Folding Kinetics and Unfolded State Dynamics of the GB1 Hairpin from Molecular Simulation

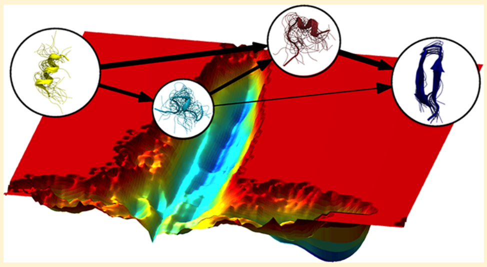
David De Sancho,[†] Jeetain Mittal,[‡] and Robert B. Best*[,§]

[†]Cambridge University, Department of Chemistry, Lensfield Road Cambridge CB2 1EW, United Kingdom

[‡]Department of Chemical Engineering, 111 Research Drive, Iacocca Hall, Bethlehem, Pennsylvania 18015, United States

[§]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, United States

S *Supporting Information*



ABSTRACT: The C-terminal β-hairpin of protein G is a 16-residue peptide that folds in a two-state fashion akin to many larger proteins. However, with an experimental folding time of ~6 μs, it remains a challenging system for all-atom, explicitly solvated, molecular dynamics simulations. Here, we use a large simulation data set (0.7 ms total) of the hairpin at 300 and 350 K to interpret its folding via a master equation approach. We find a separation of over an order of magnitude between the longest and second longest relaxation times, with the slowest relaxation corresponding to folding. However, in spite of this apparent two-state dynamics, the folding rate determined based on a first-passage time analysis depends on the initial conditions chosen, with a nonexponential distribution of first passage times being obtained in some cases. Using the master equation model, we are now able to account quantitatively for the observed distribution of first passage times. The deviation from the expected exponential distribution for a two-state system arises from slow dynamics in the unfolded state, associated with formation and melting of helical structures. Our results help to reconcile recent findings of slow dynamics in unfolded proteins with observed two-state folding kinetics. At the same time, they indicate that care is required in estimating folding kinetics from many short folding simulations. Last, we are able to use the master equation model to obtain details of the folding mechanism and folding transition state, which appear consistent with the "zipper" mechanism inferred from the experiment.

INTRODUCTION

Although the resolution of atomistic molecular dynamics (MD) simulations of protein folding is greater than that of any experimental technique, their high computational cost and the limited accuracy of molecular mechanics force fields[1−7] make benchmarking against experimental data essential.[8] Ultrafast folding peptides and mini-proteins are well suited to such validation,[9] since they can be studied using high resolution spectroscopic techniques[10] and are also amenable to atomistic molecular simulation studies;[8] additionally simplified statistical mechanics models are analytically solvable for short sequences.[11,12] From the extensive characterization of a number of ultrafast folders, a description of the folding of peptides and small proteins which is both accurate and microscopically detailed is currently emerging.[13] Here, we focus on the GB1 hairpin, the C-terminal β-hairpin of protein G[14] (see Figure 1A), which has become one of the testbeds for comparing simulations, theory, and experimental results. This 16-residue peptide was shown to fold to a native-like three-dimensional structure[15] and to exhibit two state folding from relaxation experiments using ultrafast temperature jumps, with consistent results using different experimental probes.[16,17] Since then the GB1 hairpin has been extensively characterized in multiple experimental[17−20] and simulation[21−37] studies.

Sampling of folding events in all-atom simulations of protein folding is still a major challenge, and with a folding time of ~6 μs, the GB1 hairpin is no exception. One approach to sampling folding (or unfolding) events, extensively exploited in worldwide distributed computing projects,[38] is to run many relatively short simulations, of which only a few will contain folding transitions. The folding rate can then be estimated by assuming the folding to occur according to simple two-state kinetics.[25] For example, a recent study of the GB1 hairpin used many simulations of between 0.25 and 1.5 μs each to sample folding and unfolding events, for a total of 0.7 ms of data.[39] A curious result of this work was that the apparent folding rate computed from mean first passage times depended on the choice of initial "unfolded" states. Using as starting structures a set of unfolded states chosen from the equilibrium distribution (U−A), a folding time of 59 μs was obtained, while using a subset of the unfolded state containing no α-helical structures (U−B) resulted in a folding time of 8.2 μs. This is clearly an unexpected result if the peptide indeed folds in a two-state manner: if relaxation within the unfolded state is much faster than folding, one would expect little dependence of the results

Received: November 22, 2012
Published: February 7, 2013

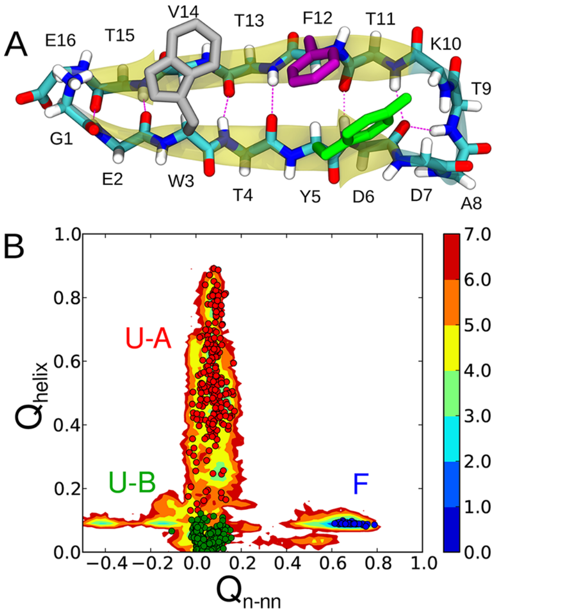


Figure 1. Folding/unfolding simulations of GB1 from different initial states. (A) Representation of the three-dimensional structure of the GB1 hairpin. Backbone atoms are shown for every residue, and side chains are shown for aromatics. The cartoon representation of the hairpin in yellow is overlaid. (B) Potential of mean force from REMD trajectories for the fraction of native minus non-native contacts Q_{n-nn} and the fraction of helical contacts Q_{helix}. Contours are in units of k_BT. Initial conformations for the simulations from the F (blue), U−A (red), and U−B (green) subensembles are shown as circles.

on the chosen initial conditions. To resolve this discrepancy, the *ad hoc* phenomenological kinetic model

$$\text{U−S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \text{U−B} \overset{k_2}{\rightarrow} \text{F} \tag{1}$$

was proposed to provide an explanation for the unexpected dependence on initial conditions. This model resolved the equilibrium unfolded state U−A into two slowly interconverting substates of the unfolded state, U−S and U−B, although it did not provide any microscopic information about the slow equilibration between these substates.

The observation of slow dynamics between the two unfolded substates of the GB1 hairpin is consistent with a number of recent studies using MD simulations[40,41] and their combination with experimental results.[42−45] Addressing the effect of unfolded state dynamics on the calculation of protein folding rates is therefore particularly important. More generally, rates are the quantities that are most easily, and most often, compared with experimental results.[13,46,47] It is therefore essential to find a way of calculating folding rates in an unbiased way that does not depend on the initial setup of the simulations.[48,49]

In this work, we present a master equation (ME) model analysis of the previously reported data set for the GB1 hairpin.[39] In this approach, we do not start by counting "folding" and "unfolding" events as monitored by projection onto a one-dimensional reaction coordinate. Instead, we derive a kinetic master equation describing transitions between a large number of microscopic states representing distinct conformations of the hairpin. In our case, we have chosen to use the backbone torsion angles to classify the states, a method which has successfully been applied to describe the kinetics of helix formation.[50] The advantage of using an ME model is that rates for exchange between the microscopic states are fast and can be accurately determined from short simulations, while the global relaxation can still be computed from the master equation. We find that we are able to recover the same slow relaxation times, within statistical error, from the master equation, irrespective of which set of simulations are used as input data for the model. The model also quantitatively explains the unusual (approximately biexponential) distribution of first-passage times obtained from the original simulations and provides a microscopic explanation for the slow events in the unfolded state, which are associated with helix−coil dynamics. The slowest time-scale of the calculated rate matrix is consistent with the earlier maximum likelihood estimate of folding and unfolding rates from first passage times, when the initial configurations are drawn from an equilibrium distribution of structures in the unfolded state.[39] This has important implications for the calculation of rates from short folding simulations. Last, we have used our model to identify putative transition states via a committor analysis of the rate matrix and performed a systematic coarse-graining in order to obtain a more intuitive description of the dynamics in the unfolded state.

■ METHODS

Simulation Data Set. We use two previously reported simulation data sets for the GB1 hairpin (see Figure 1 A) with the Amber ff03* force field, optimized to match the balance between helical and nonhelical states,[3] and the TIP3P water model.[51] The first of these is a replica-exchange molecular dynamics (REMD) study, in which both folded and unfolded states were used as initial conditions in independent REMD runs to ensure that equilibrium sampling was achieved.[36] The second data set, the most important for the present study, was an extensive set of long molecular dynamics (MD) simulations, starting from either folded or unfolded initial conditions, and at 300 and 350 K.[39]

Initial conditions for the long simulations were chosen using appropriate reaction coordinates. In Figure 1B, we illustrate the free energy landscape projected onto two such coordinates: a modified fraction of native contacts (Q_{n-nn}) and the fraction of helical contacts (Q_{helix}).[36] The exact definitions of Q_{n-nn} and Q_{helix} were given in an earlier publication,[39] but we summarize them here. Q_{n-nn} is the difference between Q_n, the fraction of native contacts ($Q_n(x)$ is the fraction of the atom−atom contacts that would be present in the native state and that are formed in configuration x), and Q_{nn}, the fraction of contacts in a misfolded state, i.e., $Q_{n-nn} = Q_n - Q_{nn}$. Defining the coordinate in this way helps to separate the misfolded state from the folding barrier. $Q_{helix}(x)$ is the fraction of the atom−atom contacts that would be present in an ideal α-helix, which are formed in configuration x. In addition, a third quantity, the least-squares root-mean-square deviation (RMSD) of the coordinates from the native structure,[14] was used.

Using these coordinates (and corresponding potentials of mean force), three regions were chosen for initiating long simulations: (i) the folded state (F), with $Q_{n-nn} > 0.7$ and RMSD < 2 Å, (ii) the equilibrium unfolded state (U−A), with $Q_n < 0.2$ and RMSD > 5 Å, which contains a large amount of helical structure, and (iii) an "unstructured" unfolded-state (U−B) chosen to have little helical structure, with $Q_{n-nn} < 0.2$, RMSD > 5 Å, and $Q_{helix} < 0.1$. We note that, although no constraint was placed on the amount of helical structure in U−

A, in practice none of the randomly selected structures had $Q_{\text{helix}} < 0.1$. In Figure 1B, we show the positions of the initial sets of configurations at 300 K in the projection on $Q_{\text{n−nn}}$, and Q_{helix}. In Table 1, we summarize the data used in this study for the construction of ME models.

Table 1. Simulation Data Used in This Study[a]

	temperature	
initial state	300 K	350 K
F	60 (50)	50 (100)
U–A	290 (250)	100 (200)
U–B	75 (300)	100 (200)

[a] We show the cumulative length of simulations in μs and the number of independent trajectories in parentheses.

Coarse State Assignment. In previous work on α-helix formation,[50,52] coarse states or "microstates" were described using the (Φ,Ψ) Ramachandran dihedral angles (see Figure 2A). In these studies, two posibilities, α-helix (A) and extended (E), were considered for defining the state of each amino acid residue. In the present case, we also explicitly represent the left-handed helix state (L), as the flipping of the Φ dihedral of Lys10 from negative to positive values (i.e., to α_L conformation) was observed to be highly correlated with folding events from more global projections[39] (see, e.g., Figure 2B). Each internal amino acid residue in every snapshot of the simulation trajectories is assigned to these A, E, and L states using transition based assignment (TBA;[52] for boundary definitions, see Figure 2A). The TBA method identifies transition paths rather than using instantaneous conformations for assigning states, which greatly reduces fast recrossings that result in non-Markovian effects in the dynamics.[52,53] The use of A, E, and L however results in a total of 3^{14} (= 4 782 969) microstates for the 16 residue uncapped peptide, an extremely large state space that would become numerically intractable. We consider only the most populated microstates in the simulations, i.e., those which cumulatively account for 95% of the data. The trajectories are then reassigned, and transitions to states outside of this most populated set are not considered. This reduces the complexity of the state space to ∼1000 microstates at 300 K and ∼2000 at 350 K while improving the statistics for the transition counts. We note that the native hairpin, i.e., state EEEEEAAALEEEEE, is the most populated microstate in the model, with a population far exceeding that of any other state.

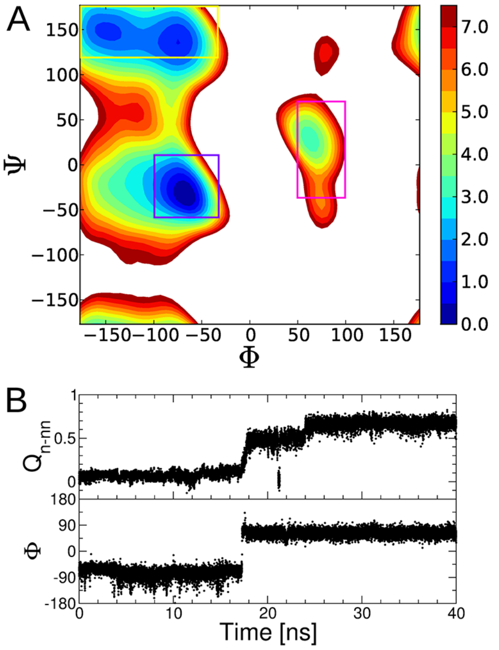


Figure 2. Transition based assignment of dihedral states. (A) Ramachandran potential of mean force averaged for the 14 (Φ,Ψ) pairs. Contours are in units of k_BT. Rectangles mark narrow regions for transition based assignment of α-helix (purple, $-100 < \Phi < -30$ and $-60 < \Psi < 10$), extended (yellow, $-180 < \Phi < -30$ and $120 < \Psi < 180$), and left-hand helix (pink, $50 < \Phi < 100$ and $-40 < \Psi < 70$) states. (B) Time series corresponding to a folding event at 300 K, as monitored in the projection on the fraction of native contacts (Q_n, top) and the Φ dihedral of Lys10.

Master Equation. We assume that the dynamics of the system can be described using a master equation of the form

$$\dot{\mathbf{P}}(t) = \mathbf{KP}(t) \tag{2}$$

where $\mathbf{P}(t)$ is the vector with the populations of each microstate at time t and $\mathbf{K}$ is a rate matrix. The solution of this equation with initial populations $\mathbf{P}_0$ is $\mathbf{P}(t) = \exp[\mathbf{K}t]\mathbf{P}_0 \equiv \mathbf{T}(t)\mathbf{P}_0$, where we have defined the transition probability matrix $\mathbf{T}(t) = \exp[\mathbf{K}t]$. In our approach, we first determine $\mathbf{T}(\Delta t)$ for a given "lag time" Δt, using the maximum likelihood estimator[54]

$$t_{ji}(\Delta t) = n_{ji}(\Delta t) / \sum_j n_{ji}(\Delta t) \tag{3}$$

Here, $n_{ji}(\Delta t)$ are the elements of the transition count matrix $\mathbf{N}(\Delta t)$, determined from the simulation trajectories as the number of transitions from microstate i to microstate j after a specified lag Δt. With thousands of states from hundreds of simulation trajectories, the resulting count matrix $\mathbf{N}$ may not be ergodic (i.e., for every pair i,j of microstates, there may not be paths from $i \rightarrow j$ and from $j \rightarrow i$). This is the case, for example, of rare microstates visited at the end of a simulation trajectory from which the peptide does not transition to any other microstates. We restrict our analysis to the largest strongly connected subset of the data, selected using Tarjan's algorithm,[55] similar to what has been done previously.[56]

An accurate estimate of $\mathbf{T}$ using eq 3 in this way requires that Δt be long enough that the dynamics appears Markovian. This is usually monitored by looking at the convergence of relaxation (or "implied") times with the chosen Δt.[57] Since we prefer to work with the rate matrix $\mathbf{K}$, we use the following approximation for $\mathbf{K}$

$$k_{ji} \approx \begin{cases} t_{ji}(\Delta t)/\Delta t & \text{for } i \neq j \\ -\sum_{j \neq i} k_{ji} & \text{for } i = j \end{cases} \tag{4}$$

The approximation in eq 4 becomes exact in the limit $\Delta t \rightarrow 0$. In Supporting Information (SI) Figure 1, we show that the relaxation times are accurate for Δt less than ∼1 ns. We have chosen to use eq 4 because, for a large number of microstates, it becomes much more computationally convenient than other methods for estimating $\mathbf{K}$, for example, stochastic optimization of $\mathbf{K}$ to maximize the likelihood of observations.[52] Although a gradient-based optimization may be feasible, we have not pursued this here. We find that we are able to get accurate results by carefully choosing Δt to be sufficiently short that the approximation in eq 4 holds but sufficiently long that the relaxation times have "converged."

We compute equilibrium populations from the right eigenvector of the stationary mode of the rate matrix (ψ_0^R) and relaxation times τ_i from its eigenvalues λ_i as $\tau_i = -1/\lambda_i$. Errors for these quantities are obtained by a bootstrap method.[58] Each bootstrap sample was generated by randomly drawing whole trajectories from the pool of simulations with repetition, until the same amount of data as in the original data set is obtained. For each sample, the analysis is done as for the original data set, and standard errors were calculated as the standard deviation of the parameter distributions over 100 bootstrap samples.

Hierarchical Clustering into Metastable States. Given the large number of microstates in the system, we produce a more intuitive description of the ME model by grouping microstates together into clusters or "macrostates". We do this clustering using the information carried by the eigenvectors of the rate matrix, namely the participation of each microstate in the aggregate transition occurring at the nth time-scale $\tau_n = -1/\lambda_n$. In practice, we gradually split the existing macrostates based on a normalized form, σ_n, of the nth left eigenvector ψ_n^L,

$$\sigma_n(i) = \frac{\psi_n^L(i) - \psi_{n,\min}^L}{\psi_{n,\max}^L - \psi_{n,\min}^L} \qquad (5)$$

We start with all the microstates grouped together in a single cluster, which is then divided into two, one formed by all microstates with $\sigma_1 \leq 0.5$ and another formed by all microstates with $\sigma_1 \geq 0.5$. This procedure is repeated for all eigenmodes with τ_n slower than a time-scale of interest by splitting into two the cluster with the largest spread of σ_n around 0.5. The sign structure of the eigenvectors could also be used for clustering, although we find σ_n to produce more metastable clusters. The approach we follow here is in fact a robust version of Perron's clustering theory.[52,59]

Committor Analysis and Flux Calculations. We use a recently developed method to calculate values for the committor (p_{fold}) and reactive fluxes from the rate matrix.[60] Among our microstates, we label a few as definitely unfolded (*UU*, with $p_{\mathrm{fold}} = 0$) and definitely folded (*FF*, with $p_{\mathrm{fold}} = 1$). For the remainder of microstates (intermediate or *I*), we define the committor as the probability that a trajectory starting from a microstate $i \in I$ will reach *FF* before reaching *UU*. This can be calculated by solving the equation

$$\sum_{i \in I} p_{\mathrm{fold}}(i) K_{ij} + \sum_{i \in FF} K_{ij} = 0 \qquad (6)$$

Using the values of p_{fold} the reactive flux is given by[60]

$$J = \sum_{i \in F^*, j \in U^*} K_{ij} p_{\mathrm{eq}}(j)[p_{\mathrm{fold}}(i) - p_{\mathrm{fold}}(j)] \qquad (7)$$

We use this procedure both for the microscopic and for the coarse-grained Master equation models.

RESULTS AND DISCUSSION

Rate Estimates from the ME Model. We have three independent sets of long MD simulations (as detailed in methods): a set of trajectories initiated from structures in the folded basin (F), a set initiated from structures chosen from an equilibrium distribution in the unfolded basin (U−A), and a set initiated from unfolded structures chosen so as to have little helical structure (U−B). We first construct an ME model using all three of these simulation data sets, one for each temperature sampled (300 and 350 K). This model includes a total of 1101 microstates at 300 K and 2326 at 350 K. In Figure 3A, we show the corresponding relaxation times (τ_i) for the 30 slowest modes, with time-scales ranging from ~10 μs to tens of nanoseconds. Here, rate matrices were constructed using a lag time Δt = 1 ns. This was chosen such that our approximation for the rate matrix is accurate (see SI Figure 1), while also being sufficiently long that the slowest relaxation times are approximately converged (see Figure 3B); i.e., the dynamics is well approximated by a Markovian model.

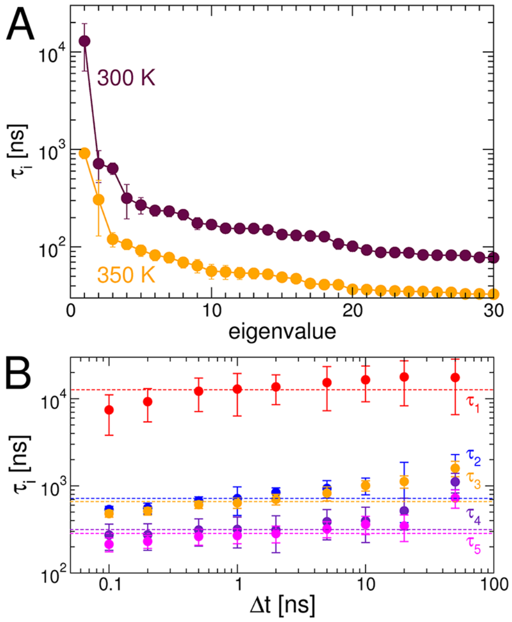


Figure 3. Eigenvalue spectrum of the rate matrix. (A) Slowest 30 relaxation times of the rate matrix at 300 and 350 K calculated for a lag time Δt = 1 ns. (B) Dependence of the relaxation times of the transition matrix $\mathbf{T}(\Delta t)$ on the lag time Δt for constructing the ME model at 300 K.

The spectrum of relaxation times (Figure 3A) reveals a large gap between the first mode (τ_1) and the next few slow modes (τ_2,τ_3,...). This separation of time scales, greater than an order of magnitude at 300 K and slightly smaller at 350 K, is characteristic of systems exhibiting two-state behavior, with the slowest time-scale being considerably greater than the others. At 300 K, the slowest relaxation time is τ_1 = 13 ± 6 μs, in reasonable agreement with the 3 μs measured in T-jump experiments at the same final temperature ($\tau_F \approx \tau_U \approx 6\ \mu s$[16]). Indeed the eigenvector for the slowest mode (ψ_1^R, see below) indicates that this process corresponds to the exchange between the native microstate and a large number of unfolded microstates.

Independence of Initial Conditions. The simplest way to estimate a folding rate k_F from a long simulation is by means of mean first passage times (MFPT), that is, the mean time τ_F^{mfpt} taken for a molecule in the unfolded state to reach the folded state. For a two-state system, the folding rate will simply be obtained as $k_F = 1/\tau_F^{\mathrm{mfpt}}$ (under the assumption that a careful assignment of the folding events will avoid counting recrossings of the barrier as transitions). For a large number N of trajectories initiated from the unfolded state, each of length t_{sim}, and where N_{fold} trajectories fold with an average folding time τ_{fold}, the maximum likelihood estimator for the MFPT is $\tau_F^{mfpt} = [N_{fold}\tau_{fold} + (N - N_{fold})t_{sim}]/N_{fold}$. A similar method can be used for unfolding times. However, when this method was applied to the long simulations used to construct our ME model (using the Q_{n-nn} coordinate to determine when the folded state was reached), there was an unexpected difference between the folding times obtained from simulations started in the U−A and U−B ensembles (59 and 8.2 μs respectively).[39] Clearly, using the "equilibrium" unfolded state U−A should be more correct for computing the first passage times. However, for a two-state protein, in which relaxation in the unfolded state ought to be fast, one might expect a similar result to be obtained for the more restricted U−B unfolded state.

We shall revisit the reason for the discrepancy in MFPT estimates for the folding time below, but first we test whether our ME model is able to overcome this limitation and yield rate estimates independent of which data set was used as input. To investigate this, we constructed the model using only the F and U−A simulation sets, or only the F and U−B simulations. We find that in both cases the slowest modes are in good agreement with the model constructed from all the data (i.e., F, U−A, and U−B; see Figure 4A). The corresponding relaxation times are in agreement within statistical error: τ_1 = 17 ± 8 μs when the rate matrix is constructed from F and U−A, and τ_1 = 10 ± 5 μs when it is built using F and U−B. These values are also consistent with the relaxation times derived from maximum likelihood of first passage times for the U−A simulation data set asuming a two-state model ($1/(k_F + k_U)$ = 18 ± 9 μs). A similar value can be obtained from the mean first passage times for the U−B data set using the slow mode of 13 ± 6 μs from the three-state model of eq 1,[39] but the ME methodology produces robust estimates of the relaxation times without the need to assume an such an *ad hoc* kinetic scheme.[39]

An even more stringent test is whether the folding and unfolding times are both in agreement with the maximum likelihood estimates obtained from first passage times for folding. In Figure 4B and C, we show a comparison of folding and unfolding times, respectively, estimated from the maximum likelihood expressions for a two-state model relaxation and from the slowest eigenvalues of the master equation. Folding and unfolding times were calculated from the relaxation times and equilibrium population of the native microstate (0.38 ± 0.12 at 300 K), assuming a two-state kinetic scheme. Again, the three ME models (considering all the data or independently treating the alternative unfolded states) result in the same rates for both folding and unfolding. Notably, the folding rate obtained from first passage times starting from U−B and assuming a two-state model yields a signficantly faster folding time. However, rate estimates from ME models are, within error, independent of the initial simulation setup.

The folding time from the full ME model (42 ± 14 μs) is much longer than the experimental estimate of ∼6 μs[16] computed using a two-state approximation from the experimentally estimated folded population and the relaxation rate after a temperature jump. There is, however, some disagreement over the exact folding midpoint of the hairpin, with some groups finding it to be ∼300 K[16,18] and others closer to ∼273 K.[19] As pointed out by Gai and co-workers,[17] using an estimated folded population of 0.5 at 298 K[16,18] would give a folding time of 6 μs, as originally reported,[16] while using a population of 0.3[19] would give a larger folding time of 10 μs. While this moves the experimental estimate closer to the simulation folding rate, there is still a significant discrepancy. Furthermore, the temperature dependence of the folding rate is also larger than in the experiment.[39] These remaining differences may be indicative of residual inaccuracies in current molecular dynamics force fields.[8] Nonetheless, this type of close comparison between experimental and simulation results should help to point the way toward future force field developments.

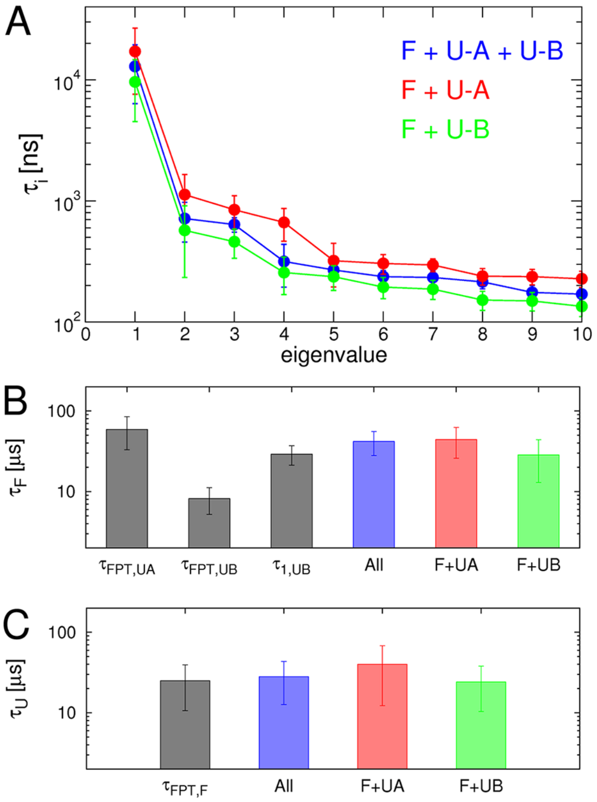


Figure 4. The ME model from different unfolded states. (A) Comparison of the slowest 10 eigenvalues of the rate matrix at 300 K calculated from different subsets of the simulation data: including the two unfolded set of runs (blue), including only F and U−A runs (red) and including only the U−B and F runs (green). (B) Folding times from maximum likelihood estimates[39] (gray) and from the ME model. $\tau_{FPT,UA}$ and $\tau_{FPT,UB}$ were calculated asuming single exponential kinetics while $\tau_{1,UB}$ was derived asuming double exponential kinetics. Estimates from the ME were calculated using all the simulation data (i.e., F + U−A + U−B, blue) or all the folded and one of the unfolded trajectory sets (F + U−A, red; F + U−B, green). (C) Same as B but for unfolding times.

Origin of First Passage Time Distributions. As discussed above, an unexpected difference in folding times was obtained from simulations started in the U−A and U−B ensembles (59 and 8.2 μs, respectively), when these were obtained from first passage times assuming a two-state model. This discrepancy can be traced to differences in the distribution of first passage times in each case, as seen in the cumulative distributions of first passage times for simulations initiated from U−A and U−B in Figure 5A, defined as time taken for the simulation to cross Q_{n-nn} = 0.7 for the first time. While the distribution for folding from U−A is single exponential, as expected for two-state folding, a faster relaxation is also present for the simulations initiated from U−B.

Can we use our ME model to explain the differences in first passage time distributions? Master equation models can be

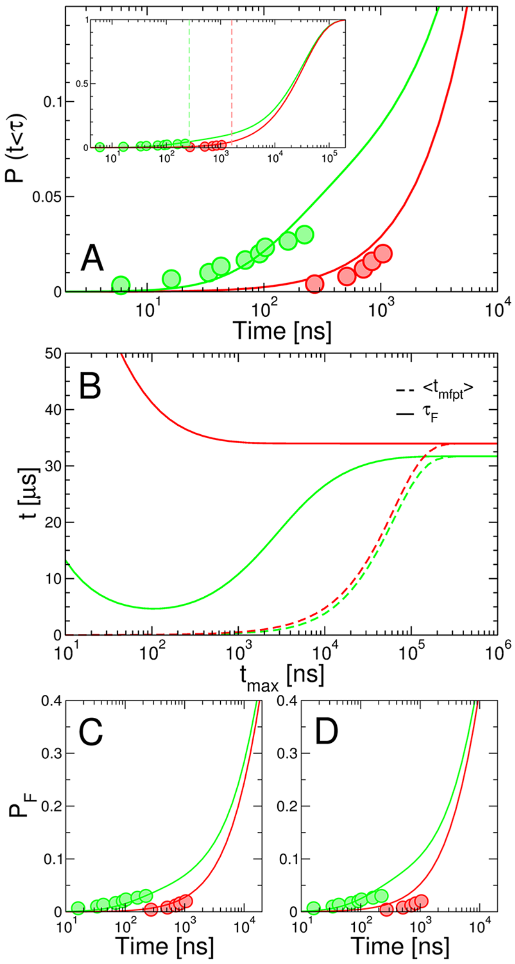


Figure 5. First passage times for folding from simulations and the ME model. (A) Cumulative distribution of unfolded state lifetimes (circles) for the simulations run from U−A (red) and U−B (green) estimated from the projection on Q_{n-nn}. Lines represent the folded population estimated from the ME model constructed using all data sets for initial populations $\mathbf{P}_{U-A}$ (red) and $\mathbf{P}_{U-B}$ (green). The inset in A shows a large range on both the time and probability axes. Dashed lines show the maximum length of the MD simulations started from U−A (light red) and U−B (light green). (B) Mean first passage times for folding, $\langle t_{mfpt}\rangle$ (broken lines) computed using only the runs which have folded, and maximum likelihood estimate of the folding time τ_F (continuous lines) estimated for a large set of runs of length t_{max} initiated from U−A (green) and U−B (green). (C) Same as A but with ME model constructed using only F+U−A. (D) Same as A but with ME model constructed using only F+U−B.

used to recover dynamics on long time scales from shorter simulations,[40] the only assumptions being that the dynamics is Markovian and that all relevant microscopic transitions are sampled. To calculate the first passage time distribution, $P_F(t)$, we construct the rate matrix $\tilde{\mathbf{K}}$, which is the same as $\mathbf{K}$, except that the folded microstate acts as a sink, so that $k_{ij} = 0$ whenever j is the folded microstate. The first passage time distribution is then given by the time evolution of the folded population, starting from the appropriate initial conditions $\mathbf{P}_{init}$ (i.e., $\mathbf{P}_{U-A}$ or $\mathbf{P}_{U-B}$), $\mathbf{P}(t) = \exp[\tilde{\mathbf{K}}t]\mathbf{P}_{init}$. To construct the initial distribution $\mathbf{P}_{init}$, we first calculate the average value of the coordinates Q_{n-nn} and Q_{helix} for each microstate in the model from snapshots of the simulation. We then define the normalized $\mathbf{P}_{U-A}$ via $P_{U-A}(i) \propto P_{eq}(i)$ if the values of Q_{n-nn} and Q_{helix} are within the boundaries of either U−A, otherwise $P_{U-A}(i) = 0$. A similar procedure was used to construct $\mathbf{P}_{U-B}$.

In Figure 5A, we show relaxations as a function of time from these initial populations. We find that the overall agreement between the ME model results and first passage times from the projection method is remarkably good. While they differ markedly at short times, the cumulative distribution of first passage times from the U−A and U−B initial conditions converge at sufficiently long times ($t \simeq 10$ µs, Figure 5A, inset). This explains the discrepancy in the first-passage-time estimate of the folding times. Mean first-passage times estimated from runs initiated in the nonequilibrium U−B distribution will be too short, if the average is computed from only the initial part of the data. This is the case for the very short (0.25 µs) simulations initiated from U−B. Only for sufficiently long trajectories would the badly chosen, nonequilibrium, set of initial conditions (U−B) approximate the true mean first passage time, as expected for a two-state model. Therefore, when using first passage time simulations to estimate rates, the individual runs need to be sufficiently long (Figure 5A, inset). To better define how long such runs need to be, we consider running a large set of short simulations of length t_{max}. After t_{max}, we expect a fraction $P_F(t_{max})$ of the runs to have folded, for which the mean first passage time would be $\langle t_{mfpt}\rangle = \int_0^{t_{max}} t\dot{P}_F(t)\,dt/\int_0^{t_{max}} \dot{P}_F(t)\,dt$. In Figure 5B, we plot $\langle t_{mfpt}\rangle$ calculated from the rate matrix $\tilde{\mathbf{K}}$, as well as the maximum likelihood estimate of the folding time, assuming a single exponential (two-state) first passage time distribution, $\tau_F = (\langle t_{mfpt}\rangle P_F(t_{max}) + t_{max}(1 - P_F(t_{max})))/P_F(t_{max})$. The maximum likelihood estimate of the folding time obtained from the U−A initial conditions converges within ∼200 ns to the correct value, as expected given the nearly single-exponential relaxation. However, the maximum likelihood estimate from U−B, while still much better than $\langle t_{mfpt}\rangle$, requires ∼20 µs, on the order of the folding time itself, to become accurate.

We note that the predictions of the ME model for the first passage time distributions are independent of the data sets used to derive the model (Figure 5C,D). Even if we calculate first passage times using the master equation constructed from only U−A and F, we can still recover the first passage time distributions started from U−B (Figure 5C). Similarly the first passage times starting from U−A can be obtained from the model built from U−B and F (Figure 5D). This test further establishes the ability of ME models to recover correct kinetics regardless of the specific setup of the MD simulations.

Microscopic Interpretation of the Kinetics. In addition to recovering the first passage time distributions, the master equation model can reveal their microscopic origin, via the modes which contribute the most to the observed decay. The evolution of population can in general be written as a sum over exponential terms, using the spectral decomposition of the matrix exponential

$$\mathbf{P}(t) = \sum_{n=0}^{N-1} \psi_n^R[\psi_n^L \cdot \mathbf{P}_{init}]e^{\lambda_n t} \tag{8}$$

In this expression, ψ_n^R and ψ_n^R are respectively the nth left and right eigenvectors of $\mathbf{K}$, and λ_n represents the corresponding eigenvalues. Thus, the evolution of the folded population can be written as $P_F(t) = \Sigma_n A_n \exp[\lambda_n t]$, with the amplitudes given by $A_n = \psi_n^R(F)[\psi_n^L \cdot \mathbf{P}_{init}]$, where $\psi_n^R(F)$ is the value of ψ_n^R for the native microstate. This can also be viewed as a special case of the "dynamical fingerprint" approach.[50,61]

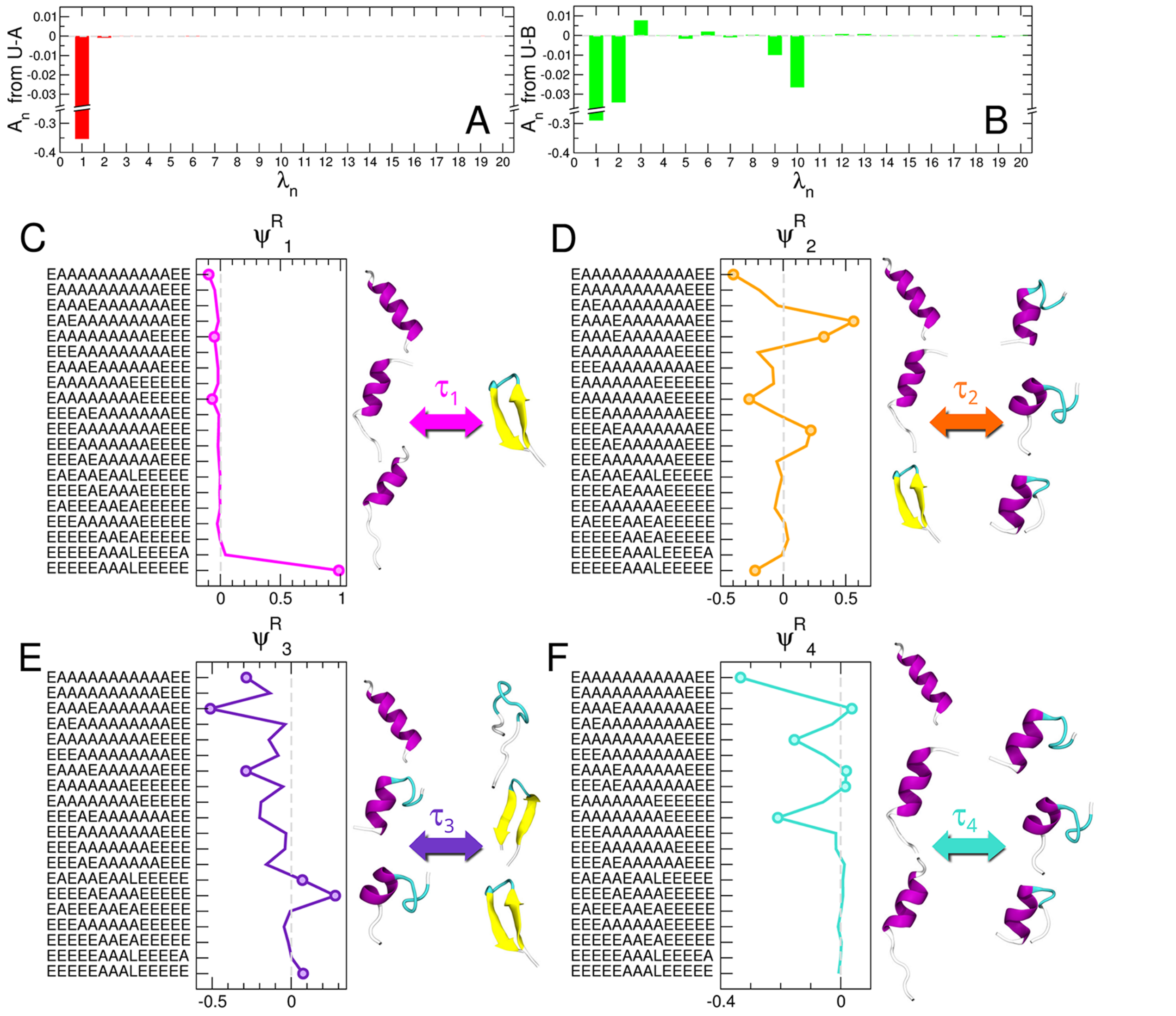


Figure 6. Contributions of different modes to the change in fraction folded. (A) Amplitudes for the 20 slowest eigenmodes when the dynamics are propagated from an initial U−A distribution. (B) Same as A but for an initial U−B distributions. (C−F) Right eigenvectors for the four slowest modes. For each mode, we show only the values of the eigenvector for the microstates with largest weights. The dashed gray line in each plot marks the value of $\psi_n^R = 0$. For each eigenmode, we show cartoon representations of some of the microstates that exchange more population, which we indicate as circles in the eigenvector plot.

In Figure 6A,B, we show the amplitudes A_n representing the contributions of the slowest modes to the first passage time distributions discussed in the previous section, i.e., starting with $\mathbf{P}_{\text{init}} = \mathbf{P}_{\text{U−A}}$ or $\mathbf{P}_{\text{init}} = \mathbf{P}_{\text{U−B}}$. We see that the amplitude of the slowest mode (λ_1) is dominant regardless of the initial condition. This is consistent with the greater importance of this mode in the proposed phenomenological three-state model[39] and as evident in Figure 5. However, when we use the nonhelical, nonequilibrium initial condition given by $\mathbf{P}_{\text{U−B}}$, we obtain significant contributions from other slow modes. Notably, the second mode λ_2 has a contribution with the same sign as that of λ_1 (Figure 6B), consistent with the fast 700-ns relaxation that was observed for the simulations started in U−B (Figure 5).

In Figure 6C−F, we show which states undergo the largest exchanges of population in each of the slow modes, as given by the right eigenvectors ψ_n^R. For the slowest mode $n = 1$ with a relaxation time of $\tau_1 = 13 \pm 6$ μs, we obtain the expected exchange corresponding to two-state folding, in which structured, helical unfolded microstates interchange with the native microstate (Figure 6C). The next few modes, with timescales between 300 and 800 ns, reveal more complex dynamics. For example, the second mode, with $\tau_2 = 710 \pm 30$ ns, corresponds to interconversion of less helical microstates, e.g., EAAAEAAAAAAAEE, with either more helical microstates, like EAAAAAAAAAAAEE, or the native microstate EEEEEAAALEEEEE. This is completely consistent with the fast component in the first passage time distribution, which can be attributed to less structured unfolded microstates either forming more helical structure or converting to the folded state. The fourth mode describes the helix−coil dynamics in the unfolded state, but as the coefficient of the folded microstate is ~0, it of course makes no contribution to the folding amplitude (see Figure 6F). Thus, we can directly resolve each component of the first-passage time distribution in terms of the microscopic dynamics.

Coarse-Graining Reveals Slow Unfolded State Dynamics. The existence of a "spectral gap" in the spectrum of

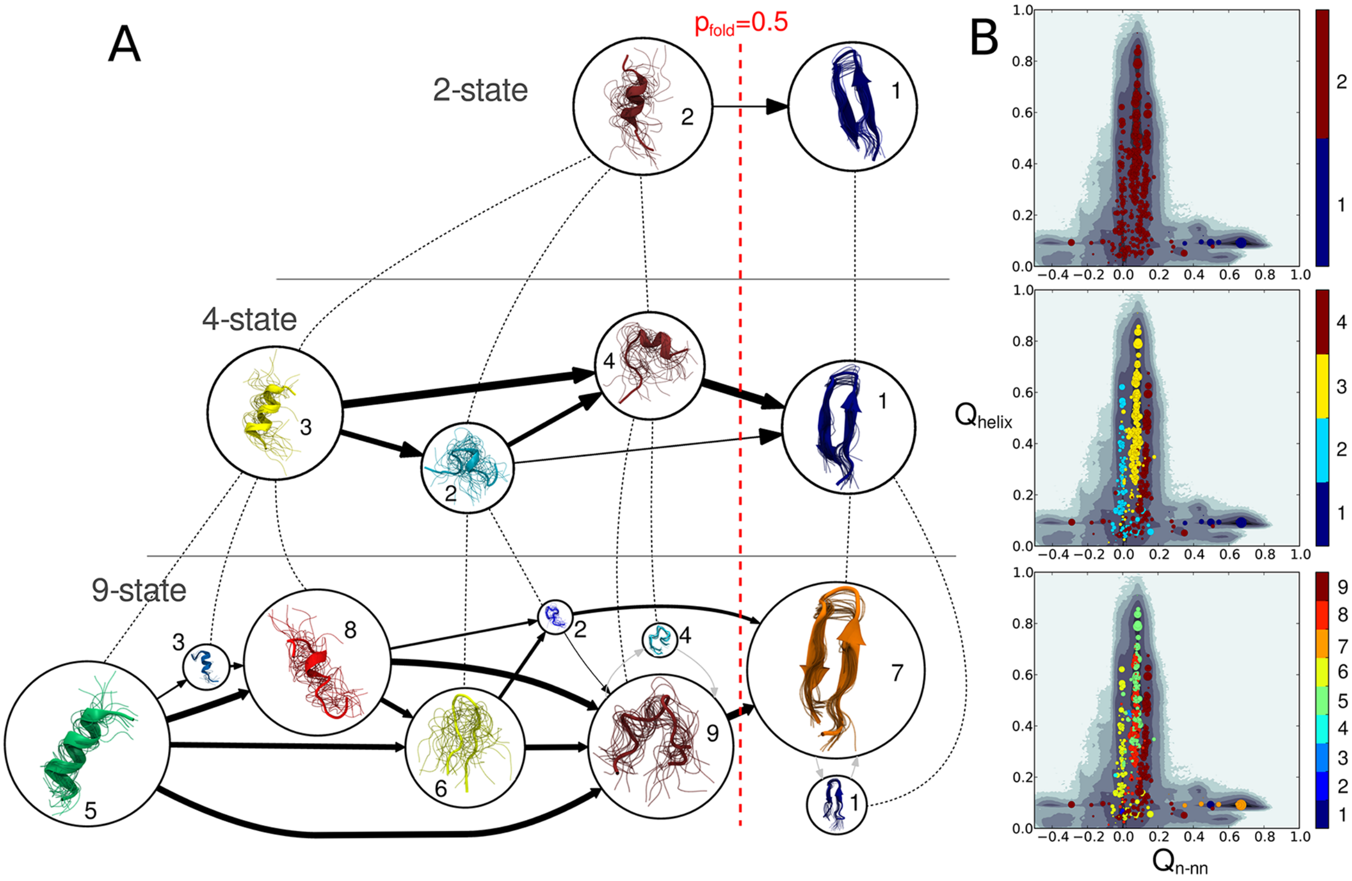


Figure 7. Hierarchical clustering of the ME model. (A) Network diagrams for two (top), four (center), and nine (bottom) clusters. For each clustering, the area of the circles is approximately proportional to the population of the macrostates they represent, on a log scale. The thickness of the black continuous lines is proportional to the logarithm of the reactive flux from unfolded to folded between different species. Gray lines are shown to mark the connectivity between clusters, where the connections do not contribute to reactive flux. Dashed lines indicate the hierarchical relationship between macrostates at different levels of coarse graining. Representative structures for each macrostate are shown in ribbon representation, with each cluster assigned a different color. A dashed red line separates macrostates with $p_{\mathrm{fold}} < 0.5$ (left) and macrostates with $p_{\mathrm{fold}} > 0.5$ (right). (B) Projection of the microstates on the Q_{n-nn} vs Q_{helix} potential of mean force for the three coarse grainings. In this case, the size of the circles representing the microstates is proportional to the population of each microstate, on a log scale. Colors correspond to the macrostate index as in A.

relaxation times in the microscopic ME model suggests that the dynamics can be well approximated by grouping the ∼1000 states in our model into relatively few clusters or macrostates. The gaps would correspond to a separation of time scales between the fast internal dynamics within each macrostate and the slow dynamics between different macrostates. Here, we present a hierarchy of coarse grainings based on gaps in the eigenvector spectrum for the GB1 hairpin. We use a simple eigenvector-based clustering, though we note that these methods are currently an area of active development.[62−65]

As explained above, the first (and largest) gap is found after the first mode λ_1 (see Figure 3), as corresponds to a two-state system. The next gap in the spectrum of eigenvalues is found after the fourth mode, with another smaller gap after the eighth mode. In Figure 7A, we represent the resulting networks, corresponding respectively to two, four, and nine macrostates. We also show a projection of the microstates on the two-dimensional PMF using average values of Q_{n-nn} and Q_{helix}, indicating which microstates are grouped into the corresponding sets of macrostates.

The first, "two-state", clustering (Figure 7, top) clearly shows that virtually all states with high and low average Q_{n-nn} values are grouped, respectively, into the native state and unfolded states. Thus, the Q_{n-nn} coordinate can cleanly separate the two stable states and so is a good coordinate, in this sense at least.

The next clustering into four macrostates (Figure 7, center) is constructed from the two-state clustering using eigenmodes with relaxation times of ≃600 ns and resolves more details of the slow unfolded state dynamics. We find that microstates with high helicity in macrostate 1 (e.g., EAAAAAAAAAAAEE, EAAAAAAAAEEEEE), do not interconvert directly with the native. Instead, transitions occur via microstates with broken helical stretches either at position 5 (e.g., EAAAEAA-AAAAAEE) or 6 (e.g., EAEAAEAALEEEEE, EEEEAEA-AAEEEEE), which are grouped respectively into macrostates 4 and 2. We note that most of the flux actually goes through macrostate 4, where helix breaking has occurred at position 5, i.e., the last residue on the first β-strand of the native structure of the GB1 hairpin. As shown above, helix breaking seems to be a slow process en route to folding.

Finally, we show the nine macrostate clustering, constructed from the four macrostate model using relaxation times in the ∼100 ns range (Figure 7, bottom). The global picture remains consistent with the gradual helix melting from the high helical states to the β hairpin and the existence of parallel pathways for folding. However, at this greater level of detail, we obtain a finer resolved model with apparent off-pathway macrostates. The interconversion between clusters 1 and 7 corresponds to exchange between the native hairpin (i.e., EEEEEEAA-ALEEEEE) and microstates like with an α-helical Thr13 (e.g., EEEEEAAALEEAEE or AEEEEAAALEEAEE) which results in hairpin fraying. The interconversion between clusters 9 and the off-pathway cluster 4 corresponds to the exchange between microstates EAAAEAAEEELEEE and EAAAEAAEEELEEA with their precursor EAAAEAAEEEEEEE, respectively, i.e. a flipping from extended to left-handed for Phe12. In both cases, the off pathway clusters are slow conformational excursions from another cluster. By virtue of being off-pathway, these macrostates have the same value of the committor as the parent macrostate to which they are connected and do not contribute to the reactive flux.

Microstate Committors and Putative Transition States. Although the coarse grained clusters help to resolve metastable macrostates, in order to obtain finer detail on the folding mechanism, including transiently populated states, it is necessary to use the microscopic master equation, including all ~1000 states. From a mechanistic point of view, the most important configurations for a two-state protein are the folding transition states, those structures with a committor, or p_{fold}, value of 0.5 (the committor is the probability a trajectory initiated with velocities randomly chosen from a Maxwell—Boltzmann distribution will first reach the folded state, rather than the unfolded state). It is possible to compute analytically the committors for each microstate in our model, based on the rate matrix, using the Berezhkovskii—Hummer—Szabo method.[60] For this purpose, we define as definitely unfolded (UU) microstates (i.e., for which $p_{\text{fold}} \equiv 0$) those from U—A with high equilibrium probabilities ($P_{\text{eq}}(i) > 5.5 \times 10^{-3}$). The native string (with $p_{\text{fold}} \equiv 1$) is defined as definitely folded (FF). Having made this choice, we compute the committors for the remaining microstates from the rate matrix.[60] In Figure 8A, we show the distribution of p_{fold} values, which is extremely bimodal, with very few microstates having intermediate values. We note that although we report results for a rate matrix derived with a lag time $\Delta t = 1$ ns, the distribution is very similar for the shortest lag time possible (100 ps). The origin of the bimodality in the p_{fold} distribution is that p_{fold} generally changes very sharply from 0 to 1 in the vicinity of the transition state. Thus, we find very few microstates with p_{fold} near 0.5. A likely reason for this is that our microstate definitions are too broadly defined to capture transition states exclusively. Achieving that goal would probably require the inclusion of other degrees of freedom, e.g., hydrogen bonds, in the definition of the microstates.

Those microstates with intermediate values of the committor (i.e., $p_{\text{fiod}} \simeq 0.5$), although not true transition states, can nonetheless be regarded as reactive states. On the unfolded side of the barrier ($0.3 < p_{\text{fold}} < 0.5$), we find two hairpin-like coarse-states that have a flipped dihedral at Phe12 (corresponding to the sequence of dihedrals EEEEEAAALEAEEE and EEEEEAA-ALEAEEA). Although quite similar to the native conformation ($\langle Q_{n-nn} \rangle = 0.5$), these microstates are stabilized by non-native hydrogen bonds (see Figure 8B and D). On the folded side of the barrier ($0.5 < p_{\text{fold}} < 0.7$), we find a microstate with all internal dihedrals in the native conformation but with frayed ends (EAAEEAAALEEEEE, see Figure 8C). In this conformation, two of the native hydrogen bonds are correctly formed. This type of conformation was also proposed as a transition state in the simple Ising-like model of GB1 proposed by Muñoz et al.[16] In fact, all the microstates with $p_{\text{fold}} > 0.3$ share in common six native-like internal dihedrals in the turn region (—EAAALE—, from Asp6 to Thr11). This finding supports the kinetic zipper mechanism proposed by Muñoz et al.[16] and supported by experimental kinetic data for mutants.[17,66,67]

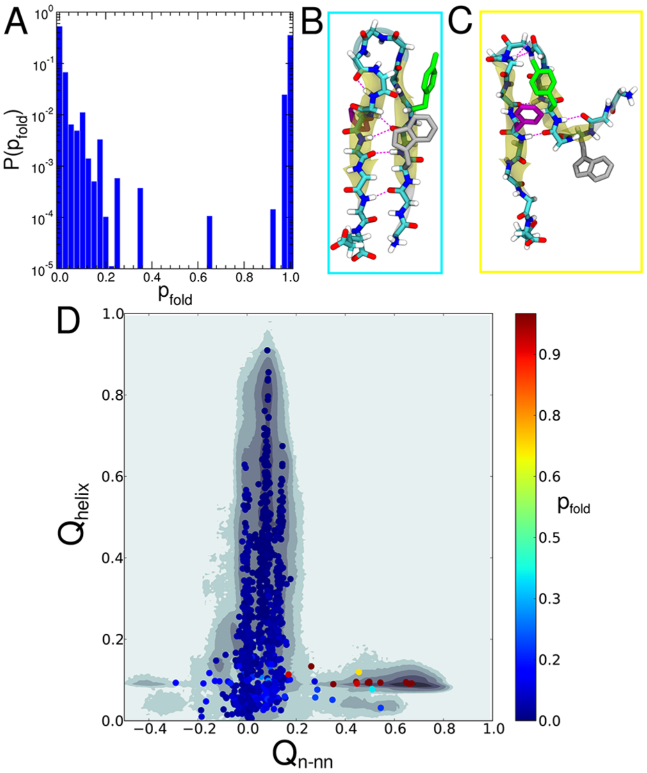


Figure 8. Commitment probabilities of the GB1 hairpin. (A) Probability distribution of p_{fold} values for individual microstates at $\Delta t = 1$ ns. (B) Most representative conformation for the EEEEEA-AALEAEEE state ($p_{\text{fold}} = 0.35$). (C) Most representative conformation for the EAAEEAAALEEEEE state ($p_{\text{fold}} = 0.67$). (D) p_{fold} values for all the coarse-states shown on the projection on Q_{n-nn} and Q_{helix}. The color of the frames in B and C correspond to the color scale in D.

■ CONCLUSIONS

In the simulations reported, as in the experiment, the GB1 hairpin is a clear two-state folder, with a separation of time-scales between the slowest and next slowest modes of over an order of magnitude. In spite of this property, attempts to estimate the folding rate based on many short folding simulations gave different results, depending on the initial conditions chosen. By constructing a master equation model with microstates defined according to backbone torsion angles, we were able to obtain folding rates without a dependence on the initial conditions chosen and were able to explain the origin of the different results obtained based on simple rate estimates using first-passage times from different initial conditions. The key point is that to obtain reliable rates from many MD simulations, one should either make sure to start from an equilibrium distribution of initial conditions or run the simulations for long enough that short initial processes do not contribute excessively to the true rate. Alternatively, using master equation, or Markov state, models, one can also directly obtain the two-state folding rate.

In this paper, we additionally provide microscopic insight into the mechanism of folding of the GB1 hairpin. We find evidence for the hairpin folding via a "zipper" mechanism, with the putative transition state from our model being very similar to that proposed by Muñoz and co-workers[16] based on Ising-like models for hairpin folding. By clustering the ME model into a network of macrostates, we are able resolve the unfolded state dynamics of the peptide into interconversions between a number of structured states. We show that folding seems to proceed only after the helical stretches of the unfolded state are broken at the center. The relaxation times within the unfolded state are relatively slow (≃ 100 ns) processes.

The importance of some of our findings, particularly those related to the high helical population in the unfolded state, are of course subject to the validity of the Amber ff03* force field, as no experimental evidence of helical states has been reported for the GB1 hairpin. Other force fields which obtain a similar folded fraction for the hairpin at 300 K can have substantially different unfolded states. For example, the helix fraction in the unfolded ensemble is much smaller with the Amber ff99SB force field[68] than with Amber ff03* and almost undetectable with the OPLS-AA/L force field[69] with SPC water.[70] These differences between the equilibrium properties as calculated for different force fields would of course be expected to result in differences in unfolded state dynamics. We note, however, that the Amber ff03* force field was adjusted to correct for an excessive helical propensity,[3] and it has been shown to fold both alpha and beta proteins[71] and, for the case of the GB1 hairpin, matched experimental NMR chemical shifts and FRET efficiencies.[36] Among such force fields with a good balance between α and β structure (e.g., Amber ff99SB*, ff03*), the differences in the unfolded ensemble are smaller.[70]

A limitation of the present kinetic model is the reliance on only backbone torsion angles to define the microstates. A more inclusive description would add also other slow degrees of freedom, for example native hydrogen bonds or native contacts. While the Ramachandran based-discretization works well for helical peptides and for the GB1 hairpin, it will probably be insufficient for larger peptides and proteins. For these larger systems, a contact-map based clustering may prove useful.[72]

ASSOCIATED CONTENT

Supporting Information

Details on the analysis of the simulations, figure showing quality of rate matrix approximation, table giving coarse states for nine-state model. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: robertbe@helix.nih.gov.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

D.D.S was supported by a FEBS Long Term Post Doctoral Fellowship and R.B.B. by a Royal Society University Research Fellowship. The authors acknowledge Prof. Victor Muñoz for granting access to computing resources. This work used the Extreme Science and Engineering Discovery Environment, which is supported by National Science Foundation grant no. MCB-120014.

REFERENCES

(1) Best, R. B.; Buchete, N.-V.; Hummer, G. *Biophys. J.* **2008**, *95*, L7−L9.

(2) Freddolino, P. L.; Park, S.; Roux, B.; Schulten, K. *Biophys. J.* **2009**, *96*, 3772−3780.

(3) Best, R. B.; Hummer, G. *J. Phys. Chem. B.* **2009**, *113*, 9004−9015.

(4) Lange, O. F.; van der Spoel, D.; de Groot, B. L. *Biophys. J.* **2010**, *99*, 647−655.

(5) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950−1958.

(6) Beauchamp, K.; Lin, Y.-S.; Das, R.; Pande, V. *J. Chem. Theory Comput.* **2012**, *8*, 1409−1414.

(7) Best, R.; De Sancho, D.; Mittal, J. *Biophys. J.* **2012**, *102*, 1462−1467.

(8) Best, R. *Curr. Opin. Struct. Biol.* **2012**, *22*, 52−61.

(9) Eaton, W.; Muñoz, V.; Hagen, S.; Jas, G.; Lapidus, L.; Henry, E.; Hofrichter, J. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 327−359.

(10) Kubelka, J.; Hofrichter, J.; Eaton, W. A. *Curr. Opin. Struct. Biol.* **2004**, *14*, 76−88.

(11) Muñoz, V.; Henry, E.; Hofrichter, J.; Eaton, W. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 5872−5879.

(12) Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 11311−11316.

(13) Lindorff-Larsen, K.; Piana, S.; Dror, R.; Shaw, D. *Science* **2011**, *334*, 517−520.

(14) Gronenborn, A.; Filpula, D.; Essig, N.; Achari, A.; Whitlow, M.; Wingfield, P.; Clore, G. *Science* **1991**, *253*, 657−661.

(15) Blanco, F.; Rivas, G.; Serrano, L. *Nat. Struct. Mol. Biol.* **1994**, *1*, 584−590.

(16) Muñoz, V.; Thompson, P.; Hofrichter, J.; Eaton, W. *Nature* **1997**, *390*, 196−199.

(17) Du, D.; Zhu, Y.; Huang, C.-Y.; Gai, F. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 15915−15920.

(18) Honda, S.; Kobayashi, N.; Munekata, E. *J. Mol. Biol.* **2000**, *295*, 269−278.

(19) Olsen, K. A.; Fesinmeyer, R. M.; Stewart, J. M.; Andersen, N. H. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 15483−15487.

(20) Davis, C. M.; Xiao, S.; Raleigh, D. P.; Dyer, R. B. *J. Am. Chem. Soc.* **2012**, *134*, 14476−14482.

(21) Klimov, D.; Thirumalai, D. *Proc. Natl. Acad. Sci. U. S. A.* **2000**, *97*, 2544−2549.

(22) Ma, B.; Nussinov, R. *J. Mol. Biol.* **2000**, *296*, 1091−1104.

(23) García, A. E.; Sanbonmatsu, K. Y. *Proteins* **2001**, *42*, 345−354.

(24) Zhou, R.; Berne, B. J.; Germain, R. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 14931−14936.

(25) Zagrovic, B.; Sorin, E. J.; Pande, V. *J. Mol. Biol.* **2001**, *313*, 151−169.

(26) Bolhuis, P. G. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 12129−12134.

(27) Wei, G.; Mousseau, N.; Derreumaux, P. *Proteins* **2004**, *56*, 464−474.

(28) Evans, D. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *121*, 1080−1090.

(29) Krivov, S. V.; Karplus, M. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 14766−14770.

(30) Andrec, M.; Felts, A. K.; Gallicchio, E.; Levy, R. M. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6801−6806.

(31) Nguyen, P. H.; Stock, G.; Mittag, E.; Hu, C.-K.; Li, M. S. *Proteins* **2005**, *61*, 795−808.

(32) Bussi, G.; Gervasio, F. L.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2006**, *128*, 13435−13441.

(33) Weinstock, D. S.; Narayanan, C.; Felts, A. K.; Andrec, M.; Levy, R. M.; Wu, K.-P.; Baum, J. *J. Am. Chem. Soc.* **2007**, *129*, 4858−4859.

(34) Yoda, T.; Sugita, Y.; Okamoto, Y. *Proteins* **2007**, *66*, 846−859.

(35) Bonomi, M.; Branduardi, D.; Gervasio, F. L.; Parrinello, M. *J. Am. Chem. Soc.* **2008**, *130*, 13938−13944.

(36) Best, R. B.; Mittal, J. *J. Phys. Chem. B.* **2010**, *114*, 8790−8798.

(37) Berteotti, A.; Barducci, A.; Parrinello, M. *J. Am. Chem. Soc.* **2011**, *133*, 17200−17206.

(38) Shirts, M.; Pande, V. *Science* **2000**, *290*, 1903−1904.

(39) Best, R. B.; Mittal, J. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 11087−11092.
(40) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011−19016.
(41) Bowman, G.; Pande, V. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 10890−10895.
(42) Waldauer, S.; Bakajin, O.; Lapidus, L. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 13713−13717.
(43) Voelz, V.; Singh, V.; Wedemeyer, W.; Lapidus, L.; Pande, V. *J. Am. Chem. Soc.* **2010**, *132*, 4702−4709.
(44) Voelz, V. A.; Jager, M.; Yao, S.; Chen, Y.; Zhu, L.; Waldauer, S. A.; Bowman, G. R.; Friedrichs, M.; Bakajin, O.; Lapidus, L. J.; Weiss, S.; Pande, V. S. *J. Am. Chem. Soc.* **2012**, *134*, 12565−12577.
(45) Lapidus, L. J. *Curr. Opin. Struct. Biol.* **2012**, *23*, in press.
(46) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102−106.
(47) Snow, C.; Sorin, E.; Rhee, Y.; Pande, V. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 43−69.
(48) Fersht, A. R. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 14122−14125.
(49) Paci, E.; Cavalli, A.; Vendruscolo, M.; Caflisch, A. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 8217−8222.
(50) De Sancho, D.; Best, R. B. *J. Am. Chem. Soc.* **2011**, *133*, 6809−6816.
(51) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(52) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057−6069.
(53) Schutte, C.; Noe, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. *J. Chem. Phys.* **2011**, *134*, 204105.
(54) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schutte, C.; Noe, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(55) Tarjan, R. *SIAM J. Comput.* **1972**, *1*, 146−160.
(56) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.
(57) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571−6581.
(58) Efron, B. *Ann. Stat.* **1979**, *7*, 1−26.
(59) Deuflhard, P.; Huisinga, W.; Fischer, A.; Schütte, C. *Linear Algebra Appl.* **2000**, *315*, 39−59.
(60) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J. Chem. Phys.* **2009**, *130*, 205102.
(61) Noé, F.; Doose, S.; Daidone, I.; Löllmann, M.; Sauer, M.; Chodera, J.; Smith, J. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 4822−4827.
(62) Noe, F.; Horenko, I.; Schutte, C.; Smith, J. C. *J. Chem. Phys.* **2007**, *126*, 155102.
(63) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J. Chem. Phys.* **2007**, *126*, 155101.
(64) Beauchamp, K. A.; McGibbon, R.; Lin, Y.-S.; Pande, V. S. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17807−17813.
(65) Bowman, G. R. *J. Chem. Phys.* **2012**, *137*, 134111.
(66) Du, D.; Tucker, M. J.; Gai, F. *Biochemistry* **2006**, *45*, 2668−2678.
(67) Culik, R. M.; Jo, H.; DeGrado, W. F.; Gai, F. *J. Am. Chem. Soc.* **2012**, *134*, 8026−8029.
(68) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712−725.
(69) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474−6487.
(70) Best, R. B.; Mittal, J. *Proteins* **2011**, *79*, 1318−1328.
(71) Mittal, J.; Best, R. B. *Biophys. J.* **2010**, *99*, L26−L28.
(72) Kellogg, E. H.; Lange, O. F.; Baker, D. *J. Phys. Chem. B* **2012**, *116*, 11405−11413.