doi:10.1016/S0022-2836(03)00324-3 J. Mol. Biol. (2003) 328, 921–926

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

What is the Minimum Number of Letters Required to Fold a Protein?

Ke Fan and Wei Wang*

National Laboratory of Solid State Microstructure and Department of Physics Nanjing University, Nanjing 210093, People's Republic of China

*Corresponding author

Experimental studies have shown that the full sequence complexity of naturally occurring proteins is not required to generate rapidly folding and functional proteins, i.e. proteins can be designed with fewer than 20 letters. This raises the question of what is the minimum number of amino acid types required to encode complex protein folds? Here, we investigate this issue from three aspects. First, we study the minimum sequence complexity that can reserve the necessary structural information for detection of distantly related homologues. Second, we compare the ability of designing foldable model sequences over a wide range of reduced amino acid alphabets, which find the minimum number of letters that have the similar design ability as 20. Finally, we survey the lower bound of alphabet size of globular proteins in a non-redundant protein database. These different approaches give a remarkably consistent view, that the minimum number of letters required to fold a protein is around ten.

Keywords: sequence complexity; reduced amino acid alphabet; protein folding; protein design; protein evolution

Introduction

There are thousands of families of proteins whose unique native structures have been determined. The total number of families in nature is much larger, and is estimated to be around 23,100.[1] The large majority of these proteins are composed of complex patterns of the 20 kinds of proteogenic amino acids. This pattern and compositional complexity has proven to be a significant hurdle to scientists attempting to crack the protein folding code or explore the sequence–structure relationship. However, is the full complexity of naturally occurring sequences necessary for proteins to encode their unique native structures?

A number of experimental and theoretical studies have suggested that the answer to the above question might be negative.[2–6] In fact, it is well known that some amino acid residues are similar in physico-chemical properties and their substitutions are tolerated in many regions of a protein sequence. This implies that it may be applicable to use fewer letters rather than 20 to design a protein. For example, Stroud and co-workers have recently published the synthesis and structure of a *de novo* designed 108 residue protein composed of just seven residue types that folds into a fully native, four-helix bundle.[7] Another example is given by Baker's laboratory.[8] They used a phage display technique to select for significantly simplified proteins that adopt an SH3 domain fold and found that this predominantly β-sheet fold can be encoded largely by a five-letter amino acid alphabet. Moreover, the folding rates of reduced alphabet proteins are very close to that of the naturally occurring SH3 domain, regardless of the dramatic changes in sequence. Very recently, a simplified 213 residue enzyme was fabricated with conservation of its *in vivo* function.[9] In the final variant, only nine amino acid types occupy 88% of its positions, and the number of total amino acid types is 13. This demonstrated that large proteins can be designed with a reduced set of amino acids. Taken together, these results show clearly that even compositionally simple sequences are likely to be able to encode foldable and functional proteins.

Since the full sequence complexity is not necessary for protein design, the question naturally arises: is there a lower bound of amino acid types required for a protein to fold itself into a stable structure and perform its function? If the answer is yes, what is the minimum number of letters? Is the minimum number of letters to fold an all-α

Abbreviation used: EPQ, error per query.

E-mail address of the corresponding author: wangwei@nju.edu.cn

protein the same as that for an all-β protein? The answers to these questions may have special significance for protein design, as well as for studies on the protein folding problem, because using fewer letters in design may greatly reduce the complexity of *de novo* design of a protein. Besides, it has been thought that for polypeptide chains consisting of fewer letters, the physics and chemistry may be sufficiently simplified for a thorough understanding of the protein folding code.[10,11] Furthermore, some unexpected sequence–structure patterns may emerge if the extra sequence complexity of naturally occurring proteins could be reduced in a reasonable way.

Here, we investigate the issue of the minimum number of letters required to fold a protein from three aspects: (1) the minimum sequence complexity required for detection of distantly related proteins; (2) the minimum number of letters required to design foldable model sequences; and (3) the lower bound of alphabet size for globular proteins. Hopefully, our systematic investigation will yield an insight into this interesting and significant question.

Results and Discussion

Minimum sequence complexity required for detection of distantly related proteins

It is well established that the structural information of a protein is encoded in its amino acid sequence. Therefore, ideally one should be able to predict the structure of a protein from its sequence alone. However, this goal will not be achieved in the near future, due to the complexity of the sequence–structure relationship. Nevertheless, one can detect the possible homologues of a protein in the database by sequence alignment methods and further determine its structure. The full sequence complexity is not required for encoding the structural information; that is, one can use a restricted code to encode the sequence with the necessary structural information still retained. Thus, we may ask to what extent can the sequence complexity of a protein be reduced with the necessary structural information still reserved?

To answer the above question, we first work out a series of reduced alphabets by clustering the most closely related residues into a group (see Figure 1 for the grouping scheme used in this work, and details about grouping residues may be found elsewhere[11] (T. P. Li *et al.*, unpublished results) and then use these reduced alphabets to simplify the sequence; say, each amino acid residue in the sequence was replaced with the representative one in its group. Finally, we compare the structural information encoded in the simplified sequence with that of the original sequence so as to obtain a compromise between the loss of structural information and the reduction of sequence complexity. Because the structural informational

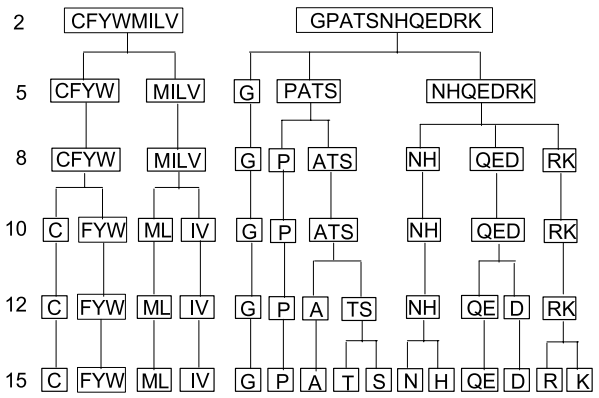

Figure 1. The grouping scheme used in this work. The single-letter representation of amino acids is used. Details of how to group amino acids will be presented elsewhere (T. P. Li *et al.*, unpublished results).

content encoded in a sequence cannot be measured directly, we turn to a somewhat indirect but more applicable method, assessing the ability of detection of distantly related proteins. Once a sequence is simplified, the structural information encoded will be partially lost. Thus, with a simplified sequence, the possibility of finding its true homologues by alignment will be decreased. Here, the ability of detection of distantly homologues is used as a measure of the structural information encoded in a sequence.

To compare the original sequences with the simplified sequences for their ability to detect distantly related homologues, we perform an all-against-all sequence alignment for a non-redundant database, SCOP40.[12] With respect to the sequence alignment method, we use an ungapped version of the Smith–Waterman algorithm,[13] because the significance of the *E*-value is strictly proven. The scoring matrix used is BLOSUM62,[14] the default choice of many sequence alignment programs. For simplified sequences, the scoring matrix was simplified correspondingly, which was constructed using the average values of the matrix elements of residues

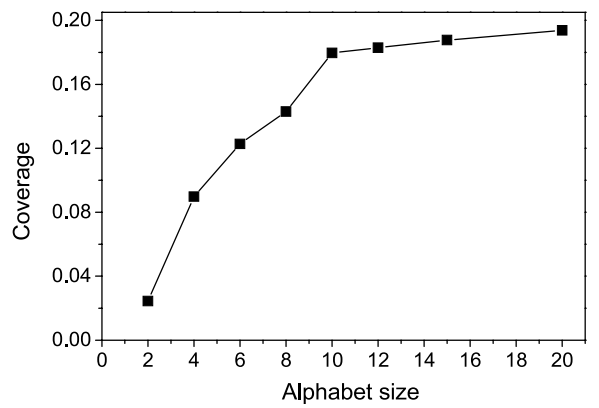

Figure 2. Coverage at EPQ = 0.001 as a function of the alphabet size for an all-against-all sequence alignment for the SCOP40 database.

in a group. We did not use the widely used BLAST[15] because that, for some reduced matrices, cannot compute the parameters K and λ.[16]

To assess the ability of detecting homologues, we use a quantity, namely coverage, defined as the fraction of structurally determined homologues that have scores above the selected threshold at a given error per query (EPQ). Here, the EPQ is an indicator of selectivity, defined as the number of non-homologous pairs above the threshold divided by the number of queries. Figure 2 shows the coverage at EPQ = 0.001 for various reduced alphabets. It can be seen that the coverage remains almost the same when the alphabet size is reduced from 20 to ten. This suggests that even though the sequence complexity is reduced greatly, the structural information encoded is still reserved to a large extent, which enables one to find its homologues with almost the same degree of accuracy. In other words, the necessary sequence complexity required to encode a protein fold is around ten. Extra sequence complexity may be a result of the long history of evolution. Recently, Murphy *et al*. performed similar sequence alignment but using BLAST and with a different grouping scheme.[17] Although the calculation of parameters K and λ for reduced alphabets is somewhat problematic,† which results in the low coverage especially for the few-letter alphabets, their general conclusion is consistent with ours.

To investigate further the performance of reduced alphabets, we extracted three subsets from the SCOP40, all-α, all-β, and α/β proteins. We found that the coverage is different for the three subsets. Specifically, it is easiest to detect the homologues of all-α proteins, and most difficult for all-β proteins. Nevertheless, the relative accuracy of detection of distantly related homologues of reduced matrices is almost the same for the three subsets, which suggests that almost the same necessary sequence complexity is required for encoding different folds.

Minimum number of letters required to design foldable model sequences

Nature uses 20 kinds of amino acids to design protein sequences. With the 20 kinds of amino acids, it is easy to generate viable (or foldable) sequences. One may wonder what will happen if the letters used in the design are reduced. A reasonable speculation is that the fraction of viable sequences in the design will decrease if fewer letters are used. However, is the decrease in the fraction of viable sequences linear with the number of letters used in the design? And to what extent can the number of letters used in the design be reduced while still keeping the fraction of viable sequences comparable with that when using 20 letters?

Currently, large-scale *de novo* protein-designing experiments are not feasible to study the above question, so we turn to a simplified lattice model. Despite its great simplification of the atomic details of a protein, the lattice model is able to capture key features of protein stability and folding, and is therefore used widely in studies of protein folding and protein structure prediction.[18,19] In our model, a 27 residue protein chain is confined to a maximally compact $3 \times 3 \times 3$ three-dimensional lattice, in which each residue occupies a lattice site. For the maximally compact $3 \times 3 \times 3$ three-dimensional lattice proteins, there are a total of 51,704 possible self-avoiding walks on this lattice, excluding rotations and reflections, which represents the 51,704 different possible conformations for the protein chain. The energy function for a sequence in a particular conformation is given by a simple pair-contact form:

$$E = \sum_{i<j} B_{ij}\Delta_{ij} \quad (1)$$

where B_{ij} is the contact energy between residues i and j, and Δ_{ij} is equal to unity if residues i and j are space neighbor and zero otherwise. For the 20 letters, we use the most widely used MJ potential.[20] While for reduced alphabets with size N, the 20×20 MJ matrix is reduced to an $N \times N$ matrix, in which the elements are the average values of the MJ matrix elements of residues in a group.

Here, we adopt the approach of previous work and presume that a sequence is viable if: (1) its lowest-energy structure is non-degenerate, i.e. there is only one unique lowest-energy structure among the 51,704 possible conformations for a viable sequence; and (2) its Z-score is lower than a threshold value. Here, the first factor is for the thermodynamic stability, and the second factor is to ensure that the sequence can fold into its native structure within a reasonable time-scale.[21] For a given sequence, its Z-score can be calculated as:

$$Z = (E_N - \langle E \rangle)/\sigma \quad (2)$$

where E_N is the energy of the native conformation, and $\langle E \rangle$ and σ are, respectively, the mean and the standard deviation of the energy distribution of the 51,704 possible conformations. The threshold of Z-score used in this work is -3.6. Please note that our result is not very sensitive to the choice of Z-score, because we focus on the comparison of design ability among different alphabets.

For each alphabet with its specific amino acid contact potential, we generate 2×10^6 sequences

† The calculation of parameters K and λ is closely related with the scoring matrix and the compositions of amino acid residues. Strictly, for reduced alphabets with N letters (or groups), the scoring matrix should also be reduced to an $N \times N$ matrix, and the composition of a group should be the sum of composition of residues in the group. However, when using BLAST, the scoring matrix and the composition are always for 20 letters. Obviously, this kind of difference will have an effect on the calculation of parameters K and λ, and further on the coverage. As shown here, this effect is especially significant for few-letter alphabets.

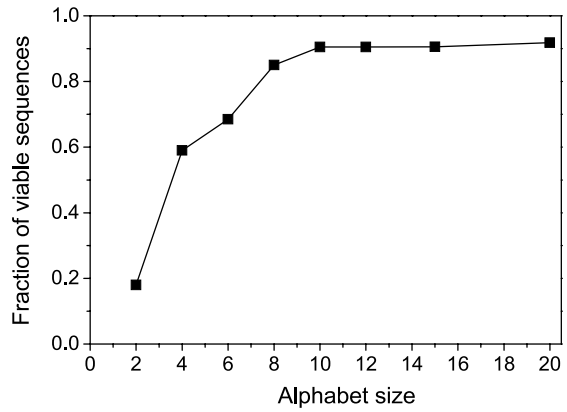

Figure 3. Fraction of viable sequences *versus* the alphabet size for a 27mer lattice protein model.

and then survey how many sequences are viable. From Figure 3, we can see that for the 20 letter alphabet, most sequences have non-degenerate native structure and low *Z*-score. With the reduction of the letters used in the design of sequences, the fraction of viable sequences begins to decrease, as expected. However, it still remains at a comparable level relative to the 20 letter alphabet in a rather wide range of alphabet size. In detail, from 20 to ten, the fraction of viable sequences decreases only slightly, from 0.92 to 0.91. After that, the fraction of viable sequences decreases sharply with the decrease of alphabet size, until it reaches 0.18 for a two-letter alphabet. This suggests that ten may be the minimum number of letters required to design foldable model sequences. To ensure that our results are robust to the choice of *Z*-score, we changed *Z*-score in a large range and perform the same survey. As expected, we observed similar results in all cases.

In addition, we found that the highly designable structures might vary from alphabet to alphabet, which suggests that extrapolating structural information from the few-letter alphabets may not be applicable to the properties of real 20-letter proteins. This is in agreement with Buchler and Goldstein's results for the two-dimensional case.[22] Specifically, the highly designable structures for two-letter alphabet are usually highly symmetric, while the highly designable structures are somewhat more complex for the 20-letter alphabet. Experimentally, it is found that the symmetrical structures are more designable by using limited sets of amino acids.[2,23] Therefore, for some particularly symmetric structures, it is possible to use a restricted code to design.

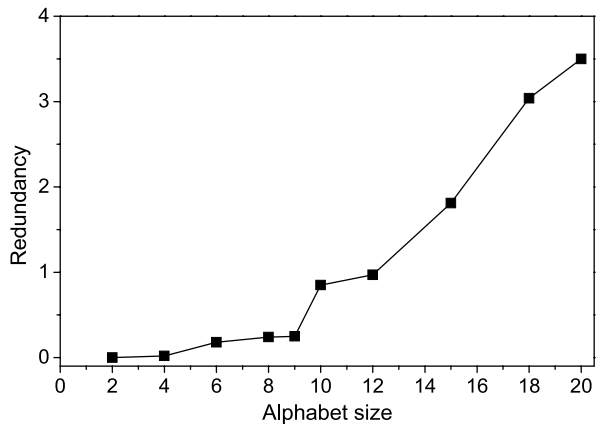

Figure 4. Redundancy *versus* the alphabet size. Here, the redundancy is defined to be the difference between the number of letters in the alphabet and the average number of letters observed in the sequence window.

Lower bound of alphabet size for globular proteins

Now we survey the number of letters that appear in protein domains in a non-redundant database, SCOP90 (version 1.59), to see whether there is a lower bound of alphabet size for globular proteins. First, we define an alphabet size, A, to be the number of different letters in any given sequence window w :

$$A(w_j) = \sum_{i=1}^{N} \delta_i \qquad (3)$$

where $\delta_i = 1$ if residue i is present in window j and 0 otherwise, and N is the number of letters in the alphabet. Here, the choice of w should be cautious. Too small w may cause large statistical fluctuations, while too large w may blur the resolution of different regions for the analysis of segments within proteins. Previous experimentation on protein sequence databases suggested that $w = 45$ provides a good starting point of complexities of different sequences.[24] So, we adopt $w = 45$ in this work. Previously, the issue of the lower bound of sequence complexity of globular proteins has been studied.[25] However, during the three years since the work was done, the number of determined protein structures in the RCSB Protein Data Bank (PDB) has grown quickly, and now it is almost twice that in 1998. Here, we re-study this issue with the latest protein database and make a more detailed analysis.

Similarly, we check the three subsets of SCOP90, all-α, all-β, and α/β proteins. It is found that there is minor difference in the average alphabet size, while the minimal alphabet size for all-α and α/β proteins is nine, slightly lower than ten for all-β proteins. We also notice that for the 20-letter alphabet, the average alphabet size $\langle A \rangle$ is 16.37, i.e. only about 16 out of 20 letters appeared in the sequence window, on average. This means that nature does not use all 20 letters in every design, and there is redundancy in the alphabet. Therefore, we perform a simplification for the sequences using the reduced alphabets, and investigate when the redundancy will vanish. In Figure 4, we see clearly that the redundancy almost vanishes when the alphabet size is reduced to nine. This suggests that at least nine kinds of amino acids are required to encode naturally occurring proteins.

Next, we did the necessary statistical analysis on a random database as a control. Given the current prevalence of the 20 amino acids, we observed 16.9 amino acids at random in a set of 45, very close to that of real protein database. However, the smallest alphabet size observed at random is 11, a little bit larger than that of SCOP. Thus, it seems that protein sequence deviates only slightly from random sequences. This is in agreement with the previous statement, that proteins are only slightly edited random copolymers.[26] The small degree of editing of protein sequences is related closely to the neutral theory of evolution. We re-do the analysis on the random database with the reduced alphabets, and observe results similar to that of the real protein database. It seems that the observed redundancy probably represents the process of statistical sampling, which is somewhat surprising. Anyway, the lower bound of alphabet size observed is nine, which is in remarkably good agreement with results obtained from other approaches.

Conclusion

In this work, we carry out a detailed study on the issue of the minimum number of letters required to encode a protein fold. The results obtained from three different approaches give remarkably consistent conclusions. It is found that the minimum number of letters required for a natural protein to encode its structure is around ten. However, we do not exclude the possibility of using fewer letters to design proteins with highly symmetric structures or some small proteins. Our results suggested that the complexity of coding required for real protein-like behavior is less than that of naturally occurring proteins, but greater than that used in the simplest computer models, which is consistent with previous statements.[8,27] The extra complexity in natural proteins may have functional or evolutionary reasons, which is difficult to study using simple models. In a different approach of extracting energy-like quantities from protein structures, Thomas and Dill found that all the training set proteins can be predicted correctly if the number of amino acid classes is larger than nine, which suggested that the minimum number of amino acid classes required to learn the training set proteins is ten.[28] This provides substantial support for our results. Researchers on protein evolution have proposed that primordial proteins might comprise a simpler set of amino acids, which involves only 7–13 amino acids, and that the existing system for protein synthesis has evolved progressively from the primordial one by gradually obtaining new amino acids for the repertoire in protein synthesis.[29–32] Our results are in agreement with such a proposal. It should be noted that our results are not sensitive to the precise grouping of amino acids, provided that the grouping method is reasonable. In fact, the difference is subtle for grouping of amino acids using different criteria, and the subtle difference has a minor effect on our results.

In summary, the full sequence complexity of naturally occurring proteins may be a result of a long history of neutral evolution, while for designing rapidly folding or even functional proteins, the full sequence complexity is not needed. This provides convenience to study the protein folding problem and the relationship of sequence and structure. Some general features may emerge as a result of the reasonable reduction of extra sequence complexity. In recent studies, unexpected sequence conservation has been found for folding purpose other than functional reasons, by dividing the 20 amino acids into six classes.[33–35] Another exciting finding identified several thousand significant sequence–structure patterns, with classifying the 20 amino acids into seven classes.[36] These unexpected patterns have been discovered, to a large extent, due to a reasonable reduction of extra sequence complexity of natural proteins. Therefore, we have reason to believe that with a more refined grouping of amino acids, say, assign different grouping for segments with different secondary structure, we will reach a better understanding of the sequence–structure relationship and the mechanism underlying folding.

Acknowledgements

We thank Ken Dill for helpful comments on the manuscript. This work was supported by the Foundation of NNSF (nos 90103031, 10204013, 10074030, 10021001), and the non-linear project (973) of the NSM.

References

1. Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature*, **372**, 631–634.
2. Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. & Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science*, **262**, 1680–1685.
3. Regan, L. & Degrado, W. F. (1988). Characterization of a helical protein designed from first principles. *Science*, **241**, 976–978.
4. Plaxco, K. W., Riddle, D. S., Grantcharova, V. & Baker, D. (1998). Simplified proteins: minimalist solutions to the "protein folding problem". *Curr. Opin. Struct. Biol.* **8**, 80–85.
5. Clarke, N. D. (1995). Sequence "minimization": exploring the sequence landscape with simplified sequences. *Curr. Opin. Biotech.* **6**, 467–472.
6. Beasley, J. R. & Hecht, M. H. (1997). Protein design: the choice of *de novo* sequences. *J. Biol. Chem.* **272**, 2031–2034.
7. Schafmeister, C. E., LaPorte, S. L., Miercke, L. J. W. & Stroud, R. M. (1997). A designed four helix bundle

protein with native-like structure. *Nature Struct. Biol.* **4**, 1039–1046.
8. Riddle, D. S., Santiago, J. V., Bray-Hill, 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.
9. Akanuma, S., Kigawa, T. & Yokoyama, S. (2002). Combinatorial mutagenesis to restricted amino acid usage in an enzyme to a reduced set. *Proc. Natl Acad. Sci. USA*, **99**, 13549–13553.
10. Chan, H. S. (1999). Folding alphabets. *Nature Struct. Biol.* **6**, 994–996.
11. Wang, J. & Wang, W. (1999). A computational approach to simplifying the protein folding alphabet. *Nature Struct. Biol.* **6**, 1033–1038.
12. Murzin, A. G., Breener, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of protein database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
13. Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
14. Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
15. Altschul, S. F., Gish, W., Miller, W., Meyers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
16. 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.
17. Murphy, L. E., Wallqvist, A. & Levy, R. M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.* **13**, 149–152.
18. Dill, K. A., Brombers, S., Yue, K., Fiebig, M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995). Principles of protein folding—a perspective from simple exact models. *Protein Sci.* **4**, 561–602.
19. Msirny, L. & Shakhnovich, E. (2001). Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 361–396.
20. Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures quasi-chemical approximation. *Macromolecules*, **18**, 534–552.
21. Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I. (1995). Evolution-like selection of fast-folding model proteins. *Proc. Natl Acad. Sci. USA*, **92**, 1282–1286.
22. Buchler, N. E. & Goldstein, R. A. (1999). Effect of alphabet size and foldability requirements on protein structure designability. *Proteins: Struct. Funct. Genet.* **34**, 113–124.
23. Davidson, A. R., Lumb, K. J. & Sauer, R. T. (1995). Cooperatively folded proteins in random sequences libraries. *Nature Struct. Biol.* **2**, 856–864.
24. Wootton, J. C. (1994). Nonglobular domains in protein sequences—automated segmentation using complexity measures. *Comput. Chem.* **18**, 269–285.
25. Romero, P., Obradovic, Z. & Dunker, A. K. (1999). Folding minimal sequences: the lower bound for sequence complexity of globular proteins. *FEBS Letters*, **462**, 363–367.
26. Ptitsyn, O. B. & Volkenstein, M. V. (1986). Protein structures and neutral theory of evolution. *J. Biomol. Struct. Dynam.* **4**, 137–156.
27. Wolynes, P. G. (1997). As simple as can be? *Nature Struct. Biol.* **4**, 871–874.
28. Thomas, P. D. & Dill, K. A. (1996). An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl Acad. Sci. USA*, **93**, 11628–11633.
29. Wong, J. T. (1975). A co-evolution theory of the genetic code. *Proc. Natl Acad. Sci. USA*, **72**, 1909–1912.
30. Osawa, S., Jukes, T. H., Watanabe, K. & Muto, A. (1992). Recent evidence for evolution of the genetic code. *Microbiol. Rev.* **56**, 229–264.
31. Baumann, U. & Oro, J. (1993). Three stages in the evolution of the genetic code. *Biosystems*, **29**, 133–141.
32. Di Giulio, M. & Medugno, M. (1999). Physicochemical optimization in the genetic code origin as the number of codified amino acids increases. *J. Mol. Evol.* **49**, 1–10.
33. Ptitsyn, O. B. & Ting, K. L. H. (1999). Non-functional conserved residues in globins and their possible role as a folding nucleus. *J. Mol. Biol.* **291**, 671–682.
34. Mirny, L. & Shakhnovich, E. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* **291**, 177–196.
35. Mirny, L. & Shakhnovich, E. (2001). Evolutionary conservation of the folding nucleus. *J. Mol. Biol.* **308**, 123–129.
36. Bradley, P., Kim, P. S. & Berger, B. (2002). TRILOGS: discovery of sequence–structure patterns across diverse proteins. *Proc. Natl Acad. Sci. USA*, **99**, 8500–8505.

Edited by J. Thornton

(Received 13 November 2002; received in revised form 5 March 2003; accepted 6 March 2003)