Supporting Information

A general mechanism of 2-state protein-folding kinetics

Geoffrey C. Rollins[†] and Ken A. Dill[*,‡]

Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, and Laufer Center for Physical and Quantitative Biology and Departments of Chemistry and Physics and Astronomy, Stony Brook University, Stony Brook, NY

E-mail: dill@laufercenter.org

Simple metric fits

Various scaling laws have been proposed for the dependence of folding rates on protein length. Early on, Thirumalai proposed that the logarithms of folding rates should scale as the square root of chain length based on polymer theory arguments.[1] Later studies have suggested a variety of functional forms for the scaling law.[2–11] Fu and Wang found an $O(2^{4.306n^{2/3}log(n)})$ time algorithm to fold 3D HP lattice models.[12] A recent manuscript from Lane and Pande argues that current available data is insufficient to infer the correct scaling law.[13] We show in Fig. S1 such correlations, using our data set of 93 proteins (detailed in Tables S3,S4).

Fig. S1 shows fits to experimental folding rates for 93 proteins, both two-state and multi-state proteins. For the multi-state proteins, we fitted to the slowest phase. We show fits

[*]To whom correspondence should be addressed
[†]University of California, San Francisco
[‡]Stony Brook University

to some of the presently most prominent metrics: relative contact order (RCO),[14] absolute contact order (ACO),[3–5] chain length (L), and the square root of chain length ($\sqrt{L}$).[6–8,15] We fit the function $log\,(k_f) = log\,(k_0) - ax$, where x is RCO, ACO, $\sqrt{L}$, or L using k_0 and a as adjustable parameters. The fit parameters are shown in Table S1.

The shading in the figures show 95% confidence intervals, which we obtain by bootstrapping[16,17] the data (resampling with replacement). Because $\sqrt{L}$ gives a reasonable fit to the data, and because the principal difference between RCO and ACO is simply that the latter contains the chain length,[3–5,7] a key conclusion is the importance of the chain length in predicting protein folding rates, more important than the topology, *per se*.

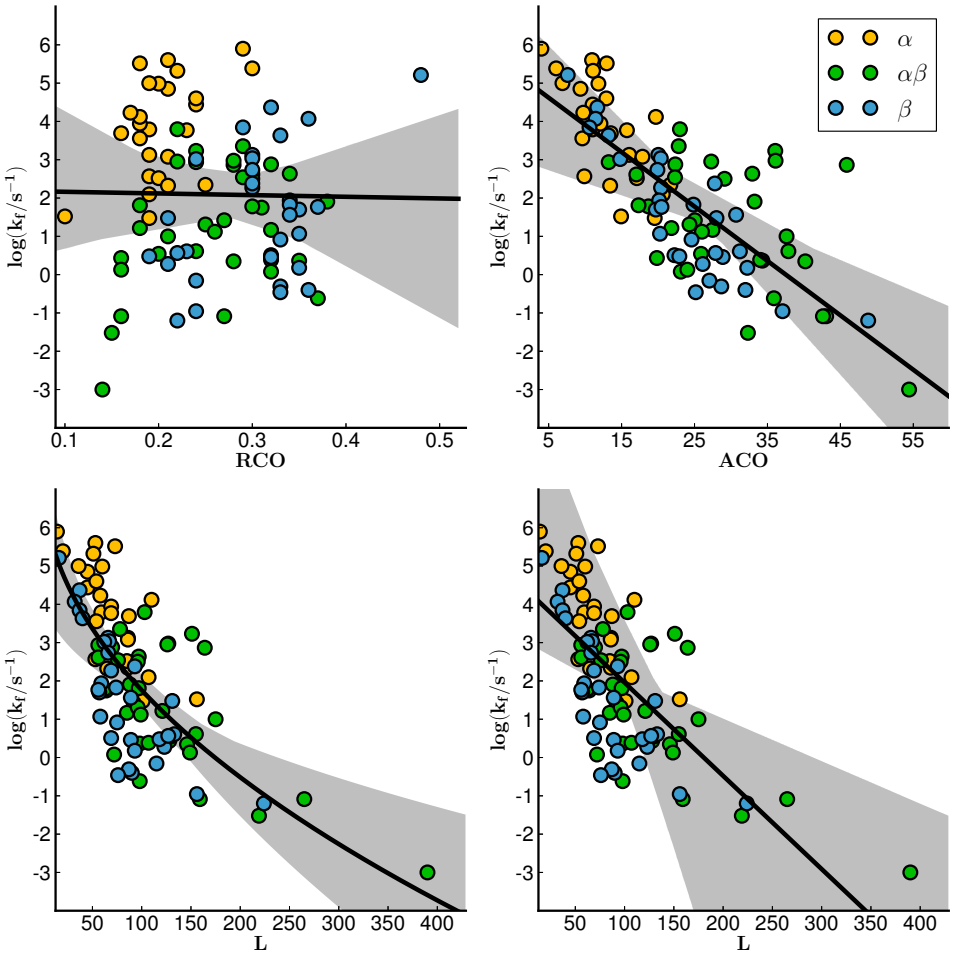

Figure S1: Folding rates vs. simple metrics (RCO, ACO, $\sqrt{L}$, and L). Proteins are colored based on structural class. The black line is the fit to the data. The gray bands represent the 95% confidence interval. Fit parameters are tabulated in Table S1.

The kinetic model

Our kinetic model is a "one-step" continuous time Markov process,[18] a process that consists of hops between adjacent sites along a 1D lattice. Each lattice site is labeled with an integer. In our folding model, we refer to the lattice sites as "states" since they correspond to configurational states of the protein. Our model has $N+1$ states, where N is the number of secondary structures in the protein. The states span the integer range $c \in [0, 1, 2, \ldots, N-1, N]$, where c represents the number of folded secondary structures.

The rate constants for forward (v) and reverse (r) hopping are

$$v_c = (N - c)\, k_1 \tag{S1}$$

$$r_{c+1} = \frac{(c+1)\, k_1}{K_2 K_3^{n_{c+1}-n_c}} \tag{S2}$$

where k_1 is a rate constant for the folding of an isolated secondary structure and $N - c$ represents the number of secondary structures still waiting to fold, given that c have already folded. The rate constant for reverse hopping, r_{c+1}, represents unfolding a secondary structure, and we derive it from detailed balance:

$$w(c)v_c = w(c+1)r_{c+1} \tag{S3}$$

$$r_{c+1} = v_c \frac{w(c)}{w(c+1)} \tag{S4}$$

Where $w(c)$ and $w(c+1)$ are the Boltzmann weights of states c and $c+1$, respectively. Substituting equation 1 from the main text for $w(c)$ and $w(c+1)$, we get

$$r_{c+1} = v_c \frac{\frac{N!}{c!(N-c)!} K_2^c K_3^{n_c}}{\frac{N!}{(c+1)!(N-(c+1))!} K_2^{c+1} K_3^{n_{c+1}}} \tag{S5}$$

$$= v_c \left[\frac{(c+1)c!(N-c-1)!}{c!(N-c)(N-c-1)!}\right] \left[\frac{1}{K_2 K_3^{n_{c+1}-n_c}}\right] \tag{S6}$$

where in the second step we divided through and rewrote $(c+1)!$ as $(c+1)c!$ and $(N-(c+1))!$

as $(N-c)\,(N-c-1)!$. Several combinatoric terms cancel, leaving:

$$r_{c+1} = v_c \left[\frac{c+1}{N-c}\right] \left[\frac{1}{K_2 K_3^{n_{c+1}-n_c}}\right] \tag{S7}$$

Finally, we replace v_c with equation S1 to get the result shown above in equation S2:

$$r_{c+1} = (N-c)\, k_1 \left[\frac{c+1}{N-c}\right] \left[\frac{1}{K_2 K_3^{n_{c+1}-n_c}}\right] \tag{S8}$$

$$= \frac{(c+1)k_1}{K_2 K_3^{n_{c+1}-n_c}} \tag{S9}$$

Escape from the folded state has an additional factor, K_f, which stabilizes the folded state:

$$r_N = \frac{N k_1}{K_2 K_3^{n_N - n_{N-1}} K_f} \tag{S10}$$

which we can rewrite in terms of Q_F, the folded partition function

$$r_N = k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} \tag{S11}$$

Given the forward and reverse rate constants, we write the kinetics of our model as a master equation:

$$\frac{dp_c}{dt} = r_{c+1} p_{c+1} + v_{c-1} p_{c-1} - (r_c + v_c) p_c \tag{S12}$$

Analytical expressions for the folding and unfolding rates

To compute the folding and unfolding rates, we follow Zwanzig[19] and posit that the rate-limiting step is the transition from the "first-excited state" ($c = N-1$) to the folded state ($c = N$). Based on Eqn. S12, we can write the rate of change of the population of the folded

state as

$$\frac{\mathrm{d}p_N}{\mathrm{d}t} = v_{N-1}p_{N-1} - r_N p_N \tag{S13}$$

When we substitute Eqn. S1 for v_{N-1} and Eqn. S11 for r_N, we get

$$\frac{\mathrm{d}p_N}{\mathrm{d}t} = k_1 p_{N-1} - k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} p_N \tag{S14}$$

We follow Zwanzig's local thermodynamic equilibrium (LTE) approximation that says, if the highest barrier is $c = N - 1$, all the states $c < N$ rapidly equilibrate, conditional on $p_N(t)$,

$$p_c(t) = \frac{w(c)}{Q_U} \left(1 - p_N\left(t\right)\right) \quad \text{for } c < N \tag{S15}$$

We substitute Eqn. S15 for p_{N-1} in Eqn. S14 to get

$$\frac{\mathrm{d}p_N}{\mathrm{d}t} = k_1 \frac{w(N-1)}{Q_U} \left(1 - p_N\left(t\right)\right) - k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} p_N \tag{S16}$$

$$= k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_U} \left(1 - p_N\left(t\right)\right) - k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} p_N \tag{S17}$$

This shows that the rate of change of the folded state population is a competition between a rate of gain, $k_1 N K_2^{N-1} K_3^{n_{N-1}} Q_U^{-1}$, and a rate of loss, $k_1 N K_2^{N-1} K_3^{n_{N-1}} Q_F^{-1}$. This is analogous to a two-state folding reaction $U \rightleftarrows F$ in which

$$\frac{\mathrm{d}[F]}{\mathrm{d}t} = k_f[U] - k_u[F] \tag{S18}$$

Comparing Eqn. S18 with Eqn. S17, we see that k_f is our rate of gain and k_u is our rate of loss:

$$k_f = k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_U} \tag{S19}$$

$$k_u = k_1 N \frac{K_2^{N-1} K_3^{n_{N-1}}}{Q_F} \tag{S20}$$

This result is shown as Eqn. 10 and Eqn. 11 in the main text.

A more general expression for k_f would be:

$$k_f = k_1 \frac{w(c^\ddagger)}{Q_D}, \tag{S21}$$

where $w(c^\ddagger)$ is the weight of the state at the top of the barrier and Q_D is the partition function for states $c = 0, 1, \ldots, c^\ddagger$ (the left side of the barrier).

Numerical integration of master equation

The master equation (Eqn. S12) can be written in matrix form as

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{A} \tag{S22}$$

where $\mathbf{P}(t)$ is a vector of the state probabilities at time t and $\mathbf{A}$ is the transition rate matrix. The matrix elements of $\mathbf{A}$ are: Each off-diagonal element $\mathbf{A}_{ij}$ represents the transition rate from state i to state j, but only adjacent states (that differ by one secondary structure) have non-zero transition rates. Each diagonal element $\mathbf{A}_{ii}$ is defined so that the rows sum to zero (satisfying the condition that total probability density should be conserved).

$$\mathbf{A} = \begin{pmatrix} -v_0 & v_0 & 0 & \cdots & 0 & 0 & 0 \\ r_1 & -(r_1+v_1) & v_1 & \cdots & 0 & 0 & 0 \\ 0 & r_2 & -(r_2+v_2) & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(r_{N-2}+v_{N-2}) & v_{N-2} & 0 \\ 0 & 0 & 0 & \cdots & r_{N-1} & -(r_{N-1}+v_{N-1}) & v_{N-1} \\ 0 & 0 & 0 & \cdots & 0 & r_N & -r_N \end{pmatrix} \tag{S23}$$

We can numerically solve equation S22 to get the state populations at time t based on the populations at $t = 0$:

$$\mathbf{P}(t) = \mathbf{P}(0)\ e^{\mathbf{A}t} \tag{S24}$$

We compute state probabilities as a function of time using equation S24 with initially all of the probability density localized to the fully unfolded state ($p_0(0) = 1$, $p_i(0) = 0$ for $i > 0$). We plot the results of one such calculation in Fig. 5 in the main text. Then, to get the folding rate, we compute the rate spectrum using the *ratespec* python package of Voelz and Pande.[20] Given a time trace, *ratespec* calculates the rate spectrum using regularized linear regression.

Eigendecomposition of rate matrix

We can also compute a folding rate from the eigen values of our rate matrix, $\mathbf{A}$. We diagonalize $\mathbf{A}$ as follows:

$$\mathbf{A} = \mathbf{B}\mathbf{D}\mathbf{B}^{-1} \tag{S25}$$

where $\mathbf{B}$ is a matrix of the eigenvectors of $\mathbf{A}$ and $\mathbf{D}$ is a diagonal matrix with the eigenvalues of $\mathbf{A}$ along the diagonal. The smallest eigenvalue is $\lambda_1 = 0$, and its corresponding eigenvector represents the equilibrium populations of the states of the model. The folding rate is obtained

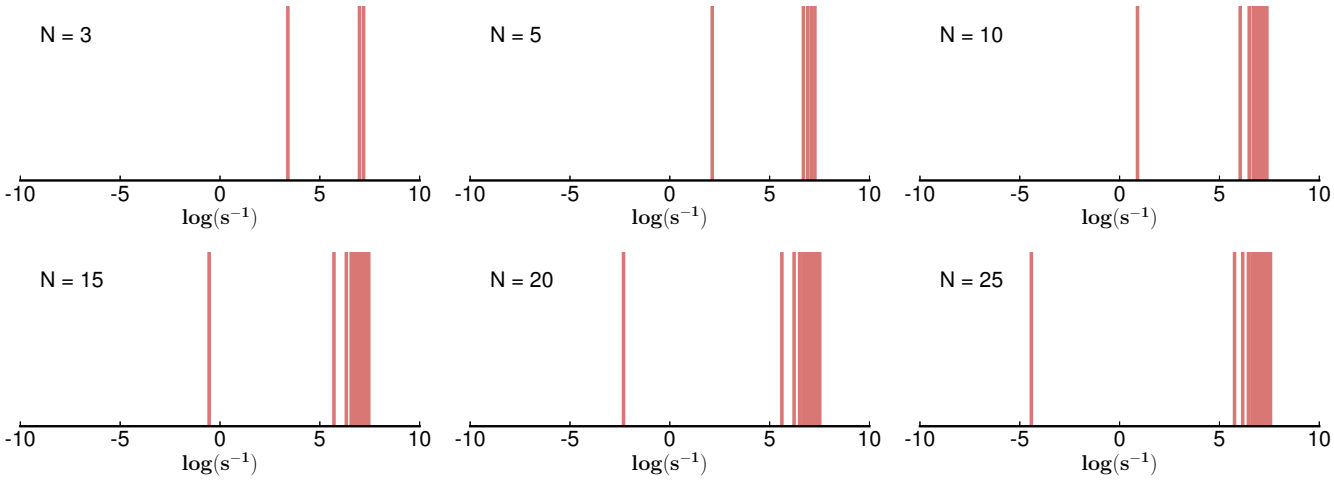

Figure S2: Eigen spectra for different values of N. At each value of N, we see two-state folding. There is a clear separation of time scales between the folding mode and faster modes.

from the smallest non-zero eigenvalue, $k_f = -\lambda_2$. The larger eigenvalues represent dynamics occuring on faster timescales. We observed two-state folding in our model: there was a clear separation of timescales (Fig. S2).

Numerical validation of analytical expression for folding rate

Here we show that our analytical expression for the folding rate (Eqn. S19) is consistent with the results of numerical simulations of our master equation, as well as eigen decomposition of the rate matrix. In Fig. S3, we plot the results from computing k_f via the three different approaches: (1) from Eqn. S19 (gray), (2) from numerical integration of master equation (Eqn. S24), and (3) from the smallest non-zero eigenvalue. The overlap between the three methods is very good. We find that the analytical expression for k_f is a good approximation for the more exact numerical evaluations of k_f.

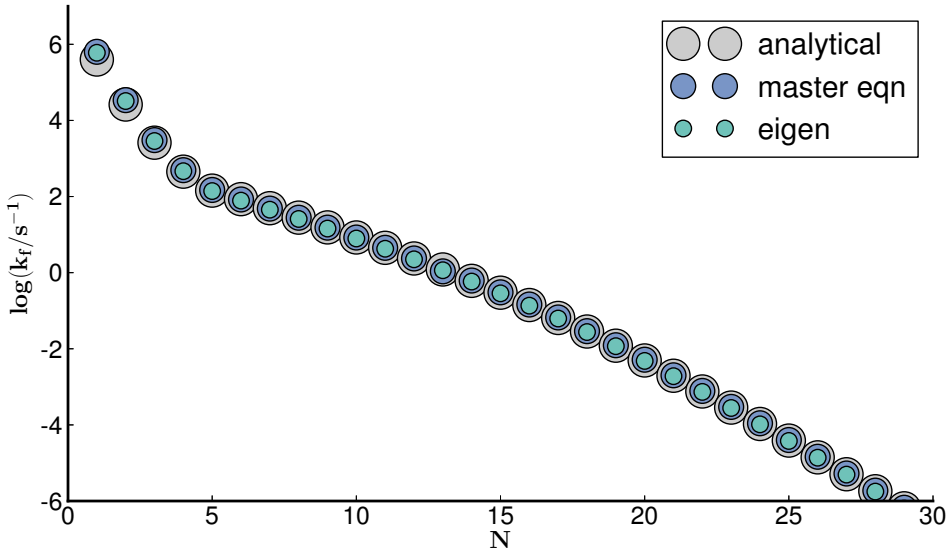

Figure S3: Comparison of three methods for computing $log(k_f)$ from our model. The agreement between the three methods is very good. We use different dot sizes so that the results from all three methods can be seen.

Fit to experimental data

Fig. S4 shows folding rates predicted from the model vs. experimental folding rates.

Uniqueness of fit to kinetic data

From our bootstrap fitting, we observed a strong inverse correlation between the best-fit values of $log(K_3)$ and $log(K_2)$ (Fig. S5A). The relationship is $log(K_3) = -0.33 * log(K_2) - 0.18$. To address the uniqueness of our fit to the folding rate data, we've plotted the RMS error of the fit vs. $log(K_2)$ and $log(K_3)$, in the vicinity of the optimal values for these parameters (Fig. S5B). The lowest contour represents RMS error of 1.5 or less, and the region of parameter space covered by this contour is roughly comparable to the 95% confidence intervals determined via bootstrapping. During the bootstrap fitting, we found that–although the values of K_2 and K_3 were randomized initially and unconstrained during the fitting procedure–the parameters always converged to the narrow region shown here.

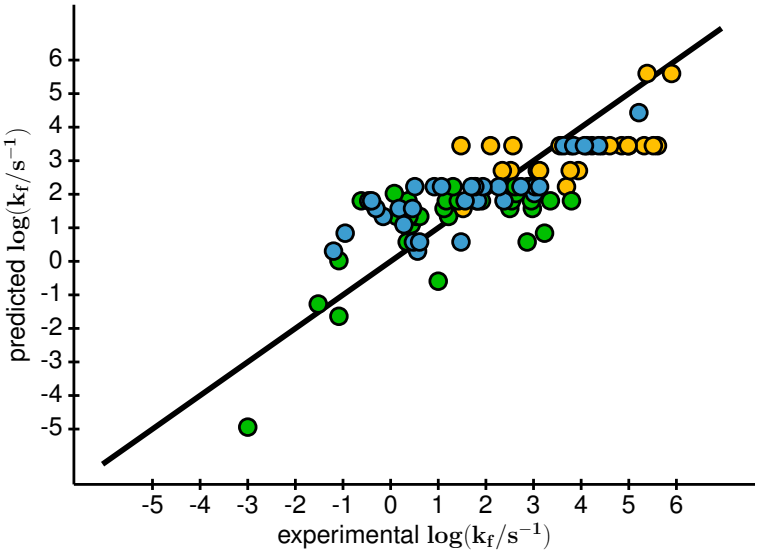

Figure S4: Predicted folding rate vs. experimental folding rate. The colored points are experimental values, and they are colored by structural class. The black line is the prediction from the model. Fit parameters (95% CI): $K_2 = 0.037$ $(0.025, 0.058)$, $K_3 = 1.96$ $(1.67, 2.23)$. We fixed $k_1 = 10^{5.6} s^{-1}$, and K_f was fitted to an equilibrium stability model, independent of the folding rate fit. Fit quality (95% CI): $R^2 = 0.63$ $(0.49, 0.72)$, rms error $= 1.30$ $(0.96, 1.65)$.

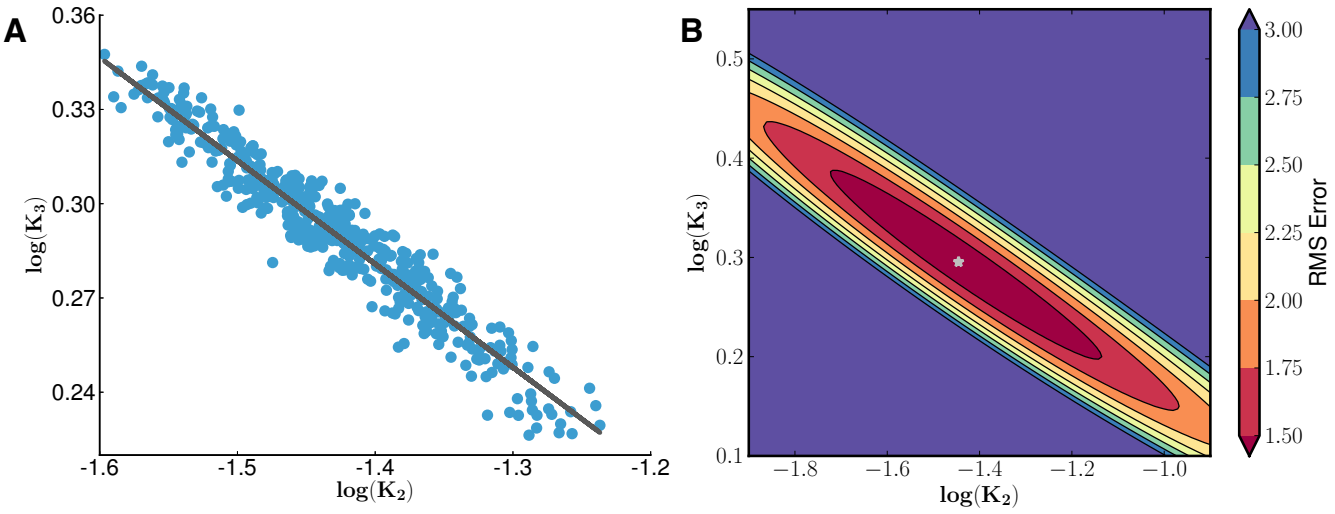

Figure S5: (A) Results of bootstrap fitting. Each data point represents the value of $log(K_3)$ and $log(K_2)$ determined during one bootstrap iteration. The gray line is the result of linear regression ($R^2 = 0.92$). The fit line represents: $log(K_3) = -0.33 * log(K_2) - 0.18$. (B) Contours in RMS error space for the global fit to the folding rate data. The best fit region occupies a relatively narrow band of parameter space.

Fitting parameter K_f to protein stability model

We used the linear relationship between chain length (L) and number of secondary structures (N) (main text Fig. 6) to compute an average chain length (L_{fit}) at each N. Then, we used the protein stability model of Dill and Ghosh[21,22] to fit K_f for each N (for each L_{fit}). The Dill & Ghosh model predicts stability as a function of L and temperature. We set $T = 300K$. The fit values are tabulated below (Table S2).

Comparison with Zwanzig model

Here, for ease of comparison, we explain how the parameters in our model map on to the original Zwanzig model.[19] We've split Zwanzig's $K = \nu e^{-\beta U}$ into our two equilibrium constants, K_2 and K_3. We don't break the equilibrium constants down into enthalpic and entropic contributions, as Zwanzig does with his parameters U and ν, respectively. Zwanzig's ν represents the number of incorrect configurations per residue. This chain entropy is one of the components of our parameters, K_2 and K_3. Zwanzig's stability gap $e^{-\beta\epsilon}$ corresponds to our K_f. Our kinetic rate constant k_1 is the same as Zwanzig's. Our order parameter c represents the number of correct secondary structures; Zwanzig's order parameter was S, the number of incorrect residues.

Equilibrium temperature dependence of model

Additional parameterization would be required for the current model to capture the temperature dependence of folding. Here, we demonstrate a possible way of doing so. We split the equilibrium constants into enthalpic and entropic components (for simplicity, we leave K_f unsplit):

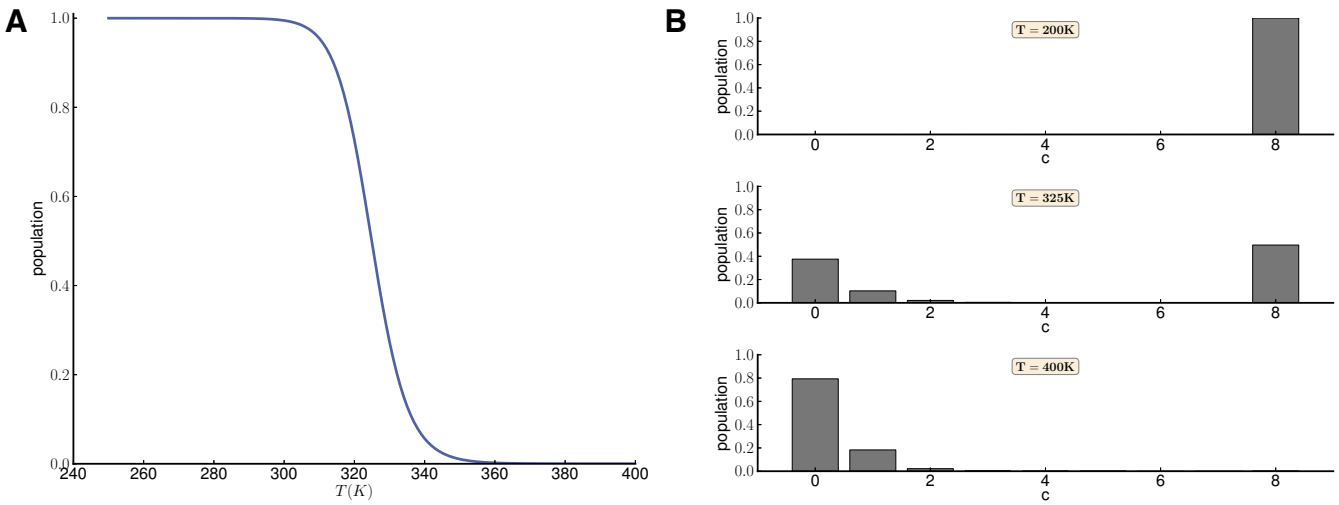

Figure S6: (A) Population of the folded state vs. temperature. (B) State populations at three temperatures. From top to bottom, they represent: below the midpoint, near the midpoint, and above the midpoint. Parameters: $N = 8$, $H_2 = -0.6$, $H_3 = -1.2$, $S_2 = -8.6$, $S_3 = -2.6$, $G_f = -10$ (with H's and G's in $kcal\ mol^{-1}$, S's in $cal\ mol^{-1}\ K^{-1}$). These parameters correspond to the equilibrium constants presented in the main text at $T = 300\ K$: $log_{10}(K_2) = -1.4$, $log_{10}(K_3) = 0.29$, and $log_{10}(K_f) = 7.46$.

$$K_2 = e^{-(H_2 - TS_2)\beta} \tag{S26}$$

$$K_3 = e^{-(H_3 - TS_3)\beta} \tag{S27}$$

$$K_f = e^{-G_f\beta} \tag{S28}$$

As a proof of concept, we've computed equilibrium state populations as a function of temperature for this modified version of our model. For this example, we used the following parameter values: $N = 8$, $H_2 = -0.6$, $H_3 = -1.2$, $S_2 = -8.6$, $S_3 = -2.6$, $G_f = -10$ (with H's and G's in $kcal\ mol^{-1}$, S's in $cal\ mol^{-1}\ K^{-1}$). These parameters correspond to the equilibrium constants presented in the main text at $T = 300\ K$: $log_{10}(K_2) = -1.4$, $log_{10}(K_3) = 0.29$, and $log_{10}(K_f) = 7.46$. We selected the ΔH and ΔS values here; they were not obtained by fitting to additional data. Our intent with this example is just to demonstrate a way of extending the model, rather than fit to a specific protein.

We plot the results in Fig. S6. Panel A shows the population of the folded state ($c = 8$) vs. T, and panel B shows the populations of all states ($c = 0, 1, \ldots, 8$) at three temperatures: below the midpoint, near the midpoint, and above the midpoint. Fig. S6A shows that the population of F undergoes a sharp unfolding transition, and Fig. S6B shows that near the midpoint, most of the partly folded states have low populations.

This section was just a test that illustrates the model gives 2-state thermal cooperativity with these parameters. We have not explored parameter space in a detailed way or attempted to fit the temperature dependence data of specific proteins.

Exploring parameter space

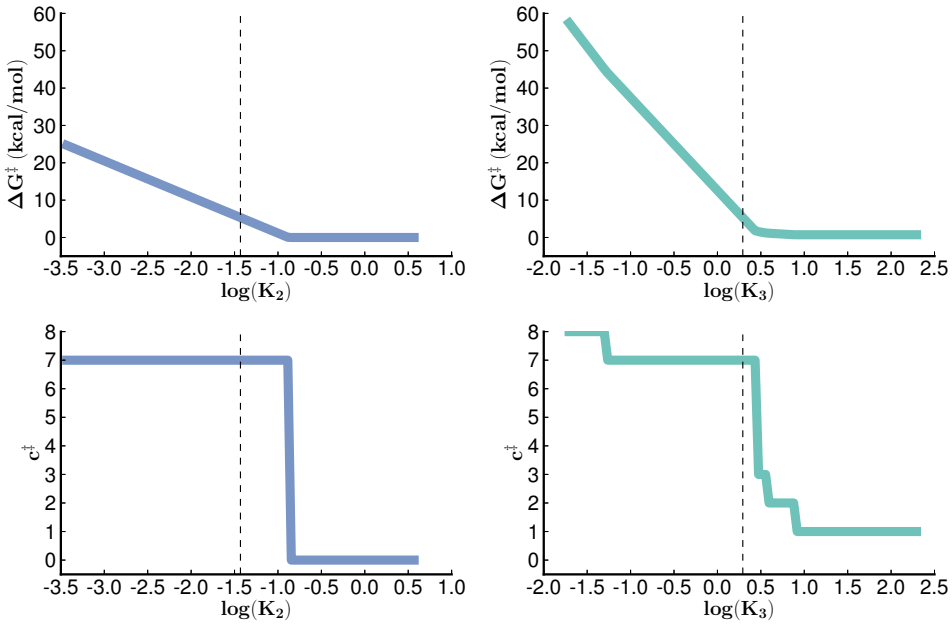

Figure S7: Barrier height (top) and barrier location (bottom) as a function of $log(K_2)$ (left panels) and $log(K_3)$ (right panels). The black dashed lines indicate the best-fit values from the main text. Parameters: $N = 8$, $log(K_f) = 7.46$. On the left, $log(K_3)$ is fixed to 0.29. On the right, $log(K_2)$ is fixed to -1.4.

Fig. S7 addresses the question of how the barrier changes with respect to K_2 and K_3. Fig. S7 shows the barrier height and barrier location as a function of $log(K_2)$ and $log(K_3)$ for a protein with $N = 8$ secondary structures. In the left panels, we fixed $log(K_3)$ to 0.29,

its best-fit value from the main text. In the right panels, we fixed $log(K_2)$ fixed to its best-fit value: $log(K_2) = -1.4$.

From looking at the top-left panel of Fig. S7, we see that barrier height increases as secondary structure formation become less favorable (left of the dashed line). As secondary structure formation becomes more favorable (right of the dashed line), the barrier decreases until the fully unfolded state ($c = 0$) becomes the state with the highest free energy (essentially, a downhill folding situation).

We see a similar pattern in the top-right panel for $log(K_3)$: barrier height increases as tertiary interactions become less favorable and decreases as tertiary interactions become more favorable. The barrier decreases until the fully unfolded state ($c = 1$) becomes the state with the highest free energy.

The bottom panels of Fig. S7 show that the location of the barrier has a non-linear relationship with the parameters K_2 and K_3. The barrier is located at $N - 1$ for our best-fit values, but at a certain point in parameter space, $c = 0$ or $c = 1$ becomes the state with the highest free energy. The parameter regions in which $N - 1$ is not the barrier also correspond to parameter values that give poor fits to the folding rate data.

Proteins in data set

This set of proteins is largely identical to the set used by Ouyang and Liang.[7] To that set, we've added some additional two-state proteins from the data set of Zou and Ozkan,[23] as well as the spectrins R15, R16, and R17,[24] the homeodomain, Pit1,[25] and the L9 helix characterized by Mukherjee et al.[26]

Table S1: Simple metric fit parameters and fit quality. Values in parentheses represent 95% confidence intervals from bootstrapping.

Model	R^2	rmse	$log(k_0)$	a
RCO	0.00 $(0.00, 0.08)$	3.46 $(2.51, 4.22)$	2.21 $(0.53, 4.09)$	0.43 $(-5.56, 7.40)$
ACO	0.59 $(0.41, 0.75)$	1.40 $(0.96, 1.88)$	5.33 $(4.60, 6.03)$	0.14 $(0.11, 0.17)$
L	0.48 $(0.32, 0.61)$	1.82 $(1.39, 8.51)$	4.39 $(3.35, 9.64)$	0.02 $(0.01, 0.07)$
$\sqrt{L}$	0.53 $(0.35, 0.66)$	1.62 $(1.19, 2.03)$	7.24 $(6.33, 8.16)$	0.55 $(0.46, 0.65)$

Table S2: Fit values of model parameter K_f as a function of N at $T = 300K$.

N	L_{fit}	$log\,(K_f)$
1	14	1.75
2	28	3.21
3	41	4.37
4	55	5.23
5	69	5.79
6	83	6.35
7	97	6.90
8	111	7.46
9	124	8.02
10	138	8.58
11	152	9.15
12	166	9.71
13	180	10.27
14	194	10.84
15	207	11.41
16	221	11.99
17	235	12.56
18	249	13.15
19	263	13.74
20	277	14.33
21	290	14.94
22	304	15.55
23	318	16.18
24	332	16.81
25	346	17.46
26	360	18.11
27	373	18.77
28	387	19.44
29	401	20.11

Table S3: List of two-state proteins in data set.

Name	PDB	Structure	Length	N	$log(k_f)$
a3D	2A3D	α	73	3	5.52
Abp1 SH3	1JO8	β	58	5	1.07
AcP	1APS	$\alpha\beta$	98	7	−0.62
AcP common	2ACY	$\alpha\beta$	98	7	0.36
ADA2h	1O6X	$\alpha\beta$	70	5	2.88
Albumin bd	1PRB	α	53	3	5.60
bACBP	2ABD	α	86	4	2.35
BBL	2WXC	α	45	3	4.85
Bc Csp	1C9O	β	66	5	3.13
c myb	1FEX	α	59	3	3.79
CheW	1K0S	$\alpha\beta$	151	11	3.23
CI2	2CI2	$\alpha\beta$	64	6	1.75
CspA	1MJC	β	69	5	2.27
CspB	1CSP	β	67	5	3.04
Cyclophilin A	1LOP	$\alpha\beta$	164	12	2.87
CTL9	1DIVC	$\alpha\beta$	92	7	1.42
E3BD WW	1W4E	α	45	3	4.44
EC298	1RYK	α	69	4	3.94
FBP WW	1E0L	β	37	3	4.37
FKBP12	1FKB	$\alpha\beta$	107	9	0.39
FNfn9	1FNF9	β	90	7	−0.40
fyn SH3	1SHF	β	59	5	1.94
hbLBD BCKD	1K8M	β	87	8	−0.31
HPr	1HDN	$\alpha\beta$	85	7	1.17
Im7	1AYI	α	86	4	3.13
Im9	1IMQ	α	86	4	3.08
L23	1N88	$\alpha\beta$	96	5	1.31
L9 helix	n/a	α	14	1	5.90
lambda	1LMB	α	87	5	3.69
MerP	2HQI	$\alpha\beta$	72	6	0.08
NTL9	1DIVN	$\alpha\beta$	56	5	2.94
P13	1QTU	β	115	9	−0.16
PI3 SH3	1PKS	β	76	7	−0.46
POB	1W4J	α	51	3	5.32
protA Y15W	1SS1	α	60	3	4.99
protG	1PGB	$\alpha\beta$	56	5	2.62
protG hairpin	1PGBb	β	16	2	5.21
protL	2PTL	$\alpha\beta$	62	5	1.78
PsaE	1PSE	β	69	5	0.51
R15	R15	α	110	3	4.12

Continued

Name	PDB	Structure	Length	N	$log(k_f)$
R16	R16	α	107	3	2.10
R17	R17	α	101	3	1.48
RafRBD	1RFA	$\alpha\beta$	78	7	3.35
Rap1	1IDY	α	54	3	3.56
S6	1RIS	$\alpha\beta$	97	6	2.64
Sho1 SH3	2VKN	β	75	5	0.92
spectrin SH3	1SHG	β	57	5	1.70
src SH2	1SPR	$\alpha\beta$	103	7	3.80
src SH3	1FMK	β	56	5	1.77
sso7d	1SSO	β	62	6	3.02
Tendamistat	2AIT	β	74	7	1.83
Tm1083	1J5U	$\alpha\beta$	127	8	2.98
Tm Csp	1G6P	β	66	5	2.74
TNfn3	1TEN	β	89	8	0.46
Trf1	1BA5	α	53	3	2.57
TrpCage	1L2Y	α	20	1	5.38
Twitchin	1WIT	β	93	8	0.18
U1A	1URN	$\alpha\beta$	96	8	2.50
Ubq	1UBQ	$\alpha\beta$	76	7	2.54
Urm1	2QJL	$\alpha\beta$	99	8	1.12
Villin	1VII	α	36	3	5.00
Villin 14T	2VIK	$\alpha\beta$	126	7	2.95
WW pin	1PIN	β	32	3	4.07
WW prototype	1E0M	β	37	3	3.84
WW YAP	1K9Q	β	40	3	3.63

Table S4: List of multi-state proteins in data set.

Name	PDB	Structure	Length	N	$log(k_f)$
AlphaLactAlb	1HMK	$\alpha\beta$	121	9	1.22
ApoPseuAz	1ADW	β	123	10	0.28
BetaLactoGlob	1BEB	β	156	11	-0.95
CheY	3CHY	$\alpha\beta$	128	10	0.43
Colicin E7	1CEI	α	85	4	2.52
CPGK	1PHPc	$\alpha\beta$	219	18	-1.52
CRBP II	1OPA	β	133	12	0.61
Cro	2CRO	α	65	5	2.32
DHFR	1RA9	$\alpha\beta$	159	14	-1.09
EnHD	1ENH	α	54	3	4.60
FF HYPA	1UZC	α	69	4	3.77
FNfn10	1FNF10	β	93	7	2.38
GFP	1B9C	β	224	13	-1.20
GroEL apical	1DK7	$\alpha\beta$	146	12	0.35
HEWL	1HEL	$\alpha\beta$	129	9	0.54
HisActPhil	1HCD	β	118	12	0.48
IFABP rat	1IFC	β	131	12	1.48
ILBP	1EAL	β	127	13	0.56
NHypF	1GXT	$\alpha\beta$	88	7	1.91
NPGK	1PHPn	$\alpha\beta$	175	16	1.00
P16	2A5E	α	156	8	1.52
Pit1	1AU7	α	58	3	4.23
RNase HI	2RN2	$\alpha\beta$	155	9	0.61
StaphNuc	1JOO	$\alpha\beta$	149	9	0.13
Suc1	1SCE	$\alpha\beta$	97	7	1.81
TrypSynthAlpha	1QOPa	$\alpha\beta$	265	19	-1.09
TrypSynthBeta	1QOPb	$\alpha\beta$	390	27	-3.00
Twitchin Ig	1TIT	β	89	7	1.56

References

(1) Thirumalai, D. *J. Phys. I* **1995**, *5*, 1457–1467.

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

(3) Galzitskaya, O.; Garbuzynskiy, S.; Ivankov, D.; Finkelstein, A. *Proteins* **2003**, *51*, 162–166.

(4) Ivankov, D.; Garbuzynskiy, S.; Alm, E.; Plaxco, K.; Baker, D.; Finkelstein, A. *Protein Sci.* **2003**, *12*, 2057–2062.

(5) Ivankov, D.; Finkelstein, A. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 8942–8944.

(6) Naganathan, A.; Munoz, V. *J. Am. Chem. Soc.* **2005**, *127*, 480–481.

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

(8) De Sancho, D.; Doshi, U.; Munoz, V. *J. Am. Chem. Soc.* **2009**, *131*, 2074–2075.

(9) De Sancho, D.; Munoz, V. *Phys. Chem. Chem. Phys.* **2011**, *13*, 17030–17043.

(10) Lane, T. J.; Pande, V. S. *J. Phys. Chem. B* **2012**, *116*, 6764–6774.

(11) Garbuzynskiy, S. O.; Ivankov, D. N.; Bogatyreva, N. S.; Finkelstein, A. V. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 147–150.

(12) Fu, B.; Wang, W. A 2(O)(n(1-1/a) log n) time algorithm for d-dimensional protein folding in the HP-model. Automata, Languages and Programming, Proceedings. 2004; pp 630–644.

(13) Lane, T. J.; Pande, V. S. *PLoS ONE* **2013**, *8*, e78606.

(14) Plaxco, K.; Simons, K.; Baker, D. *J. Mol. Biol.* **1998**, *277*, 985–994.

(15) Chang, L.; Wang, J.; Wang, W. *Phys. Rev. E* **2010**, *82*.

(16) Efron, B. *Ann. Stat.* **1979**, *7*, 1–26.

(17) Efron, B.; Tibshirani, R. *Science* **1991**, *253*, 390–395.

(18) Van Kampen, N. *Stochastic processes in physics and chemistry*, 3rd ed.; Elsevier, 2007.

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

(20) Voelz, V. A.; Pande, V. S. *Proteins* **2012**, *80*, 342–351.

(21) Ghosh, K.; Dill, K. A. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 10649–10654.

(22) Dill, K. A.; Ghosh, K.; Schmit, J. D. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 17876–17882.

(23) Zou, T.; Ozkan, S. B. *Phys. Biol.* **2011**, *8*, 066011.

(24) Wensley, B. G.; Gaertner, M.; Choo, W. X.; Batey, S.; Clarke, J. *J. Mol. Biol.* **2009**, *390*, 1074–1085.

(25) Banachewicz, W.; Religa, T. L.; Schaeffer, R. D.; Daggett, V.; Fersht, A. R. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 5596–5601.

(26) Mukherjee, S.; Chowdhury, P.; Bunagan, M. R.; Gai, F. *J. Phys. Chem. B* **2008**, *112*, 9146–9150.