THE JOURNAL OF PHYSICAL CHEMISTRY B

ARTICLE

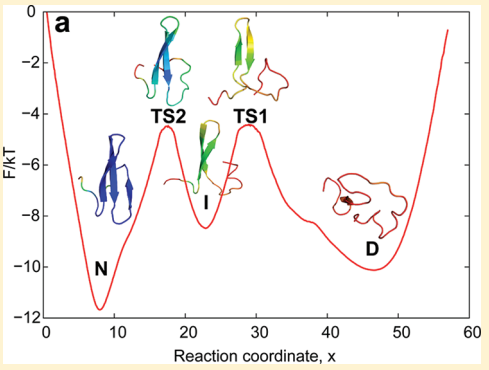
pubs.acs.org/JPCB

The Free Energy Landscape Analysis of Protein (FIP35) Folding Dynamics

Sergei V. Krivov*

Institute of Molecular and Cellular Biology, Leeds University, Leeds, United Kingdom

ABSTRACT: A fundamental problem in the analysis of protein folding and other complex reactions is the determination of the reaction free energy landscape. The current experimental techniques lack the necessary spatial and temporal resolution to construct such landscapes. The properties of the landscapes can be probed only indirectly. Simulation, assuming that it reproduces the experimental dynamics, can provide the necessary spatial and temporal resolution. It is, arguably, the only way for direct rigorous construction of the quantitatively accurate free energy landscapes. Here, such landscape is constructed from the equilibrium folding simulation of FIP35 protein reported by Shaw et al. *Science* **2010**, *330*, 341−346. For the dynamics to be accurately described as diffusion on the free energy landscape, the choice of reaction coordinates is crucial. The reaction coordinate used here is such that the dynamics projected on it is diffusive, so the description is consistent and accurate. The obtained landscape suggests an alternative interpretation of the simulation, markedly different from that of Shaw et al. In particular, FIP35 is not an incipient downhill folder, it folds via a populated on-pathway intermediate separated by high free energy barriers; the high free energy barriers rather than landscape roughness are a major determinant of the rates for conformational transitions; the preexponential factor of folding kinetics $1/k_0 \sim 10$ ns rather than 1 μs.



INTRODUCTION

The free energy landscapes are often used to represent the equilibrium and dynamic properties of a protein in a quantitatively accurate while intuitively clear way.[1,2] The free energy as a function of one or two reaction coordinates gives the equilibrium population $p(x) \sim \exp(-F(x)/kT)$ and together with the position dependent diffusion coefficient $D(x)$ describes the dynamics as diffusion.[3] In spite of their fundamental importance, the quantitatively accurate free energy landscapes of proteins are yet to be constructed. Direct determination of the free energy landscapes by the state of the art experimental techniques is hindered by limited spatial and temporal resolution. Various approaches have been developed to indirectly probe the properties of the landscapes. The ϕ value analysis can give valuable structural information about the transition state.[4−6] The barrier-less or downhill folding proteins and the notion of protein folding "speed limit" have been introduced[7−10] to estimate the preexponential factor, so that the value of the free energy barrier can be determined from the experimentally measured folding rate. The related quantities, the rate of intrachain contact formation in the unfolded state[11] or upper bound on transition path times have been measured.[12] Atomic force microscopy and optical tweezers allow one to follow folding dynamics with higher temporal resolution at the expense of biasing the sampling by an applied force.[13] The limited spatial resolution (single degree of freedom) of the single molecule experiments can be alleviated to some degree by a sophisticated analysis.[14−16]

Simulation, in principle, can provide high spatial and temporal resolution, necessary for the construction of the quantitatively accurate free energy landscapes. In practice, however, it suffers from force field inaccuracies, sampling issues, and inadequate analysis.[17,18] Due to computational constraints, the free energy landscapes have been mainly constructed for model systems of protein folding, i.e., lattice models,[2,19] Go models,[20,21] or small peptides.[22−26] While being instrumental in developing the free energy landscape methodology, it is not clear how transferable are the results to the real proteins.[27] Recently, due to advances in the hardware and simulation methodology, realistic simulation of folding of small fast-folding proteins became computationally affordable.[28−30] In particular, Shaw et al. reported a "brute-force" 200 μs equilibrium folding simulation of FIP35 protein in explicit water that contains 15 folding−unfolding events with the folding rate and the native structure in agreement with experiment.[30] The free energy landscape was constructed by optimizing the reaction coordinate around a transition state defined by the folding probability $p_{fold} \sim 0.5$.[20] They conclude that FIP35 is "an incipient downhill folder", "landscape roughness is a major determinant of the rates for conformational transitions", and that the preexponential factor is $k_0^{-1} \sim 1$ μs.

Quantitative analysis of protein dynamics in terms of the free energy landscapes is notoriously difficult. A poorly chosen reaction coordinate may hide the complexity of the free energy

Received: September 6, 2011
Published: September 08, 2011

dx.doi.org/10.1021/jp208585r | *J. Phys. Chem. B* 2011, 115, 12315–12324

landscape and associated dynamics.[17,24] Here, I perform a detailed rigorous analysis of the folding simulation of FIP35 using an alternative definition of optimal reaction coordinate: a reaction coordinate is optimal if the dynamics projected on it is diffusive. In this case, the free energy landscape and the diffusion coefficient give a complete and accurate description of the dynamics. The analysis suggests that FIP35 folds via a populated on-pathway intermediate separated by high free energy barriers, the high free energy barriers are a major determinant of the rates for conformational transitions, and that the preexponential factor (for a single transition state) is $k_0^{-1} \sim 10$ ns.

METHODS

The putative optimal coordinate with diffusive dynamics is constructed by numerically maximizing the cut free energy profile $F_C(x)$ with a penalty term to avoid overfitting. The higher the profile, the more diffusive is the projected dynamics.[31] The analysis is performed with a time resolution of $\Delta t = 0.2$ ns and employs no adjustable parameters other than the optimized reaction coordinate.

Free Energy Profiles. The reaction coordinate time series is computed as $x(i\Delta t) = R(\vec{X}(i\Delta t))$, where $\vec{X}(i\Delta t)$ is a multidimensional trajectory recorded with time interval Δt and $x = R(\vec{X})$ defines the reaction coordinate. The partition function of the conventional histogram-based free energy profile in a bin $[x_i, x_i + \Delta x]$ is equal to the density of points in the bin

$$Z_H(x) = N/\Delta x \quad \text{for} \quad x \in [x_i, x_i + \Delta x] \tag{1}$$

where N is the number of time-series points in the bin and Δx is the size of the bin. The partition function of the cut-based free energy profile equals half the total number of transition trough point x:

$$Z_C(x) = 1/2 \sum_i \Theta\{(x(i\Delta t) - x)(x - x(i\Delta t + \Delta t)\} \tag{2}$$

where $\Theta\{x\}$ is the Heaviside step function. The corresponding free energies are $F_H(x) = -kT \ln Z_H(x)$ and $F_C(x) = -kT \ln Z_C(x)$. If Z_H does not change much on the average displacement Δx, $Z_C(x) = 1/2\langle|\Delta x|\rangle Z_H(x)$, where $\langle|\Delta x|\rangle$ is the mean absolute displacement of the reaction coordinate during time interval Δt at point x. For diffusive dynamics with Gaussian increments $P(\Delta x) \sim \exp[-\Delta x^2/(4D\Delta t)]$, one finds $\langle|\Delta x|\rangle = 2(D\Delta t/\pi)^{1/2}$ and

$$Z_C(x) = \sqrt{D(x)\Delta t/\pi} Z_H(x) \tag{3}$$

The equation allows one to estimate the position dependent diffusion coefficient

$$D(x) = \pi/\Delta t Z_C^2(x)/Z_H^2(x) \tag{4}$$

or equivalently $D(x) = 4\pi\langle|\Delta x|\rangle^2/\Delta t$. For subdiffusive dynamics, when $\langle|\Delta x(\Delta t)|\rangle \sim \Delta t^{\alpha}$, one obtains $Z_C(\Delta t) \sim t^{\alpha-1}$ and

$$\alpha = 1 + \ln \frac{Z_C(\Delta t_2)}{Z_C(\Delta t_1)} / \ln \frac{\Delta t_2}{\Delta t_1} = \ln \frac{\langle|\Delta x(\Delta t_2)|\rangle}{\langle|\Delta x(\Delta t_1)|\rangle} / \ln \frac{\Delta t_2}{\Delta t_1}$$

where $\alpha(x)$ and $Z_C(x)$ might be position-dependent. The mean first passage time is estimated via the Kramers equation in the overdumped regime[3,7]

$$t_f = \int_N^D e^{\beta F_H(x)}/D(x)\, dx \int_x^D e^{-\beta F_H(y)}\, dy \tag{5}$$

where $\beta = 1/k_BT$, N, and D are the positions of the native and denatured basins and $N < D$, which can be cast to

$$t_f = \Delta t/\pi \int_N^D Z_C^{-2}(x) Z_H(x)\, dx \int_x^D Z_H(y)\, dy \tag{6}$$

The folding probability is computed from the free energy profile as[32]

$$p_{fold}(x) = \frac{\int_x^D \exp(F(x)/kT)/D(x)\, dx}{\int_N^D \exp(F(x)/kT)/D(x)\, dx}$$

which is equivalent to

$$p_{fold}(x) = \frac{\int_x^D Z_C^{-2}(x) Z_H(x)\, dx}{\int_N^D Z_C^{-2}(x) Z_H(x)\, dx} \tag{7}$$

Reaction Coordinate Optimization. The optimal reaction coordinate $R(x)$ was constructed by numerically optimizing the parameters $\{\alpha_i\}$ of reaction coordinate functional form $R(x, \alpha)$ to make the cut free energy profile $F_C(R)$ the highest. The whole coordinate was optimized by maximizing $I_{ND} = \int_N^D Z_H(x)/Z_C^2(x)\, dx$, where N and D are the positions of local minima in the native and denatured basins. Other optimization functionals, e.g., the mean first passage time,[31] produced very similar results.

The numerical optimization was performed by randomly modifying the parameters α_i, recomputing the reaction coordinate, the profiles, and the functional, and accepting the new parameter value if $\Delta I/I < 0.01$, where ΔI and I are the change and the current value of the optimization functional.

The putative reaction coordinate is taken as the (smoothed) number of contacts $R(x, \alpha) = \Sigma_{ij}\alpha_{ij}h(\alpha_{ij}^r - r_{ij})$, where α_{ij} is either 1 or -1, α_{ij}^r is a threshold when a contact is considered to be formed, and r_{ij} is the distance between atoms i and j; $h(x) = \max(0,\min(1,x))$. Different sets of atoms were considered: backbone HN and O atoms, CA atoms, backbone HN and O and CA atoms. The simplistic functional form of the reaction coordinate (the number of contacts) is chosen to illustrate the robustness of the results: even a simple coordinate, if optimized properly, provides an accurate description of dynamics. While the description is accurate only around the transition states, it is sufficient to estimate a number of characteristics. More complex forms of coordinate (e.g., neural networks[33]) are likely to extend the regions with accurate description.

Overfitting. When sampling is limited and a reaction coordinate has many parameters, it is possible to overfit the data by constructing a reaction coordinate with a free energy profile higher than the correct one. In this case, the dynamics projected on the coordinate is superdiffusive,[31] which can be used as an indicator of overfitting (see Results). A penalty term is introduced in the optimization to avoid overfitting: $\int p(x)Z_H(x)/Z_C(x)\, dx$, where $p(x) = k^{10}(x)$ if $k(x) > 1$ and zero otherwise, and $k(x) = Z_C(x,\Delta t_2)/Z_C(x,\Delta t_1)(\Delta t_2/\Delta t_1)^{1/2}$, with $\Delta t_1 = 1$, $\Delta t_2 = 8$ trajectory steps of 0.2 ns. The performance of the overfitting penalty term is illustrated at the end of the Results section.

The Optimization Procedure. The reaction coordinate optimization problem is complicated by having many competing local minima, where the numerical optimization can be stuck.

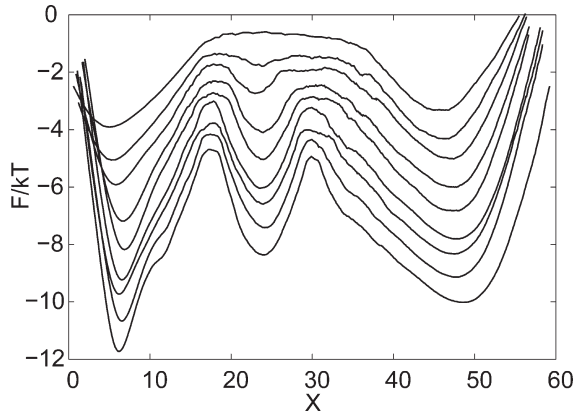


Figure 1. The free energy profiles along coordinates computed during iterative optimization with different Δt values of 128, 64, 32, 16, 8, 4, 3, 2, and 1. The profiles are rescaled along x and shifted along F/kT for visual clarity. The profile with $\Delta t = 1$ is the lowest, while that with $\Delta t = 128$ is the highest.

The following procedure was found to be useful to find the putative global minimum. The large time scale description of the dynamics should be simpler, since many metastable states are absent. The reaction coordinate optimization problem at the large time scale (large sampling interval Δ) should be simpler as well. The optimization was performed in an iterative manner starting with large sampling interval $\Delta t = 128$ and halving it at the next iteration until $\Delta t = 1$. Optimization was initialized with a seed reaction coordinate (the rmsd from the native structure here), which contribution was slowly decreased to zero.

Figure 1 illustrates the performance of the iterative algorithm. At $\Delta t = 128$, the free energy profile has just one broad transition state. At higher time resolution, $\Delta t \leq 16$, the intermediate state appears. At $\Delta t = 1$, the profile shows the intermediate and the two transition states. Figure 1 establishes the general tendency and the robustness of the approach with respect to the choice of the sampling interval (Δt). The transition state around $x \approx 30$ is somewhat lower and narrower than that in Figure 2a. The transition state is suboptimal, as indicated by subdiffusion exponent $\alpha < 0.5$.[31] To obtain uniform optimization of both the transition states (to refine the profile), it was useful to compute the functional for each transition state and optimize their product $I_{NI}I_{ID}$ rather than I_{ND}, where N, I, and D are the native, intermediate, and denatured states, respectively. Note that a single reaction coordinate is used to describe both of the transition states. Such constructed reaction coordinate is used for analysis, e.g., to compute the free energy profile shown in Figure 2a.

Analysis of a Temperature-Jump Experiment. The relaxation dynamics of the system (at temperature $T + \Delta T$) is described by the three-state model N ⇔ I ⇔ D with the native N, intermediate I, and denatured D states (Figure 2a). k_{ji} denotes the reaction rate from state i to state j, with the detailed balance equation $k_{ji}Z_i = k_{ij}Z_j$, where Z_i is the (relative) equilibrium partition function of state i at $T + \Delta T$. The relaxation dynamics is described by the system of equations

$$dP_N/dt = -k_{IN}P_N + k_{IN}Z_N/Z_IP_I$$
$$dP_D/dt = -k_{ID}P_D + k_{ID}Z_D/Z_IP_I$$
$$P_N + P_I + P_D = 1$$
$$Z_N + Z_I + Z_D = 1$$

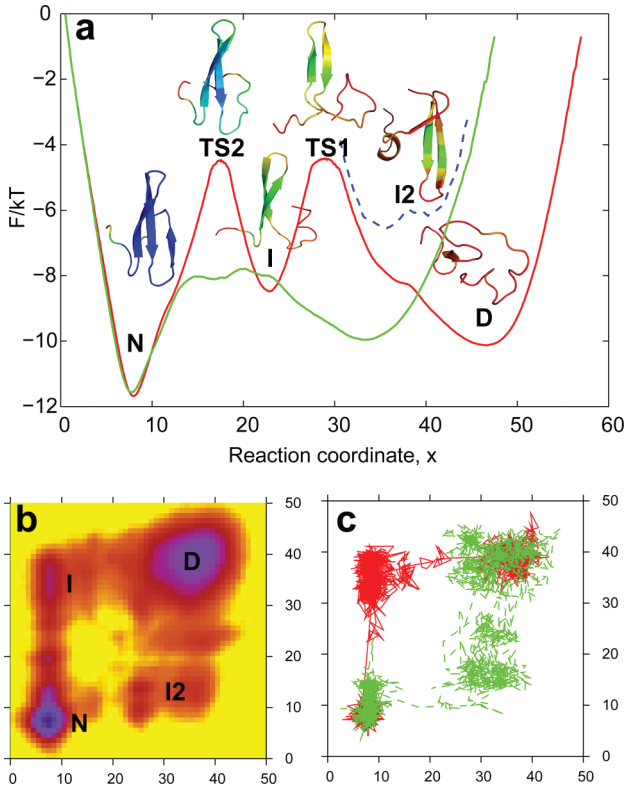


Figure 2. (a) The free energy profile along the constructed reaction coordinate (red) and that along the coordinate used by Shaw et al. (green). N, D, I, I2, and TS1, TS2 denote the native, denatured, and intermediate basins and transition states, respectively. The I2 intermediate overlaps with the denatured basin upon projection and is indicated by a dashed blue line. The representative structure for the regions of the landscape shows a trajectory snapshot closest to the average structure of the region. Colors code the root-mean-square fluctuation of atom positions around the average structure from 1.5 Å (blue) to 7 Å (red). The structures describe the (major) folding pathway.[30] The reaction coordinates are rescaled, so that the diffusion coefficient is $D(x) = 1$. While the positions of the transition state regions on the two rescaled coordinates may seem different ($13 < x < 35$ and $13 < x < 23$), they, in fact, belong to the same region of the configuration space defined by the relative partition function $0.58 < Z_A < 0.63$.[34] (b) The free energy surface as a function of the optimal reaction coordinates for the first and second hairpins. (c) Two segments of trajectory illustrate different folding pathways.

To obtain a closed form solution, it is assumed that $k_{IN} = k_{ID} = k$, which is a good approximation, since $1/k_{IN} = 5.1\ \mu s \approx 1/k_{ID} = 4.4\ \mu s$, as estimated from the trajectory. Exact numerical solution of the equations confirms the accuracy of the approximation. Initial populations (the equilibrium populations at T) are given by $P_i(t = 0) = Z_i'/(\Sigma_i Z_i')$, where $Z_i' = \exp[-(F_i/k_BT)] = \exp[-(H_i/k_BT) + kS] = \exp[-(H_i/k_BT) + (H_i/k_B(T + \Delta T))]\cdot Z_i = \exp[-(H_i\Delta T/k_BT^2)]Z_i$.

$$P_D(t = 0) = \frac{(\gamma_D + 1)Z_D}{1 + \gamma_N Z_N + \gamma_D Z_D}$$

$$P_N(t = 0) = \frac{(\gamma_N + 1)Z_N}{1 + \gamma_N Z_N + \gamma_D Z_D}$$

where $\gamma_N + 1 = \exp[-(H_N - H_I)\Delta T/(k_BT^2)]$, $\gamma_D + 1 = \exp[-(H_D - H_I)\Delta T/(k_BT^2)]$, and H_N, H_I, and H_D are the

enthalpies of the native, intermediate, and denatured states. For the population of the native state, one finds

$$P_N(t) = Z_N + C[e^{-kt}(\gamma_N - \gamma_D)Z_D + e^{-kt/Z_I}(\gamma_N Z_N + \gamma_D Z_D)Z_I]$$

where C is an (unimportant) constant. By expanding the exponents (the linear regime), the varying part (up to a normalizing factor) is written in the following form[10]

$$S(t) = A_m e^{-t/\tau_m} + (1 - A_m)e^{-t/\tau_a}$$

where $\tau_m < \tau_a$, $\tau_a = 1/k$, $\tau_m = Z_I\tau_a = Z_I/k$, and

$$A_m = \frac{((H_I - H_N)Z_N + (H_I - H_D)Z_D)Z_I}{((H_I - H_N)Z_N + (H_I - H_D)Z_D)Z_I + (H_D - H_N)Z_D}$$

The relative partition functions change with temperature as

$$Z_i(T) = \frac{e^{-\Delta\beta H_i}Z_i^0}{e^{-\Delta\beta H_N}Z_N^0 + e^{-\Delta\beta H_I}Z_I^0 + e^{-\Delta\beta H_D}Z_D^0}$$

where $\Delta\beta = 1/k_BT - 1/k_BT_0$, and index 0 denotes the reference values obtained from simulation. The following parameters are used: $H_I - H_N = 25$ kcal/mol, $H_I - H_D = -14$ kcal/mol, $H_D - H_N = 39$ kcal/mol,[30] $Z_I = 0.03$, $Z_N = 0.59$, and $Z_D = 0.38$.

The temperature dependence of the rate constant k, which is necessary to analyze the temperature dependence of τ_m, cannot be determined within the model. It was modeled as $k \sim \exp(-\alpha/T)$, with $\alpha = -1692.5$, so that $\tau_a = 1/k$ mimics the behavior of τ_a seen in experiment; i.e, τ_a is doubled at temperature when $A_m \sim 0.5$.[10]

RESULTS

The Free Energy Profile. Figure 2a shows the free energy profiles as functions of the constructed reaction coordinate and that of Shaw et al. The coordinates are rescaled so that the diffusion coefficient equals unity ($D(x) = 1$);[34] in that case, the dynamics is described by the free energy profile only. The profile shows that FIP35 is not an "incipient downhill folder".[30] It folds via a populated on-pathway intermediate (I) separated by high free energy barriers (transition states TS1 and TS2) from the denatured (D) and native (N) states. The representative structures of the states show that the constructed reaction coordinate reproduces the observed order of folding events (of the major pathway) as described in ref 30: unstructured (D), tip of the first hairpin formed (TS1), the first hairpin is formed (I), stabilization of the first hairpin and formation of the second hairpin (TS2), native state (N).

The mean folding time (mft) estimated from the free energy profile with the Kramers equation with the computed position dependent diffusion coefficient (eq 6) is 3.9 μs. It agrees reasonably with the mft of 9 μs estimated directly from the trajectories. That for the free energy profile of Shaw et al. estimated with eq 6 is 0.2 μs. Note that this estimate is obtained with analysis with a time resolution of $\Delta t = 0.2$ ns. The mean folding time of 4.0 μs is obtained for the free energy profile and diffusion coefficient constructed with a time resolution of 60 ns.[30] The total numbers of transitions between the native state and the intermediate ($x = 8 \leftrightarrow x = 23$) estimated from the profile and counted from the trajectory are 50 and 22, respectively. Those between the denatured and intermediate states ($x = 45 \leftrightarrow x = 23$) are 33 and 17. The close numbers indicate that the free energy profile describes the kinetics reasonably well up to a factor of 2.

The Free Energy Surface. Figure 2b shows the free energy as a function of the two "optimal" reaction coordinates that describe the folding dynamics of each of the hairpins. The coordinates were constructed by considering subsets of atoms from the first (residues 9−21) or second (residues 18−30) hairpin, respectively. The free energy surface confirms the existence of the intermediate (I) and suggests the existence of another intermediate with the second hairpin formed (I2) and two folding pathways differing in the order of formation of two hairpins. I2 is $\sim 1.5kT$ higher in free energy than I in agreement with ref 30 and overlaps with the denatured basin on projection onto the reaction coordinate (Figure 2a). Analysis of the trajectory shows that the major folding pathways, where the formation of the first hairpin is followed by the second, accounts for 80% folding−unfolding events (12 out of 15). The second folding pathway, with the reverse order of the hairpin formation, accounts for 20% events (3 out of 15). Figure 2c shows the segments of the trajectory which follow the two folding pathways. Multiple folding pathways for WW proteins were reported in other simulation studies as well.[29,35,36]

The Diffusivity of the Projected Dynamics. Establishing that the projected dynamics is diffusive is essential if the free energy profile is used for the quantitative analysis of the dynamics, e.g., with Kramers theory.[3] It is also an indicator of whether the reaction coordinate is optimal. Dynamics is subdiffusive when projected on a suboptimal coordinate with lower free energy profile.[31] Figure 3a shows that the mean square displacement $\langle \Delta x^2 \rangle$ for the constructed optimal coordinate and that of Shaw et al. grows significantly slower than is expected for diffusive dynamics. The dynamics projected on the reaction coordinates is subdiffusive with $\langle \Delta x^2 \rangle \sim t^{2\alpha}$ and $\alpha \sim 0.15 < 0.5$. A large number of experimental and theoretical works have shown that protein dynamics is subdiffusive when analyzed along various coordinates.[37−40] The subdiffusion is observed in the whole analyzed time range, a time scale after which the dynamics is diffusive cannot be selected. The subdiffusion indicates that the profile cannot be used for consistent quantitative description of the dynamics as diffusion, e.g., the value of the mean folding time or the diffusion coefficient, depends on the time scale chosen for the analysis.[34]

However, the results presented in Figure 3a are too coarse-grained. The reaction coordinates are usually optimized around the transition state regions—the most important parts for the kinetics.[20,31] Contribution of the transition state regions to the overall plot (Figure 3a) is negligible due to their small population. Figure 3b shows the position dependent subdiffusion exponent $\alpha(x)$ for the two coordinates[31] (see Methods and eq 8 below). The constructed coordinate is diffusive around the two transition state regions but subdiffusive on the rest of the coordinate. It is due to the simplistic functional form of the reaction coordinate (see Methods). Such a functional form is chosen to illustrate the robustness of the results: even a simple coordinate, if properly optimized, can provide an accurate (diffusive) description of dynamics. The folding dynamics can be accurately described as diffusion on the constructed free energy profile around the transition states. The description is sufficient for accurate estimation of such quantities as the mean folding time, the folding probability p_{fold}, and the preexponential factor k_0. For example, when computing the mean folding time, the value of the diffusion coefficient around the transition states

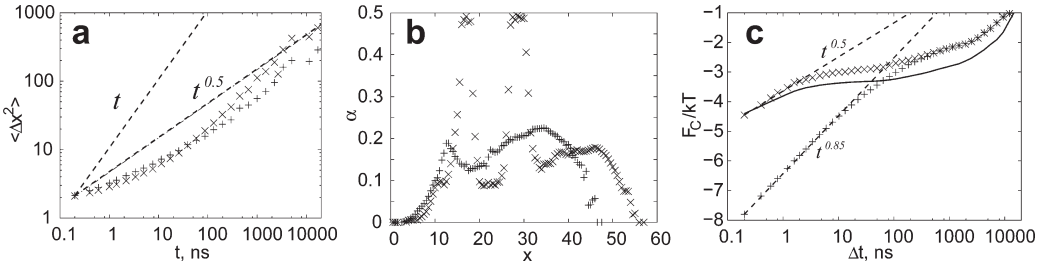


Figure 3. Analysis of whether the dynamics projected on the reaction coordinates is diffusive; crosses for the constructed coordinate and pluses for that of Shaw et al. (a) The mean square displacement $\langle \Delta x^2 \rangle$ vs time. Dashed lines show diffusive (t) and subdiffusive ($t^{0.5}$) scaling to guide the eye. (b) The position dependent subdiffusion exponent $\alpha(x)$. (c) $F_C(\Delta t)$ at the top of the transition states vs the sampling interval Δt. Dashed lines show diffusive ($t^{0.5}$) and subdiffusive ($t^{0.85}$) scaling to guide the eye. The solid line shows $F_C(\Delta t)$ for the Monte Carlo dynamics on the free energy profile (Figure 2a).

only is important; the denatured state enters through the total partition function. The coordinate of Shaw et al. is entirely subdiffusive.

According to the Mory−Zwanzig formalism,[41,42] dynamics projected on an arbitrary coordinate can be exactly described by the generalized Langevin equation with a memory kernel. In this sense, a coordinate with the highest free energy barrier is no better than a coordinate with a low barrier, or no barrier at all. Correlations described by the kernel lead to subdiffusive dynamics and compensate low barriers to preserve the kinetics.[31] To completely specify the dynamics described by the generalized Langevin equation, the memory kernel needs to be computed, which is complicated.[43,44] It is also not clear whether the kernel can be incorporated into the free energy landscape, so that the latter would provide a complete description of the dynamics. It seems reasonable to define the optimal reaction coordinate as the one where the kernel equals zero, i.e., the dynamics is diffusive and Markovian and can be completely described by the free energy profile and the diffusion coefficient.

Such a coordinate can be constructed by optimizing its cut free energy profile $F_C(x)$: the higher $F_C(x)$, the larger is the subdiffusion exponent $\alpha(x)$ and the closer is the dynamics to diffusive.[31] $F_C(x)$ is invariant to reaction coordinate rescaling and equals the conventional (histogram) free energy profile $F_H(x)$ if the diffusion constant is $D(x) = 1$ (Figure 2a). Consider the region around the top of a transition state, where the free energy profile is flat. The partition function $Z_C(x)$ is equal to the number of transitions through point x.[34] In this case, $Z_C(\Delta t) = 1/2\langle|\Delta x(\Delta t)|\rangle Z_H(\Delta t)$, $F_C/kT = -\ln Z_C$, and $F_H/kT = -\ln Z_H$.[34] Assuming $\langle|\Delta x(\Delta t)|\rangle \sim \Delta t^{\alpha}$, one obtains $Z_C(\Delta t) \sim t^{\alpha-1}$ and

$$\alpha = 1 + \ln \frac{Z_C(\Delta t_2)}{Z_C(\Delta t_1)} \Big/ \ln \frac{\Delta t_2}{\Delta t_1} = \ln \frac{\langle|\Delta x(\Delta t_2)|\rangle}{\langle|\Delta x(\Delta t_1)|\rangle} \Big/ \ln \frac{\Delta t_2}{\Delta t_1} \quad (8)$$

since $Z_H(\Delta t) \sim 1/\Delta t$. Let $Z_C(R,\Delta t)$ and $Z_C(S,\Delta t)$ be the partition functions for the two reaction coordinates computed at the top of the transition states with sampling interval Δt. At some large sampling interval Δt_2, when the local difference between the two coordinates is negligible, $Z_C(R) \approx Z_C(S)$. For example, at large Δt_2, when the dynamics of transitions over the barrier is ballistic, Z_C reaches the limit value of the total number of folding events, independent of reaction coordinate choice. α estimated from the distance between the cut profiles computed at the analysis interval of $\Delta t_1 = 1$ and Δt_2 shows that the smaller $Z_C(\Delta t_1)$ is (the higher F_C), the larger is α.

To define the smallest value of Z_C ($Z_C^{\star}$), consider dynamics in an original multidimensional configuration space. Let the configuration space be partitioned into fine-grained microstates $\{s_i\}$, so that the dynamics between them is Markovian. The equilibrium dynamics of the system can be described as a flow over a network,[45] with nodes being the microstates $\{s_i\}$ connected by the links with capacities equal to the equilibrium number of transitions between the microstates $c_{ij} = n_{s_i s_j}$. The (variational) transition state ensemble is defined as the surface that separates the denatured and native basins and has the minimum partition function ($Z_C^{\star}$). It can be found as the minimum cut of the network that separates the nodes of the denatured and native basins.[34,45] $Z_C^{\star}$ is attained by the optimal reaction coordinate, which places the microstates on the correct sides of the cut. In other words, every projection of the multidimensional free energy surface onto a single reaction coordinate is likely to decrease the height of the transition state barrier; only the optimal projection preserves the barrier height. Assume that the dynamics, in the original (full) configuration space, observed at time scale Δt_1 is diffusive. Then, the dynamics on the optimal reaction coordinate with highest F_C ($Z_C = Z_C^{\star}$) is diffusive as well, while that on the suboptimal coordinates with lower F_C ($Z_C > Z_C^{\star}$) is subdiffusive. In practice, when sampling of a transition state is limited and a reaction coordinate has many parameters, it is possible to construct a reaction coordinate with $Z_C < Z_C^{\star}$, i.e., to overfit the data. In this case, $\alpha > 0.5$ and dynamics is superdiffusive, which can be used in a penalty term to avoid overfitting (see Methods).

Figure 3c illustrates the analysis by showing $F_C(\Delta t)$ computed at the transition states for the constructed reaction coordinate and that of Shaw et al. The F_C values for the two coordinates are equal for $\Delta t \geq 200$ ns. At $\Delta t = 0.2$ ns, $F_C \sim -7.8$ for the coordinate of Shaw et al. is lower than $F_C \sim -4.3$ for the constructed coordinate, which leads to $\alpha \sim 0.15$ and $\alpha \sim 0.5$, respectively. Diffusive scaling for F_C for the constructed coordinate turns into superdiffusive $\alpha \sim 1$ for $\Delta t > 4$ ns because the dynamics over the transition state is close to ballistic at these time scales.

Non-Markovian dynamics can become Markovian at large time scale, when the memory effects vanish, i.e., subdiffusive dynamics along a suboptimal reaction coordinate can become diffusive. Shaw et al. used $\Delta t \sim 60$ ns to compute the free energy profile and the diffusion coefficient using approach of Hummer.[46] The approach fits the free energy profile and the diffusion coefficient to the dynamics observed at this time

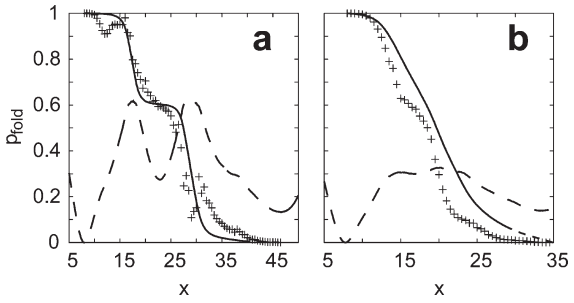


Figure 4. The folding probability computed from diffusive dynamics on the free energy profile (solid line) and directly from the trajectory (symbols); dashed lines show the corresponding free energy profiles: (a) the constructed reaction coordinate; (b) the reaction coordinate of Shaw et al.

scale by comparing computed and observed transition probabilities (rates) $P_{ij}(\Delta t)$.[46] The thus constructed free energy profile and diffusion coefficient reproduce the folding rate. However, they fail to reproduce complementary characteristics: the transition path times and the folding probability as shown below. In particular, the chosen time scale is too large to resolve the two transition states with the transition path times of 3−10 ns.

The Folding Probability. It is instructive to compare the folding probabilities computed directly from the trajectories[47] with those computed from the free energy profiles (Figure 4). For the constructed reaction coordinate, the agreement is reasonably good, taking into account the limited statistics and the fact that the profile is accurate and the dynamics is diffusive only around the transition states. Both curves show a steep decrease in the transition state regions with a plateau in the intermediate state. Non-monotonic behavior around $x \approx 12$ and $x \approx 32$ is due to intermediate basins (e.g., I2 in Figure 2b) that are projected onto the reaction coordinate suboptimally. For the coordinate of Shaw et al. (Figure 4b), the p_{fold} computed from the trajectories is similar to that in Figure 4a, suggesting the presence of the two transition states. The p_{fold} computed from the profile differs and indicates that the free energy profile and diffusion coefficient do not accurately describe the dynamics. The linear decrease of p_{fold} is characteristic for a flat free energy profile and a constant diffusion coefficient. The folding probabilities for the transition states and the intermediate, estimated from the trajectory, are $p_{fold}(TS2) \sim 0.75$, $p_{fold}(I) \sim 0.6$, and $p_{fold}(TS1) \sim 0.2$.

The reaction coordinate of Shaw et al. is constructed by using the variational approach of Best and Hummer, which maximizes the conditional probability of being on a transition path that leads from one metastable state to the other.[20] For diffusive dynamics, the maximum is attained at $p_{fold} = 0.5$, which is a reasonable definition of the transition state for systems with one dominant barrier.[19] For systems with an on-pathway intermediate separated by two similar barriers, such as found here, $p_{fold} \sim 0.5$ corresponds to an intermediate state (Figure 4a). Regions before or after $p_{fold} \sim 0.5$ (the two transition states) are not optimized by the approach.[48] The full p_{fold} reaction coordinate can be constructed by optimizing the generalized cut free energy profile.[49] The free energy profile along this coordinate is virtually identical to the one presented here.[49]

The Transition Path Times. $\langle t_{TP} \rangle$ measures the average time it takes to actually cross a barrier and is a quantity related to the preexponential factor k_0.[7,12] It is relatively insensitive to the barrier height while sensitive to the diffusion coefficient.[12] It can be used as an indicator (complementary to the reaction rate) of how accurately the free energy profile and diffusion coefficient describe the dynamics. From the trajectory, one obtains $\langle t_{TP} \rangle \sim 3$ ns to overcome TS2 ($x = 12 \leftrightarrow x = 23$) and $\langle t_{TP} \rangle \sim 10$ ns to overcome TS1 ($x = 23 \leftrightarrow x = 38$). The transitions appear as instantaneous in Figure 2c. $\langle t_{TP} \rangle \sim 160$ ns to overcome the two transition states and the intermediate ($x = 12 \leftrightarrow x = 38$), since the system has to escape the intermediate (which has a lifetime of about 150 ns). They are in agreement with $\langle t_{TP} \rangle$ estimated from the constructed free energy profile (Figure 2a): $\langle t_{TP} \rangle \sim 3.5$ ns for TS2, $\langle t_{TP} \rangle \sim 7.2$ ns for TS1, and $\langle t_{TP} \rangle \sim 100$ ns to overcome the two transition states and intermediate. The $\langle t_{TP} \rangle$ estimates for TS1 and TS2 are close to that reported for the β-hairpin of 17 ns at $T = 350$ K.[26] Shaw et al. estimated $\langle t_{TP} \rangle \sim 400$ ns to overcome the two transition states and intermediate.[30] That for each transition state (half the distance) can be estimated as $\langle t_{TP} \rangle \sim 100$ ns, using $\langle t_{TP} \rangle \sim \Delta x^2$ for the flat profile with a constant diffusion coefficient.

Estimation of the Preexponential Factor k_0. The Kramers equation estimates the mean folding time in high-friction limit as[3,7] $t_f = \int_N^D e^{\beta F(x)}/D(x)\, dx \int_x^D e^{-\beta F(y)}\, dy$, where $\beta = 1/k_B T$, $F(x)$ and $D(x)$ are the free energy profile and diffusion coefficient, respectively, and N and D are the coordinates of the native and denatured basins. For relatively high barrier, it simplifies to $t_f = \int_{TS} e^{\beta F(x)}/D(x)\, dx \int_D e^{-\beta F(x)}\, dx$, where $\int_{TS}$ and $\int_D$ denote integration over the transition state (barrier) and denatured basin, respectively. The value of the diffusion coefficient is important only at the top of the barrier. The denatured state enters via the total partition function, which justifies the usage of the constructed reaction coordinate. Note that the intermediate makes a negligible contribution to both integrals and hence to the folding rate. However its presence is important to explain why the approach of Best and Hummer has constructed a suboptimal coordinate.

Employing harmonic approximation, one obtains

$$t_f = \frac{\sqrt{2\pi/\beta}\, \exp(\beta F_{TS})}{\omega_{TS} D_{TS}} \frac{\sqrt{2\pi/\beta}\, \exp(-\beta F_D)}{\omega_D} \tag{9}$$

where ω^2 are the corresponding curvatures and subscripts TS and D denote the transition state and denatured basin, respectively. The equation can be rearranged[7]

$$t_f = \frac{2\pi k_B T}{\omega_{TS} D_{TS} \omega_D} \exp(\beta(F_{TS} - F_D)) = k_0^{-1} \exp(\beta \Delta F) \tag{10}$$

where k_0 and ΔF are the so-called preexponential factor and the free energy barrier. Knowing k_0, the free energy barrier can be estimated from the experimentally measured reaction rate. Assuming $\omega_{TS} = \omega_D$ and $D_{TS} = D_D$,[7,18,50] one obtains $k_0^{-1} = 2\pi\tau_{corr,D}$, where $\tau_{corr,D} = k_B T/(D_D \omega_D^2)$ is the decay time of the autocorrelation function in the denatured basin. The latter can be measured experimentally and leads to the following estimate $k_0^{-1} \sim 1\ \mu$s.[7,18]

Using only $\omega_D = \omega_{TS}$, one obtains instead $k_0^{-1} = 2\pi\tau_{corr,TS}$, where $\tau_{corr,TS} = k_B T/(D_{TS} \omega_{TS}^2)$ is the autocorrelation decay time in the transition state. The top of the first transition state (Figure 2a) is approximated by $\omega^2/2 = 0.13 k_B T$, which leads to $k_0^{-1} = 2\pi/0.26 \times 0.2$ ns ~ 5 ns.

To relax the assumption $\omega_D = \omega_{TS}$, eq 9 is modified

$$t_f = \frac{2\pi k_B T}{{\omega_{TS}}^2 D_{TS}} \frac{\sqrt{2\pi/\beta}\exp(-\beta F_D)/\omega_D}{\sqrt{2\pi/\beta}\exp(-\beta F_{TS})/\omega_{TS}} = k_{TS}^{-1}\frac{Z_D}{Z_{TS}} \quad (11)$$

where Z_D and Z_{TS} are the total partition function of the denatured state and transition state and $k_{TS}^{-1} = 2\pi\tau_{corr,TS}$. Considering the transition state as an intermediate state in folding dynamics N ⇔ TS ⇔ D. The folding rate is[51] $k = k_{N\leftarrow TS}k_{TS\leftarrow D}/(k_{N\leftarrow TS} + k_{D\leftarrow TS})$, where $k_{i\leftarrow j}$ denotes the reaction rate from j to i. For a symmetric barrier, $k_{N\leftarrow TS} = k_{D\leftarrow TS}$, and using the detailed balance, one finds $k_{TS} = k_{N\leftarrow TS}/2$; i.e, the prefactor equals half the reaction rate from the transition state to the native state. Such decomposition of the reaction rate onto the free energy difference and the prefactor (eq 11) have the following advantages over the conventional (eq 10): it is invariant to coordinate transformation, it does not assume that $\omega_D = \omega_{TS}$, the free energy profile should be accurate only around the top of the barrier, and k_{TS} is a property of the transition state only. If $\omega_{TS} = \omega_D$, the two decompositions are equivalent $k_{TS} = k_0$ and $Z_D/Z_{TS} = \exp(\beta\Delta F)$.

Equation 9 can be rearranged in another way

$$t = \frac{\Delta t\sqrt{2\pi/\beta}\exp(-\beta F_D)/\omega_D}{\exp(-\beta F_{TS})\sqrt{D_{TS}\Delta t/\pi}}\sqrt{\frac{2k_BT/\Delta t}{D_{TS}{\omega_{TS}}^2}} = \frac{\Delta t Z_D}{Z_C}\sqrt{2\tau_{corr,TS}/\Delta t} \quad (12)$$

where Z_C is the cut profile at the top of the transition state and Δt is the sampling interval. The mean folding time is estimated here as the total time spent in the denatured basin $Z_D\Delta t$ divided by the total number of transitions through the transition state Z_C corrected for recrossing events by factor $(2\tau_{corr,TS}/\Delta t)^{1/2}$. The estimation does not require separate values for ω_{TS} and D_{TS}. For the first transition state ($Z_C = 85$, $Z_D = 0.38 \times 10^6$) with the mean first passage time of $t = 2.3$ μs (estimated by the Kramers equation), one obtains $\tau_{corr,TS} = 0.66$ ns and $k_{TS}^{-1} \sim 4$ ns. The mean first passage time of $t = 4.4$ μs (estimated directly from the trajectory) leads to $\tau_{corr,TS} = 2.45$ ns and $k_{TS}^{-1} \sim 15$ ns.

The agreement between the estimates of k_{TS} and $\langle t_{TP}\rangle$ is encouraging and suggests that the preexponential factor k_{TS} (and k_0, assuming $\omega_{TS} = \omega_D$) is about $k_{TS}^{-1} \sim 10$ ns. It indicates that the high free energy barriers rather than "landscape roughness" are "a major determinant of the rates for conformational transitions".

A Barrier-less Variant of WW. A model of a barrier-less variant of WW protein can be constructed by removing the two transition states with the free energy profile $F'(x) = \min[F(x), -8.2kT]$ (Figure 2a). The folding time estimated by the Kramers equation ($D(x) = 1$) is 400 ns. The estimate is close to the proposed speed limit of protein folding of $N/100$ μs, where $N = 35$ is the number of residues.[7] A successful mutation of FIP35 that halves the folding time has been reported.[52]

Analysis of a Temperature-Jump Experiment. So far, the two free energy landscapes were compared by how accurately they reproduce the simulated dynamics. It is interesting to see how they agree with experiment. Note that it implicitly assumes that the simulated dynamics reproduces that in experiment, in particular, that the force-field is correct.

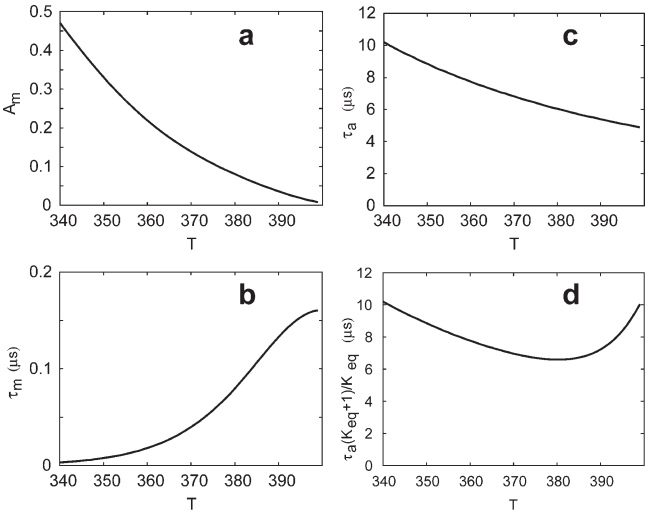


Figure 5. Analysis of a temperature-jump experiment. For details, see text.

A temperature-jump experiment[10] has shown that the relaxation dynamics of FIP35 deviates from single exponential at small times. It is approximated by

$$S(t) = A_m e^{-t/\tau_m} + (1 - A_m)e^{-t/\tau_a}$$

where τ_m and A_m are the relaxation time and amplitude of the fast "molecular rate" associated with diffusion through transition state and τ_a is the conventional relaxation time. $A_m \sim 0.3$, $\tau_m \sim 1.5$ μs, and $\tau_a \sim 10$ μs at the melting temperature of 78 °C.[10] Trajectories obtained by simulating Langevin dynamics on the free energy profile of Shaw et al. were found to contain the fast molecular phase[10,30] characterized by $\tau_m \sim 0.35$ μs and $A_m \sim 0.02$.[10]

The constructed free energy landscape (Figure 2a) agrees with the experiment and provides an alternative interpretation: the fast "molecular rate" is due to the depopulation of the intermediate. Describing dynamics by the three-state model N ⇔ I ⇔ D and assuming that $k_{IN} = k_{ID} = k$, analytical expressions for τ_m, A_m, and τ_a are obtained (see Methods). In particular, $\tau_a = 1/k$ and $\tau_m = Z_I\tau_a$, where Z_I is the partition function of the intermediate state. $1/k_{IN} = 5.1$ μs and $1/k_{ID} = 4.4$ μs (estimated from the trajectory) justify the assumption. The obtained values $A_m = 0.02$ and $\tau_m = 0.15$ μs agree with those obtained using the free energy profile of Shaw et al.[10]

A rigorous way to determine the temperature dependence of A_m and τ_m is to perform and analyze simulations at a series of temperatures. As a crude approximation, one may determine them from the model (Figure 5), assuming it is valid in some temperature range; for details, see Methods. A_m increases as the temperature decreases and reaches 0.5 at $T \sim 340$ K (Figure 5a). The temperature dependence of τ_a cannot be determined from the model. Using the one that mimics the experiment (Figure 5c), the "molecular time" τ_m (Figure 5b) as well as the "activated forward reaction time" $\tau_a(K_{eq} + 1)/K_{eq}$ (Figure 5d) are found to increase with the temperature. The results are in qualitative agreement with the experiment (Figure 4 in ref 10). Note that the model was not fine-tuned to obtain the agreement. Using the model, one can estimate the population of the intermediate state at the melting temperature from the experiment $Z_I = \tau_m/\tau_a \sim 0.15$, which differs from that obtained in

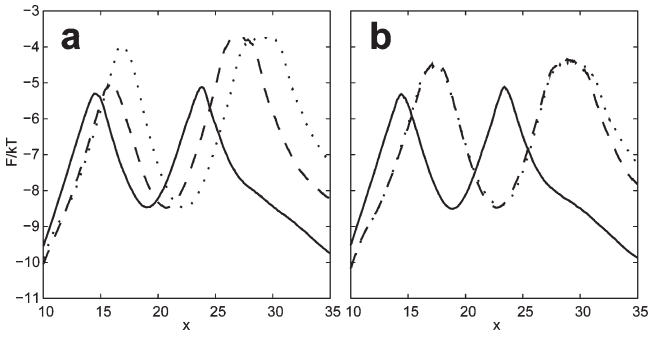


Figure 6. Optimization without (a) and with (b) the overfitting penalty term. Solid lines are for the coordinates constructed from positions of CA atoms (595 parameters), dashed that of HN and O atoms (2080 parameters), and dotted that of HN, O, and CA atoms (4950 parameters).

simulation $Z_I = 0.03$. Using $Z_I = 0.15$ improves the agreement. In particular, $A_m \sim 0.1$ at $T \sim 395$ K and reaches $A_m \sim 0.5$ at $T \sim 370$ K.

Overfitting. An important practical question is how to detect and avoid overfitting of a relatively small number of folding transitions (15 here) by the reaction coordinate with a large number of parameters (2080 here) which lead to overestimation of barrier height. Clearly, a small number of parameters gives more confidence in the result; however, at the same time, it may limit the flexibility and approximation power of the reaction coordinate and lead to suboptimal results. The total number of parameters is not a robust quantitative measure of overfitting, as illustrated by suboptimal results obtained with the principal component analysis.[24] Overfitting is position dependent; e.g., the coordinates are optimized mainly around the transitions states, which are more likely to be overfitted.

Cross-validation, the conventional approach to detect overfitting, optimizes the parameters on the training set and compares the results with that on the test set. It however requires the two sets to be homogeneous, which is problematic in the regime of limited sampling. The test set may contain a pathway or local minimum not present in the training set. The reaction coordinate constructed without the information is likely to produce suboptimal results. Transition state regions are most vulnerable because they are negligibly populated compared with any other region of the landscape that might be wrongly projected on them.

Define overfitting as when the cut free energy profile is higher than the correct one with the diffusive dynamics $Z_C^\star$ (defined as the minimum cut over the network of transitions covering the whole configuration space, see above). In such case, as follows from eq 8, the projected dynamics is superdiffusive, i.e., $\alpha(x) > 0.5$ can be used as an indicator of overfitting. A penalty term is introduced in the optimization to avoid overfitting: $\int p(x) Z_H(x)/Z_C(x)\, dx$, where $p(x) = k^{10}(x)$ if $k(x) > 1$ and zero otherwise, and $k(x) = Z_C(x,\Delta t_2)/Z_C(x,\Delta t_1)(\Delta t_2/\Delta t_1)^{1/2}$, with $\Delta t_1 = 0.2$ ns, $\Delta t_2 = 1.6$ ns. The weighting factor $Z_H(x)/Z_C(x)$ ensures that the penalty factor is invariant to coordinate rescaling. To illustrate its performance, three coordinates with increasing number of parameters have been optimized with and without the penalty term. Figure 6a shows that, without the penalty term, optimization of a coordinate with a larger number of parameters results in higher transition states. The first coordinate is inflexible and suboptimal and has $\alpha \sim 0.45$. The second and third coordinates have much higher transition states with $\alpha \sim 0.85$, indicating overfitting. Figure 6b shows optimization with the penalty term. The second and third coordinates show very similar profiles with $\alpha \sim 0.5$, i.e., free of overfitting.

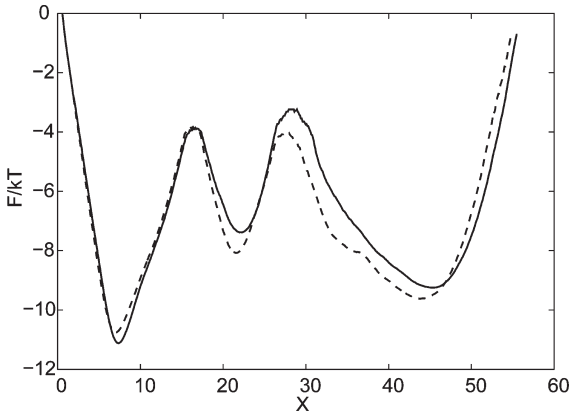


Figure 7. The free energy profiles along the optimal reaction coordinates constructed using the first and second halves of the trajectory separately.

At large sampling intervals ($\Delta t > 4$ ns in Figure 3c), when the system ballistically passes over the barrier, the dynamics is inherently superdiffusive ($\alpha \sim 1$). To show that this behavior is consistent with diffusive dynamics on the constructed free energy profile, the latter was simulated with MC. The corresponding $F_C(\Delta t)$ (Figure 3c) behaves very similar to the one from the folding trajectory. It is somewhat lower in the plateau region because the constructed profile overestimates the number of folding events by a factor of 2, leading to $\Delta F_C/kT \sim 0.7$.

The free energy profiles along the optimal reaction coordinates constructed separately from the first and second halves of the trajectory (Figure 7) are similar to that for the full trajectory (Figure 2a). The sum of the two profiles reproduces the profile for the whole simulation (Figure 2a), e.g., for TS1, exp(4.08) + exp(3.27) $\sim$ exp(4.45). The results provide a "weak" cross-validation: profiles constructed from parts of the trajectory are consistent with each other and with that computed from the whole trajectory. They indicate that increasing the length of trajectory does not change the results, it only decreases the statistical errors.

CONCLUDING DISCUSSION

The optimal reaction coordinate has been constructed for the folding trajectory of FIP35 by optimizing the cut free energy profile. FIP35 folds via the populated on-pathway intermediate separated by the two high transition states. The coordinate, associated free energy profile, and diffusion coefficient provide accurate (diffusive) description of the dynamics around the transition states. They make possible accurate estimation of the mean first passage times, the transition path times, the folding probability, and the preexponential factor. The results of the temperature-jump experiment[10] are qualitatively reproduced.

The coordinate constructed by Shaw et al. is suboptimal: it does not resolve the two transition states. The associated free energy profile and diffusion coefficient, constructed using the approach of Hummer,[46] reproduce the folding rate but fail to reproduce complementary characteristics: the transition path times (for the two transition states) and the folding probability.

The prefactor is estimated for the whole complex of both transition states and the intermediate.

When is the coarse-grained description obtained by Shaw et al. sufficient and when is it necessary to use the more fine-grained description provided by the constructed reaction coordinate? Multiple free energy barriers on a free energy landscape (roughness) can be described by an effective diffusion coefficient when "many fluctuations in roughness take place in the distance of interest".[53,54] The description is valid when the landscape consists of many similar barriers and is certainly not valid when there is a single dominant free energy barrier. The landscape with two transition states, observed here, is closer to the latter.

It seems likely that the fast-folding proteins generally have landscapes with multiple similar barriers. The highest barriers are eliminated to speed up folding until all the barriers are equally small. Such systems with multiple barriers can be approximately described in the coarse-grained manner on the distance of many barriers. One can use an effective diffusion coefficient instead of explicit consideration of barriers. However, the description is not necessarily transferable to proteins with a single dominant barrier. In particular, the preexponential factor for the single dominant barrier is likely to be close to that estimated here for the single transition state ($1/k_0 \sim 10$ ns) rather than that for the whole complex ($1/k_0 \sim 1\ \mu$s).[30] There is no direct evidence that the two state proteins have a single dominant barrier; however, results for model systems[21,24,26] and characteristic single-exponential kinetics make it very plausible.

The estimate of the preexponential factor obtained here ($1/k_0 \sim 10$ ns) is much smaller than the generally accepted value of $1/k_0 \sim 1\ \mu$s.[7,18] While the latter is supported by a large body of indirect evidence, the former is obtained in a direct way by the rigorous analysis. It is indirectly supported by the value of the contact times for short peptides which are in the range of tens of nanoseconds.[11,55,56] It remains to be seen whether the obtained estimate is universally valid for other (small) proteins.

Note that the shape and the number of dominant barriers are the robust characteristics of the optimal coordinate. In particular, introduction of every dihedral angle into the coordinate does not result in additional (dihedral) barriers (i.e., the landscape roughness). The (one-dimensional) free energy profile shows only barriers that divide the configuration space as a whole, not a particular degree of freedom. For example, the profile (Figure 2a) does not show the second intermediate (I2 in Figure 2b). It suggests that while the multidimensional potential energy surface is likely to have roughness (due to, e.g., dihedral barriers) on top of the smooth surface,[53] the free energy surface as a function of a single (or a few) reaction coordinate is inherently smooth.

It is often advocated that more than one reaction coordinate or a Markov state network are necessary to analyze the complex dynamics of protein folding.[18,27,57] Indeed, a single coordinate cannot be used to show all the complexities of the folding dynamics, e.g., the existence of the second intermediate and multiple folding pathways (Figure 2b). However, an optimally constructed reaction coordinate can be used to describe the major properties of the folding dynamics on an arbitrary complex landscape. The optimal reaction coordinate with the highest cut free energy profile should provide an accurate description of the transition state region, as illustrated here. Diffusive dynamics on the free energy profile as a function of the folding probability (p_{fold}) reaction coordinate reproduces the mean folding time for an arbitrary complex landscape.[58]

Here, a simplistic functional form of the reaction coordinate, namely, the number of contacts, was chosen to illustrate the robustness of the approach and the obtained results. The putative optimal reaction coordinate accurately describes the dynamics only around the transition states. Recently, the approach has been used to analyze the game of chess.[59] A reaction coordinate which provides an accurate, diffusive description of the game of chess for the entire coordinate has been constructed. In particular, the probabilities to win the game (corresponding to p_{fold}), estimated from the free energy profile and computed directly from the game trajectories, are in much closer agreement than those in Figure 4a. It may be considered as surprising, anecdotal evidence that the accurate free energy landscape description of the protein folding is more difficult than that of the chess game.

One often validates the quality of a constructed putative transition state ensemble by inspecting whether the committor distribution is picked around $p_{\mathrm{fold}} = 0.5$. As shown here, such validation may fail for systems with two similar transition states. The committor distribution constructed by Shaw et al. passes the test,[30] while the constructed reaction coordinate and the transition states are suboptimal. A robust indicator of the optimality of a putative reaction coordinate is the diffusivity of the projected dynamics.

With recent advances in computer hardware and simulation methodology, simulation of realistic protein folding became computationally affordable. The analysis presented here allows one to describe the complex folding dynamics in a simple and intuitive while rigorous and quantitatively accurate way as diffusion on a free energy landscape. The rigorous analysis is likely to be useful in addressing difficult questions of protein folding, e.g., the value of the preexponential factor, considered here, and others.[60]

AUTHOR INFORMATION

Corresponding Author
*E-mail: s.krivov@leeds.ac.uk.

ACKNOWLEDGMENT

I am grateful to David Shaw and his co-workers for making the folding trajectories available. The work was supported by an RCUK fellowship.

REFERENCES

(1) Onuchic, J. N.; Socci, N. D.; Luthey-Schulten, Z.; Wolynes, P. G. *Folding Des.* **1996**, *1*, 441–450.
(2) Dobson, C. M.; Sali, A.; Karplus, M. *Angew. Chem., Int. Ed.* **1998**, *37*, 868–893.
(3) Kramers, H. A. *Physica* **1940**, *7*, 284–304.
(4) Fersht, A. *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding*; W.H. Freeman: New York, 1999.
(5) Huysmans, G. H. M.; Baldwin, S. A.; Brockwell, D. J.; Radford, S. E. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 4099–4104.
(6) Naganathan, A. N.; Muñoz, V. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 8611–8616.
(7) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76–88.
(8) Yang, W. Y.; Gruebele, M. *Nature* **2003**, *423*, 193–197.
(9) Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Muñoz, V. *Science* **2002**, *298*, 2191–2195.

(10) Liu, F.; Nakaema, M.; Gruebele, M. *J. Chem. Phys.* **2009**, *131*, 195101.
(11) Krieger, F.; Fierz, B.; Bieri, O.; Drewello, M.; Kiefhaber, T. *J. Mol. Biol.* **2003**, *332*, 265–274.
(12) Chung, H. S.; Louis, J. M.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 11837–11844.
(13) Gebhardt, J. C. M.; Bornschlögl, T.; Rief, M. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 2013–2018.
(14) Baba, A.; Komatsuzaki, T. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19297–19302.
(15) Schuetz, P.; Wuttke, R.; Schuler, B.; Caflisch, A. *J. Phys. Chem. B* **2010**, *114*, 15227–15235.
(16) Hummer, G.; Szabo, A. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 21441–21446.
(17) Freddolino, P. L.; Harrison, C. B.; Liu, Y.; Schulten, K. *Nat. Phys.* **2010**, *6*, 751–758.
(18) Buchner, G. S.; Murphy, R. D.; Buchete, N.; Kubelka, J. *Biochim. Biophys. Acta* **2010**, *1814*, 1001–1020.
(19) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334–350.
(20) Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6732–6737.
(21) Allen, L. R.; Krivov, S. V.; Paci, E. *PLoS Comput. Biol.* **2009**, *5*, e1000428.
(22) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495–1517.
(23) Bolhuis, P. G.; Dellago, C.; Chandler, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877–5882.
(24) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.
(25) Krivov, S. V.; Muff, S.; Caflisch, A.; Karplus, M. *J. Phys. Chem. B* **2008**, *112*, 8701–8714.
(26) Best, R. B.; Mittal, J. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 11087–11092.
(27) Bowman, G. R.; Voelz, V. A.; Pande, V. S. *Curr. Opin. Struct. Biol.* **2011**, *21*, 4–11.
(28) Freddolino, P. L.; Liu, F.; Gruebele, M.; Schulten, K. *Biophys. J.* **2008**, *94*, L75–L77.
(29) Ensign, D. L.; Pande, V. S. *Biophys. J.* **2009**, *96*, L53–L55.
(30) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341–346.
(31) Krivov, S. V. *PLoS Comput. Biol.* **2010**, *6*, e1000921.
(32) Rhee, Y. M.; Pande, V. S. *J. Phys. Chem. B* **2005**, *109*, 6780–6786.
(33) Ma, A.; Dinner, A. R. *J. Phys. Chem. B* **2005**, *109*, 6769–6779.
(34) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13841–13846.
(35) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.
(36) Juraszek, J.; Bolhuis, P. G. *Biophys. J.* **2010**, *98*, 646–656.
(37) García, A. E.; Blumenfeld, R.; Hummer, G.; Krumhansl, J. A. *Phys. D (Amsterdam, Neth.)* **1997**, *107*, 225–239.
(38) Kneller, G. R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2641–2655.
(39) Min, W.; Luo, G.; Cherayil, B. J.; Kou, S. C.; Xie, X. S. *Phys. Rev. Lett.* **2005**, *94*, 198302.
(40) Neusius, T.; Daidone, I.; Sokolov, I. M.; Smith, J. C. *Phys. Rev. Lett.* **2008**, *100*, 188103–188104.
(41) Mori, H. *Prog. Theor. Phys.* **1965**, *33*, 423–455.
(42) Zwanzig, R. *Nonequilibrium Statistical Mechanics*; Oxford University Press: New York, 2001.
(43) Portman, J. J.; Takada, S.; Wolynes, P. G. *J. Chem. Phys.* **2001**, *114*, 5082.
(44) Darve, E.; Solomon, J.; Kia, A. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 10884–10889.
(45) Krivov, S. V.; Karplus, M. *J. Chem. Phys.* **2002**, *117*, 10894–10903.
(46) Hummer, G. *New J. Phys.* **2005**, *7*, 34–34.
(47) Rao, F.; Settanni, G.; Guarnera, E.; Caflisch, A. *J. Chem. Phys.* **2005**, *122*, 184901.
(48) Peters, B. *Chem. Phys. Lett.* **2010**, *494*, 100–103.
(49) Krivov, S. V. *J. Phys. Chem. B* **2011**in press.
(50) Best, R. B.; Hummer, G. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 1088–1093.
(51) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Chem. Phys.* **2009**, *130*, 205102.
(52) Piana, S.; Sarkar, K.; Lindorff-Larsen, K.; Guo, M.; Gruebele, M.; Shaw, D. E. *J. Mol. Biol.* **2011**, *405*, 43–48.
(53) Zwanzig, R. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2029–2030.
(54) Lifson, S.; Jackson, J. L. *J. Chem. Phys.* **1962**, *36*, 2410.
(55) Hudgins, R. R.; Huang, F.; Gramlich, G.; Nau, W. M. *J. Am. Chem. Soc.* **2002**, *124*, 556–564.
(56) Yeh, I.; Hummer, G. *J. Am. Chem. Soc.* **2002**, *124*, 6563–6568.
(57) Gruebele, M. *Methods* **2010**, *52*, 1–2.
(58) Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. *J. Chem. Phys.* **2008**, *129*, 174102.
(59) Krivov, S. V. *Phys. Rev. E* **2011**, *84*, 011135.
(60) Sosnick, T. R.; Barrick, D. *Curr. Opin. Struct. Biol.* **2011**, *21*, 12–24.