Probing the origins of two-state folding

Thomas J. Lane,¹ Christian R. Schwantes,¹ Kyle A. Beauchamp,² and Vijay S. Pande^{1,2,3}

¹*Department of Chemistry, Stanford University, Stanford, California 94305, USA*

²*Biophysics Program, Stanford University, Stanford, California 94305, USA*

³*Department of Computer Science, Stanford University, Stanford, California 94305, USA*

(Received 7 May 2013; accepted 13 September 2013; published online 11 October 2013)

Many protein systems fold in a two-state manner. Random models, however, rarely display two-state kinetics and thus such behavior should not be accepted as a default. While theories for the prevalence of two-state kinetics have been presented, none sufficiently explain the breadth of experimental observations. A model, making minimal assumptions, is introduced that suggests two-state behavior is likely for any system with an overwhelmingly populated native state. We show two-state folding is a natural consequence of such two-state thermodynamics, and is strengthened by increasing the population of the native state. Further, the model exhibits hub-like behavior, with slow interconversions between unfolded states. Despite this, the unfolded state equilibrates quickly relative to the folding time. This apparent paradox is readily understood through this model. Finally, our results compare favorably with measurements of folding rates as a function of chain length and K_{eq} , providing new insight into these relations. © 2013 AIP Publishing LLC. [<http://dx.doi.org/10.1063/1.4823502>]

I. INTRODUCTION

Most small (<100 residues), single domain proteins fold in a two-state manner.^{1–4} Specifically, protein systems appear to be thermodynamically two-state – they have only two equilibrium phases (folded and unfolded) – and also kinetically two-state, exhibiting single exponential kinetics. Protein domains that break this two-state paradigm are usually either large (>100 residues) and folding via one or more intermediates,⁵ or very small and extremely rapidly folding.^{6–17}

Simple two-state folding kinetics should be considered surprising. Protein chains have a large number of independent conformations available to them, and folding occurs via a stochastic interconversion between these conformations, often described as dynamics evolving on a “rough” potential energy function. These dynamics might be thought of as a network, where the nodes are conformations and connections are the rates of interconversion. It is then interesting to ask what the dynamics on a random network looks like,^{18,19} and how they compare to protein folding. In such random dynamical systems, two-state behavior is exceptional and rarely seen.^{20–22} Later in the paper, we will show why. The fact that two-state behavior is rare in random systems suggests that two-state kinetics cannot be accepted as a default.

Typically two-state kinetics in protein folding is rationalized in terms of a Kramer’s rate expression, which postulates the existence of a single dominant free energy barrier between folded and unfolded thermodynamic states. Postulating such a barrier – in effect, enforcing two-state kinetics – implies two thermodynamic phases. Here, we investigate the converse; we take two-state thermodynamics as a postulate and show, without ever invoking Kramer’s theory or an activated process, that we can expect such systems to typically be two-state when certain conditions are satisfied.²³

One oft-cited explanation is that two-state folding minimizes aggregation, and is therefore evolutionarily advantageous. This argument suggests that folding intermediates are more prone to aggregation than the unfolded or folded state, and therefore biology has attempted to minimize their population during folding.^{13,24,25} While intriguing and certainly possible, there is little direct evidence currently supporting this claim.

An alternative reason for the predominance of two-state folding suggests that most sequences capable of folding are intrinsically two-state. That is, the space of sequences that overwhelmingly populate a native state is enriched in systems that are thermodynamically and kinetically two- or few-state. Previous work has studied the connection between cooperativity and two-state folding using simulation,^{27–29} suggesting an inherent link between protein thermodynamics and kinetic complexity.

We call these competing theories the *aggregation* hypothesis, which dictates that two-state behavior is biologically necessary to avoid aggregation, and the *thermodynamic* hypothesis, dictating two-state folding is intimately linked to folding sequences whether those sequences were formed in the lab or in vivo.³⁰ Conclusive proof of either of these, or rejection of both in favor of an alternative, would have a major impact on the folding field. If two-state folding is necessary to avoid aggregation, this certainly has implications for understanding folding in vivo and developing therapeutics for Alzheimer’s and other aggregation-related diseases. If two-state folding is a consequence of optimizing sequence to yield a stable native state, then realizing that fact is part of a complete understanding of protein folding.

Here, we present an argument in favor of the thermodynamic hypothesis. We show that for a simple model of protein folding, if there is an overwhelmingly populated native state, two-state thermodynamics and single exponential kinetics are

FIG. 1. Illustration of three different kinetic spectra associated with three protein systems. Each blue horizontal line represents one system timescale (see Eq. (4)). The theory developed here predicts that as the folded state become more populated, the gap between the slowest and next-slowest timescales will grow. As this separation grows, so will the likelihood that a given experiment will see one dominant relaxation timescale, and classify the system as “two-state.” These spectra are idealized models that ignore the more complex situations of systems with more than two states. Note that these are *timescale* spectra, representing kinetic processes, and therefore are different from some classic work in folding on thermodynamic energy spectra.²⁶

extremely likely. This allows us to directly analyze the necessary and sufficient conditions for two-state behavior. Interestingly, we find that multi-exponential kinetics can be explained as highly perturbed or relatively unstable two-state systems (Fig. 1).

Moreover, the model provides an explanation how two-state systems might produce a “kinetic hub,”^{31–33} a dynamical system where most transitions between any two states pass through a third “hub” state.³⁴ Specifically, we introduce the concept of *slow interconversions with rapid equilibration*, which shows that a situation where the unfolded state equilibrates rapidly does not necessitate a situation where specific structures can reach each other quickly. This seeming paradox is resolved, allowing us to reconcile hub-like kinetics with two-state behavior. The presented model reproduces the bimodal mean first-passage time distribution from all-atom molecular dynamics simulations that originally inspired the hub hypothesis,³⁵ and still retains two-state folding.

Finally, and crucially, our model provides an explanation for why large proteins (> 100 residues) typically exhibit three- or more-state folding kinetics (Figs. 2 and 4). The model predicts that as protein systems get larger, additional timescales will be experimentally observable.

In the construction of the model, we postulate only that (A) protein dynamics can be represented by a master equation, (B) the system satisfies detailed balance, and (C) there exists one folded state that is highly populated compared to all other conformations. From these assumptions we build a model that quantitatively contains no additional information, or, equivalently, is the *most random*, using the maximum entropy formalism.

FIG. 2. Small proteins fold in a two-state manner, while larger proteins exhibit additional timescales. Shown are all proteins in the KineticDB,³⁶ which are classified as two-state (blue) or multi-state (green). Top panel shows folding times as a function of size (units of k_{obs} are s^{-1}), bottom panel shows a histogram of kinetic type as a function of number of residues. At ~ 100 residues there is a transition between two-state and multi-state behavior. (Insert) Two timescale spectra for the model discussed in the main text ($N = 100$ states, $\epsilon = 0.01$, left and $\epsilon = 0.05$, right) showing both two- and higher-state kinetics can be observed in the model. The model predicts more timescales will be observed experimentally as proteins get larger and the perturbation parameter ϵ increases.

The maximum entropy method is a natural choice for studying folding, because in some sense it reflects the process by which foldable sequences occur. One could conceive of evolution as a random search for a specific target – the target being a functional structure, and the search occurring through a huge number of possible sequence mutations. The one requirement for this search would be that the functional structure is overwhelmingly populated at equilibrium. Our model shows that *this thermodynamic requirement alone* is sufficient to explain two-state behavior.

II. THEORY OF DISCRETE TIME MASTER EQUATIONS

Our model is framed in the language of Markovian master equations, which give us a framework for implementing the maximum entropy model mentioned. We consider a discrete time propagator $T_{\Delta t}$ that describes the system dynamics.

Let $p(t)$ be a function describing the population density of the system at time t – we are interested in the time evolution of this function. Momentarily, we will postulate a partitioning of phase space into N discrete states. In this case, $p(t)$ is a vector, such that $p_i(t)$ is the population of a discrete state i at time t . Further, $T_{\Delta t}$ is a stochastic matrix, whose elements T_{ij} describe the probability for the probability density in state i to transfer to state j in the lag time Δt ,

$$p(n \cdot \Delta t) = p_0 T_{\Delta t}^n$$

from initial populations p_0 . Note that in what follows the specific lag time used will not be too important, so we will drop the Δt subscript and just write T .

The system dynamics can then be understood via the eigenmodes of the propagator. The system timescales τ_n are given by the eigenvalues λ_n of T ,

$$\tau_n = -\frac{\Delta t}{\log(\lambda_n)}, \quad (1)$$

while the corresponding eigenvectors describe the exchange of population between states on those timescales. The first eigenvalue is always unity ($\lambda_1 = 1$) corresponding to infinite time. Its corresponding eigenvector is the stationary distribution, denoted π .

Each row of T is a probability distribution, and therefore must be row-normalized and admits a measure of entropy, loosely speaking the information content of the probability distribution for each row. Further, the entire propagator has an associated entropy (sometimes called the “caliber”).^{37,38} Assuming that the distributions described by each row are independent (which is identical to the Markov assumption used to formulate the master equation), we can write the *propagator entropy* as

$$S_T = -\sum_{i,j} T_{ij} \log T_{ij}. \quad (2)$$

Maximizing this function, subject to some restraints describing known information, gives the model that makes the fewest assumptions about the system dynamics.

III. THE MAXIMUM ENTROPY FORMALISM ALLOWS THERMODYNAMIC POSTULATES TO RESULT IN KINETIC MODELS

Using the maximum entropy principle, let us build a model of protein folding. Our goal will be to use well-known facts about proteins as starting assumptions, but limit these as much as possible. Let us postulate:

- (A) Protein dynamics can be described as transitions between N distinct states that partition phase space, where each state is approximately a single conformation (we will call these “microstates” simply states, not to be confused with the thermodynamic folded/unfolded macrostates). We assume dynamics on this space are described by a Markovian propagator T . Finally, we expect the number of possible states to be large (famously estimated by Levinthal to be $\sim 3^{100}$), such that we will not be too hesitant in assuming N is big.

- (B) The system is ergodic and time-reversible, and therefore the detailed balance condition holds

$$\pi_i T_{ij} = \pi_j T_{ji}, \quad (3)$$

where π describes the stationary solution to T that is approached asymptotically in time.

- (C) Proteins have evolved in such a way that there exists a folded state F that has a much larger equilibrium population than all other states, i.e., $\pi_F \gg \pi_i$ for all i that are not F .

Solving for the transition matrix T that maximizes (2) subject to postulates (A, B, C) is a straightforward exercise in Lagrange multipliers. In what follows, we investigate analytical solutions for the specific case where there is one highly populated native state, and all unfolded states are equally populated at equilibrium. Mathematical detail has been relegated to the supplemental information; we present the results.

Our key result is a timescale spectrum (the eigenspectrum of T), which is simply a collection of the system timescales. For instance, in a re-folding experiment a series of exponential decays might be observed in, e.g., a trace of Trp fluorescence,

$$A(t) = \sum_i A_i e^{-t/\tau_i}, \quad (4)$$

where $A(t)$ is the observable trace, τ_i are the observed timescales.³⁹ Finally, the amplitudes $A_i = \langle p_0, \psi_i^L \rangle \langle \mathcal{O}, \psi_i^R \rangle$, with p_0 the initial populations of each state when the experiment begins, $\mathcal{O}$ a vector of the average observable value for each state, and ψ_i^L, ψ_i^R are the i th left and right eigenvectors of the propagator, respectively. This amplitude is the mathematical expression of familiar physical concepts. First, the initial populations of each state will affect what experimental response is observed; the factor $\langle p_0, \psi_i^L \rangle$ represents the extent to which a prepared sample (with populations p_0) will participate in the mode with timescale τ_i . Second, different experimental probes will report more strongly on certain states than others; the factor $\langle \mathcal{O}, \psi_i^R \rangle$ captures this uneven reporting.

In such an experiment, the “kinetic spectrum” is just the collection of τ_i (Fig. 1). There could be one observed exponential, two, or many, depending on the system and experiment. Which timescales are observed will depend on two factors: to be seen, the amplitude A_i must be large enough for a given experiment with limited sensitivity to observe it. Second τ_i must be in the appropriate range of the experiment’s temporal resolution. In this model we compare to experiment by computing the kinetic spectrum, $\{\tau_i\}$, which is invariant over different experimental probes and initial conditions; we do not compute the amplitudes $\{A_i\}$, which depend strongly on the experiment under consideration.

IV. A SINGLE LOW ENERGY STATE RESULTS IN A KINETICALLY AND THERMODYNAMICALLY TWO-STATE SYSTEM

We take $\pi_F \gg \pi_i$ and $\pi_i = \pi_j$, where i and j are unfolded states. Let us label all such states as U , and label the respective populations as π_F and π_U (π_U is the population of just one of $N - 1$ unfolded states). With this, it can be shown⁴⁰ that T

takes the form

$$T = \begin{pmatrix} T_{UU} & T_{UU} & \cdots & T_{UF} \\ T_{UU} & T_{UU} & & T_{UF} \\ \vdots & & \ddots & \vdots \\ T_{FU} & T_{FU} & \cdots & T_{FF} \end{pmatrix}$$

where T_{UU} is the probability of transition from a single unfolded state to any other, T_{UF} is the probability of going from an unfolded to a folded state, and T_{FU} is the probability of the converse. Further, if $\pi_F > \pi_U$, then $T_{FU} < T_{UU} < T_{UF}$; the differences between these get larger as the difference in population between F and U states increases.⁴⁰

This matrix is the foundation of our model. In Sec. V, we analyze models very near this maximum entropy solution, but where the symmetry in the transition probabilities is broken by a small perturbation; we begin by analyzing the current result, as it represents the simplest case.

We find three eigenvalues: one stationary (λ_1), one corresponding to the folding/unfolding reaction (λ_2), and one corresponding to unfolded state dynamics (λ_3 , with multiplicity $N - 2$),⁴⁰

$$\lambda_1 = 1,$$

$$\lambda_2 = 1 - T_{UF} - (N - 1)T_{FU},$$

$$\lambda_3 = 1 - T_{UF} - (N - 1)T_{UU},$$

showing that since $T_{FU} < T_{UU}$, $\lambda_1 > \lambda_2 > \lambda_3$. Further, for $\pi_F \gg \pi_U$, there will be a significant gap between λ_2 and λ_3 , leading to a *separation of timescales* consistent with a two-state picture, even though the model consists of $N \gg 2$ states.

The final step of our model involves solving an equation numerically, therefore we cannot write down a closed-form expression for the λ_2/λ_3 gap; it is possible, however, to plot this timescale gap as a function of K_{eq} or the number of configuration states N .⁴⁰ Figure 3 shows the scaling of the folding timescale $\tau_2 = -\Delta t / \log(\lambda_2)$ with K_{eq} , and compares this scaling with two experimental systems, lambda repressor¹¹ and BBL.⁴¹ Further, λ_3 is not a function of K_{eq} in this model. Thus, the rightmost panel of Fig. 3 demonstrates the theoretical scaling of the gap between the slowest (folding)

timescale and the next-slowest system timescale. A precise treatment of this second-slowest timescale is performed in Sec. VI.

It is critical to note that the discussion of modulating K_{eq} is restricted to changing the conditions for a *single protein sequence*. Our model contains no features that allow it to distinguish between sequences. Thus, we use the term “stability” to mean the relative population of the native state between under two different conditions, and not relative stability between mutants or different proteins (that might be compared via a ΔG of folding or melting temperature T_m).

Why is a gap between λ_2 and λ_3 a signature of two-state kinetics? In a system where the folding timescale is much slower than the rest, an experiment designed to study folding may be poorly suited (too low-resolution) to measure faster kinetics. Further, probes designed specifically to study folding may have a large amplitude response to folding kinetics, but not to other kinetic modes in the system. Mathematically, if $\langle \mathcal{O}, \psi_2^R \rangle$ is maximized (ψ_2^R is the eigenvector describing folding), then $\langle \mathcal{O}, \psi_n^R \rangle$ for $n \geq 3$ will be small, since the eigenvectors ψ_i^R are orthogonal. Therefore, experiments well-designed to study folding will measure other system modes at a much lower amplitude.

Recently, as higher resolution instrumentation has been developed,⁴³ faster timescales such as these have been found in folding systems that were previously considered two-state. Perhaps the best single example of this is the villin headpiece. In Ref. 44, the kinetics of villin were measured by NMR line-shape analysis and fit well to a two-state model with a folding time of order μs . Later, laser-induced temperature jump and ultra-fast triplet-triplet energy transfer experiments revealed additional dynamical processes, including intermediate formation at 70 ns⁴⁵ and native state locking/unlocking at 170 ns.⁴⁶ Thus, the villin timescale spectrum has one relatively slow mode, its folding time, and at least two faster modes. This is consistent with our model, which predicts proteins have one slow mode (λ_2) and a number of faster modes (λ_n , $3 \leq n \leq N$).

In systems where the separation between slow and fast modes is very large – larger than the separation for villin, for instance – it is likely that an experiment designed to measure the slow folding timescale will not measure faster timescales.

FIG. 3. A comparison of the theory and experiment: the change in observed folding timescale as a function of K_{eq} , modulated experimentally by temperature. (Left) A series of lambda repressors from Ref. 11. (Middle) BBL, from Ref. 41. (Right) This work (values of $K_{eq} < 1$ not plotted). Since τ_3 is not a function of K_{eq} , this plot also demonstrates the gap between τ_2 and τ_3 as a function of K_{eq} . (Insert) Theoretical dependence of folding times on chain length is exponential assuming $R \sim e^{\alpha N}$, which is consistent with experiment (Fig. 2), though is not the only consistent model.⁴²

FIG. 4. (Left) The calculated C_V vs β curve (for $N = 100$), showing a phase transition at the melting temperature $\beta = 1$. (Right) Experimental dependence of τ_3 with protein chain length R . Data from 12 multi-state proteins reported in the KineticDB.³⁶

Thus, only one kinetic timescale is ever seen experimentally, and we call such systems kinetically “two-state.”

We have observed the model is kinetically two-state, but it is also (by construction, due to the choice of π) thermodynamically two-state, as demonstrated by a sharply peaked C_V curve. Denote the free energy of each state by A_U for unfolded and A_F for the folded states. Then set our scale of energy such that $A_U = 0$. With this, we can write the ratio of state populations

$$\frac{\pi_F}{\pi_U} = \frac{e^{-\beta A_F}}{e^0} = e^{-\beta A_F}$$

now, scale our units of temperature such that the folding temperature, where $\pi_F = 0.5$, occurs at $\beta = 1$. Then it is clear from the above that $A_F = -\log(N - 1)$. The partition function is

$$Z = \sum_i e^{-\beta A_i} = (N - 1) + (N - 1)^\beta$$

resulting in the heat capacity

$$C_V = \beta^2 (N - 1)^{\beta+1} \left[\frac{\log(N - 1)}{(N - 1) + (N - 1)^\beta} \right]^2$$

which exhibits a first order phase transition at $\beta = 1$, consistent with what is observed in experiment (Fig. 4).

V. EFFECTS OF PERTURBATION OF THE MODEL

While this simple model demonstrates a minimal set of sufficient requirements for two-state folding, it retains an artificial symmetry – all the rates in each set $\{T_{UU}\}$, $\{T_{UF}\}$, $\{T_{FU}\}$ are identical. Such symmetry can be broken by adding random “noise” to the transition matrix elements. Robust two-state folding should not be affected by such a perturbation – experimentally, a single mutation or slight change in experimental conditions is insufficient to disrupt two-state behavior in the majority of systems.

A reasonable perturbation is the addition of a random Gaussian to each element of T , in the form of a matrix T' whose elements are derived from Gaussians

$$\tilde{T} = T + \epsilon T',$$

where primes to denote perturbing terms, tildes the resulting perturbed solution, and ϵ is a control parameter denoting the size of the perturbation.⁴⁷ In the supplementary material,⁴⁰ we

FIG. 5. The model’s eigenspectra ($\{\lambda_i\}$) under no perturbation (left), a perturbation (middle left), and a larger perturbation (middle right). These plots are for a single instance of the distribution indicated in the main paper, with ϵ set as indicated. The far right is the spectrum of one sample from the GOE (red line is origin). In the absence of any native bias ($\pi_F \approx \pi_U$), the kinetic spectrum would be qualitatively similar to just the GOE (with all of the eigenvalues positive). Parameters were $N = 100$ and $K_{eq} = 1.0$.

show that one can construct such a perturbation while ensuring $\tilde{T}$ maintains detailed balance and is a stochastic matrix, in the thermodynamic limit ($N \rightarrow \infty$). The perturbation derives from a symmetric matrix whose elements are drawn independently from a Gaussian distribution – such a matrix is known as a member of the Gaussian orthogonal ensemble (GOE).⁴⁸

From this, we find⁴⁰ the probability density of obtaining a spectrum $P(\{\tilde{\lambda}_n\})$ under a random perturbation is

$$P(\{\tilde{\lambda}_n\}) = C \prod_{n=3}^N e^{-\tilde{\lambda}_n^2 / 2\epsilon^2} \sum_{3 \leq i}^N \sum_{j > i}^N |\tilde{\lambda}_i - \tilde{\lambda}_j|,$$

where we have considered only the eigenvalues representing dynamics in the unfolded state (λ_3). C is simply a constant that normalizes the distribution.

This perturbation has the effect of splitting the degeneracy of and spreading out the previously overlapping eigenvalues (Fig. 5). One can see this by noticing that the terms of the form $|\tilde{\lambda}_i - \tilde{\lambda}_j|$ require the probability density go to zero as two eigenvalues get close together. This effect is known as *level repulsion* in random matrix theory.

What consequences does this perturbation have for protein folding? One can see that for small perturbations, the symmetry of the original degenerate model is broken, but a spectral gap between the unfolded timescale and the folding timescale (λ_2) still exists (Fig. 5).^{49,50} Once perturbations get very large, however, one expects that the timescales of unfolded state dynamics will spread sufficiently to be comparable to the folding timescale. This will destroy the two-state features of the model, and shows that, as stated in the Introduction, highly random models will not exhibit two-state

behavior. In Sec. VI, we demonstrate why we might expect two-state systems to be fairly robust to such perturbations.

VI. THREE OR GREATER STATE BEHAVIOR MAY BE OBSERVED, ESPECIALLY IN LARGE SYSTEMS

Random matrix theory also provides an estimate for the relative timescale of the slowest non-folding process,⁴⁰

$$\frac{\tau_3}{\Delta t} = - \left[\frac{\alpha}{2} R + \log \epsilon \sqrt{2} \right]^{-1}$$

as a function of the size of the random perturbation ϵ and chain length R , assuming the exponential scaling $N \sim e^{\alpha R}$.⁵¹ The specific model of rate-length scaling is not critical to understand so long as there is a rate-length relationship,⁴² the model presented here predicts that larger proteins will exhibit more measurable timescales. While the expression for τ_3 contains too many unknown parameters (ϵ , α , Δt) to make meaningful quantitative predictions of experiment, it does suggest the qualitative feature that the slowest non-folding timescales in proteins should increase with chain length. This is consistent with what is seen experimentally for the 12 multi-state proteins for which the slowest non-folding timescale is reported in the literature (Fig. 4)³⁶ – additional data will be necessary to definitively confirm this prediction.

As discussed previously, the timescale gap $\tau_2 - \tau_3$ dictates whether or not we classify a system as “two-state.” In our model, the perturbation parameter ϵ enters into this gap in the expression for $\tau_3(\epsilon)$. Thus by computing the *condition number* $\kappa = |x f'(x)/f(x)|$ we may evaluate the robustness of the model’s two-state behavior with respect to perturbation parameter ϵ . For the condition number of $\tau_3(\epsilon)$, we obtain

$$\kappa(\epsilon) = \left| \frac{\alpha}{2} R + \log \epsilon \sqrt{2} \right|^{-1}$$

which is comparable to the condition number of the natural logarithm, $\kappa(x) = \log(x)^{-1}$. We conclude that the qualitative feature of the timescale gap leading to two-state behavior can be thought of as robust, in the very specific sense that a perturbation of order ϵ will affect the qualitative structure of the model only logarithmically. This holds so long as $\epsilon \ll e^{-\alpha R/2}$, where $\kappa(\epsilon)$ exhibits a singularity. In this special regime, the eigenstructure of the propagator T becomes unstable. Thus, as proteins near the edge of non-two state behavior, we predict that their dynamics will become very sensitive to additional perturbation, and that the onset of this behavior will happen more quickly for larger proteins (bigger R).

Unfortunately, in the context of the present model, we cannot provide an absolute estimate of the absolute magnitude of ϵ upon, e.g., a protein mutation – the present model has no absolute energy or timescale, and therefore without building in these aspects through additional postulates or assumptions, we cannot directly evaluate ϵ . We note this may be a fruitful area for further work.

VII. THE MODEL EXHIBITS NATIVE HUB-LIKE BEHAVIOR

The mean first passage time (MFPT) is the expected time it takes for a walker starting at state i of a Markov

FIG. 6. MFPT distribution (normalized) for an unperturbed (black, δ -functions), slightly perturbed (blue, $\epsilon = 10^{-5}$) and significantly perturbed (red, $\epsilon = 10^{-4}$) models. The distribution is bimodal, or “hub-like.” Shown are 1000 numerical samples for $N = 100$ and $K_{eq} = 1.0$. There are some large MFPTs representing $F \rightarrow U$ transitions that are not shown here.

chain to reach state j for the first time. It is apparent from inspection that, due to the fact that $T_{UU} < T_{UF}$, we expect the MFPT from an unfolded state to any other unfolded state to be *slower* than the passage time from that state to the native state. A plot of the distributions of MFPTs from every state to every other is therefore bimodal (Fig. 6), a property that has been described as “hub-like”³⁵ after it was witnessed in all-atom molecular dynamics simulations^{31,52–54} (recently more sensitive measures of hub-like phenomenology have been proposed and employed a model very similar to this one^{34,55}). Numerical simulations show a small perturbation spreads out such a distribution, but does not destroy the bimodality (Fig. 6). The system only exhibits these hub-like behaviors when there is a single native state, $\pi_F \gg \{\pi_U\}$.

Slow MFPTs between unfolded states might seem inconsistent with the fact that λ_3 , the eigenvalues corresponding to the unfolded state dynamics, represent timescales that are much faster than the folding time. This “paradox” arises because the eigenvalue (λ_3) is a measure of the *ensemble* dynamics of the system, while MFPTs measure dynamics at the level of a single protein visiting specific states. Imagine the following two cases. If we monitor a single protein molecule as it folds, it will visit many unfolded states before folding. However, the chances that it reaches *one particular* unfolded state before folding is very small – it is much more likely to visit the native state before this single unfolded state. Quantitatively, this results in a smaller MFPT for $U \rightarrow F$ transitions than $U \rightarrow U$ transitions (Fig. 6).

Next, consider an ensemble of proteins. Because they each visit a large number of unfolded states before reaching the single native state, they are able to spread out quickly, equilibrating all of the unfolded states, before these unfolded states have a chance to equilibrate with the native state. This results in a system that exhibits *slow interconversions with rapid equilibration*.

This phenomenon is purely a result of dividing the unfolded state into many parts – in effect, increasing the resolution of non-native dynamics. Our model, being phenomenological, cannot address whether or not significant energetic or

enthalpic barriers exist in the non-native ensemble. This is an important outstanding question that will require further work to address. It is important to note that, at high spatial resolution, hub-like kinetics might be present purely due to the size of the unfolded state space, regardless of if such barriers exist or not. A corollary of this is that at low resolution, this hub-like behavior will disappear – in the limit of two states, it is by definition impossible to have any kind of network hub.

This interpretation is consistent not only with traditional views of rapid unfolded-state equilibration, but also recent reports of relatively slow interconversions between non-native conformations.^{32,35,53,56} This model provides a lens for reconciling these views.

VIII. CONCLUSIONS

What are the minimal sufficient features for a folding sequence to exhibit two-state kinetics and thermodynamics? Is two-state folding biologically advantageous, or a corollary to a highly populated native state? The model presented here suggests that simply a large enough energy gap between folded and unfolded states is enough to result in two-state behavior.

Further, this model

- Explains why two-state systems are common in small proteins, but additional timescales appear in larger proteins. The model attributes this to the nature of protein thermodynamics, and does not invoke an aggregation-based evolutionary hypothesis.
- Shows why additional fast timescales, usually not directly involved in folding, can be observed in traditionally two-state systems such as villin.
- Displays two-state kinetics without making reference to or assuming an activated process. Agnosticism in this regard allows us to analyze non-native dynamics in ways that models that begin with an activated process cannot.
- Reconciles the hub-like kinetics observed in simulation and with two-state kinetics, through the concept of rapid equilibration with slow interconversions.
- Proposes a new interpretation of multi-exponential kinetics in fast folding proteins, and this mechanism is seen to be in agreement with reported experiments. The model makes clear predictions about how folding times change with respect to K_{eq} .
- Introduced an approach based on maximizing the entropy of a dynamical propagator given known information as a way of probing protein folding theoretically. More sophisticated models, that include more detailed structural information and precise state energy structures, may yield additional experimental predictions.

Each of these items provides either new insight into empirical observations made in experiments or simulations that were previously poorly understood, or pushes the methods employed currently in the construction of analytical theories of protein folding.

The results presented in this paper suggest an intriguing connection to previous work suggesting that a native en-

ergy gap is a necessary requirement for sequence foldability.²⁶ That theory would dictate that evolved sequences would exhibit such a gap; our theory further suggests such sequences are also likely to exhibit two-state kinetics if they are short ($N < 100$), or few-state kinetics if longer. This picture appears consistent with experimental observations of the folding of single domain globular proteins thus far.

Our model brings into focus two significant research questions that remain unresolved. First, this model says little about what the topologies of realistic propagators of protein dynamics look like, and to what extent those topologies dictate protein dynamics. Second, we have so far been unable to address the nature of dynamics in the unfolded state. Whether or not these non-native dynamics are restricted by significant barriers remains an open question, one that requires a microscopic theory (as opposed to the phenomenological theory presented here) validated careful experimentation and simulation to fully understand.

ACKNOWLEDGMENTS

T.J.L. would like to acknowledge a conversation with Attila Szabo on the reconciliation of hub-like and two-state kinetics explained here. This paper has benefited greatly from discussions with Max Prigozhin, Hannah Gelman, Alex Grosberg, Geoff Rollins, and Robert McGibbon. T.J.L. was supported by the NSF GRF. Further support was provided by the NSF (NSF-MCB-0954714) and NIH (R01-GM062868).

- ¹S. Jackson and A. R. Fersht, *Biochemistry* **30**, 10428 (1991).
- ²D. Perl, C. Welker, T. Schindler, K. Schröder, M. A. Marahiel, R. Jaenicke, and F. X. Schmid, *Nat. Struct. Biol.* **5**, 229 (1998).
- ³E. Rhoades, M. Cohen, B. Schuler, and G. Haran, *J. Am. Chem. Soc.* **126**, 14686 (2004).
- ⁴D. Barrick, *Phys. Biol.* **6**, 015001 (2009).
- ⁵A. I. Bartlett and S. E. Radford, *Nat. Struct. Mol. Biol.* **16**, 582 (2009).
- ⁶J. D. Bryngelson, J. Onuchic, N. D. Socci, and P. Wolynes, *Proteins* **21**, 167 (1995).
- ⁷W. Eaton, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 5897 (1999).
- ⁸J. Sabelko, J. Ervin, and M. Gruebele, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 6031 (1999).
- ⁹M. M. Garcia-Mira, M. Sadqi, N. Fischer, J. M. Sanchez-Ruiz, and V. Muñoz, *Science* **298**, 2191 (2002).
- ¹⁰W. Y. Yang and M. Gruebele, *Nature (London)* **423**, 193 (2003).
- ¹¹W. Y. Yang and M. Gruebele, *Biochemistry* **43**, 13018 (2004).
- ¹²W. Y. Yang and M. Gruebele, *Biophys. J.* **87**, 596 (2004).
- ¹³M. Gruebele, *C. R. Biol.* **328**, 701 (2005).
- ¹⁴H. Nguyen, M. Jager, J. W. Kelly, and M. Gruebele, *J. Phys. Chem. B* **109**, 15182 (2005).
- ¹⁵M. Sadqi, D. Fushman, and V. Muñoz, *Nature (London)* **442**, 317 (2006).
- ¹⁶F. Liu and M. Gruebele, *J. Mol. Biol.* **370**, 574 (2007).
- ¹⁷F. Liu, D. Du, A. A. Fuller, J. E. Davoren, P. Wipf, J. W. Kelly, and M. Gruebele, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2369 (2008).
- ¹⁸R. Albert and A. Barabasi, *Rev. Mod. Phys.* **74**, 47 (2002).
- ¹⁹M. E. J. Newman, *Comput. Phys. Commun.* **147**, 40 (2002).
- ²⁰Complexity theory has not traditionally focused on this topic, but the tools are there for its investigation. See, e.g., **21** and **22**.
- ²¹A. N. Samukhin, S. N. Dorogovtsev, and J. F. F. Mendes, *Phys. Rev. E* **77**, 036115 (2008).
- ²²F. Chung, L. Lü, and V. Vu, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 6313 (2003).
- ²³It should be noted that we do not wish to show this converse holds in general. We suspect that there exist systems that are thermodynamically two-state, but under certain conditions will display multi-exponential kinetics.
- ²⁴T. Schindler, M. Herrler, M. A. Marahiel, and F. X. Schmid, *Nat. Struct. Biol.* **2**, 663 (1995).
- ²⁵M. Jacob, T. Schindler, J. Balbach, and F. X. Schmid, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 5622 (1997).

- ²⁶A. Sali, E. Shakhnovich, and M. Karplus, *Nature (London)* **369**, 248 (1994).
- ²⁷H. Kaya and H. S. Chan, *Proteins* **52**, 510 (2003).
- ²⁸H. Kaya and H. S. Chan, *J. Mol. Biol.* **315**, 899 (2002).
- ²⁹P. F. N. Faisca and K. W. Plaxco, *Protein Sci.* **15**, 1608 (2006).
- ³⁰One additional possibility has been suggested by Robert McGibbon: it could be that two-state systems are considered “typical” by the folding community, and non-two state systems are thought of as pathological and therefore not studied. For the purposes of this paper we ignore this possibility, but it is important that those working in the field keep it in mind.
- ³¹G. R. Bowman, V. A. Voelz, and V. S. Pande, *J. Am. Chem. Soc.* **133**, 664 (2011).
- ³²V. S. Pande, *Phys. Rev. Lett.* **105**, 198101 (2010).
- ³³T. J. Lane and V. S. Pande, *J. Phys. Chem. B* **116**, 6764 (2012).
- ³⁴A. Dickson and C. L. Brooks, *J. Chem. Theory Comput.* **8**, 3044–3052 (2012).
- ³⁵G. R. Bowman and V. S. Pande, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 10890 (2010).
- ³⁶N. S. Bogatyreva, A. A. Osypov, and D. N. Ivankov, *Nucleic Acids Res.* **37**, D342 (2009).
- ³⁷E. T. Jaynes, *Annu. Rev. Phys. Chem.* **31**, 579 (1980).
- ³⁸G. Stock, K. Ghosh, and K. A. Dill, *J. Chem. Phys.* **128**, 194102 (2008).
- ³⁹J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, *J. Chem. Phys.* **134**, 174105 (2011).
- ⁴⁰See supplementary material at <http://dx.doi.org/10.1063/1.4823502> for mathematical detail: Lagrange multiplier-based solution of the maximum entropy propagator T , analysis of the eigenvalues and eigenvectors of that propagator, perturbation-theoretic approach to the stability of that eigenstructure, calculation of the timescale gap between λ_2 and λ_3 .
- ⁴¹P. Li, F. Y. Oliva, A. N. Naganathan, and V. Muñoz, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 103 (2008).
- ⁴²T. J. Lane and V. S. Pande, preprint [arXiv:1301.4302v1](https://arxiv.org/abs/1301.4302v1) (2013).
- ⁴³M. B. Prigozhin and M. Gruebele, *Phys. Chem. Chem. Phys.* **15**, 3372–3388 (2013).
- ⁴⁴M. Wang, Y. Tang, S. Sato, L. Vugmeyster, C. J. McKnight, and D. P. Raleigh, *J. Am. Chem. Soc.* **125**, 6032 (2003).
- ⁴⁵J. Kubelka, T. K. Chiu, D. R. Davies, W. A. Eaton, and J. Hofrichter, *J. Mol. Biol.* **359**, 546 (2006).
- ⁴⁶A. Reiner, P. Henklein, and T. Kiefhaber, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 4955 (2010).
- ⁴⁷We have only carried out the perturbation to first order. The perturbations we want to consider are small, since large perturbations stray too far from the entropy maximum solution and therefore are expected to imply work done by biology.
- ⁴⁸M. L. Mehta, *Random Matrices* (Academic Press, 2004).
- ⁴⁹Because of the spreading of the eigenvalues, this gap will be smaller than before. One could calculate the probabilistic difference from the Tracy-Widom distribution,⁵⁰ but we have not undertaken this task (yet).
- ⁵⁰C. A. Tracy and H. Widom, *Commun. Math. Phys.* **177**, 727 (1996).
- ⁵¹Note that this expression requires $\epsilon \sim \mathcal{O}(1/N)$, which for large state number N makes ϵ very small.
- ⁵²F. Rao and A. Caflisch, *J. Mol. Biol.* **342**, 299 (2004).
- ⁵³T. J. Lane, G. R. Bowman, K. Beauchamp, V. A. Voelz, and V. S. Pande, *J. Am. Chem. Soc.* **133**, 18413 (2011).
- ⁵⁴R. M. Levy, W. Dai, N.-j. Deng, and D. E. Makarov, “How long does it take to equilibrate the unfolded state of a protein?,” *Protein Sci.* (published online).
- ⁵⁵A. Dickson and C. L. Brooks III, *J. Am. Chem. Soc.* **135**, 4729–4734 (2013).
- ⁵⁶S. Waldauer, O. Bakajin, and L. Lapidus, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 13713 (2010).