doi:10.1006/jmbi.1999.3663 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 298, 303–312

JMB

Evolutionary Conservation in Protein Folding Kinetics

Kevin W. Plaxco[1]*, Stefan Larson[3], Ingo Ruczinski[2], David S. Riddle[1] Edward C. Thayer[1], Brian Buchwitz[1], Alan R. Davidson[3,4] and David Baker[1]*

[1]*Department of Biochemistry and*

[2]*Department of Statistics University of Washington Seattle, WA 98195, USA*

[3]*Department of Molecular and Medical Genetics and*

[4]*Department of Biochemistry University of Toronto, Toronto Ontario, Canada, M5S 1A8*

**Corresponding authors*

The sequence and structural conservation of folding transition states have been predicted on theoretical grounds. Using homologous sequence alignments of proteins previously characterized *via* coupled mutagenesis/kinetics studies, we tested these predictions experimentally. Only one of the six appropriately characterized proteins exhibits a statistically significant correlation between residues' roles in transition state structure and their evolutionary conservation. However, a significant correlation is observed between the contributions of individual sequence positions to the transition state structure across a set of homologous proteins. Thus the structure of the folding transition state ensemble appears to be more highly conserved than the specific interactions that stabilize it.

Keywords: nucleation-condensation; phi-values; evolution

Introduction

Natural selection clearly conserves functionally important residues in proteins in order to preserve biological activity. As protein function generally requires a well-populated native state, natural selection likewise conserves residues critical to protein stability. Does evolution also maintain sequence in order to ensure rapid folding? Proteins spontaneously fold to their unique native structures many orders of magnitude more rapidly than would be possible if the folding process involved an exhaustive, random search of conformation space (Levinthal, 1968). This seemingly paradoxical behavior has led to the proposal that evolutionary pressures do select for those presumably rare sequences able to fold rapidly (Levinthal, 1968; Anfinsen, 1973), and that the identity of residues participating in the rate-limiting step of folding might be conserved (Shrivastava *et al.*, 1995; Bryngelson *et al.*, 1995; Shakhnovich *et al.*, 1996; Ptitsyn, 1998; Mirny *et al.*, 1998; Demirel *et al.*, 1998; Parker *et al.*, 1998; Kragelund *et al.*, 1999; Pouron & Mornon, 1999; Mirny & Shakhnovich, 1999).

Both experimental (reviewed by Fersht, 1997) and theoretical (reviewed by Pande *et al.*, 1998) studies suggested that protein folding occurs *via* the formation of a small region of native-like structure that serves as a nucleus upon which further residues condense in a process analogous to a phase transition. If evolution acts at the level of fine sequence details in order to produce and maintain rapid folding, conservation of residue identity and the structures of these nuclei might be expected within families of homologous proteins or across protein superfamilies (Shakhnovich *et al.*, 1996; Mirny *et al.*, 1998; Ptitsyn, 1998; Michnick & Shakhnovich, 1998; Mirny & Shakhnovich, 1999). In contrast, recent experimental reports have suggested that there may be little correlation between sequence conservation and participation in the folding transition state (Grantcharova *et al.*, 1998; Martinez *et al.*, 1998; Fulton *et al.*, 1999). Here, we report a more exhaustive and quantitative experimental examination of the hypothesized conservation of transition state sequence identity and structure.

Present addresses: K. W. Plaxco, Department of Chemistry and Biochemistry and Interdepartmental Program in Biochemistry and Molecular Biology, University of California, Santa Barbara CA 93106, USA; D. S. Riddle, Department of Immunology; Mayo Foundation, Rochester MN 55904, USA; E. C. Thayer, ZymoGenetics; 1201 Eastlake Ave. East, Seattle, WA 98102, USA.

Abbreviations used: CI2, chymotrypsin inhibitor 2; SH3, src homology 3; ADAh2, the activation domain of procarboxypeptidase A2; ACBP, acyl carrier binding protein; AcP, acyl phosphatase.

E-mail addresses of the corresponding authors: kwplaxco@u.washington.edu dabaker@u.washington.edu

0022-2836/00/020303–10 $35.00/0

The kinetic consequences of large numbers of point mutations have been reported for more than a dozen proteins. These include chymotrypsin inhibitor 2 (CI2), λ-repressor, the src homology 3 (SH3) domains of α-spectrin, src and fyn, the IgG binding domain of protein L, the activation domain of procarboxypeptidase A2 (ADAh2), acyl phosphatase (AcP), acyl carver binding protein (ACBP) and FKBP12 (Itzhaki *et al.*, 1995; Viguera *et al.*, 1996a; Burton *et al.*, 1997; Gu *et al.*, 1997; Martinez *et al.*, 1998; Grantcharova *et al.*, 1998; Villegas *et al.*, 1998; Kragelund *et al.*, 1999; Fulton *et al.*, 1999; Chiti *et al.*, 1999; A.R.D., unpublished data) which fold with two-state kinetics, and barnase, barstar, cheY and Arc repressor (Matouscheck *et al.*, 1992; Milla *et al.*, 1995; Lopez-Hernandez & Serrano, 1996; Nolting *et al.*, 1997), which fold *via* more complicated multi-state or bimolecular kinetics. With the exception of several λ-repressor mutations (Burton *et al.*, 1997), none of the reported sequence modifications appears to affect significantly the conformation of the folding transition state (as monitored by the effects of denaturants on folding), and thus the results of mutational studies can conveniently be interpreted as measuring the contributions of a residue to the structure of a relatively well-defined transition state ensemble.

Results

In order to access accurately the degree to which residues involved in the folding nucleus are conserved by evolution, a quantitative measure of the importance of a given residue to the structure of the folding nucleus is required. Some years ago Fersht and co-workers introduced Φ-value analysis as a convenient measure of this property (Fersht *et al.*, 1992). For mutations that do not significantly perturb the folding pathway, Φ corresponds to the ratio of the impact of a mutation on the stability of the transition state to its impact on the stability of the native state and is given by:

$$\Phi = \Delta\Delta G_{\ddagger-U}/\Delta\Delta G_{F-U} \quad (1)$$

While Φ-values sometimes misrepresent transition state involvement, for example, a residue with backbone atoms participating critically in the transition state might exhibit a low Φ-value if its side-chain points away from the structured regions of the transition state, it remains the only objective experimental measure of transition state participation.

Because a residue will rarely contribute more significantly to the stability of the transition state than to the native state, Φ-values rarely exceed 1. Mutations that increase the volume of a side-chain also sometimes produce negative Φ-values by increasing a protein's folding rates (by stabilizing the poorly packed transition state *via* non-native interactions) while destabilizing the tightly packed native structure. However, isosteric mutations or mutations that remove side-chain atoms that destabilize the folding transition state (i.e. decrease folding rates) also usually destabilize the native state, and thus the Φ-values of such mutations are rarely significantly less than 0. The Φ-values associated with side-chain truncating mutations that are significantly outside of the range 0-1.0 are most likely due to the large relative errors associated with measuring small changes in equilibrium free energy. We have omitted from our data set residues exhibiting Φ-values outside of the range −0.5 to 1.5 and preferentially selected conservative mutations decreasing side-chain volume (with the exception of glycine to alanine residue substitutions) in order to highlight any possible relationships between sequence conservation and Φ.

To assess the relationship between sequence conservation and Φ-values, it is essential to construct accurate and comprehensive sequence alignments. As described in Materials and Methods, exhaustive database searches employing PSI-BLAST (Altschul *et al.*, 1997) were used to ensure that alignments contained all available variants of the sequence under investigation. Once alignments were constructed, the average percentage identity of each sequence with every other sequence in the alignment was calculated. Sequences with average percentage identities below 20 % were discarded to avoid the inclusion of sequences that might not encode the correct fold. Each alignment was also carefully examined to verify that each included sequence possessed key features defining the fold (e.g. correctly positioned hydrophobic core residues). The overall diversity of each sequence alignment was determined by performing all-against-all pair-wise sequence identity calculations and determining the mean of these values. Of the six alignments, those of the SH3 domain and CheY are the largest (>265 sequences) and most diverse (mean pair-wise identities >30 %). Although some alignments contained relatively few sequences (e.g. ACBP and ADAh2), their diversity, as measured by the mean pair-wise sequence identity, was similar to some of the larger alignments (Table 1).

A means of quantifying sequence conservation is provided by informational entropy, which measures how significantly the residue identities in an alignment of homologous sequences deviate from a random distribution of amino acid residues (Shenkin *et al.*, 1991). It is given by:

$$-\sum_{j=1}^{m} p_j(i) \ln p_j(i) \quad (2)$$

where $p_j(i)$ is the frequency of residue j at positions i in the alignment and m is the number of possible amino acid types (20 for naturally occurring proteins). For studies involving sets of variant proteins derived from phage-display selection experiments (Riddle *et al.*, 1997; Kim *et al.*, 1998), relative entropies were used to take into account highly variable mutagenesis probabilities:

Table 1. Alignment, kinetic characterization and correlation statistics

Protein	Alignment size (number of sequences)	Mean pair-wise sequence identity (%)	Median entropy	Fraction of residues characterized (%)	Correlation statistics		Excluding active site	
					r	p	r	p
SH3	266	27	2.14	91	0.18	0.30	0.15	0.42
CI-2	63	40	1.44	67	−0.12	0.44	−0.12	0.48
FKBP12	264	43	1.50	33	0.32	0.15	0.27	0.25
ACBP	37	44	1.48	28	0.02	0.91	−0.15	0.56
ADAh2	21	36	1.36	25	0.09	0.72	0.16	0.55
CheY	425	28	2.08	26	−0.51	0.005	−0.41	0.04

$$-\sum_{j=1}^{m} p_j(i) \ln p_j(i)/p_j^{bg} \quad (3)$$

where p_j^{bg} is the frequency of occurrence of residue j given the background residue composition. To adequately estimate $p_j(i)$ it is necessary to construct a sufficiently large alignment, here defined as more than 20 relatively distantly related sequences (see Table 1). Of the proteins previously subjected to reasonably complete mutagenesis/kinetics studies (>20 % of positions characterized), homologous sequences sufficient to accurately estimate sequence conservation (>20 sequences) have been reported only for CI-2, the SH3 domains, ACBP, FKBP12, ADAh2 and CheY. We will thus focus on these examples which, while providing a perhaps limited view of multi-state folding, cover rather broadly the apparently two-state folding of simple, single domain proteins.

We have characterized the Φ-values of 52 of the 57 structured residues of the two-state srcSH3 domain (Grantcharova *et al.*, 1998; Riddle *et al.*, 1999), representing the most exhaustive Φ-value analysis reported to date. This SH3 domain lacks any statistically significant correlation between Φ-values and sequence conservation ($r = 0.18$, $p = 0.30$; Figure 1(a)). The highest characterized Φ-value (1.16) is associated with the eighth least conserved of the kinetically characterized residues in the domain (Gly40). There is no evidence that the putative conservation of kinetically significant residues is masked by the significant conservation observed for residues conserved for functional rather than kinetic reasons (Table 1); high-Φ residues ($\Phi \geqslant 0.5$) are, on average, no better conserved than low-Φ residues ($\Phi < 0.5$) or characterized, non-active site (defined in Lim & Richards, 1994) residues (Figure 2(a)).

Fersht and co-workers (Itzhaki *et al.*, 1995) have measured the kinetic impact of mutations of 43 of the 64 structured residues in the two-state protein CI-2. We observe no statistically significant evidence that the high Φ-value residues in the set of characterized residues are particularly well conserved (Figure 1(b)). The correlation between Φ-value and sequence entropy is statistically insignificant (correlation coefficient, $r = -0.12$, $p = 0.44$; Figure 1(b)), and while the highest Φ-value residue (Ala16, $\Phi = 1.06$) is the third most conserved residue, the second highest Φ-value residue (Lys18, $\Phi = 0.70$) is the eighth least conserved of the kinetically characterized residues in the CI-2 alignment. There is no evidence that the putative conservation of kinetically significant residues is masked by the significant conservation of residues playing functional rather than kinetic roles (Table 1). The average conservation of high Φ-value residues ($\Phi \geqslant 0.5$) in CI-2 is not significantly different from the average conservation of all kinetically characterized residues, of low-Φ ($\Phi < 0.5$) residues or of characterized, non-active site residues (Figure 2(b)) (defined by MacPhalen & James, 1988).

More recently, several groups have reported somewhat less complete Φ-value analyses for the two-state proteins ADAh2, FKBP12 and ACBP (Villegas *et al.*, 1998; Fulton *et al.*, 1999; Kragelund *et al.*, 1999). Across the characterized residues of these proteins there exist no significant correlations between Φ-values and sequence conservation (Figure 1(c)-(e); Table 1) and many of the higher Φ-value positions in these proteins are some of their most poorly conserved. Moreover, within each family the average conservation of high Φ-value residues ($\Phi \geqslant 0.5$) does not differ significantly from the average conservation of all kinetically characterized residues, of low Φ-value ($\Phi < 0.5$) residues or of characterized, non-active site residues (Figure 2(c)(e)) (defined by DeCenzo *et al.*, 1996; Garcia-Saez *et al.*, 1997; Kragelund *et al.*, 1999).

Lopez-Hernandez & Serrano (1996) reported the effects of mutations at 34 of the 129 residues of the non-two-state protein CheY. Across this data set there is a statistically significant correlation between Φ-values and sequence entropies ($r = -0.51$; $p = 0.005$; Figure 1(f)), although the two highest Φ-values observed to data ($\Phi = 0.89$, 0.75 for Val33 and Ala36, respectively) occur at the 11th and fourth least conserved of the kinetically characterized positions in the protein. The average sequence entropy of characterized, high-Φ residues ($\Phi \geqslant 0.5$) is slightly lower than that of characterized, low-Φ residues ($\Phi < 0.5$) (Figure 2(f)). The exclusion of active-site residues (defined by Bellsolell *et al.*, 1994) substantially reduces the statical significance of the correlation ($r = -0.41$; $p = 0.04$; Figure 1(f), dotted line). It should be

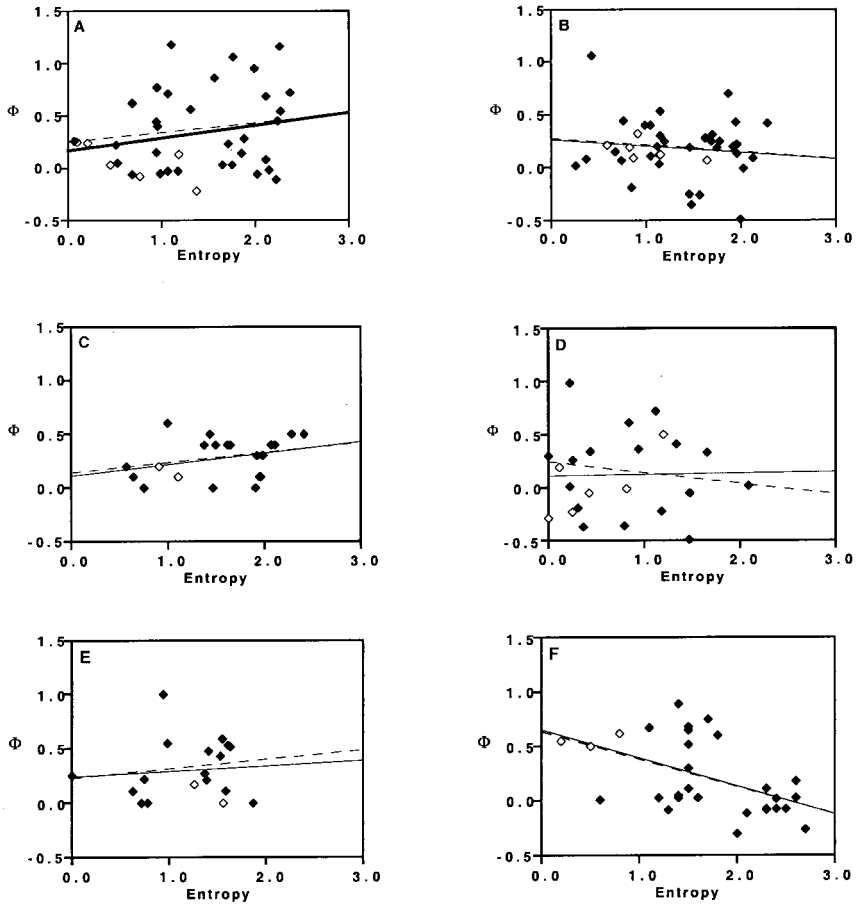

Figure 1. There is no statistically significant relationship between a residue's involvement in transition state structure (Φ) and its conservation (sequence entropy) across the (a) SH3, (b) CI-2, (c) FKBP12, (d) ACBP or (e) ADAh2 protein families ($|r| < 0.35$, $p > 0.15$). If, on the grounds that conservation for kinetic reasons might be obscured by functionally relevant conservation, active-site residues (open symbols) are removed from these data sets the correlations remain statistically insignificant ($|r| < 0.3$, $p > 0.25$; broken lines). (f) There is, however, a statistically significant correlation between Φ and conservation among the 25 % of positions in CheY that have been characterized to date ($r = -0.51$; $p = 0.005$), although the significance of this correlation is largely reduced when active-site residues (open symbols) are omitted from the analysis (broken line; $r = -0.41$; $p = 0.04$).

noted that only approximately one-quarter of the residues of CheY have been characterized kinetically and thus the extent, and conservation, of the CheY folding nucleus remains relatively poorly characterized.

We have generated two large sets of highly modified, correctly folded proteins using an *in vitro* phage display selection method in order to characterize the sequence dependence of folding kinetics (Riddle *et al.*, 1997; Kim *et al.*, 1998). In protein L, the region most highly conserved in phage selections is in the first β-hairpin (Kim *et al.*, 1998). This is also the region of the protein that exhibits the highest Φ-values, with all characterized $\Phi > 0.5$ residues appearing there (Gu *et al.*, 1998). Despite this, however, there is no statistically significant correlation between the Φ-values and relative entropy on a position by position basis (Figure 3(a)). Similarly, in a parallel experiment in which significant portions of the srcSH3 domain were "simplified" (sequences biased to Lys, Glu, Ile, Gly or Ala; Riddle *et al.*, 1997), the average sequence diversity of high-Φ residues appears indistinguishable from that of low-Φ residues (data not shown). Here, however, evidence suggests there may be a stronger correlation than implied by raw sequence entropy; several positions were under such strong selective pressure that amino acid residues ostensibly not allowed by the mutagenesis protocol were recovered in high yield (making it impossible to calculate meaningful relative entropies). Consistent with this observation, the average Φ-value of these residues and other highly conserved residues (relative entropy >2) is slightly higher than the average Φ-values of more poorly conserved residues (Figure 3(b)). As in the case of protein L, a single β-turn contains most of the high Φ-values and showed evidence of strong selective pressure. These results suggest that the conservation of high-Φ residues may be more significant during the large-scale sequence perturbations of the phage-display approach than the

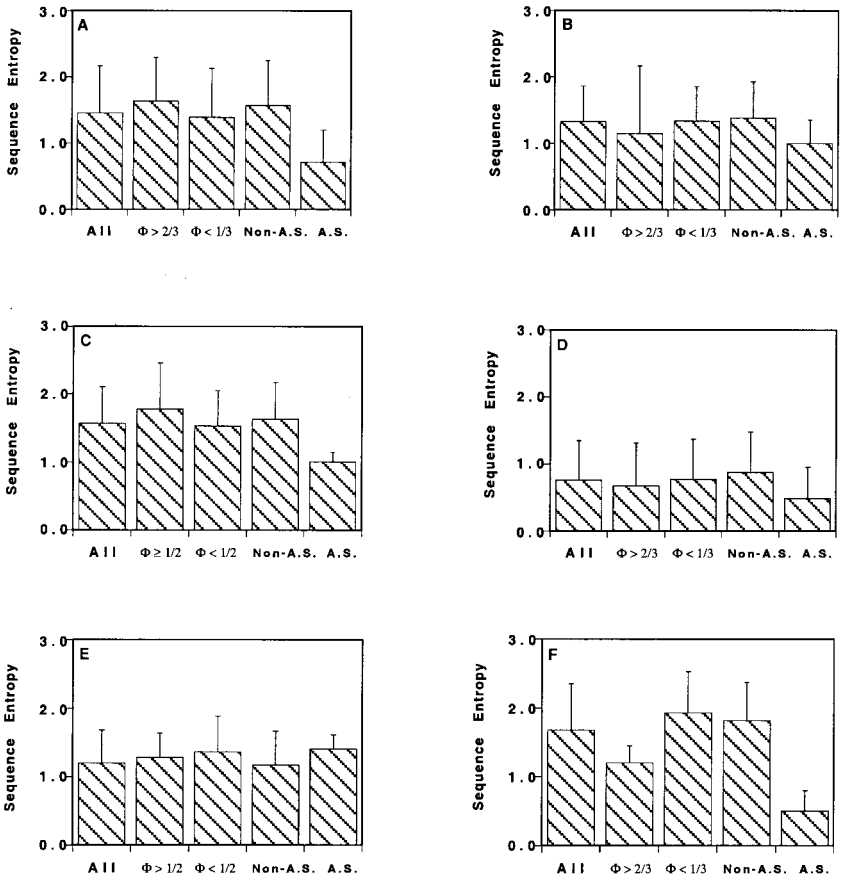

Figure 2. There are no statistically significant differences between the average entropies of high-Φ residues, all kinetically characterized residue (all), low-Φ residues or kinetically characterized, non-active-site (non-AS) residues for the proteins (a) srcSH3, (b) CI-2, (c) FKBP12, (d) ACBP or (e) ADAh2. In contrast, active-site residues (AS) are relatively well conserved across most of these proteins. (f) For the protein CheY, which exhibits a significant correlation between Φ and sequence entropy (Figure 1(f)), high-Φ residues exhibit significantly lower sequence entropies than low-Φ and non-active site residues. The error bars indicate the standard deviation of the data within each bin.

more subtle sequence changes relevant to evolutionary drift. Disruptive substitutions in the critical β-turns in protein L and SH3 may reduce the robustness of folding to the point that further large-scale sequence changes are strongly selected against.

In contrast to the relatively limited conservation of sequence identity at high-Φ positions, the available experimental evidence suggests that the Φ-value of a given position may be conserved across homologous proteins (Martinez *et al.*, 1998; Grantcharova *et al.*, 1998; Martinez & Serrano, 1999). Three homologous SH3 domains have been at least partially characterized using Φ-value analysis (Viguera *et al.*, 1996a; Grantcharova *et al.*, 1998; Martinez *et al.*, 1998; Riddle *et al.*, 1999; Martinez & Serrano, 1999; A.R.D. & J. Northey, unpublished results). The positions for which Φ-values have been determined in both the src and spectrin SH3 domains (36 % identity) are highly correlated ($r = 0.74$; $p = 0.006$; Figure 4(a)). Recently we have characterized the Φ-values of several residues in the fynSH3 domain (A.R.D. & J.N., unpublished results). The Φ-values of these residues are also highly correlated with the corresponding residues in srcSH3 (78 % sequence identity; $r = 0.82$; $p = 0.002$; Figure 4(b)), suggesting that the fynSH3 folding nucleus is also structurally related.

Discussion

We have not observed any statistically significant correlation between the role of a residue in folding transition state structures and its evolutionary conservation among naturally occurring homologs of the better-characterized, two-state folding proteins we have investigated. There is no statistically significant correlation between Φ-value and sequence entropy. The sequence entropy of high-Φ residues is not significantly less than the average entropy of all residues, of all kinetically characterized residues or of all non-active-site residues and residues with the highest characterized Φ-values are often some of the least conserved residues in these proteins. Consistent with these observations, we also observe no strong correlation between the Φ-value and sequence entropy across sets of highly modified variants of protein L and the srcSH3 domain, although specific, highly conserved regions in both proteins play important roles in the folding process. A correlation is observed between Φ and sequence entropy for the rather poorer-characterized, multi-state protein CheY, but the exclusion of residues conserved for functional reasons from the analysis reduces substantially its statistical significance.

Lattice folding simulations indicate that the sequence identity and structure of the folding

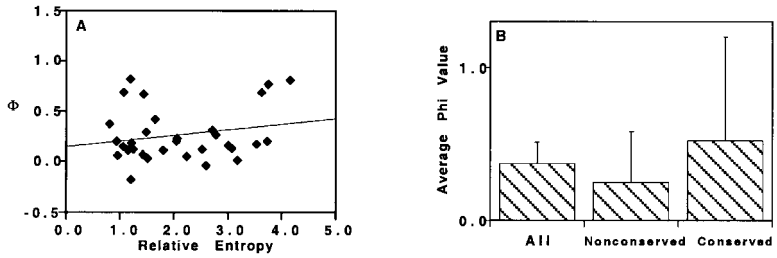

Figure 3. (a) While the first β-turn of protein L is under significant sequence constraints in random sequence selections and exhibit relatively high Φ-values (Gu *et al.*, 1998; Kim *et al.*, 1998), there is no statistically significant evidence ($r = 0.21$; $p = 0.25$) suggesting a correlation between sequence conservation and Φ across highly modified variants of protein L generated using *in vitro* phage-display selection techniques (Kim *et al.*, 1998; Riddle *et al.*, 1997). Note that due to the rather non-random sequence compositions allowed by the mutagenesis strategies used, this Figure presents relative entropy values (equation (3)); thus in contrast to Figure 1, more highly conserved residues appear to the right. (b) Some evidence of conservation of high-Φ residues is observed across a set of sequences that fold to the SH3 domain fold and that feature reduced alphabet complexity (sequence biased to Ile, Lys, Glu, Ala or Gly). The more conserved residues of these variants (relative entropies $\geqslant 2$ and positions for which non-allowed residues types were recovered) exhibit somewhat higher Φ-values than less conserved residues, although the statistical significance of this observation remains relatively weak.

nucleus of lattice polymer models are conserved when sequences are selected for stability (Shakhnovich *et al.*, 1996; Michnick & Shakhnovich, 1998) or rapid folding (Gutin *et al.*, 1998; Mirny *et al.*, 1998). Based on this observation, it has been predicted that residues involved in the folding nuclei of proteins may be conserved across naturally occurring homologous families, or exhibit intra-family conservation across superfamilies, including those of CI-2 (Shaknovich *et al.*, 1996), cytochrome *c* (Ptitsyn, 1998), ubiquitin (Michnick & Shakhnovich, 1998), CheY (Mirny *et al.*, 1998), ADAh2 (Mirny & Shakhnovich, 1999) and CD2 (Mirny & Shakhnovich, 1999). Only extremely limited mutagenesis/kinetics studies (<10% of positions characterized) have been reported for cytochrome *c* (Colo'n *et al.*, 1996; McGee & Nall, 1998), ubiquitin (Khorasanizadeh *et al.*, 1996) or CD2 (Lorch *et al.*, 1999), thus the conservation of kinetically significant residues in these proteins is difficult to address at this time. As indicated above, however, we observe no statistically significant evidence in support of the suggestion that residues forming the folding nucleus of CI-2 are highly conserved (Shaknovich *et al.*, 1996).

Using a more sophisticated measure of conservation defined by intra-family conservation of residue similarity across protein superfamilies (''conservatism-of-conservatism''), Shakhnovich and co-workers have investigated the conservation of high-Φ residues in CheY and ADAh2 (Mirny *et al.*, 1998; Mirny & Shakhnovich, 1999). While some high-Φ positions are clearly conserved by this measure, confounding conservation (due to functional constraints) and incomplete kinetic characterization preclude rigorous statistical confirmation of such a relationship. As measured by conservativism-of-conservativism, six of the ten characterized CheY residues exhibiting $\Phi \geqslant 0.5$ are among the 17 most conserved residues in the protein (probability of chance occurrence 0.3%; I.R., unpublished results). These six residues, however, include several active-site positions and it has been shown that conservation across superfamily members is often high for such functionally critical residues (Lichtarge *et al.*, 1996; Mirny & Shakhnovich, 1999). Two of the five ADAh2 residues known to exhibit $\Phi \geqslant 0.5$ are among the nine residues exhibiting the highest conservativism-of-conservativism in the alpha/beta plait superfamily (Mirny & Shakhnovich, 1999). However, while this suggests that some residues may be conserved for kinetic reasons, the limited kinetic data available for this family (Garcia-Saez *et al.*, 1997) prevent any quantitative, statistical verification of the relationship (probability of chance occurrence >10%; I.R., unpublished results). Clearly more complete kinetic characterization will be required in order to establish unequivocally the relationship between sequence conservation across superfamilies and participation in the folding nucleus.

Sequence as a determinant of folding kinetics

Sequence is a primary determinant of protein folding kinetics, as sequence alone encodes the three-dimensional structure of a protein and the rate with which that structure is formed (Anfinsen, 1973). But the observation that the identities of high-Φ residues are typically poorly conserved is consistent with numerous lines of evidence suggesting that protein folding kinetics are relatively insensitive to fine details of sequence and are, instead, defined by sequence through its effects on more global parameters such as topology (Nymeyer *et al.*, 1998; Plaxco *et al.*, 1998a; Alm & Baker, 1999) and, to a leser extent, stability (Viguera *et al.*, 1996b; Mines *et al.*, 1996; Plaxco *et al.*, 1997, 1998b).

Conservation of transition state structure

While we observe little statistically significant evidence of the predicted sequence conservation of

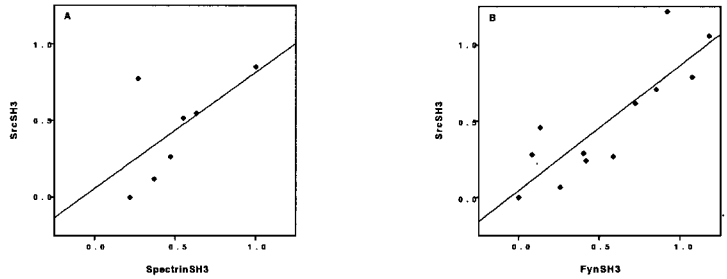

Figure 4. While the identities of high-Φ residues are not typically conserved, the Φ-values of equivalent sequence positions are highly correlated between even distantly related homologous proteins. (a) Plot of Φ-values for corresponding positions in the spectrin and srcSH3 domains (36 % sequence identity, $r = 0.74$, $p = 0.006$), (b) the src and fynSH3 domains (78 % sequence identity, $r = 0.82$, $p = 0.002$). Like those of the native proteins, the structures of the folding transition states of these homologs appear to be relatively well conserved.

folding nuclei, the limited evidence available is consistent with the predicted structural conservation of folding nuclei. With the notable exception of mutations of λ-repressor (Burton *et al.*, 1997), point mutations rarely affect the relative solvent-accessible surface area of the transition state (e.g. Itzhaki *et al.*, 1995; Grantcharova *et al.*, 1998; Kragelund *et al.*, 1999; Fulton *et al.*, 1999), suggesting that the majority of point mutations do not strongly perturb transition state structure. Consistent with this, Serrano and co-workers used Φ-value analysis to characterize the transition state structure of a mutant spectrin SH3 domain with a folding transition state some 1.8 kcal/mol more stable than that of the wild-type protein. Despite this large increase in stability, the conformation of the transition state ensemble is not significantly perturbed by the mutation (Martinez *et al.*, 1998).

Studies of homologous proteins provide further evidence that the structures of the folding transition states are relatively insensitive to fine details of sequence and are instead defined by more global parameters. As reported here (Figure 4), there is statistically significant evidence that the Φ-values of a given sequence position are maintained across homologous proteins (although the relatively high levels of sequence identity, 36 and 78 %, contrast strongly with the 27 % average pair-wise sequence identity of the SH3 data presented in Figures 1(a) and 2(a)). Analogous to this, studies of a number of sets of even distantly related, homologous proteins indicate that the relative solvent-accessible surface areas of their transition states are fairly well conserved (e.g. Kragelund *et al.*, 1996; Mines *et al.*, 1996; Plaxco *et al.*, 1997, 1998b; Reid *et al.*, 1998). More recently, Chiti *et al.* (1999) have reported a statistically significant correlation between the Φ-values of equivalent positions in the proteins AcP and ADAh2 which, while structurally similar, lack obvious sequence homology. Clearly, such correlations could arise only if neither point mutations nor the large scale sequence changes apparent across these sets of topologically identical proteins significantly perturb the structures of their folding transition states.

Conclusions

The native conformations of homologous proteins tend to be highly conserved. The conservation of the relative solvent-accessible surface area of folding transition states and the conservation of Φ-values between equivalent sequence positions in homologous proteins suggest that folding transition state conformations may be similarly conserved. The lack of significant conservation of high-Φ residues indicates that this putative structural conservation does not imply or require a high degree of sequence similarity.

Materials and Methods

The Φ-values were taken as reported in the literature. In order to reduce biases arising from poorly constrained Φ-values, values outside of the range −0.5 to 1.5 were rejected (as described above). When the kinetic impacts of several mutations have been reported for a single position, the most conservative mutation was chosen (e.g. Ile to Val, Ala to Gly, Gly to Ala, etc.) with a Φ-value within the acceptable limits.

A semi-automated exhaustive algorithm was used to assemble the sequence alignments. An initial target sequence was used to initiate a PSI-BLAST search (default values) of the non-redundant database compilation nrdb (Altschul *et al.*, 1997). The set of homologous sequences was retrieved and the stretch of residues aligned to the initial target domain was extracted from each protein sequence. Any redundant sequences (defined here as sequence identities >90 %) in this set of homologues were removed and the remaining sequences were aligned using ClustalW(1.7) (Thompson *et al.*, 1994) with manual refinement to minimize gaps in regions of secondary structure. An all-against-all comparison was used to calculate the average identity of each sequence with the rest of the alignment. Using this similarity measure as a guide, new target sequences were selected and used to initiate a new round of sequence gathering and alignment. This iterative process was continued until homology searching produced no new sequences. The final sequence alignment was weighted by the Henikoff algorithm to reduce bias due to over-representation of particular subfamilies (Henikoff & Henikoff, 1994). These sequence weights were used to calculate

weighted residue frequencies at each position in the alignment. Sequence alignments are available upon request.

Sequence entropies were calculated from final sequence alignments as described above (equation (2)). Some groups have suggested that these measures of sequence entropy underestimate conservation by ignoring chemically conservative changes. Using a ''reduced complexity amino-acid alphabet'' Mirny *et al.*, 1998) that scores chemically similar residues as equivalent, however, produces results equivalent to those reported here (data not shown). Relative sequence entropies were calculated from variant protein sequences using equation (2). Variant alignments were taken as reported in the relevant kinetics literature (Riddle *et al.*, 1997; Kim *et al.*, 1998).

It should be noted that less than one-third of the residues of FKBP12, ACBP, ADAh2 and CheY have been subjected to kinetic characterization, and thus the exact number and identities of high-Φ residues in these proteins remains relatively poorly characterized. Moreover, the one-quarter of the residues of ACBP that have been characterized kinetically sample primarily the more conserved regions of the protein (median entropies all residues, 1.48; characterized residues, 0.79) and thus the relative conservation of this protein's folding nucleus remains particularly difficult to establish quantitatively.

Single variable linear relationships were assumed. The reported correlation coefficients and *p*-values were calculated using S-plus (MathSoft, Inc). Reported *p*-values are the probability that an observation from a *t*-distribution with the appropriate number of degrees of freedom would exceed the ratio of the estimate of the gradient to its estimated standard deviation. The *p*-values thus represent the probability that, if the null hypothesis were true (i.e. that there exists no relationship and the true slope is zero), an estimate of the slope would be generated as far or farther from zero than that actually observed.

Acknowledgments

The authors are indebted to Fabrizio Chiti, Viara Grantcharova, Sophie Jackson, David Kim, Julian Northey, Jed Santiago and Luis Serrano for communicating important, pre-publication results. This work was supported by the Medical Research Council of Canada (A.R.D.) and by an NIH grant and young investigator awards from the NSF and Packard Foundation (D.B.).

References

Alm, E. & Baker, D. (1999). Matching theory and experiment. *Curr. Opin. Struct. Biol.* **9**, 189-196.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223-230.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Bellsolell, L., Prieto, J., Serrano, L. & Coll, M. (1994). Magnesium binding to the bacterial chemotaxis protein CheY results in large conformational changes involving its functional surface. *J. Mol. Biol.* **238**, 489-495.

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.

Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L. & Oas, T. G. (1997). The energy landscape of a fast-folding protein mapped by Ala → Gly substitutions. *Nature Struct. Biol.* **4**, 305-310.

Chili, 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.

Col'on, W. W., Elove, G. A., Wakem, P., Sherman, F. & Roder, H. (1996). Side-chain packing of the N and C-terminal helices plays a critical role in the kinetics of cytochrome *c* folding. *Biochemistry*, **35**, 5538-5549.

DeCenzo, M. T., Park, S. T., Jarrett, B. P., Aldape, R. A., Futer, O., Murcko, M. A. & Livingston, D. J. (1996). FK506-binding protein mutational analysis: defining the active-site residue contributions to catalysis and the stability of ligand complexes. *Protein Eng.* **9**, 173-180.

Demirel, M. C., Atilgan, A. R., Jernigan, R. L., Erman, B. & Bahar, I. (1998). Identification of kinetically hot residues in proteins. *Protein Sci.* **7**, 2522-2532.

Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.

Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771-782.

Fulton, K. F., Main, E. R. G., Daggett, V. & Jackson, S. E. (2000). Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.* **291**, 445-461.

Garcia-Saez, I., Reverter, D., Vendrell, J., Aviles, F. X. & Coll, M. (1997). The three-dimensional structure of human procarboxypeptidase A2. Deciphering the basis of the inhibition, activation and intrinsic activity of the zymogen. *EMBO J.* **16**, 6906-6913.

Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* **5**, 714-720.

Gu, H., Kim, D. & Baker, D. (1997). Contrasting roles for symmetrically disposed beta-turns in the folding of a small protein. *J. Mol. Biol.* **274**, 588-596.

Gutin, A. M., Abkevich, V. I. & Shahlanovich, E. I. (1998). A protein engineering analysis of the transition state for protein folding: simulation in the lattice model. *Fold. Des.* **3**, 183-194.

Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574-578.

Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.

Kim, D. E., Gu, H. D. & 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.

Khorasanizadeh, S., Peters, I. D. & Roder, H. (1996). Evidence for a three-state model of protein folding

from kinetic analysis of ubiquitin variants with altered core residues. *Nature Struct. Biol.* **3**, 193-205.

Kragelund, B. B., Hojrup, P., Jensen, M. S., Schjerling, C. K., Juul, E., Knudsen, J. & Poulsen, F. M. (1996). Fast and one-step folding of closely and distantly related homologous proteins of a four-helix bundle family. *J. Mol. Biol.* **256**, 187-200.

Kragelund, B. B., Osmark, P., Neergaard, T. B., Schiodt, J., Kristiansen, K., Knudsen, J. & Poulsen, F. M. (1999). The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nature Struct. Biol.* **6**, 594-601.

Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys.* **65**, 44-45.

Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**, 342-358.

Lim, W. A. & Richards, F. M. (1994). Critical residues in an SH3 domain from Sem-5 suggest a mechanism for proline-rich peptide recognition. *Nature Struct. Biol.* **1**, 221-225.

Lopez-Hernandez, E. & Serrano, L. (1996). Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Fold. Des.* **1**, 43-55.

Lorch, M., Mason, J., Clarke, A. & Parker, M. (1999). Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the I-state. *Biochemistry,* **38**, 1377-1385.

MacPhalen, C. A. & James, M. N. G. (1988). Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-C-subtilisin carlsberg and CI-2-subtilisin novo. *Biochemistry,* **27**, 6582-6598.

Martinez, J. C., Pisabarro, M. T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721-729.

Martinez, J. C. & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010-1016.

Matouschek, A., Serrano, L. & Fersht, A. R. (1992). The folding of an enzyme. IV. Structure of an intermediate in the refolding of barnase analysed by a protein engineering procedure. *J. Mol. Biol.* **224**, 819-835.

McGee, W. A. & Nall, B. T. (1998). Refolding rate of stability-enhanced cytochrome *c* is independent of thermodynamic driving force. *Protein Sci.* **7**, 1071-1082.

Michnick, S. W. & Shakhnovich, E. (1998). A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Fold. Des.* **3**, 239-251.

Milla, M. E., Brown, B. M., Waldburger, C. D. & Sauer, R. T. (1995). P22 Arc repressor: transition state properties inferred from mutational effects on the rates of protein unfolding and refolding. *Biochemistry,* **34**, 13914-13919.

Mines, G. A., Pascher, T., Lee, S. C., Winkler, J. R. & Gray, H. B. (1996). Cytochrome *c* folding triggered by electron transfer. *Chem. Biol.* **3**, 491-497.

Mirny, L. A., Abkevich, V. I. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA,* **95**, 4976-4981.

Mirny, L. A. & Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177-196.

Nolting, B., Golbik, R., Neira, J. L., Soler-Gonzalez, A. S., Schreiber, G. & Fersht, A. R. (1997). The folding pathway of a protein at higa resolution from microseconds to seconds. *Proc. Natl Acad. Sci. USA,* **94**, 826-830.

Nymeyer, H., Garcia, A. E. & Onuchic, J. N. (1998). Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proc. Natl Acad. Sci. USA,* **95**, 5921-5928.

Pande, V. S., Grosberg, A. Y., Tanaka, T. & Rokhsar, D. S. (1998). Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.* **8**, 68-79.

Parker, M. J., Dempsey, C. E., Hosszu, L. L. P., Waltho, J. P. & Clarke, A. R. (1998). Topology, sequence evolution and folding dynamics of an immunoglobulin domain. *Nature Struct. Biol.* **5**, 194-198.

Plaxco, K. W., Spitzfaden, C., Campbell, I. D. & Dobson, C. M. (1997). A comparison of the folding kinetics and thermodynamics of two homologous fibronectin type III modules. *J. Mol. Biol.* **270**, 763-770.

Plaxco, K. W., Simons, K. T. & Baker, D. (1998a). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985-994.

Plaxco, K. W., Guijarro, J. I., Morton, C. J., Pitkeathly, M., Campbell, I. D. & Dobson, C. M. (1998b). The folding kinetics and thermodynamics of the fynSH3 domain. *Biochemistry,* **37**, 2529-2537.

Pouron, A. & Mornon, J. P. (1999). Predicting the protein folding nucleus from a sequence. *FEBS Letters,* **452**, 283-289.

Ptitsyn, O. B. (1998). Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *J. Mol. Biol.* **278**, 655-666.

Reid, K. L., Rodriguez, H. M., Hillier, B. J. & Gregoret, L. M. (1998). Stability and folding properties of a model beta-sheet protein, *Escherichia coli* CspA. *Protein Sci.* **7**, 470-479.

Riddle, D. S., Santiago, J. V., BrayHall, 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.

Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016-1024.

Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature,* **379**, 96-98.

Shenkin, S. P., Erman, B. & Mastrandrea, L. D. (1991). Information-theoretical entropy as a measure of sequence variability. *Proteins: Struct. Funct. Genet.* **11**, 297-313.

Shrivastava, I., Vishveshwara, S., Cieplak, M., Maritan, A. & Banavar, J. R. (1995). Lattice model for rapidly folding protein like heteropolymers. *Proc. Natl Acad. Sci. USA,* **92**, 9206-9209.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.

Viguera, A. R., Serrano, L. & Wilmanns, M. (1996a). Different folding transition states may result in the same native structure. *Nature Struct. Biol.* **3**, 874-880.

Viguera, A. R., Virtudes, V., Aviles, F. X. & Serrano, L. (1996b). Favourable native-like helical local interactions can accelerate protein folding. *Fold. Des.* **2**, 23-33.

Villegas, V., Martinez, J. C., Aviles, F. X. & Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283**, 1027-1036.

Edited by C. R. Matthews

(Received 18 August 1999; received in revised form 9 November 1999; accepted 19 November 1999)