THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

First Passage Analysis of the Folding of a β-Sheet Miniprotein: Is it More Realistic Than the Standard Equilibrium Approach?

Igor V. Kalgin,[†] Sergei F. Chekmarev,*[,†,‡] and Martin Karplus*[,§,∥]

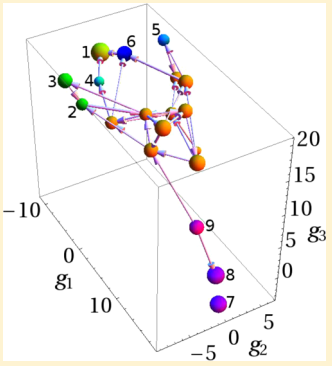
[†]Department of Physics, Novosibirsk State University, 630090 Novosibirsk, Russia

[‡]Institute of Thermophysics, SB RAS, 630090 Novosibirsk, Russia

[§]Laboratoire de Chimie Biophysique, ISIS Université de Strasbourg, 67000 Strasbourg, France

[∥]Department of Chemistry & Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, United States

ABSTRACT: Simulations of first-passage folding of the antiparallel β-sheet miniprotein beta3s, which has been intensively studied under equilibrium conditions by A. Caflisch and co-workers, show that the kinetics and dynamics are significantly different from those for equilibrium folding. Because the folding of a protein in a living system generally corresponds to the former (i.e., the folded protein is stable and unfolding is a rare event), the difference is of interest. In contrast to equilibrium folding, the Ch-curl conformations become very rare because they contain unfavorable parallel β-strand arrangements, which are difficult to form dynamically due to the distant N- and C-terminal strands. At the same time, the formation of helical conformations becomes much easier (particularly in the early stage of folding) due to short-range contacts. The hydrodynamic descriptions of the folding reaction have also revealed that while the equilibrium flow field presented a collection of local vortices with closed "streamlines", the first-passage folding is characterized by a pronounced overall flow from the unfolded states to the native state. The flows through the locally stable structures Cs-or and Ns-or, which are conformationally close to the native state, are negligible due to detailed balance established between these structures and the native state. Although there are significant differences in the general picture of the folding process from the equilibrium and first-passage folding simulations, some aspects of the two are in agreement. The rate of transitions between the clusters of characteristic protein conformations in both cases decreases approximately exponentially with the distance between the clusters in the hydrogen bond distance space of collective variables, and the folding time distribution in the first-passage segments of the equilibrium trajectory is in good agreement with that for the first-passage folding simulations.



1. INTRODUCTION

In computer simulation studies of protein folding, the folding reaction is most often considered under equilibrium conditions; i.e., one chooses a temperature at which both the unfolded and folded (native) states of the protein are populated (Shea and Brooks[1]). Under these conditions, provided that the simulated trajectory is sufficiently long, the protein experiences many folding/unfolding events. The results of the equilibrium simulations are typically organized in the form of a free energy surface, disconnectivity graph (Becker and Karplus[2]) or equilibrium kinetic network (Rao and Caflisch[3]), which describe the populations of the characteristic states of the system and the rates of transitions between them in the course of repeating folding and unfolding. To calculate the time evolution of the system through the network, the Markov process approximation is often employed (Krivov et al.,[4] Noé and Fischer,[5] and Lane et al.[6]).

Under the usual physiological conditions in the organism, the native state is stable and unfolding events are improbable. Then, the folding reaction corresponds essentially to "first-passage folding" (FPF), which can be studied with an ensemble of the trajectories that are initiated in the unfolded state of the protein (e.g., as the polypeptide comes of the ribosome) and are terminated when the native state is reached (Chekmarev et al.,[7,8] Palyanov et al.,[9] and Kalgin et al.[10,11]). This raises the question as to how the equilibrium folding/unfolding results are related to FPF. In such a comparison, it should be noted that often the environment conditions used will be different; e.g. equilibrium folding (EF) requires a higher temperature than the FPF, though of course, as we do here, it is possible to investigate FPF at the same temperature as the EF. So far, the FPF simulations have been limited to coarse-grained protein models. In an early 125-residue lattice protein model study (Dinner and Karplus[12]), the low-temperature folding pathways resembled the high-temperature unfolding pathways, but for the same temperature the pathways were different. A number of nonequilibrium folding experiments have been made, in which a reagent (e.g., GdmCl) stabilizing the unfolded state is rapidly diluted and folding (collapse) is observed by FRET (Lipman et al.[13]).

There have been a large number of studies of unfolding on the assumption that it is the inverse of folding.[14−16] Because unfolding is fast at the high temperature usually used, all-atom models in explicit solvent can be employed. Unfolding

Received: December 28, 2013
Revised: March 14, 2014
Published: March 26, 2014

   dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287−4299

simulations have been found to be most meaningful for proteins with two-state kinetics when the unfolded states and the native state are separated by a single free energy barrier,[14] though two-state kinetics can be observed even when there are multiple barriers.[17] If the folding kinetics are more complex, e.g., when a range of reaction channels are involved, unfolding is not necessarily the reverse of folding.[12]

To inquire into the relation between the FPF and the EF at a given (elevated) temperature, we examine the antiparallel β-sheet miniprotein (called beta3s, Figure 1), one of the few

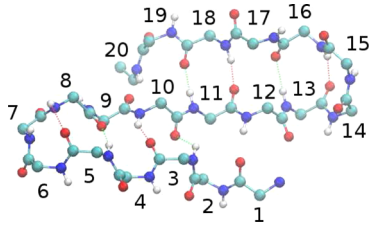


Figure 1. Native structure of beta3s. The lower part of the protein corresponds to the N-terminal hairpin, and the upper part to the C-terminal hairpin. The dashed lines indicate hydrogen bonds.

systems for which the folding reaction under equilibrium conditions has been studied in detail with an all-atom representation. The published studies are based on a set trajectories of total length of 20 μs, during which the protein experiences on the order of one hundred folding/unfolding events.[19−24] The simulations were performed using the CHARMM program[25] with an implicit solvent model. To have the denatured and native state of the protein both significantly populated, the temperature for the simulations was typically chosen to be T = 330 K, which is slightly above the melting temperature (Cavalli et al.[26]). The equilibrium folding of this system has been analyzed in many ways, which we do not review here; see ref 24 for a listing of some of the studies. In what follows we describe the first passage folding and compare it with the EF results obtained in our previous work.[24] Section 2 contains a brief survey of the methods we used to perform simulations and analyze the results. Section 3 describes the results and section 4 contains a concluding discussion.

2. METHODS

The method used to simulate and analyze folding of the beta3s miniprotein under first-passage folding (FPF) conditions is similar to that employed previously for the EF (Kalgin et al.[24]). Below, we give a brief survey of the method. The system used is the one studied by Caflisch and various co-workers.[18−23]

System and Molecular Dynamics Simulations. The designed three-stranded antiparallel 20-residue peptide (Thr1-Trp2-Ile3-Gln4-Asn5-Gly6-Ser7-Thr8-Lys9-Trp10 -Tyr11-Gln12-Asn13-Gly14-Ser15-Thr16-Lys17-Ile18-Tyr19-Thr20 with charged termini[27]), shown in Figure 1, was modeled with the CHARMM program.[25] All heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms were considered explicitly; PARAM19 force field (Neria et al.[28]) and a default cutoff of 7.5 Å for the nonbonding interactions were used. A mean field approximation based on the solvent-accessible surface (SAS) was employed to describe the main effects of the aqueous solvent (Ferrara et al.[29]). The simulations were performed with a time step of 2 fs using the Berendsen thermostat (coupling constant of 5 ps) at T = 330 K. For the present protein model, this temperature is slightly above the melting temperature.[26] Two hundred MD trajectories started in unfolded states of the protein and terminated upon reaching the native-like state were generated. The atomic coordinates ("frames") were saved every 20 ps.

The definition of the initial and final states deserves additional comments. It is expected that small proteins up to size of 10−15 kDa do not fold until they have left the ribosome (Fersht and Daggett[14]), because, as has been shown for barnase fragments and chymotrypsin inhibitor 2 (Neira and Fersht[30]), the last residues at the C terminus of the protein have to be free to allow folding. Consequently, the initial stage of folding is likely to be independent of interactions with the ribosome or with chaperones for many proteins. This circumstance is used to justify in vitro experimental studies of protein folding, where the initial states of the protein are prepared by thermal (temperature-jump experiments) or chemical (stopped-flow experiments) denaturation of the native state of a protein.[31−33] In the present study, we used the standard CHARMM[25] protocol to prepare initial conformations. More specifically, an extended conformation of the protein was first minimized (200 steps of the steepest descent followed by 300 steps of the conjugate gradient algorithm) and then heated to T = 330 K and equilibrated for 5×10^3 time steps. As will be shown below (section 3), the initial conformations thus obtained are similar to the most unfolded conformations found under equilibrium folding conditions starting with the native state at the temperature of interest (T = 330 K).

The final state where the FPF trajectory was terminated is the native state of the protein. Beta3s has a number of native-like conformations that differ not only by the hydrogen bond

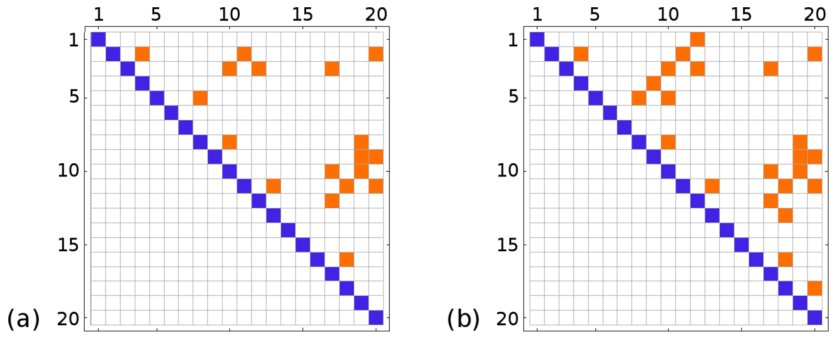


Figure 2. Contact maps for two different distances between the geometrical centers of the side chains d_{nat} that were used to determine native contacts. Panels a and b are for d_{nat} < 6.5 Å and d_{nat} < 7.5 Å, respectively.

dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287−4299

distances, which are used below to characterize the protein conformations, but also by orientation of the side chains. Though the hydrogen bond distances could be used to define the native contacts, we employed the side-chain distances. Specifically, two criteria were tested: one assumed that a native contact was formed if the distance between the geometrical centers of the side chains of two residues d_{nat} was less than 6.5 Å, and the other that d_{nat} < 7.5 Å. Excluding nearest neighbors (i.e., the pairs of residues for which $|j - i| = 1$ with i and j the residue numbers), the numbers of native contacts are $N_{nat} = 18$ at d_{nat} < 6.5 Å and $N_{nat} = 23$ at d_{nat} < 7.5 Å (Figure 2). In the latter case, five additional contacts appear, which are (1,12), (4,9), (5,10), (13,18), and (18,20) contacts. Among them, four contacts, i.e., (1,12), (4,9), (13,18), and (18,20), were listed in ref 18 as native contacts. Therefore, we considered d_{nat} < 7.5 Å to be the more suitable criterion, though the effect of the difference between the two is not large. With this definition of the final state to terminate the trajectory, the fraction of the hydrogen bonds in the final states was equal to 0.76 on average, i.e., approximately 6 hydrogen bonds among the bonds indicated in Figure 1 were present.

Conformation Space and Collective Variables. To characterize protein conformations, the hydrogen bond PCA (HB PCA) method[24] was used. In this method, the original conformation space of the protein in the form of the hydrogen bond distances is reduced to a three-dimensional space of collective variables $\mathbf{g} = (g_1, g_2, g_3)$ space with a specialized principal component analysis (PCA).[34] A distinctive feature of this method is that only the formed bonds are taken into account to make the folded states more pronounced. The first three modes corresponding to the largest eigenvalues were chosen as the variables g_1, g_2, and g_3. They account for 24% of the data variation (Figure 3). Because the collective variables

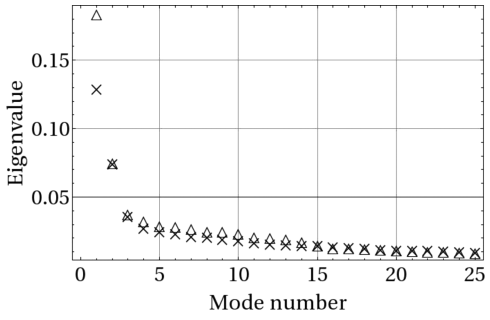


Figure 3. Spectrum of the largest eigenvalues. Triangles and crosses correspond to the equilibrium and first-passage folding, respectively. The eigenvalues are normalized so that their sum is equal to 1.

are linear combinations of the original variables, they are measured in the same units as the bond distances, i.e., in angstroms. Figure 3 also makes clear that the variables g_1, g_2, and g_3 are different from those for the equilibrium folding. This difference is due to the fact that the set of representative points from which they are calculated in the FPF is different from that for the EF.

Clustering the Conformations. To divide the representative points of the protein states in the $\mathbf{g} = (g_1, g_2, g_3)$ space into clusters, the MCLUST method by Fraley and Raftery[35] was used. In this method, the collection of points is approximated by a set of multidimensional (in our case 3D) Gaussian functions with generally different covariance matrices and different weights.

Secondary Structure Analysis. As in the previous studies,[3,20,21,24] protein conformations were discriminated according to the secondary structure strings (SSSs) encoded with the DSSP alphabet;[36] i.e., the letters H, G, I, E, B, T, S, and "-" stand for α-helix, 3_{10}-helix, π-helix, extended, isolated β-bridge, hydrogen bonded turn, bend, and unstructured segments, respectively. With this coding, the native state (Figure 1) is represented by the string "-EEEETTEEEEEET-TEEEE-".[3] The program WORDOM[37] was used to perform the analysis.

"Hydrodynamic" Description of the Folding Process. Using the first passage folding trajectories, the local probability flows (fluxes) of the transitions $\mathbf{j}(\mathbf{g})$ in the space of collective variables $\mathbf{g} = (g_1, g_2, g_3)$ are determined. They are calculated as the 2-fold (time and ensemble) averages of the local transitions. On the basis of these fluxes, the folding process is viewed as a steady flow of a folding "fluid" from the unfolded states to the native state, with the density of the fluid being proportional to the probability for the system to be found at the current point of the $\mathbf{g}$ space.[8,11] Having the fluxes $\mathbf{j}(\mathbf{g})$, the "streamlines" of the folding flows can be constructed, which are tangent to the local directions of the $\mathbf{j}(\mathbf{g})$ vectors.[38] In the case of two dimensions, e.g., for the projection of the folding flow onto the (g_1, g_2) plane, the streamlines can be calculated as the lines corresponding to constant values of the stream function,[8,24] and in the case of the three-dimensional space they are visualized with passive tracers (weightless point particles).[11,24]

3. COMPARISON OF FIRST PASSAGE FOLDING (FPF) AND EQUILIBRIUM FOLDING (EQ)

To study the first passage folding (FPF) of beta3s, two hundred folding trajectories were initiated at an extended state of the protein and terminated upon reaching a native-like state. Because of some looseness in determining the native contacts (section 2), the native-like state was considered to be reached if the number of native contacts was not less than 23, i.e., $N_{nat} - 1$. Specifically, the criterion d_{nat} < 7.5 Å was used to determine a native contact. The change of this criterion to a more "stiff" one (d_{nat} < 6.5 Å), which decreased the number of native contacts from 23 to 18, was found to have no significant effect. In particular, the first passage time distributions for the two criteria agree not only between themselves but also with the corresponding distribution obtained by Krivov et al.[21] (Figure 4). The temperature at which these simulations were performed was the same as in the equilibrium simulations, i.e., T = 330 K. The representative points were taken from these trajectories at 20 ps intervals, which resulted in the total number of points $\approx 1.2 \times 10^6$; i.e., the number of points is approximately equal to that for the EF studies (1×10^6).

Figure 5a presents the distribution of the representative points in the 3D space of the collective variables $\mathbf{g} = (g_1, g_2, g_3)$ obtained with the HB PCA method for EF, and Figure 5b shows the corresponding results for the FPF. The points are colored according to the clusters of characteristic conformations to which they belong. They are numbered in accord with Tables 1 and 2, respectively; the orange points without a number correspond to other clusters. The variables g_1, g_2, and g_3 in Figure 5a,b, as well as in similar figures below, are measured in angstroms. We note that these variables are different for the EF and FPF calculations because they are calculated from different collections of representative points. As in the EF,[24] if two points (1 and 2) in the $\mathbf{g}$ space are sufficiently distant, so that the protein conformations do not

dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287−4299

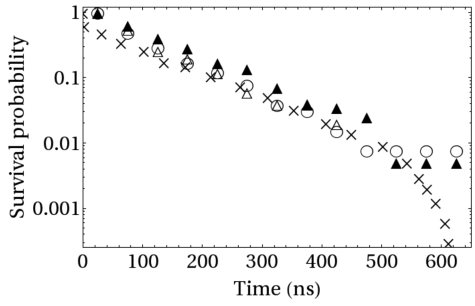


Figure 4. Survival probability distributions of the first passage time $F(t) = \int_t^\infty p(t)\,dt$, where $p(t)$ is the distribution of the first passage times. Empty and solid triangles are, respectively, for $d_{nat} < 6.5$ Å and $d_{nat} < 7.5$ Å in the present work, and the crosses present the distribution of ref 21. The number of trajectories for $d_{nat} < 6.5$ Å (the empty triangles) was 4 times smaller than for $d_{nat} < 7.5$ Å (the solid triangles). The circles show the distribution corresponding to the first-passage folding segments of the EF trajectory of ref 24.

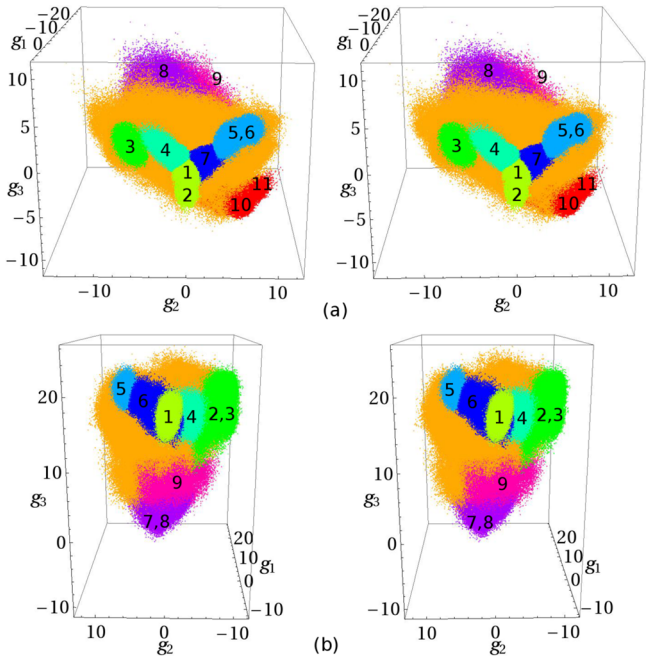


Figure 5. Stereoviews of the distribution of the representative points of beta3s in the 3D spaces of collective variables $\mathbf{g} = (g_1, g_2, g_3)$. Panel a is for the equilibrium folding (reproduced with permission from ref 24), and panel b is for the first-passage folding. In both cases, the g_1, g_2, and g_3 variables are in angstroms.

overlap in the hydrogen bond space, the distance between them $g = [\sum_{i=1}^{i=3}(g_i^{(2)} - g_i^{(1)})^2]^{1/2}$ is proportional to the all-atom RMSD between the corresponding protein conformations (Figure 6); this holds at approximately for $g > 3.6$ Å. Due to this relation, the distribution of the spatially separated clusters in the **g** space can be viewed as a distribution in the RMSD space. It should be noted that because the variables g_1, g_2, and g_3 are different for the EF and FPF (Figure 3), the scaling is different: in the former case one unit in the **g** space corresponds to approximately 0.14 Å in the RMSD space, and in the latter to approximately 0.07 Å.

Tables 1 and 2 show the clustering of the points obtained with the MCLUST program[35] for EF and FPF, respectively. In each table, the first column is the cluster number, and the second column is the relative number of points in the cluster (in percentage of the total number of 10^6 and 1.2×10^6 points, respectively). Also, these tables contain information about the protein secondary structures characteristic of each cluster. The third column presents the number of conformations that have different SSSs, the fourth column shows the SSSs of the two most populated secondary structures, and the fifth column the weight of these structures in the cluster. Finally, the last column indicates the type of representative protein conformation with which the cluster is associated according to the SSSs. The representative conformations are labeled as in the previous studies of folding of beta3s;[3,20−24] i.e., "native" stands for native-like structures, "Ns-or" for conformations in which the C-terminal hairpin is formed and the N-terminal hairpin is unstructured ("or" means "out of register"), "Cs-or" for conformations with the N-terminal hairpin formed and the C-terminal unstructured, "Ch-curl" for conformations that have a curl-like structure with the C-terminal hairpin formed, and "helical" for conformations that contain a helical region. Based on the similarity of the SSSs, the clusters for the structured conformations are grouped into five "consolidated" clusters, which represented locally stable characteristic conformations. For the EF, they consist of clusters 1 and 2 (native), cluster 3 (Cs-or), clusters 5 and 6 (Ns-or), clusters 8 and 9 (helical), and clusters 10 and 11 (Ch-curl). Also, two intermediate clusters, 4 and 7, are observed that present mixtures of the native-like conformations with the Cs-or and Ns-or conformations and are positioned between the native cluster and the Cs-or cluster and the Ns-or cluster, respectively. With these intermediate clusters joined to the native cluster, the residence probabilities of the system in the consolidated clusters is in good agreement with the results of the previous studies.[21,22] The clusters which present unstructured conformations form a pool of conformations (an "entropic" basin[21]) that connect the clusters of the structured conformations.

The main difference between FPF and EF is that in the former the Ch-curl conformations become so rare that they do not form a cluster, whereas the weight of the helical conformations drastically increases (Tables 1 and 2 and Figure 5a,b). This effect appears to be due to the fact that the ensemble of initial structures in the FPF consists of conformations that readily form helical conformations. Figure 7a shows the points in the **g** space at which the trajectories were started. It is seen that they lie on the boundary of the conformation space visited by the system, or more specifically, on the part of it that is close to the helical conformations, but they do not contain the hydrogen bonds between i and $i + 4$ residues that are characteristic of helices. Figure 8 gives typical examples of the corresponding conformations. The "secondary structure" of these conformations (i.e., a hairpin-like form with distant strands and the presence of local chain bends) suggests that because they involve short-range contacts, the formation of helical conformations is dynamically much more likely than the formation of Ch-curl conformations, because the latter require the N- and C-terminal strands to come into contact that are distant along the chain. According to the criterion we used to define the native contacts, $d_{nat} < 7.5$ Å (section 2), the average number of native contacts in the initial conformations is equal to 8, i.e., approximately 27% of the total number of native contacts ($N_{nat} = 23$). In addition, a comparable number of non-native contacts (on average, 11 contacts) is present in these conformations. The 200 conformations, which make up the initial states, are $200/(1.2 \times 10^6) \approx 0.017\%$ of the total number of the recorded conformations. For comparison, the number of

dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287−4299

Table 1. Clusters of Protein Conformations

cluster[a]	W_{clst}[b]	N_{str}[c]	most populated structure[d]	W_{str}[e]	cluster type[f]
1	21.5	523	-EEEETTEEEEEETTEEEE-	38.6	native
			-EEEETTEEEEEETTEEE--	37.0	
2	3.9	939	-EEEETTEEEEEETTEEEE-	16.2	
			-EEEETTEEEEEETTEE---	14.1	
3	2.6	2337	-EEEETTEEEEEEEEEE--	12.3	Cs-or
			-EEEETTEEEEEEEEEEEE-	9.8	
4	3.1	1173	-EEEETTEEEEE-SS-EEE-	7.2	Cs-or + native
			-EEEETTEEEEE-SS-EE--	5.6	
5	3.0	773	-EEE-SSS-EEEETTEEEE-	46.1	Ns-or
			-EEEESSSEEEEETTEEEE-	5.5	
6	2.5	631	-EEE-SSS-EEEETTEEEE-	22.3	
			-EEEESSSEEEEETTEEEE-	19.8	
7	5.0	1005	-EEEETTEEEEEETTEEE--	8.4	Ns-or + native
			-EE--SSS-EEEETTEEEE-	6.6	
8	7.6	48567	--HHHHHHHHHHHT------	0.4	helical 1
			---HHHHHHHHHHT------	0.2	
9	5.1	33302	--SS--HHHHTTT-------	0.3	helical 2
			--SS--HHHHHHHSS-----	0.3	
10	3.3	2347	-B-SSSSS--EEETTEE-B-	5.6	Ch-curl 1
			-B--SSS---EEETTEE-B-	4.5	
11	4.4	5758	-B-SSSSS-EEEETTTEEE-	3.3	Ch-curl 2
			-B-SSSS--EEEETTTEEE-	3.2	
12	4.6	13206	-EEEETTEEEE--SS-----	1.5	others
			-EEEETTEEEE-SSS-----	1.3	
13	3.2	3799	-EEEETTEEEEEETTEEEE-	7.1	
			----BTTEEEEEETTEEEE-	3.0	
14	8.4	15590	-----SS--EEEETTEEEE-	1.5	
			----SSS--EEEETTEEEE-	1.3	
15	8.7	47727	-EE-SSS-EE---SS---B-	0.7	
			-EEE-SSS-EEEEEEEEE--	0.4	
16	3.4	17009	-EEEETTEEE---SS-----	0.6	
			-B---SSS-----SSS--B-	0.5	
17	9.7	63733	-EEETTTEEEETTTEEEE--	0.3	
			----SSS-----SSS-----	0.2	

[a]Cluster number. [b]Cluster weight equal to the number of representative points in the cluster relative to the total number of the points (in %). [c]The number of conformations that have different secondary structure strings. [d]The secondary structure strings of the most populated conformations. [e]Weight of the given conformation in the cluster (in %). [f]Corresponds to Figure 4.

corresponding conformations along the 20 μs equilibrium trajectory[24] (i.e., the conformations that have the numbers of contacts not exceeding 8 native and 11 non-native contacts) is equal to 543, which is ≈0.05% of the total number of the conformations that were included in the analysis. This suggests that the conformations from which the FPF trajectories were started approximate the most unfolded conformations that occurred in EF.[24]

A closer examination of the FPF shows that the Ch-curl conformations that constitute a considerable fraction of the conformations observed in the course of EF are found only in 20 of the 200 trajectories, with the total fraction of them ($1.2 \times 10^6 \times 0.054 \times 0.014 \approx 900$, cluster 14 in Table 2) being ≈6.8 times smaller than for the EF ($1 \times 10^6 \times [0.033 \times (0.056 + 0.045) + 0.044 \times (0.033 + 0.032)] \approx 6200$, clusters 10 and 11 in Table 1). At the same time, the weight of the helical conformations increases, from ≈12.8% to ≈21.8%, which is comparable with the total weight of the Cs-or, Ns-or, and helical clusters in the EF (Tables 1 and 2). It is of interest that if the representative points for the FPF are projected onto the collective variables g_1, g_2, and g_3 for the EF, a cluster for Ch-curl conformations emerges and has a weight 2.2%. The weights of the other clusters also change but not greatly, staying within the variations of the weights of these clusters that are obtained with different methods (Table 2 of ref 24); for the native, Cs-or, and helical clusters they decrease by 20−30%, and for the Ns-or cluster it increases by 35%. These changes, and particular the appearance of the Ch-curl cluster, indicate that the principal coordinates obtained with the HB PCA method are specific to the manifold of the representative points to which the method is applied, and thus to the process that produces this manifold. We note also that the cluster of native-like conformations at which the trajectories were terminated in the FPF simulations is as significant as for the EF simulations. This is true mainly because a variety of conformations corresponding to the condition $N_{nat} - 1$ used to terminate the trajectory exists that have different coordinates in the **g** space.

The increased contribution of helical conformations also affects the hydrogen bond composition of the collective variables (Figure 9). The variables g_1 and g_2 in the FPF have the largest projections onto the same eight bonds as they had in the case of the EF, and they thus play a similar role as in the EF; i.e., g_1 serves as a good reaction coordinate for the overall description of the folding process, and g_2 discriminates between

dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287−4299

Table 2. Clusters of Protein Conformations: The First-Passage Folding

cluster[a]	W_{clst}[b]	N_{str}[c]	most populated structure[d]	W_{str}[e]	cluster type[f]
1	10.9	554	-EEEETTEEEEEETTEEEE-	36.6	native
			-EEEETTEEEEEETTEEE--	28.8	
2	3.5	6093	-EEEETTEEEE-SSS-----	3.5	Cs-or
			--EEETTEEEEETTTEEEE-	2.7	
3	4.6	1095	-EEEETTEEEEEEEEEE--	13.0	
			-EEEETTEEEEEEEEEEE-	11.8	
4	2.1	4027	--EEETTEEEE-SSS-EE-	3.9	Cs-or + native
			-EEEETTEEE--SSS----	3.1	
5	2.7	1402	-EEEESSSEEEEETTEEEE-	19.8	Ns-or
			-EEE-SSS-EEEETTEEEE-	16.1	
6	6.8	3838	-EEE-SSS-EEEETTEEEE-	10.2	Ns-or + native
			-EEEETTEEEEEETTEEEE-	8.4	
7	7.3	29 644	---HHHHHHHHHHS------	0.8	helical 1
			--HHHHHHHHHHHT------	0.7	
8	9.0	63 876	--HHHHHHHHHHHT------	0.1	
			---HHHHHHHHHHSS-----	0.1	
9	5.4	45 686	----SSS-HHHHHT------	0.1	helical 2
			-----SSS-HHHHT------	0.1	
10	6.2	28 656	-----SSS-EE-SSS--EE-	0.6	others
			-----SS--EE-SSS--EE-	0.5	
11	7.1	59 800	-----SS-SSB-SS-B----	0.1	
			-----SS-----BTTTB---	0.1	
12	5.8	40 779	----BTTB----SSS-----	0.3	
			-EEEETTEEE--SSS-----	0.3	
13	4.4	28 806	-----SSS-EEEETTEEEE-	0.6	
			--EE--BTTEEEESSS-EE-	0.4	
14	5.4	30 721	-B-SSSS--EEEETTTEEE-	0.7	
			-B-SSSSS-EEEETTTEEE-	0.7	
15	3.9	9065	----SSS--EEEETTEEEE-	1.7	
			-----SS--EEEETTEEEE-	1.6	
16	5.5	16 541	-----SS--EEEETTEEEE-	1.1	
			----SSS--EEEETTEEEE-	0.8	
17	4.0	14 991	---EETTEE---SSS-----	1.2	
			-EEEETTEEEE--SS-----	1.1	
18	1.1	6406	-----SSS-SSEETTEE---	0.7	
			---SB-SSTTTEETTEE---	0.7	
19	4.3	22 988	-EEEETTEEEEETTTEEEE-	0.6	
			-EEEETTEEEEEEEEEEE-	0.5	

[a]Cluster number. [b]Cluster weight equal to the number of the representative points in the cluster relative to the total number of the points (in %). [c]The number of conformations that have different secondary structure strings. [d]The secondary structure strings of the most populated conformations. [e]Weight of the given conformation in the cluster (in %). [f]Corresponds to Figure 5b.

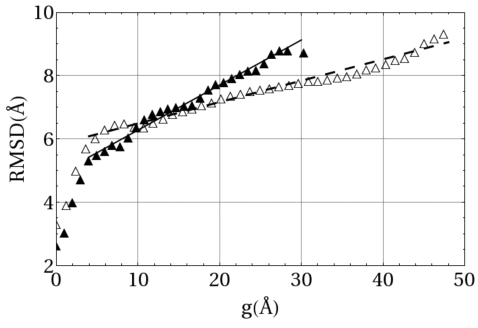


Figure 6. All-atom RMSD as a function of the distance in the **g** space. The solid and empty triangles correspond to the equilibrium and first-passage folding, respectively. The solid and dashed lines show the best fits to the data with the slopes of the lines ≈0.14 and ≈0.07, respectively.

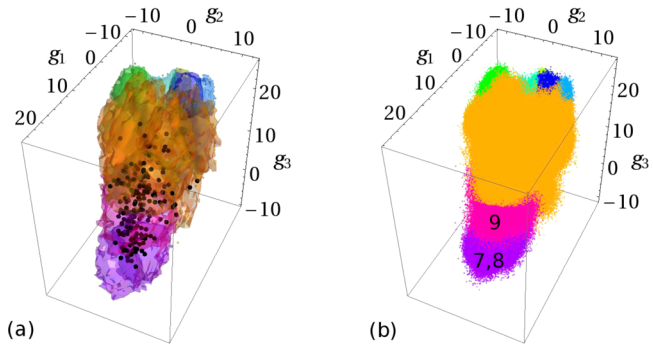


Figure 7. First-passage folding. Panel a depicts the starting points superposed on the distribution of the representative points of Figure 5b. The clusters of the representative points are colored according to this figure and, to make the starting points visible, are shown as semitransparent objects. Panel b reproduces Figure 5b in the same orientation as for (a).

dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287−4299

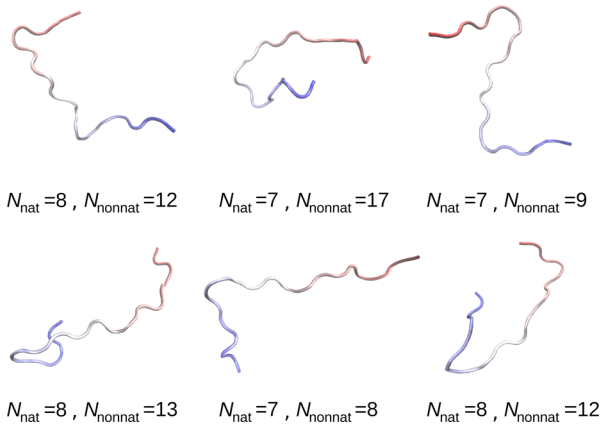


Figure 8. Examples of the initial conformations for the first-passage folding simulations.

the Ns-or and Cs-or conformations.[24] However, the role of the g_3 variable is essentially different: athough in the EF g_3 cannot be associated with any secondary structure element,[24] in the case of the FPF, it has the largest projections on the bonds characteristic of helical conformations, i.e., the hydrogen bonds between i and $i + 4$ residues. The most important bonds among them are at the N-terminal end, in agreement with their SSSs in Table 2.

Panels a and b of Figure 10 present spatial kinetic networks for the EF and FPF, which show how the clusters of protein conformations are connected in these cases. The ball volumes are proportional to the number of intracluster transitions, and the tube cross sections to the number of intercluster transitions (the latter were calculated as one-half of the total number of the forward and backward transitions between two clusters). More clearly, the difference between the cluster interconnection is seen from the paths of passive tracers (Figure 11a,b) and a directed kinetic network for the FPF (Figure 12). In Figure 11a,b the paths were initiated, respectively, at the representative points of Figure 5a,b with the largest fluxes $\mathbf{j}(\mathbf{g})$ and continued for some time (for details, see ref 24); the number of the points is equal to 900 for the EF and to 766 for the FPF. It is seen that in the FPF, in contrast to the EF, there are many tracer paths between the clusters for unstructured conformations and the native cluster, whereas the direct paths between the Ns-or (5) and Cs-or (2 and 3) clusters and the native cluster (1) are absent. Because the intensity of a tracer path is proportional to the (average) flux $\mathbf{j}(\mathbf{g})$,[24] the absence of the path can be a result of either the lack of the transitions or the presence of detailed balance. As is seen from Figure 10b, the numbers of transitions between the Ns-or and Cr-or clusters and the native cluster (the cross sections of the tubes) are comparable with those from the clusters for unstructured conformations to the native cluster. It follows that detailed balance between the Ns-or and Cr-or clusters and the native cluster exists. This is confirmed by the directed kinetic network, depicted in Figure 12, in which the tubes of the transitions between the clusters are taken to be proportional to the difference between the upward and backward transitions (to make the picture more clear, the tubes with not less than ten transitions, among the total number of transition $\sim 10^6$, are not shown). Moreover, direct counting of the numbers of transitions between the Cr-or and

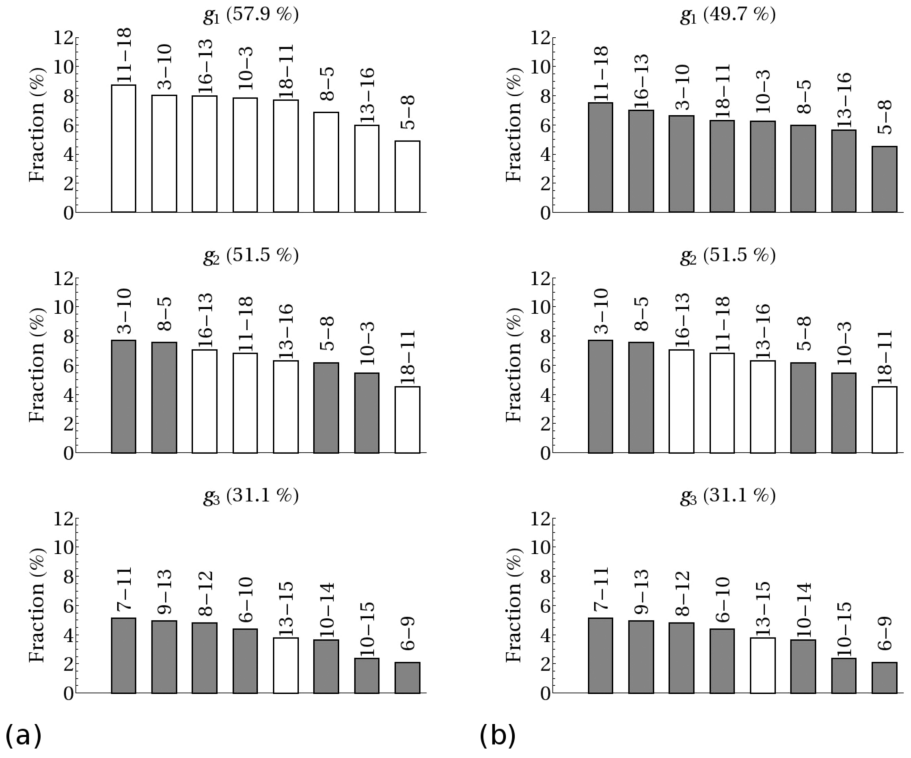


Figure 9. Fractions of the hydrogen bonds which make a major contribution to the collective variables g_1, g_2, and g_3. Panel a is for the equilibrium folding (reproduced with permission from ref 24), and panel b is for the first-passage folding. The figures at the top of each bar denote the bond; the first figure is the number of the residue with the oxygen atom, and the second figure is that with the nitrogen atom. The empty and solid bars are for the bond contributions to the negative and positive directions of the collective variable, respectively. The numbers in percentage at the top of each panel are the total contribution of the given bonds to the collective variable.

dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287–4299

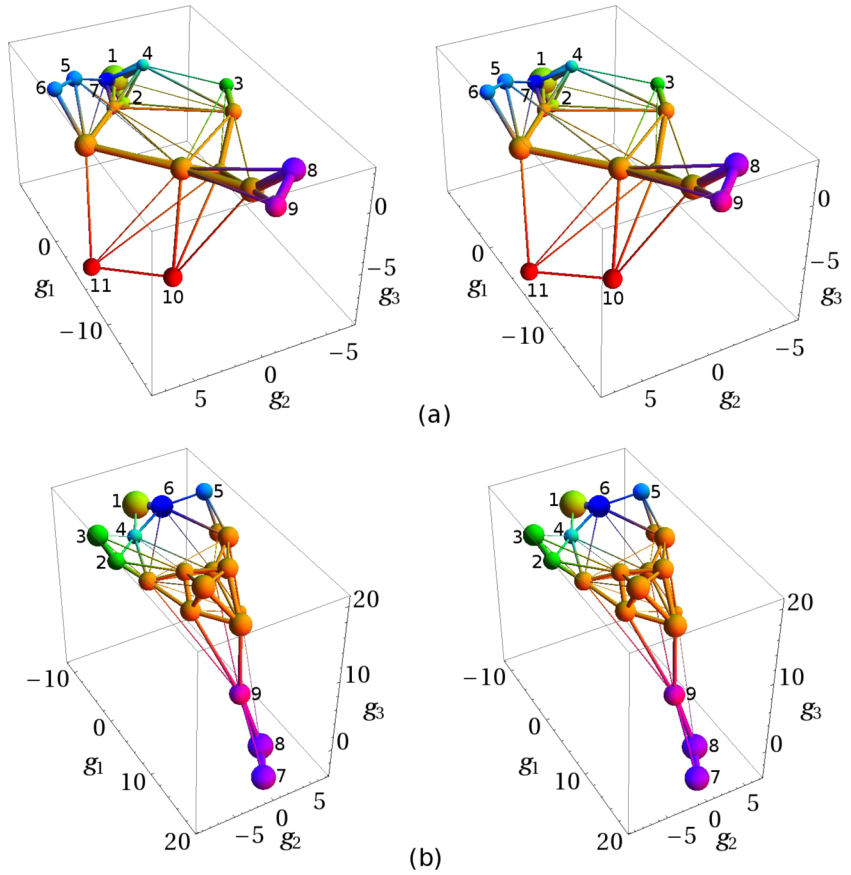


Figure 10. Stereoviews of the spatial kinetic networks. Panel a is for the equilibrium folding (reproduced with permission from ref 24), and panel b is for the first-passage folding. Clusters are numbered as indicated in the text and colored according to the palette of Figure 5. The units of the g_1, g_2, and g_3 variables are in angstroms.

Ns-or clusters and the native cluster shows that detailed balance between them is satisfied exactly. Thus, the overall flow goes from the unfolded states to the native state directly, not passing through the structured Cs-or and Ns-or conformations, which supports the conclusion that beta3s is a barrierless/low-barrier folder.[21] The most probable pathway is illustrated in Figure 13, which presents a two-dimensional kinetic network corresponding to the directed three-dimensional kinetic network of Figure 12. The red line that connects clusters 7, 8, and 9, besides which the trajectories were started (Figure 7), with cluster 1 for native-like conformations shows the shortest pathway, which was calculated using the Bellman−Ford algorithm.[39]

Figures 14 and 15 show the FES, two-dimensional streamlines and tracer paths of folding flows for the EF and FPF, respectively. For the FPF the stream function is normalized such that $\Psi = 1$ corresponds to the total folding flow from the unfolded states to the native basin, i.e., to 200 folding trajectories. For the EF, where there is no net flow from the unfolded states to the native state, the normalization of the stream function was performed by assuming that the total number of (virtual) trajectories would be less than for the FPF as the ratio of the numbers of frames in these cases, i.e., by $10^6/1.2 \times 10^6 \approx 0.83$ times. Because in the case of FPF every folding trajectory initiated at an extended state reaches and is terminated in the native basin, the total folding flow is the same in each (g_2 = constant) cross-section. As in the EF, local minima corresponding to the clusters of characteristic conformations are observed; they are the clusters indicated in Table 2 and Figures 5b, 10b, 11b, and 12. However, the flow fields are drastically different from those for the EF, in both the streamlines and tracer paths. Although small vortices are still present at the minima, similar to the EF, indicating that the system spends some time in them, there exists a pronounced overall folding flow from the unfolded states to the native state. It is represented by streamtubes that originate at the unfolded states of the protein (large values of g_1) and converge at the native state ($g_1 \approx -10$). Similarly, tracer paths connecting the unfolded and native states are present. Such a behavior of the streamtubes and tracer paths has been previously observed in the FPF simulations of an α-helical hairpin and SH3 domain (streamtubes[8,10] and tracer paths[11]). For the EF, in contrast, neither the streamtubes or tracer paths that have such properties are present.

Panels a and b of Figure 16 show the dependence of the transition rate upon the distance between the clusters of conformations in the **g** space. Although the scattering of the data for the FPF is higher than that for the EF, the reduced standard error of partial slopes is comparable—it is equal to ≈9% for the EF, and to ≈11% for the FPF. Therefore, in the FPF case the average decrease of the rates with the distance remains roughly exponential and is approximately the same as for the EF. As has been indicated in ref 24, this dependence is in accord with the fact that the distance in the **g** space is correlated with the change in hydrogen bonding required to go from one cluster to another. The robustness of this behavior is of interest; it shows that though the overall folding pictures for the FPF and EF are drastically different, the elementary rates,

dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287−4299

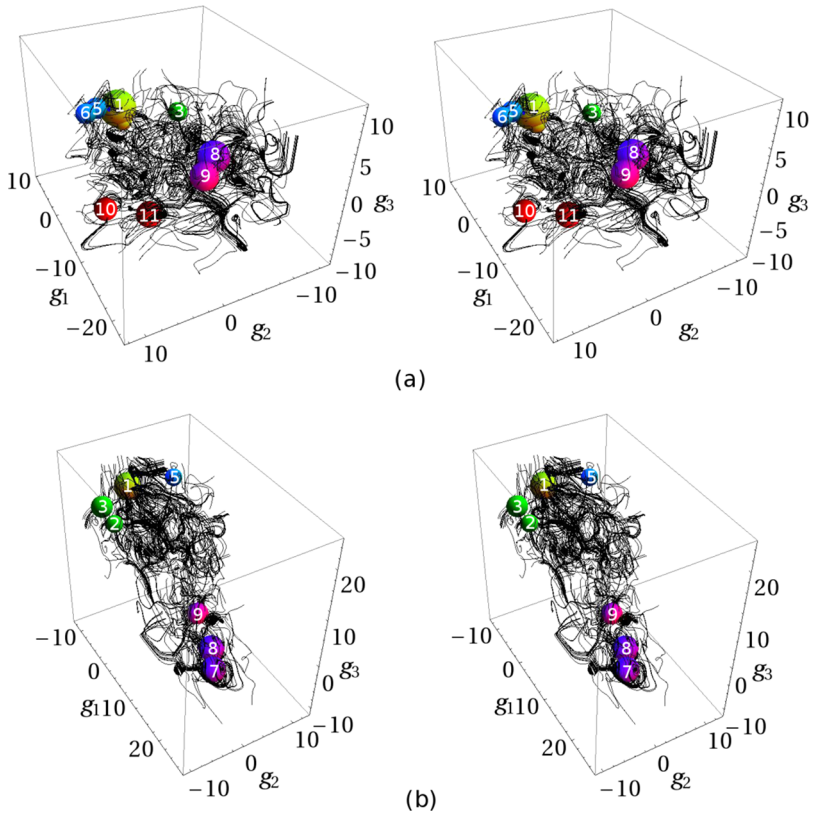


Figure 11. Stereoviews of passive tracer paths. Panel a is for the equilibrium folding (reproduced with permission from ref 24), and panel b is for the first-passage folding. The balls represent the native, Cs-or, Ns-or, Ch-curl, and helical clusters shown in the corresponding panels (a and b) of Figure 10. The radii of the balls are increased for illustrative purpose.

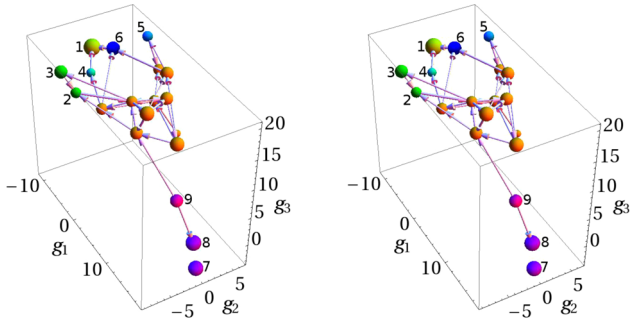


Figure 12. Stereoview of the directed kinetic network of beta3s for the first-passage folding. Clusters are numbered as in Table 2 are colored according to the palette of Figure 5. Variables g_1, g_2, and g_3 are in angstroms.

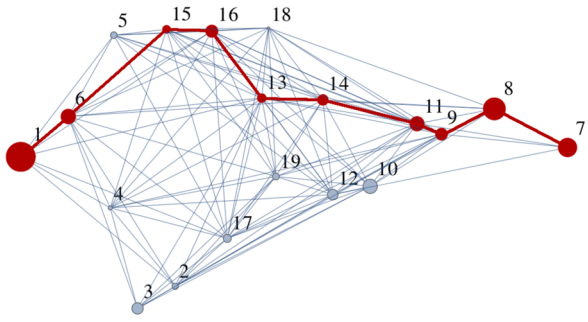


Figure 13. Two-dimensional kinetic network for the first-passage folding. The red line shows the most probable (shortest) pathway calculated with the Bellman–Ford algorithm.[39]

i.e., the rates of transitions between the clusters, remain the same at the same temperature.

It is of interest to compare the folding time distribution for the FPF with that obtained from the EF.[24] To determine the latter, we selected all segments of the equilibrium trajectory of ref 24 between two successive visits of the native state. If the considered segment contained a conformation with eight of less native contacts (similar to those we chosen for the initial conformations to start the trajectories in the FPF simulations, Figure 7a), the part of this segment from the point with the lowest number of native contacts to the native state was taken as a "first-passage trajectory". There were 130 such trajectory segments in the equilibrium simulation. The distribution of the first-passage times is very close to that for the FPF (Figure 4). Figure 17 also depicts the representative points of the system that fall into these first-passage trajectories (colored from blue to red) and the points that are outside the trajectories (black). Comparison of this figure with Figure 5a shows that the points within the first-passage trajectories are mostly related to the conformations that are distant from the native state, including the helical- and Ch-curl-like conformations and unfolded conformations. The points that are outside the first-passage trajectories are related to the conformations close to the native state, i.e., those within the Cs-or or Ns-or clusters, the intermediate clusters, and the clusters of native-like conformations.

Figure 11b suggests that the folding flows are very far from uniform. To illustrate this, Figure 18 shows the distribution of the g_1-component of the folding flux $\mathbf{j}(\mathbf{g})$ in a g_1 = const cross-section of the $\mathbf{g}$ space close to the native state. However, despite all the heterogeneity of the fluxes (Figure 19), their

dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287–4299

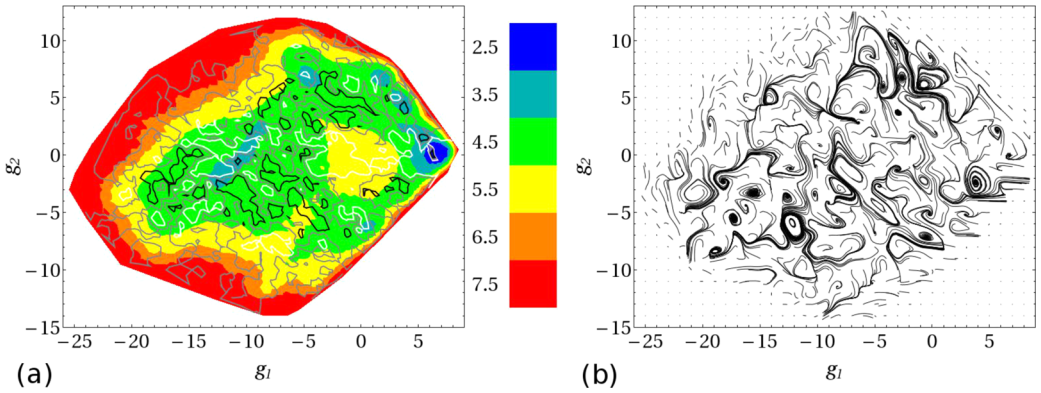


Figure 14. Protein folding in two-dimensional (g_1, g_2) space, the equilibrium folding (reproduced with permission from ref 24). Panel a shows the streamlines superimposed on the free energy surfaces (in kcal/mol). The blue local minima on the surfaces correspond to the clusters indicated in Table 1 and Figures 5a, 10a, and 11a. In panel a, the white, gray, and black lines correspond to the stream function values $\Psi = -0.01$, $\Psi = 0$, and $\Psi = 0.01$, respectively. The closed white and black streamlines restrict the vortex regions, in which the rotation of folding flows is, respectively, clockwise and anticlockwise. Panels b depicts the paths of passive tracers.

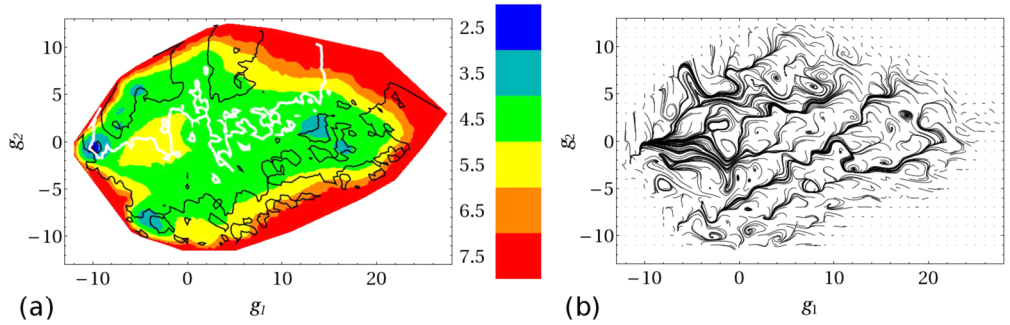


Figure 15. Protein folding in two-dimensional (g_1, g_2) space, the first-passage folding. Panel a shows the streamlines superimposed on the free energy surfaces (in kcal/mol). The blue local minima on the surfaces correspond to the clusters indicated in Table 2 and Figures 5b, 10b, and 11b. In panel a, the lower and upper black lines correspond to approximately the lower and upper bounds of the total folding flow, and the white lines to the half of the flow (the values of the normalized stream function at these lines are $\Psi = 0.01$, $\Psi = 0.5$, and $\Psi = 0.9$, respectively). Panel b depicts the paths of passive tracers.

distribution possesses a well pronounced property of self-similarity, similar to what was previously found for folding of SH3 domain.[11] To estimate the degree of the heterogeneity of the fluxes, we calculated the function $G(L) = \langle |J_{g_1,L}|/\bar{J}_{g_1} \rangle$, where $|J_{g_1,L}|$ is the absolute value of g_1 component of the flow through the square of linear size L, M is the number of elementary squares covered by the square of size L, $\bar{J}_{g_1} = (\sum_1^M J_{g_1,i}^2/M)^{1/2}$ is the average flux in g_1-direction, and the angular brackets denote the averaging over the g_1-cross sections of the $\mathbf{g} = (g_1, g_2, g_3)$ space. The linear size L is measured in units of the elementary square linear size l, which was taken to be 1 Å. It is seen that $G(L) \sim L^D$, where $D \approx 0.68$. Because D is less than 2, i.e., the Euclidean dimension expected for a homogeneous flow, the flows are fractal, with the exponent D being the fractal dimension.[40]

4. CONCLUDING DISCUSSION

In this paper we compare first-passage folding (FPF) (the process of going from the unfolded to the folded state) with folding under equilibrium conditions (EF) (i.e., when there are many folding and unfolding events). We find that there are significant differences between the two. A reason why this is of interest is that generally in living systems, the conditions are such that after the protein is synthesized on the ribosome, the folded (native) protein is stable and unfolding is a rare event.

There is considerable uncertainty concerning the initial conditions from which folding takes place.[41] It is possible, for example, that in some cases, partial folding to form helices takes place before the polypeptide chain leaves the ribosome. However, essentially all of the large number of folding simulations have been in aqueous solution in the absence of other cellular elements; exceptions are folding/unfolding studies of the role of GroEL, for example. Given that, it is reasonable to argue that the first-passage folding simulations described here are likely to be more realistic than equilibrium folding simulations.

The initial stage of the FPF occurs from nearly fully unfolded conformations, which are relatively rare in the EF simulations, even at temperatures where the folded and denatured states are both populated. When the trajectory starts to fold from an extended conformation, it first reaches either a helical conformation, which is readily formed due to the short-range contacts involved, or double hairpin Cs-or or Ns-or conformations, which consist of antiparallel β-strands. Formation of a Ch-curl conformation is less probable in FPF than in EF because it contains a parallel β-strand arrangement; it is, thus, less stable because the hydrogen bonds are distorted in comparison to those of the parallel β-strand arrangement,[42] and it is more difficult to form dynamically because it has distant N- and C-terminal strands. The Ch-curl conformations become so rare that they do not form a cluster, while the weight of the

dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287−4299

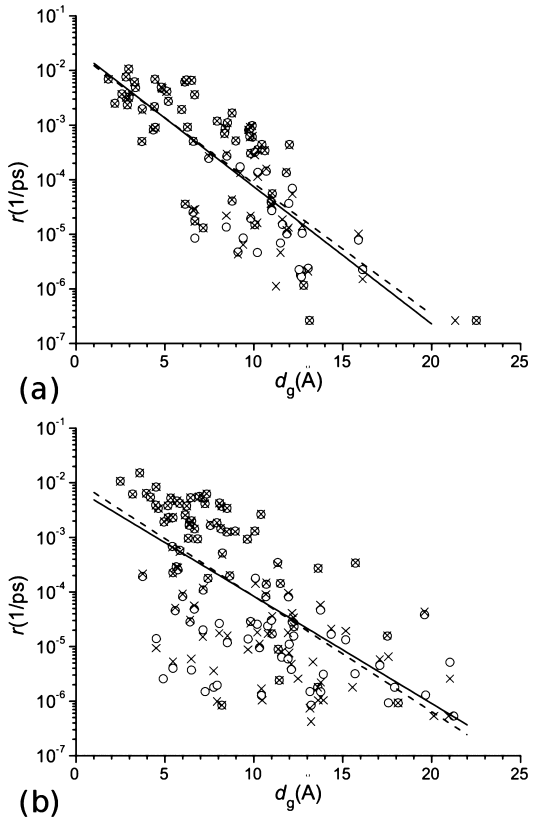


Figure 16. Rates of transitions between the clusters of conformations vs the distances between the centers of the clusters in the **g** space. Panel a is for the equilibrium folding (reproduced with permission from ref 24), and panel b is for the first-passage folding. In both cases the crosses and circles are for the transitions from smaller and larger populated clusters, respectively. In panel a the dashed line corresponds to the best fit for the crosses [$r \sim \exp(-0.55d_g)$], and the solid line to that for the circles [$r \sim \exp(-0.58d_g)$]. In panel b the corresponding fits are $r \sim \exp(-0.48d_g)$ and $r \sim \exp(-0.55d_g)$, respectively.

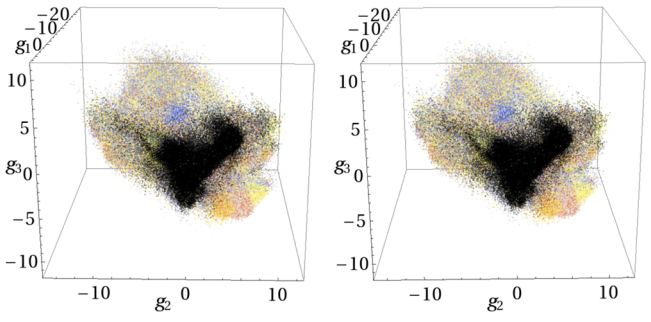


Figure 17. Stereoviews of the distribution of the representative points that fall into the first-passage segments of the equilibrium trajectory (colored from blue to red) and which are outside these trajectories (colored black).

helical conformations drastically increases, from ≈12.8% to ≈21.8%, which is comparable with the total weight of the Cs-or, Ns-or and helical clusters. The increased contribution of helical conformations also affects the hydrogen bond structure of the collective variables, changing the role of the g_3 variable: Although the variables g_1 and g_2 have the largest projections onto the same eight bonds as they had in the case of the equilibrium folding. Thus, they preserve their functions as, respectively, the principal reaction coordinate and the

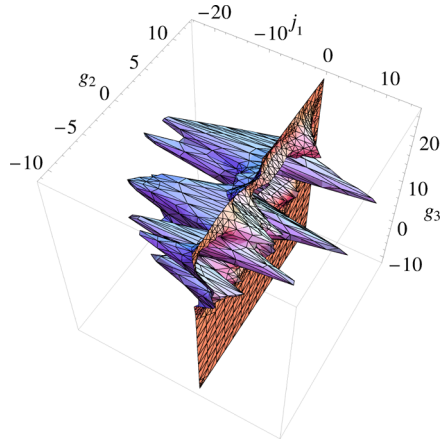


Figure 18. Distribution of the folding flux component j_1 in the cross-section $g_1 = -3.0$. The first-passage folding. The negative sign of j_1 corresponds to the direction toward the native state.

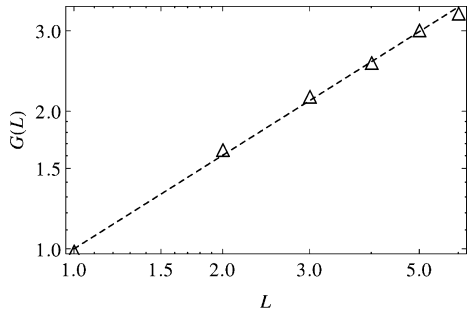


Figure 19. Heterogeneity of folding fluxes, the function $G(L)$ (see the text). The first-passage folding. The symbols show the function $G(L)$, and the dashed line the best fit to the function $G(L) \sim L^D$; $D \approx 0.68$.

coordinate that distinguishes between the Cs-or or Ns-or conformations, the largest projections of g_3, which did not relate to a specific conformation in EF, correspond in FPF to the bonds characteristic of helical conformations. In other words, the variable g_3 captures the essential difference between the first-passage and equilibrium processes. It is of interest that when the representative points for the first-passage folding are projected onto the collective variables g_1, g_2, and g_3 for the equilibrium folding, a cluster for Ch-curl conformations emerges, though with a low weight (2.2%). This indicates that the principal coordinates obtained with the HB PCA method are specific to the manifold of the representative points to which the method is applied, and thus to the process which produces this manifold.

Counting the numbers of transitions between the clusters, the 3D distribution of the representative points has been represented in the form of spatial kinetic networks, undirected and directed. These networks have shown that the folding flows do not go through the Cs-or and Ns-or structures that are conformationally close to the native state, which is consistent with the conclusion that beta3s is a barrierless/low-barrier folder.[21] Easy rearrangement of the Cs-or and Ns-or conformations into the native conformation and back leads to detailed balance between these structures and thus makes the flow through them negligible (at least, for the temperature close to the melting temperature that is used here).

Another essential difference between the first-passage and equilibrium folding is revealed by the "hydrodynamic" analysis.[8,10,11,24] The projection of the passive tracer paths

dx.doi.org/10.1021/jp412729r | J. Phys. Chem. B 2014, 118, 4287−4299

representing the "streamlines" of the folding flows onto the FESs depending on two variables shows that in the case of equilibrium folding the folding flow field consists of a variety of small vortices, not only at the minima corresponding to the clusters of protein conformations (native, Cs-or, Ns-or, Ch-curl, and helical) but also in flat regions of the PES. This indicates that the local folding flows do not follow the PES landscape. In contrast, the streamlines for the first-passage folding are mostly directed from the denatured to the native state, although they are complex and do not exactly follow the PES landscape. It is of interest that despite all the complexity of the folding flows, their distribution is self-similar and has fractal dimension ($D \approx 0.68$). A similar property of folding flows has been previously observed for folding of the SH3 domain,[11] although the fractal dimension was different, varying from $D \approx 1.5$ for the initial (almost "laminar") stage of folding to $D \approx 1$ for the final ("turbulent") stage. This suggests that the self-similarity of folding flows may be an inherent property of protein folding.

Although there are significant differences in the general picture of the folding process from the equilibrium and first-passage folding simulations, some aspects of the two are in agreement. The rate of transitions between the clusters of characteristic protein conformations in both cases decreases approximately exponentially with the distance between the clusters in the hydrogen bond distance space of collective variables, and the folding time distribution in the first-passage segments of the equilibrium trajectory is in good agreement with that for the first-passage folding simulations. Also, the first-passage segments of the EF trajectory that start at an unfolded state of the protein and converge to the native state are similar to the trajectories in the FPF simulations in that they have similar folding time distributions.

AUTHOR INFORMATION

Corresponding Authors

*Sergei F. Chekmarev: e-mail, chekmarev@itp.nsc.ru; phone, +7(383)3165048.

*Martin Karplus: e-mail, marci@tammy.harvard.edu; phone, +1 617.495.4018; fax, +33(0)368855123.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank A. Caflish for making the equilibrium trajectory available to us and A. Vitalis for help with the WORDOM program. This work was supported in part by the grant from the U.S. Civilian Research and Development Foundation (RUB2-2913-NO-07). The research at Harvard was supported in part by a grant from the National Institutes of Health. The first passage folding simulations were performed in the Supercomputer Center of the Siberian Branch of the Russian Academy of Sciences.

REFERENCES

(1) Shea, J. E.; Brooks, I. C. L. From Folding Theories to Folding Proteins: A Review and Assessment of Simulation Studies of Protein Folding and Unfolding. *Annu. Rev. Phys. Chem.* **2001**, *52* (7), 499−535.

(2) Becker, O. M.; Karplus, M. The Topology of Multidimensional Potential Energy Surfaces: Theory and Application to Peptide Structure and Kinetics. *J. Chem. Phys.* **1997**, *106* (4), 1495−1517.

(3) Rao, F.; Caflisch, A. The Protein Folding Network. *J. Mol. Biol.* **2004**, *342* (1), 299−306.

(4) Krivov, S. V.; Chekmarev, S. F.; Karplus, M. Potential Energy Surfaces and Conformational Transitions in Biomolecules: A Successive Confinement Approach Applied to a Solvated Tetrapeptide. *Phys. Rev. Lett.* **2002**, *88* (3), 038101.

(5) Noé, F.; Fischer, S. Transition networks for Modeling the Kinetics of Conformational Change in Macromolecules. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.

(6) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. Markov State Model Reveals Folding and Functional Dynamics in Ultra-Long MD Trajectories. *J. Am. Chem. Soc.* **2011**, *133*, 18413−18419.

(7) Chekmarev, S. F.; Krivov, S. V.; Karplus, M. Folding Time Distributions as an Approach to Protein Folding Kinetics. *J. Phys. Chem. B* **2005**, *109* (11), 5312−5330.

(8) Chekmarev, S. F.; Palyanov, A. Y.; Karplus, M. Hydrodynamic Description of Protein Folding. *Phys. Rev. Lett.* **2008**, *100* (1), 018107.

(9) Palyanov, A. Y.; Krivov, S. V.; Karplus, M.; Chekmarev, S. F. A Lattice Protein with an Amyloidogenic Latent State: Stability and Folding Kinetics. *J. Phys. Chem. B* **2007**, *111* (10), 2675−2687.

(10) Kalgin, I. V.; Karplus, M.; Chekmarev, S. F. Folding of a SH3 Domain: Standard and "Hydrodynamic" Analyses. *J. Phys. Chem. B* **2009**, *113* (38), 12759−12772.

(11) Kalgin, I. V.; Chekmarev, S. F. Turbulent Phenomena in Protein Folding. *Phys. Rev. E* **2011**, *83* (1), 011920.

(12) Dinner, A. R.; Karplus, M. Is Protein Unfolding the Reverse of Protein Folding? A Lattice Simulation Analysis. *J. Mol. Biol.* **1999**, *292* (2), 403−419.

(13) Lipman, E. A.; Schuler, B.; Bakajin, O.; Eaton, W. A. Single-Molecule Measurement of Protein Folding Kinetics. *Science* **2003**, *301*, 1233−1235.

(14) Fersht, A. R.; Daggett, V. Protein Folding and Unfolding at Atomic Resolution. *Cell* **2002**, *108*, 573−582.

(15) Day, R.; Daggett, V. Direct Observation of Microscopic Reversibility in Single-Molecule Protein Folding. *J. Mol. Biol.* **2007**, *366* (2), 607−686.

(16) Toofanny, R. D.; Daggett, V. Understanding Protein Unfolding from Molecular Simulations. *WIREs Comput. Mol. Sci.* **2012**, *2*, 405−423.

(17) Levy, R. M.; Dai, W.; Deng, N.-J.; Makarov, D. E. How Long Does It Take to Equilibrate the Unfolded State of a Protein? *Protein Sci.* **2013**, *22*, 1459−1465.

(18) Ferrara, P.; Caflisch, A. Folding Simulations of a Three-Stranded Antiparallel Beta-Sheet Peptide. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97* (20), 10780−10785.

(19) Settanni, G.; Rao, F.; Caflisch, A. Φ-Value Analysis by Molecular Dynamics Simulations of Reversible Folding. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102* (3), 628−633.

(20) Muff, S.; Caflisch, A. Kinetic Analysis of Molecular Dynamics Simulations Reveals Changes in the Denatured State and Switch of Folding Pathways upon Single-Point Mutation of a Beta-Sheet Miniprotein. *Proteins: Struct., Funct., Bioinform.* **2008**, *70* (4), 1185−1195.

(21) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. One-Dimensional Barrier-Preserving Free-Energy Projections of a Beta-Sheet Miniprotein: New Insights into the Folding Process. *J. Phys. Chem. B* **2008**, *112* (29), 8701−8714.

(22) Muff, S.; Caflisch, A. ETNA: Equilibrium Transitions Network and Arrhenius Equation for Extracting Folding Kinetics from REMD Simulations. *J. Phys. Chem. B* **2009**, *113* (10), 3218−3226.

(23) Qi, B.; Muff, S.; Caflisch, A.; Dinner, A. R. Extracting Physically Intuitive Reaction Coordinates from Transition Networks of a Beta-Sheet Miniprotein. *J. Phys. Chem. B* **2010**, *114* (20), 6979−6989.

(24) Kalgin, I. V.; Caflisch, A.; Chekmarev, S. F.; Karplus, M. New Insights into the Folding of a Beta-Sheet Miniprotein in a Reduced Space of Collective Hydrogen Bond Variables: Application to a Hydrodynamic Analysis of the Folding Flow. *J. Phys. Chem. B* **2013**, *117* (20), 6092−5105.

(25) Brooks, B. R.; Brooks, C. L., III; MacKerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.;

Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545−1614.
(26) Cavalli, A.; Ferrara, P.; Caflisch, A. Weak Temperature Dependence of the Free Energy Surface and Folding Pathways of Structured Peptides. *Proteins: Struct., Funct., Genet.* **2002**, *47* (3), 305−314.
(27) De Alba, E.; Santoro, J.; Rico, M.; Jiménez, M. De novo Design of a Monomeric Three-Stranded Antiparallel Beta-Sheet. *Protein Sci.* **1999**, *8* (4), 854−865.
(28) Neria, E.; Fischer, S.; Karplus, M. Simulation of Activation Free Energies in Molecular Systems. *J. Chem. Phys.* **1996**, *105* (5), 1902−1921.
(29) Ferrara, P.; Apostolakis, J.; Caflisch, A. Evaluation of a Fast Implicit Solvent Model for Molecular Dynamics Simulations. *Proteins: Struct., Funct., Genet.* **2002**, *46* (1), 24−33.
(30) Neira, J. L.; Fersht, A. R. Exploring the Folding Funnel of a Polypeptide Chain by Biophysical Studies on Protein Fragments. *J. Mol. Biol.* **1999**, *285*, 1309−1333.
(31) Gruebele, M. The Fast Pritein Folding Problem. *Annu. Rev. Phys. Chem.* **1999**, *50*, 485−516.
(32) Eaton, W. A.; Muñoz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. Fast Kinetics and Mechanisms in Protein Folding. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 327−59.
(33) Roder, H.; Maki, K.; Cheng, H. Early Events in Protein Folding Explored by Rapid Mixing Methods. *Chem. Rev.* **2006**, *106*, 1836−1861.
(34) Jolliffe, I. T. *Principal Component Analysis*; Springer Verlag: Berlin, 2002.
(35) Fraley, C.; Raftery, A. E. Model-Based Clustering, Discriminant Analysis, and Density Estimation. *J. Am. Stat. Assoc.* **2002**, *97* (458), 611−631.
(36) Andersen, C. A.; Palmer, A. G.; Brunak, S.; Rost, B. Continuum Secondary Structure Captures Protein Flexibility. *Structure* **2002**, *10* (2), 175−184.
(37) Seeber, M.; Cecchini, M.; Rao, F.; Settanni, G.; Caflisch, A. Wordom: a Program for Efficient Analysis of Molecular Dynamics Simulations. *Bioinformatics* **2007**, *23* (19), 2625−2627.
(38) Landau, L. D.; Lifshitz, E. M. *Fluid Mechanics*; Pergamon: New York, 1987.
(39) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L.; Stein, C. *Introduction to Algorithms*, 2nd ed.; MIT Press: Cambridge, MA, 2001.
(40) Moon, F. C. *Chaotic and Fractal Dynamic*; Wiley: New York, 1992.
(41) Whitford, D. *Proteins: Structure and Function*; Wiley: Hoboken, NJ, 2005.
(42) Voet, D.; Voet, J. G. *Biochemistry*, 4rd ed.; Wiley: Hoboken, NJ, 2011.