Alternation of phases of regular and irregular dynamics in protein folding

Sergei F. Chekmarev[*]
Institute of Thermophysics, SB RAS, 630090 Novosibirsk, Russia
and Physics Department, Novosibirsk State University, 630090 Novosibirsk, Russia

(Received 15 September 2018; published 19 February 2019)

The regularity of the dynamics in different phases of protein folding is investigated for a set of proteins which undergo a cooperative, two-state folding transition. To determine the degree of regularity of the dynamics, the fractal dimension of probability fluxes is calculated on the basis of simulated folding trajectories. It has been found that the phases of regular and irregular dynamics alternate as follows. In the initial (collapse) phase of folding, the dynamics are essentially regular. Then, as the protein comes to the basin of semicompact states that precedes the transition state, the dynamics become irregular. At the transition state, the dynamics are regularized again but become less regular when the nativelike states are explored. Depending on the specific conditions at which the protein folding was considered, some phases of the dynamics could not be well resolved, but no significant deviation from this general picture has been observed. The regularization of the dynamics at the transition state is discussed in relation to the recent studies of the Hamiltonian dynamics of small clusters, where both regular and chaotic dynamics were observed depending on the flatness of the energy surface at the transition state.

DOI: 10.1103/PhysRevE.99.022412

I. INTRODUCTION

Dynamics of protein folding are inherently very complex. In the course of folding, a protein generally follows many different routes and has to overcome free energy barriers of different heights. The overall sequence of folding events is well known [1–6]. In the simplest case of a cooperative, two-state folding transition, the protein first collapses to one of semicompact states, which form a basin on the free energy surface (FES), dwells in this basin and, then, finds a way to the native state by overcoming a free energy barrier that separates the semicompact states from the nativelike ones. At the same time, little is known about how the character of folding dynamics changes during the protein passage through these phases of folding. For folding of an α-helical hairpin [7], it has been found that the folding flow streamlines are regular in the initial phase of folding, fluctuate in the basin of semicompact states, and become regular again as the native state is approached. On the other hand, recent studies of folding of SH3 domain [8], beta3s miniprotein [9], and Trp-cage miniprotein [10] suggested that on the global scale, i.e., in the overall transition from the unfolded to the native state, the folding flows become less regular as the protein folds. The dynamics at the transition state are of particular interest in this respect because in a series of related studies of Hamiltonian dynamics of small atomic and molecular clusters [11–21] it has also been found that depending on the steepness of the saddle, the dynamics of transitions through the saddle can either be more regular (flat saddles) [11–20] or more chaotic (sharper saddles) [12,13,20,21].

A key element in the dynamics of any reacting system is overcoming the transition state barrier, which determines the rate of the reaction. In clusters, at least, in small ones, the transition state is represented by a saddle on the potential energy surface that separates one conformation (reactant) from another (product). In this case, the use of the Hamiltonian dynamics is appropriate, with a number of methods available to quantify the degree of the regularity of the dynamics, such as the Lyapunov exponents, local Kolmogorov entropy, and invariants of motion [11–23]. In protein folding, the protein states are conventionally projected onto a reduced space of collective variables [1–6] to form the FES, which, in contrast to the potential energy surface governing the cluster dynamics, has a statistical nature. In particular, the transition state is then represented by a free energy barrier which results from the interplay between the potential energy, directing the system toward the native state, and entropy, leading it in the opposite direction, i.e., toward a variety of less compact conformations. According to the statistical nature of the FES, it seems appropriate to use a statistical description of the protein folding dynamics as well. In particular, the probability fluxes of transitions in a reduced conformational space of the protein can be employed for this purpose [7,24–27]. Given a flux distribution, the regularity of fluxes can be evaluated by calculating the fractal dimension of the fluxes [8–10]. Specifically, if the fluxes are mostly regular, the fractal dimension should be close to the Euclidean dimension of the flow field cross section, and if the fluxes are sufficiently irregular, the fractal dimension should be much less than that Euclidean dimension.

In order to determine where the protein folding dynamics are regular and where they are irregular, we study systematically how the probability fluxes change in the course of protein folding and calculate their fractal dimension. Overall, five different proteins which undergo a cooperative, two-state

[*] chekmarev@itp.nsc.ru

folding transition have been considered, ranging in representation from coarse-grained to all-atom models with implicit and explicit solvents (a model α-helical protein, β hairpin, and a mutant of villin headpiece subdomain in the main text, and Trp-cage miniprotein and ubiquitin in the Supplemental Material [28]). It has been found that, in an ideal case, the phases of regular and irregular dynamics alternate as follows. In the initial (collapse) phase of folding, the probability fluxes, and thus the dynamics, are essentially regular. Then, as the protein comes to the basin of semicompact states that precedes the transition state, the fluxes become irregular, but at the transition state they are regularized again. After the transition state, when the basin of nativelike states is explored, the fluxes become less regular and, finally, as the native state is approached, the fluxes are regularized. Depending on the specific conditions at which the protein folding was considered, some phases of the dynamics could not be well resolved, but no significant deviation from this ideal picture has been observed.

The paper is organized as follows. Section II describes the study of a model protein: the system and simulation method (Sec. II A), collective variables and free energy surface (Sec. II B), the calculation of probability fluxes (Sec. II C), and the picture of folding and fractal dimension of probability fluxes (Sec. II D). Section III presents the results for all-atom protein models: β hairpin (Sec. III A) and a mutant of the 35-residue villin headpiece subdomain (Sec. III B). Section IV discusses the relation between protein and cluster dynamics and contains some concluding remarks.

II. MODEL PROTEIN

A. System and simulation method

To perform a detailed study of the evolution of probability fluxes, a model protein was constructed, with which all folding phases of interest were well expressed and separated. Specifically, a 35-residue villin headpiece subdomain (1wy4.pdb [29]), whose native state consists of three α helices, was used for this purpose. To have converged results at a reasonable computation cost, a coarse-grained representation of the protein in the framework of a C_α model was employed. Two C_α beads were considered to be in native contact if they were not nearest neighbors along the protein chain and had the interbead distance not longer than $d_{\text{cut}} = 7.8$ Å. This value of d_{cut} was sufficient for the correct formation of the native structure of the protein, i.e., the α helices and their mutual disposition. To govern the protein dynamics, a Gō-like potential [30] was used, which accounted for the rigidity of the backbone and the contributions of native and non-native contacts in the form of the Lennard-Jones potential [31]. The parametrization was the same as in that work [31] except for the above mentioned value of d_{cut}. The simulations were performed with a constant-temperature molecular dynamics (MD) based on the coupled set of Langevin equations [32]. The time step $\Delta t = 0.0125\tau$ and the friction constant $\gamma = 3m/\tau$ were employed, where τ is the characteristic time. At the length scale $l = 7.8$ Å and the attractive energy $\epsilon = 2.2$ kcal/mol, $\tau = (Ml^2/\epsilon)^{1/2} \approx 2.7$ ps, where $M = 110$ Da is the average mass of the residue.

Folding trajectories were initiated in an unfolded state of the protein and terminated upon reaching the native state, i.e., we considered the case of "the first-passage folding," which is expected to mimic physiological conditions in that the native state is stable [9]. More specifically, the initial states were such that the α helices were partially formed but the contacts between the helices were absent. The native state was considered to be reached when the root-mean-square-deviation (RMSD) from the native structure (σ_{nat}) was 2.5 Å or less. To have the FES with the transition state that clearly separates catchment basins for semicompact and nativelike conformations, the simulations were performed at a temperature as low as $T = 0.05$ in units of the attractive energy (with the Boltzmann constant set to unity). In total, 50 000 folding trajectories have been run. At the given temperature, the folding kinetics were close to two-state kinetics [28].

B. Collective variables and free energy surface

In contrast to the previous works [8–10], where the fractal dimension of probability fluxes was calculated in a three-dimensional space of collective variables, we consider the fluxes in a two-dimensional space of variables. In this case, the flow streamlines can be determined and superimposed on the FES, which makes the process of folding more representative. The choice of two collective variables is not straightforward because the reduction of a multidimensional conformational space with, e.g., a principal component analysis (PCA) method [33] generally does not offer two modes that would be well separated from the others and covered a dominant fraction of the data variance. This is characteristic of folding of this and other proteins we study [28]. Therefore, it was found reasonable to characterize the initial conformational space by two traditional collective variables. Since the subsequent analysis required the probability fluxes to be determined in orthogonal space, the initial valuables needed to be converted to orthogonal ones. Accordingly, the initial variables should have to be of the same dimension. Among possible pairs of such variables, the most representative are two pairs: the number of the total and the native contacts, and the radius of gyration and the RMSD from the native state. These pairs are similar in that the first variable in each pair determines how the protein compacts and the second one shows how the protein approaches the native state. Although the number of native contacts could possibly be more informative than the RMSD from the native state [34] (see the discussion of this issue in Sec. III B), the second pair of variables was chosen to represent the initial conformation space of the protein, i.e., the radius of gyration (R_g) and the RMSD from the native state (σ_{nat}), mostly because these variables are not discrete, which allowed us to use a fine, adjusted grid for accurate calculation of probability flows.

Figure 1 shows the FES as a function of σ_{nat} and R_g. The free energy was calculated as $F(\sigma_{\text{nat}}, R_g) = -T \ln P(\sigma_{\text{nat}}, R_g)$, where $P(\sigma_{\text{nat}}, R_g)$ is the probability to find the system at the point $(\sigma_{\text{nat}}, R_g)$. We note that this thermodynamic relation is of limited application in the case of first-passage simulations [35] because detailed balance does not hold [36] (see also Refs. [37,38] for a discussion of related questions). The FES thus determined is not the true FES

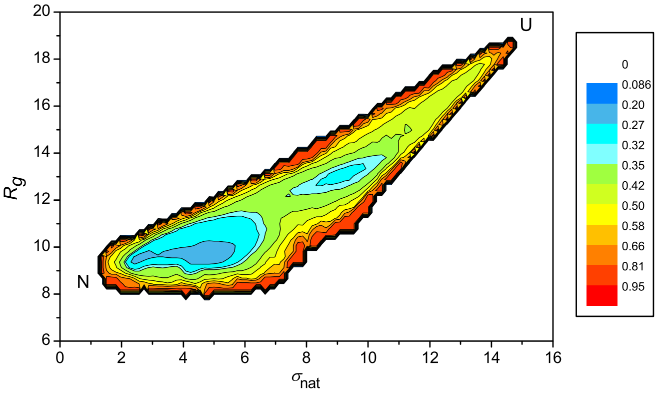

FIG. 1. Free energy surface for the model protein as a function of the RMSD from the native states ($\sigma_{\rm nat}$) and the radius of gyration ($R_{\rm g}$), both measured in angstroms. Labels U and N indicate, respectively, the unfolded states, where folding trajectories were initiated, and the native state, where they were terminated.

but rather is a means to represent the probability distribution $P(\sigma_{\rm nat}, R_{\rm g})$ in a way typical of protein folding studies. It is seen that the FES contains two basins separated by a well expressed free energy barrier; one basin is related to partially folded (semicompact) states of the protein (larger values of $\sigma_{\rm nat}$), and the other to nativelike states (smaller values of $\sigma_{\rm nat}$). To proceed further, i.e., to calculate the probability fluxes and analyze them, the space of the variables $\sigma_{\rm nat}$ and $R_{\rm g}$, which are not independent, was transformed into the space of orthogonal variables $\mathbf{g} = (g_1, g_2)$ with the PCA method. The corresponding $F(g_1, g_2)$ surface is shown in Fig. 2(a). It retains all characteristic features of the $F(\sigma_{\rm nat}, R_{\rm g})$ surface. Since the PCA is a linear transformation, the variables g_1 and g_2 are measured in the same units as the initial variables, specifically, in angstroms.

C. Probability fluxes

To determine the probability fluxes in the $\mathbf{g}$ space, the hydrodynamic description of the folding reaction was used [7]. The g_1 component of the flux at a point $\mathbf{g}$ was determined as

$$j_{g_1}(\mathbf{g}) = \left[\sum_{\mathbf{g}',\mathbf{g}''(\mathbf{g}\subset\mathbf{g}^*)}^{g_1''-g_1'>0} n(\mathbf{g}'',\mathbf{g}') - \sum_{\mathbf{g}',\mathbf{g}''(\mathbf{g}\subset\mathbf{g}^*)}^{g_1''-g_1'<0} n(\mathbf{g}'',\mathbf{g}')\right] \Big/ (N\bar{t}_{\rm f}\Delta g_2), \qquad (1)$$

where N is the number of folding trajectories, $\bar{t}_{\rm f}$ is the mean first-passage time (MFPT), $n(\mathbf{g}'', \mathbf{g}')$ is the number of transitions from state $\mathbf{g}'$ to $\mathbf{g}''$, and $\mathbf{g} \subset \mathbf{g}^*$ is a symbolic designation of the condition that the transitions included in the sum have the straight line connecting points $\mathbf{g}'$ to $\mathbf{g}''$ that crosses the line $g_1 = \text{const}$ within the segment of the length of Δg_2 centered at the point $\mathbf{g}$. The g_2 component of $\mathbf{j}(\mathbf{g})$ was determined in a similar way, except that the transitions crossing the line $g_2 = \text{const}$ within the segment of the length of Δg_1 were selected. To avoid nonphysical conformations at interpolation between two neighboring structures (points $\mathbf{g}'$ and $\mathbf{g}''$), the discretization of the (g_1, g_2) space should match the difference between the structures. In the present case, the average RMSD between neighboring structures ($\sigma_{\rm nbr}$) was equal to 0.035 Å, and the calculations were performed on a grid with $\Delta g_1 = \Delta g_2 = 0.035$ Å. The knowledge of the $\mathbf{j}(\mathbf{g})$ makes it possible to determine the streamlines of folding flows, which are tangent to the local directions of $\mathbf{j}(\mathbf{g})$. Each streamline corresponds to a constant value of the stream function $\Psi(g_1, g_2) = \int_{y=0}^{y=g_2} j_{g_1}(g_1, y)dy$. Two streamlines with $\Psi(g_1, g_2) = C_1$ and $\Psi(g_1, g_2) = C_2$, where C_1 and C_2 are constant such that $C_2 > C_1$, create a stream tube which contains a fraction $(C_2 - C_1)/G$ of the total flow G.

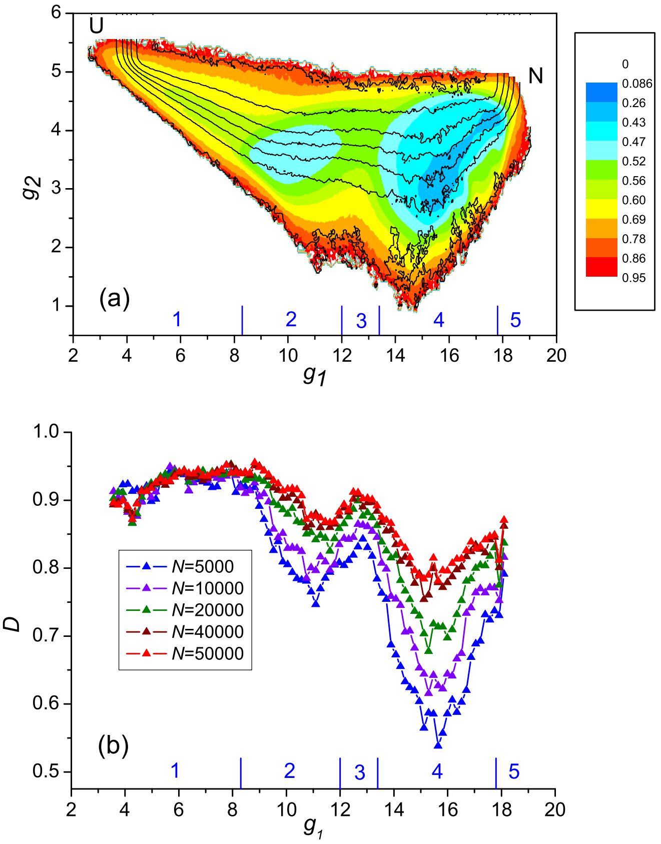

FIG. 2. (a) Free energy surface for the model protein as a function of the PCA variables g_1 and g_2. The black lines depict the streamlines of the folding flow corresponding to 0.001, 0.1, 0.3, 0.5, 0.7, 0.9, and 0.999 fractions of the total flow (from the lower to the upper line). The calculations are based on 5×10^3 folding trajectories. (b) Variation of the fractal dimension D with g_1. The triangles present the simulation results after averaging over five neighboring points along g_1, and the corresponding solid lines are to guide the eye. The labels and lines of different color correspond to the number folding trajectories indicated in the inset. The blue bars at the g_1 axis mark approximate boundaries of the characteristic phases of folding, which are indexed from 1 to 5: an almost uniform flow (1), a basin of semicompact states (2), the transition state (3), a basin of nativelike states (4), and the native state (5).

D. Picture of folding and fractal dimension of probability fluxes

The calculated folding flow streamlines are shown in Fig. 2(a) as the lines superimposed on the FES. The streamlines are in agreement with those for the previously studied α-helical hairpin [7] in that they are regular in the initial and termination phases of folding. A new essential feature of Fig. 2(a) is that the streamlines are regularized when the system crosses the transition state, similar to what was previously observed for the Hamiltonian dynamics of clusters of a few atoms [11–20] (see Sec. IV for a discussion of this). It should be noted that the regularity of the fluxes determined by Eq. (1) depends on the number of folding trajectories on the basis of which the fluxes are calculated, i.e., as the number of folding trajectories N increases, random fluctuations of the fluxes coming from different trajectories cancel each other to make flux distributions more regular. The same is for the flow streamlines. However, such a dependence of fluxes and streamlines on the number of trajectories affects only the fluctuations they are subject to but not their behavior in average [28].

To gain a closer insight into the folding dynamics, we consider the spatial distribution of probability fluxes. Since the total flow is generally directed along the g_1 axis, which plays a role of the overall reaction coordinate, the g_1 component of the fluxes, $j_{g_1}(\mathbf{g})$, is of most relevance here. To characterize the change of the probability fluxes along g_1, we introduce the function $G(g_1, L)$, which represents the average ratio of the probability flow through a segment of the g_2 axis of the length L to the average value of probability fluxes within this segment. Specifically, $G(g_1, L)$ was defined as $G(g_1, L) = \{\sum_{i=1}^{i=M}[J_{g1,i}(g_1, L)/\bar{j}_{g_1}(g_1, L)]^2/M\}^{1/2}$, where $J_{g1,i}(g_1, L)$ is the flow through the ith segment of length L, M is the number of the segments covering the cross section of the flow field at the given g_1, and $\bar{j}_{g_1}(g_1, L) = [\sum_{k=1}^{k=L} j^2_{g_1,k}(g_1)/L]^{1/2}$ is the average value of the flux within the current segment of length L (the summation index k denotes the current point of the grid along the g_2 axis). The length of the segments L is measured in the units of the number of the elementary segments. Although the results do not depend significantly on the maximum value of L, it was chosen to be not larger than 5 segments, as a compromise between the length of the segments and the convergence of the results (a larger value of L would lead to a smaller number of segments M, and thus to poorer statistics). Figure 3 shows the calculated values of $G(g_1, L)$ for the characteristic regions of the FES. The best fit of the data to the equation $G(g_1, L) \sim L^{D(g_1)}$ reveals that in all cases the flow distributions are self-similar with respect to the length of coarse-graining L. Since the exponent $D(g_1)$ is mostly smaller than unity, which is characteristic of a uniform flow, it follows that the distributions of the probability fluxes have a fractal nature [39].

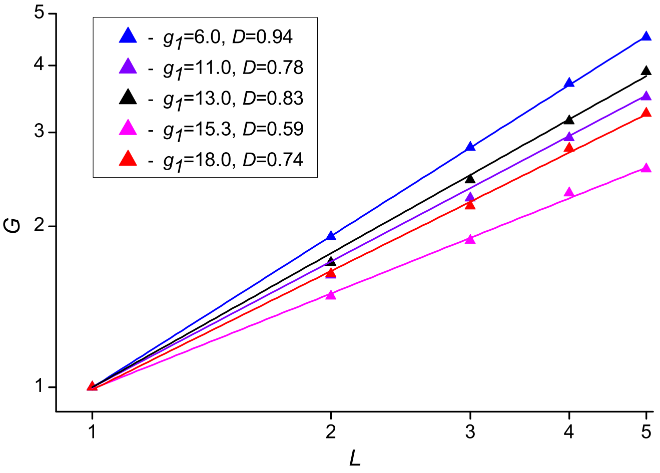

FIG. 3. $G(g_1, L)$ function for the model protein in the characteristic regions of the FES: the collapse phase ($g_1 = 6.0$), semicompact states ($g_1 = 11.0$), the transition state ($g_1 = 13.0$), nativelike states ($g_1 = 15.3$), and the approach to the native state ($g_1 = 18.0$). Labels correspond to the calculated values of $G(g_1, L)$, and the lines show the best fit of the data to the equation $G(g_1, L) \sim L^{D(g_1)}$. The calculations are based on 5×10^3 folding trajectories; cf. Fig. 2(b).

Figure 2(b) presents a detailed variation of the fractal dimension D with g_1. There are shown the results of calculation of D on the basis of several sets of folding trajectories increasing in number (N) from 5×10^3 to 5×10^4. The variations of D with g_1 for different N are similar, except that the convergence of the results is not uniform, i.e., the smaller the fractal dimension, the larger the number of trajectories required to reach convergence. Figure 4 shows the standard deviation of the g_1 component of the fluxes along the g_1 axis that is reduced by N and by $N^{1/2}$, panels (a) and (b), respectively. The standard deviation was calculated as $\sigma(g_1) = \{\int [j_{g_1}(g_1, g_2) - \bar{j}_{g_1}(g_1)]^2 dg_2/A(g_1)\}^{1/2}$, where $\bar{j}_{g_1}(g_1) = \int j_{g_1}(g_1, g_2) dg_2/A(g_1)$ is the current average value of j_{g_1} and $A(g_1)$ is the current cross section of the flow field. Please note that, in the protein collapse region (phase 1), σ changes along g_1 axis not because the fluxes are irregular in this region but because the streamlines are not parallel to the g_1 axis due to a ballistic shift in the g_2 direction [see Fig. 2(a)]. Figure 4 reveals that the standard deviation of the fluxes, and thus the fluxes themselves, scales linearly with N in the protein collapse region, which indicates that the motion is regular here. In contrast, in the basin of nativelike state (centered at $g_1 \approx 16$) it scales approximately as $\sim N^{1/2}$, which indicates that the motion is mostly random (diffusivelike). The present character of convergence of the results with the number of folding trajectories reinforces the conclusion that can be drawn from the comparison of the regularity of folding flow streamlines and the variation of the fractal dimension (Fig. 2), i.e., that the degree of fractality of probability fluxes is coordinated with the degree of regularity of the fluxes so that the higher the regularity of the fluxes, the higher the fractal dimension. Figure 2(b) thus suggests that, in the initial phase of folding (phase 1), where the fractal dimension is close to unity, the probability fluxes are essentially uniform and hence the dynamics of protein collapse are regular. As the protein comes to the basin of semicompact states (2), the fractal dimension decreases, indicating that the dynamics become irregular. However, as the transition state is approached (3), the dynamics are regularized again. Further, in the basin of nativelike states (4), irregularity of the flows increases until the protein finds a way to the native state (5) and proceeds in a more regular manner.

III. ALL-ATOM PROTEIN MODELS

More realistic, all-atom simulations for a 12-residue β-hairpin protein [40] and the Nle/Nle double mutant of the

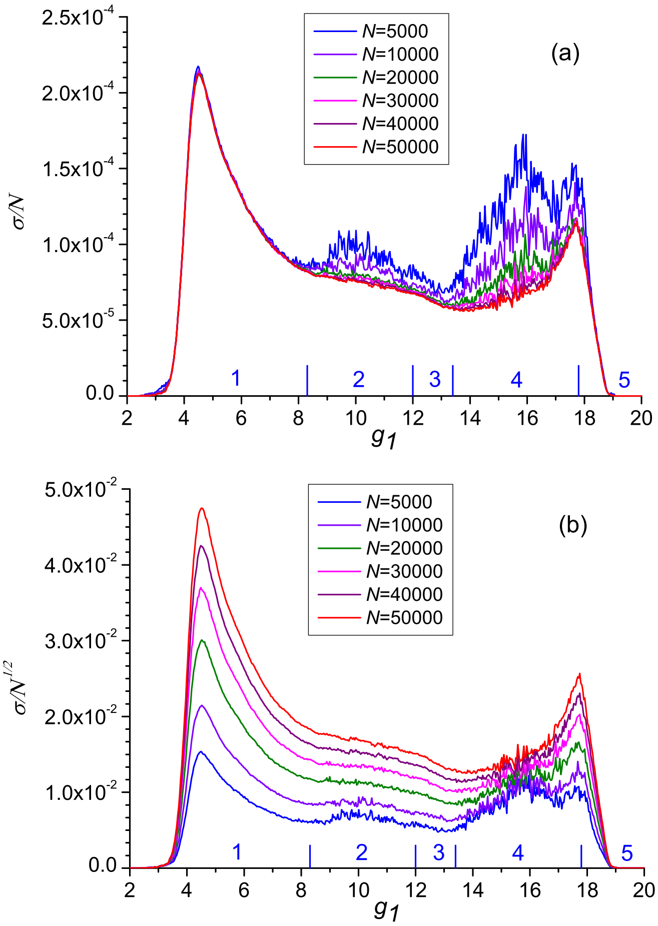

FIG. 4. Standard deviation of the g_1 component of the fluxes along the g_1 axis: (a) reduced by N and (b) reduced by $N^{1/2}$, where N is the number of folding trajectories. The blue bars at the g_1 axis mark approximate boundaries of the characteristic phases of folding: an almost uniform flow (1), a basin of semicompact states (2), the transition state (3), a basin of nativelike states (4), and the native state (5).

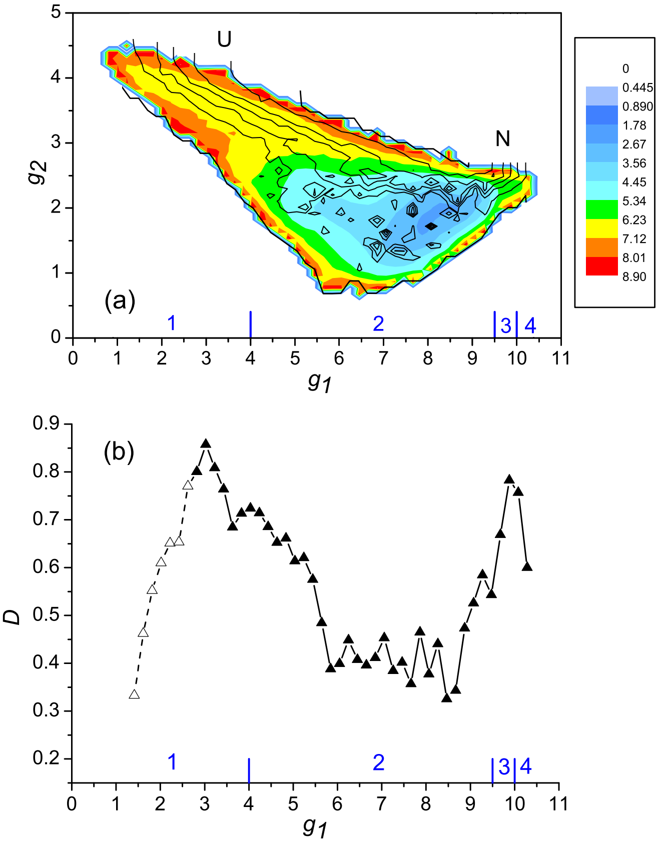

FIG. 5. (a) Free energy surface for β hairpin as a function of the PCA variables g_1 and g_2. The black lines depict the folding flow streamlines corresponding to the 0.001, 0.1, 0.3, 0.5, 0.7, 0.9, and 0.999 fractions of the total flow (from the lower to the upper line). (b) Variation of the fractal dimension D with g_1. The triangles present the simulation results after running averaging over three points, and the corresponding solid lines are to guide the eye. The solid triangles are for the region where the flow is larger by 70% of the total flow (Supplemental Material Fig. S10 [28]). The blue bars at the g_1 axis mark approximate boundaries of the characteristic phases of folding, which are indexed from 1 to 4: a collapse phase (1), a basin of semicompact states (2), the transition state (3), and the native state basin (4).

35-residue villin headpiece subdomain (HP-35 NleNle) [41] show a similar alternation of phases of regular and irregular dynamics (Figs. 5 and 6, respectively), although some phases of the dynamics are poorly resolved. The reason is twofold: (i) a restricted convergence of the simulation results due to a limited number of folding trajectories and (ii) a delocalization of the sources of folding flows due to a diversity of the initial (unfolded) states of the protein [28]. In both the proteins, the PCA reduction of the conformational space in the form of native bond distances did not yield two well-separated largest modes [28]. Therefore, similar to the model protein, the $(\sigma_{\text{nat}}, R_{\text{g}})$ space was used to introduce the (g_1, g_2) space of PCA variables (here and below, σ_{nat} and R_{g} are calculated for C_α atoms) [28].

A. β hairpin

Folding of the β hairpin (KTWNPATGKWTG; 2evq.pdb) [40] was simulated with the CHARMM program [42]. All heavy atoms and the hydrogen atoms bound to nitrogen or oxygen atoms were considered explicitly; PARAM19 force field [43] and a default cutoff of 7.5 Å for the nonbonding interactions were used. To account for the effects of aqueous solvent, a solvent-accessible surface-area (SASA) approximation [44] was employed, which has proved to be successful for β-sheet proteins. The temperature was controlled using the Berendsen thermostat with a coupling constant of 5 ps. The SHAKE algorithm [45] was applied to fix the length of the covalent bonds involving hydrogen atoms, which allowed the integration time step of 2 fs. The MD trajectories were initiated in unfolded states of the protein and terminated upon reaching the native state. The unfolded states were prepared using the standard CHARMM protocol [42]; i.e., an extended conformation of the protein was first minimized and then heated and equilibrated at the temperature of interest (for 5×10^3 time steps). The native state was considered to be reached as the C_α-RMSD from the native state σ_{nat} was <1 Å. One hundred folding trajectories were generated at $T = 330$ K. Aiming at

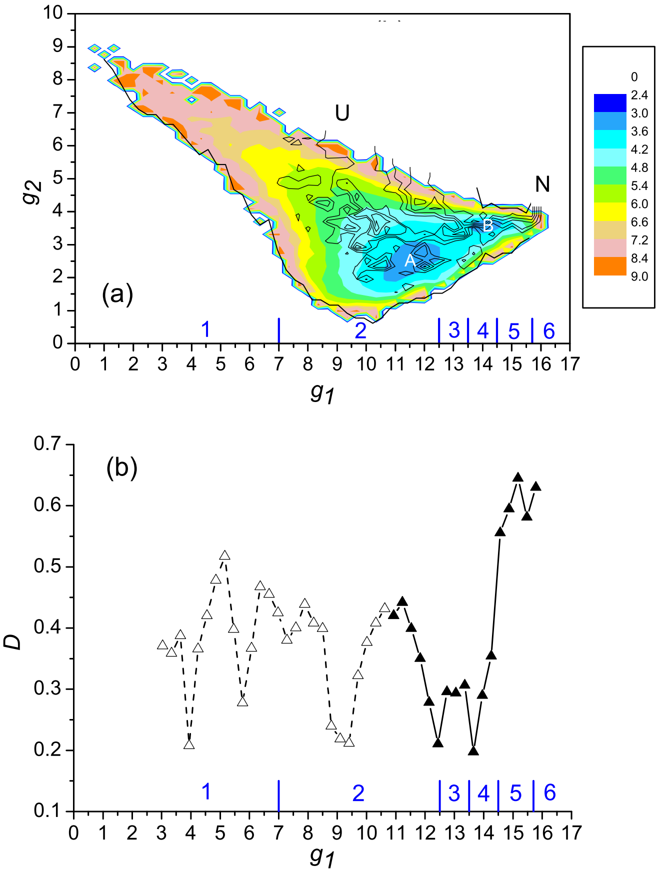

FIG. 6. (a) Free energy surface for HP-35 NleNle as a function of the PCA variables g_1 and g_2. The black lines show the folding flow streamlines with the 0.001, 0.1, 0.3, 0.5, 0.7, 0.9, and 0.999 fractions of the total flow (from the lower to the upper line). (b) Variation of the fractal dimension D with g_1. The triangles show the simulation results (averaged as in Fig. 5), and the corresponding solid lines are to guide the eye. The solid triangles are for the region where the flow is larger by 70% of the total flow (Supplemental Material Fig. S17 [28]). The blue bars at the g_1 axis mark approximate boundaries of the characteristic phases of folding, which are indexed from 1 to 6: a collapse phase (1), a part of the basin of semicompact states that includes the sub-basin A (2), the transition state between sub-basins A and B (3), the sub-basin B (4), the transition to the native state (5), and the native state (6).

a better resolution of fluxes in the transition state region, the atomic coordinates ("frames") were saved every 1 ps. Since the average values of the C_α-RMSD between the neighboring structures (σ_{nbr}) was relatively large (≈0.86 Å), a rather rough discretization of the (g_1, g_2) space was chosen ($\Delta g_1 \approx 0.2$ Å and $\Delta g_2 \approx 0.1$ Å).

Figure 5 presents the results. The general picture of folding [Fig. 5(a)] is similar to those for the equilibrium folding of short β hairpins [46–48], except that the basin of nativelike states is too small due to termination of folding trajectories in the native state. The variation of the fractal dimension with the reaction coordinate g_1 [Fig. 5(b)] shows that the dynamics becomes less regular as the system comes to the basin of semicompact states (phase 2) but is regularized at the transition state (phase 3).

B. HP-35 NleNle

The calculations were performed on the basis of the MD trajectories for folding of the HP-35 NleNle in explicit solvent [41]. The 395 μs equilibrium trajectory at 370 K was used to extract 74 first-passage trajectories. The points that are most distant from the native state, with $\sigma_{\mathrm{nat}} > 13$ Å, were taken to represent the unfolded states, and the native state was determined as the first point in the current segment of the equilibrium trajectory where the condition $\sigma_{\mathrm{nat}} < 1$ Å was satisfied. With the frames saved every 200 ps, σ_{nbr} was ≈ 2.36 Å, so that a rougher (g_1, g_2) grid, with $\Delta g_1 \approx 0.3$ Å and $\Delta g_2 \approx 0.15$ Å, was employed. It may be noted that due to subdiffusivity of folding dynamics [49], σ_{nbr} increases with the time interval between the frames much slower than the interval itself [28].

Figure 6 shows that, similar to the two previous proteins, the dynamics are irregular in the basin of semicompact states (phase 2) but becomes more regular at the transition to the native state (5). It is noteworthy that the transition state between sub-basins A and B (phase 3) also reveals an enhanced regularity of the dynamics.

As has been indicated, a poorer resolution of some protein states and phases of folding in the present all-atom simulations, as compared to the coarse-grained simulations for the model protein, can be a consequence of the limited number of folding trajectories. However, it cannot be ruled out that an inappropriate choice of initial collective variables to characterize the conformational space of the protein, i.e., the RMSD from the native state and radius of gyration we used, is the reason. It is thus of interest to see what is obtained with the other possible pair of initial collective variables (Sec. II B), which are the numbers of native and total contacts, or, what is the same, the numbers of native and nonnative contacts, with the latter determined as the difference between the total number of contacts and that of native contacts [28]. After transition to the orthogonal (PCA) variables, the results have been found very similar to those in the present case. In particular, in agreement with Fig. 6(a), two sub-basins separated by relatively low transition state are observed in the FES [28]. Also, in agreement with Fig. 6(b), the fractal index in the transition state region and in the vicinity of the native state is found to be larger than in the sub-basins, i.e., the dynamics is regularized at the transition state and when the protein approaches the native state [28].

IV. CONCLUDING DISCUSSION

In order to determine the degree of regularity of dynamics in different phases of protein folding, the dimension of probability fluxes was calculated. The fluxes were found self-similar and had a fractal dimension that varied along the reaction coordinate. Several proteins which undergo a cooperative, two-state folding transition have been studied, i.e., a model helical protein, β hairpin, and HP-35 NleNle (the main text), and Trp-cage miniprotein and ubiquitin [28]. It has been found that, in an ideal case, the phases of regular and irregular dynamics alternate as follows. In the initial (collapse) phase of folding, the probability fluxes, and thus the dynamics, are essentially regular. Then, as the protein comes to the basin of semicompact states that precedes the transition state, the fluxes become irregular, but at the transition state they are regularized again. After the transition state, in the basin for nativelike states, irregularity of the fluxes increases, and finally, as the native-state is approached, the fluxes become regular. Depending on the specific conditions at which the protein folding was considered, some phases of the folding dynamics could not be well resolved, but no significant deviation from this ideal picture has been observed.

As indicated in the Introduction, the regularization of system dynamics at the transition state has previously been observed in the studies of Lennard-Jones and Morse clusters of a few atoms [11–21] (the microcanonical ensembles of the trajectories on the potential energy surfaces were investigated). At the same time, the subsequent studies of small water clusters, described by a specific H_2O potential [50], have shown that the dynamics at the transition state can be less regular than in the preceding basin [20,21]. This has confirmed the anticipation that the regularity of dynamics depends on the transition state flatness [12,13], i.e., at flat transition states, as for the atomic clusters, the dynamics are mostly regular, and at sharp transition states, as for the water clusters, they are irregular, because the system has no sufficient time to regularize the dynamics [13,20]. To measure the degree of irregularity of the dynamics, the local Lyapunov exponents and Kolmogorov entropy were employed. Accordingly, the irregular motion was interpreted as chaotic dynamics in the conventional definition [39], i.e., as a motion highly sensitive to the initial conditions. For the present statistical characterization of dynamics, the existence of such a direct connection between the fractality of the probability fluxes and the chaoticity of the dynamics remains an open question, although it cannot be ruled out that such a connection exists. One encouraging example is the Kaplan-Yorke conjecture, in which the fractal dimension of the attractor of a dynamical system is determined in terms of the Lyapunov exponents [51], particularly, a heuristic derivation of such a relation [52]. However, for the time being, in order to consider the observed fractality of the probability fluxes to be a consequence of chaotic dynamics, we can only rely on arguments of general character, such that the protein folding dynamics are inherently chaotic [8,23,53–57], i.e., due to the sensitivity to initial conditions, the protein motion becomes unpredictable on large time scales [53,54,56], and it is accompanied by the formation of fractal [8] and chaotic [57] attractors. A more specific, though limited, argument is that the fractal dimension is well correlated with the character of convergence of the probability fluxes, i.e., the closer the process of convergence to a random process, the lower the fractal dimension (Sec. II D), which indicates that the fractal nature of the probability fluxes is not a result of complex but regular motion.

In order to see whether a similar connection between the dynamics and the flatness of the transition states, which was observed in the cluster studies [11–21], is characteristic of protein folding, the free energy profiles along the g_1 axis were calculated (Fig. 7). Comparison with the corresponding Figs. 2, 5, and 6 shows that the intermediate transition states occurring in the model protein and HP-35 NleNle, where the dynamics are more regular than in the preceding basins, are relatively flat. At the same time, at approaching the transition state preceding the native state, where the dynamics are more regular as well, the free energy rises steeply in each case (although the transition state itself is not properly resolved because, due to the termination of folding trajectories in the native state, the native state basin is too small and shallow). Thus, in contrast to the clusters, the present results suggest that the protein folding dynamics are always regularized as the protein approaches the transition state, irrespective of the degree of flatness of the underlying energy surface. The reasons why the protein dynamics we observed here differ from the cluster dynamics may range from the difference in the approaches (the Langevin versus Hamiltonian dynamics, the statistical folding fluxes versus phase-space cluster trajectories, etc.) to a simple possibility that a protein which is characterized by irregular dynamics at the transition state has not occurred (this, however, seems unlikely, taking into account how the considered proteins varied in structure and mechanism of folding). To resolve this issue, a systematic study of the dynamics, preferably of the same system, at the surfaces of varying steepness is probably required. In this context, it may be noted that because both the folding dynamics and underlying (free) energy surfaces were characterized statistically, the present, statistical approach is self-consistent and seems to be practical for application to the systems of large size such as proteins.

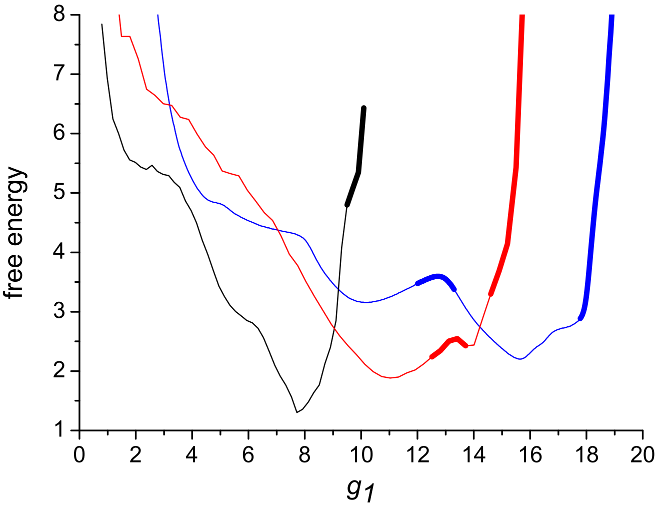

FIG. 7. Free energy profiles along g_1. The blue line is for the model protein, black line is for β hairpin, and red line is for HP-35 NleNle. The segments of the profiles shown in the thick lines indicate the regions of the FESs that correspond to either the transition states, as the inner segments of the curves for the model protein and HP-35 NleNle, or to the approach to the native states, as the terminal segments of the curves (cf. Figs. 2, 5, and 6).

ACKNOWLEDGMENTS

I am grateful to M. Karplus and R. S. Berry for stimulating discussions, and D. E. Shaw Research for sharing their folding trajectories for the villin headpiece subdomain (HP-35 NleNle). Support from the Ministry of Education and Science of the Russian Federation is acknowledged.

[1] A. R. Dinner, A. Šali, L. J. Smith, C. M. Dobson, and M. Karplus, Trends Biochem. Sci. **25**, 331 (2000).

[2] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).

[3] C. M. Dobson, A. Šali, and M. Karplus, Angew. Chem. Int. Ed. **37**, 868 (1998).

[4] W. A. Eaton, V. Muñoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter, Annu. Rev. Biophys. Biomol. Struct. **29**, 327 (2000).

[5] J.-E. Shea and C. L. Brooks III, Annu. Rev. Phys. Chem. **52**, 499 (2001).

[6] K. A. Dill, S. B. Ozkan, M. S. Shell, and T. R. Weikl, Annu. Rev. Biophys. **37**, 289 (2008).

[7] S. F. Chekmarev, A. Yu. Palyanov, and M. Karplus, Phys. Rev. Lett. **100**, 018107 (2008).

[8] I. V. Kalgin and S. F. Chekmarev, Phys. Rev. E **83**, 011920 (2011).

[9] I. V. Kalgin, S. F. Chekmarev, and M. Karplus, J. Phys. Chem. B **118**, 4287 (2014).

[10] V. A. Andryushchenko and S. F. Chekmarev, PLoS One **12**, e0188659 (2017).

[11] T. L. Beck, D. M. Leitner, and R. S. Berry, J. Chem. Phys. **89**, 1681 (1988).

[12] R. J. Hinde, R. S. Berry, and D. J. Wales, J. Chem. Phys. **96**, 1376 (1992).

[13] R. J. Hinde and R. S. Berry, J. Chem. Phys. **99**, 2942 (1993).

[14] C. Amitrano and R. S. Berry, Phys. Rev. E **47**, 3158 (1993).

[15] T. Komatsuzaki and R. S. Berry, J. Chem. Phys. **110**, 9160 (1999).

[16] T. Komatsuzaki and R. S. Berry, Phys. Chem. Chem. Phys. **1**, 1387 (1999).

[17] T. Komatsuzaki and R. S. Berry, Adv. Chem. Phys. **123**, 79 (2002).

[18] J. R. Green, J. Jellinek, and R. S. Berry, Phys. Rev. E **80**, 066205 (2009).

[19] J. C. Lorquet, J. Phys. Chem. A **115**, 4610 (2011).

[20] J. R. Green, T. S. Hofer, D. J. Wales, and R. S. Berry, Mol. Phys. **110**, 1839 (2012).

[21] J. R. Green, T. S. Hofer, D. J. Wales, and R. S. Berry, J. Chem. Phys. **135**, 184307 (2011).

[22] Y. Matsunaga, C.-B. Li, and T. Komatsuzaki, Phys. Rev. Lett. **99**, 238103 (2007).

[23] D. Nerukh, G. Karvounis, and R. C. Glen, in *CompLife 2006*, edited by M. R. Berthold, R. Glen, and I. Fischer, Lecture Notes in Bioinformatics Vol. 4216 (Springer-Verlag, Berlin-Heidelberg, 2006), pp. 129–140.

[24] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, Proc. Natl. Acad. Sci. USA **106**, 19011 (2009).

[25] A. Berezhkovskii, G. Hummer, and A. Szabo, J. Chem. Phys. **130**, 205102 (2009).

[26] W. E and E. Vanden-Eijnden, Annu. Rev. Phys. Chem. **61**, 391 (2010).

[27] T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande, J. Am. Chem. Soc. **133**, 18413 (2011).

[28] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.99.022412 for the distributions of the first-passage times, the FESs depending on the initial collective variables, the eigenvalues of covariance matrices, the sources and sinks of the folding flows, the total flows from the unfolded to the native states, and the approximation of the fractal dimension—all for the model protein, β-hairpin protein, and HP-35 NleNle with two variants of initial collective variables. Also, there are shown the FESs and variations of the fractal dimension for Trp-cage miniprotein and ubiquitin.

[29] T. K. Chiu, J. Kubelka, R. Herbst-Irmer, W. A. Eaton, J. Hofrichter, and D. R. Davies, Proc. Natl. Acad. Sci. USA **102**, 7517 (2005).

[30] N. Gō, Annu. Rev. Biophys. Bioeng. **12**, 183 (1983).

[31] T. X. Hoang and M. Cieplak, J. Chem. Phys. **112**, 6851 (2000).

[32] R. Biswas and D. R. Hamann, Phys. Rev. B **34**, 895 (1986).

[33] I. Jolliffe, *Principal Component Analysis* (Springer Verlag, New York, 2002).

[34] R. B. Best, G. Hummer, and W. A. Eaton, Proc. Natl. Acad. Sci. USA **110**, 17874 (2013).

[35] S. F. Chekmarev, S. V. Krivov, and M. Karplus, J. Phys. Chem. B **109**, 5312 (2005).

[36] S. F. Chekmarev, J. Chem. Phys. **139**, 145103 (2013).

[37] J. Wang, K. Zhang, L. Xu, and E. Wang, Proc. Natl. Acad. Sci. USA **108**, 8257 (2011).

[38] L. Xu, F. Zhang, E. Wang, and J. Wang, Nonlinearity **26**, R69 (2013).

[39] F. C. Moon, *Chaotic and Fractal Dynamics* (Wiley, New York, 1992).

[40] N. H. Andersen, K. A. Olsen, R. M. Fesinmeyer, X. Tan, F. M. Hudson, L. A. Eidenschink, and S. R. Farazi, J. Am. Chem. Soc. **128**, 6101 (2006).

[41] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, Proc. Natl. Acad. Sci. USA **109**, 17845 (2012).

[42] B. R. Brooks, C. L. Brooks III, A. D. MacKerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, J. Comput. Chem. **30**, 1545 (2009).

[43] E. Neria, S. Fischer, and M. Karplus, J. Chem. Phys. **105**, 1902 (1996).

[44] P. Ferrara, J. Apostolakis, and A. Caflisch, Proteins: Struct. Funct. Bioinf. **46**, 24 (2002).

[45] J. P. Ryckaert, G. Ciccotti, and H. J. Berendsen, J. Comput. Phys. **23**, 327 (1997).

[46] A. R. Dinner, T. Lazaridis, and M. Karplus, Proc. Natl. Acad. Sci. USA **96**, 9068 (1999).

[47] R. Zhou, B. J. Berne, and R. Germain, Proc. Natl. Acad. Sci. USA **98**, 14931 (2004).

[48] S. V. Krivov and M. Karplus, Proc. Natl. Acad. Sci. USA **101**, 14766 (2004).

[49] Y. Meroz, V. Ovchinnikov, and M. Karplus, Phys. Rev. E **95**, 062403 (2017).

[50] T. S. Mahadevan and S. H. Garofalini, J. Phys. Chem. B **111**, 8919 (2007).

[51] J. Kaplan and J. Yorke, in *Functional Differential Equations and the Approximation of Fixed Points*, edited by H. O. Peitgen and H. O. Walther, Lecture Notes in Mathematics Vol. 730 (Springer, Berlin, 1978), p. 228–237.

[52] H. Mori, Prog. Theor. Phys. **63**, 1044 (1980).

[53] H.-B. Zhou and L. Wang, J. Phys. Chem. **100**, 8101 (1996).

[54] M. Braxenthaler, R. Unger, D. Auerbach, J. A. Given, and J. Moult, Proteins: Struct. Funct. Genet. **29**, 417 (1997).

[55] V. Villani, L. D'Alessio, and A. M. Tamburro, J. Chem. Soc. Perkin Trans. **2**, 2375 (1997).

[56] L. S. D. Caves, J. D. Evanseck, and M. Karplus, Prot. Sci. **7**, 649 (1998).

[57] V. Villani, J. Mol. Struct., Theochem. **621**, 127 (2003).