

Inferring the Rate-Length Law of Protein Folding

Thomas J. Lane¹, Vijay S. Pande^{2*}

1 Department of Chemistry, Stanford University, Stanford, California, United States of America, **2** Departments of Chemistry, Computer Science, and Biophysics, Stanford University, Stanford, California, United States of America

Abstract

We investigate the rate-length scaling law of protein folding, a key undetermined scaling law in the analytical theory of protein folding. Available data yield statistically significant evidence for the existence of a rate-length law capable of predicting folding times to within about two orders of magnitude (over 9 decades of variation). Unambiguous determination of the functional form of such a law could provide key mechanistic insight into folding. Four proposed laws from literature (power law, exponential, and two stretched exponentials) are tested against one another, and it is found that the power law best explains the data by a modest margin. We conclude that more data is necessary to unequivocally infer the rate-length law. Such data could be obtained through a small number of protein folding experiments on large protein domains.

Citation: Lane TJ, Pande VS (2013) Inferring the Rate-Length Law of Protein Folding. PLoS ONE 8(12): e78606. doi:10.1371/journal.pone.0078606

Editor: Pratul K. Agarwal, Oak Ridge National Laboratory, United States of America

Received: February 13, 2013; **Accepted:** September 21, 2013; **Published:** December 5, 2013

Copyright: © 2013 Lane, Pande. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: TJL was supported by a National Science Foundation GRF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pande@stanford.edu

Introduction

A deep understanding of protein folding must involve a description of the general mechanisms involved. It is reasonable to suspect this will consist of a simple model, based on microscopic physics, expressed in a simple mathematical language. Such a model would show how biological sequences are able to employ physics to spontaneously self-assemble into intricate molecular machines.

Simple models of this sort often start by postulating a mechanism of folding, and then derive the consequences of that mechanism [1–12]. This suggests it might be possible to infer the general mechanisms of protein folding by verifying the specific predictions of these models. For a simple model of protein folding, however, there is a limited set of *general* experimental trends that can be readily predicted. One such experimental trend is the law governing how folding times scale with chain length. This is perhaps the simplest comparison of theory and experiment possible, but has not yet been unambiguously inferred despite nearly two decades of active research [2,7].

Polymer theory provides strong precedent for using chain-length scaling laws as a point of connection between experiment and theory. In polymer theory, for instance, comparing the theoretically predicted scaling of the longest relaxation time and polymer diffusion constant as a function of polymer chain length demonstrates deficiencies in the classical Rouse model [13]. Inclusion of hydrodynamic interactions – as is done *e.g.* in the Zimm model [14] – is necessary to get the correct scaling laws for these kinetic parameters [15]. Thus, comparing chain-length scaling laws in theory and experiment yields new scientific insight, namely that hydrodynamic effects contribute significantly to polymer dynamics in solution. Our hope is that by determining a rate-length law for protein folding, similar comparisons between experiment and theory will yield new insight into how proteins fold.

The rate-length law is also an interesting result in and of itself. Such a law can be viewed as a statement of the computational complexity of protein folding – given a problem of size N (residues), how does one expect the time-to-solution (folding) to scale? Levinthal pointed out that an exhaustive search would result in exponential scaling, and suggested that this would result in unreasonably large folding times [16]. Thus, in many ways, a resolution to Levinthal's paradox is likely to be phrased directly as a rate scaling law, either non-exponential (polynomial) or exponential with an explicitly small exponential factor.

A number of issues complicate inferring such a law from experiment, most importantly the fact that available kinetic data on protein folding spans a very limited range of chain lengths – about 30 to 300 residues [17]. The statistical power of the data is inherently limited by the fact that protein domain sizes barely span a single order of magnitude, and that most studies of folding have focused on small, well-behaved model systems.

Further complicating the inference of the rate-length scaling law is the fact that chain length is certainly not the only factor affecting folding rates. In fact, it has been argued that it is a fairly weak predictor of the folding time [18,19]. For instance, there seems to be some correlation of folding times with topological complexity of the native state, such that if two proteins have the same number of residues, but different folds, they may take different amounts of time to fold [18,20]. Moreover, even protein mutants with the same native structure can have at least 3 orders of magnitude variation in their folding rates [21]. Thus, we expect that experimental data on the scaling of folding time with chain length should be very noisy, and difficult to statistically estimate.

Nonetheless, as we will demonstrate, there does seem to be a significant correlation between the number of residues (N) and folding times (τ). Many different mathematical forms of this scaling law have been postulated, either from theory or empirically, but all fall into one of three basic classes. Shakhovich [2], Cieplak [3,4], and co-workers have proposed a power-law, $\tau \sim N^\nu$. We recently

constructed a model that suggested exponential scaling, $\tau \sim e^{\alpha N}$ [1], consistent with predictions made by Zwanzig, Szabo and Bagchi [5,6]. Finally, Thirumalai [7,8], Muñoz [9], Takada [10], Finkelstein [11], and co-workers have suggested a stretched exponential, $\tau \sim e^{\alpha N^\beta}$, with β as 1/2 or 2/3. Wolynes has proposed the law may conditionally change between all four suggested models [12].

Each model for the rate-length law derives from a different model of the fundamental physics of folding. Exponential rate laws have been estimated based on models of protein folding as a biased conformational search [1,5,6]. Stretched exponentials have been derived by modeling folding as a critical nucleation process [11] or random energy model [7–9]. Power laws [2–4] and stretched exponentials [10] have been posited empirically based on simulation. Finally, lack of a strong rate-length relationship was suggested based on experimental data [18,19]. Clearly, different postulates for folding mechanisms lead to different predictions for the rate-length law (or lack thereof). Thus, inferring the rate-length law of protein folding may directly inform our understanding of the mechanisms by which all proteins fold.

In what follows, we develop and apply two methods for choosing between these models and evaluate how each proposed model performs.

Modeling Methods

Below, we outline two complementary methods for inferring which proposed scaling law is the most reasonable. First, we present a method capable of fitting each model’s parameters to the known data, and examining how well each model explains the data. Next, we investigate a second discriminatory method, which proposes that folding times must be below a certain threshold value to be biologically viable. It has been demonstrated that in the crowded milieu of the cell, proteins must fold rapidly to avoid aggregation or degradation. We suggest that this implies that we can check the reasonableness of any model by seeing if its prediction for this threshold time is reasonable with what has been observed empirically in biology.

In this study we focus on single-domain globular proteins. Kinetic data for the folding times of proteins were taken from the KineticDB [17], which reports protein folding times at zero denaturant, near room temperature, and under neutral pH. Other data sets exist [22–24], but were not consistent with one another – despite this, they yielded very similar results (Fig. S1, Table S1 in File S1).

Direct Method: Likelihood Maximization

We want to estimate the parameters for each proposed form of the scaling law. In what follows, we adopt a model that accounts not only for this scaling law, but all other factors (topology, experimental conditions, etc.) via a random Gaussian component. Thus, by fitting each model, we not only learn parameters for each

proposed model, but also get an estimate for the relative importance of these other factors in determining folding times.

We assert the following model for the folding time,

$$\log \tau / \tau_0 = f(N) + X \tag{1}$$

where

$$f(N) = \begin{cases} v \log N & \text{power-law} \\ \alpha N & \text{exponential} \\ \alpha N^{1/2} & \text{stretched exp. (1/2)} \\ \alpha N^{2/3} & \text{stretched exp. (2/3)} \\ \text{const.} & \text{null} \end{cases}$$

are the proposed folding rate laws, X represents a random variable distributed (independently and identically) as a zero-mean Gaussian, $X \sim \mathcal{N}(0, \sigma^2)$, and τ_0 is a fit constant accounting for units of time. These models include the four proposed rate-length laws suggested in the literature and a null model that corresponds to the lack of a rate-length relationship. By adding X to the logarithm of the folding time (1), we model random variation in *relative* terms, and it enters as a multiplicative factor.

The random variable X models all contributions to the folding rate not accounted for by the chain length. In a first-principles model, this would possibly include effects due to specific sequences, folded state topology, experimental conditions, or perhaps other factors. Here, we lump these terms into a single random variable in order to focus on the chain length exclusively.

Equation (1) implies τ is distributed as a log-normal, with location parameter $f(N)$ and scale parameter σ . The likelihood of the entire data set (assuming n independent measurements) is

$$\mathcal{L} = \prod_{i=1}^n \frac{1}{\tau_i \sqrt{2\pi\sigma^2}} \exp \left[-\frac{(\log \tau_i / \tau_0 - f(N_i))^2}{2\sigma^2} \right] \tag{2}$$

We have three parameters for each model, σ , τ_0 , and α or v for the exponential and power-law families, respectively.

We have fit these parameters by maximizing the likelihood \mathcal{L} . Model comparison can then be performed by investigating the ratio of the likelihoods of two alternative models (Table 1). We have adopted the simple likelihood approach (versus a full-fledged Bayesian analysis) because the number of fit parameters are small and equal for each model, the models are simple and low-dimensional, and we have little prior information about the parameters. See Fig. S3 and Table S2 in File S1 for a Bayesian

Table 1. Likelihood Ratios of \mathcal{L} -Maximized Models.

Model ¹	Pr. Law	Exp.	S. E. 1/2	S. E. 2/3	Null
Power Law		1.59·10 ³	7.98·10 ⁰	3.82·10 ¹	1.17·10 ¹³
Exponential	6.30·10 ⁻⁴		5.03·10 ⁻³	2.41·10 ⁻²	7.36·10 ⁹
S. E. 1/2	1.25·10 ⁻¹	1.99·10 ²		4.79·10 ⁰	1.46·10 ¹²
S. E. 2/3	2.61·10 ⁻²	4.15·10 ¹	2.09·10 ⁻¹		3.05·10 ¹¹
Null	8.56·10 ⁻¹⁴	1.36·10 ⁻¹⁰	6.83·10 ⁻¹⁰	3.27·10 ⁻¹²	

Primary model is on the left, alternate model along the top - thus, a larger number favors the model in the leftmost column.
doi:10.1371/journal.pone.0078606.t001

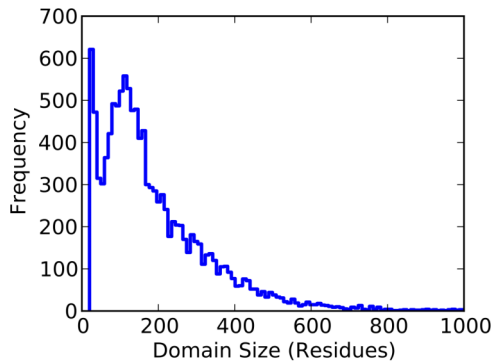


Figure 1. The size distribution of non-homologous protein domains listed in the PDB. Of the 12,151 sequences reported, 39 are larger than 1000 residues (0.32%) - while not shown here for clarity, they were obtained in subsequent analyses. Data were obtained from the NIH's VAST algorithm (<http://www.ncbi.nlm.nih.gov/Structure/VAST/nrpdb.html>) on 5/5/2012 with a dissimilarity p-value of 10^{-7} . doi:10.1371/journal.pone.0078606.g001

analysis and comparison. Note that parameters (α , ν) obtained from likelihood maximization of this model will be equivalent to those obtained by performing least-squares regression for the stated models on the logarithms of the folding times. Our maximum likelihood approach provides an estimate of the width of the fits involved (in the form of σ) and also a mechanism for rigorous model comparison via likelihood ratios.

Indirect Method: Biological Limits

We postulate that there exists a critical time, τ_c , that places a biological upper bound on folding times. Specifically, if a protein folds slower than this time (*i.e.* $\tau > \tau_c$) then that protein will be much more likely to aggregate during the course of folding, and therefore is evolutionarily selected against.

The majority of biologically observed proteins should have folding times less than τ_c , but we postulate that some proteins will have *greater* times. These proteins are those that receive help folding from chaperones or other cellular machinery. It has been

estimated that about $C \approx 10\%$ of proteins fall into this category [25].

Together, these assumptions allow us to build a model for the predicted distribution of protein chain lengths. The size distribution of domains (Fig. 1) can be roughly approximated by a Gaussian with parameters μ_N and σ_N . In that case,

$$\int_{f^{-1}(\tau_c/\tau_0)}^{\infty} \frac{1}{\sqrt{2\pi\sigma_N^2}} \exp\left[-\frac{(N-\mu_N)^2}{2\sigma_N^2}\right] = C \quad (3)$$

where $f^{-1}(\tau_c/\tau_0) = N_c$ is the chain length corresponding to τ_c for a specific model (power law, exponential, etc.), and C is the percentage of proteins with folding times slower than τ_c .

This framework is, of course, an approximation. There are undoubtedly many other factors affecting the optimal sizes of proteins beyond merely their folding times. Metabolic efficiency, structural packing constraints [26,27], and the behavior of specific proteins in their local cellular environments certainly play a role. Nonetheless, the concept of an upper limit to the folding times is reasonable, and our aim here is to simply extract some general comments about the reasonableness of predicted folding times, rather than make quantitatively accurate predictions.

Results and Discussion

Direct fitting of all proposed models to the available data yields reasonable results for each (Fig. 2). Each model reports a scale parameter (σ) of approximately 3, which indicates that 68% of proteins will have folding times within a factor of $e^\sigma \approx 20$ from the time predicted by the rate law, and 95% will be within a factor of $e^{2\sigma} \approx 400$. Thus, with knowledge of only the chain length alone, one can estimate the folding time of a protein to within about 2 orders of magnitude and be correct 95% of the time. Since the available data spans folding times of more than 9 orders of magnitude (between $1.9 \cdot 10^{-7}$ and $9.9 \cdot 10^2$ seconds), this demonstrates that chain length captures much of the variation in folding times.

To further assess the presence or absence of an overall rate-length law the likelihood of a null model, which modeled the

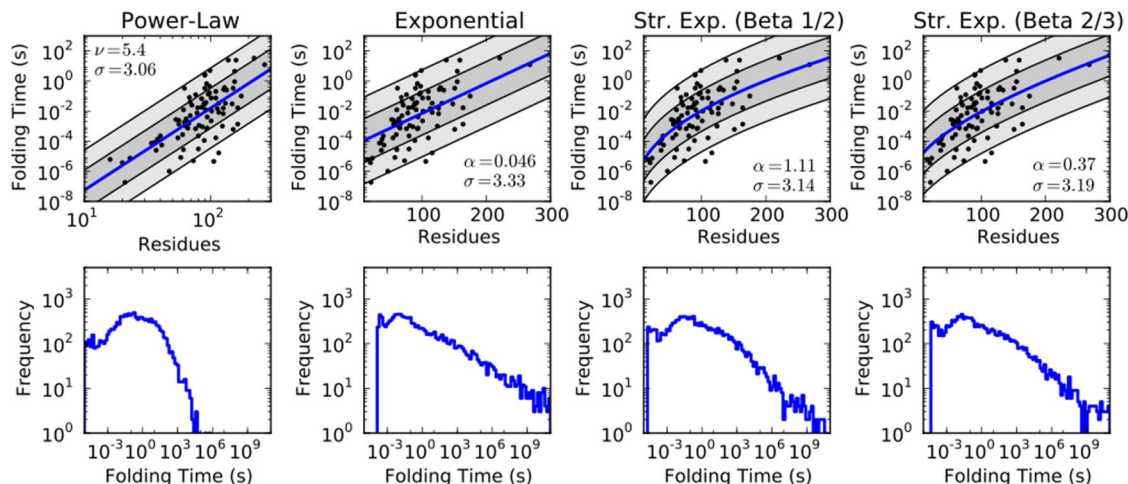


Figure 2. The predicted models for the folding rate law, overlaid with measurements of folding times (top), and the putative folding time distributions these models imply (bottom). Parameter values derived from a maximum likelihood fit are displayed, along with intervals indicating the spread in the fit probability distribution. Dark grey shading indicates a factor of e^σ , while light grey indicates $e^{2\sigma}$. doi:10.1371/journal.pone.0078606.g002

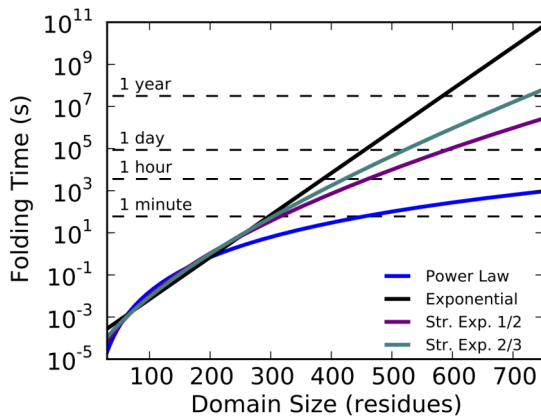


Figure 3. The predicted folding times from each model in Figure 2, in a direct comparison. Intuitive timescales are denoted for clarity.
doi:10.1371/journal.pone.0078606.g003

logarithm of folding rates as a Gaussian distributed independent of chain length ($f(N) = \text{const.}$), was computed. This null model was found to be significantly less likely (at least a factor of 10^9 less probable) given the data than any of the four rate-length models proposed (Table 1).

Do the data support any one model? The power-law model is slightly favored by comparing the likelihoods that each model generated the observed data (Table 1). In such comparisons, typically a ratio of 10^2 or greater is considered significant, and often models differ by hundreds of orders of magnitude [28] – thus, the power law model is better supported by the data, but only by a modest margin. Further, an attempt to fit the stretched exponential form with β as a variable parameter resulted in an unreasonably small value of β along with a very large value of α , resulting in a fit that is very close to the power law (Fig. S2 in File S1). Finally, the power law model has the smallest fit σ , indicating that it explains the most variation in the data, and attributes less to orthogonal factors.

It is clear, however, that there is little difference between the models in the range of available data. These models diverge significantly only for very large proteins (Fig. 3). We conclude that, with given experimental data, a direct statistical analysis cannot support any one of the theoretically proposed rate-length laws.

Despite this, each law yields significantly different predictions for the distribution of folding times, generated by transforming the known distribution of domain sizes into each of the different models (Fig. 2). The most significant differences are in the tails of these distributions, where the exponential forms predict much longer folding times for the largest proteins (Table 2). The power

Table 2. Estimated Fraction of Protein Domains with Folding Times Greater than Time Indicated.

	Hour	Day	Month	Year
Power Law	0.41%	0.01%	0.00%	0.00%
Exponential	9.56%	5.70%	3.34%	2.46%
S. E. 1/2	5.48%	2.37%	0.95%	0.57%
S. E. 2/3	7.49%	3.53%	1.74%	1.11%

doi:10.1371/journal.pone.0078606.t002

Table 3. Most Likely τ_c for Values of C (in seconds).

C	0.1	0.05	0.01	0.001
Power Law	$1.10 \cdot 10^1$	$2.88 \cdot 10^1$	$1.23 \cdot 10^2$	$4.36 \cdot 10^2$
Exponential	$3.03 \cdot 10^2$	$5.85 \cdot 10^3$	$1.51 \cdot 10^6$	$7.64 \cdot 10^8$
S. E. 1/2	$9.99 \cdot 10^1$	$6.54 \cdot 10^2$	$1.53 \cdot 10^4$	$3.47 \cdot 10^5$
S. E. 2/3	$1.68 \cdot 10^2$	$1.57 \cdot 10^3$	$7.67 \cdot 10^4$	$4.25 \cdot 10^6$

doi:10.1371/journal.pone.0078606.t003

law model predicts no proteins fold in times longer than an hour, while the exponential forms show a significant number of proteins with folding times longer than a day (Fig. 3).

An evaluation of the reasonableness of these folding time distributions is provided by the critical time τ_c for each model (Table 3). For reasonable values of C ($\approx 10\%$, predicted from experiment [25]), the power law τ_c is on the order of minutes. The exponential forms, on the other hand, predict τ_c is on the order of hours. Compare this to some cellular benchmarks, as done by Rollins and Dill (personal communication): the 16 second average time to synthesize a protein in *E. Coli* (325 residues $\times 0.05 \text{ s}^{-1}$ synthesis rate) [29], the 30 seconds it takes for GroEL to refold a protein [30,31], or the 20 min *E. Coli* doubling time.

Indeed, picking a reasonable value of τ_c and calculating the probability of observing the empirically observed domain size distribution (Fig. 4) shows that for values of $\tau_c \sim 10$ seconds, the power law model is clearly the best. However, for any values of τ_c greater than 100 seconds, the exponential laws are much better models.

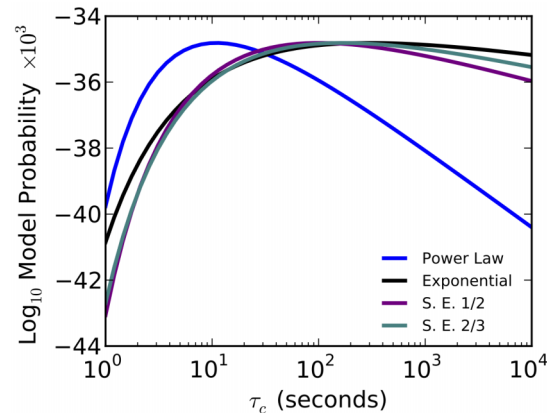


Figure 4. The probability that each of the rate-length laws reproduces the experimentally observed domain size distribution given a value of τ_c . Given a rate-length law mapping chain length to folding times, if the distribution of sizes of protein domains is modeled as a Gaussian, choosing a timescale τ_c above which proteins need assistance to fold, and an approximate fraction C of proteins that need assistance to fold fully specifies the distribution of domain sizes (eq. 3; here, $C = 0.10$). That distribution was used as a model for the generation of the empirically observed domain size distribution reported in Fig. 1. Here, we plot the probability that the empirically observed distribution resulted from the model indicated. Note the y-axis is \log_{10} and divided by 10^3 for clarity, so small differences on the plot are actually quite large.
doi:10.1371/journal.pone.0078606.g004

Conclusions

We conclude that while chain length is a statistically significant factor for predicting folding rates, current experimental data do not support any one of the proposed models for the rate-length relationship over any other. While the power law model appears to best explain the available raw data, it results in very fast predicted folding times. The exponential forms, while doing a marginally poorer job of explaining the raw data, yield a distribution of folding times more in line with what we expect from biology. Given the current available data, no clear victor emerges.

Previous theories have claimed that simply predicting one of the four laws investigated here is strong evidence in support of that theory. This is manifestly not the case – not only must the proposed law be reasonable, but it must also predict reasonable parameter estimates, and even then the supporting evidence the rate scaling law can provide given current data is limited. Conversely, the analytical theories mentioned here are not ruled out by the current available data. This is most striking in the case of the exponential form, since exponential scaling of the folding times has often been associated with Levinthal's paradox. This study shows that exponential scaling is reasonable given current experimental data, so long as the exponential scaling constant (α) is sufficiently small.

Clear evidence for any one rate law remains missing. However with a few clear examples of very large globular proteins (500 residues or larger) capable of folding unassisted *in vitro*, it might be possible to discriminate between the models proposed here. Figure 3 clearly shows the divergences between predicted folding times for large proteins, and shows how just a few data points in this extreme regime might be able to begin differentiating between the proposed models investigated here.

References

- Lane TJ, Pande VS (2012) A simple model predicts experimental folding rates and a hub-like topology. *J Phys Chem B* 116: 6764–6774.
- Gutin AM, Abkevich VI, Shakhovich EI (1996) Chain length scaling of protein folding time. *Phys Rev Lett* 77: 5433–5436.
- Cieplak M, Hoang TX (2003) Universality Classes in Folding Times of Proteins. *Biophys J* 84: 475–488.
- Cieplak M, Hoang T, Li M (1999) Scaling of folding properties in simple models of proteins. *Phys Rev Lett* 83: 1684–1687.
- Zwanzig R, Szabo A, Bagchi B (1992) Levinthal's Paradox. *Proc Natl Acad Sci USA* 89: 20–22.
- Zwanzig R (1995) Simple model of protein folding kinetics. *Proc Natl Acad Sci USA* 92: 9801.
- Thirumalai D (1995) From Minimal Models to Real Proteins - Time Scales for Protein-Folding Kinetics. *J Phys I* 5: 1457–1467.
- Li MS, Klimov DK, Thirumalai D (2002) Dependence of Folding Rates on Protein Length. *J Phys Chem B* 106: 8302–8305.
- Naganathan AN, Muñoz V (2005) Scaling of Folding Times with Protein Size. *J Am Chem Soc* 127: 480–481.
- Koga N, Takada S (2001) Roles of native topology and chain-length scaling in protein folding: A simulation study with a Gō-like model. *J Mol Biol* 313: 171–180.
- Finkelstein AV, Ivankov DN, Garbuzynskiy SO, Galzitskaya OV (2007) Understanding the folding rates and folding nuclei of globular proteins. *Curr Protein Pept Sci* 8: 521–536.
- Wolynes P (1997) Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc Natl Acad Sci USA* 94: 6170–6175.
- Rouse PE (1953) A Theory of the Linear Viscoelastic Properties of Dilute Solutions of Coiling Polymers. *J Chem Phys* 21: 1272.
- Zimm BH (1956) Dynamics of polymer molecules in dilute solution: viscoelasticity, flow birefringence and dielectric loss. *J Chem Phys* 24: 269.
- Grosberg AY, Khokhlov AR, Atanov YA (1994) Statistical physics of macromolecules.
- Levinthal C (1969) How to Fold Graciously. *Mossbauer Spectroscopy in Biological Systems Proceedings of a meeting held at Allerton House* : 22–24.

Supporting Information

File S1 Supplemental Information.

Figure S1. The rate-length data from each source used, plotted together. Even though there is some overlap in reported sequences, many identical proteins have different chain lengths or folding times reported. We combined these data for our analysis, eliminating only identical measurements.

Figure S2. The parameter fit for a stretched exponential, $\log \tau \propto \alpha N^\beta$, with β a free parameter. Notice how the fit is perfectly straight on a log-log plot, a characteristic trait of power laws.

Figure S3. The parameter posteriors for each model are sharply peaked around their modes. Plotted here is the maximum (not the marginal value) of the posterior at various values of the key parameters α or v . One can see that the likelihood has a sharply peaked value along this dimension.

Table S1. The KineticDB dataset with mutants included. Since there are only a few proteins with mutants, and there are many mutants for these few proteins, this database gives artificially more weight to those individual proteins.

Table S2. Bayes factors comparing datasets. (PDF)

Acknowledgments

We would like to thank Kings Ghosh and Kevin Dalton for discussions.

Author Contributions

Conceived and designed the experiments: TJL VSP. Performed the experiments: TJL. Analyzed the data: TJL. Wrote the paper: TJL VSP.

- Bogatyeva NS, Osypov AA, Ivankov DN (2009) KineticDB: a database of protein folding kinetics. *Nucleic Acids Research* 37: D342–D346.
- Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 277: 985–994.
- Nakamura H, Takano M (2005) Temperature dependence and counter effect of the correlations of folding rate with chain length and with native topology. *Phys Rev E* 71.
- Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, et al. (2003) Contact order revisited: Influence of protein size on the folding rate. *Protein Sci* 12: 2057–2062.
- Lawrence C, Kuge J, Ahmad K, Plaxco KW (2010) Investigation of an Anomalous Accelerating Substitution in the Folding of a Prototypical Two-State Protein. *J Mol Biol* 403: 446–458.
- Ouyang Z, Liang J (2008) Predicting protein folding rates from geometric contact and amino acid sequence. *Protein Sci* 17: 1256–1263.
- De Sancho D, Muñoz V (2011) Integrated prediction of protein folding and unfolding rates from only size and structural class. *Phys Chem Chem Phys* 13: 17030.
- Ivankov DN, Finkelstein AV (2010) Protein Folding as Flow across a Network of Folding-Unfolding Pathways. 2. The “In-Water” Case. *J Phys Chem B* 114: 7930–7934.
- Hartl FU, Hayer-Hartl M (2009) Converging concepts of protein folding in vitro and in vivo. *Nat Struct Mol Biol* 16: 574–581.
- Xu D, Nussinov R (1998) Favorable domain size in proteins. *Fold Des* 3: 11–17.
- Shen MY, Davis FP, Sali A (2005) The optimal size of a globular protein domain: A simple spherepacking model. *Chem Phys Lett* 405: 224–228.
- Kass RE, Raftery AE (1995) Bayes Factors. *J Am Stat Assoc* 90: 773–795.
- Young R, Bremer H (1976) Polypeptide-chain-elongation rate in *Escherichia coli* B/r as a function of growth rate. *Biochem J* 160: 185–194.
- Ewalt KL, Hendrick JP, Houry WA, Hartl F (1997) In vivo observation of polypeptide flux through the bacterial chaperonin system. *Cell* 90: 491–500.
- Horwich AL, Fenton WA (2009) Chaperonin-mediated protein folding: using a central cavity to kinetically assist polypeptide chain folding. *Quart Rev Biophys* 42: 83.