Scaling Behaviour of Conserved Sites in Protein Families *

LI Jie(李捷), WANG Jun(王骏), WANG Wei(王炜)**

National Laboratory of Solid State Microstructure, Institute of Biophysics, and Department of Physics, Nanjing University, Nanjing 210093

(Received 25 January 2005)

Base on the database of families of structurally similar proteins, a statistical study is made on the scaling behaviour of occupying probabilities of conserved sites (P_c) in various protein families. A power-law decrease of P_c with the increasing protein-chain length L_f is found. This is related to the power-law scaling behaviour of the occurring probabilities of local contact interactions ($P_{\rm local}$) between residues. In addition, applying residue grouping, we find the same scaling behaviour when the number of residue types is more than 12, indicating that 12 residue types are enough to present the complexity of proteins.

PACS: 87.10.+e, 87.15.Cc

It is well-known that a structural prototype of protein or protein fold can be shared by many protein sequences although every sequence can fold into a native structure uniquely. It has been argued that there are only about several thousands different protein folds which cover the whole proteins in nature.[1−8] This asymmetrical mapping between sequences and structures actually results in the complexity in proteins. To understand the relationship between different proteins, protein sequences with similar structures are classified into the same family. That is, based on the protein database bank (PDB),[9] some secondary databases of various protein families, such as the database of the structural classification of proteins (SCOP)[10] and the database of families of structurally similar proteins (FSSP) resulted from the structural alignment,[11] are constructed.

In the FSSP database, it is known that the numbers of sequences in different families are different, and in every family the types of residues at the same sites (the aligned columns) of different sequences may be different. The sites with the same type (or almost the same types) of residues in a family are defined as conserved sites, and the others as non-conserved sites. The conserved sites are sensitive to the mutations in the evolution of proteins, and at the same time, are important and irreplaceable for structural stability and biological functions. Nevertheless, the non-conserved sites are insensitive to the mutations, and may also be important, but are replaceable. The occupying probability of conserved sites in different families are different, and may depend on the protein chain lengths. The coding way of proteins with high occupying probability of conserved sites is usually related to the local model,(centralized model)[12,13] in which the fold specificity is coded by just the few critical residues, while the coding way of the proteins with low occupying probabilities of conserved sites is basically related to the global model,(distributed model)[12,13] in which the fold is formed by the interactions involving the entire sequence, and typical residues in the substitution range over the sequence have little impact on the fold. Therefore, to study the occupying probabilities of conserved sites and non-conserved sites in protein families is important for understanding the coding mechanism of proteins. In this Letter, based on the FSSP database we study the variation of occupying probabilities of conserved sites when the protein chain length is changed. A scaling behaviour is obtained and the physical origin is discussed.

The total number of families in the FSSP database (version 1.1[11]) is 2859. However, considering that too few sequences in a family may induce a bias to the identification of conserved sites, families with number of sequences less than 10 are excluded. Thus the total number of useable families is reduced to 2428. For each family, a statistical study on the types of residues at all sites of the well aligned sequences is carried out. We note that all insertions at various sites in the aligned sequences are ignored.

As the first case, the conserved site is strictly defined as the aligned site which is occupied by the same type of residues. By counting the number of conserved sites in each protein family, the occupying probabilities of conserved sites (P_c) for the various families can be obtained and then scaled with the protein chain length L_f averaged over all the sequences in each family. Here L_f represents the length features of all the sequences in a family, and can be termed as the family length similar to the definition by Wood and coworkers.[13]

Figure 1 shows the variation of P_c versus L_f. It

* Supported by the National Natural Science Foundation of China under Grant Nos 90103031, 10074030, and 10204013, and the Nonlinear Project of the National Key Basic Research Special Foundation (NKBRSF) of China.

** To whom correspondence should be addressed. Email: wangwei@nju.edu.cn

is seen that $\ln(P_c)$ decreases linearly as $\ln(L_f)$ increases. The correlation coefficient (C) is 0.76 and the slope (K) is -1.64. This suggests a power-law scaling $P_c \sim {L_f}^K$, which is consistent with the power-law scaling behaviour of the folding nucleus with sequence lengths[14] since the nucleus are related to the conserved sites.[15] It is noticed that the exponent in the power-law of the folding nucleus is -1.[14] Here our exponent for the power-law behaviour of conserved sites is $K = -1.64$. The difference in both the exponents may be due to two reasons. One is the usage of the strict definition for our conserved sites, and the other is that our conserved sites include not only kinetically important sites (i.e., the nucleus) but also functionally or thermodynamically important ones.

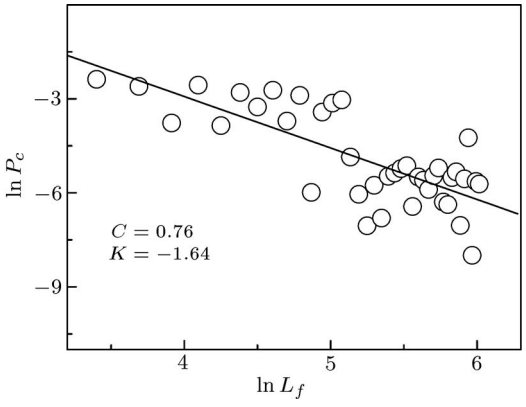

Fig. 1. The occupying probability of conserved sites P_c versus protein chain length L_f. Each point is averaged over 10 bin. The linear coefficient C is 0.76, and the slope K is -1.64.

What is the physical origin of the power-law behaviour of P_c with the increase of protein chain lengths, and what is the physical indication of the power-law behaviour to the coding of structures from sequences? Let us consider the detailed features of the three-dimensional native structures of proteins and the interactions between the residues. Figure 2 shows a certain structure of protein, saying ribosomal S6. The representative local interactions (with solid lines) and nonlocal interactions (with dashed lines) are clearly shown. Here the local interactions between residues i and j are defined for residue pairs with $i - j < 10$,[17] and the nonlocal interaction for residue pairs with $i - j \geq 10$[18] when the distance between the two C_α atoms of residues i and j is shorter than 7.0 Å. The local interactions prefer to stabilizing the structure locally, and the participated residues usually contribute to the stability of the whole protein structure individually, while the nonlocal interactions prefer to stabilizing the structure globally and cooperatively. That is to say, the local interactions describe the importance of residues individually in the sequence, and make the individual residues irreplaceable in the proteins, while the nonlocal interactions introduce the cooperativity of residues, and make the individual residues less important and replaceable in the proteins. Therefore, the scaling behaviour of the conserved sites P_c with the protein chain lengths may relate to the scaling behaviour of occurrence probabilities of local and nonlocal interactions.

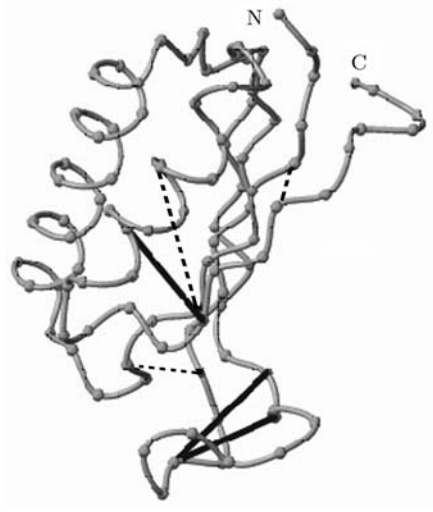

Fig. 2. The structure of protein ribosomal S6 (pdb code: 1ris). C_α atoms for the 97 residues are shown by the spheres. The solid straight lines represent the local interactions between residues, and the dashed straight lines denote the non-local interactions.

Figure 3 shows the average occurring probability of local interactions ($\langle P_{\rm local}\rangle$) of sequences in various families scaled with related L_f. Here $P_{\rm local}$ of each sequence is calculated based on its corresponding structure from the PDB. It is found that similar to the scaling behaviour of P_c, $\langle P_{\rm local}\rangle$ also decreases in a power-law way as L_f increases. To show the correlation between P_c and $\langle P_{\rm local}\rangle$ clearly, in the inset of Fig. 3, P_c versus $\langle P_{\rm local}\rangle$ is plotted. A linear relationship can be seen and a correlation coefficient $C = 0.70$ is obtained, indicating that the power-law decrease of P_c is well related to the power-law decrease of $\langle P_{\rm local}\rangle$ as L_f increases. Since the occurring probability of nonlocal interactions $P_{\rm nonlocal} = 1 - P_{\rm local}$, the decrease of $\langle P_{\rm local}\rangle$ means the increase of $\langle P_{\rm nonlocal}\rangle$ as L_f increases. The increase of $\langle P_{\rm nonlocal}\rangle$, which implies the increase of the cooperativity between residues, enables the effect of every individual residue becoming weak, thus makes more residues replaceable. Therefore, the values of P_c decrease as L_f increases. Note that the correlation coefficient in Fig. 3 is higher than that in Fig. 1. This may be due to the fact that the conserved sites include not only the structurally important site,[19] but also the functionally important sites.

The variation of conserved sites with the increase of protein chain lengths may also be related to pro-

tein coding. As discussed in a previous work,[12,13] there are two models for protein coding. One is the "global model", in which the folding features are encoded by interactions involving the entire sequence, the other is the "local model", in which the folding features are encoded in just few critical residues. The global model may be related to the sequences with low P_c which is close to zero, while the local model to the sequences with high P_c which is more than 10%–20%.[12,13] According to the scaling behaviour of P_c, it could be argued that for proteins with large sizes or longer chains, the coding mechanism presumably prefers to the global model. This interprets that the number of critical residues does not increase too much for big proteins with respect to small proteins. However, it should be noted that the conserved sites in a protein are divided into structure-dependent and functional-dependent ones. Although it is difficult to distinguish, here we believe that most of them are structure-dependent.

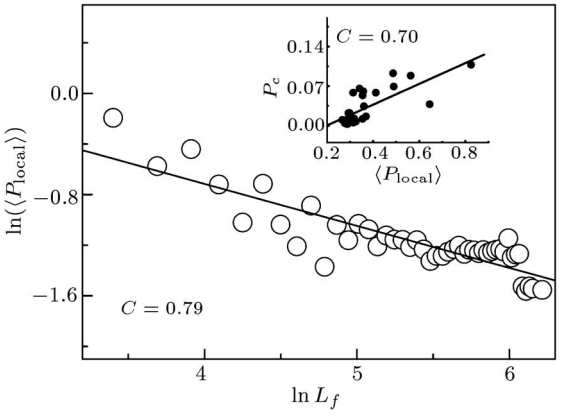

Fig. 3. The average occurring probability of local contact interactions $\langle P_{\rm local}\rangle$ scaled with the protein chain lengths L_f. Each point is averaged over 10 bin. The correlation coefficient C is 0.79. Inset: P_c versus $\langle P_{\rm local}\rangle$.

Now, let us discuss the second case based on a different definition of the conserved sites. It is well-known that the residues with similar physicochemical features can be clustered into groups in the study on the simplification of protein complexity.[20–24] Accordingly, the conserved sites can also be defined as the sites occupied by residues in the same groups. To study whether the scaling behaviour depends on the identification of conserved sites based on the similarity of 20 kinds of residues, the simplified table of residues by Li *et al.*[25] is applied. Based on the FSSP database, the scaling of P_c versus L_f for different numbers of simplified groups N are obtained, and the related correlation coefficient C and slope K are illustrated in Fig. 4. From Fig. 4(a), it can be seen that the values of C are basically around 0.75 except the case of $N = 2$ in which $C = 0.66$. From Fig. 4(b), it is found that the slope K of the scaling of P_c versus L_f decreases as the grouping number N increases, and becomes saturated when $N \geq 12$. That is to say, when the number of residue types is larger than 12, the power-law scaling behaviour of P_c versus L_f reserves. Thus, the identification of conserved sites regarding the similarity of residues takes no effect on the power-law relationship between P_c and L_f, especially when $N > 12$. This suggests that the power-law scaling behaviour of P_c is the inherent characteristic of conserved sites, and does not depend on the special definition of conserved sites as long as the number of residue types is enough to present the complexity of proteins. Moreover, in view of the complexity simplification for proteins, the occurring of the saturation suggests that an alphabet with 12 types of residues is a reasonable grouping of residues. This is consistent with the conclusion of previous studies on protein complexity that $10 \sim 13$ types of residues are enough for the encoding of proteins.[26,27]

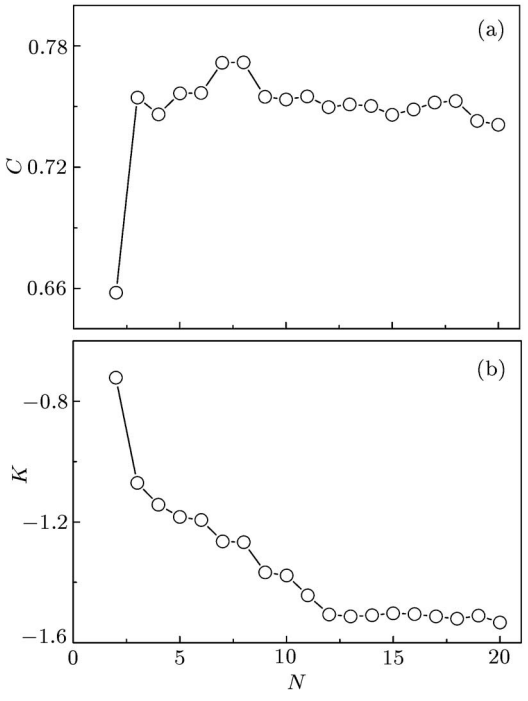

Fig. 4. The correlation coefficient C (a), and the exponent K (b) versus the number of residue types for the probability of conserved sites $P_c \sim L_f^K$.

In conclusion, we have studied the scaling behaviour of the conserved sites with the protein chain lengths in various protein families. It is found that the occupying probability of the conserved sites decreases in a power-law way as the protein chain length increases. Such behaviour is related to the power-law decreasing of the occurring probability of local contact interactions between residues in proteins. The ratio of number of local contacts to the total number of contacts does not increase significantly when the sizes of proteins are large due to the compact feature

of proteins. Furthermore, by applying residue grouping, it is found that the power-law scaling behaviour of conserved sites with the protein chain lengths is an inherent characteristic of conserved sites, and is independent of the definition of conserved sites when the number of residue types is more than 12. An alphabet with 12 kinds of residues are enough to present the complexity of proteins.

References

[1] Orengo C A, Jones D T and Thornton J M 1994 *Nature* **372** 631
[2] Liu X S, Fan K and Wang W 2004 *Proteins* **54** 491
[3] Wang Z X 1996 *Proteins* **26** 186
[4] Zhang C T 1997 *Protein Engin.* **10** 757
[5] Zhang C and DeLisi C 1998 *J. Mol. Biol.* **284** 1301
[6] Wang Z X 1998 *Protein Engin.* **11** 621
[7] Govindarajan S, Recabarren R and Goldstein R 1999 *J. Mol. Biol.* **35** 408
[8] Zhang C and DeLisi C 2001 *Cell. Mol. Life. Sci.* **58** 72
[9] Berman H M et al 2000 *Nucl. Acids. Res.* **28** 235
[10] Murzin A G et al 1995 *J. Mol. Biol.* **247** 536
[11] Holm L and Sander C 1997 *Nucl. Acids. Res.* **26** 316
[12] Lattman E E and Rose G D 1993 *Proc. Natl. Acad. Sci. U.S.A.* **90** 439
[13] Wood T C and Pearson W R 1999 *J. Mol. Biol.* **291** 977
[14] Gutin A M, Abkevich V I and Shakhnovich E I 1996 *Phys. Rev. Lett.* **77** 5433
[15] Mirny L and Shakhnovich E I 2001 *J. Mol. Biol.* **308** 123
[16] Koradi R, Billeter M and Wuthrich K 1996 *J. Mol. Graphics* **14** 51
[17] Greene L H and Higman V A 2003 *J. Mol. Biol.* **334** 781
[18] Weikl T R and Dill K A 2003 *J. Mol. Biol.* **329** 585
[19] QIN M, Wang J and WANG W 2003 *Chin. Phys. Lett.* **20** 1883
[20] Wolynes P G 1997 *Nat. Struct. Biol.* **4** 871
[21] Chan H S, and Dill K A 1998 *Proteins* **30** 2
[22] Wang J and Wang W 1999 *Nat. Struct. Biol.* **6** 1033
[23] Tiana G, Broglia R A and Shakhnovich E I 2000 *Proteins: Struct. Funct. Genet.* **39** 244
[24] Wang J and Wang W 2002 *Phys. Rev.* E **65** 041911
[25] Li T P, Fan K, Wang J and Wang W 2003 *Protein Engin.* **16** 323
[26] Giulio D M and Medugno M 1999 *J. Mol. Evol.* **49** 1
[27] Akanuma S, Kigawa T and Yokoyama S 2002 *Proc. Natl. Acad. Sci. U.S.A.* **99** 13549