Eur. Phys. J. B **47**, 431–436 (2005)
DOI: 10.1140/epjb/e2005-00333-x

THE EUROPEAN PHYSICAL JOURNAL B

Size dependent complexity of sequences in protein families

J. Li[1], J. Wang[1], and W. Wang[1,2,a]

[1] National laboratory of Solid State Microstructure, Institute of Biophysics, and Department of Physics, Nanjing University, 210093, P.R. China
[2] Interdisciplinary Center of Theoretical Studies, Chinese Academy of Sciences, Beijing 100080, P.R. China

Received 11 January 2005 / Received in final form 6 June 2005
Published online 28 October 2005 –

Abstract. The size dependent complexity of protein sequences in various families in the FSSP database is characterized by sequence entropy, sequence similarity and sequence identity. As the average length L_f of sequences in the family increases, an increasing trend of the sequence entropy and a decreasing trend of the sequence similarity and sequence identity are found. As L_f increases beyond 250, a saturation of the sequence entropy, the sequence similarity and the sequence identity is observed. Such a saturated behavior of complexity is attributed to the saturation of the probability P_g of global (long-range) interactions in protein structures when $L_f > 250$. It is also found that the alphabet size of residue types describing the sequence diversity depends on the value of L_f, and becomes saturated at 12.

PACS. 87.10+e General theory and mathematical aspects – 87.15.Cc Folding and sequence analysis

1 Introduction

Proteins with large sequence similarity or same homologies are classified into the same family, and one family usually relates to one structural prototype which is given the term fold. In a family, although the structural features of the proteins are basically the same, the sequences may be rather different. The consequence is that the mapping from protein structure to protein sequence is not unique unlike from protein sequence to protein structure, implying a complexity of the coding from sequences to structures. To understand how proteins encode their structures, sequence analysis of protein families is usually undertaken, and has become a hot topic in molecular biology [1–6].

Based on the protein primary database, i.e., Protein Data Bank (PDB) [7], several databases of protein families have been compiled into a protein secondary database, such as the database of families of structurally similar proteins (FSSP) [8], and the structural classification of proteins (SCOP) [9]. Many studies have been done on the analysis of various features for these databases of protein families [1–6]. The FSSP database resulting from structural alignment includes many families each with a certain structural prototype, and different families consist of different numbers of sequences. In the FSSP database, all the sequences of every family are aligned, i.e., all the related sites are aligned into columns. A site may have a (or have no) specific effect on protein structure and/or functions. Thus, different sites have different specificities. Some sites have a high specificity and the residues on them are irreplaceable. These sites are conserved with the same types of residues. Nevertheless, some others have a low specificity and the residues on them are replaceable. These sites are non-conserved with different types of residues. However, for some non-conserved sites, the specificity may not be very low and the number of residue types may be very small. Such specificity of site in protein sequences relates to the sequence complexity. The higher the averaged specificity over all sites in a family is, the lower the sequence complexity in this family is. Obviously, a lower complexity implies a lower diversity, i.e., less types of residues are needed to construct the protein sequences in the family. For different families the complexities are different and may depend on the averaged sequence lengths in the families, and the complexity can be characterized by sequence entropy or sequence similarity. What are the statistical features of the sequence complexities and the ratios of conservative sites for various protein families? How many types of residues on averaging are there for the sequences in different families? Do these relate to the lengths of sequences? These are important questions and have not been well clarified. The answers are also relevant to the alphabet size (or the number of residue types) for reduction of protein complexity based on the grouping of residues.

In this work, the complexity of sequences in various families of the FSSP database is characterized using the sequence entropy, similarity and identity. It is found that for these factors there is an increasing trend of the sequence entropy and a decreasing trend of the sequence similarity and identity first as the averaged length L_f of

[a] e-mail: wangwei@nju.edu.cn

sequences in the family increases, and then a saturation when $L_f > 250$. Such saturated behavior is attributed to the saturation in the probability P_g of global interactions in proteins when $L_f > 250$. The saturation of complexity indicates that the alphabet size for those families with $L_f > 250$ is about 12 types of residues. Thus when the sequence lengths are larger than 250 residues, one needs basically 12 types of residues to characterize the sequences. Our results support the suggestion that the reduction of residue types for proteins is size dependent [10].

2 Methods

2.1 Column entropy and sequence entropy

In a certain family, the column entropy [11,12] $D(i)$ at column i is defined as following:

$$D(i) = -\sum_{k=1}^{20} p_k(i) \ln p_k(i) \tag{1}$$

where $p_k(i)$ is the occupying frequency of residue k at column i among all N_s sequences in a family. Then the sequence entropy of a family is defined as:

$$D = N_a^{-1} \sum_{i}^{N_a} D(i) \tag{2}$$

where N_a is the number of columns.

2.2 Sequence similarity

The sequence similarity [13,14] of a family can be quantified by the similarity substitution matrix BLOSUM62 [15] as

$$S = \frac{1}{(N_a \times N_s^2)} \sum_{i}^{N_a} \sum_{m=1}^{N_s} \sum_{n=1}^{N_s} B(R_m(i), R_n(i)) \tag{3}$$

where $B(R_m(i), R_m(i))$ is the element of the matrix describing the score between residue $R_m(i)$ in sequence m and residue $R_n(i)$ in sequence n at column i. The close the similarity of the sequences, the higher the value of S is.

2.3 Sequence identity

Sequence identity [16,17] is a factor to characterize the information of residue types in a family, and is defined as:

$$I = \frac{1}{100 \times (N_a \times {N_s}^2)} \sum_{i}^{N_a} \sum_{m=1}^{N_s} \sum_{n=1}^{N_s} \delta(R_m(i), R_n(i)) \tag{4}$$

where $\delta(R_m(i), R_n(i))$ equals to 1 when residue $R_m(i)$ and $R_n(i)$ are of the same type, otherwise it is 0. Actually, the sequence identity I is similar to the sequence similarity S to some degree, however, it is a more exact factor to quantify the residue types in the family. A smaller value of I means less residue types in the family, and vice versa.

2.4 The alphabet size for protein family

From the nature of entropy, the entropy $D(i)$ actually describes the diversity of residues. That is, the number of residue types at column i can be translated by $e^{D(i)}$. For example, when $D(i) = 0$, there is only one residue type at column i. Therefore, the sequence entropy D for a family can be transferred to the averaged types of residues, i.e., the alphabet size for a protein family, and it is defined as the following:

$$R = e^D. \tag{5}$$

Actually, the value of R is related to the minimal number of residue types required to characterize the sequences in the family. When $D = 0$, the number of residue types for the sequences at every column is unity, while for $D = \ln 20$, i.e., the maximal value of D, 20 types of residues are needed.

2.5 Family length

To detect the scaling behavior of the sequence complexity versus protein sizes in all the families, the family length (L_f) is introduced and is defined as:

$$L_f = \frac{1}{N_s} \sum_{i=1}^{N_s} L_i. \tag{6}$$

Here L_i is the sequence length of ith sequence in a family. Similar to the study by Wood and coworkers [2], the average sequence length (not the average alignment length) is used to represent the average size of all proteins in the family since protein structure is encoded by the whole sequence which is generally longer than the aligned segment of the proteins in the family. Thus L_f represents the average sequence length of a family, and is termed as the family length.

2.6 Global interactions

Sequence sites contribute to the stability of protein structure by two kinds of interactions (or the contact interactions generally). One is the global (long-range) interactions, and the other is the local interactions [18]. Here, the interactions in a protein are identified based on the C_α model of proteins in which each residue is represented by its C_α atom. The interaction is defined for two residues when the distance between their C_α atoms is less than 7.0 Å. Then, the global (or local) interactions are classified when the sequence distance of the two interacted residues is larger than 10 (or less than 10) residues [18]. Here the sequence distance with 10 residues for a contact interaction means that there are 10 residues on the sequence between the two interacting residues. Thus, based on the structures of proteins from the database PDB, the probability P_g of global interactions of any a protein can be obtained. For every family in the FSSP, the related $\langle P_g \rangle$ is the average over all the proteins in the family.

(a) VSQEEVKKWAESLENLINHECGLAAFKAFLK
. . . . . AESLENLINHECGLAAFKAFLK
PTKMRVERWAENFSELIRDPKGRQSFQSFLR
PTKMRVERWAFNFSELIRDPKGRQSFQSFLR
PTKMRVERWAENFSELIRDPKGRQSFQSFLR

(b) RCRNEHPYIYNRPYADDLIFDFRTDVARQAD
HLEESCMYDFWHLDCCPELFYWVTGESLKDE
PGVRIMSMVAGPYVSVQQQREKADNNDKPRA
VDTPLDFGTVYCFLTETMNYANVLTPANSLG
QMGDKTYPGSNQEQRQPHVVLNSVTQGGRPA

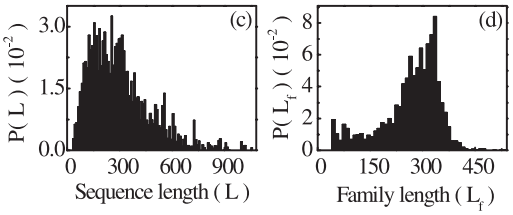

Fig. 1. (a) A segment of five sequences based on structural alignment in a family in the FSSP. Inserts in the alignment are marked as single dots, and grey columns are the conservative sites with the same residues. (b) A segment of five random sequences generated with the ratio of hydrophobic (H) to polar (P) residue as 1:1. (c) Distribution of all sequences in the FSSP over their length L. (d) Distribution of all families in the FSSP over the family length L_f. The bin sizes are 10 residues.

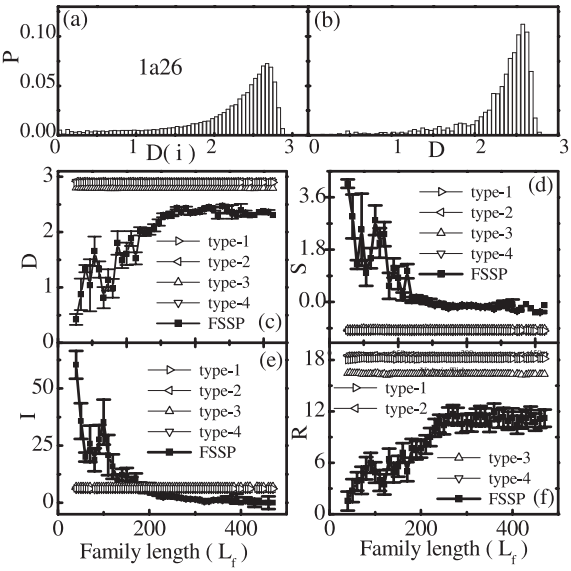

Fig. 2. Distributions of column entropy $D(i)$ for family 1a26 in the FSSP (a). Distribution of sequence entropy D for all families in the FSSP (b). The sequence entropy D (c), sequence similarity S (d), sequence identity I (e) and alphabet size $\mathcal{R}$ of residues for families (f) versus the family length L_f. Each point is averaged over 10 residues. Solid squares represent the results for sequences in the FSSP and open points for the artificial sequences.

3 Results and discussion

Figure 1a shows a segment of structural alignment for the sequences in a certain family in the FSSP database (Version 1.1) [8]. The grey columns are the conserved sites with the same type of residues, and the single dots are the blank insertions during the alignment. Figure 1b shows a segment of random sequences produced by keeping the ratio of hydrophobic residues to polar residues (H:P) following its value in natural sequences, i.e. H:P $\simeq$ 1:1. Although these random sequences have a similar ratio of H:P to natural proteins, they contain very little information on the protein structures. From Figures 1a and 1b, it is seen that the distribution of residues in Figure 1a is rather regular compared to that in Figure 1b although the residues at every column in Figure 1a are not exactly the same. This suggests that the random sequences lack structural information, and are more complex than the natural protein sequences.

Figure 1c shows a Poisson distribution of all 27 194 sequences in 2871 families in the FSSP over their sequence lengths L. It is clearly seen that most sequence lengths contain about 200 residues, which is relevant to the results for all natural protein sequences observed so far [19]. This may result from the natural evolution of proteins under the functional, kinetic and thermodynamic pressure. Proteins with too short sequences may not be able to perform their functions, while proteins with too long sequences may have difficulty in keep the stability of their structures. Figure 1d shows the distribution of families over L_f in each family. Note that the lengths of all sequences in a family are more or less the same. It is also noted that families with less than 20 sequences or with more than 50% mutated sequences by protein engineering are not included because the number of sequences in these families is too small to make a significant statistical study. From Figure 1d, it is seen that most families have values of L_f between 150 and 400 residues.

Figure 2 shows the variation of sequence entropy, sequence similarity, and sequence identity as the value of L_f increases. It is found that the column entropy $D(i)$ has a unimodal distribution for every family (see Fig. 2a), but the positions and the heights of the peaks are slightly different for different families. This implies that the features of the distribution of $D(i)$ values are more or less the same for various families. The distribution of sequence entropy D for all families is similar to that of $D(i)$, but the peak is more sharp (see Fig. 2b). For different families with different lengths L_f, the values of D are different, and show an increasing trend as L_f increases. When $L_f > 250$ the values of D become saturated (see Fig. 2c). Differently, both the sequence similarity S and sequence identity I show a decreasing trend as L_f increases and also become saturated at $L_f = 250$ (see Fig. 2d and 2e). Especially, at saturation, both the values of S and I are basically around zero.

In Figure 2f, the alphabet size R for protein families versus L_f for all families in the FSSP is plotted. It can be seen that there is an increasing trend in R when $L_f < 250$, and the values of R become saturated with $R = 12$ when $L_f > 250$. This implies that when the sequence lengths are

larger than 250 residues, one needs basically 12 kinds of residues to characterize the sequences. This supports the suggestion made by Akanuma and coworkers [10] that the reduction of residue kinds for proteins is size dependent. When the sequence lengths are smaller than 150 residues, one needs about 6 kinds of residues. However, this may relate to proteins with simple structural prototypes such as proteins of all-α or all-β types. The experiment done by Baker and co-workers for the protein SH3 domain using 5 kinds of residues is an example [20].

Do the above mentioned results reveal the complexity or the information encoded in the natural protein sequences? Here, we make a further study on the artificial families which only contain random sequences with different lengths according to the related 2871 families. That is, for a certain family in the FSSP with N_s sequences and average sequence length L_f, a corresponding artificial family is constructed by generating N_s random sequences with sequence length L_f. Four types of such families are produced with different rulers: 1) the ratio of the hydrophobic residues to polar residues (H:P) in the random sequences is taken as 1:1 (see Fig. 1b), 2) the ratio of the hydrophobic residues to polar residues is taken as 2:3, 3) the ratios of the 20 kinds of natural residues follow the statistical values of natural protein sequences [15], 4) the ratios of the 20 kinds of residues are taken uniformly. Here the hydrophobic residues (the H-group) include residues C, M, F, I, L, V, W and Y, and the polar residues (the P-group) includes residues A, G, T, S, N, Q, D, E, H, R, K and P[21,22]. The value of H:P for sequences in type-1 families is similar to that of protein sequences in nature [23], and the value of H:P for sequences in type-2 families is selected to make a comparison with type-1 families. Note that during the generation of type-1 and type-2 families, residues belonging to the H-group or P-group are selected uniformly. The ratios of residues for sequences in type-3 families refer to the situation of natural proteins, while the ratios of residues for sequences in type-4 families are statistically uniform. Note that all these random sequences lack structural information although some sequence information of proteins is included in sequences in type-1 and type-3 families.

For these four types of artificial families, the related factors D, S, I and R are worked out and plotted in Figure 2c to Figure 2f, respectively. It is seen that the values of D, S and I are nearly the same for all four types of families. The values of D, S, and I are about 2.8, –1, and 6 which are close, but not equal, to 3.0, –4, and 0 (the related values in an extreme condition that all the numbers of 20 kinds of residues are set to be absolutely equal). This is because for each type of families, these values of D, S, and I are the statistical average over 2871 random families. The sequence entropy D is obvious larger and the sequence similarity S is smaller than those in the FSSP case, suggesting that the sequence complexity of the sequences in the artificial families is high due to the lack of the structural information. However, the sequence identity I of the artificial families is higher than that of the natural protein families when $L_f > 250$. This does not mean that the sequence complexity of the artificial families is lower than that in natural protein families contrary to what is suggested by the values of D and S. The reason is that the description of the sequence complexity using the factor I is coarse. According to the definition of I, it can be seen that I actually describes the sequence complexity only considering the types of residues. In a different way, the sequence similarity S and sequence entropy D characterize the sequence complexity based on the similarity between the residues and the partition of various residue types on all sites. Thus the factors S and D provide a more accurate measure of the complexity than that of the factor I.

The alphabet sizes R for the artificial families are independent on L_f too, and are higher than those for the FSSP case. The alphabet sizes R are 18 for type-1, type-2 and type-4 families, and is 16 for type-3 families. The same values of R for type-1, type-2 and type-4 families indicate that these sequences which are generated only considering the ratio between the H-type and P-type residues contain so little information of protein sequences that their values of R are the same as those for sequences generated completely randomly. The values of R for type-3 families is lower since the sequences in this type of families contain more sequence information of proteins based on a detailed consideration of the ratio of the 20 residues. Comparing the length dependent behavior for the families in FSSP, the values of D, S, I and R for four types of artificial families are all independent of L_f. This suggests that some length dependent information is encoded in the natural protein sequences in the FSSP database due to the natural optimization in the functions, thermodynamic stability and kinetic accessibility.

What is the origin of the dependence of the family lengths on the alphabet size? Why is there a saturation of the sequence complexity characterized by the sequence entropy, the sequence similarity, and the sequence identity? Since the sequences in artificial families generated without structural information do not show the length dependent behavior of sequence complexity, we seek the answer to the above questions from the topological features of protein structures.

As is known, to realize biological functions natural proteins should satisfy two requirements. One is the kinetic requirement with which the protein folds to its native structure rapidly, the other is the thermodynamic requirement with which the native structure of protein is stable. Previous studies suggested that the kinetic requirement and thermodynamic requirement of proteins are correlated with each other [24]. That is, the kinetic requirement can be satisfied as long as the thermodynamic requirement is well satisfied. For a protein, to satisfy the stability of its native structure, a constraint in residue composition in the sequence is required. This is the structural constraint under which the complexity of natural sequences is smaller than that of random sequences (see Figs. 2c, 2d, 2e, and 2f). However, this structural constraint is not strict since different sequences could fold into a similar native structure. This means that residue mutations on few sites in

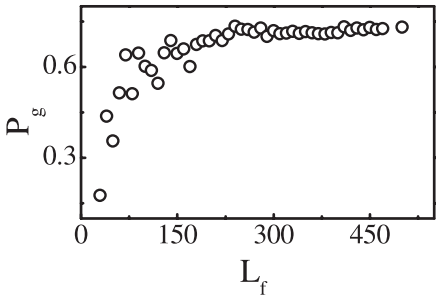

Fig. 3. Observation probability $\langle P_g \rangle$ of global interactions in families versus L_f.

the sequences will still retain the original native structure. But, these mutations are not arbitrary since different sites in the sequence contribute to the stability of the native structure differently. Some mutations on those sites with low specificity keep the stability of the native structure, thus the native structure could be retained. While mutations on sites with high specificity will change the native structure. Thus the greater the number of sites with low specificity in a family, the larger the sequence complexity is, and the greater the number of sites with high specificity in a family, the smaller the sequence complexity is.

In another aspect, different sites in a sequence play different roles in stabilizing the protein structure by contact interactions between residues. Two kinds of contact interactions, divided into the global interactions and the local interactions, are included. The global interactions stabilize the structure globally, and make the sites in the sequence contribute to the structural stability cooperatively rather than specifically, while the local interactions behave in the opposite manner. That is to say, the global interactions weaken the individual specificity of the related sites to the stability for the native structure, and make the residues on these sites replaceable. On the other hand, the residues on the sites with local interactions are irreplaceable since the effect of the individual specificity becomes significant. Therefore, a high probability P_g of the global interactions induces a low specificity of the sites in the structure, while a low P_g (i.e. high probability of local interactions) induces a high specificity of sites. These indicate that a family with a high value of $\langle P_g \rangle$ has a high sequence complexity, while a family with a low value of $\langle P_g \rangle$ has a low sequence complexity. In Figure 3, the variation of $\langle P_g \rangle$ for various families with different values of L_f is plotted. It is seen that there is first an increasing trend in $\langle P_g \rangle$ when $L_f < 250$, and then a saturation when $L_f > 250$. Clearly, this size dependent behavior of $\langle P_g \rangle$ results in the size dependent behavior of sequence complexity in various families as shown in Figures 2c, 2d, 2e, and 2f, suggesting that the size dependent features of protein sequences comes from the size dependent features of protein structures.

Finally, the above mentioned sequence complexity could also be detected based on various residue groupings [22]. Suppose that 20 kinds of natural occurring residues are grouped into five groups [21] as $X1 = \{C, F, Y, W\}$, $X2 = \{M, L, I, V\}$, $X3 = \{G\}$, $X4 = \{P, A, T, S\}$ and $X5 = \{N, H, Q, E, D, R, K\}$. Then the residues in the sequences in a family are substituted by five letters $\{X1, X2, X3, X4, X5\}$ and the related sequence identity I_5 can be calculated. A plot of I_5 versus I is shown in Figure 4a. Here I relates to the case of 20 kinds of residues. The more the points deviate from the diagonal, the less is the similarity of the sequences due to the substitution. The correlation coefficient C and slope K of a linear fitting are used to depict the nature of the complexity reduction by residue substitution. The closer the values of C and K to 1, the less the effect of the reduction for the sequences in the families is. For different groupings with N groups, the sequence identity is I_N, so the related C and K, can be obtained. To compare the results with different L_f, all families are divided into three subsets according to their lengths: 1) $L_f < 250$, 2) $L_f > 250$, 3) all values of L_f. In Figures 4b and 4c, the related values of K and C, normalized between 0 to 1, versus the number of groups of residues are plotted. The groupings of residues in Figures 4b and 4c is taken from the simplified table of residues made by Li et al. [21]. As a common feature, it is found that both K and C show a plateau for each subset as marked by the arrows. These plateaus indicate a minimal number of residue types or a minimal alphabets size for reduction. When the number of residue types is smaller than such a minimal number, the reduction is not reasonable and the information in the sequences is lost. For example, for protein sequences with $L_f > 250$, a large alphabet sizes with 12 kinds of residues is needed to build their structures. Besides, these minimal alphabet sizes are quantitatively compatible with previous results obtained from the values of R (see Fig. 2f).

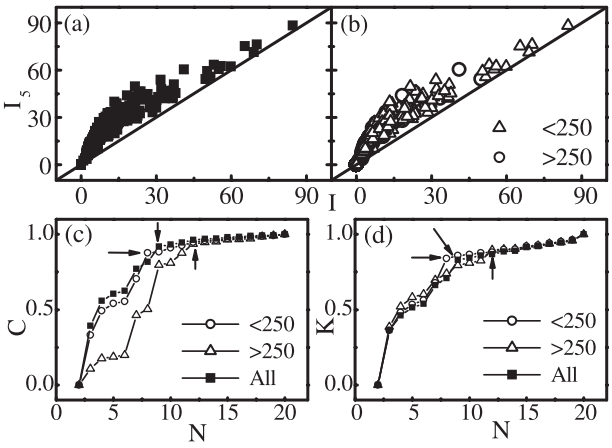

Fig. 4. Sequence identity I_5 using 5 groups of residues versus sequence identity I using 20 kinds of residues for families with all values of L_f (a), with $L_f < 250$ and $L_f > 250$ (b). The correlation coefficient C vs. N (c), and the slope of linear fitting K vs. N with N groups of residues (d). The arrows indicate the plateaus.

4 Conclusion

Our study on the FSSP database shows that the complexity of sequences in various families depends on the family length L_f. When $L_f < 250$, the complexity shows

an increasing trend as L_f increases. When $L_f > 250$, all the features become saturated and an alphabet size with 12 kinds of residues may be sufficient to characterize the complexity for proteins.

This work was supported by the National Natural Science Foundation of China (Nos. 90403120, 10474041, 10204013 and 10021001), and the Nonlinear Project (973) of the NSM.

References

1. I.V. Grigoriev, S.H. Kim, Proc. Natl. Acad. Sci. USA **96**, 14318 (1999)
2. T.C. Wood, W.R. Pearson, J. Mol. Biol. **291**, 977 (1999)
3. O. Lichtarge, Nat. Struct. Biol. **8**, 918 (2001)
4. G. Yona, M. Levitt, J. Mol. Biol. **315**, 1257 (2002)
5. N.V. Dokholyan, B. Shakhnovich, E.I. Shakhnovich, Proc. Natl. Acad. Sci. USA **99**, 14132 (2002)
6. M.L. Sierk, W.R. Pearson, Protein Sci. **13**, 773 (2004)
7. H.M. Berman, J. Westbrook, Z.K. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucl. Acids. Res. **28**, 235 (2000)
8. L. Holm, C. Sander. Nucl. Acids. Res., **26**, 316 (1997)
9. A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, J. Mol. Biol. **247**, 536 (1995)
10. S. Akanuma, T. Kigawa, S. Yokoyama, Proc. Natl. Acad. Sci. USA **99**, 13549 (2002)
11. S.P. Shenkin, B. Erman, L.D. Mastrandrea, Proteins: Struct. Funct. Genet. **11**, 297 (1991)
12. S.M. Larson, J.L. England, J.R. Desjarlias, V.S. Pande, Protein Sci. **11**, 2804 (2002)
13. P. Koehl, M. Levitt, J. Mol. Biol. **323**, 551 (2002)
14. O. Teodorescu, T. Galor, J. Pillardy, R. Elber, Proteins: Struct. Funct. Genet. **54**, 41 (2004)
15. S. Henikoff, J.G. Henikoff, Proc. Natl. Acad. Sci. USA **89**, 10915 (1992)
16. C. Chothia, L.M. Lesk, EMBO J. **5**, 823 (1986)
17. H.H. Gan, R.A. Perlow, R. Roy, J. Ko, M. Wu, J. Huang, A. Yan, S.X. Nicoletta, J. Vafai, D. Sun, L.H. Wang, J.E. Noach, S. Pasquali, T. Schlick, Biophys J. **83**, 2781 (2002)
18. L.H. Greene, V.A. Higman, J. Mol. Biol. **334**, 781 (2003)
19. J.Z. Zhang, Trends Genet. **16**, 107 (2000)
20. D.S. Riddle, J.V. Santiago, S.T. Bray, N. Doshi, V.P. Grantcharova, Q. Yi, D. Baker, Nat. Struct. Biol. **4**, 805 (1997)
21. T.P. Li, K. Fan, J. Wang, W. Wang, Protein Eng. **16**, 323 (2003)
22. J. Wang and W. Wang, Nat. Struct. Biol. **6**, 1033 (1999)
23. Fiebig K.M. Thomas P.D. Chan H.S. Shakhnovich E.I. Yue, K., K.A. Dill, Proc. Natl. Acad. Sci. USA **92**, 325 (1995)
24. E.I. Shakhnovich, A.M. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993)