doi:10.1016/j.jmb.2004.02.055 J. Mol. Biol. (2004) 338, 573–583

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Characterization of the Folding Energy Landscapes of Computer Generated Proteins Suggests High Folding Free Energy Barriers and Cooperativity may be Consequences of Natural Selection

Michelle Scalley-Kim[1] and David Baker[2]*

[1]*Molecular and Cellular Biology Program University of Washington Seattle, WA, USA*

[2]*Department of Biochemistry Howard Hughes Medical Institute, University of Washington, Seattle WA 98195, USA*

**Corresponding author*

To determine the extent to which protein folding rates and free energy landscapes have been shaped by natural selection, we have examined the folding kinetics of five proteins generated using computational design methods and, hence, never exposed to natural selection. Four of these proteins are complete computer-generated redesigns of naturally occurring structures and the fifth protein, called Top7, has a computer-generated fold not yet observed in nature. We find that three of the four redesigned proteins fold much faster than their naturally occurring counterparts. While natural selection thus does not appear to operate on protein folding rates, the majority of the designed proteins unfold considerably faster than their naturally occurring counterparts, suggesting possible selection for a high free energy barrier to unfolding. In contrast to almost all naturally occurring proteins of less than 100 residues but consistent with simple computational models, the folding energy landscape for Top7 appears to be quite complex, suggesting the smooth energy landscapes and highly cooperative folding transitions observed for small naturally occurring proteins may also reflect the workings of natural selection.

Keywords: protein folding; natural selection; folding kinetics; protein design; free energy landscape

Introduction

Over the past 15 years the protein folding field has been revolutionized by the acquisition of experimental data on the folding of a large number of small, single domain proteins with much simpler folding mechanisms than the larger proteins studied earlier.[1–4] In parallel, there has been a flourishing of theoretical work seeking to explain the broad trends and generalities found in the experimental studies.[5–8] However, the application of theory to the experimental data is complicated by the fact that proteins are the results of the natural evolutionary process, and it is difficult to disentangle the properties of naturally occurring proteins that are general consequences of the physical chemistry of polypeptide chains that fold from those that reflect the workings of natural selection. It is not clear whether a theory which seeks to account for the experimental data on naturally occurring proteins can focus exclusively on the physical chemistry of polypeptide chains and their dynamics, or whether it is necessary to also model the effects of natural selection. The question of the extent to which evolutionary history shapes biophysical and biological phenomena is, in fact, quite general to the theoretical and computational modeling of biological systems.

A property of particular interest is the rate of protein folding. It is clear that proteins are under selective pressure for both stability and function, but to what extent does nature select for proteins that are able to fold quickly into their native states? What benefits may be conferred to the cell by fast protein folding kinetics? It is clearly necessary that proteins are able to fold in a biologically relevant time-scale, and slow folding proteins may have an increased susceptibility towards proteolysis and non-specific aggregation.

Abbreviations used: ΔG_{UNF}, free energy of unfolding; RMSD, root-mean-squared deviation.

E-mail address of the corresponding author: dabaker@u.washington.edu

The only way to disentangle the features of the folding of small naturally occurring proteins that reflect the fundamental properties of folded polypeptide chains from those that reflect the workings of natural selection is to study the folding of proteins that have not been generated by the natural evolutionary process. For more complex biological processes the study of analogous processes not influenced by natural selection is generally not possible, providing no clear route for extricating evolutionary history from more fundamental constraints. However, in the case of protein folding there are two alternative approaches. The first is to generate novel proteins by random sequence generation followed by selection for the very small subset of sequences that fold to stable states (one cannot simply study the properties of randomly generated sequences without selection, as only an infinitesimally small fraction of these will adopt stable discrete structures).[9] Our laboratory used such an approach several years ago and generated variants of two small proteins, protein L and the src SH3 domain, by randomizing portions of the proteins' sequences and selecting for variants which still folded to the native structures using a phage display selection strategy (in the case of the SH3 domain, only a five letter amino acid alphabet was used, I, K, E, A, and G).[10,11] For both proteins, the novel variants were found to fold as fast or faster than the naturally occurring proteins, suggesting that the sequences of small proteins are not optimized for fast folding. While powerful, the random sequence generation and selection strategy is limited. Because of the finite size of the random libraries that can be constructed (less than 10^9 distinct variants) there is an upper limit to the number of residues that can be changed simultaneously and the selection process, itself, may have biases that are difficult to isolate.

A second approach is to use computational protein design methods to identify sequences that are compatible with a given three-dimensional structure. The first complete redesign of the sequence of a naturally occurring protein was by Mayo and co-workers of a small zinc finger protein.[12] More recently, as part of a large-scale test of design methods, our laboratory has completely redesigned the sequences of nine small, naturally occurring protein structures.[13] In these computational protein redesign approaches, the starting sequence is completely random and either a deterministic algorithm, such as dead end elimination, or a stochastic method, such as a Monte Carlo search, is used to identify low energy sequences for a given structure. In either case, the optimization process is focused entirely on the stability of the native state and is completely ignorant of the folding process. Since the only selection operating is for the stability of the native state, the possible biases due to the evolutionary history present for naturally occurring proteins or due to the selection strategy for proteins identified in large combinatorial libraries are completely absent. There has been one study of the folding of computationally redesigned homeodomain sequences, and it was found that the folding rates of the redesigned proteins were quite similar to that of the naturally occurring homeodomain.[14]

Here, we first extend the previous kinetic study of the redesigned, all α-helical homeodomain proteins by investigating the folding rates and folding energy landscapes of three α-helical/β-sheet containing proteins and one all β-sheet containing protein produced by complete sequence redesign using the structures of naturally occurring proteins as design templates. Second, we go beyond the limitation of using a naturally occurring protein structure as a design template and describe the folding kinetics of a protein with both a computer-generated sequence and a computer-generated topology, neither of which have been found in nature. The results of these studies suggest that folding rates are not subject to evolutionary optimization, one of the computer redesigned proteins folds a striking 10,000-fold faster and another tenfold faster than their naturally occurring counterparts, but that other properties of the folding energy landscape may indeed be under selective pressure.

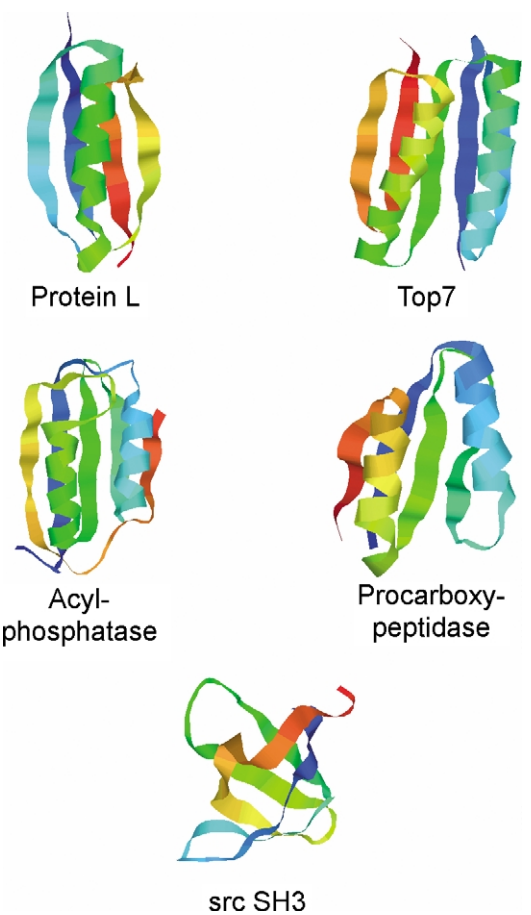

Figure 1. Models of the redesigned proteins. The PDB codes for the parent sequences are: protein L (1hz5, 1–62), acylphosphatase (2acy, 1–98), pro-carboxypeptidase (1aye, 10–79), src SH3 (1fmk, 83–142), and Top7 (1qys).

Results

The protein folding kinetics of five proteins that have never been exposed to natural selection were examined (Figure 1). Three of the proteins, pro-carboxypeptidase, protein L, and acylphosphatase, come from a previous study in which the sequences of nine globular proteins were redesigned using RosettaDesign, a computer program developed for protein design.[13] Starting with randomly chosen conformations (rotamers) for randomly chosen amino acids at each position, RosettaDesign uses a Monte Carlo optimization method in which a single move consists of changing an amino acid rotamer at a randomly chosen position to a randomly chosen rotamer of the same or different amino acid. Moves are accepted according to the standard Metropolis criterion using an energy function dominated by a Lennard–Jones potential, an orientation dependent hydrogen bonding potential, and an implicit solvation model. The procedure converges on very low energy sequences after approximately 10^6 attempted moves (about 5–10 cpu minutes). Starting from the experimentally determined structures of the naturally occurring proteins, RosettaDesign identified low energy sequences for pro-carboxypeptidase, protein L, and acylphosphatase that were, on average, 35% identical to the wild-type sequences (Table 1). The proteins were expressed, purified, and found to display circular dichroism (CD) spectra, thermal and chemical denaturation behavior, and NMR spectra consistent with well-folded proteins, making them excellent candidates for this study.[13]

The large scale design efforts of Dantas *et al.* did not yield a folded redesigned protein containing solely β-sheet secondary structure.[13] In order to add diversity to the protein folds examined here, a second, and this time successful attempt was made to redesign the src SH3 fold (see Materials and Methods). The redesigned sequence (Table 1) shares 52% overall identity and 86% core identity with the wild-type sequence. The redesigned SH3 domain was well expressed and found to be a monomer in gel filtration experiments (results not shown). While circular dichroism spectra of the SH3 domains are difficult to interpret, a change in intrinsic fluorescence of the redesigned SH3 domain is observed with increasing denaturant concentration (Figure 2), allowing for stability measurements. While the new design is quite destabilized compared to the wild-type SH3 domain (the free energy of unfolding (ΔG_{UNF}) is 4.1 kcal mol^{-1} and 1.3 kcal mol^{-1} for wild-type src SH3 and the redesigned SH3, respectively), the 1D NMR spectrum of the redesigned SH3 domain is compatible with a well-folded, rigid structure with chemical shifts that are consistent with β-sheet structure (Figure 2). The redesigned SH3

Table 1. Alignments of the wild-type and redesigned sequences

	10 20 30 40 50
Acylphosphatase, wild type	AEGDTLISVDYEIFGKVQGVFFRKYTQAEGKKLGLVGWVQNTDQGTVQGQLQGPASKVRH
Acylphosphatase, designed	PTGDSYIQVKWQVKGDVTGNNFRKMVAEFAEALGLVGKVTYTDNGTVSGQVEGPAEQVLK
	70 80 90
Acylphosphatase, wild type	MQEWLETKGSPKSHIDRASFHNEKVIVKLDYTDFQIVK
Acylphosphatase, designed	FLEWLARSGSPNADIKQTVFTNMTRIDRLTMETFKIDE
	10 20 30 40 50 60
Procarboxypeptidase, wild type	DQVLEIVPSNEEQIKNLLQLEAQEHLQLDFWKSPTTPGETAHVRVPFVNVQAVKVFLESQGIAYSIMIED
Procarboxypeptidase, designed	KTIFVIVPTNEEQVAFLEALAKQDELNFDWQNPPTEPGQPVVILIPSDMVEWFLEMLKAKGIPFTVYVEE
	10 20 30 40 50
src SH3, wild type	VTTFVALYDYESRTETDLSFKKGERLQIVNNTEGDWWLAHSLSTGQTGYIPSNYVAPSDS
src SH3, designed	KVRFVASDSYTSTGDNDLSFTKGEKLWIKDNAEGDYWFAVSDQTGKTGYIPSDKVYPDGK
	10 20 30 40 50 60
Protein L, wild type	EVTIKANLIFANGSTQTAEFKGTFEKATSEAYAYADTLKKDNGEWTVDVADKGYTLNIKFAG
Protein L, designed	DTTVRVIFIFADGKTTTIEFTGSEEAAKKQAQEYAQSLRDNYGDYSIDYQNGGELIKIVFSG

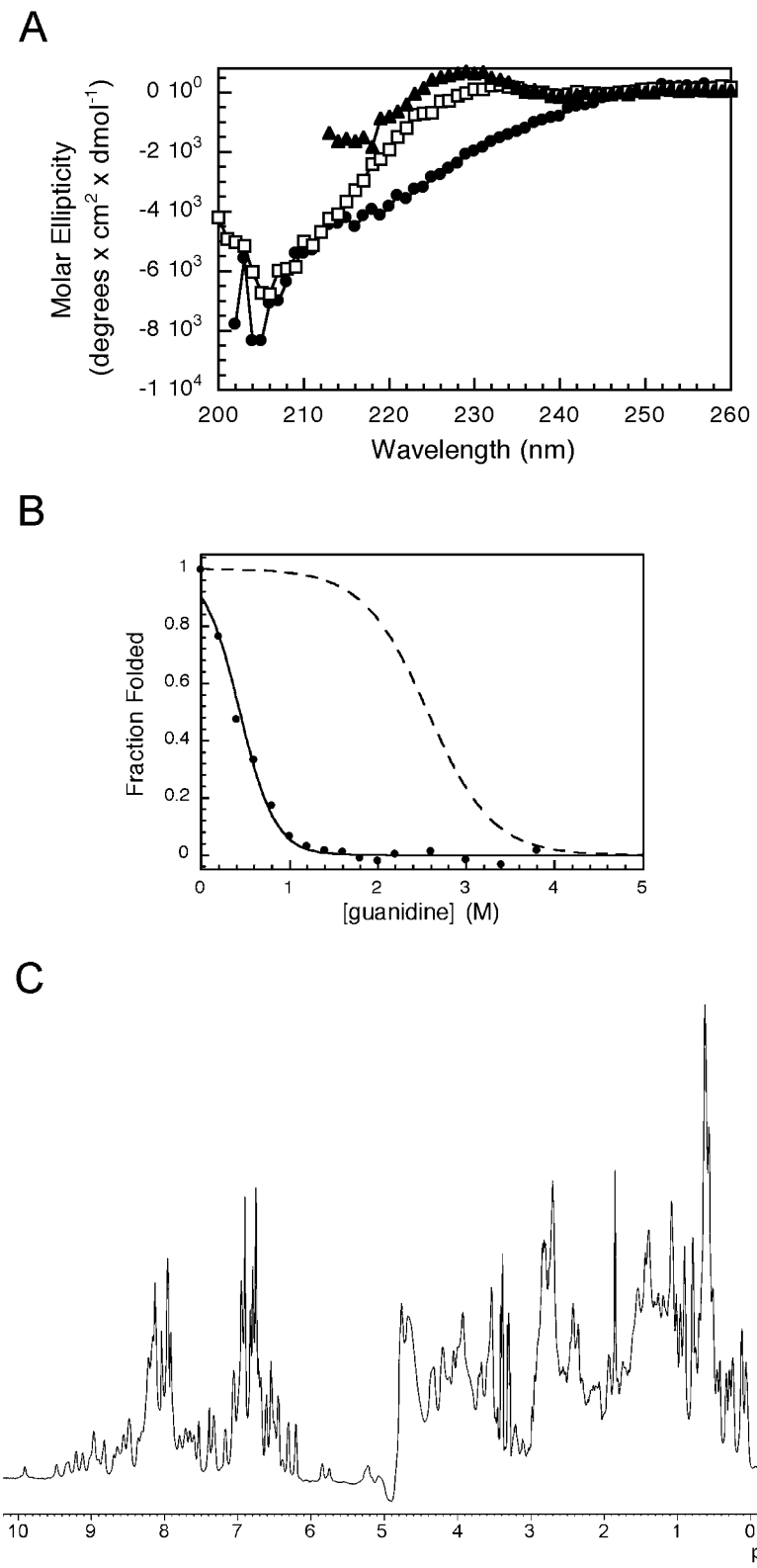

Figure 2. Characterization of the redesigned SH3 domain. A, CD spectra of the redesigned SH3 domain in 50 mM sodium phosphate (pH 7) at 25 °C (open squares) and 80 °C (filled circles) and in 7.5 M guanidine, 50 mM sodium phosphate (pH 7) at 25 °C (filled triangles). The spectrum taken under native conditions has a minimum at 208 nm, atypical for an all β structure, and no change is observed upon the addition of denaturant. Similar behavior has been observed for other SH3 domains and it has been suggested that local interactions between two tryptophan residues may be responsible for the aberration. B, Guanidine-induced denaturation of the redesigned SH3 domain was followed by fluorescence emission at 341 nm. The continuous line represents the best fit of the data to a two-state model with a linear dependence of the free energy of folding on the denaturant concentration. Since the 1D NMR spectra of the redesigned SH3 domain suggested that the protein was fully folded in the absence of denaturant, the folded baseline value for the guanidine denaturation melt was fixed at the fluorescence value observed in 0 M denaturant. The broken line is a representation of the guanidine-induced denaturation of the wild-type SH3 domain according to the values published by Grantcharova & Baker.[31] C, 1D NMR spectra of the redesigned SH3 domain.

domain is the fourth of the five proteins examined in this study.

The fifth protein studied here, called Top7, was the result of *de novo* flexible backbone computational design calculations aimed at generating a novel fold that has yet to be observed in nature.[15] The computational procedure combined the sequence optimization method described above with the Rosetta structure prediction methodology to iteratively optimize both the amino acid sequence and backbone coordinates of the designed protein. Such an iterative approach was necessary as it is unlikely that for any fixed, arbitrarily chosen structure there exists a very low

energy sequence. The Top7 protein was found to be well folded and unusually stable; the guanidine unfolding transition begins at ~6 M guanidine. X-ray crystallography studies of Top7 indicated that the sequence adopts a structure remarkably close to the design model: the root-mean-squared deviation (RMSD) between the Top7 design model and the crystal structure is only 1.17 Å.[15] As both the sequence and the structure of Top7 are the result of computer optimization for stability and have not been influenced by the natural selection process in any way, Top7 is particularly relevant for our study.

To facilitate kinetic measurements using stopped-flow fluorescence, mutations were made to Top7 and the redesigned protein L and acylphosphatase sequences. The mutation of a partially buried phenylalanine in Top7 and a partially buried tyrosine in the protein L redesign to tryptophan residues (Top7 F81W and pL2 Y34W) resulted in a change in intrinsic fluorescence for both proteins upon unfolding. The mutations did not result in significant changes in stability; ΔG_{UNF} is 13.2 kcal mol^{-1} and 13.6 kcal mol^{-1} for Top7 and the Top7 F81W variant, respectively, and 4.6 kcal mol^{-1} and 5.1 kcal mol^{-1} for the protein L

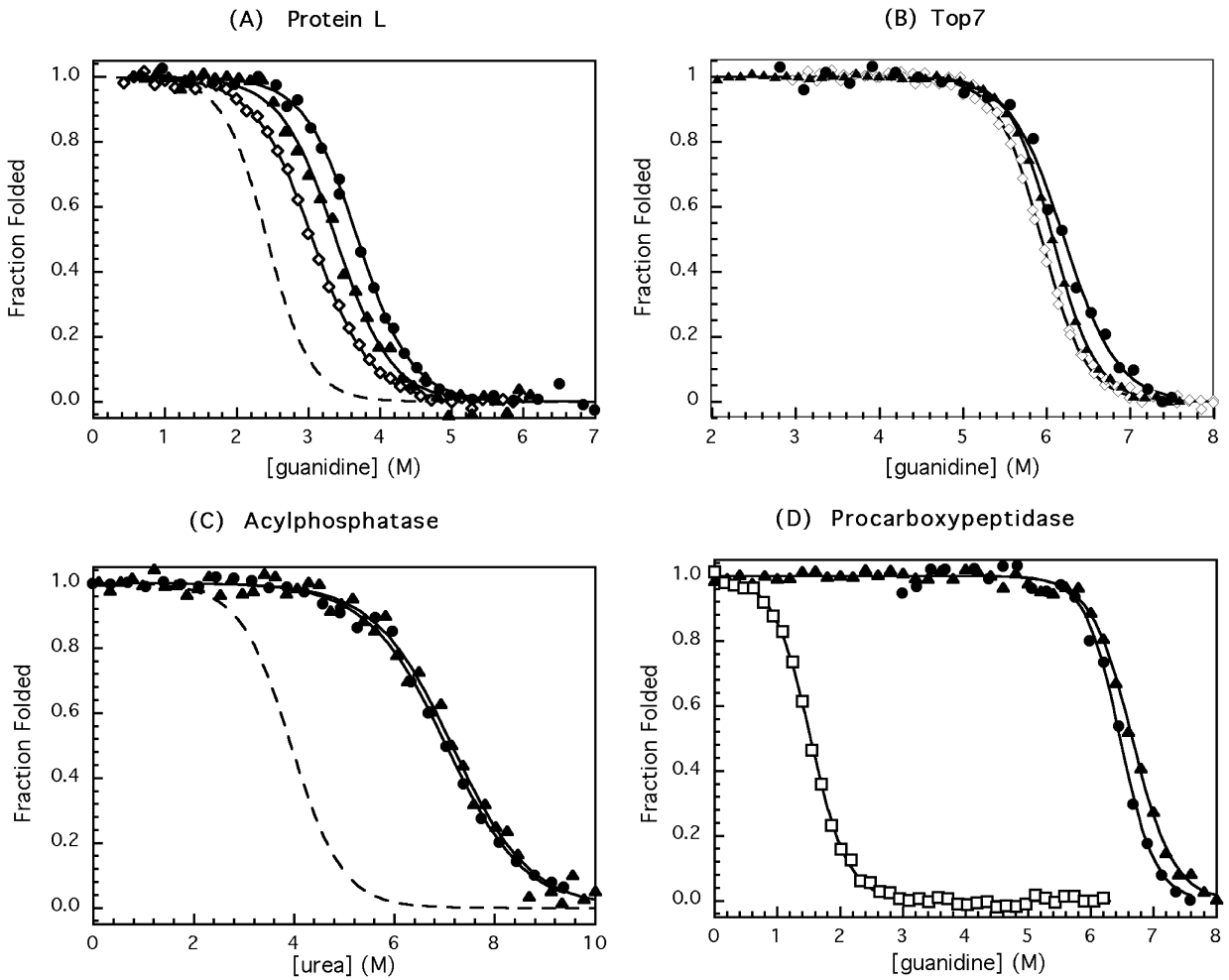

Figure 3. Equilibrium denaturation of the redesigned proteins. (A) The guanidine induced denaturation of the protein L redesign was followed by CD at 220 nm (open diamonds) and of the Y34W mutation by CD at 220 nm (filled triangles) and fluorescence emission at 350 nm (filled circles). The broken line is a representation of the guanidine-induced denaturation of wild-type protein L according to the values published by Scalley *et al.*[30] (see Table 2 for values). (B) The guanidine induced denaturation of Top7 was followed by CD at 220 nm (open diamonds) and of the F81W mutation by CD at 220 nm (filled triangles) and fluorescence emission at 370 nm (filled circles). (C) The urea induced denaturation of the acylphosphatase redesign containing the W64L mutation was followed by CD at 220 nm (filled triangles) and fluorescence emission at 320 nm (filled circles). The broken line is a representation of the urea-induced denaturation of wild-type acylphosphatase according to the values published by Chiti *et al.*[32] (see Table 2 for values). (D) The guanidine induced denaturation of wild-type pro-carboxypeptidase was followed by CD at 220 nm (open squares) and of the pro-carboxypeptidase redesign by CD at 220 nm (filled triangles) and fluorescence emission at 345 nm (filled circles). To allow for comparison, all buffers were chosen to match those used in the previous characterization of the wild-type proteins; the protein L, pro-carboxypeptidase, and Top7 redesigns were studied in 50 mM sodium phosphate (pH 7) and the acylphosphatase redesign was studied in 50 mM sodium acetate (pH 5.5). All data were fit using a two-state model and the resulting free energy of unfolding values are given in Table 2.

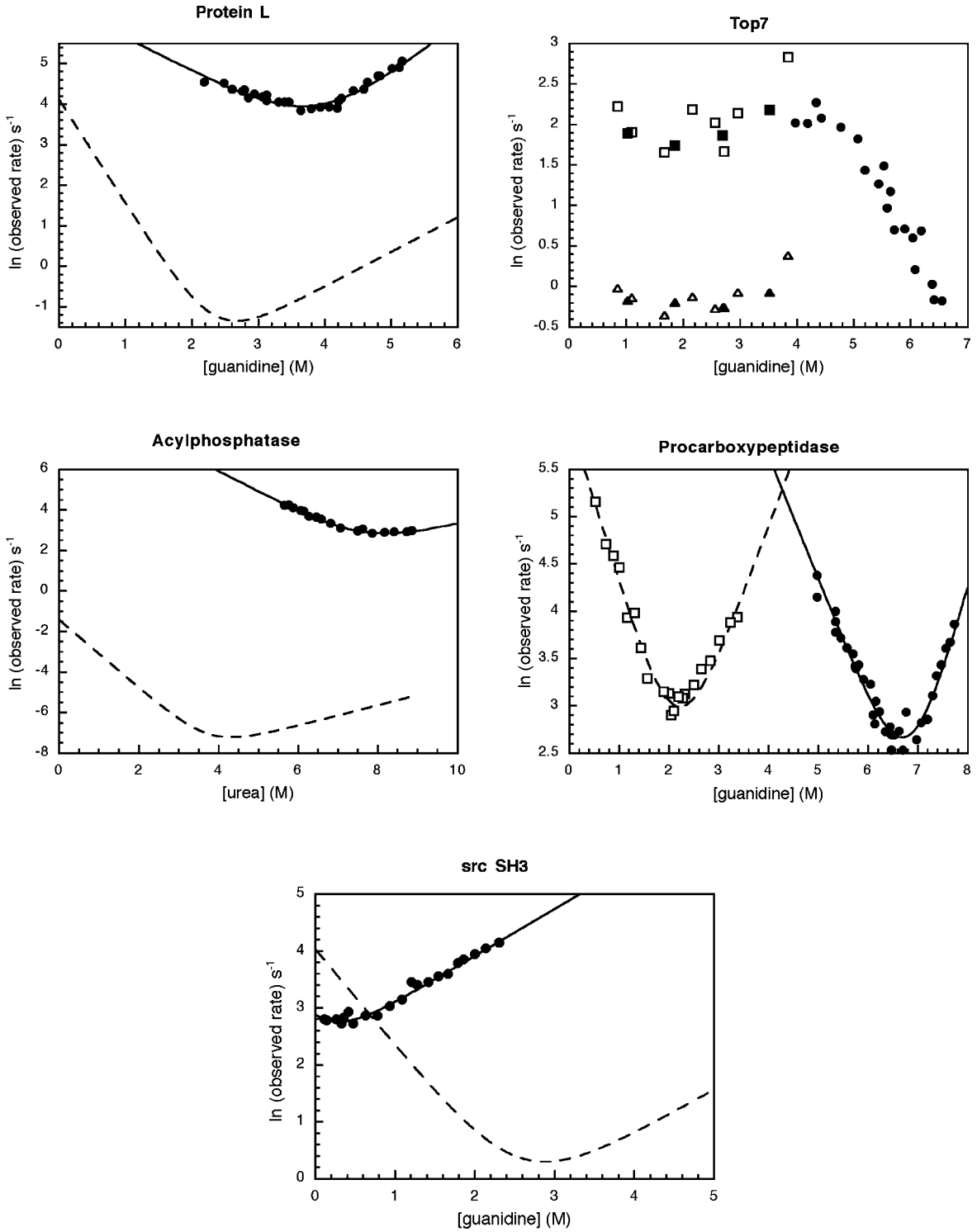

Figure 4. Denaturant dependence of folding kinetics of the redesigned proteins. The observed folding and unfolding rates for the protein L, acylphosphatase, SH3 and pro-carboxypeptidase redesigns (filled circles) at various denaturant concentrations were fit to a two-state model (continuous lines; see equation (1)). The chevron plots for wild-type proteins were recreated using previously published m_f,m_u,k_f^{H2O}, and k_u^{H2O} values (broken lines; see Table 2 for values), except for the wild-type pro-carboxypeptidase (open squares) whose folding and unfolding rates were experimentally determined as described in Materials and Methods. To allow for comparison, all buffers were chosen to match those used in the previous characterization of the wild-type sequences; the protein L, pro-carboxypeptidase, SH3 and Top7 redesigns were studied in 50 mM sodium phosphate (pH 7) and the acylphosphatase redesign was studied in 50 mM sodium acetate (pH 5.5). The folding rates of Top7 were fit well to a single exponential between 4 M and 6.5 M guanidine (filled circles). Below 4 M guanidine, the folding kinetics became bi-exponential (slow phase, open triangles; fast phase, open squares). To monitor the dependence of the folding rates of Top7 on protein concentration, the folding rates were also monitored at higher protein concentrations (final concentration ~45 μM, filled squares and filled triangles for the fast and slow phases, respectively).

redesign of the protein Y34W redesign variant, respectively. Additionally, one of the two tryptophan residues in the acylphosphatase redesign was changed to a leucine residue (W64L) in order to reduce the high amount of background fluorescence observed for the protein. Again, this mutation had little effect on stability; ΔG_{UNF} is 5.6 kcal mol^{-1} and 5.4 kcal mol^{-1} for the acylphosphatase redesign and the W64L variant, respectively. These variants were used for the remainder of the experiments described here and are referred to by the names of the parent sequences.

To determine whether secondary and tertiary structure losses during unfolding were synchronous, we monitored the equilibrium denaturation of the sequences using both fluorescence and CD (Figure 3). Typically, superimposability of equilibrium denaturation curves from CD and fluorescence experiments is a good indicator that a protein follows a two-state folding mechanism. Unfortunately, this analysis could not be performed for the redesigned SH3 sequence: like other SH3 domains, there was little change in the CD signal upon unfolding. The melts of the other four designed proteins are, for the most part, superimposable, with the greatest deviation present for the redesigned protein L sequence (as the *m*-values associated with the two transitions are very similar, the difference in midpoint is likely the result of subtle variations in experimental conditions). As shown here and previously, 1D NMR spectra for three of the redesigned sequences, SH3, protein L, and pro-carboxypeptidase, display good dispersion and sharp peaks, indicating that the redesigned proteins have well-packed, native-like cores, while the peaks of the redesigned acylphosphatase 1D NMR spectra are somewhat broader.[13] The ID NMR spectra of Top7 also displayed good dispersion and sharp peaks and the high-resolution crystal structure confirms the close, complementary packing of side-chains in the core of the protein.[15]

Folding and unfolding rates for the designed proteins were measured at different guanidine concentrations using stopped-flow fluorescence (see Materials and Methods; Figure 4). The folding rates at the mid-point of the folding transition and the extrapolated values for folding and unfolding rates in the absence of denaturant are listed in Table 2. For most of the designs, the denaturant dependence of both the folding and unfolding rates could not be fully determined as the folding and unfolding rates exceeded the limits of the stopped-flow instrument at either very low or very high denaturant concentrations, necessitating long extrapolations of the folding and unfolding rates in the absence of denaturant. However, agreement between the free energy estimates obtained from the extrapolated kinetic data (ΔG_{kin}) and from the independent equilibrium data (ΔG_{eq}) supports the validity of the extrapolations (Table 2).

For both the protein L and acylphosphatase redesigns, the folding rates in the absence of

Table 2. Equilibrium and kinetic parameters for wild-type and redesigned proteins

Protein	k_f^{H2O} (s^{-1})	k_u^{H2O} (s^{-1})	$k_{transition}$ (s^{-1})	m_f (kcal mol^{-1} M^{-1})	m_u (kcal mol^{-1} M^{-1})	$m_f/(m_f + m_u)$	ΔG_{kin} (kcal mol^{-1})	ΔG_{eq} (kcal mol^{-1})[a]
Acylphosphatase, redesigned[b]	2.2×10^4	0.3	26.4	0.6	0.3	0.7	6.6	5.4
Acylphosphatase, wild-type[c]	0.2	6.5×10^{-5}	8.9×10^{-4}	1	0.3	0.8	4.8	5.1
Pro-carboxypeptidase, redesigned	5.1×10^4	1.9×10^{-5}	13.8	0.8	1.1	0.4	12.7	11.8
Pro-carboxypeptidase, wild-type	408	0.5	34.9	1.0	0.8	0.6	4.0	3.0
Src SH3, redesigned	6.8	9.6	15.6	–	0.8		−0.2	1.3
Src SH3, wild-type[d]	56.7	0.1	1.6	1.0	0.5	0.7	3.8	4.1
Protein L, redesigned[e]	555	0.2	52.3	0.5	0.7	0.4	4.6	5.1
Protein L, wild-type[f]	60.6	0.02	0.3	1.5	0.5	0.8	4.7	4.6

[a] All reported ΔG_{eq} values are from CD experiments, except for SH3 wild-type and redesigned sequences, where reported values are from a fluorescence experiment.
[b] Containing W64L mutation.
[c] Chiti *et al.*[32]
[d] Grantcharova & Baker.[31]
[e] Containing Y34W mutation.
[f] Scalley *et al.*[30]

denaturant (k_f^{H2O}) are much faster than those of the wild-type proteins; k_f^{H2O} of redesigned protein L and acylphosphatase are enhanced approximately tenfold and 10^5-fold, respectively. As both of these redesigned proteins have stabilities similar to the wild-type proteins, they necessarily unfold faster; the unfolding rates in the absence of denaturant (k_u^{H2O}) of redesigned protein L and acylphosphatase are enhanced approximately tenfold and 4×10^4-fold, respectively. Consistent with its increase in stability, the pro-carboxypeptidase redesign folds much faster and unfolds more slowly than the wild-type protein; $k_f^{H2O} = 5.1 \times 10^4\ s^{-1}$, ~1000-fold faster than wild-type, and $k_u^{H2O} = 1.9 \times 10^{-5}\ s^{-1}$, ~20,000-fold slower than wild-type. Conversely, compatible with its decrease in stability, the SH3 redesign folds slower and unfolds faster than wild-type; $k_f^{H2O} = 6.8\ s^{-1}$, ~10-fold slower than wild-type, and $k_u^{H2O} = 9.6\ s^{-1}$, ~100-fold faster than wild-type. A decrease in the ratio $m_f/(m_f + m_u)$, an indicator of the amount of hydrophobic burial in the transition state compared to the native state, for the acylphosphatase, pro-carboxypeptidase, and protein L redesigns suggests that the transition state ensemble may be shifted towards the unfolded state for the redesigned sequences.

Since the redesigned proteins in this study have different stabilities than their wild-type counterparts, it is useful to compare folding rates at the transition mid-point where the free energies are equivalent (Table 2; this also eliminates the need for long extrapolations). The observed rates at the transition mid-point are significantly faster for the acylphosphatase, protein L, and SH3 redesigns compared to the wild-type proteins, suggesting a preferential stabilization of the redesigned proteins' folding transition states. Only the pro-carboxypeptidase redesign folds at a similar rate at the transition mid-point as the wild-type protein.

Top7's folding behavior is markedly different from that of the other designed proteins (Figure 4). Due to the high concentration of denaturant necessary to unfold Top7, unfolding rates were not measured. At higher denaturant concentrations, the folding rates are fit well by a single exponential and decrease linearly as the denaturant concentration increases. However, at lower denaturant concentrations the folding kinetic traces become double exponential. Both of the rate constants are independent of denaturant concentration with the faster rate leveling off at $\sim 6\ s^{-1}$ and the slower rate at $\sim 0.8\ s^{-1}$. To determine whether the double exponential behavior was caused by protein aggregation under near-native conditions, we investigated the protein concentration dependence of the folding rates at lower guanidine concentrations. Increasing the protein concentration by ~4-fold had no effect on either of the kinetic phases (Figure 4, filled squares and triangles), suggesting that the unusual behavior was not due to protein aggregation.

There are at least three possible explanations for the dramatic decrease in the denaturant dependence at low guanidine concentrations. First, under the more stabilizing conditions, partially folded intermediates or kinetic traps with hydrophobic surface area burial comparable to that of the transition state ensemble may become populated, resulting in a loss of denaturant dependence, since little change in hydrophobic burial would be experienced going from the partially folded intermediates to the transition state. The population of such intermediates would contrast with the high degree of cooperativity that is usually observed in the folding of small, naturally occurring proteins. Two other possibilities are that the position of the folding transition state moves significantly toward the unfolded state at low denaturant concentrations or that there is a significant increase in internal friction at low denaturant concentrations.[16]

Discussion

We find that naturally occurring proteins are not highly optimized for fast folding. Three of the four redesigned proteins, protein L, acylphosphatase, and pro-carboxypeptidase, are found to fold faster than the wild-type proteins in the absence of denaturant. More importantly, accounting for stability differences, three out of the four redesigned proteins, protein L, acylphosphatase, and src SH3, are found to fold faster at their transition mid-point than the wild-type proteins. These results are consistent with those of the studies discussed in Introduction.[10,11,14]

Why do the designed proteins fold faster than their naturally occurring counterparts? The finding that the SH3, acylphoshatase, and protein L redesigns fold faster than the wild-type proteins at the transition mid-point, suggests that the transition state ensemble is preferentially stabilized in the redesigned proteins. The design method favors an increase in hydrophobic core volume through peripheral core build-up, potentially leading to alternate ways to stabilize partially folded conformations and, hence, broadening the transition state ensemble. Serrano and co-workers observed a similar increase in both folding and unfolding rates upon mutation of polar surface residues to non-polar residues in the α-spectrin src SH3 domain.[17] Furthermore, hydrophobic core mutations in the α-spectrin src SH3 domain, resulting in overpacked cores, were found to both fold and unfold faster than the wild-type protein, suggesting a similar preferential stabilization of the transition state.[18]

The slower folding and unfolding rates of native proteins may be an indirect consequence of selection against aggregation and for population of a single functional state rather than a more molten ensemble of states. Selection for both properties may have increased the amount of buried polar interactions and reduced the number of peripheral

hydrophobic residues, both of which could increase folding/unfolding free energy barriers. The design process in contrast in selecting for the stability of the native structure typically disfavors buried polar interactions and favors peripheral hydrophobic residues. Hence evolutionary selection for functional non- aggregating proteins could produce proteins with folding behaviors very different from those produced by computational optimization of native state stability.

The folding rates of naturally occurring proteins are correlated with the average sequence separation between contacting residues in the three-dimensional structure (the contact order).[19] The folding rates of all of the wild-type proteins examined here are close to the values expected from the contact order-folding rate correlation. The folding rate in the absence of denaturant for Top7, extrapolated using the folding rates between 5 M and 6.5 M guanidine (where the folding rate is linearly dependent on guanidine concentration), is similar to the value expected from the contact order-folding rate correlation (results not shown). Interestingly, the folding rates in the absence of denaturant of two of the redesigned proteins, procarboxypeptidase and acylphosphatase, are significantly faster than expected given their contact order. In the adaptation of Zwanzig's simple model of folding that was used to initially rationalize the observation of the contact order-folding rate correlation,[19] the rate of folding is correlated with the contact order for proteins with the same overall stability, but increases with increasing stability for proteins with the same contact order (this will be the case in any model in which only native interactions are favorable, increasing stability necessarily also lowers the free energy of the folding transition state relative to the unfolded state). The contact order—folding rate correlation is presumably observed for naturally occurring proteins because evolutionary selection pressure has led to a fairly narrow range of stabilities so that the contact order, rather than the strength of the attractive interactions favoring folding, is the dominant factor contributing to the observed rate variance. The increase in folding rates for the redesigned proteins is likely to reflect the increase in the extent of favorable native interactions beyond what is typical in naturally occurring proteins: the folding rates of both redesigned proteins at the denaturation midpoints (where the free energy of folding is zero) are close to those of naturally occurring proteins with similar contact orders at their denaturation midpoints (data not shown).[20]

The folding kinetics of Top7 were found to be considerably more complex than those of naturally occurring, small, single domain proteins. Two kinetic phases are observed during refolding which are independent of guanidine concentration and protein concentration over the tested range. This complexity is likely to reflect the population of partially structured intermediates or kinetic traps. While denaturant independence of folding rates at low denaturant concentrations has been observed for other proteins, barnase,[21,22] ribonuclease A,[23] and hen lysozyme,[24] it is a fairly rare observation and is typically found over a much narrower range of denaturant (0–1 M) concentrations than is seen for Top7 (0–4 M). These results suggest that the striking cooperativity of the folding reactions of small, naturally occurring proteins may not be a necessary feature of protein folding, but rather an evolutionary advantageous adaptation for reducing aggregation of partially folded species during folding.

There is an interesting parallel between our comparison of the folding kinetics of computer generated and naturally occurring proteins and recent comparisons of the folding of pair additive *versus* more highly cooperative, native state centric (Go) computational models of folding.[25,26] The pair additive models, like Top7, show significant rollover in the folding rate and increasing kinetic complexity under more stabilizing conditions, while the more highly cooperative models have the simple exponential kinetics and linear relationships between stability and folding rates typically found for native proteins.[26] One interpretation of this parallel is that the design methodology captures pair additive but not higher-order contributions to native state stability and specificity, and indeed the energy function used in the design process is completely pair additive (this is essential to the efficient sampling of sequence and conformational space needed for the design of a novel protein like Top7). Alternatively, since Top7 is exceptionally stable, partially folded forms of Top7 may be relatively stable, and population of such states during folding could contribute to the complex kinetics. It is possible that natural selection has operated to destabilize partially folded forms of naturally occurring proteins because of the possibility for inappropriate interactions with other proteins; in the computational design process there is clearly no such negative selection.

In summary, by examining the folding kinetics of proteins generated through computational design rather than the natural evolutionary process we have identified selective pressures that may be acting on the folding processes of naturally occurring proteins. Our results suggest that proteins generated through selection for function in the natural evolutionary process may have higher folding free energy barriers than proteins generated by computational optimization for stability. Furthermore, in the first characterization of the folding of a globular protein with a topology not found in nature, we find complex folding kinetics almost never observed for small, single domain proteins, suggesting that rough energy landscapes with partially folded intermediates and kinetic traps are possible for small proteins, but may be under negative selection pressure, perhaps because of the potential for inappropriate and possible deleterious interactions of partially folded species with other cellular components.

These conclusions, however, should be viewed as somewhat tentative as even the relatively small number of designed proteins whose folding has been examined to date exhibit a great diversity of folding behaviors. For example, among the redesigns of naturally occurring scaffolds, three (protein L, acylphosphatase, and pro-carboxypeptidase) are more stable and one (SH3) less stable than their wild-type counterparts; two (protein L and acylphosphatase) have lower free energy barriers, one has a similar barrier (SH3) and one has a higher free energy barrier (pro-carboxypeptidase) than the naturally occurring proteins, and in two cases the degree of solvation of the denatured state has changed (protein L and acylphosphatase) and in one it has not (pro-carboxypeptidase). More clear delineating of the differences in the consequences for folding of evolutionary selection for function *versus* computational selection for stability will require continued characterization in the coming years of the folding of the whole new world of designed proteins which may someday grow to match the diversity found in nature.

Materials and Methods

Protein design, mutagenesis, and purification

The src SH3 sequence was redesigned using an iterative method described.[13,27] In the first step, 1000 low free energy sequences for the SH3 backbone were produced in independent Monte Carlo searches in which all 20 amino acid residues were allowed at each position, except position 37 which was restricted to a tryptophan residue, allowing for fluorescence measurements. The rotamers sampled in the first step were restricted to the standard Dunbrack library.[28] In a second set of 1000 independent Monte Carlo searches, an extended rotamer library was used and the amino acid residues allowed at each site were restricted to those observed in the first step, producing 1000 new sequences. Additionally, proline residues were only allowed in native positions. The top scoring sequence from the second set was chosen for further study and a synthetic gene was obtained from BlueHeron Technologies (Seattle, WA).

Previously reported equilibrium and kinetic data for wild-type pro-carboxypeptidase were obtained using urea denaturation.[29] Since sufficiently high urea concentrations could not be obtained to denature the redesigned pro-carboxypeptidase sequence, we obtained a synthetic gene for wild-type pro-carboxypeptidase from BlueHeron Technologies, allowing for expression, purification, and characterization of the wild-type pro-carboxypeptidase folding kinetics in guanidine.

All of the redesigned genes and the synthetic pro-carboxypeptidase gene, were cloned into the pet29b(+) expression vector (Novagen). Point mutations were introduced into Top7 and the redesigned protein L and acylphosphatase sequences using Stratagene's QuikChange kit. The proteins were expressed and purified as described.[13]

NMR

The 1D spectrum of ~800 μM redesigned SH3 protein in 90% 50 mM sodium phosphate (pH 6.0), 10% $^2H_20$ was recorded at 298 K at 500 MHz.

Equilibrium measurements

Fluorescence denaturation melts were taken in a Spex Fluorolog 2 spectrofluoremeter using a 1 cm cuvette. The excitation wavelength was 280 nm and the emission wavelength was monitored using the constant wavelength averaging function (see Figure legends for specific emission wavelengths). Slit widths of 1 mm were used for both excitation and emission monochrometers. All measurements were taken at 295(±1) K.

Circular dichroism denaturation melts were monitored using an Aviv CD spectrometer 16A DS and a Hamilton syringe titrating device. A 1 cm cuvette was used and the observation cell was thermostated to 295(±0.2) K using a Peltier device.

Kinetic measurements

All stopped flow kinetic data were obtained using a BioLogic SFM4/QFM4, and fit to a single exponential using the Biokine analysis software. All folding and unfolding reactions were carried out at 295(±1) K using a circulating water bath. Fluorescence measurements were made with an excitation wavelength of 280 nm using a 0.8 mm cuvette. The fluorescence emission was measured using a 309 nm cutoff filter (Oriel) for the redesigned acylphosphatase, SH3, protein L, and pro-carboxypeptidase sequences and a 380 nm cutoff filter (Oriel) for Top7. Folding and unfolding experiments were, otherwise, conducted as described.[30]

The folding and unfolding data were fit according to a two-state model:

$$\ln k_{\text{obs}} = \ln\left[k_{\text{f}}^{\text{H}_20}\exp\left(\frac{-m_{\text{f}}[\text{denaturant}]}{RT}\right) + k_{\text{u}}^{\text{H}_20}\exp\left(\frac{-m_{\text{u}}[\text{denaturant}]}{RT}\right)\right] \quad (1)$$

The observed relaxation rate, k_{obs}, at any given guanidine concentration is the sum of the folding and unfolding rates (k_{f} and k_{u}); the logarithms of these rate constants are assumed to be linear functions of the denaturant concentration with coefficients m_{f} and m_{u}, respectively. $k_{\text{f}}^{\text{H2O}}$ and $k_{\text{u}}^{\text{H2O}}$ are the rates of folding and unfolding in the absence of denaturant.

Acknowledgements

We thank Brain Kuhlman, Tania Kortemme, Hue Sun Chan, and Kevin Plaxco for thoughtful review of the manuscript. We also thank Gabriele Varani for aid in obtaining NMR spectra. This work was supported in part by a grant from the National Institutes of Health (to D.B.) and by the PHS NRSA T32 GM07270 from NIGMS (to M.S.-K.).

References

1. Jackson, S. E. (1998). How do small single-domain proteins fold? *Fold. Des.* **3**, R81–R91.

2. Bilsel, O. & Matthews, C. R. (2000). Barriers in protein folding reactions. *Advan. Protein Chem.* **53**, 153–207.
3. Radford, S. E. (2000). Protein folding: progress made and promises ahead. *Trends Biochem. Sci.* **25**, 611–618.
4. Eaton, W. A., Munoz, V., Hagen, S. J., Jas, G. S., Lapidus, L. J., Henry, E. R. & Hofrichter, J. (2000). Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 327–359.
5. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995). Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins: Struct. Funct. Genet.* **21**, 167–195.
6. Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995). Principles of protein folding—a perspective from simple exact models. *Protein Sci.* **4**, 561–602.
7. Shakhnovich, E. I. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29–40.
8. Onuchic, J. N., Nymeyer, H., Garcia, A. E., Chahine, J. & Socci, N. D. (2000). The energy landscape theory of protein folding: insights into folding mechanisms and scenarios. *Advan. Protein Chem.* **53**, 87–152.
9. Scalley-Kim, M., Minard, P. & Baker, D. (2003). Low free energy cost of very long loop insertions in proteins. *Protein Sci.* **12**, 197–206.
10. Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. & Baker, D. (1997). Functional rapidly folding proteins from simplified amino acid sequences. *Nature Struct. Biol.* **4**, 805–809.
11. Kim, D. E., Gu, H. & Baker, D. (1998). The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl Acad. Sci. USA*, **95**, 4982–4986.
12. Dahiyat, B. I. & Mayo, S. L. (1997). *De novo* protein design: fully automated sequence selection. *Science*, **278**, 82–87.
13. Dantas, G., Kuhlman, B., Callender, D., Wong, M. & Baker, D. (2003). A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* **332**, 449–460.
14. Gillespie, B., Vu, D. M., Shah, P. S., Marshall, S. A., Dyer, R. B., Mayo, S. L. & Plaxco, K. W. (2003). NMR and temperature-jump measurements of *de novo* designed proteins demonstrate rapid folding in the absence of explicit selection for kinetics. *J. Mol. Biol.* **330**, 813–819.
15. Kuhlman, B., Dantas, G., Ireton, G., Varani, G., Stoddard, B. & Baker, D. (2004). *De novo* design of a novel globular protein fold with atomic level accuracy. *Science*, **302**, 1364–1368.
16. Kaya, H. & Chan, H. S. (2002). Towards a consistent modeling of protein thermodynamic and kinetic cooperativity: how applicable is the transition state picture to folding and unfolding? *J. Mol. Biol.* **315**, 899–909.
17. Viguera, A. R., Vega, C. & Serrano, L. (2002). Unspecific hydrophobic stabilization of folding transition states. *Proc. Natl Acad. Sci. USA*, **99**, 5349–5354.
18. Ventura, S., Vega, M. C., Lacroix, E., Angrand, I., Spagnolo, L. & Serrano, L. (2002). Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nature Struct. Biol.* **9**, 485–493.
19. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994.
20. Ivankov, D. N., Garbuzynshiy, S. O., Alm, E., Plaxco, K. W., Baker, D. & Finkelstein, A. V. (2003). Contact order revisted: influence of protein size on the folding rate. *Protein Sci.* **12**, 2057–2062.
21. Matouschek, A., Kellis, J. T., Jr, Serrano, L., Bycroft, M. & Fersht, A. R. (1990). Transient folding intermediates characterized by protein engineering. *Nature*, **346**, 440–445.
22. Chu, R. A. & Bai, Y. (2002). Lack of definable nucleation sites in the rate-limiting transition state of barnase under native conditions. *J. Mol. Biol.* **315**, 759–770.
23. Houry, W. A., Rothwarf, D. M. & Scheraga, H. A. (1995). The nature of the initial step in the conformational folding of disulphide-intact ribonuclease A. *Nature Struct. Biol.* **2**, 495–503.
24. Kiefhaber, T. (1995). Kinetic traps in lysozyme folding. *Proc. Natl Acad. Sci. USA*, **92**, 9029–9033.
25. Kaya, H. & Chan, H. S. (2003). Solvation effects and driving forces for protein thermodynamic and kinetic cooperativity: how adequate is native-centric topological modeling? *J. Mol. Biol.* **326**, 911–931.
26. Kaya, H. & Chan, H. S. (2003). Simple two-state protein folding kinetics requires near-levinthal thermodynamic cooperativity. *Proteins: Struct. Funct. Genet.* **52**, 510–523.
27. Kuhlman, B. & Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl Acad. Sci. USA*, **97**, 10383–10388.
28. Dunbrack, R. L., Jr & Cohen, F. E. (1997). Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661–1681.
29. Villegas, V., Azuaga, A., Catasus, L., Reverter, D., Mateo, P. L., Aviles, F. X. & Serrano, L. (1995). Evidence for a two-state transition in the folding process of the activation domain of human pro-carboxypeptidase A2. *Biochemistry*, **34**, 15105–15110.
30. Scalley, M. L., Yi, Q., Gu, H., McCormack, A., Yates, J. R., III & Baker, D. (1997). Kinetics of folding of the IgG binding domain of peptostreptococcal protein L. *Biochemistry*, **36**, 3373–3382.
31. Grantcharova, V. P. & Baker, D. (1997). Folding dynamics of the src SH3 domain. *Biochemistry*, **36**, 15685–15692.
32. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005–1009.

Edited by M. Levitt

(Received 11 November 2003; received in revised form 4 February 2004; accepted 4 February 2004)