Atomistic folding simulations of the five helix bundle protein $\lambda_{6\text{-}85}$ supporting information

Gregory R. Bowman[1], Vincent A. Voelz [1], and Vijay S. Pande[1,2*]

[1]Department of Chemistry and [2]Biophysics Program, Stanford University, Stanford, CA 94305

Materials & Methods

Simulation Details

Six initial starting conformations covering a range of 0 to 13 Å C_α RMSD to the crystal structure were drawn from replica exchange simulations in implicit solvent from Bill Swope and Jed Pitera at the IBM Almaden Research Center.[1] These conformations were energy minimized using a steepest-descents algorithm in the Gromacs simulation package[2] with the AMBER03 force field.[3] They were then solvated in an octahedral box of dimensions 6.53 nm by 6.15 nm by 5.33 nm with 6,754 tip3p waters (with one Cl^- atom to neutralize the charge). The relatively small box size was chosen based on previous studies showing that this size should give identical results to a much larger box.[4] This equivalence is supported by the fact that only ~0.004% of the conformations sampled interact with their periodic image. The solvent was equilibrated at 300 K with the protein coordinates held fixed. Finally, simulations were run on the Folding@home distributed computing platform using an MPI-enabled version of Gromacs[4] at both 300 and 370 K. These simulations used a 2 fs time step. They used a grid based neighbor list with a 0.7 nm cutoff that was updated every 10 steps. All h-bonds were constrained with the SHAKE algorithm.[5] The system was held at constant volume during production runs but the equilibration simulations for each starting structure used a Parrinello-Rahman barostat[6] at 1 bar with a time constant of 10 ps and a compressibility of $4.5*10^{-5}$ bar^{-1}. Reaction field with a continuum dielectric of 78 was used for long-range electrostatics. A cutoff of 0.8 nm was used for both Coulombic and Van der Waals interactions. A switch at 0.7 nm was also used for Van der Waals interactions. The protein and solvent were coupled separately to a Nose-Hoover thermostat[7,8] with an oscillation period of 0.5 ps. The linear center-of-mass motion of the system was removed every 10 steps. Random initial velocities were drawn from a Maxwell-Boltzmann distribution at 370 K. Protein conformations were stored every 50 ps.

Most of the results described in this work are from the 370 K data as this should best correspond to the experimental temperature. The experimental T-jump study of lambda repressor was conducted at 334 K,[9] just short of the melting temperature of 347 K. However, simulations with fixed-charge force fields are known to have poor temperature dependence. For example, simulations of systems like lambda repressor tend to over-estimate melting temperatures by about 10%. [1] Thus, the experimental temperature of 334 K will be best modeled by a simulation temperature of ~370 K.

Structures were rendered with PyMOL.

MSM Construction and Analysis

We used the MSMBuilder package[10,11] to construct a microstate model with 30,000 states and a coarse-grained macrostate model with 5,000 states. The microstate model was generated by clustering conformations stored at 5 ns intervals based on their C_α RMSDs using the k-centers algorithm in MSMBuilder. The remaining data (50 ps spacing) was then assigned to these clusters and used to construct a transition count matrix (C_{ij} = the number of observed transition from state i at time t to state j at time t+τ, where τ is the lag time of the model) and corresponding transition probability matrix (P_{ij} = probability of transitioning from state i at time t to state j at time t+τ, where τ is the lag time of the model). The PCCA+ algorithm[12-14] was then used to lump kinetically related microstates into 5,000 macrostates and these state definitions were used to construct macrostate level transition count/probability matrices.

The lag time for each model was selected by computing the implied timescales of the model

$$k = \frac{-\tau}{\ln(\mu)}$$

where μ is an eigenvalue, τ is the lag time, and k is a rate. This equation comes from the equivalence between discrete time MSMs and continuous time master equations (see Refs [15] and [16] for details). By plotting the implied timescales as a function of the lag time one can identify the lag time at which they begin to level-off (satisfy the Chapman-Kolmogorov test), indicating that the model is Markovian.[17] Based on this analysis, we chose a lag time of 5 ns for our microstate model (Figure S2), where all the kinetic analyses in this work were performed.

To calculate the relaxation of the fraction folded as measured by some observable we used the procedure from Ref 4 to distinguish folded and non-native states and the procedure from Ref [11] to propagate the fraction folded. For example, with the experimental surrogate (Trp22-Tyr33 quenching) we calculated the average and standard deviation of the distance between these residues ($\mathrm{Native}_{\mathrm{ave}}$ and $\mathrm{Native}_{\mathrm{std}}$ respectively) in native-state simulations started from a model of D14A based on the 1LMB crystal structure. Five random conformations were drawn from each state and used to calculate the average distance between these residues for that state ($\mathrm{State}_{\mathrm{ave}}$). A state was considered to be native if $\mathrm{State}_{\mathrm{ave}} < \mathrm{Native}_{\mathrm{ave}} - \mathrm{Native}_{\mathrm{std}}$ and non-native otherwise. The fraction-folded can then be calculated as the dot product between a vector with 1's for folded states and 0's for non-native ones with the state populations. To mimic an ensemble T-jump we used two starting populations: 1) all states equally populated and 2) all microstates in non-native macrostates (i.e. outside the most populated macrostate) equally populated. The relaxation of these starting ensembles was modeled by propagating the populations forward in time with the transition probability matrix and calculating the fraction folded at each time step. The same procedure was used for the fraction folded determined by the RMSD to the crystal structure, which was examined to determine whether or not the Trp22-Tyr33 distance could be measuring a more local rearrangement than full folding, as proposed for villin.[4] Figure S10 shows that these two observables gave similar timescales for the full MSM and, while differences are apparent when the simulations started from β–sheet structures are ignored, the timescales do not appear to be substantially slower for the RMSD relaxation (Figure S11). The fast and slow timescales (τ_f and τ_s respectively) were obtained by fitting to the biexponential

$$Ae^{\left(-t/\tau_f\right)} + Be^{\left(-t/\tau_s\right)} + C$$

where t is the time and A, B, and C are constants. Fitting to a single exponential requires three parameters. A biexponential fit only requires two extra parameters but improves the agreement between the model and raw data by about a factor of two. Adding a third exponential (two extra parameters beyond a biexponential) does not significantly improve the agreement between the model and data. For example, a single exponential fit to the relaxation of the RMSD has a root-mean-squared (RMS) deviation of $2.21*10^{-6}$ from the raw data while both biexponential and triexponential fits have an RMS deviation of $1.38*10^{-6}$. The experimental results were also fit with biexponentials, so using them here facilitates comparing to experiment.

The states participating most strongly in a given transition mode are specified by the corresponding left eigenvector (states with negative components are interconverting with those with positive components, and the magnitude of the eigenvector component gives the degree of participation).[18] The highest flux pathways between sets of states were calculated as in Refs [19] and [20]. Mean First Passage Times (MFPTs) between states and P_{folds} were calculated as in Ref [21].

Given our finite sampling, one can estimate the kinetic connectivity of a state by counting the number of edges connecting it to other states (effectively a way of counting the number of edges with probabilities above some threshold since all connections would be made with infinite sampling).

Two residues are considered to be in contact if their C_α atoms are within 7 Å and they are at least 3 residues apart in the sequence. Native contacts are those formed in the energy-minimized model based on the crystal structure 1LMB.[22,23] The distance between two residues is the distance between the centroids of their side chains.

Relative contact orders (RCOs) were calculated to quantify the degree of local versus non-local contacts in various states.[24] The RCO is defined as

$$RCO = \frac{1}{L \cdot N} \sum_{i<j}^{N} \Delta S_{ij}$$

where N is the total number of contacts, L is the number of residues in the protein, and ΔS_{ij} is the sequence separation (in residues) between contacting residues i and j. Here, two residues are considered to be in contact if their C_α atoms are within 7 Å regardless of their sequence separation.

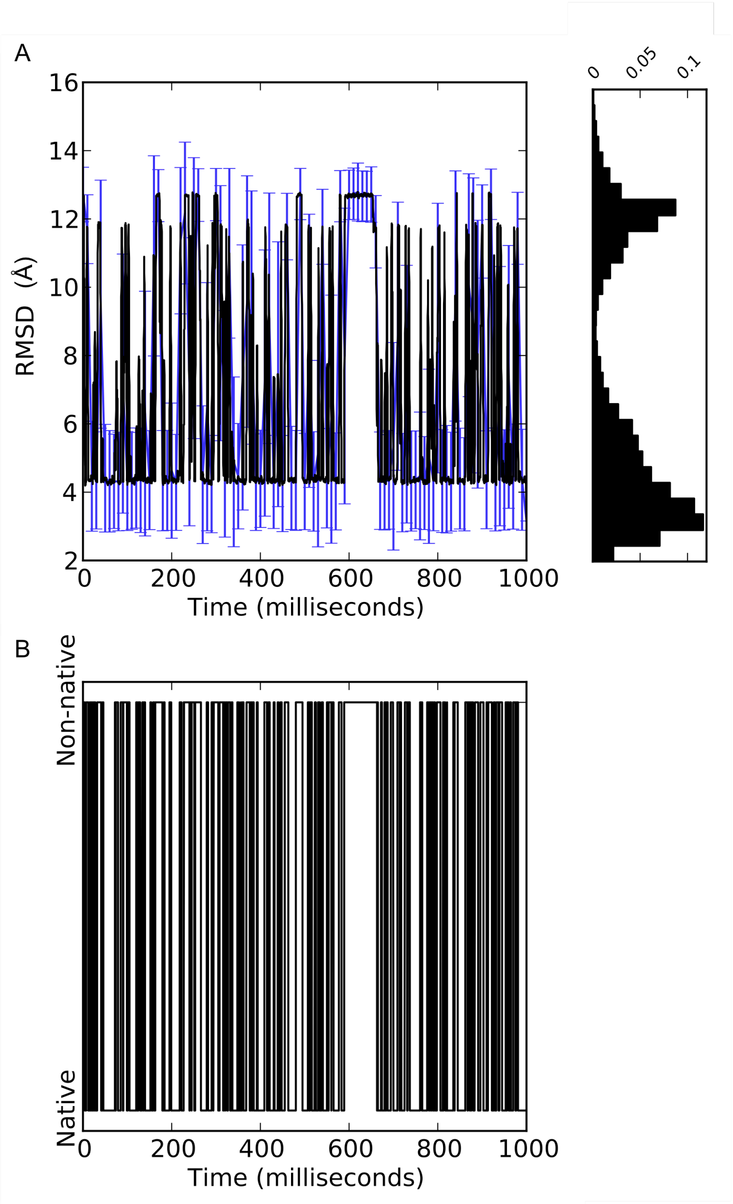

Figure S1. A 1 second simulation generated with our microstate MSM demonstrates reversible folding on 10 millisecond timescales. (A) RMSD versus time with a histogram (probability distribution) of RMSDs on the right showing apparent two-state behavior despite the hub-like character we discuss in the main text. The thick black line and blue error-bars correspond to the mean and standard deviation of the RMSD over 2 millisecond windows. (B) Plot of the simulation jumping back and forth between non-native states and the native state, supporting our interpretation of part (A).

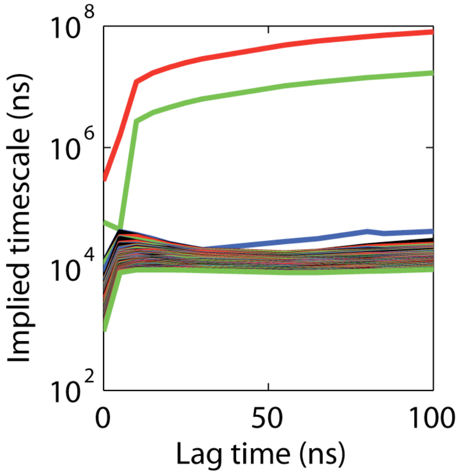

Figure S2. Implied timescales for the full 370 K dataset.

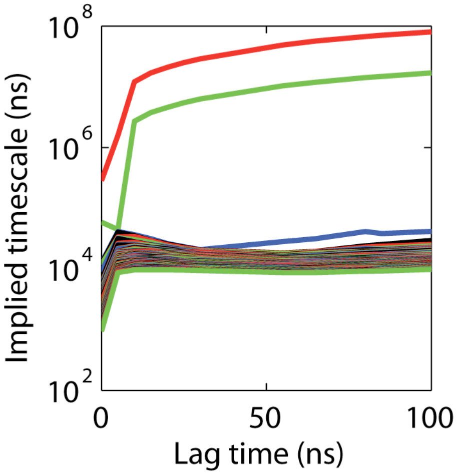

Figure S3. Implied timescales for the 300 K dataset.

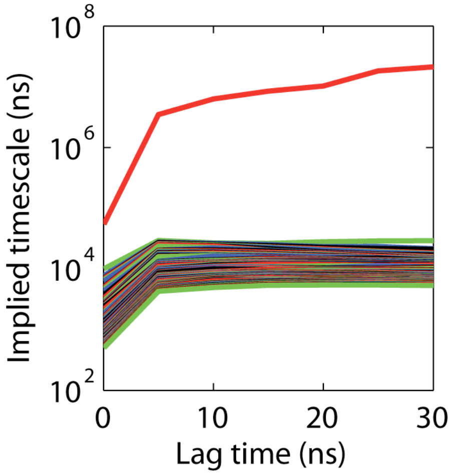

Figure S4. Implied timescales for ¾ of the 370 K dataset selected at random.

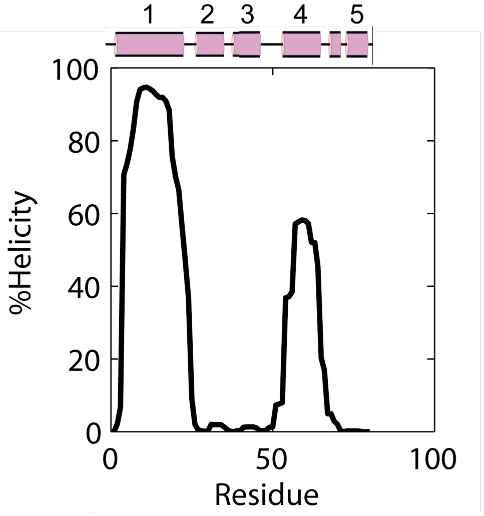

Figure S5. The helicity of each residue predicted from Agadir.[25] The purple, numbered bars show where the five helices are (the extra purple block between helices 4 and 5 is a turn).

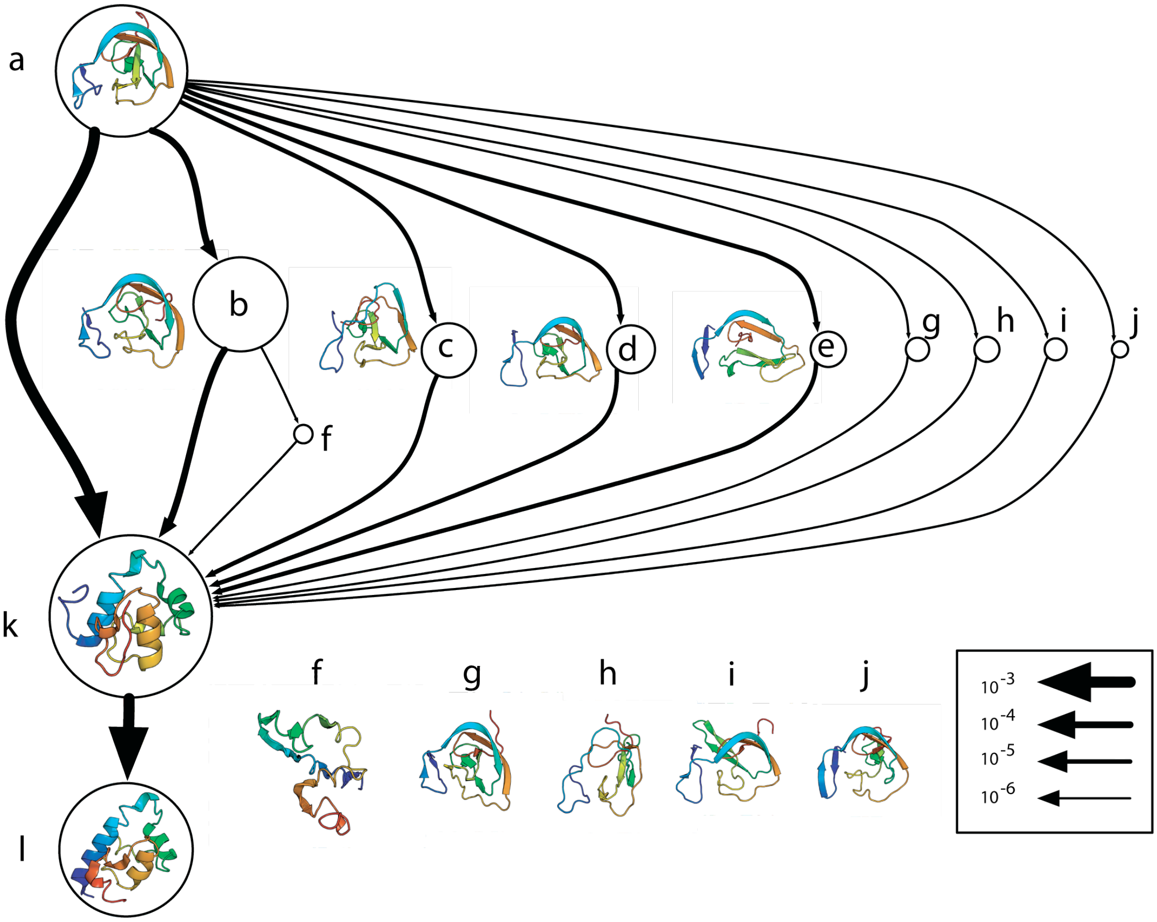

Figure S6. A coarse-grained view of the slowest transition with state sizes proportional to the free energy and arrow widths proportional to the flux (see key in figure).

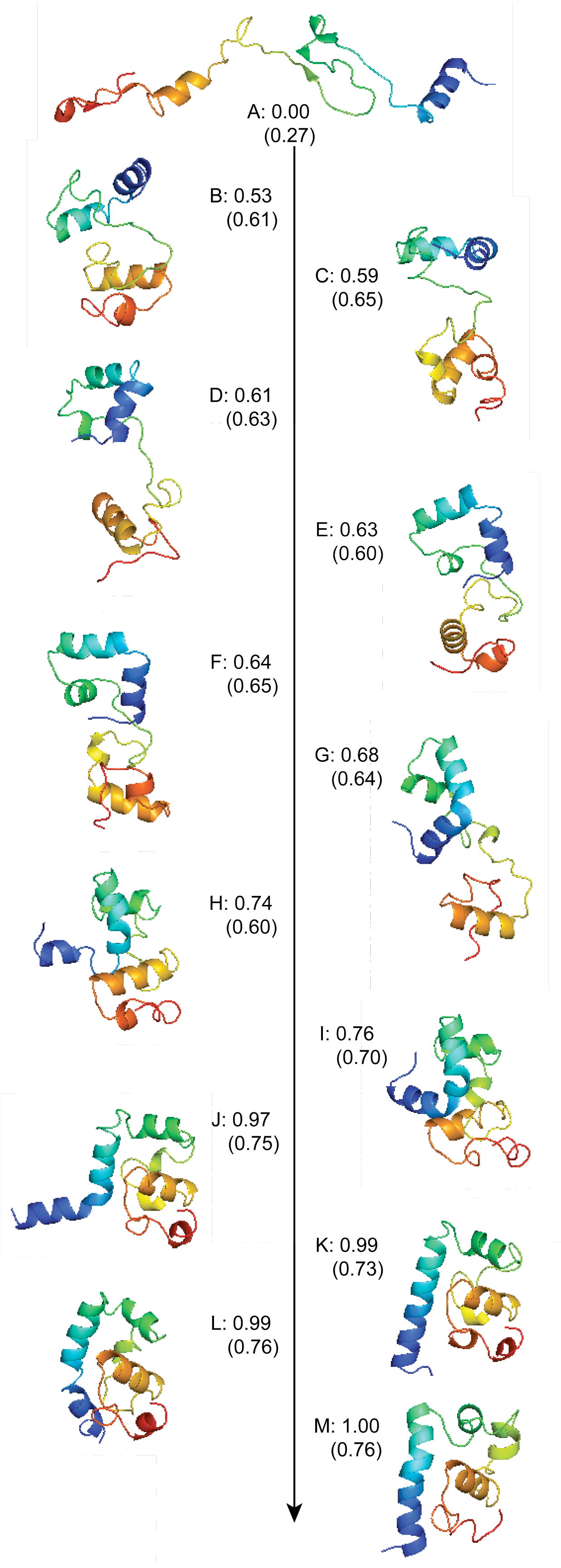
A: 0.00
(0.27)
B: 0.53
(0.61)
C: 0.59
(0.65)
D: 0.61
(0.63)
E: 0.63
(0.60)
F: 0.64
(0.65)
G: 0.68
(0.64)
H: 0.74
(0.60)
I: 0.76
(0.70)
J: 0.97
(0.75)
K: 0.99
(0.73)
L: 0.99
(0.76)
M: 1.00
(0.76)

Figure S7. Representative structures from a more detailed view of one pathway from the extended state in Figure 4E to the native state (Figure 4G) with p_{fold} values corresponding to the probability of reaching state M before state A. The proportion of native contacts is also given in parentheses as an estimate of how native-like the topology is. At the beginning, helices 1 and 4 are already partially formed, consistent with their high intrinsic helical propensity (Figure S5). Then the chain quickly collapses (B). From this point on, helix 5 has a strong propensity to pack against the hydrophobic residues of helix 4 (see the main text for more discussion of this point). Along with collapse, another segment of helix 1 forms (but there is a discontinuity between the two segments until later on) (B). Then helices 2 and 3 begin to form, with each forming ~1 helical turn (C-E). Helix 3 then remains in a primarily coil state while helix 2 forms (F-I). Once helix 2 is complete, the rest of helix 3 forms (J-M). Helix 4—which starts off tilted slightly downwards relative to helix 1—also flips a little upwards (J) and helix 1 straightens out more (though in the native macrostate there is still some flexibility around the middle of this helix) (K). The native topology (approximated by the proportion of native contacts formed) generally increases along the pathway but not completely monotonically, indicating it is not a perfect reaction coordinate for this path. The relatively high values even early in the pathway indicate the backbone quickly forms a native-like topology. Relative contact orders for each state are given in Table S2.

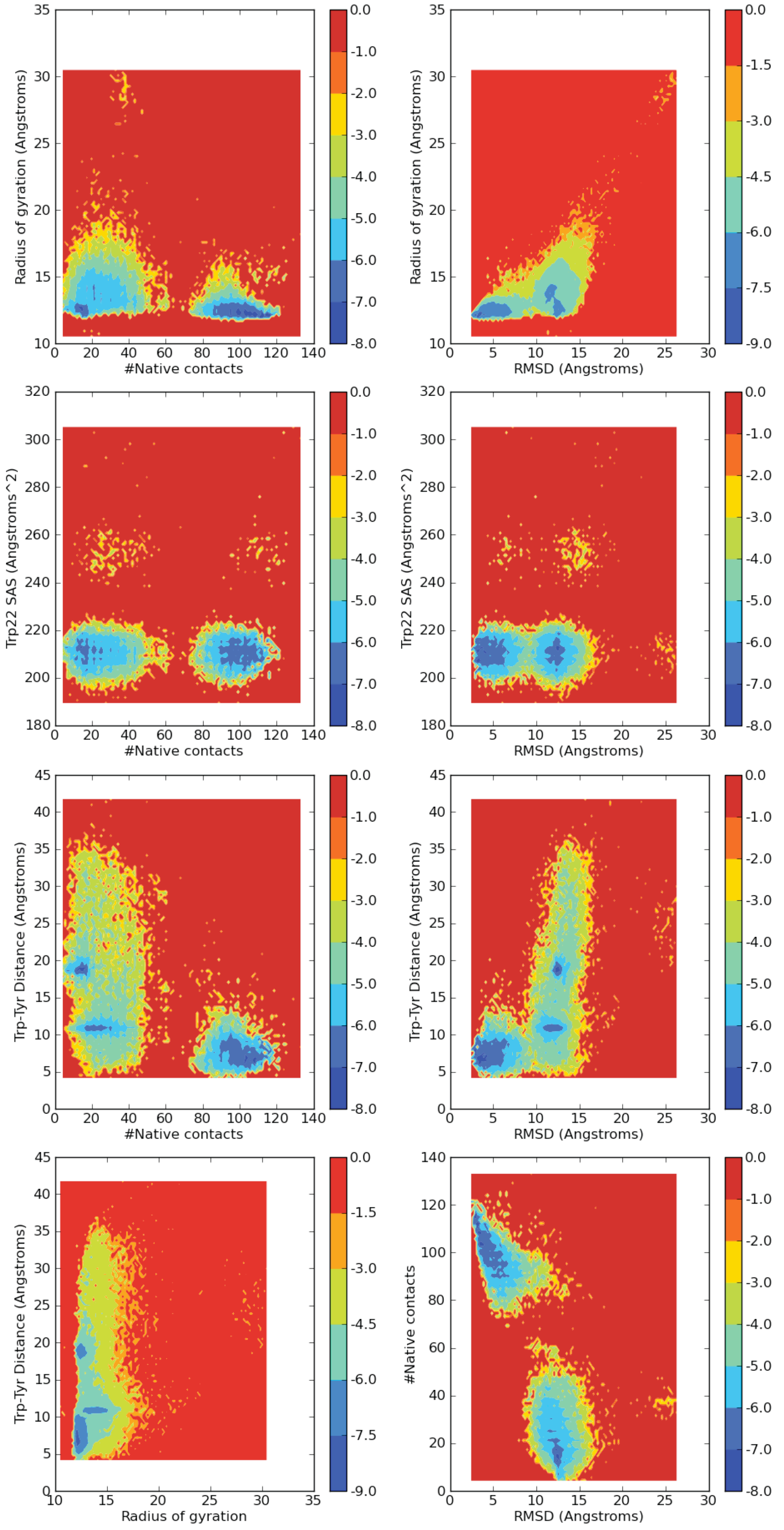
Radius of gyration (Angstroms)
35
30
25
20
15
10
#Native contacts
0
20
40
60
80
100
120
140
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
-7.0
-8.0
Radius of gyration (Angstroms)
RMSD (Angstroms)
0
5
10
15
20
25
30
0.0
-1.5
-3.0
-4.5
-6.0
-7.5
-9.0
Trp22 SAS (Angstroms^2)
320
300
280
260
240
220
200
180
#Native contacts
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
-7.0
-8.0
Trp22 SAS (Angstroms^2)
RMSD (Angstroms)
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
-7.0
-8.0
Trp-Tyr Distance (Angstroms)
45
40
35
30
25
20
15
10
5
0
#Native contacts
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
-7.0
-8.0
Trp-Tyr Distance (Angstroms)
RMSD (Angstroms)
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
-7.0
-8.0
Trp-Tyr Distance (Angstroms)
Radius of gyration
10
15
20
25
30
35
0.0
-1.5
-3.0
-4.5
-6.0
-7.5
-9.0
#Native contacts
140
120
100
80
60
40
20
0
RMSD (Angstroms)
0.0
-1.0
-2.0
-3.0
-4.0
-5.0
-6.0
-7.0
-8.0

Figure S8. Free energy projections of the microstate MSM onto typical order parameters like the radius of gyration (Rg), the C_α RMSD to the crystal structure, and the distance between the Trp22 and Tyr33 residues. Units are kcal/mol. Differences between the panels highlight the difficulty in interpreting such projections. In particular, some projections appear two-state while others look more three-state.

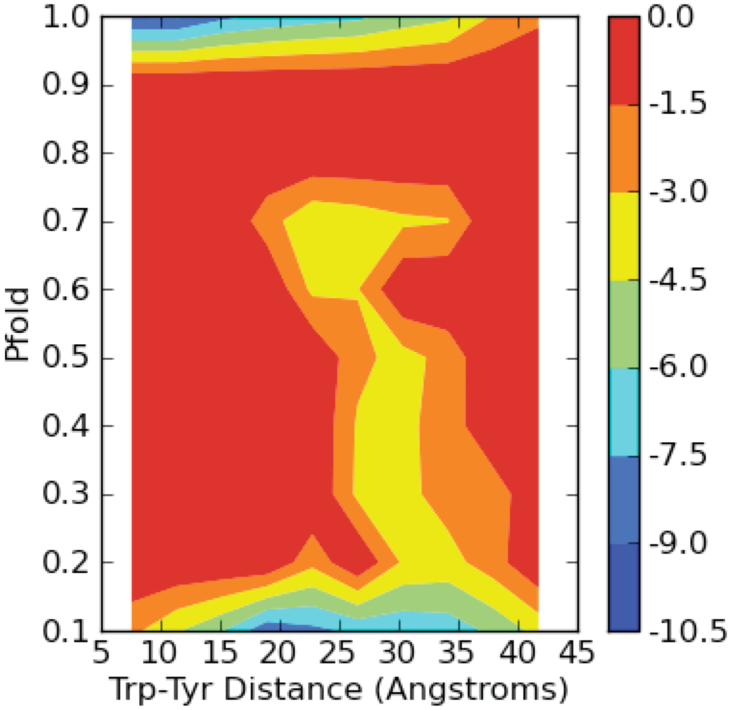

Figure S9. Free energy projection of the microstate MSM onto Pfold and the distance between the Trp22 and Tyr33 residues. Units are kcal/mol. Obtaining projections onto kinetic order parameters like Pfold is greatly simplified with MSMs. In this case Pfold refers to the probability of reaching the crystallographic state before reaching the compact β-sheet state (i.e. the slow transition from Figures 3 and 4). Unlike the projections in Figure S8, this one hints that D14A may not be well described by a simple two- or three-state model or that the Trp22-Tyr33 distance is not a good reaction coordinate, since there are a broad range of Pfold values possible for a given Trp-Tyr distance. Indeed, analysis of the MSM reveals that D14A is best described by a native hub.

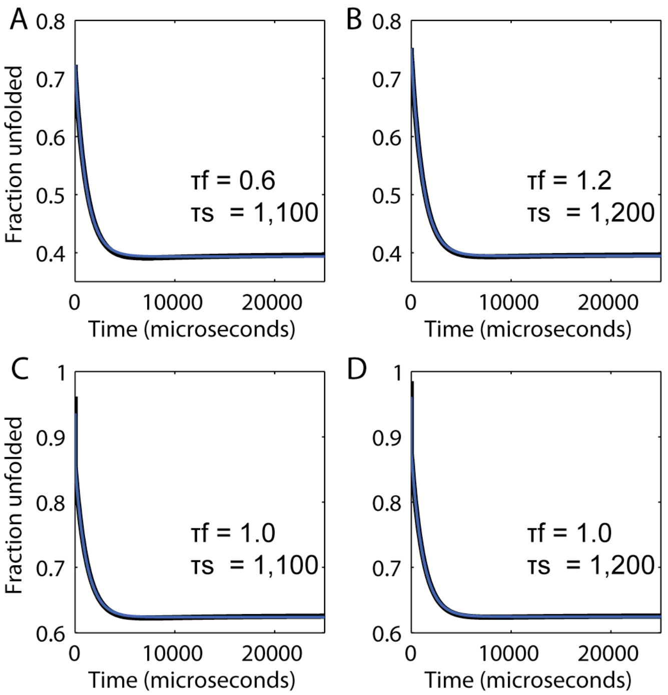

Figure S10. Relaxation of the fraction unfolded with different observables and initial population distributions. The thick black curves come from the MSM and the thin blue curves from biexponential fits to the MSM relaxation. The top row shows relaxation of the fraction unfolded measured by the Trp22-Tyr33 distance (A) starting from all states being equally populated and (B) starting from all non-native states being equally populated. The bottom row shows relaxation of the fraction unfolded measured by the C_α RMSD to the crystal structure (C) starting from all states being equally populated and (D) starting from all non-native states being equally populated. Fitting parameters are given in the figure (in units of microseconds). In this case, the fitting parameters are relatively independent of the observable and starting distribution.

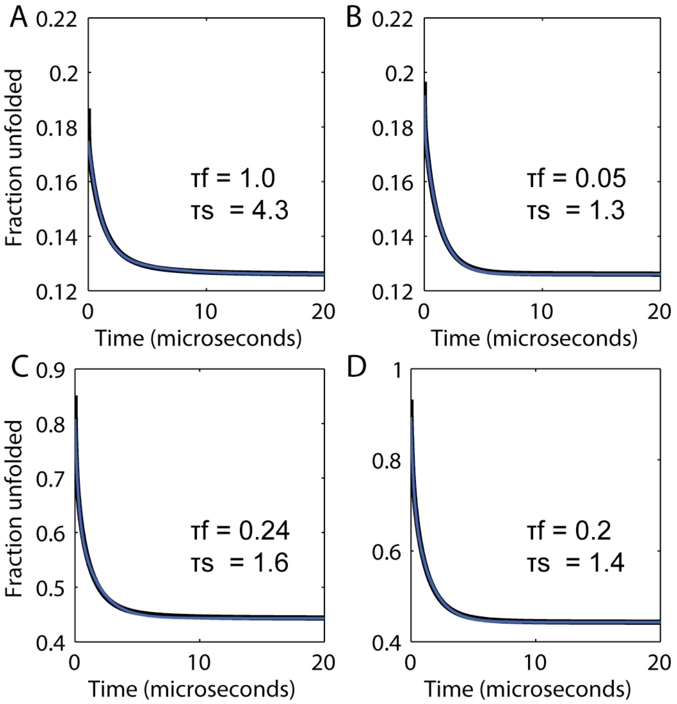

Figure S11. Relaxation of the fraction unfolded with different observables and initial population distributions from an MSM built without the trajectories started from β-sheet structures. The thick black curves come from the MSM and the thin blue curves from biexponential fits to the MSM relaxation. The top row shows relaxation of the fraction unfolded measured by the Trp22-Tyr33 distance (A) starting from all states being equally populated and (B) starting from all non-native states being equally populated. The bottom row shows relaxation of the fraction unfolded measured by the C_α RMSD to the crystal structure (C) starting from all states being equally populated and (D) starting from all non-native states being equally populated. Fitting parameters are given in the figure (in units of microseconds). In this case the fitting parameters are more dependent on the observable, consistent with the experimental observation of probe dependent kinetics.

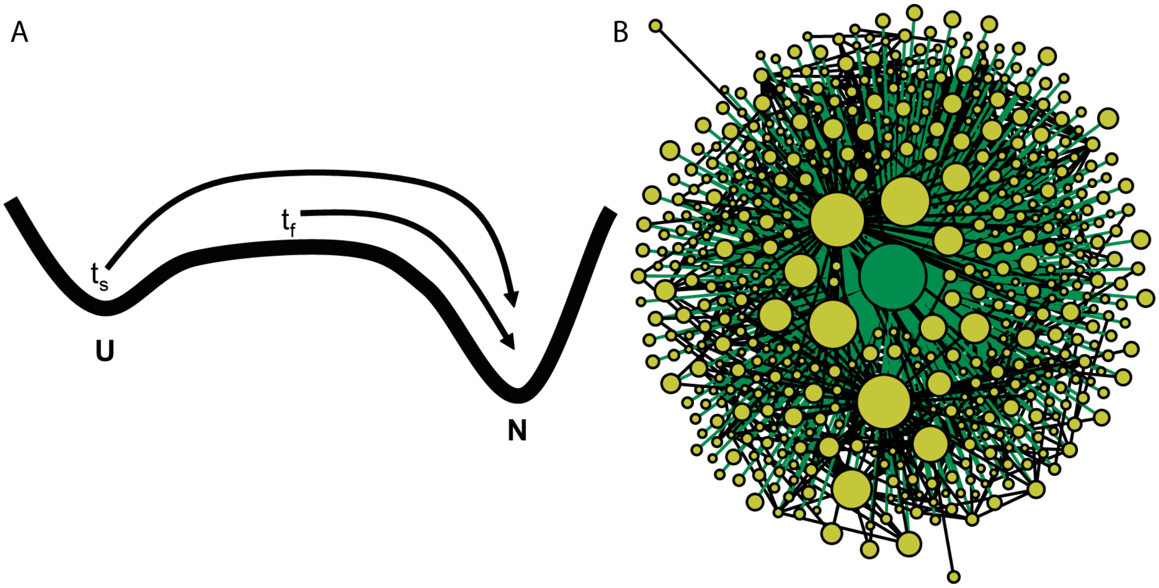

Figure S12. Two different models leading to biexponential relaxation. (A) Incipient downhill folding is like two-state folding but has a lower barrier. As a result, there is a reasonably sized population of proteins on top of the barrier that can slide downhill into the native state (N), leading to a fast phase (t_f). There is also a population of proteins in the unfolded basin (U) that has a slower transition rate toe the native state (t_s). (B) The native hub for D14A. This panel shows the 100 most populated macrostates with sizes proportional to their equilibrium populations. The native state and connections to it are colored green, highlighting the large number of connections to the native state. While the large number of native state connections does not prove that it is a kinetic hub, it is a hint that is confirmed by comparing the distribution of MFPTs to the native state and between non-native states.

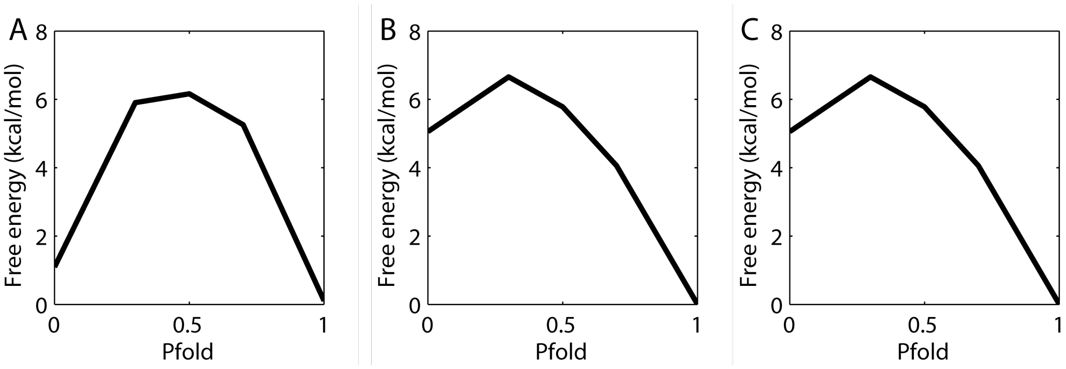

Figure S13. Projection of the free energy onto p_{fold} (A) from the compact β-sheet state in Figure 4A to the native state in Figure 4H, (B) from the extended state in Figure 4E to the crystallographic state in Figure 4H, and (C) from the extended state in Figure 4E to the model native state in Figure 4G. None are purely downhill, though some may be consistent with incipient downhill folding (i.e. have sufficiently low barriers that there is a reasonable population at the barrier top that can fold in a downhill manner in addition to activated folding across the barrier). In particular, the two projections with the extended structure from Figure 4E as a starting point have much lower barriers (i.e. are more consistent with incipient downhill folding) than when the compact β-sheet structure from Figure 4A is used as a starting point. We also note that the presence of parallel pathways and a kinetic hub may mean that no single order parameter can serve as a good reaction coordinate. For example, the compact β-sheet structures often have p_{fold} values near one when using the extended and crystal structures as starting and end points respectively (panel B). This happens because β-sheet structures can fold through other pathways. Even though going through this particular extended structure is one of the most probable pathways, the sum of the probabilities through other pathways is still higher than the probability of this particular path.

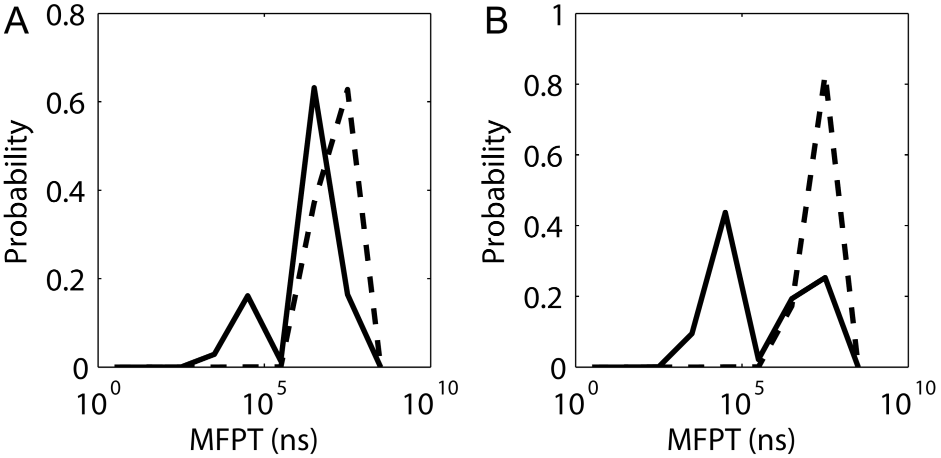

Figure S14. Distributions of mean first passage times (MFPTs) between sets of microstates (A) without weighting the distribution and (B) weighting each MFPT by the equilibrium probability of the starting state. The solid line is the distribution of MFPTs from non-native to native microstates and the dashed line is the distribution of MFPTs between non-native states. The average MFPT from non-native states to native ones is about 10 times faster than that between non-native states in (A) and the difference is even greater in (B). This difference is smaller than in previously studied systems,[26] but likely because of the higher temperature used in this work. Native microstates were defined as those in the most populated macrostate. All other microstates were considered non-native.

State	RCO
A	0.17
B	0.11
C	0.11
D	0.07
E	0.05
F	0.06
G	0.07
H	0.08

Table S1. Relative contact orders (RCOs) of the states from Figure 4 show that the contacts formed in the β-sheet states are much more non-local than those formed once the chain extends and then begins to collapse into native-like states.

State	RCO
A	0.05
B	0.05
C	0.05
D	0.05
E	0.07
F	0.07
G	0.05
H	0.05
I	0.05
J	0.05
K	0.05
L	0.07
M	0.07

Table S2. Relative contact orders (RCOs) of the states from Figure S7 show that the contacts formed throughout folding from an extended state are relatively local.

References

(1) Larios, E.; Pitera, J. W.; Swope, W.; Gruebele, M. *Chem Phys* **2006**, *323*, 45-53.

(2) Lindahl, E., B. Hess, and D. van der Spoel. *J. Mol. Modeling.* **2001**, *7*, 306-317.

(3) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J Comp Chem* **2000**, *21*, 1049-1074.

(4) Ensign, D. L.; Kasson, P. M.; Pande, V. S. *J Mol Biol* **2007**, *374*, 806-816.

(5) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comp. Phys.* **1977**, *23*, 327-341.

(6) Parrinello, M.; Rahman, A. *Journal of Applied Physics* **1981**, *52*, 7182-7190.

(7) Nose, S.; Klein, M. L. *Molecular Physics* **1983**, *50*, 1055-1076.

(8) Nose, S. *Molecular Physics* **1984**, *52*, 255-268.

(9) Yang, W. Y.; Gruebele, M. *Nature* **2003**, *423*, 193-197.

(10) Bowman, G. R.; Huang, X.; Pande, V. S. *Methods* **2009**, *49*, 197-201.

(11) Bowman, G. R.; Beauchamp, K. A.; Boxer, G.; Pande, V. S. *J Chem Phys* **2009**, *131*, 124101.

(12) Deuflhard, P.; Weber, M. *Lin. Alg. Appl.* **2005**, *398*, 161-184.

(13) Weber, M.; Kube, S. *Computational Life Sciences, Proceedings* **2005**, *3695*, 57-66.

(14) Roblitz, S., thesis, Freie Universitat Berlin, 2008.

(15) Chodera, J. D.; Singhal, N.; Pande, V. S.; Dill, K. A.; Swope, W. C. *J Chem Phys* **2007**, *126*, 155101.

(16) Noe, F.; Fischer, S. *Curr Opin Struct Biol* **2008**, *18*, 154-162.

(17) Swope, W. C.; Pitera, J. W.; Suits, F. *J Phys Chem B* **2004**, *108*, 6571-6581.

(18) Schütte, C.; Fischer, A.; Huisinga, W.; Deuflhard, P. *J Comput Phys* **1999**, *151*, 146–168.

(19) Berezhkovskii, A.; Hummer, G.; Szabo, A. *J Chem Phys* **2009**, *130*, 205102.

(20) Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc Natl Acad Sci U S A* **2009**, *106*, 19011-19016.

(21) Singhal, N.; Snow, C. D.; Pande, V. S. *J. Chem. Phys.* **2004**, *121*, 415-425.

(22) Pabo, C. O.; Lewis, M. *Nature* **1982**, *298*, 443-447.

(23) Clarke, N. D.; Beamer, L. J.; Goldberg, H. R.; Berkower, C.; Pabo, C. O. *Science* **1991**, *254*, 267-270.

(24) Plaxco, K. W.; Simons, K. T.; Baker, D. *J Mol Biol* **1998**, *277*, 985-994.

(25) Munoz, V.; Serrano, L. *Nat Struct Biol* **1994**, *1*, 399-409.

(26) Bowman, G. R.; Pande, V. S. *Proc Natl Acad Sci U S A* **2010**, *107*, 10890-10895.