Correction

BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Correction for “Protein folded states are kinetic hubs,” by Gregory R. Bowman and Vijay S. Pande, which appeared in issue 24, June 15, 2010, of *Proc Natl Acad Sci USA* (107:10890–10895; first published June 1, 2010; 10.1073/pnas.1003962107).

The authors note that, due to a printer’s error, on page 10892, left column, paragraph 3, line 8, “approximately 370 ns” should instead appear as “approximately 370 μs.”

www.pnas.org/cgi/doi/10.1073/pnas.1012332107

Protein folded states are kinetic hubs

Gregory R. Bowman[a] and Vijay S. Pande[a,b,1]

[a]Biophysics Program and [b]Department of Chemistry, Stanford University, Stanford, CA 94305

Communicated by Hans C. Andersen, Stanford University, Stanford, CA, March 24, 2010 (received for review November 2, 2009)

Understanding molecular kinetics, and particularly protein folding, is a classic grand challenge in molecular biophysics. Network models, such as Markov state models (MSMs), are one potential solution to this problem. MSMs have recently yielded quantitative agreement with experimentally derived structures and folding rates for specific systems, leaving them positioned to potentially provide a deeper understanding of molecular kinetics that can lead to experimentally testable hypotheses. Here we use existing MSMs for the villin headpiece and NTL9, which were constructed from atomistic simulations, to accomplish this goal. In addition, we provide simpler, humanly comprehensible networks that capture the essence of molecular kinetics and reproduce qualitative phenomena like the apparent two-state folding often seen in experiments. Together, these models show that protein dynamics are dominated by stochastic jumps between numerous metastable states and that proteins have heterogeneous unfolded states (many unfolded basins that interconvert more rapidly with the native state than with one another) yet often still appear two-state. Most importantly, we find that protein native states are hubs that can be reached quickly from any other state. However, metastability and a web of nonnative states slow the average folding rate. Experimental tests for these findings and their implications for other fields, like protein design, are also discussed.

Markov state model | network | protein folding

Molecular kinetics has fascinated biophysicists and biochemists for decades. From a biophysical point of view, it remains a mystery how systems with so many possible configurations can self-organize with such specificity and rapidity, carry out catalysis, and trigger signaling cascades. From a biomedical standpoint, protein misfolding causes many debilitating diseases, including Alzheimer's, Huntington, and Parkinson diseases (1). Understanding how proteins fold is a logical first step in understanding how they misfold and, more importantly, how to prevent or recover from misfolding; indeed, this approach is already proving valuable (2). Furthermore, a better understanding of protein folding mechanisms could lead to more efficient structure prediction (3, 4), for use in high throughput proteomics and studies of systems that defy experimental characterization, and better models for molecular kinetics could aid in computational drug and protein design.

What would the ultimate theory of molecular kinetics look like though? A natural way of answering this question is by analogy to well-established theories, such as Schrodinger's equation in the successful field of quantum mechanics. On the one hand, computational solutions to Schrodinger's equation have yielded quantitative agreement with and prediction of experimental observables. However, equally important is this theory's ability to yield insight into simple systems, such as the particle in a box, for the purposes of gaining an intuition for fundamental principles, like the quantization of energy and the role of molecular orbitals. Likewise, the ultimate theory of molecular kinetics should be capable of scaling from sophisticated models capable of quantitatively predicting experiments to simple models that yield mechanistic insight. At even the most fundamental levels of this hierarchy, such a theory ought to be at least qualitatively consistent with experimental observations and be capable of generating experimentally testable hypotheses. In particular, such a theory ought to provide insight into protein folding as success in describing such drastic conformational changes would be evidence for the theory's ability to describe less extreme ones.

We propose that networks of metastable, or long-lived, states (5–8) could fulfill this role because they are implicit in even the most simple protein folding models; examples include U ↔ N and U ↔ I ↔ N where U is the unfolded state, I is an intermediate, and N is the native state. Networks called Markov state models (MSMs) make these implicitly considered properties explicit and have the potential to provide complete maps of a protein's free energy landscape, with nodes corresponding to metastable states (or free energy basins) and edges representing the probabilities of transitioning between pairs of these states (5–10).

A number of recent works have provided validation for these networks by showing that they can yield quantitative agreement with experimentally derived structures and folding rates (6, 11–13). In particular, the predicted native state from our villin model (based on calculated free energies) had an rmsd to the crystal structure of approximately 1.8 Å (6). The model also correctly predicted quantitative details of the kinetics, such as the absolute folding rate (to logarithmic accuracy). This degree of accuracy in predicted free energies, structures, and rates is crucial as all experimental measurements are functions of these properties. In all, the agreement between theory and experiment leads us to the conclusion that our models provide a sufficiently accurate reflection of reality.

To further flesh out this potential theory of molecular kinetics, we have delved into the nature of the free energy landscapes of the villin headpiece (HP-35 NleNle) (14) and a 39 residue fragment of NTL9 (15). Furthermore, because complex networks for real systems are difficult to comprehend, we construct simple, generic models that capture qualitative phenomena like apparent two-state folding and provide an intuition for molecular kinetics. Together, these models allow us to assess existing theories, which describe folding as a two-state process characterized by cooperative transitions across a dominant free energy barrier separating a rapidly mixing unfolded ensemble from the native state (16, 17).

The remainder of this paper will be organized around three key results. First, protein free energy landscapes can yield apparent two-state behavior even in the absence of a single dominant barrier. Second, protein unfolded states are heterogeneous, having multiple basins that interconvert more rapidly with the native state than one another. Third, protein native states are kinetic hubs: it is possible to reach them relatively quickly from anywhere in a network but it is also possible to get stuck in a web of nonnative states.

Results and Discussion

Apparent Two-State Behavior Can Occur In the Absence of a Kinetically Relevant Two-State Decomposition. Many proteins appear to fold via a single cooperative transition from a rapidly mixing

Author contributions: G.R.B. and V.S.P. designed research; G.R.B. performed research; G.R.B. contributed new reagents/analytic tools; G.R.B. analyzed data; G.R.B. and V.S.P. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. E-mail: pande@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1003962107/-/DCSupplemental.

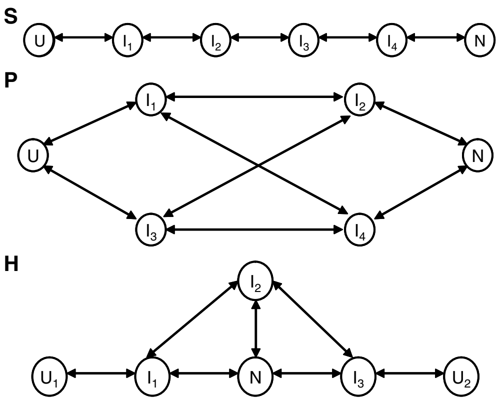

Fig. 1. Three representative networks each having unfolded state(s) (U and U_i), intermediates (I_i), and a native state (N). S has a single pathway, P has parallel pathways, and H has a heterogeneous unfolded state.

ensemble of unfolded conformations to a well defined native structure (16, 18). However, based on chemical intuition, one would expect to find many more metastable states, corresponding to the numerous favorable interactions that could form in the absence of the full native structure as well as dynamics within the native state. To reconcile these points, one typically assumes a single dominate free energy barrier that serves as the rate limiting step for folding. Other barriers are often assumed to be small relative to the thermal energy (or at least to the dominant barrier) and the equilibrium probability of any intermediate is assumed to be too small to detect.

However, in some cases modeling experimental data requires the use of at least three states (19–21) and simple toy models have shown that even three-state systems can yield apparent two-state behavior (22). Thus, it is natural to hypothesize that many systems may have more complex arrangements of metastable states (5, 23, 24) yet still exhibit apparent two-state behavior.

To test this hypothesis, we first turn to an MSM for the villin headpiece. This MSM was recently built from atomistic simulations and, by assuming stochastic jumps between its states, was shown to give quantitative agreement with experimental structures and folding rates in addition to recapitulating the raw simulation data (6). Thus, the presence of numerous metastable states in this model would be strong evidence for their actual existence and the stochastic nature of transitions between them. Indeed, with a lagtime on the order of 10ns (Fig. S1), analysis of this MSM reveals the existence of at least 500 metastable states. At least 2,000 are found for NTL9 (13). The free energy barriers between our villin states have an average height of about $5.9(+/-2.5)$ kT (*SI Text*), indicating that they are nontrivial and potentially detectable. Moreover, no single dominant barrier is apparent.

To better understand the system specific results from our all-atom models, we now consider three simple models for dynamics capable of providing insight into protein folding in general. Each of these networks has six metastable states and is depicted in Fig. 1. These models have a single folding pathway (S), parallel folding pathways (P), and a heterogeneous unfolded state (H, with multiple unfolded basins that each interconvert more rapidly with the native state than with one another) as discussed in *Materials and Methods*.

One may be tempted to associate the states in these models with folding nuclei (25), preorganized secondary structure (26), foldons (27), or the elements of some other model of protein folding (28). However, we simply require that they all be metastable. That is, a system within one state is more likely to stay there than to transition to a different state. Moreover, we propose that the concept of metastability unifies many of the previously proposed folding mechanisms, each of which describes some systems better than others, as all consist of basic units that are stable on some timescale.

We can now imagine monitoring stochastic transitions within each of these representative systems (or ensembles thereof) with a device that can only detect the native state. This hypothetical setup is equivalent to experiments wherein unfolded molecules are allowed to relax to an observable folded state where they are trapped to prevent unfolding and refolding. Fig. 2 shows that such an experiment yields the exponential behavior typical of an ideal two-state system. In fact, exponential fits to the data after the initial lag phase only give slight underestimates of the true mean first passage times (MFPTs) between the unfolded and folded states (Table S1). Thus, even these simple systems are qualitatively consistent with both stochastic jumps between numerous metastable states and apparent two-state behavior. This is particularly surprising for model H because it cannot be divided into a single, rapidly mixing unfolded basin separated from the native state by one dominant barrier (i.e., it is not two-state).

A kinetic perspective on our simple networks helps to explain why two-state behavior is often observed even when there are many large barriers. As discussed previously, when there is a single dominant rate then faster transitions will tend to be lost in the noise. Multiple slow rates will also be lost in the noise if they are too similar. Moreover, this same logic applies even when there are multiple folding routes from different starting points (and thus no kinetically relevant two-state decomposition). Thus, observing anything other than mainly single exponential kinetics requires a delicate balance wherein the slowest rates differ sufficiently to distinguish them but not so much that one dominates the rest, not to mention extremely precise measurements.

Fortunately, there is ample evidence that achieving this balance and the precision necessary to detect it are possible. Multiexponential behavior is often consistent with the experimental data, but fit to stretched exponentials (29, 30). Increasing the temporal resolution of single molecule pulling experiments has also steadily revealed more metastable states and kinetic measurements can be probe dependent (31, 32). We propose that the ability to simultaneously monitor multiple degrees of freedom (such as extension and FRET) in single molecule experiments would reveal even more metastable states, particularly if MSMs

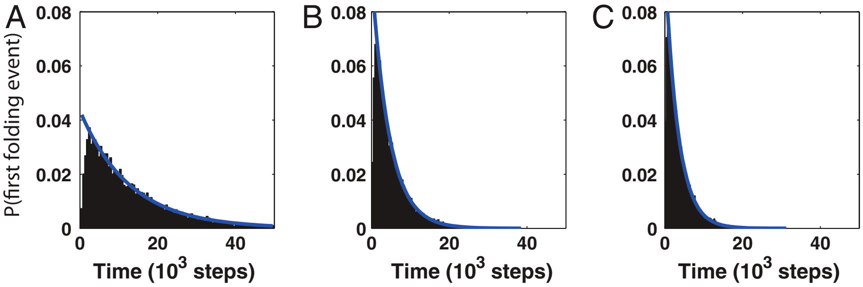

Fig. 2. Distributions of the first folding times for the simple networks S, P, and H are shown in (*A*), (*B*), and (*C*), respectively. The blue lines are exponential fits to the data after the initial lag phase.

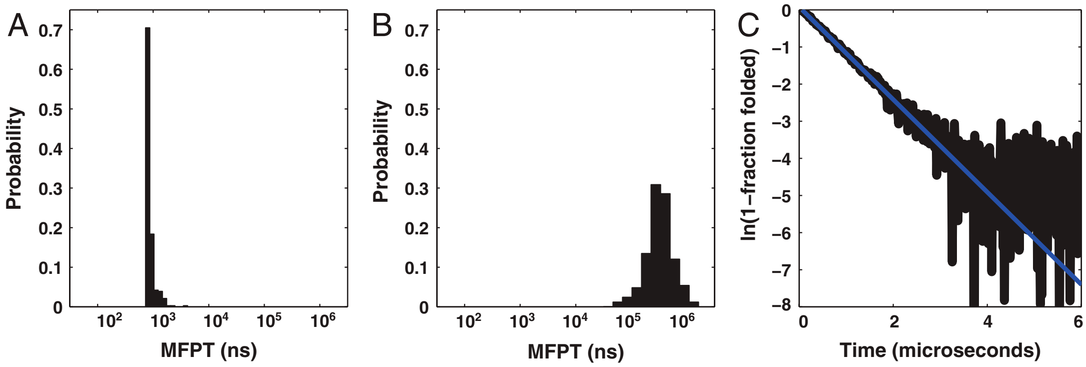

Fig. 3. Relaxation of villin from 500 state model. Distributions of the MFPTs from (*A*) unfolded states to the native state and (*B*) between unfolded states. (*C*) Relaxation kinetics with a 10:1 signal-noise ratio (black curve with Gaussian noise) and a single exponential fit (blue curve with $\tau = 810$ ns).

were used to choose the number of probes employed and their placement.

Proteins Have Heterogeneous Unfolded States with Multiple Basins that Interconvert More Rapidly with the Native State than Each Other. We now investigate which of the simple network topologies is most representative of real protein free energy landscapes. As a first step, we have calculated that every state can reach the native basin of our villin model in one or two steps. This eliminates the possibility of a single pathway because states with that topology could require up to 499 steps to reach the native basin.

Determining whether the parallel pathway model (17, 33, 34) or the heterogeneous unfolded state model is more representative of villin requires a definition of the unfolded state(s). Because every nonnative state can reach the native basin in one or two steps it is natural to label every state that is not directly connected to the native state (332 in all) as unfolded and all other nonnative states (167 in all) as intermediates.

Taking this definition, we can now examine the distribution of MFPTs from each unfolded state to the native state as well as the distribution of MFPTs between all pairs of unfolded states. Doing so reveals that the average MFPT to the native state is $880(+/-270)$ ns, in reasonable agreement with the experimentally predicted folding time of 720 ns (14). Moreover, this value is much lower than the average MFPT between pairs of unfolded states (approximately 370 ns), as shown in Fig. 3 *A* and *B*. Considering every nonnative state as part of the unfolded ensemble also gives similar distributions (Fig. S2), implying that these results are robust to the exact definition of the unfolded state. Similar results are found for NTL9 as well (Fig. S3). Thus, we can conclude that the heterogeneous unfolded state model is most representative of our villin and NTL9 models and possibly proteins in general. This result is in contrast to existing theories of protein folding, which assume rapid equilibration within the unfolded ensemble (17, 35, 36).

Examination of representative structures suggests that nonnative interactions (often in the context of relatively compact conformations) and the enormity of conformational space are responsible for slow transitions between unfolded basins (Fig. S4). Nonnative contacts can easily have free energies on the order of native contacts, making nonnative states reasonably metastable. Once a set of nonnative contacts is broken, the probability of forming a particular set of other nonnative contacts is quite small due to the large number of other possibilities. This small probability is equivalent to a slow rate. In contrast, evolutionary pressure to fold makes transitioning to the native state reasonably probable, which equates to fast folding relative to slower transitions between unfolded basins.

The tight distribution of MFPTs to the native state is also consistent with our explanation of apparent two-state behavior. Due to experimental noise, it is difficult to justify using more than one or two exponentials to fit the relaxation of our coarse-grained villin model with 500 states, as shown in Fig. 3*C*. Only with an extremely high signal to noise ratio can one accurately identify the deviations from single exponential relaxation shown in Fig. S5. We also note that more fine-grained models for villin can capture the burst phase in its relaxation (Fig. S6) but here we emphasize the ability of our coarse-grained model to capture the apparent two-state behavior that dominates this system's relaxation (14).

Our ability to reconcile our model with existing experimental data on the nature of the unfolded ensemble (specifically under native conditions, as opposed to the more rapidly mixing denatured state) indicates that more experiments will be required to definitively falsify or support our conclusions. For example, Nettels et al. have reported a 50 ns global relaxation time within the unfolded ensemble (37). Our model, however, would suggest that this may be due to relaxation within individual unfolded basins, not between them. This hypothesis is consistent with recent measurements of slow dynamics in the unfolded ensemble from the Lapidus lab (38, 39). Therefore, we suggest that this may be an interesting direction for future experimental work. In addition to existing methodologies for probing the unfolded ensemble, single molecule experiments monitoring multiple degrees of freedom could help to falsify or support our conclusions.

If our heterogeneous unfolded state model is indeed generally true then protein folding kinetics cannot be accurately described by two-states separated by a single barrier. Instead, folding must be understood in terms of multiple pathways starting from a number of distinct states. Mixing between pathways adds another layer of complexity to the folding process. Modeling the effects of mutations will thus require considering changes in the relative free energies of numerous states and barrier heights. Understanding the global effects of small changes on networks will likely also be important for protein design.

A Native Hub Allows Rapid Folding but Proteins Can Still Get Stuck In a Web of Nonnative States. The accessibility of villin's native state implies the hub-like connectivity characteristic of small-world and scale-free networks (40, 41). We can test this hypothesis by counting the number of connections observed between states because only those transitions with probabilities above some threshold are observed with our finite sampling (all transitions would be observed with infinite sampling). Examining subsets of the states independently, one finds that the average degree (or number of connections) increases as one moves from the unfolded states to the native basin. The unfolded states have an average degree of 12 whereas the intermediate states have an average degree of 25. The native state acts as a hub, connected to 167 other states. Similar results are found for a small β-sheet peptide (42) and NTL9.

Reduced connectivity between nonnative states results in slow dynamics within the unfolded ensemble. This connectivity contradicts other models, which predict bottlenecks close to the native state and high connectivity in nonnative regions (17, 34, 36, 43), as depicted in Fig. 4*A*. A more thorough discussion of the similarities and differences between our model and those proposed previously is given in the next section.

The native hub explains how villin folds so quickly. Just as there are only about six degrees of separation between people in the United States (44), it is possible to reach villin's native state in one or two jumps (each 15 ns). Therefore, it is possible to fold from anywhere in the landscape in 30 ns or less. This result is consistent with recent experimental work showing that the transition path time between the unfolded and native ensembles can be as much as four orders of magnitude faster than the average folding time (45) and likely results from evolutionary pressure to fold quickly.

Due to the kinetic proximity of the native state with a 15 ns lagtime, we see that villin can fold in just 30 ns; however, such trajectories are rare because the metastability and connectivity of nonnative states makes taking a direct route to the native state improbable. Instead, villin will often spend considerable time in a web of nonnative states before finally folding, resulting in an average folding time on the microsecond timescale. In the future, it will be interesting to test whether slower folding proteins have unfolded states further from the native one or just more strongly metastable states, which equates to higher barriers and slower transitions between states. Preliminary analysis of NTL9 suggests every basin can reach the native state in 5 steps (approximately 100 ns) or less.

We have also found a rough correlation between the connectivity of states and their equilibrium probabilities. The average probabilities of unfolded and intermediate states are approximately 0.0005 and approximately 0.004, respectively. The native state has an equilibrium probability of approximately 0.2. Fig. 4*B* shows a schematic of a protein folding network that attempts to capture all of these observations in a humanly comprehendible manner. All of these observations are in qualitative agreement regardless of the degree of lumping; that is, whether one uses smaller and more numerous states to capture more local minima in the landscape or fewer and more voluminous states to obtain an even more coarse-grained model. Whereas one may be tempted to consider Fig. 4*B* merely an alternative depiction of a funnel, we emphasize that the kinetic connectivity of the native state and lack of connectivity within the unfolded ensemble are important qualitative deviations from traditional funnel theory (17).

An important methodological consequence of the network topology found here is that many short, parallel simulations (or experiments) started from arbitrary initial points are an excellent way of exploring the entire free energy landscape. In the extreme case of using a single starting point, one could still reach every free energy basin despite the presence of numerous metastable states so long as each simulation was longer than the diameter of the network (the minimal time that allows one to reach any state from an arbitrary starting point). However, reaching every state would be impossible with simulations that were shorter than the diameter of the network. Thus, our network theory provides an alternate explanation for the previously noted need to have simulations longer than some minimal lag phase, which was then attributed to the need to equilibrate within the unfolded state before folding in two-state systems (46).

Another simple but more efficient strategy would be to start simulations from multiple conformations dispersed throughout phase space and run them long enough to ensure mixing between them and coverage of the entire space. In fact, Fig. 5 and Fig. S7 show that such a scheme is actually more valuable than a few long trajectories, using a relative entropy metric for MSMs from ref. 47

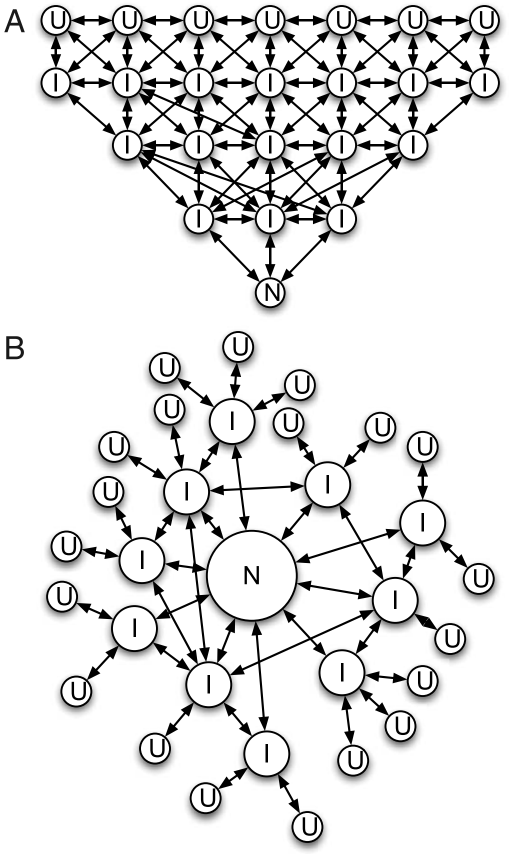

Fig. 4. Schematic diagrams of funnel and native hub models having unfolded states (U), intermediates (I), and native states (N). (*A*) A network description of a folding funnel with nodes corresponding to individual conformations and a bottleneck near the native state. (*B*) A native hub model with metastable nodes. The size of each node in (*B*) is correlated with its equilibrium probability and the connectivity falls off as one moves away from the native state.

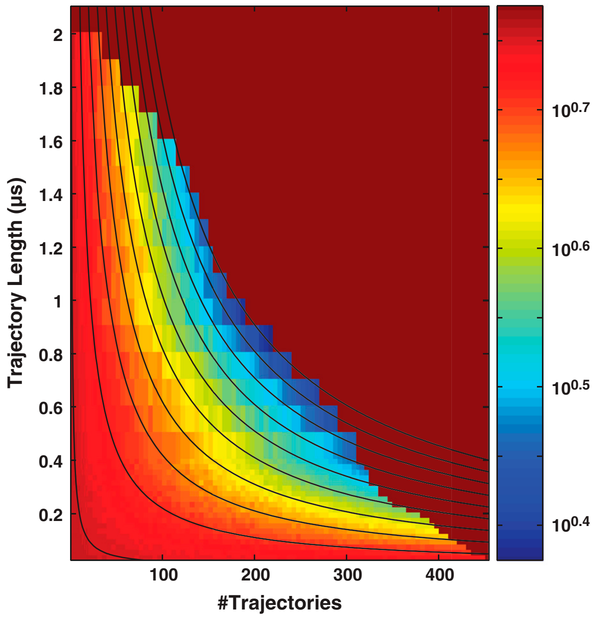

Fig. 5. Distance between the final villin MSM and MSMs constructed from subsets of the data (varying trajectory length and number of trajectories). Distance is measured by a relative entropy metric (*SI Text*). Black lines are contours of equal amounts of data. No data was available for the upper-right portion of the graph.

to measure the information content of different datasets relative to our validated villin model. However, this trend can be seen to break down for simulations that are insufficiently long or too few as they are unlikely to reach every state or traverse every possible pathway between pairs of states. The simulation length at which this breakdown occurs decreases as the number of simulations increases though. Even better performance can be obtained using adaptive sampling algorithms (47, 48), which direct sampling to where it is needed most to improve a model.

Comparison to Previous Theories for Protein Folding. There is a long history of theoretical models for protein folding (28) so it is important to put our work in the context of these previous theoretical approaches. In particular, folding funnel models (17, 36, 43) have dominated much of how the field currently conceptualizes protein folding and hence it is natural to compare our model to such theories. One of the most similar funnel categories is type0B, which is characterized by overall downhill folding interrupted by a glass transition along the reaction coordinate (17). Whereas this regime does include slow dynamics between compact states, it also results in a small number of folding pathways relative to higher connectivity in the unfolded ensemble. In addition, this and other previous funnel-based models have explicitly described rapidly interconverting unfolded states, as reflected in the "bottleneck" discussed in previous works (34, 35), as well as the choice of structurally based reaction coordinates like the number of native contacts (Q) (17, 35), which directly requires that dynamics along orthogonal degrees of freedom, such as interconversion between unfolded conformations, is rapid compared to folding. In contrast, we find a large number of folding pathways, slow dynamics between unfolded states relative to folding, and no glass transition. Our folding rates are also quite similar, rather than the different rates characteristic of the folding pathways in type0B folding.

Other funnel models have recognized the possibility of a large number of folding pathways (17, 33, 43), but still in the context of fast dynamics within the unfolded basin relative to slower transitions to the folded state. Some have even gone so far as to assume global connectivity (49, 50); however, even these emphasize that local connectivity would dominate in the full dimensional conformational space and global connectivity only arises when projecting onto a few order parameters. Furthermore, they argue global connectivity will not give an activation barrier and, therefore, these models are primarily intended for studies of downhill folding or the early activationless stages of folding. Our model, on the other hand, has a native hub and slow dynamics in the unfolded state relative to faster folding regardless of the degree of coarse-graining one employs. We also demonstrate that this can result in apparent two-state folding (i.e., activated kinetics) and that this occurs in nondownhill folding proteins, such as the millisecond folding NTL9.

Conclusions

Many biological systems, ranging from signaling pathways to social networks, can be most naturally described as networks. As a field, we have now established an additional level to this hierarchy: a network theory for molecular kinetics that is able to map out the free energy landscapes of proteins and other macromolecules in their entirety.

Previous work has demonstrated that this network theory is capable of quantitative agreement with experiments (6, 11–13) and we have now shown that it can also scale down to simple, generic models. Using this theory at both the quantitative and qualitative levels, we have provided an intuition for conformational changes as drastic as protein folding and this intuition has led to experimentally testable insights into the nature of protein free energy landscapes.

We have focused on three insights from these network models, which appear to hold regardless of the degree of coarse-graining one employs and can be reconciled with current experiments. First, even models that defy a kinetic decomposition into two states often give rise to apparent two-state behavior. Second, proteins have heterogeneous unfolded states (multiple basins that each interconvert more rapidly with the native state than with one another, preventing a kinetic decomposition into two states). Third, proteins have a native hub. Thus, it is possible to fold quickly from anywhere in the landscape but proteins often get stuck in a web of nonnative states before finally folding, greatly increasing the average folding time.

These properties are a natural result of reasonably strong nonnative interactions and the enormous number of nonnative conformations a protein can adopt, in combination with evolutionary pressure to fold quickly (for example, to avoid aggregation). Therefore, we suggest that these conclusions are likely true of proteins in general. Our approach also unifies other models for protein folding by recognizing that each of them builds upon elements, whether they are called folding nuclei (25) or foldons (27), which correspond to different types of metastable states.

We look forward to a fruitful future of drawing on network theory to better understand molecular kinetics and guide experiments probing both general properties and system specific details. In particular, can one reinterpret the many experiments that have been analyzed under a two-state assumption? If so, that could shed light on the chemistry of the underlying structures that leads to the network topology and dynamics described here. Moreover, can further experiments be designed to directly probe the unfolded state under native conditions (rather than with denaturant or high temperature, where mixing is more rapid) to directly test the predictions made here? We also hope to explore how the methodologies developed for building and understanding biomolecular networks may be applicable to other types of networks, especially as network theorists attempt to develop a general framework for understanding network dynamics.

Materials and Methods

Atomic Resolution Protein Folding Simulations and Networks. Ref. 6 describes the use of the MSMBuilder package (https://simtk.org/home/msmbuilder/) (23) to construct an MSM with 10,000 microstates for the villin headpiece (HP-35 NleNle). This model was based on approximately 450 all-atom, explicit solvent simulations, each up to 2 μs in length, for a total simulation time of 354 μs (51). Whereas the longest timescale transitions in the model from ref. 6 were found to be Markovian, implying memoryless transitions between metastable states, not every state was metastable. We used MSMBuilder to lump kinetically related microstates into 500 metastable macrostates to ensure a direct correlation between states in the MSM and free energy basins, as described in *SI Text*. This is equivalent to common experimental analyses in which the potential is smoothed and the friction is rescaled. We note, however, that the free energy landscape for this system is actually a hierarchy of basins so it is possible to build many valid MSMs with different numbers of states. As a result, one would not expect there to be exactly 500 experimentally detectable states. Regardless of the resolution at which one examines this hierarchy, however, requiring that each state is metastable ensures that they are directly related to a free energy basin. Thus, our networks of metastable states are an important step beyond previously described networks, which often used simpler approximations to define state boundaries and the transition rates between states (17, 34, 42, 52, 53). An additional 40,000 simulations, each up to 400 ns in length (for a total simulation time of 14 milliseconds), were also assigned to this MSM to explore the effect of using more simulations.

Preliminary results for a 39-residue fragment of NTL9 are based on an MSM built from approximately 1.5 milliseconds of simulation in implicit solvent with a different force field (13). Similarities between these two systems thus suggest our results are not a force field artifact.

Simple Models. We have designed three simple networks, depicted in Fig. 1, that capture the essence of various protein folding mechanisms. Each of these models has six metastable states with approximately the same equilibrium and transition probabilities so that differences between their behaviors may be attributed to differences in their topologies (*SI Text*).

The first model (S) has a single folding pathway. This model is a natural extension of the common U ↔ I ↔ N model (19, 54) and is often used to justify the expense of running long simulations as shorter ones could fail to reach every state.

The second model (P) has parallel folding pathways. Parallel folding pathways have been proposed for a number of systems (20, 21, 33, 51). In addition, this model emphasizes the need to observe numerous folding and unfolding transitions to obtain sufficient statistics on the entire process. The increased connectivity relative to S also results in faster timescales.

The third model (H) has a heterogeneous unfolded state—multiple unfolded basins that each interconvert more rapidly with the native state than with one another. Thus, there is no kinetic decomposition of this model into two states, one folded and one unfolded. This model was inspired by a growing body of work on the presence of deep minima and gutters in unfolded regions of conformational space (38, 39, 55–57).

ACKNOWLEDGMENTS. We thank V. Voelz, K. Beauchamp, and D. Ensign for helpful discussion. This work was funded by National Institutes of Health Grant R01-GM062868. Computer resources were provided by the Folding@home users and National Science Foundation Award CNS-0619926. G.R.B. was supported by the National Science Foundation Graduate Research Fellowship Program.

1. Uversky VN (2009) Intrinsic disorder in proteins associated with neurodegenerative diseases. *Front Biosci* 14:5188–5238.
2. Kelley NW, Vishal V, Krafft GA, Pande VS (2008) Simulating oligomerization at experimental concentrations and long timescales: A Markov state model approach. *J Chem Phys* 129:214707.
3. Bowman GR, Pande VS (2009) The roles of entropy and kinetics in structure prediction. *PLoS One* 4:e5840.
4. Ozkan SB, Wu GA, Chodera JD, Dill KA (2007) Protein folding by zipping and assembly. *Proc Natl Acad Sci USA* 104:11987–11992.
5. Schutte C (1999) Conformational dynamics: Modeling, theory, algorithm, and application to biomolecules. PhD thesis (Freie Universitat, Berlin).
6. Bowman GR, Beauchamp KA, Boxer G, Pande VS (2009) Progress and challenges in the automated construction of Markov state models for full protein systems. *J Chem Phys* 131:124101.
7. Buchete NV, Hummer G (2008) Coarse master equations for peptide folding dynamics. *J Phys Chem B* 112:6057–6069.
8. Yang S, Banavali NK, Roux B (2009) Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proc Natl Acad Sci USA* 106:3776–3781.
9. Noe F, Fischer S (2008) Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr Opin Struct Biol* 18:154–162.
10. Chodera JD, Singhal N, Pande VS, Dill KA, Swope WC (2007) Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J Chem Phys* 126:155101.
11. Noe F, Schutte C, Vanden-Eijnden E, Reich L, Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci USA* 106:19011–19016.
12. Huang X, Bowman GR, Bacallado S, Pande VS (2009) Rapid equilibrium sampling initiated from nonequilibrium data. *Proc Natl Acad Sci USA* 106:19765–19769.
13. Voelz VA, Bowman GR, Beauchamp KA, Pande VS (2010) Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39). *J Am Chem Soc* 132:1526–1528.
14. Kubelka J, Chiu TK, Davies DR, Eaton WA, Hofrichter J (2006) Sub-microsecond protein folding. *J Mol Biol* 359:546–553.
15. Horng JC, Moroz V, Raleigh DP (2003) Rapid cooperative two-state folding of a miniature alpha-beta protein and design of a thermostable variant. *J Mol Biol* 326:1261–1270.
16. Jackson SE, Fersht AR (1991) Folding of chymotrypsin inhibitor 2.1. Evidence for a two-state transition. *Biochemistry* 30:10428–10435.
17. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG (1995) Funnels, pathways, and the energy landscape of protein folding: A synthesis. *Proteins* 21:167–195.
18. 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* 6:15001.
19. Spudich GM, Miller EJ, Marqusee S (2004) Destabilization of the *Escherichia coli* RNase H kinetic intermediate: Switching between a two-state and three-state folding mechanism. *J Mol Biol* 335:609–618.
20. Radford SE, Dobson CM, Evans PA (1992) The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature* 358:302–307.
21. Kamagata K, Sawano Y, Tanokura M, Kuwajima K (2003) Multiple parallel-pathway folding of proline-free Staphylococcal nuclease. *J Mol Biol* 332:1143–1153.
22. Ma H, Gruebele M (2006) Low barrier kinetics: Dependence on observables and free energy surface. *J Comput Chem* 27:125–134.
23. Bowman GR, Huang X, Pande VS (2009) Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* 49:197–201.
24. Wales DJ, Scheraga HA (1999) Global optimization of clusters, crystals, and biomolecules. *Science* 285:1368–1372.
25. Wetlaufer DB (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci USA* 70:697–701.
26. Myers JK, Oas TG (2001) Preorganized secondary structure as an important determinant of fast protein folding. *Nat Struct Biol* 8:552–558.
27. Krishna MM, Maity H, Rumbley JN, Lin Y, Englander SW (2006) Order of steps in the cytochrome C folding pathway: Evidence for a sequential stabilization mechanism. *J Mol Biol* 359:1410–1419.
28. Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. *Ann Rev Biophys* 37:289–316.
29. Volk M, et al. (1997) Peptide conformational dynamics and vibrational stark effects following photoinitiated disulfide cleavage. *J Chem Phys* 101:8607.
30. Sabelko J, Ervin J, Gruebele M (1999) Observation of strange kinetics in protein folding. *Proc Natl Acad Sci USA* 96:6031–6036.
31. Liu F, Gruebele M (2007) Tuning lambda6-85 towards downhill folding at its melting temperature. *J Mol Biol* 370:574–584.
32. Liu F, et al. (2009) A one-dimensional free energy surface does not account for two-probe folding kinetics of protein alpha(3)D. *J Chem Phys* 130:061101.
33. Ghosh K, Dill KA (2007) The ultimate speed limit to protein folding is conformational searching. *J Am Chem Soc* 129:11920–11927.
34. Betancourt MR, Onuchic JN (1995) Kinetics of protein like models: The energy landscape factors that determine folding. *J Chem Phys* 103:773.
35. Cho SS, Levy Y, Wolynes PG (2006) P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes. *Proc Natl Acad Sci USA* 103:586–591.
36. Leopold PE, Montal M, Onuchic JN (1992) Protein folding funnels: A kinetic approach to the sequence-structure relationship. *Proc Natl Acad Sci USA* 89:8721–8725.
37. Nettels D, Gopich IV, Hoffmann A, Schuler B (2007) Ultrafast dynamics of protein collapse from single-molecule photon statistics. *Proc Natl Acad Sci USA* 104:2655–2660.
38. Waldauer SA, et al. (2008) Ruggedness in the folding landscape of protein L. *HFSP J* 2:388–395.
39. Voelz VA, Singh VR, Wedemeyer WJ, Lapidus LJ, Pande VS (2010) Unfolded state dynamics and structure of protein L characterized by simulation and experiment. *J Am Chem Soc* 132:4702–4709.
40. Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. *Nature* 393:440–442.
41. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512.
42. Rao F, Caflisch A (2004) The protein folding network. *J Mol Biol* 342:299–306.
43. Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. *Nat Struct Biol* 4:10–19.
44. Milgram S (1967) The small world problem. *Psychol Today* 1:61–67.
45. Chung HS, Louis JM, Eaton WA (2009) Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proc Natl Acad Sci USA* 106:11837–11844.
46. Fersht AR (2002) On the simulation of protein folding by short time scale molecular dynamics and distributed computing. *Proc Natl Acad Sci USA* 99:14122–14125.
47. Bowman GR, Ensign DL, Pande VS (2010) Enhanced modeling via network theory: Adaptive sampling of Markov state models. *J Chem Theory Comput* 6:787–794.
48. Hinrichs NS, Pande VS (2007) Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics. *J Chem Phys* 126:244101.
49. Saven JG, Wang J, Wolynes PG (1994) Kinetics of protein folding—The dynamics of globally connected rough energy landscapes with biases. *J Chem Phys* 101:11037–11043.
50. Wang J, Saven JG, Wolynes PG (1996) Kinetics in a globally connected, correlated random energy model. *J Chem Phys* 105:11276–11284.
51. Ensign DL, Kasson PM, Pande VS (2007) Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *J Mol Biol* 374:806–816.
52. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES (1999) On the role of conformational geometry in protein folding. *J Chem Phys* 111:10375.
53. Andrec M, Felts AK, Gallicchio E, Levy RM (2005) Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc Natl Acad Sci USA* 102:6801–6806.
54. Kim PS, Baldwin RL (1990) Intermediates in the folding reactions of small proteins. *Annu Rev Biochem* 59:631–660.
55. Shan B, Eliezer D, Raleigh DP (2009) The unfolded state of the C-terminal domain of the ribosomal protein L9 contains both native and non-native structure. *Biochemistry* 48:4707–4719.
56. Kuzmenkina EV, Heyes CD, Nienhaus GU (2005) Single-molecule Forster resonance energy transfer study of protein dynamics under denaturing conditions. *Proc Natl Acad Sci USA* 102:15471–15476.
57. McLeish TC (2005) Protein folding in high-dimensional spaces: hypergutters and the role of nonnative interactions. *Biophys J* 88:172–183.