ACCELERATED COMMUNICATIONS

How long does it take to equilibrate the unfolded state of a protein?

Ronald M. Levy,[1]* Wei Dai,[2] Nan-Jie Deng,[1] and Dmitrii E. Makarov[3]

[1]Department of Chemistry and Chemical Biology, Rutgers, the State University of New Jersey, Piscataway, New Jersey 08854
[2]Department of Physics and Astronomy, Rutgers, the State University of New Jersey, Piscataway, New Jersey 08854
[3]Department of Chemistry and Biochemistry and Institute for Computational Engineering and Sciences, University of Texas at Austin, Austin, Texas 78712

Received 22 July 2013; Revised 9 August 2013; Accepted 12 August 2013
DOI: 10.1002/pro.2335
Published online 21 August 2013 proteinscience.org

Abstract: How long does it take to equilibrate the unfolded state of a protein? The answer to this question has important implications for our understanding of why many small proteins fold with two state kinetics. When the equilibration within the unfolded state U is much faster than the folding, the folding kinetics will be two state even if there are many folding pathways with different barriers. Yet the mean first passage times (MFPTs) between different regions of the unfolded state can be much longer than the folding time. This seems to imply that the equilibration within U is much slower than the folding. In this communication we resolve this paradox. We present a formula for estimating the time to equilibrate the unfolded state of a protein. We also present a formula for the MFPT to any state within U, which is proportional to the average lifetime of that state divided by the state population. This relation is valid when the equilibration within U is very fast as compared with folding as it often is for small proteins. To illustrate the concepts, we apply the formulas to estimate the time to equilibrate the unfolded state of Trp-cage and MFPTs within the unfolded state based on a Markov State Model using an ultra-long 208 microsecond trajectory of the miniprotein to parameterize the model. The time to equilibrate the unfolded state of Trp-cage is ~100 ns while the typical MFPTs within U are tens of microseconds or longer.

Keywords: protein unfolded state; protein folding; mean first passage time; Markov state model

Introduction

How long does it take to equilibrate the unfolded state of a protein? The answer to this question has important implications for our understanding of why many small proteins fold with two state kinetics.[1–28] The protein folding funnel picture provides key insights.[3,5,6] When a protein folds along multiple pathways as suggested by the funnel picture, the folding kinetics will still be two-state regardless of differences in the intrinsic barriers along each pathway if the equilibration within the unfolded state ensemble is much faster than the time it takes to fold. Yet the mean first passage times (MFPTs) between different regions of the unfolded state ensemble are typically much longer than the folding time; this suggests that the time to equilibrate the unfolded state ensemble is much longer than the

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Institutes of Health; Grant number: GM30580; Grant sponsor: National Science Foundation; Grant number: CHE-1266380.

*Correspondence to: Levy Ronald, Department of Chemistry and Chemical Biology, Rutgers, the State University of New Jersey, Piscataway, NJ 08854. E-mail: ronlevy@lutece.rutgers.edu

time to fold.[29,30] So there is a paradox: the single exponential kinetics can be explained by very fast equilibration within the unfolded state U relative to folding, but the long MFPTs within U seem to imply that the equilibration of the unfolded state is slow relative to folding. In this communication we resolve this paradox. It arises when the average time for a single molecule trajectory to hit a specific location (the MFPT to state *i*) within U, is compared with the time for population fluctuations within the unfolded state to relax. This relaxation time provides a quantitative measure of the time to equilibrate the unfolded state. We will show that the MFPT to any state within the unfolded ensemble is approximately equal to the time to equilibrate the unfolded state divided by the population of the target state. The smaller the size of the target state, the longer the MFPT to that state, even though the equilibration of the unfolded state ensemble is very fast. For the Trp-cage example we use for discussion, MFPTs between different regions of the unfolded state ensemble are 10s to 100s of microseconds, while the time to equilibrate the unfolded state is of the order of 100 ns. These times are to be compared with the folding time for Trp-cage, which is 5.5 microseconds.

An estimate of the time required to equilibrate the protein unfolded state is also needed to understand the implications of the recently introduced kinetic hub model of protein folding.[29,31,32] In this model, the folded state F acts as a hub, so that most paths, which connect pairs of unfolded states U1 and U2 pass through F.[33,34] Hub like behavior appears to imply that the unfolded state partitions into subspaces, which largely fold along different pathways, but we have shown that this is not the case for Trp-cage.[28] Furthermore, when the time to equilibrate within the unfolded state ensemble is much faster than the folding time, the hub like behavior simply reflects the fact that the F state has sufficient population to have a high probability of being on most paths between typical points U1 and U2 within the unfolded state ensemble. It has recently become clear that hub like behavior is consistent with a smooth folding funnel.[28]

We use the integral of the time correlation function, which quantifies how the population fluctuations within the unfolded state relax to equilibrium as the measure of the time to equilibrate the unfolded state.[35] There are two contributions to the relaxation of population fluctuations within the unfolded state ensemble of a protein, or equivalently the equilibration of the unfolded state. The first corresponds to relaxation of fluctuations, which originate and propagate entirely within the U state and the second to relaxation within U, which arises from the equilibration between the unfolded and folded states. When the former relaxation process is much faster than the later, the protein folding is two-state. In this communication we mostly focus on the fast relaxation processes entirely within U. For our analysis we use a discrete master equation model of Trp-cage with 20 states parameterized on a 208 microseconds all atom molecular dynamics simulation of this mini-protein in water provided by the D.E. Shaw group.[23] The kinetics is characterized by the implied timescale spectrum of the transition matrix, which contains all the information about the relaxation times of the states within the discrete time Markov State Model (MSM). The Trp-cage implied timescale spectrum has a substantial gap between the longest implied timescale, which is associated with folding and the others, therefore the intra U state fluctuations can be separated from the folding and the mini-protein folds in a two-state manner with single exponential kinetics. That the remaining eigenmodes correspond to intra U-state relaxation can be verified by comparing the spectrum with the corresponding implied timescale spectrum obtained using reflecting boundary conditions at F, as we do in the following section.

Results and Discussion

We use a master equation to study the timescales over which the unfolded state equilibrates. The formal solution to the master equation is:

$$\vec{\mathbf{P}}(\mathrm{t}) = \mathrm{T}(\mathrm{t}) \cdot \vec{\mathbf{P}}(0) \tag{1}$$

where **P** is a vector of state probabilities and the transition matrix **T** (also called the propagator) contains all the information about the kinetics of the system (see Supporting Information). The propagator matrix element $T_{ij}(t)$ is the probability that the system is in state j at time t given that it was in state i at time zero. All observables of the system can be calculated in terms of functions of the $T_{ij}(t)$. The $T_{ij}(t)$ in turn can be expressed in terms of the eigenvalues and eigenvectors of **T**. Figure 1 shows the spectrum of implied timescales for the Trp-cage transition matrix constructed from the Shaw trajectory and for a modified transition matrix with a reflecting boundary added at F. Imposing the reflecting boundary condition here provides a model for the dynamics of the unfolded state alone. It can be seen that the spectrum is very similar except that the largest nonzero eigenvalue is missing from the spectrum with reflecting boundary condition at F; this eigenmode corresponds to the relaxation between the unfolded (U) state ensemble and the folded (F) state. The large gap between the largest implied timescale and the others means that the folding is two-state and the implied timescale ($\sim$1.2 μs) is the inverse of the sum of the folding plus unfolding rates.

In Figure 2 we show a typical propagator matrix element $T_{ij}(t)$ from state i to state j, both within U,

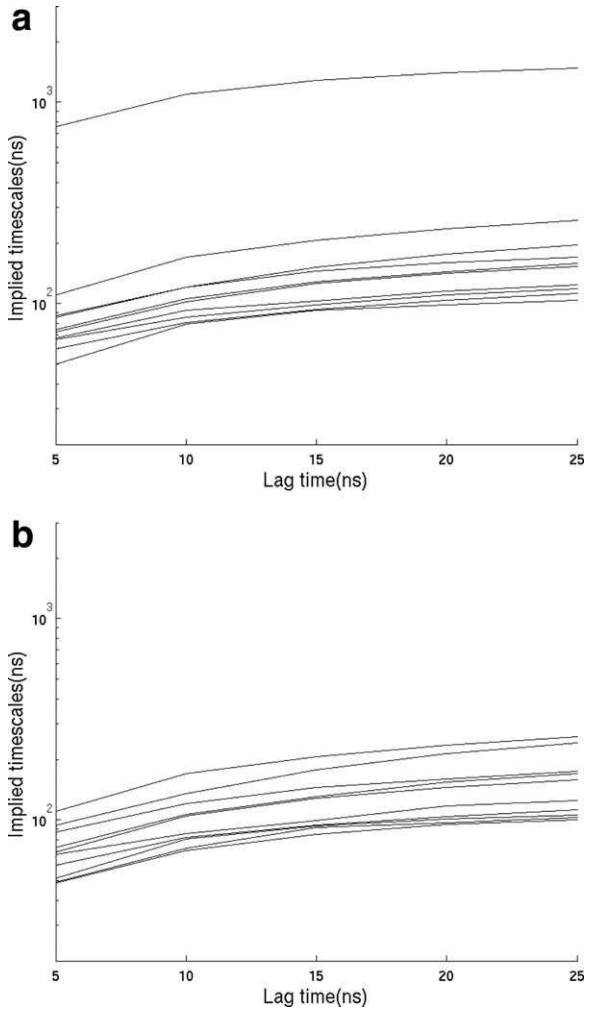

Figure 1. (a) The implied timescales corresponding to the 10 slowest decaying eigenmodes using transition matrices $T(\tau)$, with different boundary conditions, (a) unmodified equilibrium and (b) reflecting at F and I states. The optimal lag time 10 ns is chosen for further analysis based on the tradeoff between the network being Markovian and the resolution being sufficient for studying folding mechanism.

calculated three ways; using absorbing, unmodified equilibrium, and reflecting boundary conditions at *F*. The time dependence of $T_{ij}(t)$ describes the relaxation process following an initial point perturbation at state *i*. On a timescale of a few hundred nanoseconds they look very similar. Each rises rapidly to a plateau value which "overshoots" the equilibrium population of state *j* by a small amount. When added up over all the states in U, the excess corresponds to the equilibrium population of F that folds from U to F on the slower timescale of ~5 μs. After a few hundred nanoseconds, the $T_{ij}(t)$ matrix elements shown in Figure 2 have the following longer time behavior. Under reflecting boundary conditions $T_{ij}(t)$ is approximately constant, the unmodified transition matrix $T_{ij}(t)$ relaxes to the equilibrium population with a relaxation time ~1.2 μs, while under absorbing boundary condition the matrix elements relax to zero with a relaxation time ~5 μs. The results shown in Figure 2 are suggestive as to the timescales for equilibrating the unfolded state, but the full relaxation involves all the elements $T_{ij}(t)$ of the propagator. We consider the full expression for the relaxation now.

The way to estimate the time it takes to equilibrate a system from equilibrium statistical mechanics is to calculate an integral of the appropriate time correlation function.[35] The correlation function of interest here corresponds to the decay of the population fluctuations in the unfolded state. After some manipulation (see Supporting Information), this correlation function can be expressed as:

$$\widehat{C}_{tot} = \sum\nolimits_i \mathrm{P_{eq}(i)} \frac{<\Delta \mathrm{P_i}(0) \cdot \Delta \mathrm{P_i}(\mathrm{t}) >}{<\Delta \mathrm{P_i}(0)^2 >} \tag{2a}$$

$$\widehat{C}_{tot} = \sum\nolimits_i \frac{(\mathrm{T_{ii}(t)} - \mathrm{P_{eq}(i)})\mathrm{P_{eq}(i)}}{1 - \mathrm{P_{eq}(i)}} \tag{2b}$$

$$\widehat{C}_{tot} = \sum\nolimits_{n=2}^{N} \left[\sum\nolimits_i \frac{\mathrm{P_{eq}(i)} \psi_N^R(i) \psi_N^L(i)}{1 - \mathrm{P_{eq}(i)}} \lambda_n \right] \tag{2c}$$

where $\psi_n{}^R(i)$ and $\psi_n{}^L(i)$ are the ith element of the nth right and left eigenvectors of the **T** matrix. λ_n is the nth eigenvalue of the **T** matrix. $\Delta P_i(t) = P_i(t) - P_{eq}(i)$. $P_{eq}(i)$ is the equilibrium population of state *i*. $P_i(t)$ is an indicator function, which is 1 when the trajectory is on state *i* and 0 otherwise at time *t*.

In Figure 3 we show the unfolded state population fluctuation correlation function. When the motions are restricted to the unfolded state, the time to equilibrate the unfolded state is estimated from the time integral of $\widehat{C}_{tot}(t)$ to be ~100 ns; when the additional relaxation of U due to the much slower equilibration between U and F is also considered, the time to equilibrate the unfolded state is increased to ~540 ns. The separation of timescales between the equilibration within U and the folding is implicit in the folding funnel model of protein folding.[3,5] While folding on a flat "golf-course" landscape,[11] which lacks the energy bias can also produce a separation of timescales, the very fast equilibration (~100 ns) within the unfolded state is a feature of the funneled landscape.

Our estimate of the time to equilibrate the protein unfolded state based on the decay of fluctuations of the U state population (eq. 2b) is independent of the kind of experiment chosen to monitor the system. Any particular experiment will measure the time evolution of the population fluctuations reweighted by how sensitive that particular probe is to the different modes by which the population fluctuations relax. If for example, the experiment is sensitive to the fluctuations of some property *f*, then the experimental relaxation time measured for that probe of the unfolded state dynamics would be:

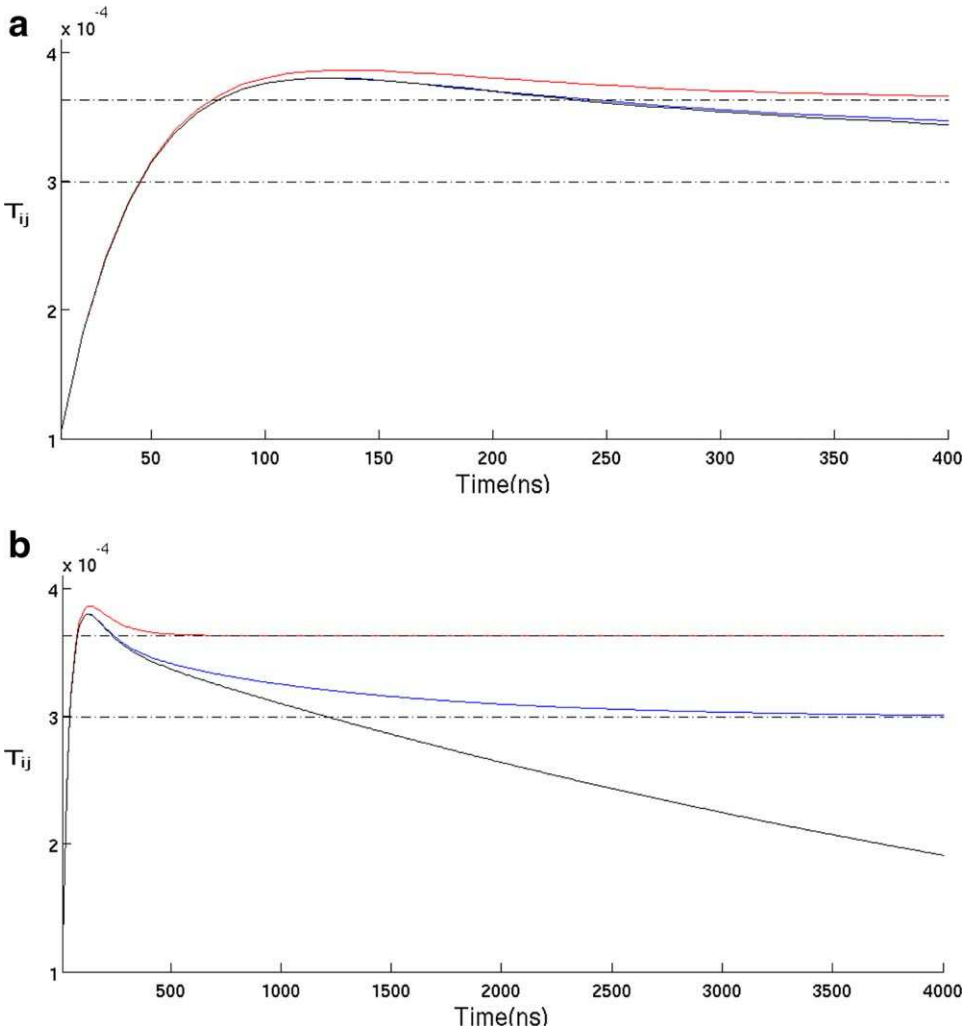

Figure 2. (a) A typical propagator matrix element *Tij*(*t*) from state *i* to state *j*, both within U, calculated three ways; using absorbing at F (black), unmodified equilibrium (blue) and reflecting boundary conditions at F (red). The upper and lower dash and point lines in each subplot are correspondingly the equilibrium population of state *j* under reflecting and unmodified equilibrium boundary conditions. All the propagator elements are calculated from spectral decomposition using all the 20 eigenmodes at lag time of 10 ns. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$\widehat{C}_f = \frac{\sum_{i,j} f(i) f(j) P_{eq}(j) (T_{ij}(t) - P_{eq}(i))}{\sum_{i,j} f(i) f(j) P_{eq}(j) (T_{ij}(0) - P_{eq}(i))} \quad (3)$$

where $f(i)$, $f(j)$ are the values of the experimental observables in state i and j.

A common choice of the experimental observable f is the FRET efficiency, which is a nonlinear function of the distance between two particular residues within the protein. The relaxation time thus determined depends on the choice of those residues.[25]

We turn now to an analysis of the MFPTs between different states within the unfolded state ensemble. From MSMs, the MFPTs between unfolded states have been reported to be tens of microseconds or longer.[29,31,32] For the Trp-cage model we studied it extends to ~200 microseconds. The MFPT to an unfolded state i can be obtained from the formula:

$$MFPT_i = \left\langle \int_0^\infty t . \frac{dt_{ji}^{abs \to i}}{dt} dt \right\rangle = \sum_j \frac{P_{eq}(j)}{1 - P_{eq}(i)} \int_0^\infty t . \frac{dT_{ji}^{abs \to i}}{dt} dt \quad (4a)$$

$$MFPT_i = \sum_j \left\{ \frac{P_{eq}(j)}{1 - P_{eq}(i)} \cdot \sum_{n=2}^{N} \psi_N^R(j) \psi_N^L(i) (-\mu_n) \right\} \quad (4b)$$

where $\psi_{jN}^{R}(i)$ and $\psi_{jN}^{L}(i)$ are the ith element of the nth right and left eigenvectors of the transition matrix with an absorbing boundary at i $\mathbf{T}^{abs \to i}$. μ_n is its the nth implied timescale (see Supporting Information).

The average shown in eq. 4a is taken over all the other states j in U and includes a sum over all the eigenmodes n. In Figure 4(a) we show the implied timescale spectrum of the transition matrix with absorbing boundary at a typical unfolded state i. The

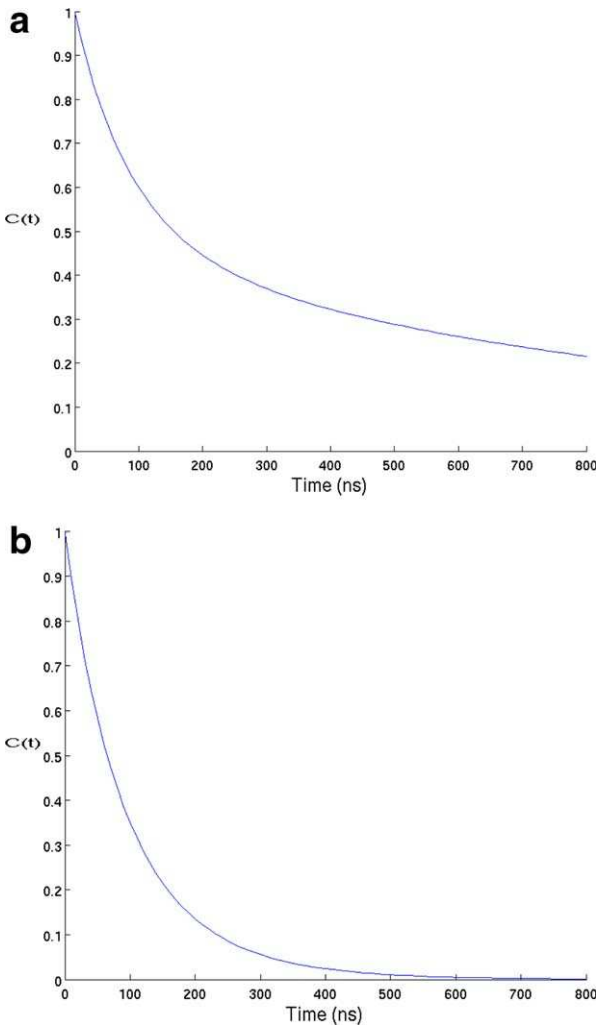

Figure 3. (a) The population fluctuation relaxation functions [Eq. (2)] of the 20-node network at a lag time of 10 ns, using two different boundary conditions, (a) unmodified equilibrium and (b) reflecting on F. The integrals of the functions are the relaxation times, which are 543 and 100 ns correspondingly. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

large gap between the largest implied timescale and the rest is the signature of the exponential distribution of first passage times to unfolded state i. The longest implied timescale is of the order of ~100 microseconds. Because the unfolded state ensemble relaxes on a timescale a hundred to a thousand times faster than the time it takes on average to reach state i, the MFPT to state i does not depend on the starting point within U. The kinetics involving the transitions between any specific state i and all the other states taken collectively is then effectively two state and the MFPT to state i can be written as:

$$\mathrm{MFPT}_i \approx \sum_j \frac{P_{\mathrm{eq}}(j)}{1-\mathrm{P_{eq}(i)}} \cdot \psi_2^R(j)\psi_2^L(i)\mu_2 \qquad (5)$$

The MFPT to the unfolded state i chosen for the example shown in Figure 4(a) is found to be 106 microseconds.

To understand why the MFPTs to states within U are so long, we consider the relationship between the average lifetime of a state i within U and the average lifetime of the collective state consisting of the remainder of U excluding state i:

$$t_{\mathrm{U-i}} = t_i\left(\frac{1}{\mathrm{P_{eq}(i)}} - 1\right) \qquad (6)$$

where t_i is the average lifetime of state i and $t_{\mathrm{U}\text{-}i}$ is the average lifetime of the collective state U-i consisting of the remainder of U excluding state i. Here we define the lifetime distribution of a state as the distribution of times recorded upon entering a state when the clock starts and then leaving it when the clock stops, during a single very long trajectory when the state is visited many times [see Supporting Information for the derivation of Eq. 6].

In Figure 4(b) we plot the MFPT to state i [Eq. (4b)] against the average lifetime of the collective state, $t_{\mathrm{U}\text{-}i}$ [Eq. 6] for each of the unfolded states in the 20-state model. It can be seen that these times are almost equal. This is true when the time to equilibrate within the unfolded state (U-i) is much shorter than the average lifetime of (U-i). Under these circumstances, the MFPT to any unfolded state i is proportional to the average lifetime of the state t_i divided by the population, and there is an equality involving Eqs 4, 5 and 6. Because the average lifetimes of the unfolded states decay on the same timescale as the decay of the population fluctuations, we find that the MFPT to any state within U is approximately equal to the time to equilibrate U divided by the population of the target state. Importantly, the MFPTs depend on the resolution of the model for the unfolded state, the more fine grained the model, the longer the MFPTs to an individual state. On the other hand, the time to equilibrate the unfolded state is a characteristic of the macrostate, which depends only weakly on the resolution. For the 20-state model of Trp-cage studied here, the longest MFPT (~200 μs) is to the state with the smallest population 0.003, while the average lifetime of that state is 48 ns, comparable to the time to equilibrate the unfolded state.

In this communication we have resolved a paradox about kinetics within the unfolded state of proteins, which leads to a better understanding of why most small proteins fold with two-state kinetics. When the equilibration of the unfolded state ensemble is very fast as it is for most small proteins, the protein will fold with single exponential kinetics. While it seems paradoxical that the time to equilibrate the unfolded state can be orders of magnitude shorter than MFPTs within U, we have shown there is no inconsistency. Using a time-correlation function approach, we have presented a general formula for the timescale of population relaxation within U [Eq. (2c)]. Applying this formula to the folding of the two-state mini-protein Trp-cage, we found that the

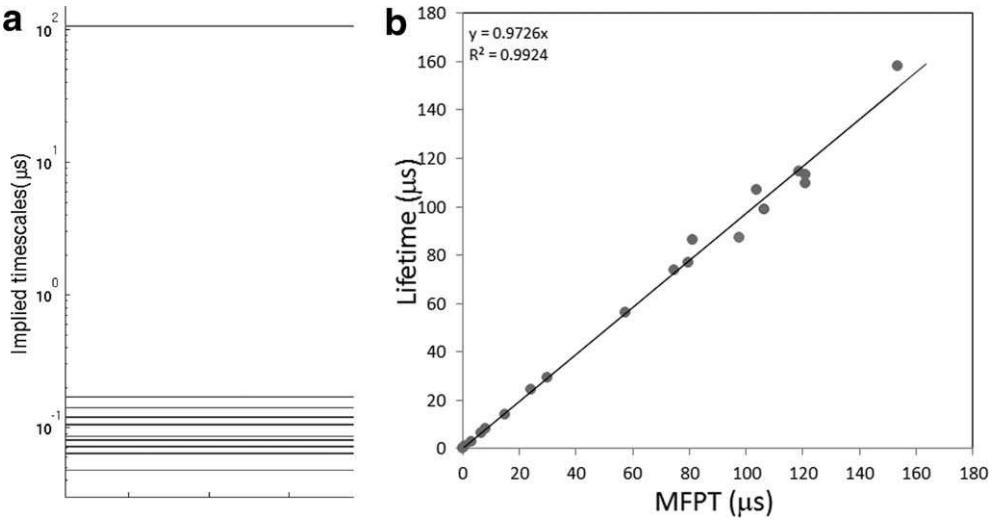

Figure 4. (a) The implied timescale spectrum to the state, which is highlighted as red in Figure 4(b). (b) The average lifetime of the collective state (U-*i*), which excludes the state *i* versus the MFPT to state *i* using Eq. (4b). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

folding follows a two-step process: starting from an arbitrary nonequilibrium conformational distribution within the unfolded region the protein population will quickly relax to a pre-equilibrium within the unfolded state on timescales (~100 ns for Trp-cage) much faster than folding. From this time forward, while the relative populations of all the unfolded microstates remain constant, the "excess" population within U, which will populate the folded state at equilibrium, folds with single exponential kinetics (rate ~1/5.5 μs). It should be noted that as we reported in a recent article, an individual Trp-cage folding trajectory only visits a fraction (e.g., ~25%) of the unfolded state space.[28] The key to reconciling this with the rapid equilibration in the U-state is to realize that while any one trajectory explores only a small part of U before folding, an ensemble of such trajectories starting from the same initial condition within U will explore all of the U states with a probability that is close to the equilibrium population of that state before folding.[11,28,30] The methodology developed in this study is also well suited for studying the kinetics of larger and more complex proteins where the timescales to equilibrate within U and to fold may overlap and the folding is no longer two state.

Materials and Methods

A MD trajectory of Trp-cage, which contains 1 million snapshots and saved at every 200 ps, was obtained from D.E. Shaw Research.[23] The simulation length is 208 μs using a modified CHARMM22 all-atom force field in the TIP3P explicit solvent. A 25000-node fine-grained network and a 20-node coarse-grained network were generated from the trajectory (see Supporting Information for detailed descriptions of how the fine-grained network was generated).

Acknowledgments

Some of the calculations were performed using the XSEDE allocation TG-MCB100145. The authors thank Dr. Attila Szabo for very helpful discussions. ND would like to thank Dr. Kyle Beauchamp from Dr. Vijay Pande group for help with the MSMBuilder2.[36]

References

1. Creighton TE (1988) Toward a better understanding of protein folding pathways. Proc Natl Acad Sci U S A 85: 5082–5086.
2. Jackson SE, Fersht AR (1991) Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. Biochemistry (Mosc.) 30:10428–10435.
3. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins 21:167–195.
4. Eaton WA, Thompson PA, Chan C-K, Hage SJ, Hofrichter J (1996) Fast events in protein folding. Structure 4:1133–1139.
5. Onuchic JN, Luthey-Schulten Z, Wolynes PG (1997) Theory of protein folding: the energy landscape perspective. Annu Rev Phys Chem 48:545–600.
6. Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. Nat Struct Biol 4:10–19.
7. Zwanzig R (1997) Two-state models of protein folding kinetics. Proc Natl Acad Aci U S A 94:148–150.
8. Perl D, Welker C, Schindler T, Schröder K, Marahiel MA, Jaenicke R, Schmid FX (1998) Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. Nat Struct Biol 5:229–235.
9. Jackson SE (1998) How do small single-domain proteins fold? Fold Des 3:R81–R91.
10. Cieplak M, Henkel M, Karbowski J, Banavar J (1998) Master equation approach to protein folding and kinetic traps. Phys Rev Lett 80:3654–3657.

11. Bicout DJ, Szabo A (2000) Entropic barriers, transition states, funnels, and exponential protein folding kinetics: a simple model. Protein Sci 9:452–465.
12. Dinner AR, Šali A, Smith LJ, Dobson CM, Karplus M (2000) Understanding protein folding via free-energy surfaces from theory and experiment. Trends Biochem Sci 25:331–339.
13. Mirny L, Shakhnovich E (2001) Protein folding theory: From lattice to all-atom models. Annu Rev Biophys Biomol Struct 30:361–396.
14. Makarov DE (2002) How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. Proc Natl Acad Sci U S A 99:3535–3539.
15. Yang WY, Gruebele M (2003) Folding at the speed limit. Nature 423:193–197.
16. Kaya H, Chan HS (2003) Simple two-state protein folding kinetics requires near-levinthal thermodynamic cooperativity. Proteins 52:510–523.
17. Weikl TR (2004) Cooperativity in two-state protein folding kinetics. Protein Sci 13:822–829.
18. Rhoades E, Cohen M, Schuler B, Haran G (2004) Two-state folding observed in individual protein molecules. J Am Chem Soc 126:14686–14687.
19. Ellison PA, Cavagnero S (2006) Role of unfolded state heterogeneity and en-route ruggedness in protein folding kinetics. Protein Sci 15:564–582.
20. Barrick D (2009) What have we learned from the studies of two-state folders, and what are the unanswered questions about two-state protein folding? Phys Biol 6:015001.
21. Zheng W, Andrec M, Gallicchio E, Levy RM (2009) Recovering kinetics from a simplified protein folding model using replica exchange simulations: A kinetic network and effective stochastic dynamics. J Phys Chem B 113:11702–11709.
22. Best RB, Hummer G (2009) Coordinate-dependent diffusion in protein folding. Proc Natl Acad Sci U S A 107:1088–1093.
23. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. Science 334:517–520.
24. Karplus M (2011) Behind the folding funnel diagram. Nat Chem Biol 7:401–404.
25. Soranno A, Buchli B, Nettels D, Cheng RR, Muller-Spath S, Pfeil SH, Hoffmann A, Lipman EA, Makarov DE, Schuler B (2012) Quantifying internal friction in unfolded and intrinsically disordered proteins with single-molecule spectroscopy. Proc Natl Acad Sci U S A 109:17800–17806.
26. Zhang Z, Chan HS (2012) Transition paths, diffusive processes, and preequilibria of protein folding. Proc Natl Acad Sci U S A 109:20919–20924.
27. De Sancho D, Mittal J, Best RB (2013) Folding kinetics and unfolded state dynamics of the GB1 hairpin from molecular simulation. J Chem Theory Comput 9:1743–1753.
28. Deng N, Dai W, Levy RM (2013) How kinetics within the unfolded state affects protein folding: An analysis based on Markov state models and an ultra-long MD trajectory. J Phys Chem B 130524112618003.
29. Bowman GR, Pande VS (2010) Protein folded states are kinetic hubs. Proc Natl Acad Sci U S A 107:10890–10895.
30. Lane TJ, Schwantes CR, Beauchamp KA, Pande VS. Probing the origins of two-state folding. Physicsbio-Ph Arxiv13050963.
31. Voelz VA, Bowman GR, Beauchamp K, Pande VS (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1−39). J Am Chem Soc 132:1526–1528.
32. Bowman GR, Voelz VA, Pande VS (2011) Taming the complexity of protein folding. Curr Opin Struct Biol 21: 4–11.
33. Dickson A, Brooks CL (2012) Quantifying hub-like behavior in protein folding networks. J Chem Theory Comput 8:3044–3052.
34. Dickson A, Brooks CL (2013) Native states of fast-folding proteins are kinetic traps. J Am Chem Soc 135: 4729–4734.
35. Chandler D (1987) Introduction to modern statistical mechanics. New York: Oxford University Press.
36. Beauchamp KA, Bowman, GR, Lane TJ, Maibaum L, Haque, IS, Pande VS (2011) MSMBuilder2: Modeling conformational dynamics on the picosecond to millisecond scale. J Chem Theory Comput 7:3412–3419.

How Long Does it Take to Equilibrate the Unfolded State of a Protein?

Ronald M. Levy, Wei Dai, Nan-jie Deng, Dmitrii E. Makarov

Supporting Information

A. Markov State Models

A Markov State Model of the kinetics of Trp-Cage was built using an ultra-long Molecular Dynamics trajectory to parameterize the model. A Markov State Model consists of a graph which represents clusters of Trp-Cage conformations that are the nodes of the graph, and a set of directed edges that connect the nodes.

In a Markov State Model, the time evolution of a vector $\vec{\mathbf{P}}$ representing the population of each node can be described by a master equation. In the continuous time case,

$$\frac{d\vec{\mathbf{P}}(t)}{dt} = \mathbf{K} \cdot \vec{\mathbf{P}}(t) \tag{S1}$$

where **K** is the rate matrix of which element represents the rate constant between two connected nodes. However, the rate matrix usually is not accessible in a real system. A Markov State Model can also be represented by a transition probability matrix,

$$\vec{\mathbf{P}}(\Delta t) = e^{\mathbf{K}\cdot\Delta t} \cdot \vec{\mathbf{P}}(0) = \mathbf{T}(\Delta t) \cdot \vec{\mathbf{P}}(0) \tag{S2}$$

where **T** is the column-stochastic transition probability matrix, and Δt is the lag time which is the resolution of the model. By repeatedly multiplying the transition matrix to the population vector, we can obtain the time evolution of the population.

The eigenvalues λ_k of the transition matrix **T** imply a timescale, which is called the implied timescale,

$$\tau_k = -\frac{\Delta t}{\ln(\lambda_k)} \tag{S3}$$

The implied timescale should approach a constant when the system is Markovian. The longest implied timescale corresponds to the slowest relaxation process which is the equilibration between the folded and unfolded states.

B. The construction of the coarse-grained network

An MD simulation of Trp-cage was performed by Shaw and coworkers on the Anton computer for 208μs using a modified CHARMM22 all-atom force field in the TIP3P explicit solvent. The MD trajectory contains 1.044×10^6 snapshots saved at every 200 ps. We use MSMBuilder2 to cluster all the snapshots into 31284 microstates with a hybrid of k-centers and k-medoids algorithm. A

20-node macrostate model is constructed from a PCCA+ lumping of the microstate model. The transition matrix $T_{ij}(\Delta t)$ is estimated by projecting the MD trajectory on the nodes and counting the number of transitions from node i to node j in a lag time of Δt. Based on the RMSD distribution of the nodes, at the 20-node level the model contains a native state with population of 17% and 19 unfolded states. Note that the definition of the macrostates is different from those in the study by Shaw et al., where the macrostates are defined by the fraction of native contacts Q. As a result, there are some quantitative differences between the kinetic properties calculated in this work and those found by Shaw et al.. This does not affect the interpretation of the main results in the main text.

To verify that the longest implied timescales of the 20-node Markov State Model of Trp-Cate are consistent with a higher resolution model, we established another fine-grained network (~25,000 nodes) which was generated by geometrically clustering the snapshots according to their mutual RMSD using a k-means clustering method. The implied timescales (Figure S1) and population relaxation time (Figure S2) of the fine-grained network have shown good consistency with the 20-node network.

C. The population fluctuation relaxation function

Let $\delta P_i(t)$ denote the instantaneous deviation in $P_i(t)$ of node i from its time-independent equilibrium average, $<P_i>$,

$$\delta P_i(t) = P_i(t) - < P_i(t) > \tag{S4}$$

Let s(t) be the location of the trajectory at time t. The indicator function is defined as follows,

$$P_i(t) = \begin{cases} 1, & s(t) = i \\ 0, & s(t) \neq i \end{cases} \tag{S5}$$

If we write population fluctuation, $\delta P_i(t) = P_i(t) - < P_i(t) >$, where $<P_i(t)>$ is the population of node i, $P_{eq}(i)$. The population fluctuation relaxation function of state i can be written in terms of $\delta P_i(t)$.

$$\begin{aligned} C_i(t) &= < \delta P_i(0) \cdot \delta P_i(t) > \\ &= (1 - P_{eq}(i))P_{eq}(i) \times (T_{ii}(t) - P_{eq}(i)) \\ &\quad + (0 - P_{eq}(i))(1 - P_{eq}(i)) \times \left(\sum_{j \neq i} T_{ji}(t) \cdot \frac{P_{eq}(j)}{1 - P_{eq}(i)} - P_{eq}(i)\right) \\ &= (1 - P_{eq}(i))P_{eq}(i) \times \left(T_{ii}(t) - \sum_{j \neq i} T_{ji}(t) \cdot \frac{P_{eq}(j)}{1 - P_{eq}(i)}\right) \\ &= (1 - P_{eq}(i))P_{eq}(i) \times \left(T_{ii}(t) - \frac{P_{eq}(i)(1 - T_{ii}(t))}{1 - P_{eq}(i)}\right) \\ &= (T_{ii}(t) - P_{eq}(i))P_{eq}(i) \end{aligned} \tag{S6}$$

The normalization condition and detailed balance condition are applied.
The normalized population fluctuation relaxation function of state i is,

$$\hat{C}_i(t) = \frac{T_{ii}(t) - P_{eq}(i)}{1 - P_{eq}(i)} \tag{S7}$$

We choose to normalize the total population fluctuation relaxation function so that it can be expressed as a weighted sum of each state i,

$$\hat{C}_{tot}(t) = \sum_i \frac{(T_{ii}(t) - P_{eq}(i))P_{eq}(i)}{1 - P_{eq}(i)} \quad \text{(S8)}$$

The propagator elements can be written as a sum of eigenvalues multiplied by corresponding eigenvectors,

$$T_{ii}(t) = \sum_n \psi_n^R(i)\psi_n^L(i)\lambda_n^t \quad \text{(S9)}$$

where $\psi_n^R(i)$ and $\psi_n^L(i)$ are the i^{th} element of the n^{th} right and left eigenvectors of the propagator matrix. λ_n is the n^{th} eigenvalue of the propagator matrix.

D. The mean first passage times (MFPT) from solving linear equations

The MFPTs from U to F and from U-i to i can be calculated in two ways. In the main text the MFPTs are expressed in terms of the time integral of propagator elements (eq. 4a). They can also be determined by solving a set of linear equations[1]. We let m_{ij} denote the MFPT in going from state i to state j. We may compute the MFPTs by solving the system of linear equations.

$$m_{ij} = \tau + \sum_{k \neq j} T_{ik} m_{kj} \quad \text{(S10)}$$

Then, the average MFPT to state j is

$$\text{MFPT}_j = \sum_i \frac{P_{eq}(i)}{1 - P_{eq}(j)} m_{ij} \quad \text{(S11)}$$

E. Average lifetime and the mean first passage time (MFPT)

It is straightforward to show that the relationship between the average lifetimes in eq.6 of the main text is an identity. The unfolded state space is divided into two states i and U-i. In an exhaustively long trajectory, we have

$$P_{eq}(i) = \frac{\sum_{k=1}^{N_i} t(k,i)}{t(tot)} \text{ and } P_{eq}(U-i) = \frac{\sum_{k=1}^{N_{U-i}} t(k,U-i)}{t(tot)} \quad \text{(S12)}$$

Here t(k,i) is the time span of the k-th segment when the trajectory is in state i; there are N_i such segments. Similarly t(k,U-i) is the time span of the k-th segment in state U-I and there are N_{U-i} such segments. Note that $|N_i - N_{U-i}| \leq 1$, i.e. they differ by at most 1, since the system is in one of the two states at any moment. Since N_i and N_{U-i} are very large, we can write $N_i = N_{U-i} = N$. Then

$$\frac{P_{eq}(i)}{P_{eq}(U-i)} = \frac{\sum_{k=1}^{N} t(k,i)}{\sum_{k=1}^{N} t(k,U-i)} \quad \text{(S13)}$$

Dividing both the numerator and denominator of the right hand side of Eq. S13 by N, we obtain

$$\frac{P_{eq}(i)}{P_{eq}(U-i)} = \frac{t_i}{t_{U-i}} \quad \text{(S14)}$$

where t_i and $t_{U\text{-}i}$ are the average lifetimes of the state i and U-i, respectively. Eq. 6 is obtained by rearranging the above eq. S14 in terms of $t_{U\text{-}i}$.

From eq. S14 after reorganization,

$$t_{U-i} = t_i \cdot \frac{P_{eq}(U-i)}{P_{eq}(i)}$$

$$= \frac{1}{\sum_{j\neq i} k_{ij}} \cdot \frac{1-P_{eq}(i)}{P_{eq}(i)}$$

$$= \frac{1-P_{eq}(i)}{\sum_{j\neq i} k_{ji} P_{eq}(j)}$$

The denominator of the equation above is the flux from U-i to i, $J_{U\text{-}i}$. Therefore, we have shown that the average lifetime of state U-i can be determined by

$$t_{U-i} = \frac{P_{eq}(U-i)}{J_{U-i}} \tag{S15}$$

Generally, the average lifetime $t_{U\text{-}i}$ in eq. 6 and the MFPT in eq. 4a of the main text are not the same unless the equilibration within the collective state U-i is much faster than the MFPT to state i. We can verify this analytically in a 3-node toy model (see Figure S4). Suppose we are going to calculate the average lifetime of the collective state which consists of node 1 and node 3. In the following steps, we will show that the average lifetime of the collective state t_{1+3} and the MFPT to node 2 $MFPT_2$ are equal to each other when the equilibration is rapid within the collective state, in other words, $k_u >> k_{slow}$ and $k_u >> k_{fast}$. From eq. 6 of the main text,

$$t_{1+3} = t_2\left(\frac{1}{P_{eq}(2)} - 1\right)$$

$$= \frac{2}{k_{slow}+k_{fast}} \tag{S16}$$

where the average lifetime of node 2, t_2, is the reciprocal of the sum of all the outgoing rates of node 2. When the rate constants are not available, the average lifetime of node 2 can also be estimated by $t_2 = \frac{\Delta t}{1-T_{22}(\Delta t)}$. To calculate the MFPT to node 2 we can use the eq. 4a in the main text.

$$MFPT_2 = \frac{k_{slow}+k_{fast}+4k_u}{2(k_{slow}k_{fast}+k_u(k_{slow}+k_{fast}))} \tag{S17}$$

When k_u goes to zero and $k_{fast} >> k_{slow}$,

$$MFPT_2 = \frac{1}{2k_{slow}} + \frac{1}{2k_{fast}} \approx \frac{1}{2k_{slow}} \tag{S18}$$

As we can see in general the average lifetime of the collective state is not equal to the MFPT to node 2.

However, when $k_u >> k_{fast} >> k_{slow}$ the $MFPT_2$ converges to t_{1+3}.

$$MFPT_2 = \frac{2}{k_{slow}+k_{fast}} \approx \frac{2}{k_{fast}} \quad \text{(S19)}$$

which equals the average lifetime of the collective state (1+3). So depending on whether the equilibration within the collective state (1+3) is very slow or very fast, the mean first passage time to state 2 goes to very different limits. It can also be shown for this model that when the equilibration of the collective state (1+3) is very fast, the first passage time distribution to state 2 is single exponential with a rate equal to the smaller non-zero eigenvalue of the rate matrix.

As shown analytically here using the 3-node model, there is agreement between eq. 4a, eq. 5, and eq. 6 in the main text when the equilibration within the collective state is very rapid. We have also shown the agreement numerically between eq. 4a, eq. 5, and eq. 6 for the 20-node Trp-Cage network where the equilibration within the unfolded state is much faster than the folding. A similar result is reported in reference 2 below.

References

1. Theodore J. Sheskin (1995) Computing mean first passage times for a Markov chain. International Journal of Mathematical Education in Science and Technology, 26-5, 1995.

2. Ellison PA, Cavagnero S (2006) Role of unfolded state heterogeneity and en-route ruggedness in protein folding kinetics. Protein Sci. 15:564–582.

Figure S1a

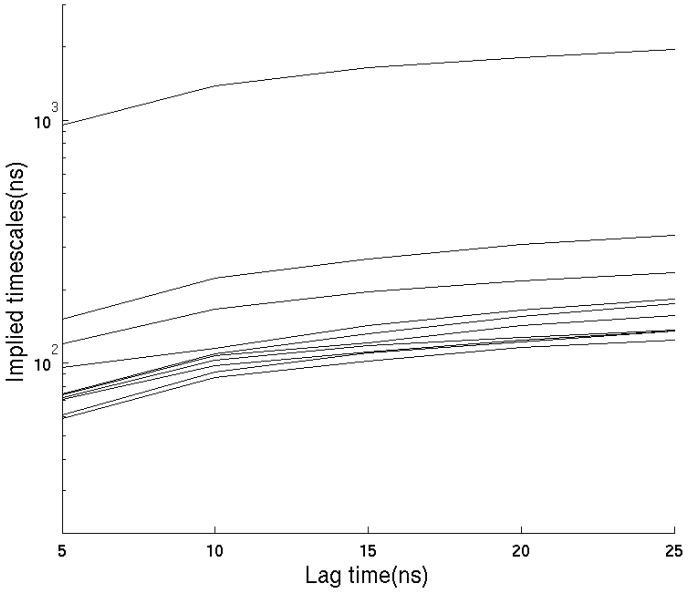

Figure S1b

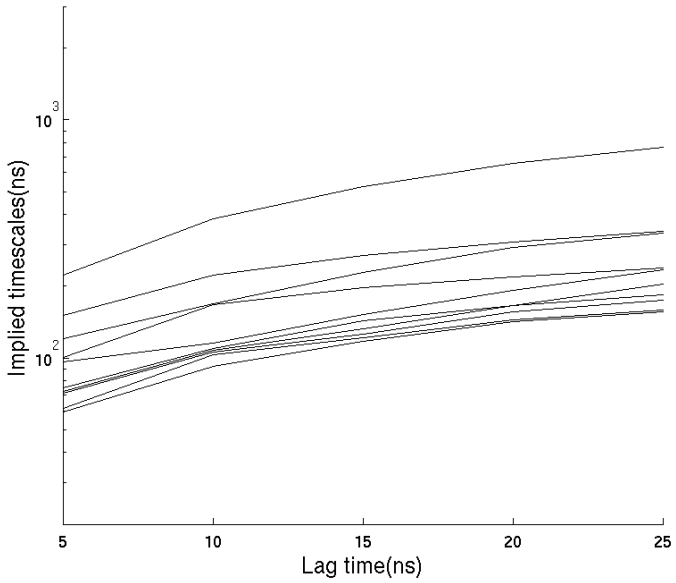

Figure S1. The implied timescales of the fine-grained network corresponding to the 10 slowest decaying eigenmodes using transition matrices **T**(Δt), with different boundary conditions, equilibrium(1a) and reflecting at F state(1b). The results are similar to those in Fgure 1 in the main text.

Figure S2a

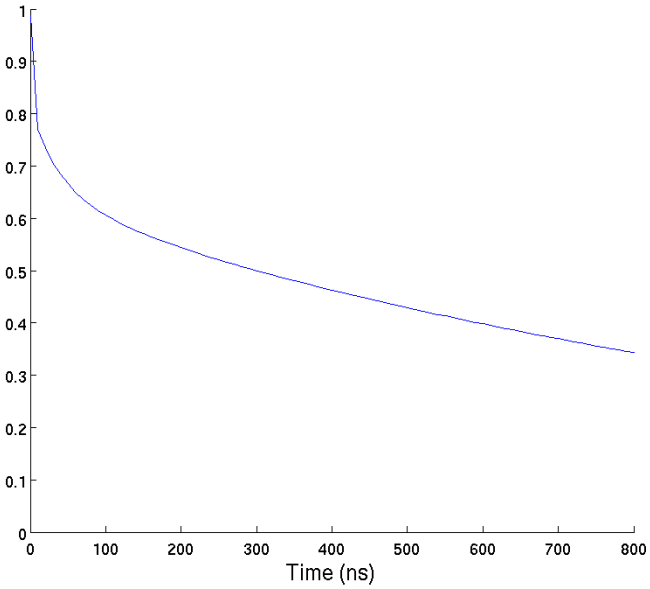

Figure S2b

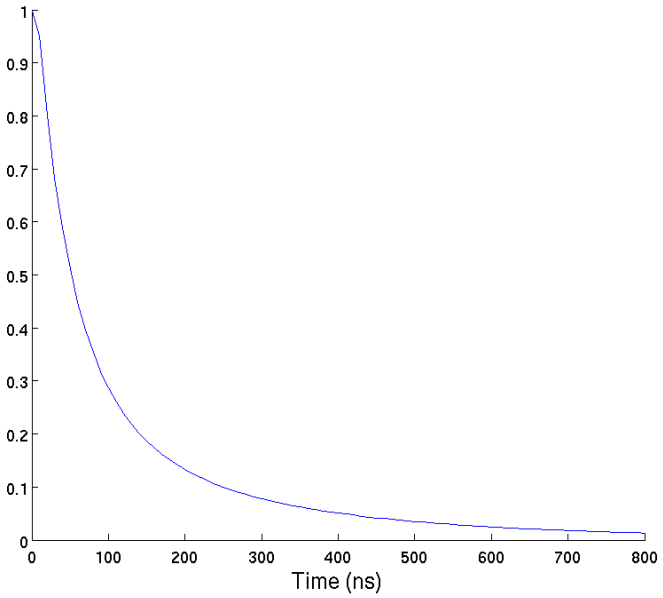

Figure S2. The population fluctuation relaxation functions of 25000-node network at lag time of 10ns, using two different boundary conditions, equilibrium (2a) and reflecting (2b). The relaxation functions are evaluated by spectral decomposition of the transition matrices. Only the 100 largest eigenvalues and their eigenvectors are used. Besides, the functions are renormalized by the population of the states which have a relaxation time larger than 8.2ns since we don't have enough statistics to estimate the relaxation time which is smaller than 8.2ns. The corresponding relaxation time is 825ns and 108ns.

Figure S3

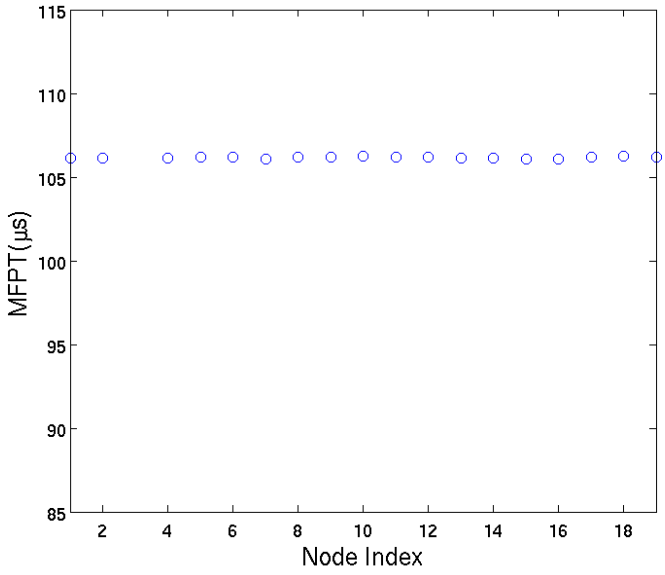

Figure S3. The MFPT to node 3(which is marked as red in Figure 4b of the main text) of 20-node network is shown. We can see the MFPT to node 3 does not strongly depend on the starting point. This implies a rapid equilibration within the unfolded state.

Figure S4

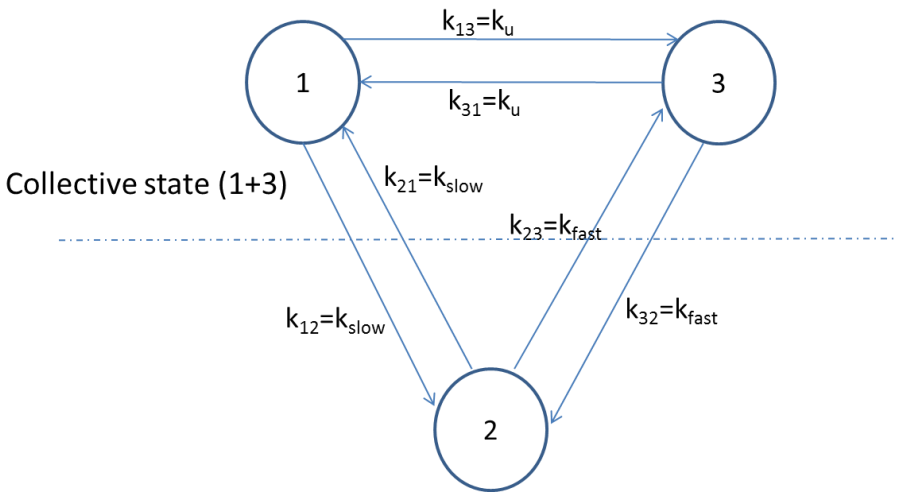

Figure S4. The schematic plot of a 3-node model with rate constants labeled.