THE JOURNAL OF
PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

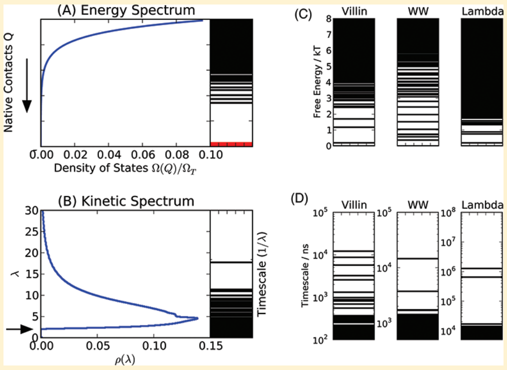
A Simple Model Predicts Experimental Folding Rates and a Hub-Like Topology

Thomas J. Lane and Vijay S. Pande*

Department of Chemistry, Stanford University, Stanford, California 94305, United States

Supporting Information

ABSTRACT: A simple model is presented that describes general features of protein folding, in good agreement with experimental results and detailed all-atom simulations. Starting from microscopic physics, and with no free parameters, this model predicts that protein folding occurs remarkably quickly because native-like states are kinetic hubs. A hub-like network arises naturally out of microscopic physical concerns, specifically the kinetic longevity of native contacts during a search of globular conformations. The model predicts folding times scaling as $\tau_f \sim e^{\xi N}$ in the number of residues, but because the model shows ξ is small, the folding times are much faster than Levinthal's approximation. Importantly, the folding time scale is found to be small due to the topology and structure of the network. We show explicitly how our model agrees with generic experimental features of the folding process, including the scaling of τ_f with N, two-state thermodynamics, a sharp peak in C_V, and native-state fluctuations.



INTRODUCTION

Understanding how proteins fold remains one of the great outstanding questions of biophysics. Since Anfinsen demonstrated that proteins adopt one unique structure with overwhelming probability, physical insight into this last step of the central dogma has remained elusive.[1] An adequate physical picture of protein folding must resolve Levinthal's paradox: were folding a random search in conformation space, it would take an astronomical time for proteins to fold.[2,3] We know biology must employ physics to reduce the folding time to biologically relevant time scales—the question is, how?

This intriguing question has led to many theories attempting to explain folding.[4] The "classical view" was that proteins likely follow a specific stepwise path, progressing from extended, unfolded structures to the native state, gaining native structure along the way.[5,6] In the late 1980s and early 1990s, this view of folding was called into question by computational models, primarily lattice models, that demonstrated that folding could occur via many heterogeneous pathways over a single free energy barrier.[5−9] This was consistent with two key experimental observations that came to light at the same time. First, many single domain proteins fold in a two-state manner, in a single cooperative step with first-order kinetics.[7,10] Second, high resolution experimental techniques, especially NMR-based hydrogen exchange, showed many time scales during folding, indicative of heterogeneity.[5,11]

These realizations lead to the "new view" of protein folding, which dictated that proteins could fold via many independent, parallel pathways that might be highly heterogeneous.[5,6,12] This theory was later expanded by directly considering the free energy landscapes of proteins. Specifically, it was realized that parallel paths leading to quick folding must result from a significant energy gradient biased toward a single, dominant global minimum, the native state. Further, the free energy landscape must be minimally frustrated, that is, as smooth and "funnel-like" as possible.[8,9,13−16] Any bumps or dead-ends would slow folding, so this theory dictated that biology must have evolved sequences that eliminated these barriers. Amino acid sequences would be designed such that contacts in the native conformation would be energetically favorable and non-native contacts would interact weakly, or even be unfavorable, removing kinetics traps in the landscape.[9]

While energy landscape theories have been successful in explaining some features of folding, they rely on assuming the structure of a landscape, rather than describing its physical origin. Here, we present a new model that connects kinetics directly to microscopic, residue-level physics, that allows us to expand the conclusions that can be reached from simple models of folding. Rather than beginning with thermodynamical considerations and attempting to derive kinetics from an energy landscape, we begin by considering kinetics. This has a number of advantages. Primarily, we need not find a reaction coordinate to describe the kinetics of our model, which not only simplifies modeling but prevents errors that might occur by projecting onto poor reaction coordinates.[17,18] Furthermore, our kinetic model allows for direct comparison to experiment

Special Issue: Harold A. Scheraga Festschrift

Received: December 21, 2011
Revised: March 20, 2012
Published: March 27, 2012

and simulation, allowing for theoretical verification and falsifiable predictions.

Inspired by recent results from all-atom molecular simulation, we have adopted the view that protein kinetics can be concisely represented as a set of states and the rates of exchange between those states. Known as master equation models, or more recently Markov state models (MSMs),[19−21] these models consist of a state space and transition matrix characterizing the rates of exchange between each state. This is equivalent to a graphical model where nodes represent protein conformations, and edges represent rates. The representation is powerful enough to describe complex phenomena while simple enough for mathematical investigation. Further, the concept of metastable states and rates is also familiar to both physicists and chemists, and therefore provides an appealing foundation for understanding protein folding.

All-atom simulations have recently provided an empirical view of how proteins could fold. Computer technology and analysis methodology have progressed to the stage where, to date, the folding kinetics of many proteins have been described in atomic detail, with sizes up to 86 residues and folding times of ∼10 ms.[22−29] First realized by Rao and Caflisch,[30] and later expanded upon by Bowman and Pande,[31] a key generality of MSMs parametrized from all-atom simulations is the hub-like nature of these models. That is, when representing protein kinetics as a graph, the native state is highly connected and central in the network, and the connectivity of states increases with native content.[24,27] This means that, from any non-native state, there is a direct route to the native state involving few "hops" between nodes. This suggested that any arbitrary structure might be able fold in a small, finite time, since it would be close to the native state in a kinetic sense based on the network topology alone.

After this development, Pande derived a model showing how, if non-native interactions in proteins were favorable, a hub-like structure could emerge in folding kinetics.[32] Further, such a hub did not result if non-native contacts were unfavorable. In that work, all non-native states were equivalent, and thus the model was limited in scope. Here, we build on that work. We adopt the view of microscopic kinetics derived previously but expand the state space to include a diverse set of states with various degrees of native content. The resulting model mirrors all-atom MD simulation, displaying a topology with strong hub-like features around the native state.

Further, we show how a hub-like topology provides an explanation for why proteins fold quickly. The model reproduces essential experimental observations, such as the scaling of the folding rate with chain length, the correct energy-gap structure predicted by previous theory and demonstrated by experiment, and a peak in the heat capacity at the melting temperature. Our simple model is not the only attempt to explain folding rates,[8,12,16,33−48] nor are we the only ones to have described protein folding as a hub-like network.[30,39,49] However, our model is new in its ability to predict protein folding times without fit parameters.

We begin by outlining the theory. This first section should provide a nonmathematical introduction for a general audience, and provide a basic understanding of the model that will be useful context for the rest of the manuscript. The reader uninterested in mathematical details should be able to read this section and skip to the discussion. Following this, we begin constructing the model by recalling the results and notation of previous work. Then, we define a state space describing protein conformations, and calculate some interesting thermodynamic properties of that space. Next, we derive the kinetics of proteins from the state space, with an emphasis on the folding process. Finally, we discuss the implications of our model, and draw comparisons to experiment and simulation. Some mathematical detail has been relegated to the Supporting Information in the interest of conciseness and readability.

■ MODEL SUMMARY

In what follows, we consider protein kinetics as a set of states and the rates of exchange between those states, and investigate the dynamics that arise from these considerations. From this kind of model, many interesting new insights emerge.

To make this approach tractable, we make a number of physical approximations, enumerated here.

(1) In light of the mounting evidence from experiment,[50−54] simulation (see the Supporting Information, Figure S1),[55,56] and theory[57] that, in the absence of denaturant, unfolded proteins are globular, we consider only protein conformations that are globular. While extended conformations of proteins are certainly interesting, here we restrict ourselves to single-domain proteins in the globular phase.

(2) Protein conformations can be accurately represented as contact maps.

(3) The energetics of this map can be captured by only two kinds of interactions, native and non-native contacts. Native contacts are by definition more energetically favorable than non-native contacts.

(4) Each contact map represents a metastable state; that is, each map is energetically stable enough such that transitions between individual maps will be much faster than times spent in an arbitrary contact geometry. This makes the dynamics approximately Markovian, since degrees of freedom orthogonal to the residue−residue contacts will rapidly equilibrate.

(5) Finally, states are either "kinetically connected" and can interconvert at some rate that will be the same for all pairs of connected states, or cannot directly interconvert.

In what follows we also make a few mathematical approximations, but will be clear to note where these are introduced.

From this simplified start, we can deduce a number of interesting things. First, since all contacts can be effectively sorted into native and non-native types, the fact that *both are attractive follows from the globular assumption*. To overcome the entropy of extended conformations, polymer theory leads us to conclude that, if the non-native states of proteins are in fact globular (in the absence of denaturant), non-native contacts must be favored with respect to solvent contacts. Moreover, the contact energies used here are derived through a consensus of experiment, simulation, and previous theory (see the Supporting Information).

On the basis of these energetics, we show that the heat capacity exhibits a first-order transition between non-native and native states at one unique temperature, T_m, which is in the appropriate range of protein melting temperatures.

Next, we calculate the kinetics of the model. A natural consequence of the model is that native states have many more kinetic neighbors than non-native states, resulting naturally in a *kinetic hub*. Further, this hub has a remarkable structure, that of a scale-free network, where the number of connections of each state follows a power law.

The scale-free structure of the network determines certain dynamical properties of the entire system. It is possible to show that the longest time scale in the system must be fast—much faster than a random search through conformations. This folding rate scales exponentially in chain length, $\tau_f \sim e^{\xi N}$, but nonetheless, folding occurs remarkably quickly because ξ is small. This ensures that the rate of folding is much faster than the estimate of Levinthal, and is in good agreement with experimental data.

A simple picture emerges from this model, where proteins fold by searching through globular conformations for their native state. This search is largely random but is aided by the fact that native contacts are more favorable than non-native contacts, a fact that has been predicted to lead to fast folding.[36,37] This means that native structure is *statistically persistent*—native contacts can break during the search but are more likely to be retained than non-native contacts. As long as native contacts are strong enough, this alone is enough to ensure that proteins fold quickly in a cooperative manner.

■ FORMALISM

Here, we briefly recall the notation and key results of previous work,[32] providing a pedagogical introduction to the formalism used. We represent a protein state as a contact map $C_{ij} \in \{0,1\}$, a binary value for each residue pair $\{ij\}$ indicating if residues i and j are in contact or not. Each contact, when formed, contributes some favorable (negative) energy. One of the key aspects of this model is that non-native contacts have a favorable energy contribution (denoted ε_{NN}), while native contacts (ε_N) are more favorable. This is captured by the microscopic Hamiltonian, for some arbitrary state α,

$$\mathcal{H}_\alpha = \varepsilon_N \sum_{ij} C_{ij}^\alpha C_{ij}^N + \varepsilon_{NN} \sum_{ij} C_{ij}^\alpha (1 - C_{ij}^N)$$

where the superscripts of C_{ij} represent either state α or the native state, N.

Here, we have written ε to represent an average potential of interaction. Consider the physical reaction of taking a residue from being solvent exposed (denoted R(aq)) to in contact with another residue (R_2(protein)), and its associated free energy change ΔF,

$$2\mathrm{R(aq)} \rightarrow \mathrm{R_2(protein)}: \Delta F = \phi + kT\sigma$$

where ϕ captures the free energy change upon contact formation and σ represents the entropy cost of loop closure. To be clear, ϕ represents enthalpic and entropic effects intrinsic to the contact, where $\sigma = \Delta S_{\mathrm{loop}}/k > 0$ accounts only for the entropy gained by loop formation, capturing the configurational entropy of the protein and associated solvent. σ is taken to be constant for all residues, consistent with previous work,[32,58] and is estimated in what follows from precise caliometric experiments ($\sigma \approx 4.0 \pm 3.0$ cal mol^{-1} K^{-1}).[59] This simplifies the reaction taking a non-native contact to a native contact

$$\mathrm{R_2(non\text{-}native)} \rightarrow \mathrm{R_2(native)}: \Delta F = \phi_N - \phi_{NN}$$

where the loop entropy terms cancel. These equations assume two-body interactions are dominant, i.e., the residue-pair formation is approximately independent of other residues. While this will introduce some error, since, e.g., contacts among adjacent residues will be loosely correlated, we can justify this approximation by noting that three-body interactions will still be much less important than three-body interactions.

To a first-order approximation, we consider the effective potential of this process to be independent of temperature. Then, for native and non-native contacts,

$$\varepsilon_N = \phi_N/kT \quad \text{and} \quad \varepsilon_{NN} = \phi_{NN}/kT$$

where ε represents a unitless interaction energy. Additionally, it will be convenient to specify the "excess" energy of a native contact, i.e., $\varepsilon_x = \varepsilon_N - \varepsilon_{NN}$ and $\phi_x = \phi_N - \phi_{NN}$.

This microscopic Hamiltonian allows us to not only define states but rates of transition between those states. We assume that the transition from one state α to some other state β must pass through a transition state that breaks all contacts not shared by these two structures, i.e., $C_{ij}^\ddagger = C_{ij}^\alpha C_{ij}^\beta$.[32] The free energy of the transition state allows one to compute the rate for $\alpha \rightarrow \beta$ using Kramer's approximation[32]

$$k_{\alpha\beta} = \tilde{k} e^{-\Delta F_{\alpha\beta}^\ddagger/kT}$$

where $\tilde{k}$ is the microscopic rate of interconversion. Now, equipped with a way to define states and derive the rates connecting them, we have a sufficient formalism to describe protein kinetics.

Let us calculate the free energies (F) of the reactant and transition states, and the associated barrier for the reaction $\alpha \rightarrow \beta$

$$\text{State } \alpha: \quad F_\alpha/kT = \varepsilon_x q^{\alpha,N} + \varepsilon_{NN} q^{N,N}$$

$$\text{TS } (\ddagger): \quad F_{\alpha\beta}^\ddagger/kT = \varepsilon_x q^{\alpha,\beta,N} + \varepsilon_{NN} q^{\alpha,\beta} - \sigma(q^{N,N} - q^{\alpha,\beta})$$

$$\Delta F_{\alpha\beta}^\ddagger = F_{\alpha\beta}^\ddagger - F_\alpha$$

$$\Delta F_{\alpha\beta}^\ddagger/kT = \varepsilon_x(q^{\alpha,\beta,N} - q^{\alpha,N}) + \varepsilon_{NN} (q^{\alpha,\beta} - q^{N,N}) - \sigma(q^{N,N} - q^{\alpha,\beta})$$

where $q^{\alpha,\beta}$ is the number of contacts in common between α and β and other q variables with superscripts represent analogous values (with N always denoting the native contact map). Also note that the free energy for β would be exactly analogous to the energy for α but does not explicitly enter Kramer's approximation. Further, one can see that the only place the loop entropy enters into our picture is in the transition state. Physically, this represents compact structures interconverting between each other via states with short free loops.

From this model, it is possible to intuitively see how hub-like kinetics could arise. To transition between two states, one must break all the contacts those states do not have in common. Since native contacts have a more favorable energy, breaking native contacts raises the energy of the transition state and slows the reaction, compared to breaking non-native contacts. However, if the product structure has a high number of native contacts, then there is a good chance of retaining those contacts in the transition state, lowering the barrier, and speeding conversion to that native-like state. At the same time, the favorable nature of non-native contacts prevents rapid rearrangement of non-native structure. This leads to hub-like characteristics, where the interconversion between arbitrary non-native states is intrinsically slow. This process favors kinetic connections to native-like states, a feature that will be a key component of the model detailed below.

■ MODEL AND THEORY

Given the contact map formalism and the corresponding method of calculating rates between states, we have all the tools needed to describe protein folding. Next, we must define a set of relevant states to consider, and then analyze the global kinetic properties resulting from the dynamics between these microstates.

Defining a State Space. To begin, we consider only states that are globular. While the issue of collapse has been a debated topic in the protein folding field, recent results from experiment,[50−54] theory,[57] and simulation[55,56] all indicate that nonspecific collapse precedes folding in the absence of any denaturant. The total number of contacts in all-atom explicit water MD simulations of protein folding show that the distribution of contacts is peaked around the native value, and that there are few conformations with few or no contacts, even in the unfolded state (see Supporting Information Figure 1).

We take the globular approximation to mean all states have the same number of total contacts and the loop entropy of each state is equivalent. This implies that all states of the model, independent of their degree of native content, consist of conformations with the same number of contacts as the native state. We denote this number of contacts by q; notice $q = q^{\alpha,\alpha} = q^{\mathrm{N,N}}$ for any state α. Following this notation, the free energy of a state is equal only to its internal energy, a function of the fraction of native contacts, $F_\alpha = E_\alpha = \varepsilon_x q^{\alpha,\mathrm{N}} + \varepsilon_{\mathrm{NN}} q$.

Now, we calculate the entropy of our state space by simply counting the number of possible contact maps, with each unique map corresponding to a unique state. We write the entropy of the system, S, for each possible number of native contacts, n, assuming that contacts can be formed between any two residues separated by at least three residue links. For a chain of length N residues

$$S(n) = k \log \Omega(n)$$

$$\Omega(n) = \binom{\frac{1}{2}(N-3)^2 - q}{q - n}\binom{q}{n}$$

where $\Omega(n)$ is the number of states with n native contacts. The first binomial bracket represents the number of ways to choose $q - n$ non-native contacts from the contact map (omitting the q native contact positions), and the second bracket is the way to choose n native contacts from q possible native contacts. For conciseness, we define $a_\Omega \equiv (N-3)^2/2$. Therefore, for a chain of $N \approx 100$ residues, a_Ω is a constant of order 10^4.

This estimate of $\Omega(n)$ in its general form has two issues. First, it is somewhat intractable analytically, involving many factorials. Second, it certainly overcounts the number of states, some of which will not be allowed due to, e.g., excluded volume effects. Through a series of mathematical approximations, we derive (Supporting Information)

$$\Omega(n) \approx e^{c(q-n)}$$

$$S(n) \approx ck(q-n)$$

where c is a constant of $O(1)$. Numerical results verify this approximation captures the scaling of the function, and this result is consistent with the exponential scaling estimated by Levinthal-like arguments. Further, exponential scaling is consistent with previous estimates from polymer theory and other simple models.[47,60] This state space definition results in a huge number of states; the total state count is $\Omega_{\mathrm{T}} = \sum_n \Omega(n) \sim O(10^{200})$.

The free energy of this state space projected onto the fraction of native contacts, $Q = n/q$, is

$$\frac{F(Q, T)}{q} = Q(\phi_x + ckT) + \phi_{\mathrm{NN}} - ckT$$

Below a certain temperature $T_{\mathrm{m}} = -\phi_x/kc$, the native state $(Q = 1)$ has the lowest free energy, and is therefore highly populated. Above that temperature, highly non-native states $(Q = 0)$ will have the most population and this population will be distributed among the many states in this regime. It is clear that this temperature corresponds to the folding phase transition but does not say anything about how that transition occurs.

Thermodynamics. From this state space, it is straightforward to calculate the thermodynamic properties of the model. We can write the partition function in terms of the degeneracy of states with n native contacts, g_n, and the energy of those states $(E_\alpha = \varepsilon_x q^{\alpha,\mathrm{N}} + \varepsilon_{\mathrm{NN}} q)$, obtaining

$$\begin{aligned} Z &= \sum_n g_n e^{-\beta E_n} \\ &= \sum_{n=0}^{q} e^{c(q-n)} e^{-(\phi_x n + \phi_{\mathrm{NN}} q)/kT} \\ &= e^{-(\varepsilon_x + \varepsilon_{\mathrm{NN}})q} \frac{e^{(c+\varepsilon_x)(q+1)} - 1}{e^{c+\varepsilon_x} - 1} \end{aligned}$$

where the last step is evaluated as a difference of geometric sums. Note that $e^{(c-\varepsilon_x)(q+1)} \gg 1$, such that we can simplify the above by dropping the unity term in the numerator. Making this approximation and recalling from statistical mechanics that

$$C_V = \frac{1}{kT^2}\frac{\partial^2 \log Z}{\partial \beta^2}$$

we obtain

$$C_V = \frac{1}{4}\frac{\phi_x^2}{kT^2} \sinh\left[\frac{1}{2}\left(\frac{\phi_x}{kT} + c\right)\right]^{-2}$$

This function has a singularity at $T_{\mathrm{m}} = -\phi_x/kc$, at the melting temperature (Figure 1). Previously, both theory and experiment[7,61] have indicated folding is a first-order phase transition, consistent with this observation. Note that this value of T_{m} matches exactly the one obtained from our free energy profile in the previous section.

Estimation of T_{m}, based on typical values of ϕ_x (see the Supporting Information), indicates that it is near 370 K for a typical 100-residue protein. Interestingly, C_V does not depend directly on the non-native or native contact energies but is a function only of the excess native energy ε_x, and scales only weakly with chain length ($c \sim \log N^2$, see the Supporting Information). The lack of strong scaling is consistent with experimental observations,[62] which show no detectable correlation of T_{m} with length, stability, or other common thermodynamic or kinetic parameters. We hypothesize from this model that protein melting temperatures are likely dominated by specific structural and physical features of individual proteins, rather than chain length, thermodynamic stability, or other general features.

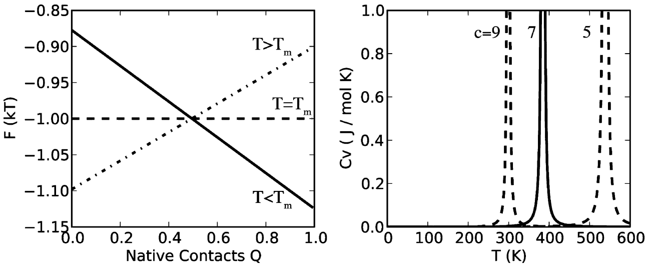


Figure 1. Thermodynamic properties of the protein folding model. (left) Free energy profiles along the order parameter Q (fraction of native contacts). There is no free energy barrier, but at low temperatures folded states are populated (solid line), whereas at higher temperatures unfolded states become more populated (dashed lines). (right) Heat capacity as a function of temperature. For a typical 100-residue protein, $c \approx 7$ and $T_m \approx 370$ K. There is a singularity corresponding to a highly cooperative first-order phase transition at T_m. The free energy profiles here verify this corresponds to the folding/unfolding transition.

Master Equation Approach. Now, we consider the statistics of transitions between states. The task at hand is taking the microscopic definition of kinetics provided by Kramer's approximation, $k_{\alpha\beta} = \tilde{k}\exp(-\beta\Delta F^{\ddagger}_{\alpha\beta})$, and applying it to the entire state space. To do this, we employ a Markovian master equation, a technique recently used to understand molecular simulations.[19−21,63] Under the master equation formalism, kinetics are defined by a rate matrix, $\mathbf{K} = \{k_{ij}\}$, where each k_{ij} represents the rate of transfer from state i to state j.

The master equation is simply an extension of first-order kinetics to many states, describing the time evolution of a set of populations, $P_i(t)$, corresponding to the probability of being in some state i at time t

$$\mathbf{K} = \begin{cases} k_{ij} \geq 0 \text{ if } i \neq j \\ k_{ii} = -\sum_{j \neq i} k_{ij} \end{cases}$$

$$\frac{\mathrm{d}P_i}{\mathrm{d}t} = \sum_j k_{ij}P_j$$

This formalism is very powerful, and able to capture complex phenomena in a straightforward manner. The theory of master equations is quite advanced, and many excellent reviews are available detailing recent advances.[19−21,63] Here, we restate many old results that are especially pertinent to our derivations.

In general, $\mathbf{K}$ represents a weighted graphical model, where nodes represent the conformations the protein could adopt (i.e., each node is a contact map), and the edges represent the rates of exchange between each node. To simplify analytical computations, here we will derive an unweighted model describing the topology of the network kinetics, and then weight the model to recover thermodynamic properties. The unweighted rate matrix $\mathbf{K}^{\mathrm{SYM}}$ is symmetric and is the opposite of the graph Laplacian $\mathbf{L}$

$$\mathbf{K}^{\mathrm{SYM}} = -\mathbf{L} = \mathbf{A} - \mathbf{D}$$

where $\mathbf{D} = \mathrm{diag}(d_i)$ is the diagonal matrix of degrees (the number of connections a node has to other nodes) and $A_{ij} \in \{0, 1\}$ is the adjacency matrix, a sparse symmetric square matrix with a 1 if ij share an edge and 0 otherwise.

Our task then is to derive $\mathbf{A}$. If one state converts to another at a rate faster than some critical rate, k_c, then we say those states are *kinetically connected*, and those vertices are connected by an edge ($A_{ij} = 1$). Otherwise, set $A_{ij} = 0$. It is worth mentioning that this equal weighting of edges in $\mathbf{K}^{\mathrm{SYM}}$ is consistent with what has been observed in detailed simulation, most significantly work showing that the edge weights are robust to perturbation.[64]

This simplified representation will give us a symmetrical matrix we denote $\mathbf{K}^{\mathrm{SYM}}$ for emphasis. The eigenvectors of the symmetric matrix do not represent the dynamics of the actual, unsymmetric master equation based on $\mathbf{K}$ (see, e.g., ref 63). However, it is required by detailed balance that $\mathbf{K}$ has a symmetric form

$$\mathbf{K} = \mathbf{P}_{\mathrm{eq}}^{1/2}\mathbf{K}^{\mathrm{SYM}}\mathbf{P}_{\mathrm{eq}}^{-1/2}$$

where $\mathbf{P}_{\mathrm{eq}} = \mathrm{diag}[P_i(t = \infty)]$ is a diagonal matrix of the equilibrium populations, which we have from our previous considerations of thermodynamics. $\mathbf{P}_{\mathrm{eq}}$ is also the eigenvector corresponding to the first eigenvalue of $\mathbf{K}$, $\lambda_1 = 0$, and represents the stationary distribution

$$\mathbf{P}_{\mathrm{eq}}\mathbf{K} = 0$$

All the dynamical processes are given by the left eigenvectors, ψ_n, $n \geq 2$

$$\psi_n\mathbf{K} = \lambda_n\psi_n$$

and the time scales of these dynamics are given by the eigenvalues, λ_n. Specifically, there are n time scales $\tau_n/\tau_0 = \lambda_n$, where τ_0 is a constant depending on the unit of time used. Note that the eigenvalues of $\mathbf{K}$ and $\mathbf{K}^{\mathrm{SYM}}$ are identical and only their eigenvectors are distinct.

We proceed by evaluating the simple case of the unweighted model by statistically analyzing which states should be connected, and then set $k_{ij} = 1$ for those states, and 0 otherwise, giving $\mathbf{K}^{\mathrm{SYM}}$. We could then recover the exact kinetics by weighting this matrix by $\mathbf{P}_{\mathrm{eq}}$. However, since in what follows we never calculate an eigenvector for the system but only the eigenvalues, we work exclusively with $\mathbf{K}^{\mathrm{SYM}}$.

Kinetic Topology of the State Space. Now, from the microscopic Kramer equation, we compute which states should be kinetically close, and investigate the resulting master equation. As mentioned previously, we are considering a simplified representation of the kinetics on our state space where two states are either "kinetically close" and rapidly interconvert or are distant and exchange slowly. The precise time scales determining these regimes, as we will soon see, are not important. If two states are kinetically close, we connect them by an edge in our graph, and increment each of their degrees.

To determine the kinetic degree of some state β, consider all states $\{\alpha\}$ with a certain number of contacts differing from β. Denote this distance $\Delta q \equiv q - q^{\alpha,\beta}$, which is the number of contacts that are broken during the transition $\alpha \rightarrow \beta$. The rate of interconversion from $k_{\alpha\beta}$ will be a monotonically decreasing function of Δq. This means that, to differentiate between kinetically close and distant structures, we need to define some cutoff rate, k_c, such that if $k_{\alpha\beta} \geq k_c$ the states are kinetically connected, and if $k_{\alpha\beta} < k_c$ they are not.

Because we are considering only an unweighted model, the topology of the network is the sole determinant of the time scales involved. One concise, computationally friendly way to describe this topology is the network degree distribution, that is, the distribution of the number of connections each node has. The degree distribution alone ignores correlations among specific nodes, and thus captures only one aspect of the topology, but because we have no direct access to this information in our current model, we take the *ansatz* that degree correlations are negligible, and therefore that the degree distribution contains all relevant information about the network topology.

This degree will simply be the number of states in the model that are kinetically close. Let us compute the degree of state β, which will be representative of the degree distribution for any state with $q^{\beta,\mathrm{N}}$ native contacts

$$d_\beta = \sum_{q^{\alpha,\mathrm{N}}} \sum_{\Delta q} \mathcal{N}(\Delta q)\theta(k_{\alpha,\beta} \geq k_c)$$

where $\mathcal{N}(\Delta q)$ is a function counting the number of structures within Δq broken contacts of β, and $\theta(\cdot)$ is the Heaviside step function asserting that the states are kinetically close. This expression simply counts the number of states in our model satisfying the condition $k_{\alpha,\beta} \geq k_c$.

To evaluate this, recall our rate for $\alpha \rightarrow \beta$

$$\begin{aligned} k_{\alpha,\beta} &= \tilde{k}\exp[-\varepsilon_x(q^{\alpha,\beta,\mathrm{N}} - q^{\alpha,\mathrm{N}}) - \varepsilon_{\mathrm{NN}}(q^{\alpha,\beta} - q) \\ &\quad + \sigma(q - q^{\alpha,\beta})] \\ &\approx \tilde{k}\exp(\varepsilon_x q^{\alpha,\mathrm{N}} - \varepsilon_x b q^{\beta,\mathrm{N}} + \varepsilon_{\mathrm{NN}}\Delta q + \sigma\Delta q) \end{aligned}$$

where we have taken the linear approximation $q^{\alpha,\beta,\mathrm{N}} \approx bq^{\beta,\mathrm{N}}$, which will be valid if native contacts are distributed in some regular way in contact space. This approximation is mathematically and physically motivated in the Supporting Information, and the value for $b \approx O(10^{-1})$ is derived.

The expression for $k_{\alpha,\beta}$ shows explicitly how the rate is faster when entering a highly native state (bigger $q^{\beta,\mathrm{N}}$) and slow when leaving a native-like state (bigger $q^{\alpha,\mathrm{N}}$). Now, for each $q^{\alpha,\mathrm{N}}$, we find some $\Delta q_{\max}$ satisfying $k_{\alpha,\beta} \geq k_c$ for all $\Delta q \leq \Delta q_{\max}$

$$\begin{aligned} k_c &= \tilde{k}\exp(\varepsilon_x q^{\alpha,\mathrm{N}} - \varepsilon_x b q^{\beta,\mathrm{N}} + (\varepsilon_{\mathrm{NN}} + \sigma)\Delta q_{\max}) \\ \Delta q_{\max} &= -\frac{\varepsilon_x}{\varepsilon_{\mathrm{NN}} + \sigma} q^{\alpha,\mathrm{N}} + \frac{b\varepsilon_x}{\varepsilon_{\mathrm{NN}} + \sigma} q^{\beta,\mathrm{N}} + \frac{\log k_c/\tilde{k}}{\varepsilon_{\mathrm{NN}} + \sigma} \\ \Delta q_{\max} &= -Eq^{\alpha,\mathrm{N}} + bEq^{\beta,\mathrm{N}} + \frac{\log k_c/\tilde{k}}{\varepsilon_{\mathrm{NN}} + \sigma} \end{aligned}$$

where in the last line we have defined a new value, $E \equiv \varepsilon_x/(\varepsilon_{\mathrm{NN}} + \sigma)$. Note that, as long as $\varepsilon_x < 0$ and $\varepsilon_{\mathrm{NN}} + \sigma < 0$, i.e., both native and non-native contacts are sufficiently favorable, $E > 0$. Our choice of k_c should not matter to the final solution, and we will see proof of this later. Therefore, choose the simplest k_c, specifically $k_c = \tilde{k}$ such that the last term here drops out.

It follows from the way we choose $\Delta q_{\max}$ that we can rewrite our degree evaluation with specific definite bounds to enforce the Heaviside function

$$d_\beta = \sum_{n=0}^{q} \sum_{\Delta q=1}^{\Delta q_{\max}} \mathcal{N}(\Delta q)$$

where to lighten the notation we have written $n = q^{\alpha,\mathrm{N}}$. Now, we must evaluate $\mathcal{N}(\Delta q)$. Physically, this is the total number of states kinetically accessible when one state breaks and re-forms Δq contacts. Strict combinatorial evaluation, while possible, will surely overcount this number without explicit consideration of, e.g., excluded volume effects and the restriction that nearby residues must form nearby contacts due to the polymer backbone.

We can estimate the number of accessible states from a simple physical argument. Let the value z represent the number of choices each residue in the loop has to form a new contact. Then, if each residue broken has approximately z choices that allow the other residues to form their contacts, we obtain

$$\mathcal{N}(\Delta q) = z^{\Delta q}$$

or exponential scaling in the number of possible conformations with the number of broken contacts.

We can estimate z as follows. Contacts will almost certainly be broken on the surface of the protein, where loops can form. Locally, this surface will be approximately planar. Empirically, the coordination number of residue−residue contacts is about 6,[65,66] consistent with traditional folding lattice models.[42,67−69] This means that we can approximate the surface of the protein as a lattice with coordination number 6, and envision a loop settling on this surface. Consider now one residue on the lattice surface, an anchor site that is the beginning of the loop. Now place a residue adjacent to this one on the lattice. Clearly, there are four options. Now, place another residue adjacent to the residue just placed. There will only be three options for this residue, and all residues after that one will have three, two, or one placement possibilities (Figure 2). Therefore, we conclude that $1 < z < 4$ but is most likely about 2, considering steric and energetic restraints. Luckily, we will see that the final folding rate scales only linearly with z while growing exponentially with other factors.

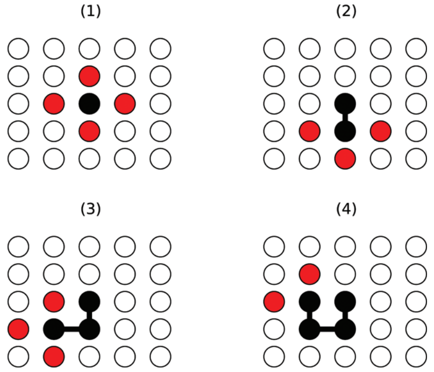


Figure 2. Illustration of the process used to estimate z. Black dots represent placed residues on a plane of lattice sites, red dots represent potential sites for the next residue along a chain. (1) through (4) show the sequential placement of four residues, demonstrating situations where two, three, and four sites are available for the next residue. These considerations lead to an estimate of $1 < z < 3$.

With the expression for $\mathcal{N}$, evaluation of d_β is straightforward. Note d_β is a function of $q^{\beta,\mathrm{N}}$ only. Writing $\delta \equiv bEq^{\beta,\mathrm{N}}$ to further lighten the notation, we obtain

$$\begin{aligned} d_\beta(q^{\beta,\mathrm{N}}) &= \sum_{n=0}^{q} \sum_{\Delta q=1}^{\Delta q_{\max}} z^{\Delta q} \\ &= \frac{z}{z-1} \frac{z^\delta(z^{E(q+1)} - 1) - (z^E - 1)(q+1)}{z^E - 1} \\ &\propto \frac{z^{E(q+1)+\delta} - z^\delta}{z^E - 1} \end{aligned}$$

where we have dropped constant values.

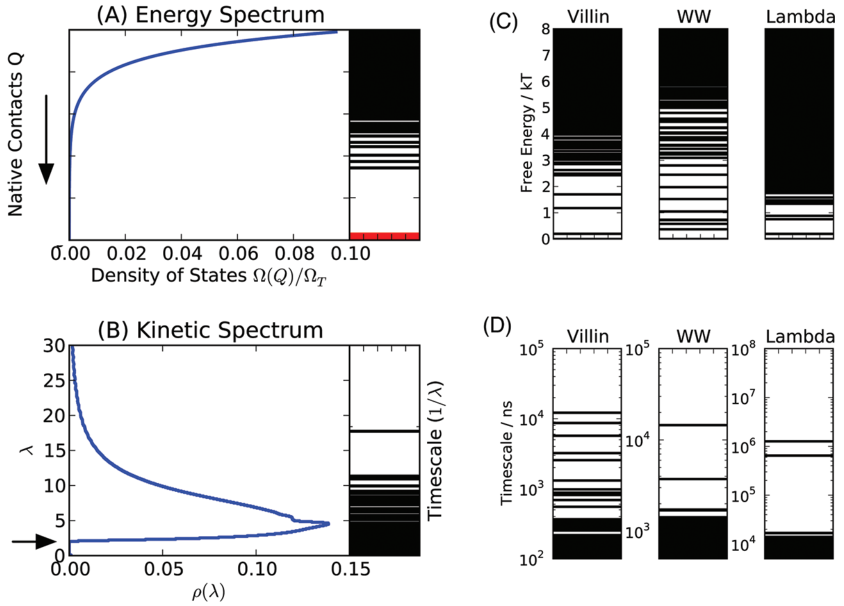


Figure 3. Spectra of the model for typical proteins. (A) The energy spectrum of the states. An exponential density leads to a large energy gap in the spectrum, previously predicted to be important for the stability of folded structures.[12] The left panel is the density of states, and the right is a spectrum of 1000 states drawn at random from that density for illustration. (B, left panel) Density of states of a typical scale-free eigenspectrum. Plotted is the average over 580 simulations of uncorrelated scale-free networks with 10^4 nodes, $\mu = 3.0$, $d_{\min} = 5$. The first time scale is necessarily bounded away from long times by the spectral edge at λ_c (arrow), leading to fast folding. Also shown are (right panel) the longest 50 time scales $(1/\lambda)$ from the first simulation, plotted on a log-scale. (C and D) The energetic and kinetic spectra of MSMs parametrized from explicit water all-atom MD simulations,[23,24,27] for comparison.

This expression can be simplified even further. Note $z^{E(q+1)} \ll -1$, so we can write

$$d_\beta(q^{\beta,\mathrm{N}}) \propto z^\delta = z^{bEq^{\beta,\mathrm{N}}}$$

giving us directly the number of kinetic connections state β has. Notice from the evaluation of $\Omega(n)$ that the probability of picking a state with $q^{\beta,\mathrm{N}}$ native contacts uniformly at random is $P(q^{\beta,\mathrm{N}}) \propto \exp[c(q - q^{\beta,\mathrm{N}})]$. The one-to-one correspondence between $q^{\beta,\mathrm{N}}$ and d_β, simple substitution, gives

$$\begin{aligned} P(d_\beta) &\propto \exp\left[-\frac{c}{bE\log z}\log(d_\beta)\right] \\ &\propto d_\beta^{-c(bE\log z)^{-1}} \\ &\propto d_\beta^{-\mu} \end{aligned}$$

Revealing that the topology of our model exhibits a power-law degree distribution with exponent $\mu = c(bE\log z)^{-1}$. This results in a *native hub*. This is a key result describing the structure of protein kinetics in a discrete configuration space. Our work up to this point has shown that this space has a hub-like structure; from here, we will see if this structure has any consequences for the system kinetics.

One can see that, if $E < 0$, then native states with large $q^{\beta,\mathrm{N}}$ will have connections to many states (large Δq) and therefore have high degrees. If $E > 0$ (non-native contacts become energetically unfavorable), a phase transition occurs where μ becomes negative, and we get many non-native states that have a high degree. The model presented here does not accurately represent this regime, since the lack of attractive non-native interactions will cause the unfolded state to consist of coil-like (as opposed to globular) conformations. We hope to examine this regime in future work, especially considering the importance of coil-like conformations of proteins under denaturing conditions.

Graphs with a power-law degree distribution, commonly known as scale-free networks,[70] are well studied, because they appear in many real world networks (e.g., WWW, Internet, protein interactions, etc.) and have many interesting properties. In what follows, we focus on the eigenspectrum of this structure, which reveals the dynamics involved in protein folding.

The Kinetic Spectrum. Here, we evaluate the first dynamical eigenvalue of the graph Laplacian, and use it to calculate the folding time. The Laplacian has n eigenvalues $\lambda_1 = 0 < \lambda_2 < ... < \lambda_n$ that describe the time scales of the dynamical processes of the system. These eigenvalues will be distributed according to an eigenspectrum specific to the graph in question. We are especially interested in λ_2, which corresponds to the slowest relaxation in the kinetic system—in the case of proteins, we assume this corresponds to folding. This eigenvalue is the inverse of the folding time scale, $\lambda_2 = \tau_0/\tau_f$. Therefore, showing λ_2 is large is equivalent to showing that proteins fold quickly.

It has been shown that λ_2 depends primarily on the nodes with the smallest degrees, $d_{\min}$.[72]

There exist proven bounds[73−75] on the value of λ_2 in terms of $d_{\min}$ for the large-graph limit, which we recall in the Supporting Information and present here

$$md_{\min} < \lambda_2 \leq \frac{V}{V-1}d_{\min},\ 0 < m < 1$$

where V is the number of nodes in the graph and m is some constant. Since both of these bounds are linear, they indicate that $\lambda_2 \sim d_{\min}$. Numerical simulations of uncorrelated scale-free matrices confirm linear scaling (Figure 4). Using the expression

$$d_\beta \propto \frac{z^{E(q+1)+\delta} - z^\delta}{z^E - 1}$$

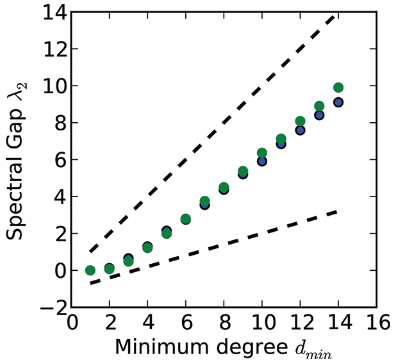


Figure 4. The scaling of the first dynamical eigenvalue λ_2 as a function of minimum degree $d_{\min}$ in scale-free graphs. Presented are the results for 10^3 node and 10^4 node graphs, in blue and green, respectively. Graphs were simulated so as to have uncorrelated degrees, using the modified configuration model of Catanzaro, Boguna, and Pastor-Satorras,[71] with parameter $\mu = 3$. Graphs with $\mu = 2$, $\mu = 2.5$, and $\mu = 5$ were also investigated and showed similar results. Our presented bounds, both upper and lower ($m = 0.3$), are plotted as dashed lines. The lower bound is shifted down by 1, for clear illustration (it is not applicable for $d_{\min} = 1$, where there is no guaranteed giant component).

and recalling $\delta \propto q^{\beta,\mathrm{N}}$, we can see that the minimum degree is obtained when $q^{\beta,\mathrm{N}}$ is minimized. Taking $q^{\beta,\mathrm{N}} = 0$, we have

$$d_{\min} \propto z^{Eq}$$

Then, applying our scaling $\lambda_2 \sim d_{\min}$, we obtain

$$\begin{aligned} \lambda_2 &\sim z^{Eq} = z^{E\rho_\mathrm{N}\mathrm{N}} \\ &\sim z^{-E\rho_\mathrm{N}\mathrm{N}} \end{aligned}$$

$$\tau_\mathrm{f}/\tau_0 \sim \mathrm{e}^{\xi\mathrm{N}} \quad \xi = \varepsilon_x\rho_\mathrm{N} \log z/(\varepsilon_\mathrm{NN} + \sigma)$$

where τ_f is the folding rate and we have noted that the number of contacts is expected to scale linearly with chain length, $q = \rho_\mathrm{N}N$. Here, we have denoted the contact density, simply a constant of proportionality, by ρ_N. Empirical data suggest $\rho_\mathrm{N} \approx 0.6$ (Supporting Information, Figure S2). Estimates of the energetic parameters (ε_N, ε_NN, and σ) from simulation,[23−27] theory,[36,37,44] and experiment[S9,65,66] yield values of $\varepsilon_\mathrm{N} \approx -3$, $\varepsilon_\mathrm{NN} \approx -2$, and $\sigma \approx 1$ (see the Supporting Information). From these approximations, the exponential factor is $\xi \approx O(10^{-1})$. This functional form is consistent with the derivation of Zwanzig,[47] who used a different model to arrive at a similar result.

The computation of τ_f is the central result of our model. It shows explicitly why proteins fold quickly, and how the folding time scales as a function of chain length. The fact that scale-free networks have a spectral edge that is bounded away from $\lambda_1 = 0$, the stationary distribution, means that the first dynamical process occurs very quickly, much more quickly than the naive prediction of Levinthal. Furthermore, though we have derived exponential scaling in folding time as a function of chain length, the exponential constant is very small, so the scaling is quite weak. This shows how proteins could grow to moderate chain lengths (10^3 residues) before reaching biologically intractable folding times.

■ DISCUSSION

We have presented a model that is quite simple—it represents an attempt at a minimal description of protein kinetics. Despite this simplicity, interesting and subtle features emerge. The master equation predicts that the native state is a kinetic hub, highly connected to many diverse nodes, while non-native states interconvert slowly, similar to results from detailed all-atom simulations. This hub has a regular structure, that of a scale-free network, the topology of which dictates the time scales of dynamics in the system.

Considering the physical origins of the hub, a simple picture of folding emerges. The model depicts protein folding as a search through globular structures, converting from globule to globule through transition states with short loops. Native contacts form heterogeneously and stochastically, and once these favorable contacts are formed, they are likely to persist, since it is energetically costly to break them. We call this tendency *statistical persistence*. Eventually, the globule accrues more and more native content, until it is completely folded. While searching through globular states, the protein hops between compact conformations that involve only moderate rearrangements, forming temporary short loops that then collapse again (Figure 6).

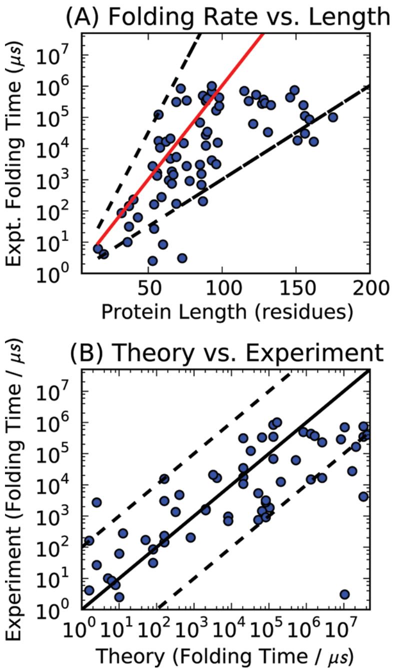


Figure 5. Theoretical prediction of the folding rate compared with 78 experimental values, with no fit parameters. (A) The folding time scaling as a function of chain length (red solid line), for the estimated value of the exponential scaling parameter, $\xi = 0.14$ ($z = 2$, $\phi_\mathrm{NN} = 1.0$ kcal mol^{-1}, $\phi_\mathrm{N} = 1.5$ kcal mol^{-1}, $\Delta S_\mathrm{loop} = 3.5$ cal mol^{-1} K^{-1}). For comparison, a least-squares fit gives $\xi = 0.10$ ($R^2 = 0.53$). The dashed black lines show extremal values of ξ, the estimated value increased and decreased by 50%. (B) The same data, comparing theory and experiment, but including the specific number of contacts of each protein in the native state, q, rather than the average contact density ρ_N. This results in a negligible difference in the model's approximation quality. The black line is an exact comparison. The dashed lines indicate 2 orders of magnitude variation, which might be accounted for by, e.g., mutations and temperature variations, factors not explicitly included in the model. Data from ref 76. The severe outlier in the bottom right of part B is α_3D (PDB: 2A3D), a *de novo* designed protein.

One interesting point is that favorable non-native contacts play a key role in this picture. These contacts ensure that unfolded states are compact globules. Further, these non-native contacts are directly responsible for the strong hub-like features of the model and the fast folding predicted (recall, without

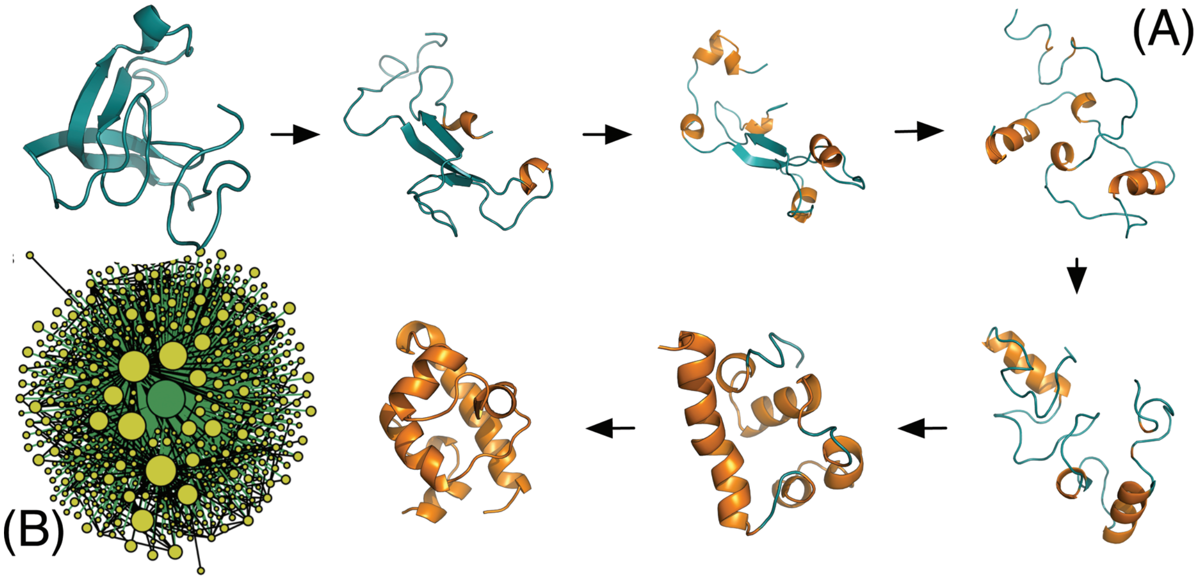


Figure 6. Exemplar pathway, taken from a highly traveled pathway in all-atom explicit water simulations of lambda repressor, and a representation of the MSM built from the simulations, illustrating the process captured by the simple model. (A) Protein folding occurs as globular structures interconvert, slowly gaining native content (orange) in a stochastic manner. (B) This process leads to a kinetic hub that dictates the time scales of the system. Represented here is the network of an MSM for this protein. Reproduced with modification from Bowman and Pande (2010),[24] with permission.

favorable non-native contacts, $E > 0$). Without favorable non-native contacts, the model breaks down. The properties of a model similar to this one in the unfavorable non-native regime is still an open question, and it is difficult to speculate on what such a model would look like.

Analysis of these kinetics predicts that the folding time scales weakly exponentially with chain length, $\tau_f \propto e^{\xi N}$, with $\xi \approx O(10^{-1})$, in strong agreement with experiment (Figure 5), though chain length is certainly not the only variable affecting folding time. Though exponential scaling might be thought as a return to Levinthal's paradox, the model explicitly shows that the exponential constant is small. We hypothesize that exponential scaling may provide an explanation for why protein domain sizes cover less than 2 orders of magnitude (10−1000 residues) and why large proteins often need chaperones to fold.[77]

Our model shines light on a number of previous experimental and theoretical observations, which we review here. Interestingly, the thermodynamics of the model support two-state behavior, exhibiting a highly cooperative transition between native and non-native states at the melting temperature (T_m). This is surprising considering that the model also exhibits many time scales arising from heterogeneous pathways. The discrepancy may indicate why the folding process appears to be quite simple, either two or three state, when viewed with a low-resolution experimental probe but appears complex when examined with high resolution techniques. An example is the small protein villin,[78] a traditional two-state folder that exhibits complex kinetics when viewed at high resolution, either in experiment[79] or simulation.[26] The apparent two-state nature of proteins is likely augmented by the fact that, while the time scales and mechanisms of folding might be heterogeneous, the thermodynamics can still dictate strong two-state behavior.[80−83]

We should note that currently whether or not this model predicts two-state behavior in a kinetic sense remains an open question. Experiments show that many (but not all) proteins not only exhibit two dominant thermodynamic states but single exponential relaxation between those states. In our current formalism, this would take the form of a separation of time scales; i.e., the third eigenvalue should be much larger than the second $\lambda_3 \gg \lambda_2$. This does not follow immediately from the topology alone, suggesting that something deeper is at play.

The hub-like nature of the model predicts folding pathways should be highly heterogeneous, without any clear global transition state. While the issue of heterogeneity is not a resolved topic,[22,25,29,38,84] our model is consistent with experimental observations indicative of heterogeneity, such as the fact that the vast majority of ϕ-values are intermediate between 0 and 1.[85−88] Furthermore, heterogeneity was a key insight of energy landscape theory, which proposed that multiple pathways helped to resolve the Levinathal paradox.[6,8,9] Our model not only agrees with this aspect of that theory but explicitly shows how parallel pathways arise from microscopic considerations, and how they lower the folding time.

Moreover, our model predicts that a few native-like states will be accessible at equilibrium, and that these states will retain a high degree of native content (many native contacts) but be considerably higher in free energy than the native state (Figure 3). These states have been observed in such experiments such as equilibrium hydrogen−deuterium exchange[38,89] and have been previously predicted by simple theories and simulations.[12,61]

Themodel's key feature, the kinetic network topology, could be directly observable in experiments. We predict that, under native conditions, a single molecule experiment able to distinguish the native state and any two or more non-native states, as well as the rates of interconversion between them, should observe many more transitions from non-native states to the native state than between non-native states. At the time of publication, we are unaware of any such experiment. Observation of this kind of behavior for a large number of states on a diverse set of systems would, in our opinion, constitute good evidence of the kinetic hub theory derived here.

CONCLUSIONS

By treating kinetics directly, we derived a simple model that captures the essential features of protein folding kinetics and thermodynamics. The model exhibits qualitative fidelity with experiment and simulation, and explicitly shows how native and non-native residue contacts could lead to a kinetic hub displaying time scales appropriate for protein folding. Despite these successes, our model is quite different from how protein folding has been previously viewed. We emphasize that there is no global transition state in the model. Folding kinetics are a consequence of the scale-free properties of the hub, rather than from the crossing of one dominant free energy barrier.

Why does our model look so different from previous theory? There are two main reasons. First, we have dealt directly with kinetics, rather than energy landscapes. This allows us to retain the full dimensionality of conformation space, without having to project onto a reaction coordinate. In high-dimensional systems, these projections can lead to significant errors.[17,18] Second, here we have not presumed the principle of minimal frustration. Indeed, we see that attractive non-native contacts are an important part of the model, and that without them the kinetic hub is not present. The experimentally supported prediction that water is a poor solvent for protein chains indicates that attractive non-native contacts are appropriate.[50−57]

While fundamentally different from previous theory, this model captures nearly all the key qualitative features of protein folding kinetics, and we hope it will provide a foundation for understanding those kinetics in future studies of protein folding.

ASSOCIATED CONTENT

Ⓢ Supporting Information

(1) The detailed derivation of $\Omega(n)$, (2) the approximation of $q^{\alpha,\beta,N}$, (3) the derivation of the bounds on λ_2, (4) a histogram showing collapse in protein folding simulations, (5) a brief discussion of the scaling of the number of residue−residue contacts with chain length, and (6) details of the estimation of contact energies and entropies. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pande@stanford.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The history and literature of protein folding is extensive. We would like to acknowledge the fact that many scholars have contributed work to the field, and while we did our best to cite previous work, we may have regrettably missed the contributions of some. We would like to thank Christian Schwantes, Gregory Bowman, Alexander Samukhin, Jesus Izaguirre, and Ken Dill for insightful discussions. Thanks also to Gregory Bowman, Kyle Beauchamp, and Vincent Voelz for generously sharing data. T.J.L. was supported by an NSF GRF. This study was made possible through funding from the NSF (MCB-0954714) and NIH (R01-GM062868).

REFERENCES

(1) Anfinsen, C. *Science* **1973**, *181*, 223−230.
(2) Levinthal, C. J. *Med. Phys.* **1968**, *65*, 44−45.
(3) Levinthal, C. *Mossbauer Spectroscopy in Biological Systems* **1969**, 22−24.
(4) Dill, K. A.; Ozkan, S.; Shell, M.; Weikl, T. R. *Annu. Rev. Biophys.* **2008**, *9*, 289−316.
(5) Baldwin, R. *J. Biomol. NMR* **1995**, *5*, 103−109.
(6) Dill, K.; Chan, H. *Nat. Struct. Biol.* **1997**, *4*, 10−19.
(7) Shakhnovich, E. I.; Finkelstein, A. V. *Biopolymers* **1989**, *28*, 1667−1680.
(8) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545−600.
(9) Onuchic, J. N.; Wolynes, P. G. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70−75.
(10) Jackson, S. E.; Fersht, A. R. *Biochemistry* **1991**, *30*, 10428−10435.
(11) Radford, S.; Dobson, C. *Nature* **1992**, *358*, 302−307.
(12) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248−251.
(13) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins* **1995**, *21*, 167−195.
(14) Bryngelson, J.; Wolynes, P. G. *Biopolymers* **1990**, 177−188.
(15) Bryngelson, J. D.; Wolynes, P. G. *J. Phys. Chem.* **1989**, 6902−6915.
(16) Bryngelson, J. D.; Wolynes, P. G. *Biophysics* **1987**, *84*, 7524−7528.
(17) Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S. *J. Chem. Phys.* **1998**, *108*, 334.
(18) Chandler, D. Finding Transition Pathways: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Classical and Quantum Dynamics in Condensed Phase Simulations*; 1998; pp 51−66.
(19) Noé, F.; Fischer, S. *Curr. Opin. Struct. Biol.* **2008**, *18*, 154−162.
(20) Prinz, J.-H.; Keller, B.; Noé, F. *Phys. Chem. Chem. Phys.* **2011**, *13*, 16912−16927.
(21) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105.
(22) NoÃl’, F.; SchÃijtte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci.* **2009**, *106*, 19011−19016.
(23) Voelz, V. a.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526−1528.
(24) Bowman, G. R.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, 12−15.
(25) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341−346.
(26) Beauchamp, K.; Ensign, D.; Das, R.; Pande, V. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 12734.
(27) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 18413−18419.
(28) Voelz, V. A.; Jager, M.; Yao, S.; Chen, Y.; Zhu, L.; Waldauer, S. A.; Bowman, G. R.; Friedrichs, M.; Bakajin, O.; Lapidus, L. J.; Weiss, S.; Pande, V. S. *J. Am. Chem. Soc.* **2012**, submitted for publication.
(29) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. *Science* **2011**, *334*, 517−520.
(30) Rao, F.; Caflisch, A. *J. Mol. Biol.* **2004**, *342*, 299−306.
(31) Bowman, G.; Pande, V. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 10890.
(32) Pande, V. S. *Phys. Rev. Lett.* **2010**, *105*, 1−4.
(33) Plaxco, K. W.; Simons, K. T.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985−994.
(34) Plaxco, K. W.; Simons, K. T.; Ruczinski, I.; Baker, D. *Biochemistry* **2000**, *39*, 11177−11183.
(35) Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311−11316.
(36) Linse, S.; Linse, B. *J. Am. Chem. Soc.* **2007**, *129*, 8481−8486.
(37) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 20.
(38) Maity, H.; Maity, M.; Krishna, M. M. G.; Mayne, L.; Englander, S. W. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 4741−4746.
(39) Ravasz, E.; Gnanakaran, S.; Toroczkai, Z. Arxiv preprint arXiv:0705.0912, 2007, 1−15.

(40) Bruscolini, P.; Pelizzola, A. *Phys. Rev. Lett.* **2002**, *88*, 1–4.

(41) Chan, H. S.; Dill, K. A. *Proteins* **1998**, *30*, 2–33.

(42) Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. *Protein Sci.* **1995**, *4*, 561–602.

(43) Ghosh, K.; Ozkan, S. B.; Dill, K. A. *J. Am. Chem. Soc.* **2007**, *129*, 11920–11927.

(44) Kubelka, J.; Henry, E. R.; Cellmer, T.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18655–18662.

(45) Portman, J.; Takada, S.; Wolynes, P. *Phys. Rev. Lett.* **1998**, *81*, 5237–5240.

(46) Shakhnovich, E. I.; Gutin, a. M. *Biophys. Chem.* **1989**, *34*, 187–199.

(47) Zwanzig, R. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9801–9804.

(48) Thirumalai, D. *J. Phys. (France)* **1995**, *5*, 1457–1467.

(49) Gfeller, D.; De Los Rios, P.; Caflisch, A.; Rao, F. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 1817–1822.

(50) Hoffmann, A.; Kane, A.; Nettels, D.; Hertzog, D. E.; Baumgärtel, P.; Lengefeld, J.; Reichardt, G.; Horsley, D. A.; Seckler, R.; Bakajin, O.; Schuler, B. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 105–110.

(51) Best, R. B.; Merchant, K. A.; Gopich, I. V.; Schuler, B.; Bax, A.; Eaton, W. a. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 18964–18969.

(52) Schuler, B.; Lipman, E.; Eaton, W. *Nature* **2002**, *419*, 743–747.

(53) Ziv, G.; Haran, G. *J. Am. Chem. Soc.* **2009**, *131*, 2942–2947.

(54) Waldauer, S. A.; Bakajin, O.; Lapidus, L. J. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 13713–13717.

(55) Bowman, G. R.; Pande, V. S. Manuscript in preparation, 2011.

(56) Voelz, V. A.; Singh, V. R.; Wedemeyer, W. J.; Lapidus, L. J.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 4702–4709.

(57) Dill, K. A.; Shortle, D. *Annu. Rev. Biochem.* **1991**, *60*, 795–825.

(58) Pande, V. S.; Grosberg, AYu,; Tanaka, T. *Folding Des.* **1997**, *2*, 109–114.

(59) Makhatadze, G.; Privalov, P. *Adv. Protein Chem.* **1995**, *47*, 307–425.

(60) Flory, P. *Statistical mechanics of chain molecules*; Interscience Publishers: 1969.

(61) Shakhnovich, E. I. *Curr. Opin. Struct. Biol.* **1997**, *7*, 29–40.

(62) Franzosa, E.; Lynagh, K.; Xia, Y. *Experimental Standard Conditions of Enzyme Characterizations* **2010**, 99–106.

(63) Buchete, N.-V.; Hummer, G. *J. Phys. Chem. B* **2008**, *112*, 6057–6069.

(64) Weber, J.; Pande, V. *Biophys. J.* **2012**, *102*, 859–867.

(65) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623–644.

(66) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, 534–552.

(67) Cieplak, M.; Hoang, T.; Li, M. *Phys. Rev. Lett.* **1999**, *83*, 1684–1687.

(68) Gutin, A. M.; Abkevich, V. I.; Shakhnovich, E. I. *Phys. Rev. Lett.* **1996**, *77*, 5433–5436.

(69) Leopold, P. E.; Montal, M.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 8721–8725.

(70) Barabási, A.; Albert, R. *Science* **1999**, *286*, 509–512.

(71) Catanzaro, M.; Boguñá, M.; Pastor-Satorras, R. *Phys. Rev. E* **2005**, *71*, 1–4.

(72) Samukhin, A. N.; Dorogovtsev, S. N.; Mendes, J. F. F. *Phys. Rev. E* **2008**, *77*, 1–19.

(73) Mohar, B. *Graph. Combinator.* **1991**, *7*, 53–64.

(74) Cohen, R.; Havlin, S. *Phys. Rev. Lett.* **2003**, *90*, 5–8.

(75) Bollobas, B.; Riordan, O. *Combinatorica* **2004**, *24*, 5–34.

(76) Ouyang, Z.; Liang, J. *Protein Sci.* **2008**, *17*, 1256–1263.

(77) Hartl, F. U.; Hayer-Hartl, M. *Nat. Struct. Biol.* **2009**, *16*, 574–581.

(78) Kubelka, J.; Chiu, T. K.; Davies, D. R.; Eaton, W. A.; Hofrichter, J. *J. Mol. Biol.* **2006**, *359*, 546–553.

(79) Reiner, A.; Henklein, P.; Kiefhaber, T. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 4955–4960.

(80) Pirchi, M.; Ziv, G.; Riven, I.; Cohen, S. S.; Zohar, N.; Barak, Y.; Haran, G. *Nat. Commun.* **2011**, *2*, 493.

(81) Sridevi, K.; Lakshmikanth, G. S.; Krishnamoorthy, G.; Udgaonkar, J. B. *J. Mol. Biol.* **2004**, *337*, 699–711.

(82) Rhoades, E.; Cohen, M.; Schuler, B.; Haran, G. *J. Am. Chem. Soc.* **2004**, *126*, 14686–14687.

(83) Wright, C. F.; Lindorff-Larsen, K.; Randles, L. G.; Clarke, J. *Nat. Struct. Biol.* **2003**, *10*, 658–662.

(84) Englander, S. W.; Mayne, L.; Krishna, M. M. G. *Q. Rev. Biophys.* **2007**, *40*, 287–326.

(85) Sánchez, I. E.; Kiefhaber, T. *J. Mol. Biol.* **2003**, *325*, 367–376.

(86) Li, L.; Mirny, L. a.; Shakhnovich, E. I. *Nat. Struct. Biol.* **2000**, *7*, 336–342.

(87) Ozkan, S. B.; Bahar, I.; Dill, K. a. *Nat. Struct. Biol.* **2001**, *8*, 765–769.

(88) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. *Nature* **2001**, *409*, 641–645.

(89) Baldwin, A. J.; Kay, L. E. *Nat. Chem. Biol* **2009**, *5*, 808–814.