Supplemental Information for "A Simple Model Predicts Experimental Folding Rates and a Hub-Like Topology"

Thomas J. Lane and Vijay S. Pande*

Department of Chemistry, Stanford University

E-mail: pande@stanford.edu

Methods

Construction and Analysis of MSMs. Detailed presentations of the construction and analysis of Markov State Models are presented in the original reports of those models,[1–3] as well as in current manuscripts detailing the general state of the art in Markov model construction.[4–8]

Definition of a Contact. In what follows, we present a number of connections to simulations where it is necessary to define a contact. Often, results can vary significantly from small changes in this definition, due to the nature of a hard cutoff. Here, a contact was said to be formed if two residues had any atoms within 5 Angstroms of each other, and were separated by at least 3 residues in sequence, which was relatively robust and produced reasonable native contact maps of the proteins investigated.

*To whom correspondence should be addressed

Derivation of $\Omega(n)$

We begin with the exact expression for the number of contact maps with n native contacts

$$\Omega(n) = \frac{(a_\Omega - q)!}{[(a_\Omega - q) - (q-n)]!(q-n)!} \frac{q!}{n!(q-n)!}$$

for ease of calculation, convert to logarithms and apply Stirling's approximation. Next, we note $a_\Omega \gg q > n$, so $a_\Omega - q \approx a_\Omega$ and $a_\Omega - n \approx a_\Omega$. Making this substitution, after a straight-forward calculation we arrive at

$$\begin{aligned}
\log\Omega &\approx -q\log a_\Omega + q\log q - q\log(q-n) + 2q \\
&\quad -n\log a_\Omega - n\log n + n\log(q-n) - 2n \\
&= n\log\left(\frac{n}{a_\Omega}\right) - (q-n)\log(q-n) - 2n + c_0
\end{aligned}$$

with c_0 a constant. Note that the logarithms involving n will vary much more slowly than n, so we approximate them by their expectations, obtaining

$$\begin{aligned}
\log\Omega &\approx -c_1 n + c_2(n-q) - 2n + c_0 \\
\Omega(n) &= \exp\left[-(c_1 - c_2 + 2)n - c_2 q + c_0\right]
\end{aligned}$$

where our constants are

$$\begin{aligned}
c_0 &= -q\log a_\Omega + q\log q + 2q \\
c_1 &= \log(\langle n\rangle / a_\Omega) \\
c_2 &= \log(q - \langle n\rangle)
\end{aligned}$$

We could, in principle, easily evaluate these constants for our model, however, our forms here have taken a number of approximations, thus our derived values of these constants may not be those best suited to recapitulate the original function. Furthermore, certain properties for $\Omega(n)$ are desired expected based on simple physical arguments. Therefore, we have adopted the approach outlined below.

Of primary importance is the fact that we expect a single native state, both physically and from our original expression. This dictates $\Omega(n=q)=1$, so the argument of the exponential in the expression for $\Omega(n)$ should be identically 0 for $n=q$. It follows that a factor $(q-n)$ must therefore appear in the exponent. It is straightforward to show that, since the argument of the exponent in $\Omega(n)=\exp\left[-c_2 q-(c_1-c_2+2)n+c_0\right]$ is linear in both q and n, the simplest form of this expression consistent with this requirement is

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

where a quick calculation shows $1<c<10$ for realistic values of q (note that $q \approx 0.6N$, see "Linearity of Contacts and Chain Length" below). Graphical analysis shows that this approximation is consistent with the original expression, but does introduce some minor error. We note that since the original counting of the contact maps would over count states, small numerical errors may not decrease the fidelity of the model. More important is that we obtain an expression that leads to a large (exponential) number of states, consistent with Levinthal-style reasoning.

Approximation of $q^{\alpha,\beta,N}$

Consider $\langle q^{\alpha,\beta,N}\rangle \approx bq^{\beta,N}$, i.e. we reason that the average number of native contacts two states share is directly proportional to the number of native contacts one of those states has. We choose the product state β as our reference, since to calculate the connectivity (degree) of a state, it seems prudent to consider transitions *to* that state, rather than from the state. The approximation will be valid if the native contacts are distributed throughout contact space in some sort of regular fashion, an assumption that seems reasonable for real proteins.

Recall $a_\Omega \approx 10^4$, $E \approx -0.5$, $c \approx 7$ for a 100 residue protein, and the power exponent of the scale-free kinetic network is

$$\mu = -c(bE \log 2)^{-1}$$

The vast literature on scale-free networks show nearly all such networks have an exponent $2 < \mu < 3$.[9] We expect that our network is not too far from these “typical" values, and so μ is likely $\mathscr{O}(1)$, leading to an estimate of $b \approx \mathscr{O}(10^{-1})$. We note, however, that b does not feature prominently in the model, and its exact numerical value is relatively inconsequential.

Bounds on λ_2

The mathematical investigation of the spectral properties of graphs, especially scale-free graphs, is an active area of research. While the general properties of the spectra of scale-free graphs remain largely unknown (though progress is being made, see e.g. the work of Samukhin, Dorogovtsev, and Mendes[10]), there exist some easy to prove bounds on the second-smallest eigenvalue of the Laplacian (λ_2) for graphs in general.[11] Here, we do not explicitly present any proofs, but simply recall bounds from sources where they are discussed in detail.

One well-known bound from above is

$$\lambda_2 \leq \frac{V}{V-1} d_{min}$$

where V is the size of the graph (the number of nodes) and d_{min} is the minimum degree of any node. Of course, this approaches $\lambda_2 \leq d_{min}$ as $V \to \infty$. We employ this bound directly.

The second bound is

$$\lambda_2 \geq \frac{4}{\rho V}$$

where ρ is the graph diameter, or the largest number of edges needed to connect any two nodes in as direct a manner as possible. It has been shown that scale-free graphs are *ultra-small*, i.e. that the diameter scales as $\rho \sim \log \log V$ as $V \to \infty$.[12,13] Using this scaling, we can evaluate this bound in terms of our model. Recall

$$V = \Omega_T = \frac{e^{cq+c} - 1}{e^c - 1} > e^{cq}$$

gives

$$\begin{aligned}\lambda_2 &> \frac{4}{e^{cq}\log(cq)} \\ &> me^{-cq}\end{aligned}$$

where we have approximated the slowly varying $4/\log(cq)$ as a constant, m. Note that since protein structures are not expected to exceed values of $q \approx 1000$, we can retain the bound for all proteins by choosing m to be the worst case, specifically $m = 4/\log(100 \times 1000) \approx 0.3$. Now, recalling from the text

$$d_{min} = 2^{Eq} = e^{Eq/\log 2}$$

we obtain

$$\begin{aligned}\lambda_2 &> m(d_{min} + 2 + e^{-c/E}) \\ &> md_{min}\end{aligned}$$

this is our second bound. Note that this bound is not truly exact, since we have presumed the limit $V \to \infty$ and that q is bounded from above. Since our network is very large, and the bound on q holds physically, we expect this bound to be applicable for all proteins considered by this model.

Both upper and lower bounds, linear in d_{min}, suggest that λ_2 should also scale linearly with d_{min}. Numerical simulations, presented in the manuscript, support this linear scaling.

Collapse in MD Simulations of Protein Folding

All-atom molecular dynamics simulations of protein folding indicate that the vast majority of protein conformations are collapsed in the absence of denaturant, even in the unfolded state ensemble.[14,15] By simply counting the frequency of the total number of contacts, it becomes clear that the distribution of the number of contacts is peaked just below its maximal possible value (Figure 1). Further, this peak in the number of contacts is close to the number of contacts in the native state. This directly supports the globular approximation employed in the derivation of this simple model.

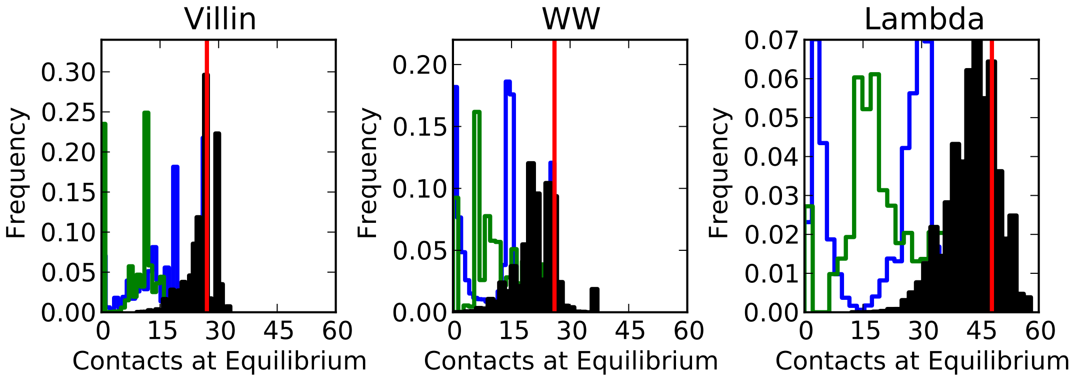

Figure 1: Distributions of contacts in converged MSMs of protein folding, for three different systems.[1–3] Each system was simulated in explicit water. Three histograms are shown, (black) total number of contacts, (green) non-native contacts, and (blue) native contacts. Vertical red lines show the number of contacts in the native state.

Linearity of Contacts and Chain Length

A simple plot of the 78 proteins examined experimentally shows clearly that, for two-state folders, the number of contacts is linear in the chain length (Figure 2).

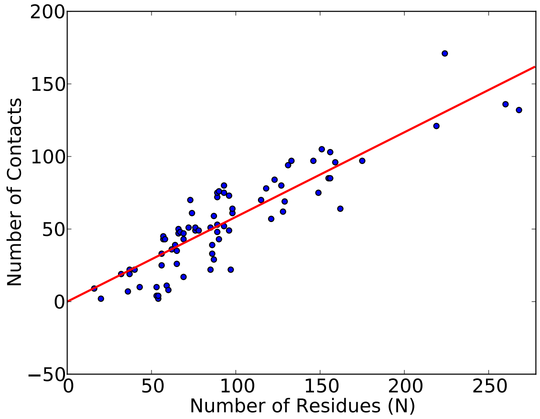

Figure 2: Scaling of the number of residue-residue contacts in the protein native state with chain length. The relationship is linear with a slope of $\rho_N = 0.58$, as determined from a least-squares fit ($R^2 = 0.71$). Data is the 78 proteins from Ref.[16]

Estimation of Energetic Parameters

In the main text, we report the estimates $\varepsilon_x \approx -1$ $\varepsilon_N \approx -3$, $\varepsilon_{NN} \approx -2$, and $\sigma \approx 1$. These values were estimated from three independent considerations: simple thermodynamic concerns, all-atom simulations, and literature reports of similar values from other simple models and calorimetric experiments. Below, we detail our investigations into each of these sources and justify the final reported ranges.

Estimation from Expected Thermodynamics

A very simple evaluation of ε_x can be quickly obtained from a back of the envelope calculation. We preform this quickly here and use it to check the consistency of the more advanced methods that follow.

Consider an archetypical 100 residue protein. Proteins of this size have thermodynamic stabilities of approximately $\Delta F_{fold} = 10$ kcal mol^{-1}. From the previous section, we estimate that the native state will have about $q = 60$ contacts. Assuming that ΔF_{fold} corresponds, under the globular approximation, to taking the protein from 60 non-native to 60 native contacts, we can write

$$\begin{aligned}\Delta F_{fold} &= \phi_N q - \phi_{NN} q \\ &= \phi_x q\end{aligned}$$

inserting our estimates for q and ΔF_{fold}, we obtain $\phi_x \approx 0.18$ kcal mol^{-1}.

Estimation from Simulation

We have estimated the effective potential of interaction for native and non-native contacts from three well-studied Markov models parametrized from all-atom MD with explicit water.[1–3] Each state i of the Markov model has an equilibrium probability $P_{eq}(i)$ and a number of native contacts n_i and non-native contacts q_i. The states are small enough that n_i and q_i vary negligibly within a state. This allows us to write down the fraction of contacts in the system that are native, non-native, or broken at equilibrium

$$\begin{aligned}x_N &= \sum_i n_i P_{eq}(i)/n_{max} \\ x_{NN} &= \sum_i q_i P_{eq}(i)/q_{max} \\ 1 &= x_N + x_{NN} + x_L\end{aligned}$$

where each x represents the fraction of contacts that are in a native contact, non-native contact, or in a loop (no contact) at equilibrium. Here, n_{max} and q_{max} denote the maximum possible native and non-native contacts that can form, respectively, and were estimated from that states in the simulation that had the most contacts.

Recall we defined ε as corresponding to the average energy associated with the reaction

$$R(\text{aq}) \rightarrow \frac{1}{2}R_2(\text{protein}) \quad \Delta F/kT = \varepsilon + \sigma$$

so ε represents the potential gained in going from a loop conformation to a contact-formed conformation. We can define a convenient reference free energy by setting the free energy of the a loop conformation, ε_L, to be 0. Te remaining energy values can then be calculated by Boltzmann factors relative to this energy. Since $\exp(-\varepsilon_L) = \exp(0) = 1$, this allows us to easily find the effective energies of non-native and native contacts

$$\begin{aligned} \frac{x_i}{x_L} &= \frac{\exp(-\varepsilon_i - \sigma)}{\exp(-\varepsilon_L)} \\ \varepsilon_i - \sigma &= -\log(x_i/x_L) \end{aligned}$$

where we have written $i \in \{N, NN\}$ to concisely represent either native or non-native contacts. The results of this procedure are documented in Table 1.

Table 1: Parameters Estimated From Simulation

Protein	N	Contacts	T (K)	x_N	x_{NN}	x_L	$\phi_N - \sigma$	$\phi_{NN} - \sigma$	ϕ_x
Villin	35	28	300	0.59	0.4	0.01	-2.40	-2.17	-0.23
WW (FiP35)	35	27	395	0.42	0.31	0.27	-0.33	-0.09	-0.24
Lambda	80	89	370	0.44	0.41	0.15	-0.85	-0.78	-0.07

Units of ϕ are kcal mol^{-1}
Value of sigma is estimated to be 1 to 2 at 300K

There are a number of difficulties with this procedure. First, it implicitly assumes that contact formation is independent, contact energies between distinct residues are pairwise additive, and that the errors due to MSM construction and contact definition are small. In particular, we noticed that small changes in the contact definition (specifically the distance cutoff) could yield significant changes in the absolute values of ϕ_N and ϕ_{NN}, but that the relative value ϕ_x was much more robust. This is apparent from the direct comparison between the values obtained for separate proteins. A further issue here is the small sample size, and the bias of the sample towards very small proteins. At the time of publication, this data represented all the MSMs of protein folding in explicit water available.

Despite these potential sources of error, for the proteins studied here, we obtain a value for ϕ_x very similar to what we might expect from our back of the envelope calculation above, indicating that this value is indeed $\mathscr{O}(10^{-1})$. This level of precision is probably all that is needed for the model presented, considering its approximate nature. To complete the model, however, we need accurate estimates of σ and either ϕ_N or ϕ_{NN}.

Estimations in the Literature

Other simple models of protein folding, as well as very careful calorimetric experiments, provide a third way to estimate energetic values. The entropy of loop formation, in particular, is robustly determined by calorimetric experiments to be in the range of 1 to 7 entropy units, depending on the protein. It is also possible to estimate the effective value of ϕ_N from the frequencies of contacts found in the PDB, following the classic work of Miyazawa and Jernigan. Estimation of non-native contact energies are very difficult experimentally, and we found no reports systematically evaluating these energies in the literature. Other authors of simple models have estimated the excess energy, ε_x, however, and these provide an indirect route to non-native energies.

Table 2 reports the values corresponding to our thermodynamic contact potentials found in the literature, along with our estimates from all-atom simulation. Below, we briefly detail where each estimate comes from, and any processing or interpretation we had to perform.

Table 2: Parameters from Literature

Method	ϕ_x	ϕ_N	ϕ_{NN}	ΔS_{loop}	Comments
All-atom Simulation[1–3]	-0.2	-1.2	-1.0	-	Averages of Villin, WW, Lambda
Calorimetry[17]	-	-1.5 ± 1.5	-	4.0 ± 3.0	Experiment
PDB Statistics[18,19]	-	-1.5	-	-	MJ Potential Statistics
Levinthal Estimate[20]	-1.0	-	-	-	Simple non-protein like model
Monte Carlo Algo.[21]	-3.0	-	-	-	Protein-like simple simulation
Ising-Like Model[22]	-0.6	-	-	3.7	Fit parameters to experiment

Units of ϕ are kcal mol^{-1}
Units of ΔS are cal mol^{-1} K^{-1}

Calorimetry. Makhatadze and Privalov have provided a detailed summary of calorimetric experiments detailing the energetics of protein contact interactions.[17] By surveying a large sample of protein data, they have compiled reliable estimates for native contact enthalpies the the entropies.

These values represent our most confident estimates for energetics in the model.

PDB Statistics. In their classical work, Miyazawa and Jernigan have calculated effective contact potentials for each residue-residue pair based on the statistics of the PDB, generating a potential for use in various simple models.[18,19] By averaging the frequency of residue-residue contacts report with the energies of each of those contacts, we obtained an averaged native contact energy of -3.5 kT, when compared to random contact formation. Taking these random interactions as a zero of energy, we obtain an effective native contact potential.

Levinthal Estimate. In 1992 Zwanzig, Szabo, and Bagchi showed that a simple bias in energy towards native contacts over non-native contacts was enough to dramatically increase the rate of folding.[20] While their considerations were very simple, inspired by the canonical monkeys-with-typewriters problem, and did not contain any explicit protein-like features, their argument is a form of the idea of statistical persistence derived in our model. They predicted that a native energy bias of 2 kT per contact would be sufficient to lead to biologically relevant folding times.

Monte Carlo Algorithm. Linse and Linse have reported a simple Monte-Carlo algorithm that models protein folding in a manner very similar to the analytical model presented here.[21] Similar to the model of Zwanzig, they model native contacts as slightly more attractive than non-native contacts, and show that this leads to rapid folding in numerical simulations. While they investigate a range of different contact energies, a native bias of -11 kJ mol^{-1}, along with a small cooperativity energy, is reported as their most realistic value and used for many of the reported results.

Ising-Like Model. Eaton and Munoz have reported many results on an Ising-like model of protein folding, with parameters for the energy (-0.63 kcal mol^{-1}) and entropy (3.7 cal mol^{-1} K^{-1}) of native contact formation obtained by fitting to experimental data.[22–24]

Variation in the energetic estimates between these diverse techniques is fairly small, and there does appear to be some consensus, allowing us to estimate these energetic parameters with some confidence.

References

(1) Bowman, G. R.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, 12–15.

(2) Voelz, V. a.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526–8.

(3) Lane, T. J.; Bowman, G. R.; Beauchamp, K.; Voelz, V. A.; Pande, V. S. *J. Am. Chem. Soc.* **2011**, *133*, 18413–9.

(4) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comp.* **2011**, *7*, 3412–3419.

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

(6) Buchete, N.-V.; Hummer, G. *The J. Phys. Chem. B* **2008**, *112*, 6057–69.

(7) Noé, F.; Fischer, S. *Curr. Op. Struct. Biol.* **2008**, *18*, 154–62.

(8) Prinz, J.-H.; Keller, B.; Noé, F. *Phy. Chem. Chem. Phys.* **2011**, *13*, 16912–27.

(9) Albert, R.; Barabási, A.-L. *Rev. Mod. Phys.* **2002**, *74*, 47–97.

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

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

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

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

(14) Bowman, G. R.; Pande, V. S. *In Preparation* **2011**.

(15) Voelz, V. A.; Singh, V. R.; Wedemeyer, W. J.; Lapidus, L. J.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 4702–9.

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

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

(18) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623–44.

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

(20) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci.* **1992**, *89*, 20.

(21) Linse, S.; Linse, B. *J. Am. Chem. Soc.* **2007**, *129*, 8481–6.

(22) Kubelka, J.; Henry, E. R.; Cellmer, T.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci.* **2008**, *105*, 18655–62.

(23) Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci.* **1999**, *96*, 11311–6.

(24) Godoy-Ruiz, R.; Henry, E. R.; Kubelka, J.; Hofrichter, J.; Muñoz, V.; Sanchez-Ruiz, J. M.; Eaton, W. A. *J. Phys. Chem. B* **2008**, *112*, 5938–49.