Published in final edited form as:

J Am Chem Soc. 2014 August 13; 136(32): 11420–11427. doi:10.1021/ja5049434.

General Mechanism of Two-State Protein Folding Kinetics

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

Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, California 94143, United States, and Laufer Center for Physical and Quantitative Biology and Departments of Chemistry and Physics and Astronomy, Stony Brook University, Stony Brook, New York 11790, United States

[†]University of California, San Francisco

[‡]Stony Brook University

Abstract

We describe here a general model of the kinetic mechanism of protein folding. In the Foldon Funnel Model, proteins fold in units of secondary structures, which form sequentially along the folding pathway, stabilized by tertiary interactions. The model predicts that the free energy landscape has a volcano shape, rather than a simple funnel, that folding is two-state (single-exponential) when secondary structures are intrinsically unstable, and that each structure along the folding path is a transition state for the previous structure. It shows how sequential pathways are consistent with multiple stochastic routes on funnel landscapes, and it gives good agreement with the 9 order of magnitude dependence of folding rates on protein size for a set of 93 proteins, at the same time it is consistent with the near independence of folding equilibrium constant on size. This model gives estimates of folding rates of proteomes, leading to a median folding time in *Escherichia coli* of about 5 s.

Introduction

Is there a general mechanism of protein folding kinetics? On the one hand, different types of protein molecules adopt different native structures – having distinctive secondary and tertiary structures and packing details. On the other hand, remarkably, essentially all small soluble globular proteins tend to reach their different atomically detailed native structures rapidly (often milliseconds) and with the simplest possible kinetics (single-exponential), independent of initial conditions. And, while folding rates span 9 orders of magnitude,[1] folding equilibria are quite insensitive to protein structure. Is there a *folding mechanism*, that is, a single narrative description that rationalizes the rates and sequences of folding events in common across different amino acid sequences and initial conditions?

[*]To whom correspondence should be addressed dill@laufercenter.org.

Supporting Information Available

Details of our kinetic model are given, along with a comparison to other simple metric ts, like contact order and chain length. This material is available free of charge via the Internet at http://pubs.acs.org/.

Beginning about forty years ago, major insights on the thermodynamics and kinetics of protein folding have emerged from experiments, computer simulations, correlational studies and theoretical modeling.[2–10] First, an early and important view has been that proteins fold kinetically through the rapid formation and assembly of secondary structures.[6,8,11–19] Second, Plaxco et al. had the pioneering insight that a protein's folding rate depends on properties that are evident from its native structure.[20] They found that helical proteins tend to fold faster than β-sheet proteins, and, in general, that local structures tend to form faster than nonlocal ones. In a more detailed discussion in Supporting Information, we describe the current consensus[1,21–24] that folding rates are better correlated with a protein's size — its chain length (L), its number of secondary structures (N), or its absolute contact order (ACO) — than they are with other metrics, like the relative contact order (RCO), that only consider the topology of a protein's native structure (Fig. S1, Table S1). Such folding rate datasets have also been fitted by statistical models.[21–23,25–29] Third, important insights have emerged from Ising models of folding. Zwanzig, Szabo & Bagchi (ZSB) used Ising models to show how the funnel shapes of energy landscapes lead to fast folding.[30,31] Muñoz, Eaton, Baker, Finkelstein, and others[32–38] have further developed and applied the ZSB approach, adding more detailed residue-level information in the form of contacts, hydrogen bonds, buried surface area, and loop entropies. Barrick and co-workers have used Ising models to explain their exceptionally comprehensive folding energy landscapes of linear repeat proteins, such as ankyrin, which fold along parallel paths.[39–42] In their model, each folding unit is an individual repeat, whereas in our model, each unit is a secondary structure. Regan and co-workers have also used an Ising-like framework to study the thermodynamic properties of repeat proteins.[43] Fourth, previous work has shown that equilibrium protein folding cooperativities can be explained as a combination of weak propensities of peptide chains to form secondary structures and stronger propensities of tertiary interactions to stabilize the secondary structures.[44]

However, as far as we are aware, there is not yet a quantitative model for a general folding mechanism that predicts folding rates and routes from only a protein's amino acid sequence. Here, we develop a model that does not require prior knowledge of native topologies, structural propensities, native geometric details or initial conditions. An associated purpose of the model is to reconcile the ‘pathway view’ that folding follows well-defined sequential events with the ‘funnel view’ that folding follows combinatoric microscopic routes.[45] A general mechanism should account for why simple-protein folding is single-exponential, the nature of the folding transition state, the sequence of formation of secondary and tertiary structures, the relative speeds of formation of the different substructures and the nature of cooperativity in kinetics.

Model

To express a protein's folding equilibrium and kinetics, we adapt the Ising-like approach of Zwanzig,[30,31] except that instead of independent amino acids, the individual units of folding in our model are the secondary structures. We represent a protein's N secondary structural units as a 1-D string of symbols fffuffuufffff . . . where f indicates that a particular secondary structure is in its folded native-like conformation, and u indicates that it is in an unfolded non-native conformation. Let c represent the number of f's, the correct secondary

structures in the string. Therefore, $c = N$ represents the folded native state. $c = N - 1$ describes the state in which the protein is native in all but one of its secondary structures, so there is one u somewhere in the string. And $c = 0$ corresponds to the fully unfolded molecule.

The model folding process is shown in Fig. 1 for a 4-helix-bundle (hb). Secondary structures form independently. Secondary structures can pair together, stabilized by tertiary interactions between them. Adopting literature terminology, we call these secondary structure elements *foldons*.[15,18,19] The main folding routes entail increasingly native structure that is assembled through the sequential addition of one secondary structure at a time. The routes are combinatoric and stochastic: different molecules fold via different sequences of secondary structural events. We call this the Foldon Funnel Model.

Thermodynamics of the model

In this model, the Boltzmann weight $w(c)$ for any non-native protein configuration having $c = 1, 2, 3, \ldots N - 1$ correct secondary structures includes the equilibrium constant K_2 for each of the c secondary structures formed at a particular stage of folding progress, the equilibrium constant K_3 for each of the n_c tertiary pairings of secondary structures, and the numbers of different ways the folded and unfolded units can be arranged in a 1-D string at a particular state c of the folding progress. Hence, we define

$$w(c) = \frac{N!}{c!\,(N-c)!} K_2^c K_3^{n_c}. \quad (1)$$

Correspondingly, the Boltzmann weight for the native configuration $c = N$ is:

$$w(N) = K_2^N K_3^{n_N} K_f. \quad (2)$$

The microscopic basis for K_2 is the same as in helix-coil theory: hydrogen bonds stabilize secondary structures, and local chain entropy opposes them. Similarly, K_3 arises from contact interactions among pairs of secondary structures and includes hydrophobic, steric and hydrogen bonding interactions. K_f accounts for the extra stabilization of a protein that steps from the next-to-native to the native structure because of the final packing and assembly (see below). That is, we envision loose packings of the secondary structures in the early steps up to $c = N - 1$, and then native-like tight packing only in the final folding step, from $N - 1$ to N. The weight $w(0) = 1$ accounts for the fully unfolded protein; $w(1) = NK_2$ accounts for the formation of any one of the N individual secondary structures; $w(N-1) = NK_2^{N-1} K_3^{n_{N-1}}$ accounts for the formation of the next-to-native state; and $w(N) = K_2^N K_3^{n_N} K_f$ accounts for the fully folded protein. The quantity n_c is the total number of tertiary interactions made by a secondary structure. From our protein data set, we find that n_c saturates; see Fig. 2. That is, for simple reasons of steric geometric exclusion, a secondary structure cannot typically be surrounded by more than about 4–5 other neighboring secondary structures: n_c is defined as a discrete function of c

$$n_c = \begin{cases} 0 & \text{if } c=0,1 \\ 1 & \text{if } c=2 \\ 3 & \text{if } c=3 \\ 4c-10 & \text{if } c>3 \end{cases}$$

where n_c is defined by the combinatorics of nearest-neighbor interactions. When there are zero or one secondary structures (c = 0 or 1), there can be no tertiary interactions. When there are c = 2 secondary structures, there is n_c = 1 tertiary interaction between them. When there are c = 3 secondary structures, there are n_c = 3 pairwise tertiary interactions between them. For c > 3, each additional secondary structure gains four tertiary neighbor interactions upon folding because this is approximately the maximum that is sterically possible; see Fig. 2.

From the weights $w(c)$, the equilibrium population for any point c along the reaction coordinate is given by $p(c) = w(c)/Q$, where $Q = Q_U + Q_F$ is the partition function, that is, the sum of weights Q_U over all the non-native states and over Q_F of the folded state:

$$Q = Q_F + Q_U = w(N) + \sum_{c=0}^{N-1} w(c) \quad (3)$$

The free energy landscape is given by $\Delta G(c) = -RT \ln[p_c(eq)]$. Fig. 4 shows a funnel representation of $\Delta G(c)$. The radial distance from the center of the funnel is the reaction coordinate c, the number of folded secondary structures. The outer flat region of the landscape corresponds to unfolded conformations (c = 0). As the protein moves uphill, secondary structures are formed. Surmounting the barrier at $c = N - 1$ leads to the folded state, which is the global free energy minimum.

Kinetics of the Model: The Folding and Unfolding Rates

The folding and unfolding dynamics of the model can be described by a continuous-time Markov process. On the one hand, we can compute the full dynamics of the model (for details, see Supporting Information). On the other hand, in an important limit, we can compute the dynamics in a very simple analytical way. In particular, because we find that the best fits to the folding rate data are when the highest barrier is at $c = N - 1$, the folding and unfolding rates are well-approximated by

$$k_f = k_1 \frac{w(N-1)}{Q_U} \quad (4)$$

$$k_u = k_1 \frac{w(N-1)}{Q_F} \quad (5)$$

where k_1 is a rate constant for the folding of a single secondary structure. In the Supporting Information, we show that these analytical expressions capture with negligible error, for appropriate ranges of parameter values, the results of the full master equation as computed by numerical integration and as found by eigen decomposition of the rate matrix. In the Supporting Information, we also show that there is a gap in the eigen value spectrum, which means that the model predicts a single dominant slowest exponential relaxation time, characteristic of two-state kinetics, which is the general behavior seen for the folding of small globular proteins.

Experimental Data Set and the Model Parameters

For comparison with experiments, we considered a data set of 93 globular proteins for which the folding rates are known; see (Tables S3,S4). This data set includes both two-state and multistate folders. For the multistate proteins, we considered only the slowest folding phase. We use this data set to fit the two parameters of the model, K_2 and K_3. To do this, we first fixed the value of the speed-limit parameter k_1 to $10^{5.6}s^{-1}$, the mean value of the folding rates of the two elementary secondary structures, $N = 1$: the mini-protein Trp Cage and the central helix of ribosomal protein L9. Also, because prior modeling gives protein folding equilibrium constants $K(L)$ as a function of the chain length L, we used that data to fix the value of K_f in our quantity Q_F/Q_U[46,47] K_f ranges from 1.75 for $N = 1$ to 19.4 for $N = 30$ (Table S2). This ensures consistency with the known database of protein stabilities, and therefore ensures roughly correct unfolding rates as well. We bootstrapped the folding rate data and fitted each resampled data set in order to generate a confidence interval. The 95% confidence interval bands are plotted in gray. $R^2 = 0.63$ for this fit.

Results and discussion

Folding Landscape Is Shaped Like a Volcano

Fig. 3 compares the model predictions from eq 4 to experimental data on folding rates of the 93 globular proteins using best-fit values: $K_2 = 0.037$ and $K_3 = 1.96$. From this modeling, we draw a few conclusions.

Outer Landscape Is Sloped Uphill Because Secondary Structures Are Not Stable—Because the parameters we obtain are $K_2 << 1$ and $K_3 > 1$, we infer that secondary structures are unstable alone and that they are stabilized by tertiary interactions. This prediction is consistent with experiments indicating that most protein secondary structures are unstable on their own.[48,49] The prediction is also consistent with measured protein equilibrium cooperativities.[44] So, the predicted folding landscape for two-state folders is shaped like a *volcano* when plotted vs. the 1D mesoscale reaction coordinate c that we use here; see Fig. 4. That is, folding is a series of uphill steps in free energy as the earliest secondary structures form and assemble into increasingly native-like tertiary structure; only the last step from $c = N - 1$ to N is downhill in free energy. Forming the first helix (i.e. the step from $c = 0$ to $c = 1$) is the most costly step. Forming the second helix (from $c = 1$ to $c = 2$) is less costly because the second helix is stabilized by assembling onto the first helix as folding proceeds. So, the slope of the free energy landscape vs. c is steep for small c but

Folding Is Two-State (Single-Exponential) Because the Global Bottleneck Is the Last Step in Folding—The highest free energy on the volcano landscape is at $c = N - 1$, the structure just before the native state. Hence, all earlier steps are effectively in *pre-equilibrium*. This divides conformational space into the two kinetic states: native ($c = N$) versus all others. Single-exponential behavior would not have intrinsically been expected for such a heterogeneous and complex process as protein folding. Indeed, there may be other parameter regimes, K_2 and K_3, that do not lead to 2-state kinetics.

What Is the Transition State?—The present model resolves a puzzle. Does the transition state appear early or late along the folding pathway? The present model gives an explanation for the ambiguity. As noted above, the global TS in the model is the last step in folding; it is the point of highest free energy on the landscape. Said differently, the last step is responsible for the dominant slowest exponential of the kinetics. On the other hand, further insight is available from looking at the full dynamics of folding, shown in Fig. 5 for a 4-helix bundle (4hb). It shows that the full kinetics entails *nested transition states* for the individual folding steps. While the 3hb is the TS for final step of folding, the 2hb is also the TS for the prior step (from single helix to the 3hb). In short, each partial structure along the folding pathway is a transition state for propagating earlier structures to later ones. Also, in our model, the transition state for folding is a loose association of the native secondary structures that occurs prior to native-like tighter packing, consistent with the view from the nucleation-condensation hypothesis that the TS is large diffuse nucleus.[50]

Does Folding Follow a Single Sequential Pathway or Parallel Heterogeneous Routes?—The present model is consistent with both the funnel landscape view that folding is a disorder-to-order transition through many different microscopic routes[9,45,51,52] and the view of folding based on sequential pathways and “foldons”, wherein secondary structural elements fold via particular path-like sequences of events.[15,18] Funnels and foldon paths are not mutually exclusive; they are just different perspectives at different levels of resolution, functions of different degrees of freedom, and focused on different parts of the landscape. Funnels express free energies in terms of microscopic degrees of freedom. Pathways express free energies in terms of macroscopic reaction coordinates. Here, our modeling is mesoscale. We express our free energy landscape in terms of a single reaction coordinate c. Some aspects of the foldon path perspective are evident in the present model: the reaction coordinate is one-dimensional, and there is a clear sequential order of folding events through the formation of $c = 1, 2, 3, \ldots N$ folding units. On the other hand, the funnel perspective is also evident: Fig. 1 shows the combinatorics of the many different routes of assembling the secondary structures (There are additional route combinatorics that arise from the many microscopic routes for forming each secondary structure, but those are below the resolution of the present model). The folding of any particular protein entails more subtle aspects: not all secondary structures form at the same rate, for example, but we believe the present model captures the essential physics with a minimum of parameters.

Recently, Hu et al. have performed comprehensive pulsed HX experiments to identify the folding pathways of RNase H.[19] Consistent with our model, Hu et al. found that RNase H folds in units no smaller than secondary structural elements (foldons), that those elements

form and assemble into ever larger and more native-like structures, and that some individual foldons form concurrently (D/5 and BC/loop), while other foldons assemble with each other in series. Yet, one difference is apparent: the free energy landscape of Hu et al. is a down-staircase with sequentially stabilized intermediates, while ours is an up-staircase. The essential difference here is that our model only pertains to two-state folding, so it does not address questions of stable intermediates, such as the one observed by Hu et al. Hu et al. found a large final barrier for the folding of 123/E and its assembly with A/4, D/5, and BC/loop. We note that the experiments of Guinn et al.[53] and the modeling of Adhikari et al.[54] also give up-staircase landscapes in two-state proteins.

Why Does the Folding Speed of a Protein Correlate with the 'Localness' of its Native Structure?—Previous studies, starting with Plaxco, Simons and Baker[20] have found that protein folding rates are correlated with a native protein's *Contact Order* (CO), a measure of the protein's numbers of local versus nonlocal contacts.[22,23] For example, helical proteins, which contain contacts that are mostly local in the sequence, tend to fold faster than β proteins. Others have compared the logarithm of the folding rate to linear[29,55] or square-root functions of the chain length.[1,21–24,29,56] In Supporting Information, we show a few such correlations on our test set of 93 proteins. The present work gives a mechanistic explanation for such observations. In the Foldon Funnel Model, secondary structures form fast (but they unravel even faster, since secondary structures are unstable on their own), but, because they form sequentially, more secondary structures take more time. Hence the folding time τ, $1/k_f(L)$, increases with chain length L (because L and N are linearly related; see Fig. 6). We believe that the CO is simply a surrogate for the effect of protein size (L or N) because folding rates only correlate with the *absolute contact order* (which is proportional to L) and not with the *relative contact order* (which is independent of L); see Supporting Information for further discussion.

What Is the Nature of Folding Cooperativity?—The present model recognizes three types of folding cooperativity: from the formation of secondary structures (in K_2), from the additional stabilization when secondary structures assemble into tertiary structures (in K_3), and from packing into the native state (in K_f). Fig. 7 shows the model prediction that small proteins tend to be more stabilized by packing, while larger ones are more stabilized by tertiary interactions.

The present model only treats how folding rates depend on protein size and not otherwise on the protein's amino acid sequence. However, it is well-known that the effects of the sequence can be large. This can be seen from the broad scatter around the fit line in Fig. 3. Some structurally similar proteins (identical N) have folding rates that differ by orders of magnitude. An example is the spectrin superfamily; these proteins have very different folding rates despite nearly identical chain lengths, secondary structure counts, and topologies.[57] Another example is the homeodomain superfamily.[58] Our data set includes both the spectrin and homeodomain helix bundles, but our focus on protein size and global fitting prevents us from predicting the sequence-dependent variation of rates within each family.

Estimating the Folding Kinetics of Proteomes

Finally, we are able to place a cell's protein folding kinetics in the context of other rate processes in the cell. As noted above, protein folding times can be estimated simply from the length L of the protein chain. Because protein-length distributions are known for many cellular proteomes, we can estimate the folding time distribution of whole cellular proteomes. Of course, such an estimate must be crude at the present time. For one thing, many proteins have multiple domains,[59] yet only very little is yet known about multidomain folding rates.[60] Nevertheless, we combined our Foldon Funnel Model prediction for folding rates, $k_f(L)$, with the known protein-length distribution $p(L)$ for the *Escherichia coli* proteome, and made our best estimate of the effects of domains to compute the approximate distribution of folding times for the *E. coli* proteome shown in Fig. 8. [1] It shows that the median protein in *E. coli* folds on the 5 s timescale, and also that there is a large variance. Fig. 8 also indicates a few other timescales that are relevant to the cell: the left line (dark blue) indicates the roughly 16 s that is required to synthesize an average *E. coli* protein (325 amino acids × 0.05 seconds to add each amino acid in translation[62]); the middle line (orange) indicates the roughly 30 s it takes for *E. coli*'s GroEL chaperones to refold a protein (a protein spends about 10 s in the chaperone cavity, and takes about 3 recycling events to fold[63,64]); and the right line (teal) indicates *E. coli*'s minimum doubling time of 20 minutes. Until the folding of larger and multidomain proteins is better understood, this distribution should be regarded as nothing more than just a simple estimate of folding times relative to other cellular landmark timescales.

However, the figure also illuminates a huge gap in our current knowledge–how do large domains fold? Over 600 of the proteins are predicted to fold on timescales slower than the doubling time, due to large, slow-folding domains (> 400 amino acids). One explanation is that these large domains may actually be made up of subdomains that fold independently, even though current domain annotations treat them as single domains. It also seems likely that many factors may mitigate problems from slow folding times, including chaperones, folding on the ribosome, and kinetic cooperativity between protein domains.

Conclusions

We have developed a simple but general mechanistic model of protein folding kinetics. The Foldon Funnel Model posits that secondary structures are the units of folding assembly, that they are relatively unstable, that isolated units flicker in and out of structure, and that individual secondary structures are stabilized and escorted along the folding route by neighboring secondary structures that lead to tertiary structure. It predicts that the free energy landscape of two-state folders is volcano-shaped: uphill for the first structures formed, and only downhill in the last step to the native state. Transition states are found to be nested: later structures are bottlenecks for earlier structures. The model is consistent with

[1] We use the domain annotations from the SUPERFAMILY database,[61] which contains domain annotations for 3003 out of the 4228 proteins in the *E. coli* proteome. In the absence of better information, we assume that each domain folds as an independent unit. (Because domains stabilize each other, the principal error introduced here will be to underestimate the folding rates of multi-domain proteins.) We approximate the folding time of each of the 3003 annotated proteins as the folding time of its largest (and slowest) domain. Here, we use the rate of the slowest domain as an approximation because we just want a rough orders-of-magnitude comparison of folding times.

both general observations on small proteins, namely that they are two-state (single exponential), that tertiary contacts give stability and cooperativity to equilibrium native structures, that localness of the native structure correlates with folding speed, and with the observed nonlinear dependence of the logarithm of folding rate on number of secondary structures on a test set of 93 proteins.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship and by NSF grant PHY 1205881. The authors thank the Laufer Center for support and Sarina Bromberg for creating Fig. 1 and the helix bundle graphics for Fig. 4 and Fig. 5. The authors also thank Doug Barrick, Walter Englander, Daniel Farrell, Kingshuk Ghosh, Adam de Graff, T.J. Lane, Jie Liang, Justin MacCallum, Susan Marqusee, and Bob Matthews for helpful discussions.

References

1. Garbuzynskiy SO, Ivankov DN, Bogatyreva NS, Finkelstein AV. Proc. Natl. Acad. Sci. U. S. A. 2013; 110:147–150. [PubMed: 23251035]
2. Levinthal CJ. Chim. Phys.-Chim. Biol. 1968; 65:44–45.
3. Ikai A, Tanford C. Nature. 1971; 230:100–102. [PubMed: 4927005]
4. Anfinsen C. Science. 1973; 181:223–230. [PubMed: 4124164]
5. Privalov P. Adv. Prot. Chem. 1979; 33:167–241.
6. Kim P, Baldwin R. Annu. Rev. Biochem. 1982; 51:459–489. [PubMed: 6287919]
7. Segawa S, Sugihara M. Biopolymers. 1984; 23:2473–2488. [PubMed: 6518262]
8. Kim P, Baldwin R. Annu. Rev. Biochem. 1990; 59:631–660. [PubMed: 2197986]
9. Wolynes P, Onuchic J, Thirumalai D. Science. 1995; 267:1619–1620. [PubMed: 7886447]
10. Finkelstein A, Badretdinov A. Fold. Des. 1997; 2:115–121. [PubMed: 9135984]
11. Karplus M, Weaver D. Biopolymers. 1979; 18:1421–1437.
12. Karplus M, Weaver D. Protein Sci. 1994; 3:650–668. [PubMed: 8003983]
13. Burton R, Myers J, Oas T. Biochemistry. 1998; 37:5337–5343. [PubMed: 9548914]
14. Baldwin R, Rose G. Trends Biochem.Sci. 1999; 24:26–33. [PubMed: 10087919]
15. Rumbley J, Hoang L, Mayne L, Englander S. Proc. Natl. Acad. Sci. U. S. A. 2001; 98:105–112. [PubMed: 11136249]
16. Islam S, Karplus M, Weaver DJ. Mol. Biol. 2002; 318:199–215.
17. Islam S, Karplus M, Weaver D. Structure. 2004; 12:1833–1845. [PubMed: 15458632]
18. Englander SW, Mayne L, Krishna MMG. Q. Rev. Biophys. 2007; 40:287–326. [PubMed: 18405419]
19. Hu W, Walters BT, Kan Z-Y, Mayne L, Rosen LE, Marqusee S, Englander SW. Proc. Natl. Acad. Sci. U. S. A. 2013; 110:7684–7689. [PubMed: 23603271]
20. Plaxco K, Simons K, Baker DJ. Mol. Biol. 1998; 277:985–994.
21. Galzitskaya O, Garbuzynskiy S, Ivankov D, Finkelstein A. Proteins. 2003; 51:162–166. [PubMed: 12660985]
22. Ivankov D, Garbuzynskiy S, Alm E, Plaxco K, Baker D, Finkelstein A. Protein Sci. 2003; 12:2057–2062. [PubMed: 12931003]
23. Ouyang Z, Liang J. Protein Sci. 2008; 17:1256–1263. [PubMed: 18434498]
24. Naganathan A, Munoz V. J. Am. Chem. Soc. 2005; 127:480–481. [PubMed: 15643845]
25. Gong H, Isom D, Srinivasan R, Rose G. J. Mol. Biol. 2003; 327:1149–1154. [PubMed: 12662937]

26. Gromiha M. J. Chem Inf. Model. 2005; 45:494–501. [PubMed: 15807515]
27. Zou T, Ozkan SB. Phys. Biol. 2011; 8:066011. [PubMed: 22146659]
28. Rustad M, Ghosh K. J. Chem. Phys. 2012:137.
29. Lane TJ, Pande VS. PLoS ONE. 2013; 8:e78606. [PubMed: 24339865]
30. Zwanzig R, Szabo A, Bagchi B. Proc. Natl. Acad. Sci. U. S. A. 1992; 89:20–22. [PubMed: 1729690]
31. Zwanzig R. Proc. Natl. Acad. Sci. U. S. A. 1995; 92:9801–9804. [PubMed: 7568221]
32. Munoz V, Eaton W. Proc. Natl. Acad. Sci. U. S. A. 1999; 96:11311–11316. [PubMed: 10500173]
33. Alm E, Baker D. Proc. Natl. Acad. Sci. U. S. A. 1999; 96:11305–11310. [PubMed: 10500172]
34. Galzitskaya O, Finkelstein A. Proc. Natl. Acad. Sci. U. S. A. 1999; 96:11299–11304. [PubMed: 10500171]
35. Alm E, Morozov A, Kortemme T, Baker D. J. Mol. Biol. 2002; 322:463–476. [PubMed: 12217703]
36. Henry E, Eaton W. Chem. Phys. 2004; 307:163–185.
37. Bruscolini P, Naganathan AN. J. Am. Chem. Soc. 2011; 133:5372–5379. [PubMed: 21417380]
38. De Sancho D, Munoz V. Phys. Chem. Chem. Phys. 2011; 13:17030–17043. [PubMed: 21670826]
39. Mello C, Barrick D. Proc. Natl. Acad. Sci. U. S. A. 2004; 101:14102–14107. [PubMed: 15377792]
40. Barrick D, Ferreiro DU, Komives EA. Curr. Opin. Struct. Biol. 2008; 18:27–34. [PubMed: 18243686]
41. Street TO, Barrick D. Protein Sci. 2009; 18:58–68. [PubMed: 19177351]
42. Aksel T, Barrick D. Methods Enzymol. 2009; 455:95–125. [PubMed: 19289204]
43. Cortajarena AL, Mochrie SGJ, Regan L. Protein Sci. 2011; 20:1042–1047. [PubMed: 21495096]
44. Ghosh K, Dill KA. J. Am. Chem. Soc. 2009; 131:2306–2312. [PubMed: 19170581]
45. Dill K, Chan H. Nat. Struct. Biol. 1997; 4:10–19. [PubMed: 8989315]
46. Ghosh K, Dill KA. Proc. Natl. Acad. Sci. U. S. A. 2009; 106:10649–10654. [PubMed: 19541647]
47. Dill KA, Ghosh K, Schmit JD. Proc. Natl. Acad. Sci. U. S. A. 2011; 108:17876–17882. [PubMed: 22006304]
48. Wright P, Dyson H, Lerner R. Biochemistry. 1988; 27:7167–7175. [PubMed: 3061450]
49. Chakrabartty A, Kortemme T, Baldwin RL. Protein Sci. 1994; 3:843–852. [PubMed: 8061613]
50. Fersht A. Curr. Opin. Struct. Biol. 1997; 7:3–9. [PubMed: 9032066]
51. Ghosh K, Ozkan SB, Dill KA. J. Am. Chem. Soc. 2007; 129:11920–11927. [PubMed: 17824609]
52. Dill KA, MacCallum JL. Science. 2012; 338:1042–1046. [PubMed: 23180855]
53. Guinn EJ, Kontur WS, Tsodikov OV, Shkel I, Record MT. Proc. Natl. Acad. Sci. U. S. A. 2013; 110:16784–16789. [PubMed: 24043778]
54. Adhikari AN, Freed KF, Sosnick TR. Proc. Natl. Acad. Sci. U. S. A. 2012; 109:17442–17447. [PubMed: 23045636]
55. Lane TJ, Pande VS. J. Phys. Chem. B. 2012; 116:6764–6774. [PubMed: 22452581]
56. Thirumalai D. J. Phys. I. 1995; 5:1457–1467.
57. Wensley BG, Gaertner M, Choo WX, Batey S, Clarke J. J. Mol. Biol. 2009; 390:1074–1085. [PubMed: 19445951]
58. Gianni S, Guydosh N, Khan F, Caldas T, Mayor U, White G, DeMarco M, Daggett V, Fersht A. Proc. Natl. Acad. Sci. U. S. A. 2003; 100:13286–13291. [PubMed: 14595026]
59. Han J-H, Batey S, Nickson AA, Teichmann SA, Clarke J. Nat. Rev. Mol. Cell Biol. 2007; 8:319–330. [PubMed: 17356578]
60. Arora P, Hammes GG, Oas TG. Biochemistry. 2006; 45:12312–12324. [PubMed: 17014084]
61. Gough J, Karplus K, Hughey R, Chothia C. J. Mol. Biol. 2001; 313:903–919. [PubMed: 11697912]
62. Young R, H, B. Biochem J. 1976; 160:185–194. [PubMed: 795428]
63. Ewalt K, Hendrick J, Houry W, Hartl F. Cell. 1997; 90:491–500. [PubMed: 9267029]
64. Horwich AL, Fenton WA. Q. Rev. Biophys. 2009; 42:83–116. [PubMed: 19638247]

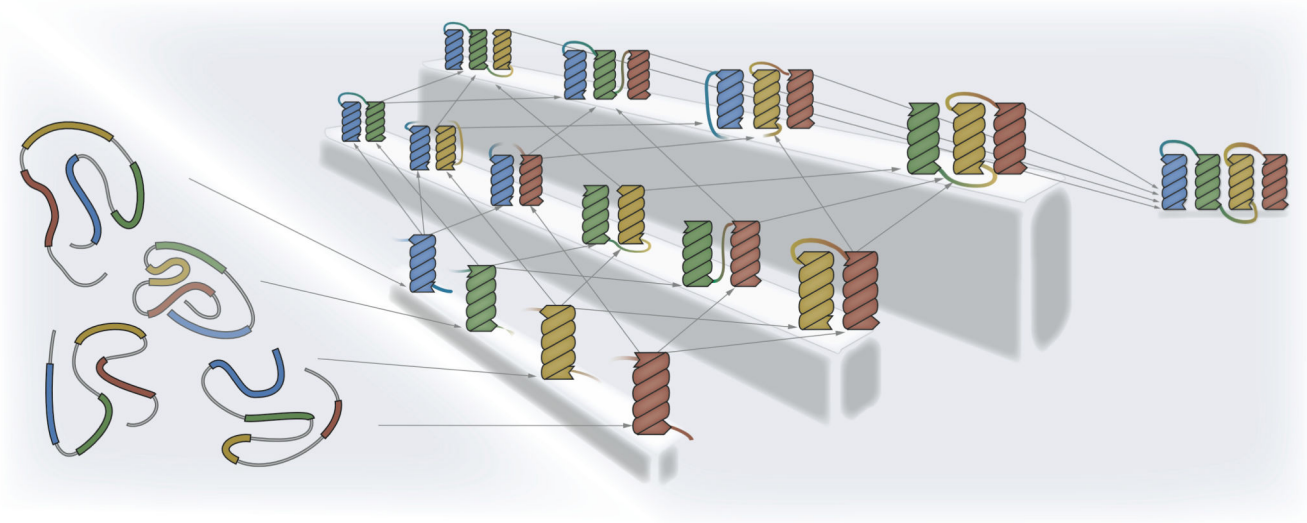

Figure 1.

Premises of the model: (1) Each secondary structure folds independently of others. (2) Tertiary structure forms as pairs of secondary structures. The folding process is a sequential accretion of secondary structure elements. (3) Routes are combinatoric: different secondary structures assemble along different trajectories.

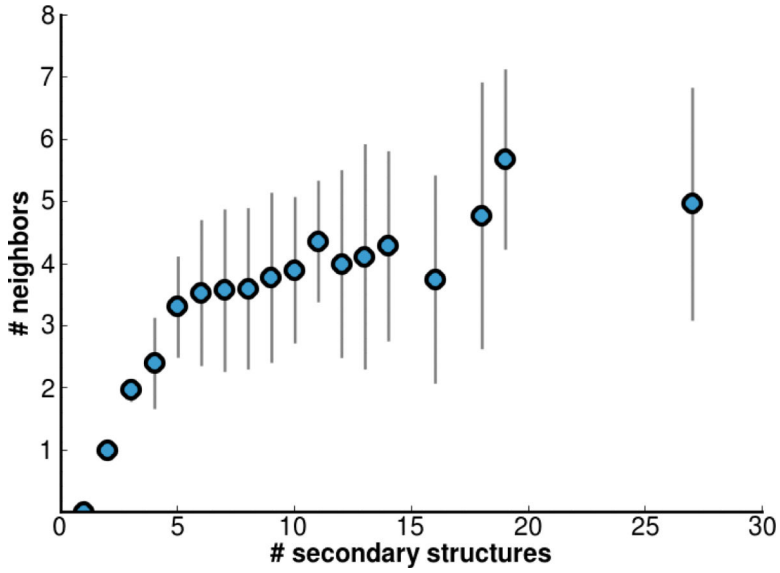

Figure 2.

Distribution of nearest-neighbors of a given secondary structural element in a protein, as a function of the total number of secondary structures in that protein. A pair of secondary structures are neighbors if they have at least 1 residue-residue contact. Residue contacts were determined from a centroid for each residue with a cutoff of 8 Å. The plot is based on the 93 proteins in our data set.

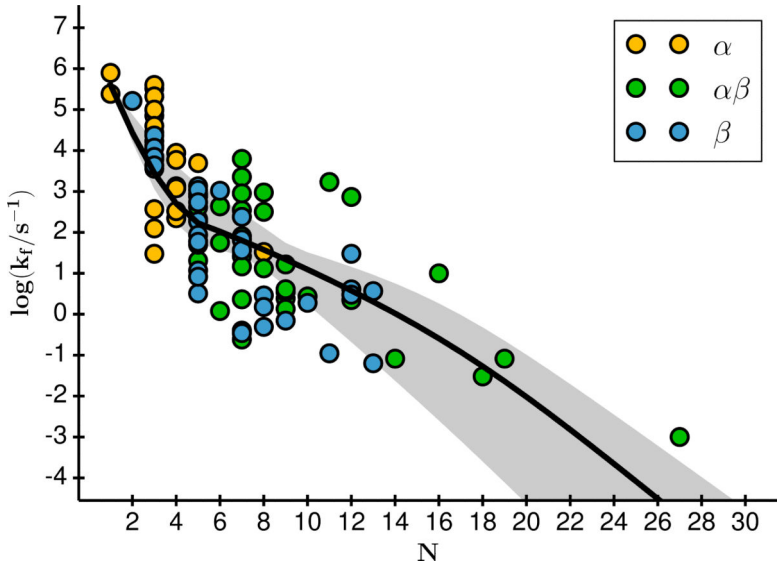

Figure 3.

Folding rates predicted by model. Folding rate vs number of secondary structures, *N*. The colored points are experimental values, and they are colored by structural class. The black line is the prediction from the model, and the gray bands represent the 95% confidence interval. The black line represents a perfect fit to the data. 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).

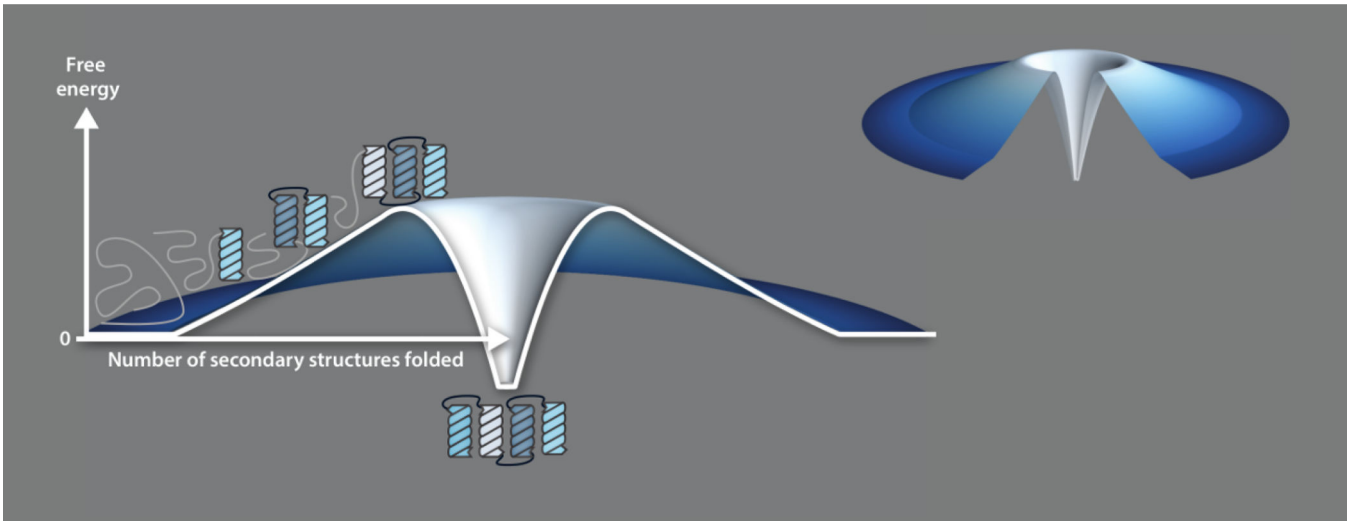

Figure 4.

Free energy landscape of Foldon Funnel Model. The radial distance corresponds to c, the number of folded secondary structures. The height of the landscape corresponds to free energy ($\Delta G(c) = -RT\ln[p_c(eq)]$). The initial flat region on the outer edge represents $c = 0$, and the start of the climb represents the $c = 0$ to $c = 1$ transition. The center of the landscape represents the folded state, $c = N$. The landscape was computed from the best-fit parameters described in the Results section for an $N = 4$ protein. The slope of the volcano is relatively linear. On the one hand, K_3 reduces the steepness at each step relative to only K_2 terms alone, but the combinatoric term essentially compensates that increase.

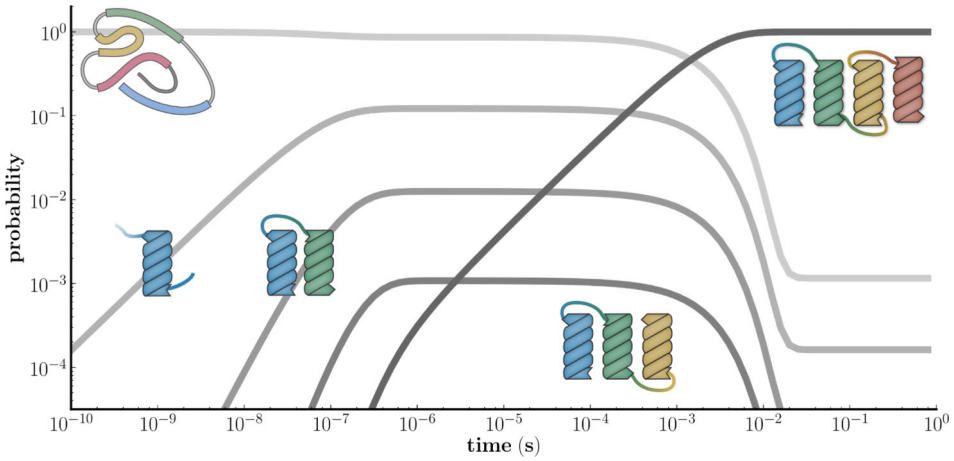

Figure 5.

Dynamics of a four-helix bundle. As the folding reaction proceeds, the probability of occupying each intermediate state rises and falls as the protein traverses its free energy landscape from the unfolded $c = 0$ state to the folded $c = 4$ state. The folding trajectory was computed by numerically integrating the kinetic master equation (see SI) for an $N = 4$ protein, using parameters: $K_2 = 0.037$, $K_3 = 1.96$, $k_1 = 10^{5.6} s^{-1}$, and $K_f = 5.23$.

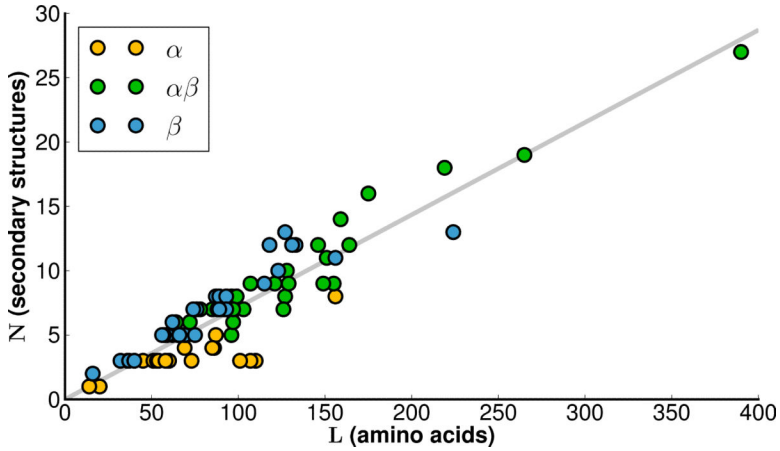

Figure 6.

Number of secondary structures vs chain length for the 93 proteins in our data set. Our fit line is $N = \gamma L$, where $\gamma = 0.0718$ secondary structures per amino acid. $R^2 = 0.85$. The slope of the line corresponds to an average of ≈ 14 amino acids per secondary structure. However, this fit includes loops, so it represent an overestimate of average secondary structure length.

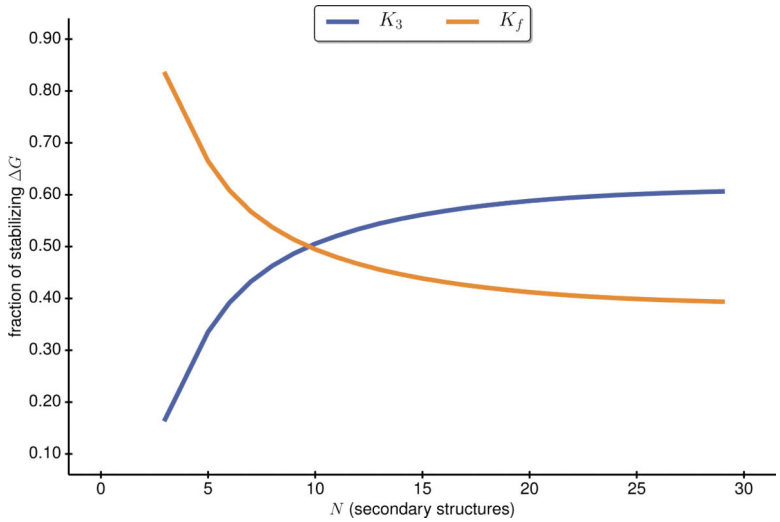

Figure 7.

Fractional contribution of K_3 and K_f to stabilization of folded state vs. number of secondary structures. For small proteins, the packing term K_f stabilizes the folded state more than the tertiary interction term K_3. For larger proteins, the reverse is true: K_3 contributes more than K_f.

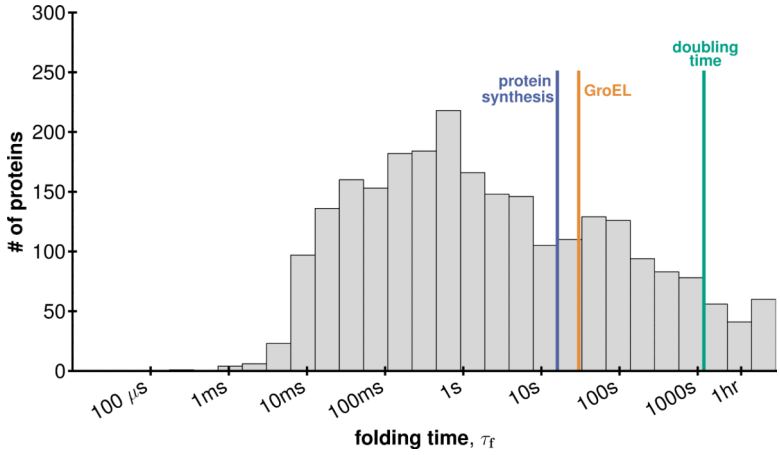

Figure 8.

E. coli folding time distribution. Colored lines indicate time scales for key cellular processes: (dark blue) ribosomal protein synthesis, (orange) GroEL refolding, (teal) doubling time.