doi:10.1006/jmbi.2001.4602 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 308, 123–129

JMB

COMMUNICATION

Evolutionary Conservation of the Folding Nucleus

Leonid Mirny* and Eugene Shakhnovich*

Harvard University
Department of Chemistry and
Chemical Biology, 12 Oxford
Street, Cambridge
MA 02138, USA

**Corresponding authors*

Here, we present statistical analysis of conservation profiles in families of homologous sequences for nine proteins whose folding nucleus was determined by protein engineering methods. We show that in all but one protein (AcP) folding nucleus residues are significantly more conserved than the rest of the protein. Two aspects of our study are especially important: (i) grouping of amino acid residues into classes according to their physical-chemical properties and (ii) proper normalization of amino acid probabilities that reflects the fact that evolutionary pressure to conserve some amino acid types may itself affect concentration of various amino acid types in protein families. Neglect of any of those two factors may make physical and biological "signals" from conservation profiles disappear.

Keywords: protein folding; folding kinetics; folding nucleus; protein evolution; evolutionary conservation

It is now widely accepted that folding of small single-domain proteins follows a "nucleation-condensation" mechanism[1–6] where the rate-limiting transition state resembles an expanded and distorted native structure which is built around a specific folding nucleus (SFN). Considerable experimental[2,7–10] and theoretical[1,11–16] effort has been devoted to identification of folding nuclei in real proteins and various models as well as factors that determine its location in structure and in sequence.

One of the most intriguing aspects of the nucleation-condensation mechanism of protein folding is its relation to protein evolution. Indeed residues constituting the folding nucleus can be metaphorically considered "accelerator pedals" of folding[17] since mutations in those positions affect folding rate to a much greater extent than elsewhere in a protein. It can be concluded that if there is evolutionary control of folding rate it should have resulted in additional pressure applied on folding nucleus residues, and such pressure can be manifested in noticeable additional conservation of nucleus residues.

This idea was first proposed in[18] where it was applied to prediction of nucleus residues from protein structure. Many sequences were designed to fit the structure of chymotrypsin inhibitor 2 (CI2) with low energy. Positions conserved among the designed sequences were identified as a putative nucleus. This way blind predictions of folding nucleus in CI2 were made that were verified in independent experiments.[2]

In related studies, Ptitsyn studied conservatism in distant yet related by sequence homology members of cytochrome *c*[19] and myoglobin[20] families. In both cases, he found conserved clusters of residues without an obvious functional role which he suggested to belong to the folding nucleus of those proteins. Michnick & Shakhnovich[21] carried out an analysis of conservation in natural and designed sequences for families of three structurally related proteins, ubiquitin, raf and ferredoxin, and predicted possible folding nucleus for those proteins.

Nevertheless, the notion of folding nucleus conservation has drawn some controversy in the literature. While earlier papers[18–21] suggested conservation of folding nucleus in some proteins, a more recent paper by Plaxco and co-authors[22] argued to the opposite. These authors looked at conservatism profile in several protein families for which protein engineering analysis of folding transition states has been carried out, and did not observe correlation between conservation and experimentally measured ϕ-values. This made them conclude that there is no evolutionary pressure to control the folding rates.

Here, we study evolutionary conservation of the folding nucleus for several homologous proteins.

Abbreviations used: SFN, specific folding nucleus; CI2, chymotrypsin inhibitor 2; H0, null hypothesis.

E-mail address of the corresponding author:
leonid@origami.harvard.edu; eugene@belok.harvard.edu

0022-2836/01/020123–7 $35.00/0

Conservation of the folding nucleus is systematically compared with the conservation in the rest of the protein sequence. In contrast to previous studies, we perform rigorous statistical tests to assess significance of higher conservation in the folding nucleus. The main result of this study is that for all studied proteins, except AcP, folding nucleus is significantly more conserved than the rest of the protein. We explain the difference between our thorough statistical analysis and that of Plaxco *et al.*[22] by pointing out some technical shortcomings in the earlier work.[22]

Results and Discussion

To study evolutionary conservation of the folding nucleus we turn to nine proteins for which the nucleus has been experimentally identified from protein engineering analysis: CI2, FKBP12, ACBP, CheY, Tenascin, CD2.d1, U1A, AcP and ADA2 h. For each of them we obtain a multiple sequence alignment from HSSP database[23] (or PFAM.[24] database if HSSP contains too few sequences). We compute variability at position l of the alignment as:

$$s(l) = -\sum_{i=1}^{6} p_i(l) \log p_i(l) \qquad (1)$$

where $p_i(l)$ is the frequency of residues from class i in position l. We use six classes of residues to reflect physical-chemical properties of amino acids and their natural pattern of substitutions: aliphatic (A V L I M C), aromatic (F W Y H), polar (S T N Q), basic (K R), acidic (D E), and special (reflecting their special conformational properties) (G P). As a result of this classification mutations within a class are ignored (e.g. $V \rightarrow L$), while mutations that change the class are taken into account. Figure 1 presents variability profile for studied proteins with nucleation positions marked by filled circles. Importantly, we defined the folding nucleus as it was identified by the original experimental groups (Table 1).

Figure 2 shows clearly that nucleus residues are almost always among the most conserved ones for all studied proteins. It also shows that nucleus residues are not the only conserved ones: many other residues (predominantly in the cores of the proteins) are also conserved.

In order to evaluate the statistical significance of nucleus conservation we compare evolutionary conservation of the folding nucleus with the conservation of all residues in the protein using the following statistical test. We start from the null hypothesis (H0) that nucleus residues are no more conserved than the whole protein sequence. To test this hypothesis we compute median of variability of the nucleus residues (med[s_{nuc}]) and compare it with the distribution of medians of variability of the same number of residues randomly chosen in the same protein (f(med[s_{rand}])). The distribution f(med[s_{rand}]) is obtained by choosing 10^5 random sets of n residues (n is the number of residues in the nucleus). Then the fraction of instances with med[s_{rand}] < med[s_{nuc}] gives the probability P_0 of accepting H0. In other words, P_0 is the probability that observed lower variability of the folding nucleus is obtained by chance. Hence, $P_0 \leqslant \alpha$ indicates statistically significant strong evolutionary conservation of the folding nucleus. Below we use confidence level $\alpha = 2\,\%$.

Table 2 presents computed P_0 values. The main result of this work is that in all proteins, except AcP, residues in the folding nucleus are significantly more conserved than the rest of the protein.

Next, we study how the obtained results depend on the way amino acid residues are grouped into classes (see Table 2). When the classification scheme from.[25] (BT) is used, still all proteins except AcP exhibit significant conservation of the folding nucleus. This demonstrates clearly that the observed conservation of the folding nucleus is not a consequence of a particular choice of the classification scheme.

However, when amino acid residues are not grouped into classes, the nucleus exhibits significant conservation only in four out of nine proteins. Taken together these results indicate that substitutions in the folding nucleus may occur, but they are limited to residues that belong to the same class (i.e. have similar physical-chemical properties).[26]

To study what physical-chemical properties are conserved in the folding nucleus we used various classification schemes. Starting from all 20 amino acid residues, we grouped some of them into classes and repeated the analysis, including the statistical tests (see Table 2). The goal is to find a

Table 1. Folding nuclei as identified by the authors

Protein	PDB	Folding nucleus	Reference
CI2	2ci2I	A35 L68 I76	Itzhaki *et al.*[2]
Tenascin	1ten	I821 Y837 I860 V871	Hamill *et al.*[35]
CD2.d1	1hnf	L19 I21 I33 A45 V83 L94 W35	Lorch *et al.*[36]
CheY	3chy	D12 D13 D57 V10 V11 V33 A36 D38 A42 V54	Lopez-Hernandez & Serrano[37]
ADA2h	1aye	I15 L26 F67 V54 I23	Vilegar *et al.*[38]
AcP	1aps, 2acy	Y11 P54 F94 Y25 A30 G45 V47	Chiti *et al.*[9]
U1A	1urn	I43 V45 L30 F34 I40 I14 L17 L26	Ternstrom *et al.*[39]
ACBP	1aca	F5 A9 V12 L15 Y73 I74 V77 L80	Kragelund *et al.*[40]
FKBP12	1fkj	V2 V4 V24 V63 I76 I101 L50	Main *et al.*[7]

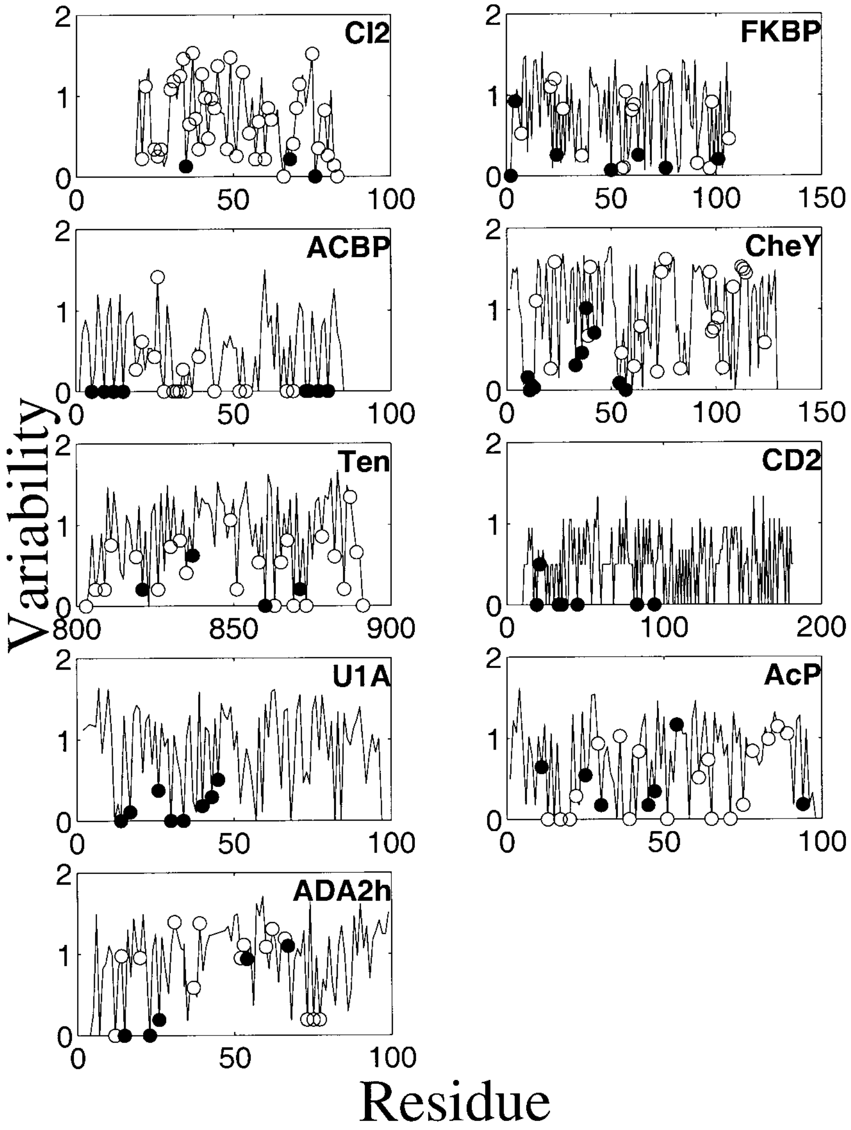

Figure 1. Variability profiles (sequence entropy) for nine different proteins computed using MS residue classes. Circles indicate positions at which ϕ-values have been experimentally measured. Residues forming the folding nucleus are shown by filled circles.

minimal classification (i.e. grouping the minimal number of amino acids together) that provides statistically significant conservation of the folding nucleus. Our results show that classification where only I, L, and V are grouped in one class while all other amino acid residues each represent their own

Table 2. Probability P_0 of nucleus being as conserved as the whole protein (see the text for details) computed for all nine proteins and seven different classification schemes

	MS	BT	No grouping	[I,L,V], [W,F,Y], [R,X], [D,E]	[I,L,V], [W,F,Y]	[I,L,V], [W,F]	[I,L,V]
N_{class}	6	5	20	14	16	17	18
CI2	**0.0041**	0.01	0.0382	0.007	0.002	0.004	0.0044
FKBP12	**0.0187**	0.02	0.1585	0.044	0.047	0.053	0.0363
ACBP	**$<10^{-5}$**	$<10^{-5}$	0.0216	0.022	0.008	0.0080	0.0067
CheY	**$<10^{-5}$**	$<10^{-5}$	0.0011	0.0040	0.0050	0.0020	0.0022
Ten	**0.008**	0.018	0.2477	0.0260	0.0220	0.0130	0.0197
CD2.d1	**$<10^{-5}$**	$<10^{-5}$	$<10^{-5}$	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	$<10^{-5}$
U1A	**0.0009**	0.001	0.0029	$<10^{-3}$	$<10^{-3}$	$<10^{-3}$	0.0002
AcP	**0.089**	0.086	0.0126	0.025	0.021	0.009	0.0136
ADA2h	**0.0023**	0.01	0.2674	0.02	0.022	0.018	0.0147

MS as in,[31,41] BT as in[25]: hydrophobic [A V F P M I L], polar [S T Y H C N Q W], basic [R K], acidic [D E], gly [G]), N_{class} number of groups in each classification.

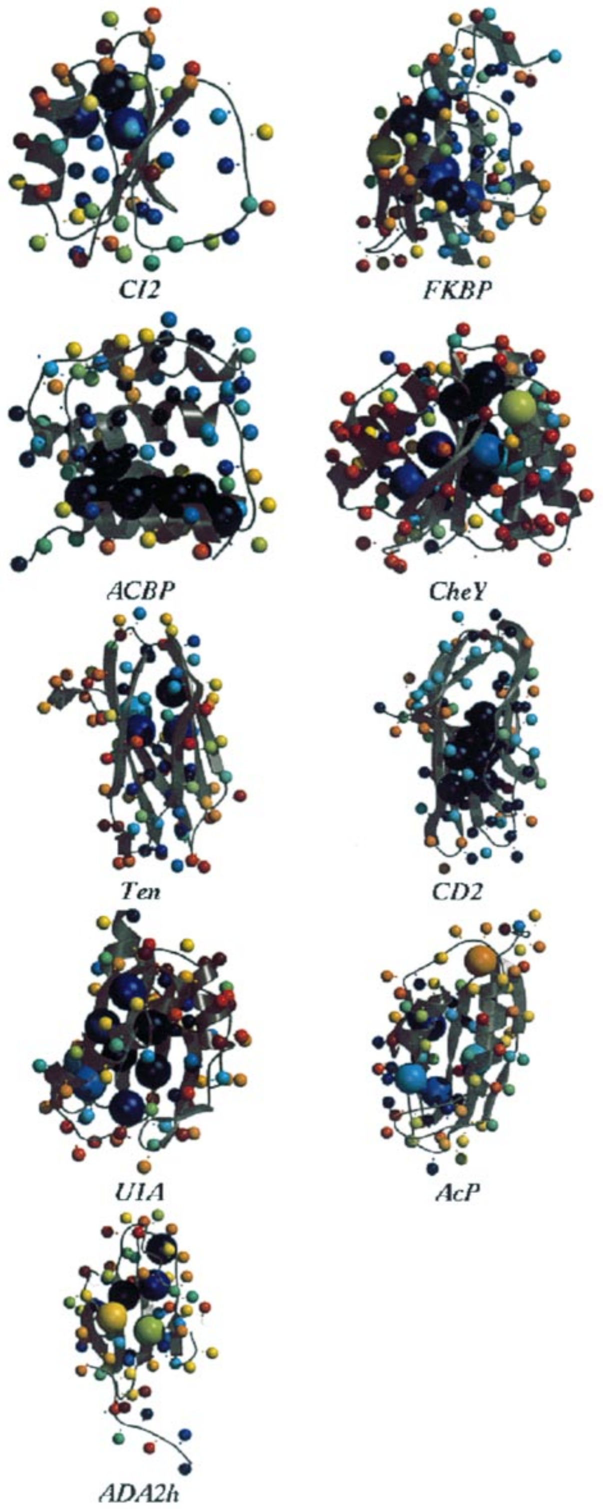

Figure 2. Nine studied proteins with C^{β} atoms colored according to the degree of their conservation (evaluated in Figure 1): from blue (high conservation) to light-blue, green, yellow and red (no conservation). Folding nucleus residues are shown by twice as large spheres. Notice conserved (blue) cores of the proteins and non-conserved (yellow and red) surfaces. Also notice several conserved non-nucleus residues in the protein core.

class satisfies this requirement (see Table 2).This classification provides significant conservation of the nucleus for all proteins except AcP with $\alpha = 5\,\%$, and for all proteins except AcP and FKBP12 with $\alpha = 2\,\%$. This result demonstrates that $I \rightleftharpoons L \rightleftharpoons V$ are the most common substitutions in the nucleus (and in the protein core in general-[27,28]). These substitutions are tolerated in the nucleus as they do not change much neither stability of the native fold nor the folding rate. Analysis of available experimental data (Supplementary Material) shows that changes in stability upon $I \rightleftharpoons L \rightleftharpoons V$ mutations are in average $\langle \Delta\Delta G_{N-D} \rangle = 1.0(\pm 0.4)$ kcal mol^{-1} for the native state and $\langle \Delta\Delta G_{\ddagger-D} \rangle = 0.2(\pm 0.3)$ kcal mol^{-1} for the transition state.

Note that grouping of residues into classes to assess conservation is similar to the use of substitution matrices in sequence alignment techniques. The underlying idea for both methods is to take into account natural physical-chemical similarity between amino acids and their substitution patterns. Plaxco *et al.* used all 20 types of amino acid residues and failed to identify strong conservation of the folding nucleus.[22] In similar way, a method that relies on simple sequence identity cannot detect distant homology. However distant homology between sequences can be detected using proper substitution matrices.[29,30] The use of substitution matrices is physically meaningful since they weight, e.g. $I - V$ match higher then $I - D$, while a method that relies on percentage of sequence identity weights $I - V$ and $I - D$ equally. Likewise, our amino acid classification scheme does not count $I \rightarrow V$ as a mutation, while it certainly considers substitutions like $I \rightarrow D$ as mutations to be counted.

Although, on average, the nucleus is more conserved than the rest of the protein, not all nucleating residues are strongly conserved. For example, in CheY two out of ten nucleation residues are not conserved. In ADA2 h two out of five and in tenascin one out of four residues are not conserved. Some nucleus residues may be less conserved because they belong to ''extended nucleus''[31] or because of limitation of our residue classification scheme that puts aromatic and aliphatic residues into two different groups, while aromatic-aliphatic substitutions may occur in the core of some proteins (i.e. tenascin, ADA2 h) usually as a result of correlated mutations that are not treated properly in this approach (but are taken into account in the conservation-of-conservation approach[31]). Another interesting observation is that the only protein that exhibits no preferential conservation of the folding nucleus is AcP, which is the slowest folding protein among all studied two-state folding proteins ($k_f^{H_2O} = 0.23\ s^{-1}$). Perhaps, this protein did not undergo evolutionary selection for faster folding and hence its folding nucleus is under no additional pressure to be conserved.

Importantly, several residues that do not belong to the nucleus are as conserved as the nucleating ones (see Figure 2). Those are the residues of the active site, a few core hydrophobic residues responsible for stabilization of the native structure and possibly some other residues whose conservatism is not fully understood. Residues of the folding nucleus are more conserved than the rest of the protein, but they may not be more conserved than the active site or key residues from the hydrophobic core that stabilize the protein. This result suggests that although the folding nucleus is conserved, it cannot be uniquely identified by analysis of a single protein family as residues conserved for various reasons (function, stability and folding nucleus) all contribute to the pattern of conservation (see Mirny & Shakhnovich[32]).

However excessive conservation of the nucleating residues can be detected in the cases when nucleus residues also belong to the hydrophobic core, i.e. when they are under dual pressure for folding and stability. Such excessive conservation can be revealed only when several families of proteins sharing the same fold are considered. By averaging the conservation over unrelated families one eliminates the functional conservation of the active site residues as they are typically located in the different parts of the fold in different families (with the exception of the super-site). The contribution of the protein stability can also be "subtracted" from the pattern of average conservation. This method reveals an excessive conservation of several folding nucleus residues (e.g. in the immunoglobulin domain).[31] However, from the conservation pattern of a single family one can hardly discriminate between the folding nucleus, the hydrophobic core and the active site. Nevertheless, analysis at the level of single families presented here clearly points to excessive conservation of the folding nucleus as compared to the rest of the protein.

Why do the results of our analysis differ from those of Plaxco *et al.*?[22] First, we took into account physical-chemical properties of amino acids and their natural substitution patterns to group amino acids into classes. As we showed, substitutions of large aliphatic residues (I,L,V) are frequent in folding nuclei and this confused the previous analysis that did not apply any amino acid classification scheme. While Plaxco *et al.* claimed in their paper[22] (without providing supporting evidence) that grouping of amino acid residues into classes did not change their conclusions, our analysis shows that proper classification of amino acid residues is crucial for detecting conservation in the folding nucleus.

Second, Plaxco *et al.* used a different method to compute sequence variability:

$$s_2(l) = -\sum_i p_i(l)\log[p_i(l)/p_i^0] \qquad (2)$$

This equation differs from equation (1), used in this study, in normalization by p_i^0, the "background" frequency of residue type i in all proteins. Although the difference may seem technical, equations (1) and (2) are based on two different models of evolution. We argue that while equation (2) may be adequate for DNA sequence analysis[33] it is not appropriate for analysis of protein evolution.

Equation (2) implicitly assumes that amino acid composition p_i^0 is fixed *a priori* in each protein. Hence, equation (2) tends to underestimate conservation of "frequent" amino acid residues (L,A,S etc), while overestimating conservation of less frequent amino acid residues (W,C,H etc). In contrast, equation (1) assumes that conservation requirement itself affects the composition, i.e. higher conservation of an amino acid leads to its higher frequency in proteins.

To illustrate this point consider a toy protein that consists of two types of residues: hydrophobic H and polar P. Assume that 70 % of amino acid residues in this protein are in the core and 30 % are in the loops. Also, assume that in the toy-world selection for stability requires a 100 % conservation of H amino acid residues in the core, while loops are under no evolutionary pressure and H and P are equally probable in the loops. Then $p_H^0 = 1 \times 0.70 + 0.5 \times 0.3 = 0.85$ and $p_P^0 = 0.5 \times 0.3 = 0.15$. At conserved core positions $s_2(\text{core}) = -1 \log 1/0.85 \approx -0.16$, while in the loops $s_2(\text{loops}) = -0.5 \log 0.5/0.85 - 0.5 \log 0.5/0.15 \approx -0.34$. Hence, the use of equation (2) leads to a counterintuitive and apparently wrong result $s_2(\text{core}) > s_2(\text{loops})$, i.e. that loops are more conserved than 100 % conserved core! It is clear that this result shows inadequacy of equation (2) as applied to protein evolution with unconstrained composition. In a similar way, application of equation (2) to real proteins leads to unreasonably low conservation of the hydrophobic core as compared to exposed loops (data not shown).

A possible way to compensate for variations in amino acid composition of proteins is to define the sequence entropy as in:[34]

$$s(l) = -\sum_i p_i(l)\log p_i(l) + \sum_i p_i^0 \log p_i^0 \qquad (3)$$

where the second term gives the background variability due to amino acid composition. This term however does not depend on l and hence does not change the relative variability.

It is interesting that the use of equation (2) by Plaxco *et al.*[22] gave rise to a surprising result that active sites in proteins are generally no more conserved than the rest of the protein (see Figure 2 of[22]). Conservation of known active sites was used as a control in[22] for their method of analysis based on equation (2), which it apparently failed.

Finally, Plaxco *et al.* did not study conservation of the folding nucleus. Instead, they focused on the residues that featured high ϕ-values in protein engineering experiments and compared them with low ϕ-value residues. As we explained above, residues in the folding nucleus do not necessarily exhibit high ϕ-values, and many low ϕ-value residues are conserved in evolution as they contribute to stabilization of the native structure. Comparison with low ϕ-value residues instead of comparison with the whole protein also confused the previous analysis since most of ϕ-values have been measured for amino acids located in the the core of a protein and hence these amino acid residues are on average more conserved. Here, in contrast, we used the folding nucleus as it was identified for each protein by the original experimental group and compared its conservation with the conservation of all amino acid residues in the protein.

In summary, we showed that the folding nucleus is indeed conserved in most of the proteins whose folding transition states are known from protein engineering analysis. That does not mean that folding nucleus residues are the only conserved ones in any family of homologous proteins. That also may not mean that folding nucleus is more conserved than other residues in the protein core, as the nucleus is equally important for protein stability and for fast folding. Our results show that the folding nucleus is more conserved than the rest of the protein. As stated earlier it is difficult to uniquely identify the folding nucleus by looking at a conservation profile in just one family of homologous sequences. Nevertheless conservation of folding nucleus found in this paper and in other works[12,31] points to an exciting possibility that folding rates may be of biological significance. Biological significance of this fact needs to be assessed in future studies.

Acknowledgments

We are grateful to Lewyn Li for useful discussions of experimental studies and for providing us with the data of ϕ-value analysis collected from the literature. L.M. is supported by Milton Fund and Harvard Society of Fellows. E.S. is supported by NIH GM52126.

References

1. Abkevich, V., Gutin, A. & Shakhnovich, E. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry,* **33**, 10026-10036.
2. Itzhaki, L., Otzen, D. & Fersht, A. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
3. Fersht, A. (1997). Nucleation mechanism of protein folding. *Curr. Opin. Struct. Biol.* **7**, 10-14.
4. Shakhnovich, E. (1997). Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.
5. Guo, Z. & Thirumalai, D. (1995). Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labelling experiments. *Biopolymers,* **35**, 137-139.
6. Pande, V., Grosberg, A. Y., Rokshar, D. & Tanaka, T. (1998). Pathways for protein folding: is a "new view" needed? *Curr. Opin. Struct. Biol.* **8**, 68-79.
7. Main, E., Fulton, K. & Jackson, S. (1999). Folding pathway of fkbp12 and characterisation of the transition state. *J. Mol. Biol.* **291**, 429-444.
8. Martinez, J., Pissabarro, T. & Serrano, L. (1998). Obligatory steps in protein folding and the conformational diversity of the transition state. *Nature Struct. Biol.* **5**, 721-729.
9. Chiti, F., Taddei, N., White, P., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Struct. Biol.* **6**, 1005-1009.
10. Grantchanova, V., Riddle, D., Santiago, J. & 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.
11. Klimov, D. & Thirumalai, D. (1998). Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J. Mol. Biol.* **282**, 471-492.
12. Li, L., Mirny, L. & Shakhnovich, E. (2000). Kinetics, thermodynamics and evolution of non-native interactions in protein folding nucleus. *Nature. Struct. Biol.* **7**, 336-341.
13. Alm, E. & Baker, D. (1999). Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl Acad. Sci. USA,* **96**, 11305-11310.
14. Munoz, V. & Eaton, W. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA,* **96**, 11311-11316.
15. Galzitskaya, O. & Finkelstein, A. (1999). A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl Acad. Sci. USA,* **96**, 11299-11304.
16. Dokholyan, N., Buldyrev, S., Stanley, H. & Shakhnovich, E. (2000). Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* **296**, 1183-1188.
17. Mirny, L., Abkevich, V. & Shakhnovich, E. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA,* **95**, 4976-4981.
18. Shakhnovich, E., Abkevich, V. & Ptitsyn, O. (1996). Conserved residues and the mechanism of protein folding. *Nature,* **379**, 96-98.
19. Ptitsyn, O. (1998). Protein folding and protein evolution: common folding nucleus in different subfamilies of *c*-type cytochromes? *J. Mol. Biol.* **278**, 655-666.
20. Ptitsyn, O. & Ting, K. (1999). Non-functional conserved residues in globins and their possible role as a folding nucleus. *J. Mol. Biol.* **291**, 671-682.
21. Michnick, S. & Shakhnovich, E. (1998). A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Fold. Des.* **3**, 239-251.
22. Plaxco, K., Larson, S., Ruczinski, I., Riddle, D., Buchwitz, B., Davidson, A. & Baker, D. (2000). Evolutionary conservation in protein folding kinetics. *J. Mol. Biol.* **298**, 303-312.
23. Dodge, C., Schneider, R. & Sander, C. (1998). The hssp database of protein structure-sequence align-

ments and family profiles. *Nucl. Acids Res.* **26**, 313-315.

24. Bateman, A., Birney, E., Durbin, R., Eddy, S., Howe, K. & Sonnhammer, E. (2000). The pfam protein families database. *Nucl. Acids Res.* **28**, 263-266.
25. Branden, C. & Tooze, J. (1998). *Introduction to Protein Structure*, Garland Publishing, Inc., New York.
26. Thompson, M. & Goldstein, R. (1996). Constructing amino acid residue substitution classes maximally indicative. *Proteins: Struct. Funct. Genet.* **25**, 28-37.
27. Henikoff, S. & Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.
28. Benner, S., Cohen, M. & Gonnet, G. (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng.* **7**, 1323-1332.
29. Abagyan, R. & Batalov, S. (1997). Do aligned sequences share the same fold? *J. Mol. Biol.* **273**, 355-368.
30. Brenner, S., Chothia, C. & Hubbard, T. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA,* **95**, 6073-6078.
31. Mirny, L. & Shakhnovich, E. (1999). Universally conserved residues in protein folds. Reading evolutionary signals about protein function, stability and folding kinetics. *J. Mol. Biol.* **291**, 177-196.
32. Mirny, L. & Shakhnovich, E. (2001). Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biophys. Chem.* **30**. In the press.
33. Stormo, G. (1998). Information content and free energy in dna-protein interactions. *J. Theoret. Biol.* **195**, 135-137.
34. Schneider, T. (1999). Measuring molecular information. *J. Theoret. Biol.* **201**, 87-92.
35. 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.
36. 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.
37. 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, ci2. *Fold. Des.* **1**, 43-55.
38. Vilegas, V., Martinez, J., Avilez, F. & Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase a2 activation domain. *J. Mol. Biol.* **283**, 1027-1036.
39. Ternstrom, T., Mayor, U., Akke, M. & Oliveberg, M. (1999). From snap-shot to movie: phi-value analysis of protein folding transition states taken one step further. *Proc. Natl Acad. Sci. USA,* **96**, 14854-14859.
40. Kragelund, B., Osmark, P., Neergaard, T., Schiodt, J., Kristiansen, K., Knudsen, J. & Poulsen, F. (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.
41. Mirny, L., Abkevich, V. & Shakhnovich, E. (1998). How evolution makes proteins fold quickly. *Proc. Natl Acad. Sci. USA,* **95**, 4976-4981.

Edited by A. R. Fersht

(Received 4 December 2000; received in revised form 9 March 2001; accepted 9 March 2001)

Supplementary Material comprising a Table of phi values is available on IDEAL