doi:10.1006/jmbi.2001.5344 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2002) **316**, 225–233

JMB

COMMUNICATION

Residues Participating in the Protein Folding Nucleus do not Exhibit Preferential Evolutionary Conservation

Stefan M. Larson[1], Ingo Ruczinski[2], Alan R. Davidson[3]*, David Baker[4]* and Kevin W. Plaxco[5]*

[1]*Department of Chemistry and Biophysics Program, Stanford University, Stanford CA 94305, USA*

[2]*Department of Biostatistics Bloomberg School of Public Health, Johns Hopkins University, Baltimore MD 21205, USA*

[3]*Department of Biochemistry and Department of Molecular and Medical Genetics, University of Toronto, Toronto Ontario M5S 1A8, Canada*

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

[5]*Department of Chemistry and Biochemistry and Interdepartmental Program in Biomolecular Science and Engineering, University of California, Santa Barbara CA 93106, USA*

**Corresponding authors*

To what extent does natural selection act to optimize the details of protein folding kinetics? In an effort to address this question, the relationship between an amino acid's evolutionary conservation and its role in protein folding kinetics has been investigated intensively. Despite this effort, no consensus has been reached regarding the degree to which residues involved in native-like transition state structure (the folding nucleus) are conserved. Here we report the results of an exhaustive, systematic study of sequence conservation among residues known to participate in the experimentally (Φ-value) defined folding nuclei of all of the appropriately characterized proteins reported to date. We observe no significant evidence that these residues exhibit any anomalous sequence conservation. We do observe, however, a significant bias in the existing kinetic data: the mean sequence conservation of the residues that have been the subject of kinetic characterization is greater than the mean sequence conservation of all residues in 13 of 14 proteins studied. This systematic experimental bias gives rise to the previous observation that the median conservation of residues reported to participate in the folding nucleus is greater than the median conservation of all of the residues in a protein. When this bias is corrected (by comparing, for example, the conservation of residues known to participate in the folding nucleus with that of other, kinetically characterized residues) the previously reported preferential conservation is effectively eliminated. In contrast to well-established theoretical expectations, both poorly and highly conserved residues are apparently equally likely to participate in the protein-folding nucleus.

Keywords: nucleation; homology; sequence entropy; phi-value; two-state

It has been predicted on theoretical grounds that protein folding kinetics are under evolutionary control[1–7] and asserted[8,9] that ''such pressure can be manifested in noticeable additional conservation of nucleus residues''. We have previously investigated this assertion by studying the sequence conservation of residues participating in the native-like folding nucleus of all of the appropriately characterized proteins then described.[10] In contrast to theory-based

Abbreviations used: ACBP, acylcarrier binding protein; AcP, acylphosphatase; ADA2h, the activation domain of human carboxypeptidase; CI2, chymotrypsin inhibitor 2; FKBP, FK506 binding protein; FNfn10, 10th fibronectin type III domain of human fibronectin; SH3, src homology region 3 of src tyrosine kinase.

E-mail addresses of the corresponding authors: kwp@chem.ucsb.edu; dabaker@u.washington.edu; alan.davidson@utoronto.ca

0022-2836/02/020225–9 $35.00/0

expectations,[3,4,7,8] we observed little statistically significant evidence in favor of a relationship between a residue's Φ-value[11] (an objective, quantitative measure of the extent to which a residue participates in native-like interactions during the rate-limiting step in folding†) and the degree to which the residue is conserved across homologous proteins.

In seemingly sharp contrast to our results, Mirny and Shakhnovich have recently reported that ''the folding nucleus is more conserved than the rest of the protein'' for eight of nine proteins investigated.[8,9] They also suggest that two specific shortcomings in our analysis account for the apparent disagreement between our experimental results and both their results and the results of prior theoretical studies. Here we demonstrate that the proposed differences in analysis do not account for the reported disparity. We suggest instead several additional differences that account for this discrepancy and demonstrate that when they are taken into consideration there remains no significant evidence in favor of the hypothesized preferential conservation of the Φ-value defined folding nucleus.

Absolute *versus* relative entropy

Mirny & Shakhnovich partially attribute the reported discrepancy to the use of an inappropriate measure of sequence conservation termed relative sequence entropy.[8,9] Counter to this explanation, however, both studies employed precisely the same measure of sequence conservation. The measure employed in both studies, termed absolute sequence entropy,[12] is given by:

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

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 classes (e.g. the 20 proteogenic amino acids). The confusion might have arisen because, as clearly described, we employed relative sequence entropy to take into account the highly variable mutagenesis probabilities of two sets of variant proteins derived from phage-display selection experiments..[13-14] Relative sequence entropy, given by:

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

where p_j^{bg} is the frequency of occurrence of residue j given the background residue composition; this tends to underestimate the conservation of common residues and thus we did not employ it to compute the conservation among alignments of naturally occurring sequences.

Sequence similarity *versus* sequence identity

Mirny & Shakhnovich also state that the amino acids must be grouped ''into classes according to their physical-chemical properties'' in order for the putative conservation of the nucleus to be detected and partially attribute the reported discrepancy to a failure to employ the necessary clustering scheme. In previous studies we addressed the effects of employing a ''reduced sequence entropy'' based on their scheme of grouping the amino acids into six classes: (AVLIMC), (FWYH), (STNQ), (KR), (DE) and (GP). However, because we felt they provided no additional insights we did not present the relevant supporting data.[10] Here we reproduce these data for the six proteins in our original test set and for several additional proteins for which the appropriate data have recently become available.

Little or no evidence that folding nuclei are preferentially conserved

Even when employing the proscribed, cluster-based, reduced sequence entropy we observe little if any evidence in favor of a statistically significant correlation between the role of a residue in folding transition state structures and its evolutionary conservation among naturally occurring homologs (Figure 1). For 12 of the 14 proteins investigated the correlation between Φ-value and reduced sequence entropy is statistically insignificant ($p > 0.13$; Table 1), and residues with the highest reported Φ-values are often some of the least conserved in the homologous family. Statistically significant correlations between reported Φ-values and reduced sequence entropy are observed for two proteins: CheY ($r^2 = 0.352$, $p = 7 \times 10^{-4}$) and TI-I27 ($r^2 = 0.244$, $p = 0.017$). The correlation observed for TI-I27, however, is opposite the direction predicted by theory. Consistent with the generally poor correlation between Φ and reduced sequence entropy, the conservation of high Φ residues ($\Phi > 0.5$) is poorer than the mean conservation of all the kinetically characterized residues and poorer than the mean conservation of low Φ residues across seven of the 13 proteins for which high Φ residues have been defined (Figure 2).

We have also tested the hypothesized preferential conservation of the folding nucleus using a statistical approach analogous to that employed by Mirny & Shakhnovich.[8] To do so we compare the median conservation of the n known high Φ ($\Phi > 0.5$) residues in a given protein with the distribution of medians obtained from 10^5 sets of n residues randomly chosen from among all of the characterized residues of the protein. The fraction

† 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: $\Phi = \Delta\Delta G_{\ddagger\text{-U}}/\Delta\Delta G_{\text{F-U}}$.

of this distribution representing greater than median conservation than that of the known high Φ residues, P_0, is the probability that conservation greater than that of the nucleus would be obtained by chance given the distribution of sequence entropies (Table 1). The results of this analysis are telling: the median conservation of high Φ residues is indistinguishable ($P_0 \geqslant 0.05$) from that expected for a set of residues randomly selected from among all characterized amino acids for all 13 of the proteins for which high Φ residues have been identified. Against a background of functional and structural pressures that constrain the identities of both low and high Φ residues, any selective pressures arising due to kinetic constraints are apparently too weak to produce measurable additional conservation in the Φ-value defined folding nucleus.

The origins of the discrepancy

We observe no significant evidence in favor of the hypothesis that selective pressures aimed at controlling folding kinetics are ''manifested in noticeable additional conservation of nucleus residues''.[8] This observation appears to contrast sharply with the claim that ''residues in the folding nucleus are considerably more conserved than the rest of the protein''.[9] We note, however, that the manner in which the relevant data were collected, measured and analyzed differs significantly between the two studies and provides several reasons why this apparent discrepancy may have arisen.

The data set

The studies in question focus on somewhat different sets of test proteins. In order to produce as unbiased and representative a test set[10] as possible, we exhaustively included every protein for which at least moderately complete and accurate Φ-value analysis had been reported (>20 % of positions characterized, no significant evidence of transition state movement upon mutation) and for which sufficient distantly homologous sequences were available for alignment (>20 sequences with less than 90 % pairwise identity). Mirny & Shakhnovich, in comparison, do not describe the criteria by which they selected the nine examples included in their analysis,[8,9] and thus it is more difficult to address the degree to which their data set is representative. For example, they omit the srcSH3 domain, which at 91 % sequence coverage is the most exhaustively characterized protein reported to date.[16] In contrast, they include in their test set two very sparsely characterized proteins: CD2.d1 (precise Φ-values reported for six out of 193 positions[17] with five of six exhibiting perfect conservation by their measure[8]) and U1A (eight out of 95 residues characterized,[18] all highly conserved by their measure[8]). As discussed below, such sparsely characterized proteins are particularly prone to experimental biases that render them difficult to employ in studies of folding nucleus conservation.

The Φ-values of a number of additional proteins have been reported since our initial study, of which six meet our original criteria for inclusion and are thus presented here. These are muscle acylphosphatase[19] (AcP), villin 14T[20] (villin), the 27th Ig domain of titin[21] (TI-I27), FN3 domains excised from tenascin[22] (tenascin) and fibronectin[23] (FNFn10), and the WW domain.[24] In addition, the relatively sparsely characterized proteins CD2.d1 and U1A[17,18] are included to allow for a more complete comparison between our study and that of Mirny & Shakhnovich.

The construction of sequence alignments

In order to obtain the best possible estimates of sequence entropy it is critical to obtain the highest quality sequence alignments. Our approach was to carefully construct alignments using exhaustive BLAST searches and manual inspection to ensure proper alignment (with the exception of TI-I27, for which the deep, pre-existing PFAM immunoglobulin alignment[25] was employed following refinement using the criteria described here). We also eliminated redundant sequences[26] (no two sequences in the finished alignment share >90 % identity) and weighted[27] the alignment to get the truest possible picture of sequence entropy. Because the quality of the alignment is a critical parameter in studies of the type reported here, we have provided a detailed description of how our alignments are constructed[28] and statistics describing their completeness (Table 1; alignments available as Supplementary Material). Mirny & Shakhnovich, in contrast, employed existing HSSP alignments or, if the ''HSSP alignments contained too few sequences'',[8] PFAM alignments. Because HSSP alignments are sometimes relatively shallow and PFAM alignments relatively redundant, it is difficult to assess their suitability to the question at hand. This is especially true given that other authors have noted that the relevant HSSP alignments are not without potentially significant errors (see, e.g. Hamill *et al*.[23]).

Defining the folding nucleus

Perhaps the most straightforward means of defining kinetically critical residues is to measure the extent to which a residue's mutation alters folding rates. This measure would, presumably, more directly reflect the selective pressures arising if evolution acts in order to ensure rapid folding. Unfortunately, such a measure is mutation dependent; mutations that do not significantly alter the chemistry of the side-chain tend to affect folding kinetics less than more dramatic side-chain alterations at the same position. This renders it difficult to assign a single value to the kinetic effects of

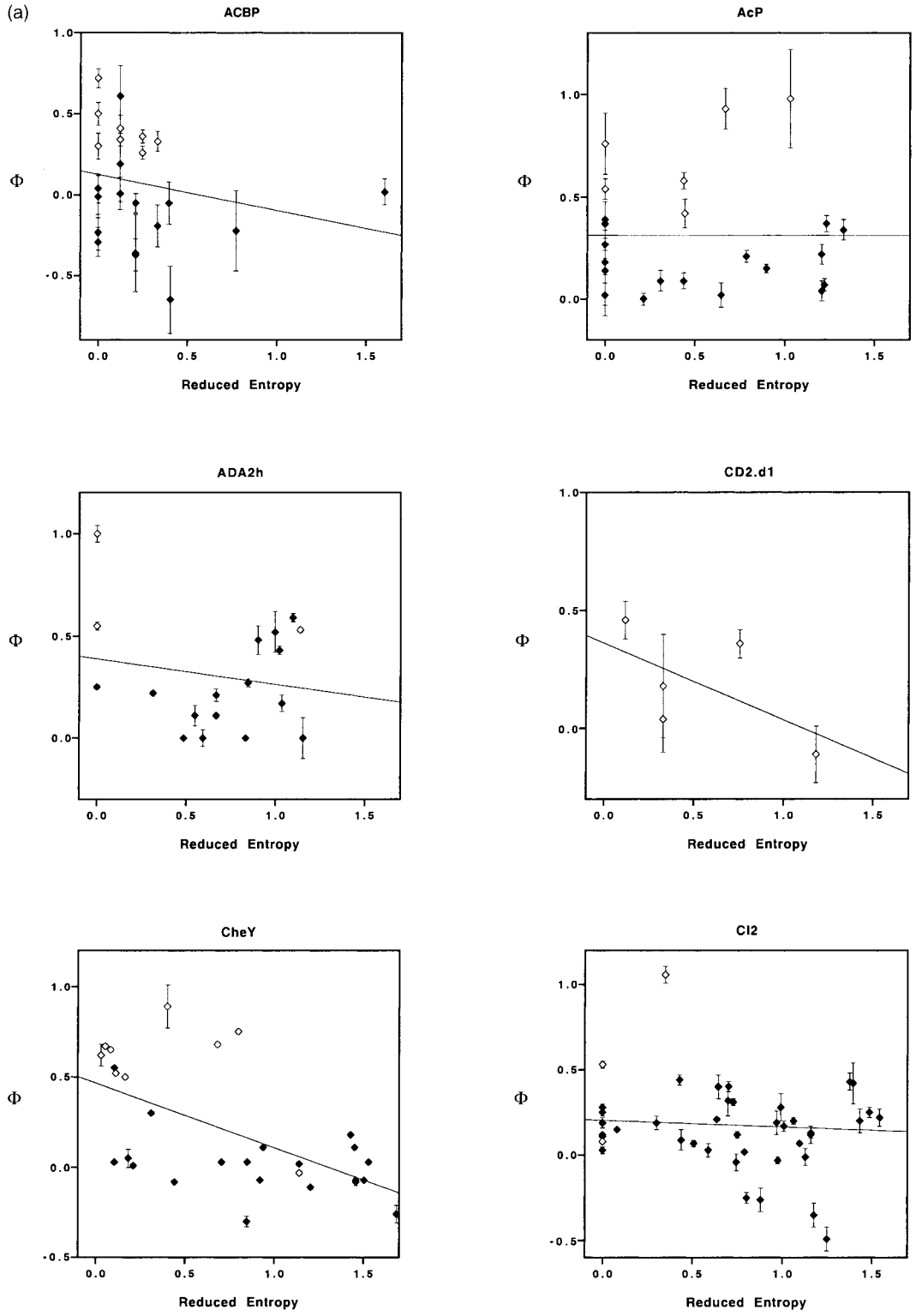

Figure 1 (*legend shown on page 230*)

mutations at a given site. In part because of this ambiguity, theoretical studies of the protein folding nucleus have predominantly focused on Φ-values,[6–9] a relatively mutation-independent[11] (J.G.B. Northey, K.L. Maxwell & A.R.D., unpublished results) measure of the extent to which the side-chain of a position participates in native-like transition state interactions. It is participation in

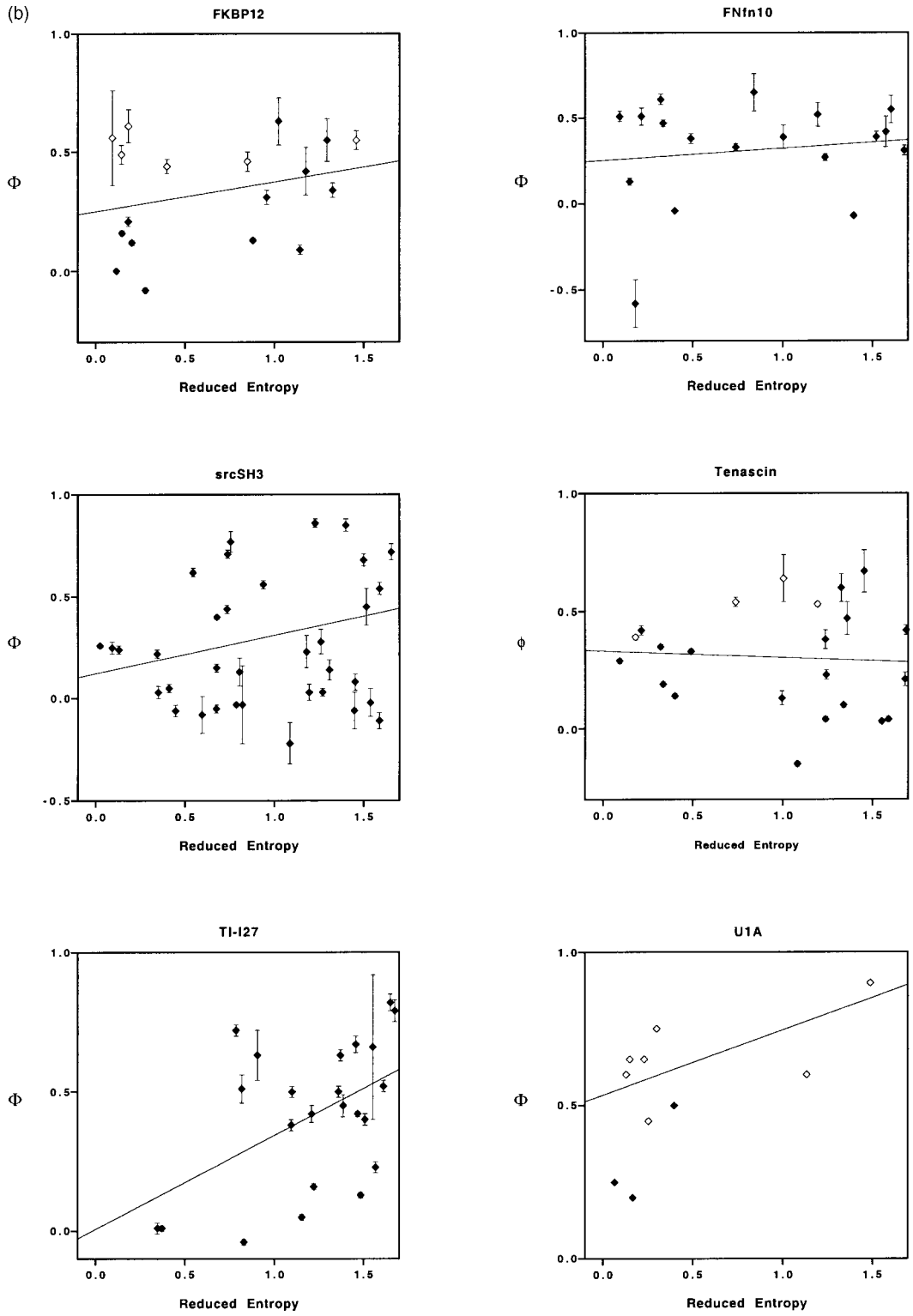

Figure 1 (*legend shown on page 230*)

these native-like interactions, rather than a residue's absolute thermodynamic contribution to the folding transition-state, that is hypothesized to give rise to kinetics-linked sequence conservation.[3,4, 7–9]

Experimentally determined Φ-values provide a readily obtainable, objective means of quantifying participation in the native-like transition-state interactions that define the kinetic folding nucleus.[11] By comparing sequence conservation

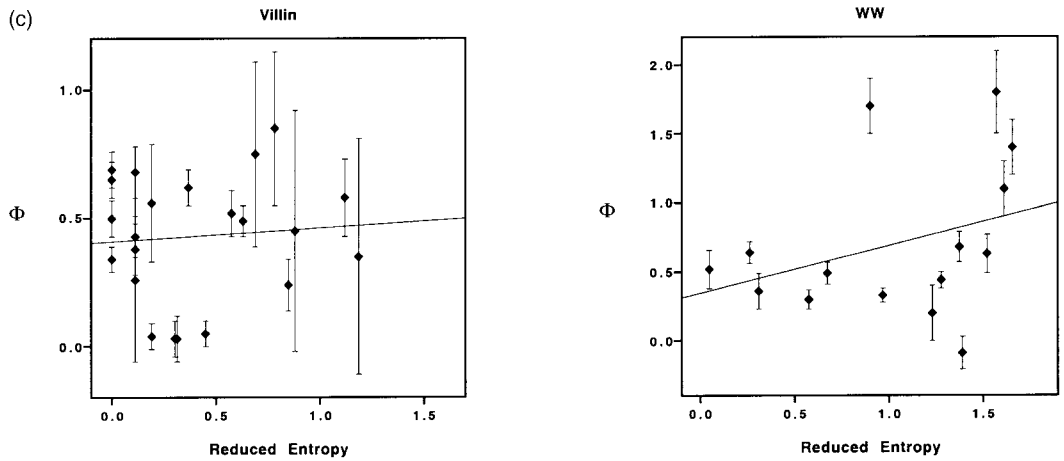

Figure 1. We observe no statistically significant correlation (all $p > 0.13$) between sequence conservation and Φ-value for 12 of the 14 proteins for which the appropriate data are available. We observe statistically significant correlations for the proteins CheY ($r^2 = 0.352$, $p = 7 \times 10^{-4}$) and TI-I27 ($r^2 = 0.244$, $p = 0.017$). The correlation observed for TI-I27, however, is opposite the direction predicted by theory. Kinetically characterized residues reported by Mirny & Shakhnovich to be in the folding nucleus are denoted with open symbols. Note that these include a number of residues with very low Φ-values and exclude some relatively high Φ positions. The proteins lacking open symbols were not included in the previous analysis.[8] Φ-Values and confidence intervals were adopted from the literature.[16–24,30–33] The data for protein U1A were those determined for $\beta_T = 0.7$;[18] data defined for other values of β_T do not alter the results significantly. Estimated confidence intervals are not available for U1A.

directly against Φ-value, we thus have employed a consistent impartial standard in our quest to understand the relationship between the folding nucleus and evolutionary conservation. Mirny & Shakhnovich, in contrast, defined ''the nucleus as it was identified by the original experimental groups''.[8] Unfortunately, this definition is subjective: few if any experimental authors state unambiguously that the folding nucleus is comprised of a specific set of residues. This metric thus leaves open to interpretation the exact composition of the nucleus and allows the inclusion of residues (such as L16, A40 and L95 in CD2.d1) that ''do not make a measurable contribution to the rate-limiting transition state'' (M. Lorch & A. Clarke, personal communication). Ambiguity aside, this standard is also arbitrary as it leads to the exclusion of some known high Φ residues (e.g. A36 in CheY at $\Phi = 0.75(\pm 0.01)$; K61 in ACBP at $0.61(\pm 0.19)$) while often including residues with very low Φ-values (e.g. I76 in CI2 at $\Phi = 0.08(\pm 0.01)$; I18 in CD2.d1 at $0.18(\pm 0.22)$; D38 in CheY at $-0.03(\pm 0.01)$). The inclusion of these low Φ residues is particularly puzzling; it seems unlikely that putative selective pressures aimed at optimizing folding kinetics would conserve the identities of residues whose side-chains do not measurably contribute to folding kinetics. Lastly, because it includes residues for which accurate experimental Φ-values have not been reported (W32 in CD2.d1; I23 in ADA2 h), this metric may be subject to a significant observer bias: it is possible that these residues are believed to be in the folding nucleus, in part, precisely because they are highly conserved. The rigorous and consistent use of experimentally determined Φ-values provides an unbiased standard and avoids most if not all of these potential pitfalls.

Of course, even the most objective standard available sometimes requires a level of interpretation. Ours is as follows. First, we only considered residues that have been subjected to kinetic characterization and for which sufficient and meaningful sequence alignments are available. Second, we consider only those Φ-values that have been determined with at least a moderate degree of precision (here defined as reported confidence intervals tighter than ± 0.5). Third, for positions for which multiple mutations have been characterized, we selected the least perturbative mutation (as defined by Plaxco *et al.*[10]) that met the above criterion. The question remains, however, given a set of experimentally valid Φ-values how does one define the residues that participate in the nucleus? As such participation is rarely if ever a binary event (i.e. Φ-values are never precisely zero or one) our preferred approach is to test for correlations between Φ-value and sequence conservation. This approach is based on the assumption that, if participation in the folding nucleus provides the predicted selective pressure to maintain sequence similarity, then residues participating more strongly in the nucleus will be relatively better conserved. As noted, we observe effectively no such correlation (Figure 1). We have also explored conservation of the nucleus by defining participation in it as coinciding with $\Phi > 0.5$. This definition of the folding nucleus (and those using lower and higher Φ-value cutoffs; data not shown) also fails

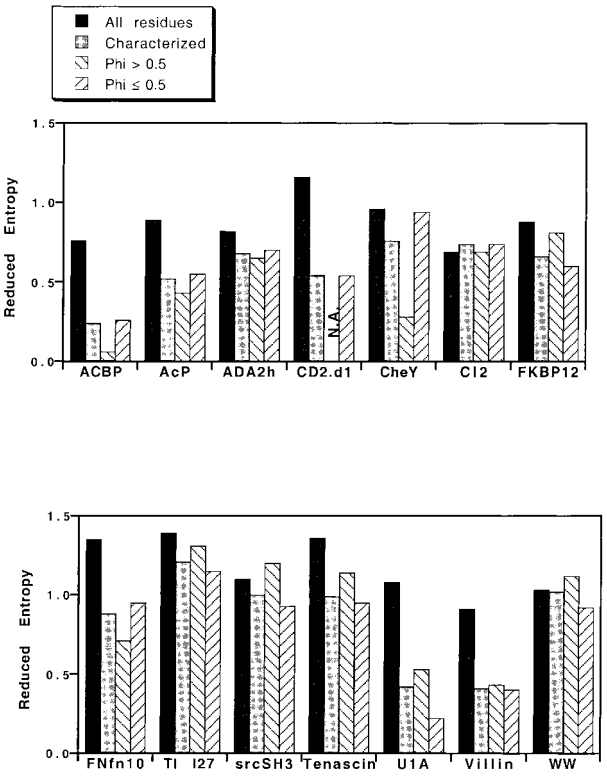

Figure 2. The mean reduced sequence entropy of all residues, all characterized residues, high Φ residues and low Φ residues. A notable, systematic bias is observed: for 13 of the 14 proteins the mean conservation of the kinetically characterized residues is greater than the mean conservation of all residues in the protein. When this bias is accounted for, little or no evidence is observed in favor of preferential conservation of the folding nucleus (as defined by $\Phi > 0.5$): for six proteins these residues are more well conserved than low Φ residues, for seven they are less well conserved. (Note, of the few residues that have been characterized in CD2.d1 exhibit a Φ-value of >0.5.)[17]

to produce any significant indication of statistically significant ''excess'' conservation (Figure 2).

Sample bias in Φ-value analysis

A very substantial bias also arises due to the approach used to define the ''statistically significant'' conservation of the folding nucleus. Mirny & Shakhnovich compared the median conservation of residues reported to participate in the nucleus with the median conservation of sets of residues randomly selected from the entire protein. This approach would prove valid if Φ-values were available for either every position in the protein or for a randomly chosen or otherwise representative subset. Unfortunately, however, Φ-value analysis is neither complete nor random: when faced with the unappealing prospect of producing and characterizing the 50 to 100 or more mutations required for an exhaustive Φ-value analysis, experimentalists understandably select a subset of 20 to 40 ''interesting'' positions for characterization. Because ''interesting'' is often correlated with conserved, the mean conservation of the kinetically characterized residues is higher than the mean conservation of all residues for 13 of 14 proteins studied (Figure 2). As a consequence of this bias the conservation of residues reported to participate in the nucleus is likely to be greater than the median conservation of the entire protein, irrespective of whether or not selective pressures act to maintain the identity of the native-like nucleus.

Analysis of the structural context of characterized residues suggests that this experimental bias arises due to the disproportionate number of hydrophobic core residues for which precise Φ-values have been determined. Because of this bias, residues reported to participate in the folding nucleus are almost invariably within the hydrophobic core (fully 80 % of the residues identified[8] by Mirny & Shakhnovich as reported to participate in the nucleus are both hydrophobic and less than 10 % solvent exposed; unpublished calculations). The preferential conservation of core residues for thermodynamic reasons[15] thus leads to the previously noted[9] ''excess'' conservation of most of the kinetically characterized residues. The disproportionate number of core residues for which accurate Φ-values have been reported apparently stems from two sources. First, experimentalists tend to select core mutations preferentially over mutations at solvent-exposed positions. For example, the mean relative solvent accessibility of the positions in tenascin for which Φ-values analysis has been attempted (15.1 %) is significantly lower than the 31 % mean of the entire protein.[23] Second, mutations at solvent-exposed sites typically produce relatively small changes in folding free energy. This often produces unacceptably high errors in Φ-value and leads to the preferential exclusion of these sites from Φ-value tabulations (e.g. none of the three most exposed positions characterized in tenascin produced a reportable result, while only one of 28 less exposed positions proved similarly recalcitrant[23]). Indeed, residues contributing significantly to protein stability are usually well conserved,[29] and thus it is generally much more likely that the Φ-value of a conserved residue will be measurable than that of a poorly conserved residue.

This significant experimental bias can, fortunately, be corrected by comparing the mean conservation of residues known to participate in native-like transition state structure with the mean conservation of the other kinetically characterized residues. The validity of this approach stems from the observation that, even against a background of functional and structural selections that conserve both low and high Φ residues, additional selections aimed at maintaining the folding nucleus would increase the average conservation of high Φ residues. Relative to this standard, residues that participate in the folding nucleus are not noticeably well conserved. As described above, for example,

Table 1. Alignment, kinetic characterization and correlation statistics

Protein	Statistical significance of relationship between Φ and reduced sequence entropy[a,b]			Fraction of residues characterized[c] (%)	Number of homologs in alignment	Mean pair-wise sequence identity[d] (%)	Mean reduced sequence entropy[a] (all positions)
	r^2	p	P_0[d]				
ACBP	0.050	0.31	0.09	28	37	43.6	0.76
AcP	10^{-5}	0.99	0.52	26	43	39.6	0.89
ADA2 h	0.044	0.43	0.67	25	21	37.1	0.82
CD2.d1	0.359	0.29	N.A.	5	39	26.7	1.16
CheY	0.352	10^{-3}	0.05	26	425	28.0	0.96
CI2	0.002	0.77	0.26	67	63	40.3	0.82
FKBP	0.079	0.26	0.67	33	264	42.2	0.88
FNfn10	0.021	0.56	0.25	30	1647	18.3	1.35
srcSH3	0.064	0.13	0.59	91	267	27.3	1.10
Tenascin	0.005	0.76	0.48	36	1647	18.3	1.35
TI-I27	0.244	0.02	0.65	29	3056	17.8	1.39
U1A	0.103	0.34	0.50	12	406	27.1	1.08
Villin	0.007	0.71	0.59	21	42	36.7	0.91
WW	0.112	0.22	0.85	47	147	31.9	1.05

[a] Calculated using the reduced alphabet by Mirny & Shakhnovich,[8] which clusters the residues into six classes: (AVLIMC), (FWYH), (STNQ), (KR), (DE) and (GP).

[b] Patterson correlation coefficients for relationships are illustrated in Figure 1. The p-value represents the probability that, if the null hypothesis were true (i.e. that there exists no relationship among the data in Figure 1), an estimate of the slope would be generated as far or farther from zero than that actually observed. P_0 is the probability that median conservation greater than that observed for the set of residues with $\Phi > 0.5$ would be obtained by random chance given the distribution of sequence entropies of the characterized residues. If the mean (rather than median, as reported here for comparison)[8] is employed, the results are effectively equivalent (data not shown) save that the relationship for CheY becomes significant.

[c] Fraction of positions for which Φ-values have been reported.

[d] Calculated using residue identity.

known high Φ residues are, as often as not, more poorly conserved than the average characterized residue (Figure 2) and the median conservation of high Φ residues is indistinguishable from that expected for a set of residues randomly selected from among all characterized amino acids (Table 1). In contrast to well-established theoretical expectations,[3,4,7–9] both poorly and highly conserved residues are apparently equally likely to participate in the Φ-value defined folding nucleus.

Conclusions

It has been predicted on theoretical grounds that protein folding kinetics are under evolutionary control and asserted that such pressure will lead to the preferential conservation of residues participating in native-like transition state structures. The observation of greater than median conservation in residues reported to participate in the native-like folding nucleus, however, arises due to significant, uncorrected biases in the experimental data and does not provide support for this hypothesis. Here we describe, in contrast, the results of a study of conservation among all of the appropriately characterized proteins reported to date, using deep, high quality sequence alignments, the prescribed measure of sequence conservation and a variety of statistical analyses that avoid critical experimental biases. This exhaustive study produces no significant evidence in support of preferential conservation of the folding nucleus. If protein folding kinetics are under evolutionary control, the selective pressures arising from that control are apparently insufficient to generate measurable conservation among the currently characterized protein folding nuclei.

Acknowledgments

The authors acknowledge helpful, clarifying discussions with Eugene Shakhnovich, Leonid Mirny, Anthony Clarke and Mark Lorch. We also acknowledge Vijay Pande for the use of his facilities and Pande, Michael Gross, Richard Goldstein and Sophie Jackson for valuable commentary regarding the manuscript. S.M.L. is a James H. Clark Fellow of the SGF program.

References

1. 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.
2. 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.
3. Shaknovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature,* **379**, 96-98.

4. Mirny, L. A., Abkevich, V. I. & Shakhnovich, E. I. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA,* **95**, 4976-4981.
5. 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.
6. Poupon, A. & Mornon, J. P. (1999). Predicting the protein folding nucleus from a sequence. *FEBS Letters,* **452**, 283-289.
7. 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.
8. Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123-129.
9. Mirny, L. & Shakhnovich, E. (2001). Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biomol.* **30**, 361-396.
10. Plaxco, K. W., Riddle, D. S., Larson, S., Ruczinski, I., Thayer, E. C., Buchwitz, B. *et al*. (2000). Evolutionary conservation and protein folding kinetics. *J. Mol. Biol.* **298**, 303-312.
11. Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3-9.
12. 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.
13. 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.
14. 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.
15. Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225-270.
16. 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.
17. 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.
18. Ternstrom, T., Mayor, U., Akke, M. & Oliveberg, M. (1999). From snapshot to movie: phi analysis of protein folding transition states taken one step further. *Proc. Natl Acad. Sci. USA,* **96**, 14854-14859.
19. 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.
20. Choe, S. E., Li, L., Matsudaira, P. T., Wagner, G. & Shakhnovich, E. I. (2000). Differential stabilization of the transition state of the villin 14T folding reaction. *J. Mol. Biol.* **304**, 99-115.
21. Fowler, S. B. & Clarke, J. (2001). Mapping the folding pathway of an immunoglobulin domain: structural detail from phi value analysis and movement of the transition state. *Structure,* **9**, 355-356.
22. Cota, E., Steward, A., Fowler, S. B. & Clarke, J. (2001). The folding nucleus of a fibronectin type III domain is composed of core residues of the conserved immunoglobulin-like fold. *J. Mol. Biol.* **305**, 1185-1194.
23. Hamill, S., Steward, A. & Clarke, J. (2000). The folding of an immunoglobulin-like Greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165-168.
24. Jäger, M., Nguyen, H., Crane, J. C., Kelly, J. W. & Gruebele, M. (2001). The folding mechanism of a β-sheet: the WW domain. *J. Mol. Biol.* **311**, 373-393.
25. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. & Sonnhammer, E. L. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 263-266.
26. Holm, L. & Sander, C. (1998). Removing near-neighbor redundancy from large protein sequence collections. *Bioinformatics,* **14**, 423-442.
27. Henikoff, S. & Henikoff, J. G. (1994). Position-based sequence weights. *J. Mol. Biol.* **243**, 574-578.
28. Larson, S. M. & Davidson, A. R. (2000). The identification of conserved interactions within the SH3 domain by alignment of sequences and structures. *Protein Sci.* **9**, 2170-2180.
29. Maxwell, K. L. & Davidson, A. R. (1998). Mutagenesis of a buried polar interaction in an SH3 domain: sequence conservation provides the best prediction of stability effects. *Biochemistry,* **37**, 16172-16182.
30. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition-state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering. *J. Mol. Biol.* **254**, 260-288.
31. López-Hernández, 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.
32. 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.
33. 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.
34. Fulton, K. F., Main, E. R. G., Daggett, V. & Jackson, S. E. (1999). Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.* **291**, 445-461.

Edited by C. R. Matthews

(Received 9 August 2001; received in revised form 28 November 2001; accepted 7 December 2001)

http://www.academicpress.com/jmb

Supplementary Material for this paper is available on IDEAL