze simulation data to extract relevant structural as well as dynamical information. The nonlinear dimensionality reduction technique known as diffusion maps automatically embeds the high-dimensional folding trajectories in a lower-dimensional space from which one can more easily visualize folding pathways, assuming the data lie approximately on a lower-dimensional manifold. The eigenvectors that parametrize the low-dimensional space, furthermore, are determined systematically, rather than chosen heuristically, as is done with phenomenological order parameters. We demonstrate that diffusion maps can effectively characterize the folding process of a Trp-cage miniprotein. By embedding molecular dynamics simulation trajectories of Trp-cage folding in diffusion maps space, we identify two folding pathways and intermediate structures that are consistent with the previous studies, demonstrating that this technique can be employed as an effective way of analyzing and constructing protein folding pathways from molecular simulations. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4913322]

I. INTRODUCTION

Proteins, such as enzymes, antibodies, and hormones, play crucial roles in numerous cellular processes. In order to function properly, proteins often need to fold into their correct three-dimensional structures; misfolding or aggregation can cause a serious loss of function or alter their activities, leading to diseases such as Alzheimer's or Parkinson's.[1] Therefore, understanding the detailed mechanisms of protein folding has long been an active area of research. Many fundamental aspects of protein folding, however, are still in veil.[2] Complementing experimental studies, molecular dynamics (MD) simulations can be employed to explore protein dynamics at atomic resolution. Miniproteins and small peptides have been designed with folding times in the microsecond range, and recent advances in computing hardware have enabled all-atom MD simulations on microsecond and even millisecond time scales.[3] In addition, many advanced sampling techniques[4–9] have been developed to study rare events, such as protein folding, that are otherwise challenging to sample on conventional simulation time scales. Because of the ever-increasing volume and complexity of simulation data, it is necessary to analyze systematically and extract automatically the relevant structural and dynamical information from the simulation trajectories.

Phenomenological observables, such as root-mean-square deviation (RMSD) and radius of gyration (Rg), are frequently used to describe protein folding. However, these variables are often inadequate to fully characterize the folding free energy landscape because they do not contain enough information regarding the complex dynamics of the systems under investigation.[10,11] As a result, many techniques have been developed for effective and informative dimensionality reduction. Principal component analysis (PCA),[12] a linear dimensionality-reduction method, projects the data onto a linear subspace that captures the maximum variance within the data. PCA, however, can fail to properly analyze complex systems such as large proteins.[10,13] The local dynamics of proteins can be sufficiently simple and linear to be accurately captured through PCA, but the complex and nonlinear overall energy landscapes of proteins create challenges in parsimoniously describing the global dynamics using linear techniques. Nonlinear dimensionality reduction techniques, such as isometric feature map (Isomap),[14] local linear embedding (LLE),[15] sketch-map,[16] and diffusion maps,[17–19] are therefore required to obtain an effective low-dimensional embedding of the complex system dynamics.

In diffusion maps, proximity between configurations in the embedded space represents the ease of the dynamic evolution from one to the other.[17] Therefore, diffusion maps can help to characterize the underlying folding/unfolding pathways of proteins with a few relevant coordinates, under

a)Electronic mail: pdebene@princeton.edu

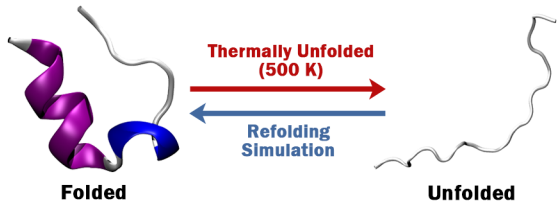

FIG. 1. Folded and unfolded Trp-cage miniproteins. The folded Trp-cage was thermally denatured at 500 K in order to create the unfolded configuration, from which the MD simulations of refolding were initiated. For the folded structure, the α-helix and 3_{10}-helix are colored in purple and blue, respectively. Images were rendered through Visual Molecular Dynamics (VMD).[49]

the assumptions that system dynamics can be described locally as diffusion processes and the short-time diffusive motions can be captured by structural similarities.[10,20] If such a description exists, a low-dimensional characterization of the folding process can be systematically obtained from diffusion maps, which can effectively describe the important conformational changes of proteins using only a small number of variables. For example, the conformational spaces of the β-hairpin,[11] β-sheet miniprotein,[21] src homology 3 domain,[22,23] promicrocin J25,[24] and n-alkane chains[10] have been successfully parametrized in only two or three dimensions using diffusion maps.

In this work, we study the folding mechanisms of the Trp-cage miniprotein using MD simulation and diffusion maps. Comprised of 20 residues, Trp-cage[25] is a designed miniprotein with a fast folding time of approximately 4 μs at room temperature.[26,27] The native state of Trp-cage, which is shown in Fig. 1, contains an α-helix (residues 2-8), a 3_{10}-helix (residues 11-14), and a polyproline II helix, which "cage" the hydrophobic tryptophan residue in the center of the protein. Due to its structural simplicity and rapid folding dynamics, Trp-cage has been extensively studied both experimentally[25,26,28–31] and computationally[32–43] in order to elucidate its folding kinetics and thermodynamics.

To the best of our knowledge, this is the first study of folding pathways of a helical protein utilizing diffusion maps with unbiased MD simulations. We make explicit comparisons to experimental and simulation studies in order to validate the approach. From the diffusion maps embedding of folding trajectories collected from unconstrained MD simulation, we successfully identified two clear folding pathways for Trp-cage. Furthermore, the embedding coordinates were successfully correlated with order parameters commonly used to describe protein folding, yielding physical interpretations of the folding pathways and intermediates. The identified pathways correspond to the two widely accepted mechanisms of protein folding: nucleation-condensation[44,45] and diffusion-collision.[46] In the folding pathway that follows the nucleation-condensation theory, a hydrophobic collapse occurs first, followed by a simultaneous formation of secondary structures and tertiary contacts. In the alternative diffusion-collision mechanism, the α-helix comprised of residues 2-8 is formed first, followed by packing of the rest of the residues to form the native structure. The pathways and folding intermediates identified using diffusion maps are consistent with the findings of other experimental[28,29,47] and simulation[35,42] studies.

The paper is structured as follows. In Sec. II, we describe the methodological details of the MD simulations and the diffusion maps technique. In Sec. III, we present the diffusion maps embedding of the Trp-cage folding simulation, along with the identification of folding pathways via a free energy surface. Interpretations of the folding mechanism are illustrated in terms of physical order parameters. We end with a summary of our work in Sec. IV.

II. METHODS

A. Molecular dynamics simulation

A folded Trp-cage, whose NMR structure was taken from the RCSB protein data bank (PDB ID: 1L2Y),[25] was solvated with 2910 molecules of the TIP4P/2005 model of water[48] in a cubic box with side length of 4.466 11 nm. The net charge of solvated Trp-cage is approximately $+1e$ at neutral pH, and so one water molecule was randomly replaced by a chloride ion. The Trp-cage was then thermally denatured through 100 ns of NVT MD simulation at 500 K. The folded and unfolded configurations are shown in Fig. 1. Starting with the unfolded configuration, we performed 25 folding simulations. Before each simulation, a 500 ps NVT MD run was performed at 500 K with the initial velocities randomly assigned (both in magnitude and direction) to each atom according to Maxwell distribution, in order to randomize the initial unfolded configuration. The system was then equilibrated by a brief 500 ps NVT MD simulation at 300 K, followed by a 500 ps NPT MD stage at 300 K and 1 bar. We then performed a NPT MD simulation at 300 K and 1 bar until the Trp-cage folded, saving configurations every 50 ps. The Trp-cage was considered folded if the RMSD of alpha carbons (C-α RMSD) with respect to the folded structure was less than 0.22 nm.[36,38,42]

The MD simulations were performed using the GROMACS[50–53] simulation package. Equations of motion were integrated using the leap-frog algorithm with a time step of 2 fs. Temperature was maintained at 300 K and 500 K during folding and unfolding simulations, respectively, using the Nosé-Hoover thermostat[54,55] with a time constant of 0.2 ps. Isotropic pressure coupling was applied using the Parrinello-Rahman barostat[56,57] with a 2 ps time constant to maintain the pressure at 1 bar. The short-range interactions were smoothly truncated at 1 nm, and the standard long-range dispersion corrections were applied for the energy and pressure.[58] The smooth-particle mesh Ewald method[59] was used to compute the reciprocal-part of the Ewald sum for the long-range electrostatic interactions. All bonds in the protein and water molecules were constrained using the linear constraint solver (LINCS) algorithm[60,61] and SETTLE,[62] respectively. Proteins were modeled using the modified version (Amber ff03w)[63] of the Amber ff03 force field.[64] The accuracy of this protein force field with the TIP4P/2005 water model[48] has been validated by other simulation studies on miniproteins.[63,65–67]

In order to characterize intermediate states and study their conformational densities, free energy surfaces were constructed using

$$\beta G = -\ln(p) + C, \tag{1}$$

where $\beta = 1/k_B T$, k_B is Boltzmann's constant, T is the temperature, C is a normalization constant, G is the Gibbs free energy, and p is a histogram approximation to the configurational density. Since we stopped each simulation shortly after the Trp-cage folded to its native structure, the configurations with fully folded Trp-cages were not sampled sufficiently to show a true free energy surface of Trp-cage at 300 K. Nonetheless, our "pseudo" free energy surfaces well serve the purpose of identifying the important folding intermediates.

B. Diffusion maps

Diffusion maps,[17,18] a nonlinear dimensionality reduction technique, was used to analyze the simulation data. Although the simulation data lie in a high-dimensional ambient space, the working assumption is that data (approximately) lie on a much lower-dimensional manifold. Diffusion maps uncovers a parametrization of this manifold (when it in fact exists), and projecting the data onto this low-dimensional subspace highlights patterns and trends that might be difficult to extract and visualize in the high-dimensional ambient space.

Let $x_1, \ldots, x_n \in \mathbb{R}^{m\times 3}$ denote the simulation data, where n is the number of sampled configurations, and m is the number of atoms (in this work, we only used α-carbons for the diffusion maps analysis) in the protein; x_i is the ith sampled configuration, where the jth row of x_i contains the x-, y-, and z-position of atom j in configuration i. The distance between two configurations, $d_{RMSD}(x_i, x_j)$, was defined as the root-mean-square deviation between configurations x_i and x_j *after* they had been optimally aligned with respect to translations and rotations. Translational degrees of freedom were removed by mean-centering both configurations, and rotational degrees of freedom were removed by aligning the two configurations using the Kabsch algorithm.[68,69] To compute the diffusion maps embedding of the data, the symmetric matrix $W \in \mathbb{R}^{n\times n}$, with

$$W_{ij} = \exp\left(-\frac{d^2_{RMSD}(x_i, x_j)}{\epsilon^2}\right), \tag{2}$$

is constructed, where ϵ is a characteristic distance in the data set. This allows us to retain only the short pairwise distances based on ϵ, while discarding the large distances that do not provide meaningful information regarding the dynamical proximity of data. According to Coifman *et al.*,[70] the range of suitable ϵ values lies in the linear region of a $\log(\sum_{i,j} W_{ij})$ vs. $\log(\epsilon^2/2)$ plot. For the results presented here, we used the median pairwise distance for the value of ϵ (1.89 nm), which falls in the appropriate range. The diagonal matrix $D \in \mathbb{R}^{n\times n}$ and the symmetric matrix $\tilde{W} \in \mathbb{R}^{n\times n}$, with

$$D_{ii} = \sum_{j=1}^{n} W_{ij} \tag{3}$$

$$\tilde{W} = D^{-1} W D^{-1} \tag{4}$$

are calculated. The matrix $\tilde{W}$ is analogous to W, but multiplication by D^{-1} makes the resulting parametrization invariant to changes in the sampling density (this corresponds to the $\alpha = 1$ density normalization in Ref. 17). The diagonal matrix $\tilde{D} \in \mathbb{R}^{n\times n}$ and the matrix $A \in \mathbb{R}^{n\times n}$, with

$$\tilde{D}_{ii} = \sum_{j=1}^{n} \tilde{W}_{ij} \tag{5}$$

$$A = \tilde{D}^{-1} \tilde{W} \tag{6}$$

are also computed. The eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_{n-1}$ and eigenvectors $\phi_0, \phi_1, \ldots, \phi_{n-1}$ of A are calculated and ordered such that $|\lambda_0| \geq |\lambda_1| \geq \cdots \geq |\lambda_{n-1}|$ (because A is similar to the symmetric matrix $\tilde{D}^{-1/2} \tilde{W} \tilde{D}^{-1/2}$, it is guaranteed to have real eigenvalues and real, orthogonal eigenvectors). The Armadillo C++ linear algebra library[71] was used for the computation of leading eigenvalues and eigenvectors. Because A is a row-stochastic matrix, the first eigenvector, ϕ_0, is always a constant eigenvector corresponding to the trivial $\lambda_0 = 1$ eigenvalue. The subsequent eigenvectors $\phi_1, \phi_2, \ldots, \phi_{n-1}$ provide coordinates to parametrize the underlying manifold, such that $\phi_j(i)$, the ith entry of ϕ_j, gives the jth embedding coordinate for configuration x_i. The eigenvalue λ_j provides a measure of the "importance" of coordinate ϕ_j. It is not uncommon to observe a "spectral gap" in the eigenvalue spectrum, such that $|\lambda_0| \geq |\lambda_1| \geq \cdots \geq |\lambda_k| \gg |\lambda_{k+1}| \geq \cdots \geq |\lambda_{n-1}|$. In this case, the dimensionality of the manifold is at most k, and the coordinates $\phi_1, \ldots, \phi_k$ provide a lower-dimensional description of the data. One must check for nonlinear correlations among these k coordinates, as it is possible that several eigenvectors are harmonics of each other and parametrize/describe the same direction in the data (see Ferguson *et al.*[10] for a more detailed discussion).

III. RESULTS AND DISCUSSION

A. Simulation results

All 25 Trp-cage folding simulations successfully folded (C-α RMSD < 0.22 nm) with an average folding time of 3.73 μs, which is in good agreement with the experimentally determined folding times of 3.56 μs at 300 K (inferred from published graph of Qiu *et al.*[26] by Byrne *et al.*[72]) and 3.7 μs at 298 K[27]. Fig. 2 shows free energy surfaces constructed from the Trp-cage folding trajectories, as functions of C-α RMSD, α-helix RMSD, and Rg. The α-helix RMSD is defined as the C-α RMSD of α-helix forming residues 2-8 with respect to an ideal α-helix, calculated using g_helix program of the GROMACS[50–53] simulation package. "F" denotes the folded state. Three potential intermediate states, labeled as "I," "J," and "K," are identified, but information obtained from the free energy surfaces using the physical order parameters is not sufficient to effectively characterize how Trp-cage folds or how the intermediates are related to each other mechanistically or dynamically.

B. Folding pathways

We then computed the diffusion maps embedding using a total of 46 398 configurations, sampled at an interval of 2 ns from the folding trajectories. We chose the sample size to be large enough to capture the folding pathways, within the limitations in computer memory needed for the analysis. Fig. 3

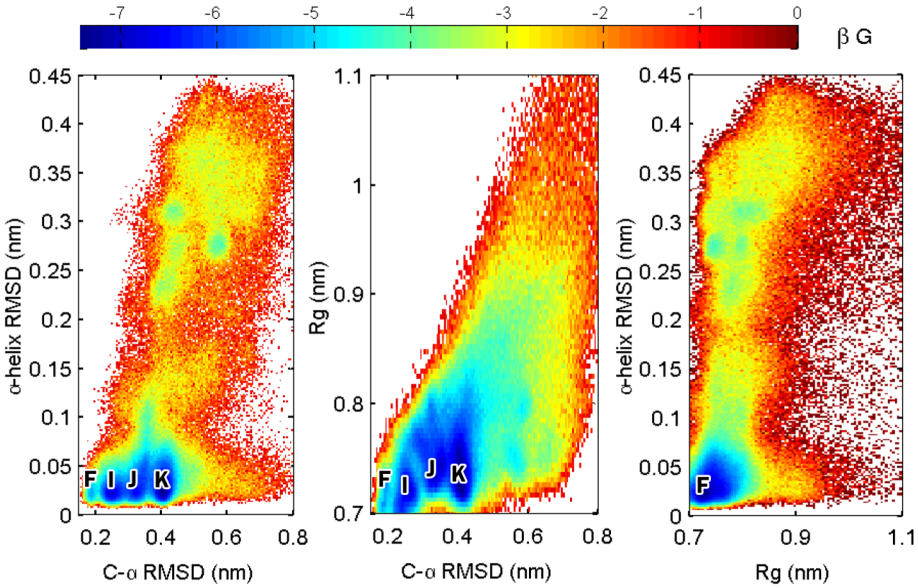

FIG. 2. Free energy surfaces of Trp-cage folding as functions of C-α RMSD, α-helix RMSD, and Rg. For each plot, the basin corresponding to the folded structure is labeled "F" and three intermediates are labeled as "I," "J," and "K" (no intermediate could be detected using α-helix RMSD and Rg as order parameters). Note that the folded basin is relatively small and has a low conformational density because the folding simulations were stopped shortly after the Trp-cage was successfully folded to its native structure.

shows the diffusion maps eigenvalue spectrum. Excluding the trivial eigenvalue ($\lambda_0 = 1$), large "spectral gaps" exist between λ_1, λ_2, and λ_3, after which the eigenvalue spectrum plateaus. Therefore, it was determined that eigenvectors ϕ_1 and ϕ_2 provide an adequate description of the low-dimensional manifold and confirm that they are not correlated.

Fig. 4 shows the two-dimensional diffusion maps embedding of the Trp-cage folding trajectories, colored by the C-α RMSD. Due to the high density of points, the C-α RMSD values were locally averaged by dividing the embedding space into a grid of 125×125 cells and computing the average C-α RMSD within each cell. For visual clarity, only cells containing two or more data points are displayed. There exists a wide region of configurations with C-α RMSD values similar to that of the folded structure, which corresponds to intermediates "I" and "J" found in Fig. 2. These configurations are nearly superimposable with the folded structure but lack the 3_{10}-helix, whose instability around 300 K has also been observed in the previous studies.[27,36,41] In diffusion maps, the top (nontrivial) eigenvector correlates with an embedding of the slowest dynamical motion.[20,73] It can be seen that ϕ_1 parameterizes the progression of folding process, since there is a gradual decrease in C-α RMSD with decreasing ϕ_1.

From Fig. 4, we can conclude that the folding of Trp-cage does not exhibit a simple two-state behavior. Rather, ϕ_2 distinguishes two main folding pathways. Fig. 5(a) shows the free energy surface over the diffusion maps embedding, from which we can identify two major folding pathways for Trp-cage. Along pathway A, Trp-cage begins folding with a hydrophobic collapse, followed by local rearrangements to form native contacts (nucleation-condensation). In contrast, the α-helix forms first along pathway B, which is then followed by correct packing of the 3_{10}-helix and polyproline II helix (diffusion-collision). This early formation of the α-helix has been reported in several studies.[36,72,74] These two Trp-cage folding pathways are in agreement with the previous simulation studies,[35,42] where similar folding mechanisms were reported. Also shown in Fig. 5 are the contracted free energies,

$$\beta G(\phi_i) = -\ln \int \exp\left[-\beta G(\phi_i, \phi_j)\right] \, d\phi_j. \tag{7}$$

Based on the free energy surface, intermediate structures were identified for each pathway. In pathway A, LOOP-I (corresponding to intermediate "K" in Fig. 2) is an intermediate

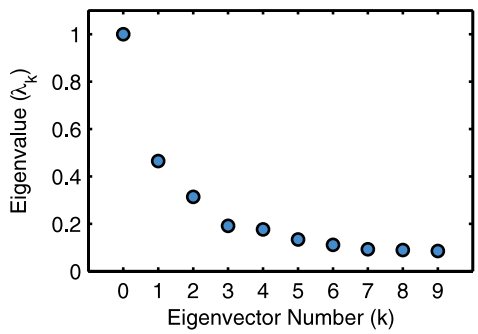

FIG. 3. First 10 diffusion maps eigenvalues. The first eigenvalue $\lambda_0 = 1$ is trivial.

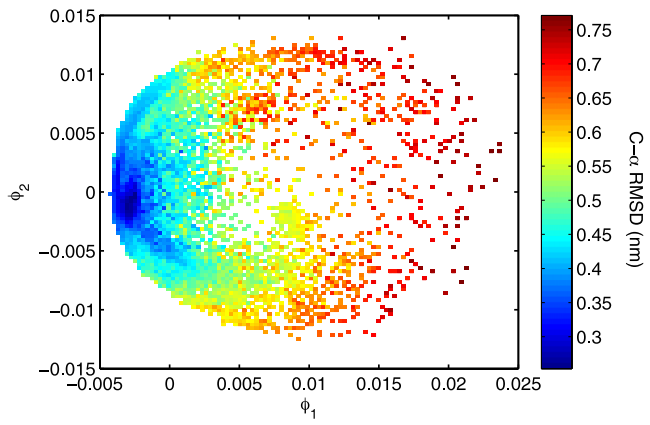

FIG. 4. Two-dimensional diffusion map embedding of Trp-cage folding trajectories, using the first two nontrivial eigenvectors (ϕ_1 and ϕ_2). The coloring is based on locally averaged C-α RMSD values with respect to the folded configuration.

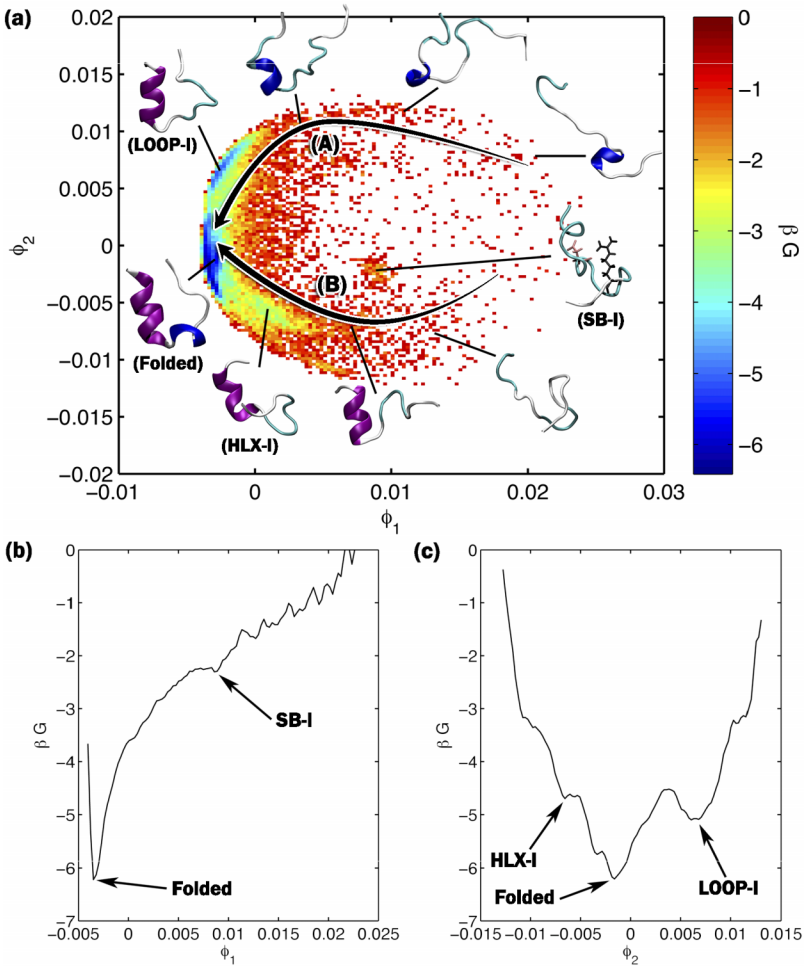

FIG. 5. Free energy surface on the two-dimensional diffusion map embedding with ϕ_1 and ϕ_2. Two folding pathways of the Trp-cage are highlighted: (A) tertiary contacts form first, followed by the main α-helix and (B) early formation of the α-helix followed by the condensation of the tertiary structure. Representative configurations are shown along each pathway. The free energy surface is shown in three different representations: (a) ϕ_2 vs. ϕ_1 with heat map, (b) βG vs. ϕ_1, and (c) βG vs. ϕ_2.

that has a very similar structure to the native Trp-cage; the only difference is the incorrect packing of the 3_{10}-helix (residues 11-14). Along folding pathway B, two intermediates are present, HLX-I and SB-I. Intermediate HLX-I contains a fully folded α-helix, but the rest of the chain is incorrectly packed. Both HLX-I and LOOP-I contain a folded α-helix, but in contrast with LOOP-I, HLX-I lacks the overall similarity of tertiary structure to the native Trp-cage. SB-I is an incorrectly folded structure that is stabilized by the formation of a salt bridge between residues Asp-9 and Arg-16. HLX-I and SB-I were not detected in Fig. 2. These intermediate structures have also been identified in other studies (Refs. 37 and 41 for LOOP-I, Refs. 35 and 36 for HLX-I, and Ref. 47 for SB-I).

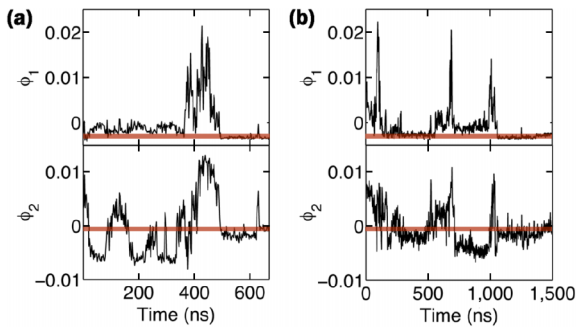

FIG. 6. The folding processes of two of the trajectories (each corresponding to (a) and (b)) are shown by the time evolution of the two diffusion maps coordinates (ϕ_1 and ϕ_2). The red lines indicate the ϕ_1 and ϕ_2 values corresponding to the folded Trp-cage configuration. Positive values of ϕ_2 generally correspond to the folding pathway A, while the negative correspond to the pathway B (see Fig. 5). As shown for both trajectories, the "switchings" of folding pathways can occur during a single folding event. For example, for trajectory (a), a switching in structure from intermediate HLX-I to LOOP-I occurs at approximately 100 ns. In addition, the Trp-cage can also unfold from an intermediate structure and start refolding, as seen at approximately 400 ns of (a) and 750 ns of (b). The high values of ϕ_1 at these time points signify the unfolding of the Trp-cage.

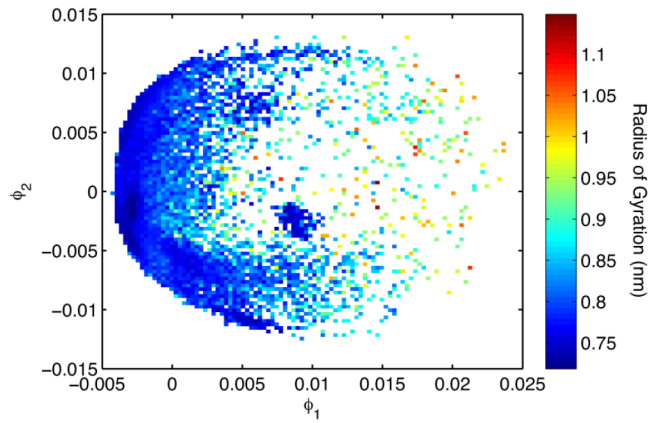

FIG. 7. Diffusion maps embedding colored by radius of gyration. As folding progresses, the radius of gyration becomes smaller as the Trp-cage residues become more closely packed. The folded structure and nearby intermediates are not distinguishable by radius of gyration.

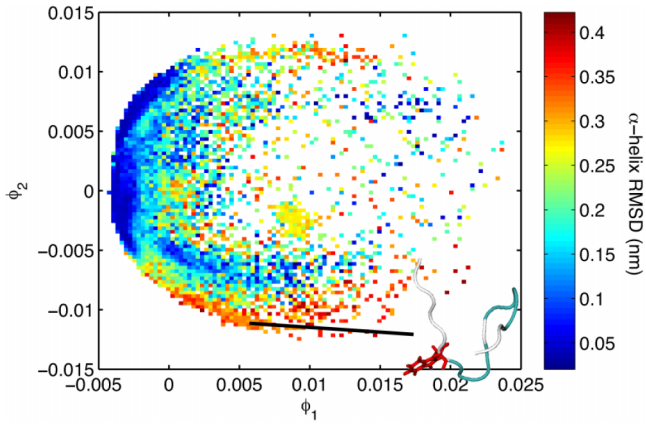

FIG. 8. Diffusion maps embedding colored by α-helix RMSD with respect to an ideal α-helix. A representative configuration of the concentrated region near folding pathway B with high values of α-helix RMSD is shown. The Trp-6 residue is explicitly drawn in red to highlight its complete exposure to the solvent.

The two folding pathways and intermediates we identified are generally consistent with the simulations of Juraszek and Bolhuis,[35] and Deng *et al.*[42] However, a discrepancy exists between our LOOP-I intermediate and the intermediate structure found by Juraszek and Bolhuis,[35] which contains no α-helix but correct tertiary contacts. In addition, Juraszek and Bolhuis found the initial collapse of tertiary structure (pathway A) to be a dominant folding pathway over the one with early α-helix formation (pathway B), while the opposite was observed in this work and other studies (Paschek *et al.*[36] and Deng *et al.*[42]). This is most likely due to the different force fields used, as suggested by Paschek *et al.*[36] Findings of notable dependencies of stability and intermediate structures of Trp-cage on the choice of protein force field[41] and water model[75] support this explanation. Juraszek and Bolhuis[35] used the OPLSAA protein force field with the SPC water model, while Amber ff94 with TIP3P, CHARMM22 with TIP3P, and Amber ff03w with TIP4P/2005 were used in the works of Paschek *et al.*,[36] Deng *et al.*,[42] and ours, respectively. Despite these discrepancies, the folding mechanisms found are in good general agreement with each other.

Consistent with the study of Juraszek and Bolhuis,[35] switching from one folding pathway to the other was observed. The switching occurred primarily between the HLX-I and LOOP-I structures, as both intermediates require correct rearrangement of the non-α-helix regions in order to fold eventually to the native structure. While rearranging, one intermediate can transition to the other instead of directly folding to the native structure. Furthermore, we observed that Trp-cage can also unfold from an intermediate and then refold, rearranging its structure as necessary. An example of this behavior is shown in Fig. 6; similar behavior has also been observed in another all-atom MD simulation.[76]

C. Physical interpretations of the folding mechanisms

Although the low-dimensional embedding can capture the folding progression and pathways well, the diffusion maps coordinates (ϕ_1 and ϕ_2) do not have physical meanings by themselves. In order to understand the mechanisms and intermediates of Trp-cage folding, we therefore correlated the diffusion maps coordinates with a set of physical parameters commonly used to describe protein folding. This was done by computing the local average order parameter in the diffusion maps embedding, as illustrated in Fig. 4.

Fig. 7 shows the diffusion maps embedding colored by the radius of gyration. As Trp-cage folds to the native state, the radius of gyration decreases as the stretched chain of residues packs into the globular, folded structure. Despite this correlation, the radius of gyration is inadequate for distinguishing the folded configuration from many of the intermediate states shown in Fig. 5. This suggests that the intermediates are incorrectly folded structures with a similar degree of packing as the folded structure.

One key feature that distinguishes two folding pathways of Trp-cage is that, for pathway A, the α-helix does not form until the LOOP-I intermediate is reached. Fig. 8 shows the diffusion maps embedding colored by the α-helix RMSD. The configurations along pathway A have higher values of α-helix RMSD compared to the ones along pathway B. As described in Sec. III B, the intermediates LOOP-I and HLX-I

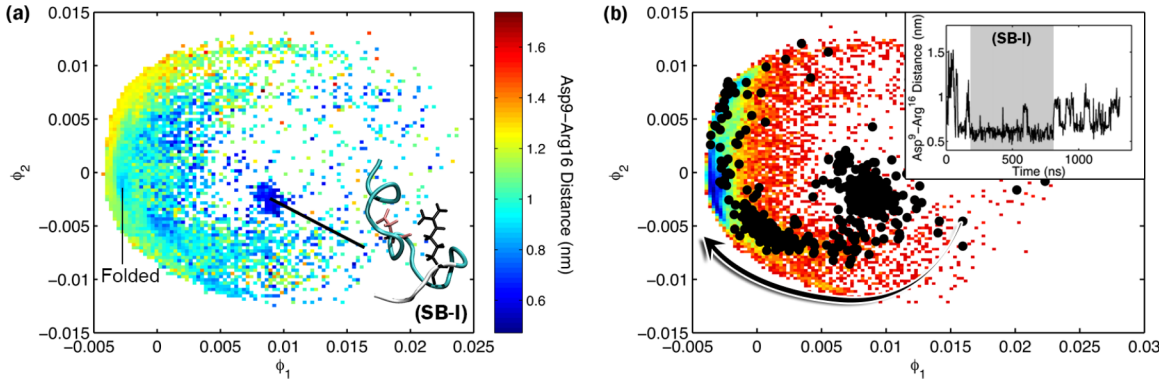

FIG. 9. (a) Diffusion maps embedding colored by the distance between residues Asp-9 and Arg-16 that form salt bridges. A representative configuration of a metastable intermediate is depicted, where the Asp-9 (red) and Arg-16 (black) residues are explicitly shown. (b) Configurations explored by one of the trajectories overlaid with black dots on the free energy surface from Fig. 5. Arrow indicates the direction of folding from the SB-I intermediate. ((b)-inset) The Asp-9—Arg-16 distance of the same folding trajectory as a function of time. The portion of trajectory in the SB-I intermediate structure is colored gray.

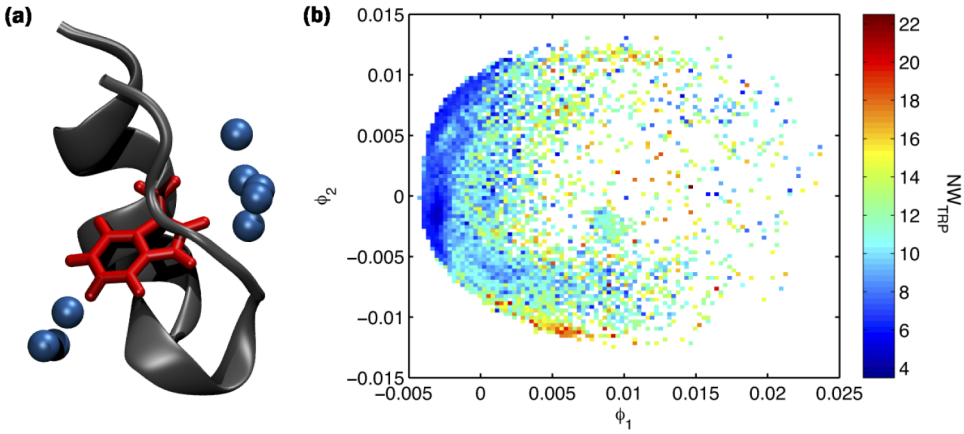

FIG. 10. (a) A graphical representation of the burial of the Trp-6 residue away from water molecules. The side chain of the Trp-6 residue is shown in red, and water molecules within 0.4 nm from the Trp-6 residue are shown as blue spheres. (b) Diffusion maps embedding colored by the number of water molecules within 0.4 nm from the Trp-6 residue (NW_{TRP}).

contain fully formed α-helices, as their α-helix RMSD values are similar to that of the folded Trp-cage.

For pathway B, a region of high α-helix RMSD values exist ($\phi_2 \approx -0.01$). The configurations in this region contain Trp-6 residues which are completely exposed to the solvent. Formation of the α-helix occurs as the hydrophobic Trp-6 residue rearranges to point to the rest of the Trp-cage chain, following the pathway B from there on.

Two oppositely charged residues, Asp-9 and Arg-16, are known to form a salt bridge due to an ionic attraction, and salt bridges are known to stabilize the folded structure of proteins.[77] Fig. 9(a) shows the diffusion maps embedding colored by the distance between the oxygen atom in Asp-9 and the nitrogen atom in Arg-16. The salt-bridge-stabilized intermediate SB-I near pathway B is rather turn-rich than helical (Fig. 9(a)), which is consistent with a salt-bridge-stabilized intermediate found experimentally.[47] For this intermediate to correctly fold, the salt bridge must first be broken and then reform to stabilize the folded structure. The large Asp-6—Arg-16 distance of LOOP-I intermediate suggests that LOOP-I is not stabilized by the salt bridge.

Fig. 9(b) shows configurations visited by the Trp-cage for one of the folding trajectories, along with the time evolution of the Asp-9—Arg-16 distance. After an early folding to LOOP-I structure and unfolding (200 ns), the Trp-cage rapidly folds to SB-I (as shown in Fig. 6). The Trp-cage is then locally trapped in the SB-I configuration for approximately half of the total folding time for that trajectory (1.31 μs), which is short compared to the average folding time of Trp-cage (3.73 μs) obtained from our simulations. This is consistent with the previous findings that the early formation of the salt bridge can either expedite[32] or impede[33,78] the Trp-cage folding kinetics.

One key feature that stabilizes the folded structure of Trp-cage is the "burial" of the hydrophobic Trp-6 residue. When the Trp-cage is folded, the Trp-6 residue is "caged" from water molecules by the surrounding residues, as shown in Fig. 10(a). In order to form a stable hydrophobic core, the Trp-6 residue must be correctly buried inside with expulsion of water from it. Fig. 10(b) shows the diffusion maps embedding colored by the number of water molecules within 0.4 nm from any of the atoms in the Trp-6 residue. The Trp-6 residues in structures along both pathways are highly exposed to water until the intermediate structures LOOP-I and HLX-I are formed, which suggests that the hydrophobic packing contributes to the stabilization of those intermediates. This is consistent with the experimental[29] and computational[35,76] findings, where the expulsion of water from the hydrophobic core happens near the end of the folding process.

IV. CONCLUSIONS

We have demonstrated how diffusion maps can systematically extract important order parameters that can characterize protein folding pathways. We used the Trp-cage miniprotein as our model system and applied diffusion maps to folding trajectories generated by all-atom MD with explicit solvent. To the best of our knowledge, this is the first study where the folding pathways of a helical protein are uncovered by applying diffusion maps to unbiased MD trajectories. Phenomenological order parameters were correlated with the diffusion maps embedding, in order to provide physical interpretations of the folding mechanisms and intermediates. Through two-dimensional embedding of the trajectories, we identified two distinct folding pathways. These pathways are consistent with two widely accepted mechanisms of protein folding, nucleation-condensation[44,45] (pathway A) and diffusion-collision[46] (pathway B). We observed that switching of the pathways also occurs within a single trajectory, which indicates that Trp-cage can fold via multiple folding mechanisms during a single folding event.

Diffusion maps embedding allows clear visualization of the dynamic evolution of Trp-cage folding, uncovering mechanisms that are consistent with experimental[28,29,47] and computational findings.[35,42] We note that some discrepancies exist among computational studies due to differences in force fields employed,[35,36,41,42,75] but the general folding mechanisms are in agreement. The main advantage of diffusion maps is the capability of low-dimensional characterization of the underlying dynamics without any prior knowledge of appropriate physical order parameters that describe the folding pathway. The simulation studies of Trp-cage folding pathways by Juraszek and Bolhuis,[35] and Deng *et al.*[42] utilized transition path sampling[6] and a Markov state model, respectively. Transition path sampling needs pre-determined order parameters that can properly define the folded and unfolded states, while a Markov state model requires selection of a number of states *a priori* with protein configurations

assigned to corresponding states. Additionally, diffusion maps provides us with a more parsimonious description of our data; for the study presented here, no pair of physical order parameters captures all the features highlighted in the two-dimensional diffusion maps embedding. Diffusion maps alone, however, is not suitable for a systematic extraction of kinetic information. Therefore, one possible avenue for future study is to combine diffusion maps with a Markov state model for a more robust characterization of both mechanisms and kinetics of protein folding, as demonstrated by Nedialkova *et al.*[79] using alanine pentapeptide.

In this work, we used traditional MD simulations to obtain the folding trajectories. However, MD simulations on microsecond time scales (or longer) can be computationally expensive. To circumvent this issue, diffusion maps can be combined with advanced sampling techniques, such as umbrella sampling[80] (successfully implemented with diffusion maps by Ferguson *et al.*[81]), transition path sampling,[6] forward flux sampling,[9] diffusion-map-directed MD,[82] and coarse reverse integration.[83] In addition, clustering in diffusion maps space can enhance the extraction of meaningful folding intermediates, especially for high-dimensional embedding where the construction and visualization of a free energy surface can be difficult.[79] With these attractive features in mind, diffusion maps can be employed in potentially interesting studies of Trp-cage folding in the presence of cosolvents or with mutated residues, in order to elucidate and visualize the effect of these perturbations on folding mechanisms.

ACKNOWLEDGMENTS

P.G.D. gratefully acknowledges financial support from the National Science Foundation (Grant Nos. CBET-1263565 and CHE-1213343). C.J.D. acknowledges support from the Department of Energy Computational Science Graduate Fellowship (Grant No. DE-FG02-97ER25308) and the National Science Foundation Graduate Research Fellowship (Grant No. DGE 1148900). I.G.K thanks the National Science Foundation (Grant No. CS&E-1310173) for its support. The computations were performed at the Terascale Infrastructure for Groundbreaking Research in Engineering and Science (TIGRESS), at Princeton University.

[1]C. Soto, Nat. Rev. Neurosci. **4**, 49 (2003).
[2]T. R. Sosnick and D. Barrick, Curr. Opin. Struct. Biol. **21**, 12 (2011).
[3]D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang, ACM SIGARCH Comput. Archit. News **35**, 1 (2007).
[4]D. Frenkel and B. Smit, *Understanding Molecular Simulations* (Academic, London, 2002).
[5]Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).
[6]C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
[7]M. R. Sørensen and A. F. Voter, J. Chem. Phys. **112**, 9599 (2000).
[8]A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).
[9]R. J. Allen, P. B. Warren, and P. R. Ten Wolde, Phys. Rev. Lett. **94**, 018104 (2005).
[10]A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, Proc. Natl. Acad. Sci. U.S.A. **107**, 13597 (2010).
[11]M. Duan, J. Fan, M. Li, L. Han, and S. Huo, J. Chem. Theory Comput. **9**, 2490 (2013).
[12]I. Jolliffe, *Principal Component Analysis* (Wiley Online Library, 2005).
[13]P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi, Proc. Natl. Acad. Sci. U.S.A. **103**, 9885 (2006).
[14]J. B. Tenenbaum, V. De Silva, and J. C. Langford, Science **290**, 2319 (2000).
[15]S. T. Roweis and L. K. Saul, Science **290**, 2323 (2000).
[16]M. Ceriotti, G. A. Tribello, and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **108**, 13023 (2011).
[17]R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, Proc. Natl. Acad. Sci. U.S.A. **102**, 7426 (2005).
[18]R. R. Coifman and S. Lafon, Appl. Comput. Harmonic Anal. **21**, 5 (2006).
[19]M. Belkin and P. Niyogi, Neural Comput. **15**, 1373 (2003).
[20]A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti, Chem. Phys. Lett. **509**, 1 (2011).
[21]W. Zheng, B. Qi, M. A. Rohrdanz, A. Caflisch, A. R. Dinner, and C. Clementi, J. Phys. Chem. B **115**, 13065 (2011).
[22]P. Das, T. A. Frewen, I. G. Kevrekidis, and C. Clementi, *Coping with Complexity: Model Reduction and Data Analysis* (Springer, 2011), pp. 113–131.
[23]M. A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi, J. Chem. Phys. **134**, 124116 (2011).
[24]A. L. Ferguson, S. Zhang, I. Dikiy, A. Z. Panagiotopoulos, P. G. Debenedetti, and A. J. Link, Biophys. J. **99**, 3056 (2010).
[25]J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, Nat. Struct. Biol. **9**, 425 (2002).
[26]L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, J. Am. Chem. Soc. **124**, 12952 (2002).
[27]R. M. Culik, A. L. Serrano, M. R. Bunagan, and F. Gai, Angew. Chem. **123**, 11076 (2011).
[28]Z. Ahmed, I. A. Beta, A. V. Mikhonin, and S. A. Asher, J. Am. Chem. Soc. **127**, 10943 (2005).
[29]H. Neuweiler, S. Doose, and M. Sauer, Proc. Natl. Acad. Sci. U.S.A. **102**, 16650 (2005).
[30]K. H. Mok, L. T. Kuhn, M. Goez, I. J. Day, J. C. Lin, N. H. Andersen, and P. J. Hore, Nature **447**, 106 (2007).
[31]A. Halabis, W. Zmudzinska, A. Liwo, and S. Oldziej, J. Phys. Chem. B **116**, 6898 (2012).
[32]C. D. Snow, B. Zagrovic, and V. S. Pande, J. Am. Chem. Soc. **124**, 14548 (2002).
[33]R. Zhou, Proc. Natl. Acad. Sci. U.S.A. **100**, 13280 (2003).
[34]J. Pitera and W. Swope, Proc. Natl. Acad. Sci. U.S.A. **100**, 7587 (2003).
[35]J. Juraszek and P. G. Bolhuis, Proc. Natl. Acad. Sci. U.S.A. **103**, 15859 (2006).
[36]D. Paschek, S. Hempel, and A. E. García, Proc. Natl. Acad. Sci. U.S.A. **105**, 17754 (2008).
[37]S. Kannan and M. Zacharias, Proteins: Struct., Funct., Bioinf. **76**, 448 (2009).
[38]R. Day, D. Paschek, and A. E. Garcia, Proteins: Struct., Funct., Bioinf. **78**, 1889 (2010).
[39]C. Velez-Vega, E. E. Borrero, and F. A. Escobedo, J. Chem. Phys. **133**, 105103 (2010).
[40]W. Zheng, E. Gallicchio, N. Deng, M. Andrec, and R. M. Levy, J. Phys. Chem. B **115**, 1512 (2011).
[41]Q. Shao, J. Shi, and W. Zhu, J. Chem. Phys. **137**, 125103 (2012).
[42]N. Deng, W. Dai, and R. Levy, J. Phys. Chem. B **117**, 12787 (2013).
[43]H. W. Hatch, F. H. Stillinger, and P. G. Debenedetti, J. Phys. Chem. B **118**, 7761 (2014).
[44]V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10026 (1994).
[45]A. R. Fersht, Curr. Opin. Struct. Biol. **7**, 3 (1997).
[46]M. Karplus and D. L. Weaver, Nature **260**, 404 (1976).
[47]P. Rovó, V. Farkas, O. Hegyi, O. Szolomájer-Csikós, G. K. Tóth, and A. Perczel, J. Pept. Sci. **17**, 610 (2011).
[48]J. L. F. Abascal and C. Vega, J. Chem. Phys. **123**, 234505 (2005).
[49]W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[50]B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[51]D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, J. Comput. Chem. **26**, 1701 (2005).
[52]E. Lindahl, B. Hess, and D. van der Spoel, J. Mol. Model. **7**, 306 (2001).
[53]H. J. Berendsen, D. van der Spoel, and R. van Drunen, Comput. Phys. Commun. **91**, 43 (1995).

[54]S. Nosé, J. Chem. Phys. **81**, 511 (1984).
[55]W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).
[56]M. Parrinello and A. Rahman, J. Appl. Phys. **52**, 7182 (1981).
[57]S. Nosé and M. Klein, Mol. Phys. **50**, 1055 (1983).
[58]M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, USA, 1989).
[59]U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, J. Chem. Phys. **103**, 8577 (1995).
[60]B. Hess, H. Bekker, H. J. C. Berendsen, and J. Fraaije, J. Comput. Chem. **18**, 1463 (1997).
[61]B. Hess, J. Chem. Theory Comput. **4**, 116 (2008).
[62]S. Miyamoto and P. A. Kollman, J. Comput. Chem. **13**, 952 (1992).
[63]R. B. Best and J. Mittal, J. Phys. Chem. B **114**, 14916 (2010).
[64]Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. M. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. M. Wang, and P. Kollman, J. Comput. Chem. **24**, 1999 (2003).
[65]K. A. Beauchamp, Y.-S. Lin, R. Das, and V. S. Pande, J. Chem. Theory Comput. **8**, 1409 (2012).
[66]P. Kührová, A. De Simone, M. Otyepka, and R. B. Best, Biophys. J. **102**, 1897 (2012).
[67]R. B. Best, D. de Sancho, and J. Mittal, Biophys. J. **102**, 1462 (2012).
[68]W. Kabsch, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **32**, 922 (1976).
[69]W. Kabsch, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **34**, 827 (1978).
[70]R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, IEEE Trans. Image Process. **17**, 1891 (2008).
[71]C. Sanderson, "Armadillo: An open source c++ linear algebra library for fast prototyping and computationally intensive experiments," Technical Report (NICTA, Australia, 2010), http://espace.library.uq.edu.au/view/UQ:224609.
[72]A. Byrne, D. V. Williams, B. Barua, S. J. Hagen, B. L. Kier, and N. H. Andersen, Biochemistry **53**, 6011 (2014).
[73]M. A. Rohrdanz, W. Zheng, and C. Clementi, Annu. Rev. Phys. Chem. **64**, 295 (2013).
[74]H. Meuzelaar, K. A. Marino, A. Huerta-Viga, M. R. Panman, L. E. Smeenk, A. J. Kettelarij, J. H. van Maarseveen, P. Timmerman, P. G. Bolhuis, and S. Woutersen, J. Phys. Chem. B **117**, 11490 (2013).
[75]D. Paschek, R. Day, and A. E. García, Phys. Chem. Chem. Phys. **13**, 19840 (2011).
[76]S. Chowdhury, M. C. Lee, and Y. Duan, J. Phys. Chem. B **108**, 13855 (2004).
[77]D. E. Anderson, W. J. Becktel, and F. W. Dahlquist, Biochemistry **29**, 2403 (1990).
[78]Z. Hu, Y. Tang, H. Wang, X. Zhang, and M. Lei, Arch. Biochem. Biophys. **475**, 140 (2008).
[79]L. V. Nedialkova, M. A. Amat, I. G. Kevrekidis, and G. Hummer, J. Chem. Phys. **141**, 114102 (2014).
[80]G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).
[81]A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis, J. Chem. Phys. **134**, 135103 (2011).
[82]W. Zheng, M. A. Rohrdanz, and C. Clementi, J. Phys. Chem. B **117**, 12769 (2013).
[83]T. A. Frewen, G. Hummer, and I. G. Kevrekidis, J. Chem. Phys. **131**, 134104 (2009).