Protein Folding Rates Correlate with Heterogeneity of Folding Mechanism

B. Öztop,[1,3] M. R. Ejtehadi,[1,2] and S. S. Plotkin[1,*]

[1]*Department of Physics and Astronomy, University of British Columbia, Vancouver, BC V6T-1Z1, Canada*
[2]*Department of Physics, Sharif University of Technology, Tehran 11365-9161, Iran*
[3]*Department of Physics, Bilkent University, Bilkent, Ankara 06800, Turkey*

(Received 14 June 2004; published 12 November 2004)

By observing trends in the folding kinetics of experimental 2-state proteins at their transition midpoints, and by observing trends in the barrier heights of numerous simulations of coarse-grained, C_α model Gō proteins, we show that folding rates correlate with the degree of heterogeneity in the formation of native contacts. Statistically significant correlations are observed between folding rates and measures of heterogeneity inherent in the native topology, as well as between rates and the variance in the distribution of either experimentally measured or simulated ϕ values.

DOI: 10.1103/PhysRevLett.93.208105 PACS numbers: 87.15.Aa, 87.14.Ee, 87.15.He, 87.15.Ya

Protein folding is a relaxation process driven by a first-order-like fluctuation of a critical nucleus [1]. Because proteins are evolutionarily designed to fold to a particular structure, frustrating interactions are minimized and the folding process can be projected onto one or a few reaction coordinates without too much loss of information [2]. This projection yields a free energy surface whose structure is subject to much interest.

What factors determine the height of the folding free energy barrier for the various proteins? As one would expect, the barrier decreases as the energetic stability of the folded structure increases [3]. Moreover, folding rates tend to increase with energetic discrimination measures between the folded state and the unfolded or misfolded decoys [4], indicating the importance of minimizing frustration [2]. As one might also expect, the barrier increases for native structures that have longer polymer loops formed during folding. A property capturing this effect, dubbed absolute contact order (ACO), measures the mean sequence separation between amino acids in close proximity (and thus interacting) in the native structure [5]: ACO $\equiv \overline{\ell} = (1/M)\sum_{i<j}|i-j|\Delta_{ij}^N$, where i and j label amino acid index, $\Delta_{ij}^N = 1$ (or 0) if amino acids i and j are (or are not) interacting in the native structure, and M is the total number of contacts in the native structure determined by either heavy side chain atoms or C_α atoms within a cutoff distance of 4.8 Å [6].

In what follows, we first reexamine the trend of rates with $\overline{\ell}$ in light of theoretical predictions [7,8]; then we will further examine higher-order aspects of native topology (and energetics) that act as predictors of folding rate.

If we take data that first corrects for the effects of differing native stabilities for different proteins by adjusting denaturant concentration to conditions at the transition midpoint, and then plot the log folding rate vs $\overline{\ell}$, we find a statistically significant correlation for a representative set of 19 2-state proteins (and P^{13-14} circular permutant of S6) [Fig. 1(a)] [9]. Observations similar to this led the folding community to accept the idea that properties of native topology strongly determine folding rate [10]. Moreover, if one simulates off-lattice C_α Gō models [6] to 18 structures of known 2-state folders [11], one also finds a statistically significant correlation between barrier height and absolute contact order [Fig. 1(b)]. One also notices from Fig. 1 that there must be more to the story then absolute contact order in determining folding rates, since the fluctuations around the best fit line are significant.

The effects of native topology (and energetics) should be describable analytically as well. To this end a free energy functional approach was developed [7,8] within which it was shown that the free energy barrier may be written in terms of an expansion involving moments of distributions of native contact interaction energies $\{\epsilon_{ij}\}$, and native contact sequence separations $\{\ell_{ij}\} \equiv \{|i-j|\}$. The lowest order corrections to the mean-field barrier are [8]

$$\frac{\Delta F^\ddagger}{MT}(\{\epsilon_{ij}\},\{\ell_{ij}\}) = \frac{\overline{\Delta F}^\ddagger}{MT} - A\frac{\overline{\delta\epsilon^2}}{T^2} - B\frac{\overline{\delta\epsilon\delta\ell}}{\overline{\ell}T} - C\frac{\overline{\delta\ell^2}}{\overline{\ell}^2}, \tag{1}$$

where A, B, C are all positive and of order unity. The lowest order mean-field term $\overline{\Delta F}^\ddagger \equiv \Delta F^\ddagger(\overline{\epsilon}, \overline{\ell})$, where $\overline{\epsilon}$, $\overline{\ell}$ are the first moments (mean) of the distributions, indeed increases as $\overline{\ell}$ increases, consistent with the observed trend. The theory gives the slope m_{MF} of the mean-field barrier vs $\overline{\ell}$ as [8] $m_{\mathrm{MF}} \equiv \partial(-\overline{\Delta F}^\ddagger/T)/\partial\overline{\ell} \approx -(3/2)\times(M/\overline{\ell}^2)\ln(\overline{\ell}^{1/2}/2)$. Calculating m_{MF} for all proteins used in Fig. 1(a) gives $\langle m_{\mathrm{MF}}\rangle = -0.41 \pm 0.09$, which is consistent with the slope of the best fit line -0.36. The mean-field slope for the proteins in Fig. 1(b) is -0.42 ± 0.08, which is almost twice the slope of the best fit line -0.19. A feasible explanation for these facts is that the degrees of freedom are significantly reduced for simulations, and many-body interactions are also neglected [12], both of

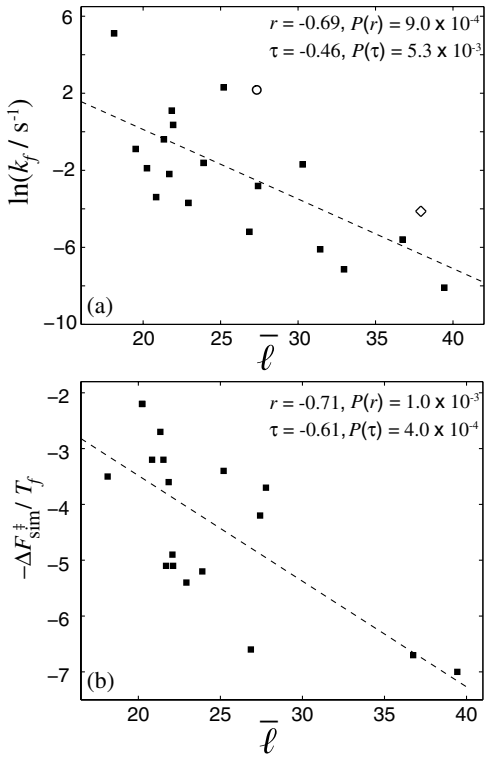

FIG. 1. (a) Logarithm of experimental folding rate (in sec^{-1}) at the transition midpoint vs absolute contact order or mean sequence separation between interacting residues in the native structure, $\overline{\ell}$. (b) The equivalent measure in Gō simulations is $-\Delta F^{\ddagger}_{\mathrm{sim}}/T_f$ (simulations are performed at the folding temperature T_f), again plotted vs $\overline{\ell}$. Both show a statistically significant anticorrelation: r (or τ) is the correlation coefficient (or Kendall's tau). Statistical significance is defined here by the probability $P(r)$ $[P(\tau)]$ to observe a given correlation coefficient or greater by chance. If $P(r)$ $[P(\tau)] < 0.05$, the dependence is typically deemed statistically significant. Shown in (a) are 19 proteins, and P^{13-14} (∘), a circular permutant of S6 (◇) [13] for which experimental rate data are available at various denaturant concentrations [9], and in (b), 18 simulated Gō model proteins [11].

which reduce the slope in Fig. 1(b). Meanwhile, the above theoretical estimates are taken only from the mean-field term [first term in Eq. (1)], and so may overestimate barriers, giving fortuitously good agreement in Fig. 1(a). Inconsistencies between the slope of Fig. 1(b) and the theoretical estimate also suggest that other factors in addition to $\overline{\ell}$ may be governing rates.

Second-order terms in Eq. (1) involving the fluctuations of native energies and loop lengths contact-to-contact all tend to decrease the barrier, leading to the notion that proteins with more heterogeneous folding mechanisms should fold faster [8]. We note that here a more heterogeneous folding mechanism corresponds to a more specific, polarized folding nucleus; i.e., the heterogeneity here refers to contact formation probability, not conformational diversity of the transition state. Earlier lattice-simulation studies (c.f. Abkevich *et al.* in [1]) as well as more recent experimental studies of circular permutants [13] support the notion that a more polarized nucleus results in a faster folding protein.

We can readily check if the second moment of the loop length distribution has an observable effect on rates, even if we ignore variations due to different $\overline{\ell}$ values protein-to-protein, as well as the terms with coefficients A and B in Eq. (1). The functional theory gives coefficient $C \approx Q^{\ddagger}$ in Eq. (1) [8], so the change in barrier height due to the presence of structural variance is

$$(\Delta F^{\ddagger} - \overline{\Delta F}^{\ddagger})/MT \equiv \delta\Delta F^{\ddagger}/MT \approx -Q^{\ddagger}\overline{\delta\ell^2}/\overline{\ell}^2. \quad (2)$$

Here, Q is the overall fraction of native contacts, and $Q^{\ddagger}$ is the value of Q at the barrier peak.

Plots of experimental log folding rate and simulated barrier heights (over MT) both show statistically significant correlation with $\overline{\delta\ell^2}/\overline{\ell}^2$ (Fig. 2); however, there are large fluctuations present, and the slope of the best fit line is only about a tenth of the theoretical prediction. Neglecting trends due to contact order and energetic variance introduces large fluctuations.

Experimentally measured ϕ values [14] involve both energetics and entropics and should better capture the effects of heterogeneity in folding mechanism. The vari-

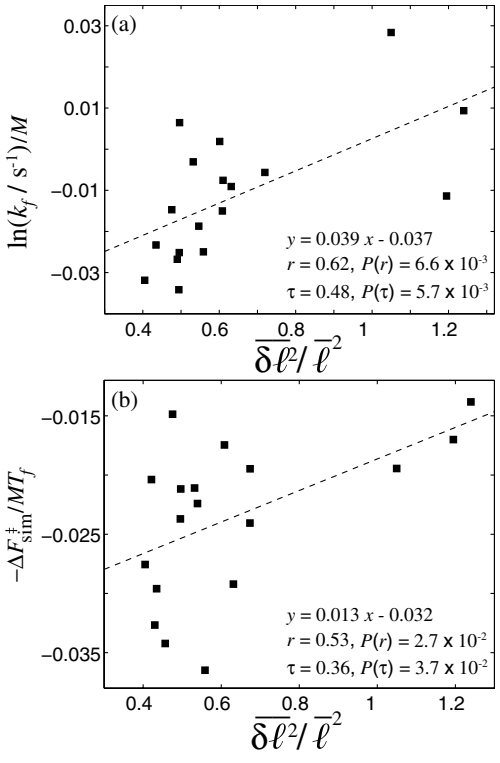

FIG. 2. Plotted in (a) are log experimental rate data (at the transition midpoints) and in (b), simulated barriers (at T_f), as a function of the measure of structural heterogeneity that appears in the functional theory in Eqs. (1) and (2). Both show a moderate but statistically significant correlation with structural variance [16]. Three α/β proteins (λ-repressor chain 3, cytochrome c, yeast iso-1-cytochrome c) tend to have both large structural variance and fast folding rates.

ance in ϕ values couples together the last three terms in Eq. (1). To facilitate a comparison of rates with ϕ variance, the free energy barrier may be recast in terms of the variance in native contact formation probabilities (Q_{ij}) [8]

$$\delta\Delta F^{\ddagger}/MT \approx -\overline{\delta Q^2}/2Q^{\ddagger}. \quad (3)$$

Equation (3) only includes the effects of heterogeneity in polymer loop length; however, energetic heterogeneity can be incorporated as well, which only changes the coefficient $(1/2Q^{\ddagger})$ in Eq. (3) to $(3/2Q^{\ddagger})$. The simulations have no variance in native contact energies; moreover, statistics arguments suggest that this native variance may be significantly reduced with respect to the variance in collapsed random structures [2].

ϕ values may be defined analytically as [12,15] $\phi_i = \sum_{j\neq i}(Q^{\ddagger}_{ij} - Q^{\mathrm{U}}_{ij})\Delta^N_{ij}/\sum_{j\neq i}(Q^{\mathrm{F}}_{ij} - Q^{\mathrm{U}}_{ij})\Delta^N_{ij}$, where Q^{U}_{ij}, $Q^{\ddagger}_{ij}$, and Q^{F}_{ij} are the probabilities of native contact formation between residues i and j in the unfolded, transition, and folded states, respectively. It follows that in the approximation that all contacts are fully formed in the native structure ($Q_{\mathrm{F}} = 1$), and unformed in the unfolded structures ($Q_{\mathrm{U}} = 0$), ϕ_i is the mean of Q_{ij} values in the transition state. Further approximating the same number of nearest neighbors z for all residues, the variances are related by $\overline{\delta\phi^2} \approx (1/z)\overline{\delta Q^2}$. If we make no approximations and simply plot $\overline{\delta Q^2}$ vs $\overline{\delta\phi^2}$ (for the simulation data), the quantities correlate extremely well (see Table I) with a slope of ≈ 1.2 and an intercept -0.04.

The above arguments indicate $\overline{\delta Q^2}$ and $\overline{\delta\phi^2}$ are within a factor of approximately unity, so we rewrite Eq. (3) in the form

$$\delta\Delta F^{\ddagger}/MT \approx -D\overline{\delta\phi^2}, \quad (4)$$

with D a parameter of order unity. Thus more polarized nuclei have lower free energy barriers. Plots of $-\Delta F^{\ddagger}/MT$ vs $\overline{\delta\phi^2}$ for experiments and simulations are shown in Fig. 3. Here we see a strong statistically significant correlation of both rates and barriers with ϕ variance. Moreover, the slopes of the best fit lines (≈ 0.3) compare somewhat more favorably with the theoretically predicted values (≈ 0.8) than was the case for structural variance. A precise comparison with experimental data is more difficult since the coordination number z as well as the numbers Q_{U} and Q_{F} are not accurately known for all proteins. Taking the slope from Fig. 3(a) and using the approximations mentioned above allows us to infer the residue-residue coordination number: $z \approx 4$ if energetic heterogeneity is negligible [Eq. (3)], and $z \approx 11$ if it is

TABLE I. Correlation coefficient and statistical significance for various quantities.

y versus:	x	r	$P(r)$[a]	τ	$P(\tau)$[a]
$\ln(k_f)$	$\overline{\ell}$	-0.69	9×10^{-4}	-0.46	5.3×10^{-3}
$-\Delta F^{\ddagger}_{\mathrm{sim}}/T_f$	$\overline{\ell}$	-0.71	10^{-3}	-0.61	4×10^{-4}
$\ln(k_f)/M$[b]	$\overline{\delta\phi^2}_{\mathrm{exp}}$	0.78	2.8×10^{-3}	0.52	2×10^{-2}
$-\Delta F^{\ddagger}_{\mathrm{sim}}/MT_f$	$\overline{\delta\phi^2}_{\mathrm{sim}}$	0.67	2.3×10^{-3}	0.47	7.2×10^{-3}
$\ln(k_f)/M$	$\overline{\delta\ell^2}/\overline{\ell}^2$	0.62	6.6×10^{-3}	0.48	5.7×10^{-3}
$-\Delta F^{\ddagger}_{\mathrm{sim}}/MT_f$	$\overline{\delta\ell^2}/\overline{\ell}^2$	0.53	2.7×10^{-2}	0.36	3.7×10^{-2}
$\overline{\ell}$	$\overline{\delta\ell^2}/\overline{\ell}^2$[c]	-0.14	0.52	-0.07	0.7
$\overline{\ell}$	$\overline{\delta\phi^2}_{\mathrm{exp}}$	-0.64	2.5×10^{-2}	-0.43	5.5×10^{-2}
$\overline{\ell}$	$\overline{\delta\phi^2}_{\mathrm{sim}}$	0.16	0.52	0.15	0.38
$\overline{\delta\phi^2}_{\mathrm{sim}}$	$\overline{\delta\ell^2}/\overline{\ell}^2$	0.71	10^{-3}	0.32	6.4×10^{-2}
$\overline{\delta\phi^2}_{\mathrm{exp}}$	$\overline{\delta\ell^2}/\overline{\ell}^2$	0.29	0.37	0.18	0.41
$\overline{\delta\phi^2}_{\mathrm{exp}}$	$\overline{\delta\phi^2}_{\mathrm{sim}}$	-0.16	0.8	0.2	0.63
$\overline{\delta\phi^2}_{\mathrm{sim}}$	$\overline{\delta Q^2}_{\mathrm{sim}}$	0.94	$<10^{-6}$	0.77	9×10^{-6}

[a]Two-sided statistical significance has been used.
[b]Here we divide by the number of native contacts M. Dividing instead by chain length N gives correlations within 10%. M and N correlate very strongly ($r = 0.94$).
[c]Data from both simulated and experimental proteins used.

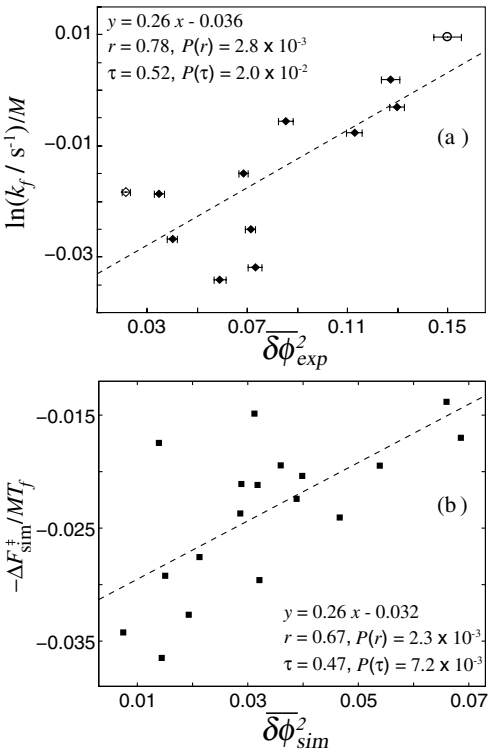

FIG. 3. (a) Plots of log experimental folding rate (over M) for a subset of proteins in Fig. 1(a) for which experimental ϕ values are available. (b) Minus free energy barrier (over MT) for simulated proteins vs ϕ variance. Both show strong statistically significant correlation. In particular, the trend in experimental data is strong even though the number of proteins with available data for both ϕ variance and transition midpoint rate is not large. Experimental data for wild type S6 ($\diamond$) and a circular permutant P^{13-14} ($\circ$) [13] fit very well to the rest of the data and increase the correlation. The strong correlation remains upon dividing by chain length N instead of total number of contacts M. Here the experimental rates at the transition midpoint have been compared to the variance in ϕ's typically measured in water or stabilizing conditions. Error bars in the experimental data are obtained by assuming a typical error of $\Delta\phi = 0.05$ for each ϕ value [9], giving $\Delta(\delta\phi^2) \approx 2\sqrt{\delta\phi^2}\Delta\phi/\sqrt{m}$, where m is the number of data points (ϕ values) for each protein.

substantial [Eq. (3) with coefficient $3/2Q^{\ddagger}$]. Residuals of $-\Delta F^{\ddagger}/MT$ vs $\overline{\ell}$, when plotted against $\overline{\delta\ell^2}/\overline{\ell}^2$ and $\overline{\delta\phi^2}$, show comparable correlations (within 10%) of those in Figs. 2 and 3.

Interestingly, experimental folding mechanisms tend to be more polarized than uniform Gō models (abscissae in Fig. 3). In the case of the simulations, the correlation between $\overline{\delta\phi^2}$ vs $\overline{\delta\ell^2}/\overline{\ell}^2$ is strong as expected, since there is no variance in native contact energies, by construction of the model. For experimental data, however, the correlation is poor, implying that there may be substantial energetic heterogeneity present in native contact energies of real proteins. It is not too surprising, then, that there is no correlation between the variance of experimental ϕ values and simulation ϕ values (see Table I). Thus in the analysis, simulated barriers were plotted against simulated ϕ variance, and experimental rates were plotted against experimental ϕ variance. We note that including S6 and its permutant does not change the correlation in Fig. 3(a), but decreases the correlation in Fig. 1(a) by 8%.

We did not find any significant correlation between rates and structural variance $\overline{\delta\ell^2}/\overline{\ell}^2$ for 3-state folders. Here there is the intriguing picture that (on-pathway) intermediates in 3-state folders are in fact induced by structural or energetic heterogeneity, so that there is no *a priori* reason for folding rates to continue to increase with increasing heterogeneity.

We showed here that both experimental rates and simulated free energy barriers for 2-state proteins depend on the degree of heterogeneity present in the folding process. The results compared quite well with the predictions of the free energy functional theory [8]. Heterogeneity due to variance in the distribution of native loop lengths, as well as variance in the distribution of ϕ values, were both seen to increase folding rates and reduce folding barriers. The observed effect due ϕ variance was the most statistically significant (as expected), because ϕ variance captures both heterogeneity arising from native topology as well as that arising from energetics.

S. S. P. acknowledges support from the Natural Sciences and Engineering Research Council and the Canada Research Chairs program. We thank Sebastian Cogswell, Kevin Plaxco, and Mikael Oliveberg for helpful discussions.

*Electronic address: steve@physics.ubc.ca

[1] D. B. Wetlaufer, Proc. Natl. Acad. Sci. U.S.A. **70**, 697 (1973); R. R. Matheson and H. A. Scheraga, Macromolecules **11**, 819 (1978); J. D. Bryngelson and P. G. Wolynes, Biopolymers **30**, 177 (1990); V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10026 (1994); A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **92**, 10869 (1995); P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **94**, 6170 (1997); D. K. Klimov and D. Thirumalai, J. Mol. Biol. **282**, 471 (1998); O. V. Galzitskaya, D. N. Ivankov, and A. V. Finkelstein, FEBS Lett. **489**, 113 (2001).

[2] S. S. Plotkin and P. G. Wolynes, Phys. Rev. Lett. **80**, 5015 (1998); S. S. Plotkin and J. N. Onuchic, Q. Rev. Biophys. **35**, 111 (2002); Q. Rev. Biophys. **35**, 205 (2002); G. Hummer, A. E. García, and S. Garde, Phys. Rev. Lett. **85**, 2637 (2000).

[3] A. R. Dinner and M. Karplus, Nat. Struct. Biol. **8**, 21 (2001).

[4] R. Mélin, H. Li, N. S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).

[5] Initially $\overline{\ell}/N$ was used to predict rates in water [K. W. Plaxco *et al.*, J. Mol. Biol. **277**, 985 (1998)], while $\overline{\ell}$ is a better predictor for both 2- and 3-state proteins [D. N. Ivankov *et al.*, Protein Science **12**, 2057 (2003)]. Here we remove effects due to varying stability by considering only rates at the transition midpoint vs $\overline{\ell}$. Stability in fact correlates with chain length for 2-state proteins in water [$r = 0.58$, $P(r) = 0.012$]. This may be in part why $\mathrm{RCO} \equiv \mathrm{ACO}/N$ acts as a better predictor of rates than $\overline{\ell}$ under these conditions. Other measures such as cliquishness [C. Micheletti, Proteins: Struct., Funct., Genet. **51**, 74 (2003)] or chain length [J. Kubelka *et al.*, Curr. Opin. Struct. Biol. **14**, 76 (2004)] can aid the prediction of rates in water, or rate limits.

[6] For a detailed description of the simulation model see, for example, Z. Guo, D. Thirumalai, and J. D. Honeycutt, J. Chem. Phys. **97**, 525 (1992); J. E. Shea, Y. D. Nochomivitz, Z. Guo and C. L. Brooks III, J. Chem. Phys. **109**, 2895 (1998), or C. Clementi, H. Nymeyer, and J. N. Onuchic, J. Mol. Biol. **298**, 937 (2000).

[7] B. A. Shoemaker, J. Wang, and P. G. Wolynes, J. Mol. Biol. **287**, 675 (1999); Proc. Natl. Acad. Sci. U.S.A. **94**, 777 (1997).

[8] S. S. Plotkin and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **97**, 6509 (2000); J. Chem. Phys. **116**, 5263 (2002).

[9] The Protein Data Bank (PDB) codes of these proteins are: 1AEY, 1APS, 1BF4, 1FKB, 1HRC, 1LMB, 1MJC, 1NYF, 1PGB, 1RIS, 1SRL, 1TEN, 1TIT, 1UBQ, 1YCC, 2AIT, 2CI2, 2PTL, 2VIK. The various references containing experimental data for rates and ϕ values for these proteins can be found at www.physics.ubc.ca/~steve/exptl.html..

[10] S. Takada, Proc. Natl. Acad. Sci. U.S.A. **96**, 11698 (1999); D. Baker, Nature (London) **405**, 39 (2000).

[11] PDB codes: 1AB7, 1AEY, 1APS, 1CSP, 1FKB, 1HRC, 1LMB, 1MJC, 1NMG, 1NYF, 1SHG, 1SRL, 1UBQ, 1YCC, 2AIT, 2CI2, 2PTL, 2U1A.

[12] M. R. Ejtehadi, S. P. Avall, and S. S. Plotkin, Proc. Natl. Acad. Sci. U.S.A. **101**, 15088 (2004).

[13] M. Lindberg, J. Tangrot, and M. Oliveberg, Nat. Struct. Biol. **9**, 818 (2002).

[14] A. R. Fersht, *Structure and Mechanism in Protein Science* (W. H. Freeman and Co., New York, 1999).

[15] J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten, and P. G. Wolynes, Fold. Des. **1**, 441 (1996).

[16] S6 displays significant correlation between native contact energies and native loop lengths [13]. For this reason we did not include it here, since $\overline{\delta\ell^2}/\overline{\ell}^2$ is only a structural measure of heterogeneity; if it is included the correlation in Fig. 2(a) decreases to $r = 0.57$, $P(r) = 9.6 \times 10^{-3}$.