Vol 18 No 4, April 2009
1674-1056/2009/18(04)/1684-07
Chinese Physics B
© 2009 Chin. Phys. Soc.
and IOP Publishing Ltd

Structural statistical properties of knotted proteins*

Wang Xiang-Hong(王向红)[a], Shen Yu(沈 瑜)[a], and Zhang Lin-Xi(章林溪)[b]†

[a] *Department of Physics, Zhejiang University, Hangzhou 310027, China*

[b] *Department of Physics, Wenzhou University, Wenzhou 325027, China*

(Received 17 April 2008; revised manuscript received 13 July 2008)

The composition and residue–residue interactions of knotted proteins, compared with those of other proteins, can provide considerable insight into the driver of the knots in proteins. In this paper, we calculate the probabilities of 20 amino acids in 273 knotted entries from the Protein Data Bank (PDB). The collection of 273 entries contains all knotted structures in the PDB, and it is not a subset. With an appropriate value of R_c, the numbers of all residue–residue contacts are counted in all 273 knotted structures. To make an accurate comparison, we count up to 9000 other entries from the PDB as well, and these entries spread over all sorts. In knotted structures, Leu occupies a maximal proportion of 9.62% among all 20 amino acids, and Leu, Phe, Trp, Gly, His, Gln, Asp, Lys and Pro may all play a more important role. Also, we analyse the effects of amino acid residues on the long-range contacts. We observe a larger average number of long-range contacts in the knotted structures than that in other ones, implying their important role in achieving the knots. Accordingly, the average number of short-range contacts becomes small when the structure becomes knotted because it depends mainly on the short-haul sequence of amino acids to form the short-range contact. In addition, the shape distribution of knotted proteins and the contrast with the other proteins are also presented. A comparison shows that the knots may make structures more globular because the average shape factor is 0.059 for the knotted proteins, which is only about 1/3 of the average shape factor for the other proteins.

Keywords: amino acid, knotted proteins, short-range and long-range contacts
PACC: 8120S, 8200

1. Introduction

Although knots are abundant and complex in globular homopolymers,[1−3] they are rare and simple in proteins.[4−8] Liang and Mislow[9] reported on their finding early in 1994, the result of a survey of structures deposited in the Brookhaven Protein Data Bank, in which several metalloproteins contain trefoil knots as well as catenated substructures. However, before then, it had not been recognized that those protein structures might contain such non-trivial topological features. Indeed, the question whether the polypeptide chains of native proteins were ever tied into knots or links (catenated loops) has been repeatedly addressed over the past three decades.[4,10−21] Sixteen methyltransferases in bacteria and viruses can be combined into the α/β knot superfamily,[19] and several isozymes of carbonic anhydrase (I, II, IV, V) are known to be knotted. Apart from these two folds only a few insular knots have been investigated[5,6,20,21] Virnau *et al* [22] provided the first comprehensive review of knots in proteins, and they examined knots in homologous proteins. Their analysis reveals several new knots, all in enzymes. In particular, they discovered that the most complicated knot found to date (5_2) was in human ubiquitin hydrolase, and suggested that its entangled topology protected it from being pulled into the proteasome. Besides, they also noticed that knots usually resided among structural homologues.

Some properties and behaviours for unknotted proteins have been well studied.[23−25] A general consensus on the existence of knots and links in proteins has been reached, however, intensive studies are still lacking, and there have been few reports so far to show the correlation between the inner ingredient and the knotted structure. One part of our present study is devoted to a comparison in inner ingredient such as amino acid ratio and contact probability between

*Project supported by the National Natural Science Foundation of China (Grant Nos 20574052 and 20774066), the Program for New Century Excellent Talents in University, China (Grant No NCET-05-0538) and the Natural Science Foundation of Zhejiang Province, China (Grant No R404047).
†Corresponding author. E-mail: Lxzhang@zju.edu.cn

knotted proteins and unknotted proteins based on the Protein Data Bank (PDB). Proteins are heterogeneous chain molecules composed of sequences of amino acids. There are 20 different amino acids in protein molecules, and their effects may be different in knot. In the protein knots, the amino acid residue–residue contacts play an important role. The knots and entanglements in proteins may influence some function or behaviour such as protein folding.

Few proteins with knotted native state conformations have been identified so far, therefore, it is tempting to analyse the composition and the structure (such as residue–residue contacts, size and shape) of knotted structures and their difference from those of the other proteins. The statement that knots are rare in native proteins means that they emerge far less frequently than randomly. In the present paper we count all 273 knotted structures in the 32,853 entries that contain proteins (Table 1) in the PDB (http://www.pdb.org)[26] in the version issued on January 3, 2006. Such a comparison has not been made explicitly and systematically in previous studies yet.

Table 1. List of knotted Protein Data Bank entries.

12ca	1a42	2abe	1aiv	1ajc	1am6	1aov	1avn	2ax2	1azm	2b63	1b90	1bcd	5bca	1bic
1bn1	1bn3	1bn4	1bnm	1bnn	1bnq	1bnt	1bnu	1bnv	1bnw	1bv3	1bzm	1c5m	1ca2	1ca3
1cah	1cai	1caj	1cak	1cal	1cam	1can	1cao	1cay	1caz	2ca2	2cab	3ca2	4ca2	4cac
5ca2	5cac	6ca2	7ca2	8ca2	9ca2	2cba	2cbb	2cbc	2cbd	2cbe	1ccs	1cct	1ccu	1cil
1cim	1cin	1cnb	1cnc	1cng	1cnh	1cni	1cnj	1cnk	1cnw	1cnx	1cny	1cqx	1cra	1crm
1cva	1cvb	1cvc	1cvd	1cve	1cvf	1cvh	2cx8	1czm	1dca	1dcb	1dmx	1dmy	1dqu	1eou
1f2w	1flj	1fnm	1fql	1fqm	1fqn	1fqr	1fr4	1fr7	1fsn	1fsq	1fsr	1fug	1g0e	1g0f
1g1d	1g3z	1g45	1g46	1g48	1g4j	1g4o	1g52	1g53	1g54	1g6v	1gz0	1h4n	2h4n	1h9n
1h9q	1hca	1hcb	1hea	1heb	1hec	1hed	1hug	1huh	1hva	1hys	1i50	1i6h	1i8z	1i90
1i91	1i9l	1i9m	1i9n	1i9o	1i9p	1i9q	1if4	1if5	1if6	1if7	1if8	1if9	1ipa	1itc
1j0y	1j0z	1j10	1j11	1j12	1j2o	1j85	1j9w	1jcz	1jd0	1js1	1jv0	1jzu	1k3r	1keq
1kop	1koq	1kwq	1kwr	1lg5	1lg6	1lgd	1lug	1lzv	1moo	1mua	1mxi	1nik	1ns5	1nt9
1nxz	1o6d	1okl	1okm	1okn	1oq5	1osn	1ovt	1p35	1p73	1p75	1p7c	1p7l	1p9p	1pmd
1poq	1qmg	1r9s	1r9t	1ray	1raz	1rg9	1rj5	1rj6	1rkn	1rza	1rzb	1rzc	1rzd	1rze
1s1h	1sfo	1t9n	1tb0	1tbt	1te3	1teq	1teu	1tg3	1tg9	1th9	1thk	1to0	1ttm	1uaj
1uak	1ual	1uam	1uga	1ugb	1ugc	1ugd	1uge	1ugf	1ugg	1urt	1v2x	1v6z	1v9e	1v9i
1vh0	1vhy	1wcm	1x7o	1x7p	1xd3	1xeg	1xev	1xpz	1xq0	1y1w	1y77	1y7w	1y8a	1yda
1ydb	1ydc	1ydd	1yh0	1yh1	1yo0	1yo1	1yo2	1yrl	1yve	1z93	1z97	1ze8	1zjr	1znc
1zsa	1zsb	1zsc												

2. Method of calculation

We study 273 knotted protein structures. The cognition of these proteins is derived from Ref.[22] (Table 1). The coordinates of all the knotted protein structures are obtained from the PDB.

Each residue in a protein molecule is represented by the centre of its side chain atom positions, and the position of C^{α} atom is used for a glycine residue. A contact is defined in two residues whose centres are close, i.e. the separation between their centres is less than R_c. This kind of simple method to evaluate the number of residue–residue contacts in proteins has often been used[27−31] Using the C^{α} coordinates and assuming a sphere of radius R_c to be fixed around each residue, the composition of surrounding residues can be calculated. A limiting value of R_c for contacts is chosen to be 0.65, i.e. $R_c = 0.65\,\mathrm{nm}$, and in the previous papers, $R_c = 0.65\,\mathrm{nm}$ is also chosen in estimation of effective inter-residue contact energies by Miyazawa and Jernigan[27,28] For a given residue, the composition of surrounding residues is analysed in terms of the location at the sequence level. A contact existing in two residues between which there are more than 4 residues along the sequence is treated as a long-range contact, and the other contacts are regarded as short-range contacts[32−34] which are the same as the long-range and the short-range interactions in the rotational-isomeric-state model.

A contact consists of two residues A and B. Sometime, residues A and B may be the same. If residues A and B have a high tendency to appear in the proteins, they maybe have a large probability of forming contacts. Here all the 20 amino acid residues to

form contacts are preferentially investigated. The average probability $P_{\mathrm{A-B}}$ of residue–residue contacts for residues A and B is defined as[35−37]

$$P_{\mathrm{A-B}} = \frac{N_{\mathrm{A-B}}}{\sqrt{N_{\mathrm{A}} N_{\mathrm{B}}}}, \tag{1}$$

where $N_{\mathrm{A-B}}$ represents the number of the contacts between residues A and B, and N_{A} and N_{B} are the total numbers of the residues of types A and B in the proteins, respectively.

Average number of contacts per residue indicates the ability of this amino acid residue to form contacts. Here we define the average number of short-range contacts per residue, C_S, and the average number of long-range contacts per residue, C_L, as

$$C_{\alpha,\eta} = \frac{\sum\limits_{\beta=\mathrm{Ala,Asp,Cys,G\,ln,\cdots,Tyr}} N_{\alpha-\beta,\eta}}{N_\alpha}$$
$$(\eta = S,\ \mathrm{or},\ L;\quad \alpha = \mathrm{Ala, Asp}, \cdots, \mathrm{Tyr}). \tag{2}$$

That residue A has a large value of C_L means that residue A has a high tendency to form long-range contacts. These calculations can provide some information on the mechanism of the protein knotting, and also help us study what plays an important role in protein folding.

In order to investigate the influence of a knot on the shape of a protein chain, we consider the radius of gyration tensor S, which is defined as

$$S = \frac{1}{N+1}\sum_{i=0}^{N} S_i S_i^T = \begin{pmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{pmatrix}, \tag{3}$$

where $S_i = \mathrm{col}(x_i,\ y_i,\ z_i)$ is the position of monomer i in a frame of reference with its origin being at the centre of a chain. The tensor S can be diagonalized to form a diagonal matrix with three eigenvalues L_1^2, L_2^2 and L_3^2 ($L_1^2 \le L_2^2 \le L_3^2$). Solc and Stockmayer first used the ratio of these parameters $\langle L_1^2\rangle : \langle L_2^2\rangle : \langle L_3^2\rangle$ to measure the shape of a flexible polymer chain in Refs.[38, 39], and they estimated the ratio to be 1:2.7:11.7 based on a random walk of 100 bonds on a simple cubic lattice by using the Monte Carlo (MC) technique. According to three eigenvalues from Eq.(3), another valuable parameter[40,41] for the shape of chains may be obtained by combining the reduced components as a single quantity, i.e.

$$\delta^* = 1 - 3\frac{L_1^2 L_2^2 + L_2^2 L_3^2 + L_1^2 L_3^2}{(L_1^2 + L_2^2 + L_3^2)^2}, \tag{4}$$

where δ^* varies between 0 (sphere) and 1 (rod).

3. Results and discussion

We count all the 20 amino acids and there exist aggregately 83057 residues in the 273 knotted proteins. The probability distribution of the amino acids is shown in Fig.1. To obtain the differences in composition between knotted proteins and other ones, we randomly count 9000 other entries from the PDB, and these entries spread all over various sorts. The results are also shown in Fig.1. The comparison indicates that in knotted structures, Leu, Phe, Trp, Gly, His, Gln, Asp, Lys and Pro may play a more important role. On the other hand, for the knotted proteins, Leu occupies a maximal composition of 9.62% among all the 20 amino acids, and the composition of the Cys amino acid is 0.84%, which is a minimal value. The maximal composition in the knotted proteins is larger than that in other proteins, and the minimal composition in the knotted proteins is smaller than that in other proteins. Those results can help us know the distribution of the amino acids in the knotted proteins clearly.

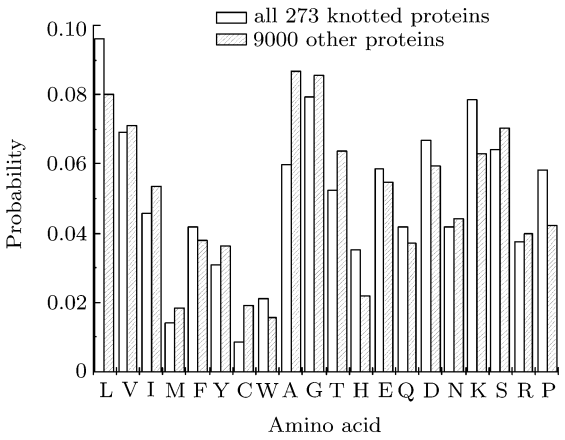


Fig.1. Comparison of amino acid composition between knotted proteins (all 273 entries in the PDB, and not just a subset) and 9000 other proteins in the PDB.

All residue–residue contacts in 273 knotted proteins are counted here. However, it is difficult to judge whose residue is easy to form contacts only based on the total number of the contacts because different residues have different compositions. Therefore it is important to take into account the probability distribution of the amino acid residues and the total number of contacts simultaneously. The average number of long-range (short-range) contacts per residue $C_L(C_S)$ and the average probability of long-range (short-range) residue–residue contacts for residues A

and B $P_{\mathrm{A-B},L}(P_{\mathrm{A-B},S})$ can account for the ability of residue to form contacts.

A prominent variance between knotted structures and unknotted structures lies in the average number of short-range contacts per residue C_S and the average number of long-range contacts per residue C_L. Taking $R_c = 0.65\,\mathrm{nm}$ for example, the results of C_S are given in Fig.2(a), and C_L in Fig.2(b). Here C_S and C_L are calculated according to expression (2). The value of C_S for different all amino acid residues in knotted proteins is remarkably less than that in unknotted proteins. The average value of C_S in knotted proteins is 2.28, which is 19% less than that in unknotted proteins. The maximum of C_S is 2.675 (Ala), and the minimum is 1.647 (Pro) in knotted proteins, while the maximum of C_S is 3.354 (Ala), and the minimum is 2.019 (Pro) in unknotted proteins. Although the values of maximum and minimum in knotted proteins are different from those in unknotted proteins, the corresponding residues are the same, i.e. Ala and Pro, respectively. In Fig 2(b), the value of C_L in knotted proteins

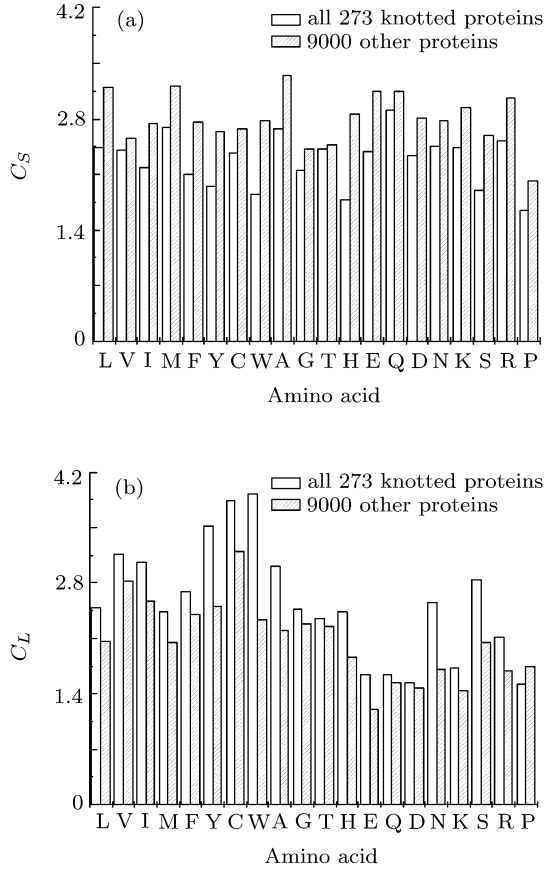


Fig.2. Average numbers of short-range contacts C_S (a) and long-range contacts C_L (b) per residue for the knotted proteins and 9000 other proteins, where $R_C = 0.65\,\mathrm{nm}$.

is larger than that in unknotted proteins except Pro, which indicates that the long-range interaction becomes furthermore significant in order to keep the knotted structure The average value of C_L for all residues in knotted proteins is 2.55, but it becomes 2.07 in unknotted proteins. The residue with the largest value of C_L is Trp, and the residue with the smallest value of C_L is Asp in knotted proteins. However, the residues with the largest and smallest values of C_L in unknotted proteins are different, and they are Cys and Glu. If there exists no long-range contact, the globular or knotted protein structure cannot form. Of course, the attractive interactions between atoms lead to the formation of long-rang residue–residue contacts. Pro has a lowest long-range contact in knotted proteins due to the fact that it is a helix breaker in globular proteins[42] It depends mainly on the short-haul sequence of amino acids in proteins to form the short-range contact, therefore, the short-range contacts consequentially decrease when the structure is knotted. From the view of chemic quality, the residues with larger values of C_L have a higher tendency to form hydrophobic clusters and disulfide bridges. However, for the short-range interactions, the amino acid effect mostly is related to coordinates.

We use the average probability $P_{\mathrm{A-B}}$ of residue–residue contacts for different amino acids A and B to scale the ability to form the contacts between A and B. Considering that A–B and B–A represent the same contact, we show only the results of A–B contact in Table 2, where $R_c = 0.65\,\mathrm{nm}$ is chosen. We find that Trp–Ala has the largest value 0.413 of average probability of long-range contacts. Those residue pairs with a larger value of $P_{\mathrm{A-B}}$ have a higher tendency to form long-range contacts. The 15 topmost long-range contacts are Trp–Ala, Val–Ile, Leu–Val, Val–Ser, Tyr–Lys, Leu–Phe, Trp–Gly, Tyr–Ala, Thr–Asn, Phe–His, Leu–Ser, Val–Gly, Leu–Ile, Val–Ala, and Leu–Ala. Comparing these results of knotted proteins with the results with 9000 other proteins in the PDB (see Table 3), we find some differences. The 15 topmost long-range contacts in unknotted proteins are Cys–Cys, Val–Ile, Val–Ala, Leu–Val, Val–Gly, Leu–Ala, Leu–Ile, Ile–Ala, Ala–Gly, Val–Thr, Val–Phe, Gly–Thr, Leu–Gly, Gly–Ser, and Val–Tyr. In these 15 topmost long-range contacts, several long-range contacts have changed. Without these long-range contacts, it is difficult to become globular and knotted, and we suppose that these several changed long-range contacts in knotted proteins play a special role in the protein

folding. Obviously, the values of average probability for knotted proteins are generally larger than those for unknotted proteins, which reflects again the fact that long-range contacts make more contributions to the protein folding and play an active role in the stability of protein molecules. On the other hand, in knotted proteins, Trp–Trp long-range contact has a minimum value, and the average probability $P_{\mathrm{A-B}}$ of Trp–Trp long-range contact is 0.01. Values of Cys–His, Met–Met, Phe–Phe, Met–Trp, Glu–Asp, Glu–Glu, Arg–Arg, Gln–Asp, Tyr–Tyr, Trp–Gln and His–Gln contacts are all less than 0.02, much smaller than those of 15 topmost long-range contacts, implying the central role of those topmost contacts in protein folding.

Table 2. Values of average probability $P_{\mathrm{A-B}}$ of residue–residue contacts between residues A and B in knotted proteins. Upper triangle region counts the long-range contacts and bottom triangle region counts the short-range contacts in protein samples, where R_{C}=0.65 nm.

	Leu	Vai	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	
	0.145	0.308	0.243	0.122	0.280	0.173	0.133	0.134	0.230	0.214	0.153	0.180	0.109	0.066	0.094	0.135	0.125	0.248	0.078	0.101	Leu
Leu	0.059	0.158	0.313	0.100	0.139	0.205	0.137	0.206	0.235	0.245	0.142	0.188	0.064	0.208	0.152	0.043	0.135	0.296	0.191	0.083	Val
Val	0.249	0.103	0.063	0.086	0.194	0.180	0.047	0.145	0.151	0.123	0.198	0.107	0.161	0.029	0.142	0.128	0.146	0.133	0.066	0.119	Ile
Ile	0.160	0.134	0.050	0.020	0.145	0.052	0.046	0.010	0.065	0.047	0.057	0.022	0.033	0.025	0.03	0.130	0.026	0.070	0.158	0.032	Met
Met	0.086	0.085	0.07	0.047	0.019	0.144	0.033	0.083	0.126	0.086	0.104	0.250	0.153	0.086	0.055	0.129	0.059	0.159	0.083	0.115	Phe
Phe	0.146	0.137	0.105	0.057	0.016	0.015	0.020	0.218	0.262	0.171	0.155	0.095	0.041	0.102	0.026	0.149	0.282	0.192	0.082	0.087	Tyr
Tyr	0.094	0.100	0.040	0.062	0.035	0.026	0.153	0.027	0.050	0.136	0.031	0.018	0.041	0.130	0.111	0.043	0.031	0.144	0.042	0.120	Cys
Cys	0.146	0.043	0.043	0.055	0.030	0.019	0.053	0.010	0.413	0.267	0.089	0.086	0.070	0.018	0.068	0.164	0.069	0.152	0.110	0.080	Trp
Trp	0.042	0.144	0.027	0.029	0.143	0.025	0.015	0.007	0.126	0.222	0.132	0.209	0.056	0.110	0.135	0.095	0.155	0.216	0.132	0.079	Ala
Ala	0.143	0.106	0.133	0.109	0.262	0.194	0.042	0.037	0.104	0.124	0.131	0.116	0.083	0.107	0.095	0.225	0.180	0.196	0.099	0.115	Gly
Gly	0.231	0.127	0.119	0.076	0.043	0.087	0.058	0.024	0.190	0.115	0.080	0.024	0.079	0.029	0.150	0.255	0.133	0.131	0.085	0.115	Thr
Thr	0.197	0.073	0.201	0.070	0.108	0.098	0.068	0.087	0.142	0.152	0.076	0.067	0.066	0.019	0.062	0.132	0.021	0.086	0.086	0.083	His
His	0.055	0.067	0.047	0.050	0.085	0.023	0.008	0.023	0.101	0.167	0.046	0.067	0.012	0.063	0.019	0.161	0.054	0.167	0.173	0.170	Glu
Glu	0.240	0.170	0.133	0.144	0.084	0.073	0.038	0.087	0.123	0.159	0.148	0.053	0.125	0.057	0.012	0.068	0.088	0.122	0.027	0.065	Gln
Gln	0.243	0.225	0.086	0.061	0.055	0.097	0.059	0.035	0.180	0.172	0.098	0.028	0.129	0.036	0.045	0.030	0.127	0.171	0.169	0.027	Asp
Asp	0.176	0.154	0.107	0.075	0.098	0.104	0.041	0.136	0.212	0.118	0.193	0.122	0.068	0.227	0.059	0.022	0.217	0.123	0.043	0.025	Asn
Asn	0.176	0.119	0.091	0.076	0.110	0.163	0.029	0.037	0.188	0.136	0.071	0.145	0.176	0.100	0.097	0.066	0.032	0.154	0.023	0.038	Lys
Lys	0.209	0.268	0.134	0.082	0.163	0.035	0.057	0.078	0.119	0.181	0.201	0.060	0.114	0.201	0.223	0.153	0.115	0.129	0.092	0.048	Ser
Ser	0.141	0.171	0.074	0.042	0.051	0.054	0.034	0.031	0.144	0.074	0.111	0.079	0.126	0.183	0.211	0.070	0.124	0.057	0.012	0.088	Arg
Arg	0.304	0.100	0.145	0.056	0.065	0.067	0.053	0.030	0.145	0.121	0.097	0.024	0.141	0.087	0.113	0.077	0.160	0.117	0.047	0.044	Pro
Pro	0.201	0.047	0.052	0.055	0.068	0.078	0.032	0.111	0.124	0.142	0.080	0.111	0.066	0.126	0.055	0.040	0.138	0.108	0.040	0.023	
	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	

We also discuss the probabilities in forming short-range contacts, and the results of knotted proteins are given in the bottom triangle region of Table 2. The value of $P_{\mathrm{A-B}}$ for short-range contacts ranges from 0.007 to 0.304. Most of them except Gln–Asp, Leu–Gln, Gln–Lys, Leu–Glu, Asp–Ser, Leu–Gly, Leu–Arg, Leu–Val, Val–Lys, Leu–Pro, Phe–Ala, Val–Gln, Asp–Lys, Ala–Asp, Leu–Lys, Ile–Thr and Thr–Lys short-range contacts, have a value smaller than 0.2 each. The fact that values of $P_{\mathrm{A-B}}$ for short-range contacts are generally smaller than those for long-range contacts in knotted proteins while values of short-range contacts in knotted proteins are generally smaller than those in other proteins (in the lower triangle of Table 3), reveals the fact that the knotted structure leads to a remarkable reduction in short-range contact, which is consistent with the conclusion obtained from Fig.2. This is comprehensible because it depends mainly on the short-haul sequence of amino acids to form a short-range contact. For the same reason, we guess if the amino acid has a large probability in all the 20 amino acids, it may have a large value of the short-range contacts.

Table 3. Values of average probability P_{A-B} of residue–residue contacts between residues A and B in other 9000 proteins. Upper triangle region counts the long-range contacts and bottom triangle region counts the short-range contacts in protein samples, where R_C=0.65 nm.

	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	
	0.107	0.246	0.203	0.096	0.155	0.147	0.094	0.093	0.212	0.172	0.149	0.077	0.086	0.082	0.082	0.087	0.093	0.136	0.113	0.099	Leu
Leu	0.166	0.165	0.260	0.118	0.183	0.168	0.116	0.097	0.254	0.214	0.187	0.100	0.113	0.102	0.110	0.103	0.124	0.166	0.126	0.115	Val
Val	0.227	0.083	0.112	0.096	0.149	0.144	0.093	0.082	0.201	0.165	0.153	0.078	0.089	0.083	0.093	0.088	0.101	0.136	0.102	0.083	Ile
Ile	0.217	0.153	0.077	0.029	0.082	0.073	0.050	0.045	0.095	0.083	0.074	0.051	0.042	0.040	0.043	0.048	0.046	0.070	0.050	0.051	Met
Met	0.168	0.114	0.114	0.051	0.065	0.120	0.082	0.070	0.148	0.139	0.117	0.071	0.071	0.068	0.069	0.079	0.073	0.116	0.092	0.078	Phe
Phe	0.182	0.127	0.113	0.103	0.059	0.074	0.097	0.074	0.148	0.140	0.125	0.063	0.062	0.072	0.073	0.085	0.101	0.113	0.094	0.087	Tyr
Tyr	0.163	0.114	0.112	0.084	0.102	0.046	0.284	0.065	0.090	0.106	0.093	0.054	0.047	0.055	0.060	0.064	0.058	0.098	0.060	0.066	Cys
Cys	0.107	0.084	0.070	0.047	0.061	0.063	0.044	0.029	0.083	0.083	0.071	0.045	0.039	0.049	0.041	0.055	0.053	0.071	0.066	0.063	Trp
Trp	0.101	0.090	0.070	0.053	0.061	0.064	0.051	0.035	0.121	0.188	0.155	0.085	0.084	0.086	0.098	0.095	0.097	0.143	0.106	0.115	Ala
Ala	0.322	0.239	0.217	0.161	0.176	0.160	0.120	0.107	0.196	0.121	0.179	0.083	0.089	0.104	0.130	0.126	0.098	0.169	0.111	0.142	Gly
Gly	0.206	0.163	0.147	0.101	0.133	0.127	0.089	0.080	0.228	0.099	0.082	0.082	0.076	0.088	0.100	0.110	0.090	0.150	0.099	0.098	Thr
Thr	0.183	0.142	0.132	0.094	0.110	0.102	0.074	0.071	0.198	0.165	0.068	0.026	0.043	0.042	0.063	0.051	0.041	0.075	0.054	0.052	His
His	0.155	0.107	0.109	0.076	0.093	0.089	0.061	0.063	0.140	0.101	0.094	0.045	0.019	0.042	0.039	0.053	0.086	0.072	0.080	0.065	Glu
Glu	0.277	0.184	0.190	0.132	0.153	0.136	0.078	0.092	0.267	0.164	0.164	0.108	0.116	0.022	0.047	0.058	0.054	0.076	0.057	0.059	Gln
Gln	0.221	0.138	0.148	0.108	0.115	0.112	0.068	0.079	0.197	0.134	0.123	0.089	0.165	0.073	0.026	0.086	0.089	0.102	0.090	0.077	Asp
Asp	0.208	0.166	0.147	0.099	0.138	0.126	0.074	0.079	0.222	0.164	0.156	0.106	0.159	0.144	0.076	0.051	0.068	0.106	0.066	0.071	Asn
Asn	0.184	0.139	0.134	0.093	0.118	0.116	0.077	0.087	0.186	0.136	0.134	0.086	0.150	0.134	0.156	0.075	0.027	0.091	0.048	0.059	Lys
Lys	0.253	0.166	0.169	0.110	0.141	0.130	0.077	0.088	0.232	0.162	0.136	0.100	0.257	0.149	0.197	0.154	0.087	0.079	0.092	0.094	Ser
Ser	0.207	0.147	0.152	0.098	0.132	0.130	0.084	0.088	0.208	0.166	0.141	0.093	0.161	0.136	0.173	0.138	0.151	0.085	0.033	0.073	Arg
Arg	0.241	0.159	0.157	0.114	0.129	0.119	0.080	0.097	0.246	0.150	0.131	0.092	0.231	0.155	0.178	0.131	0.142	0.136	0.082	0.041	Pro
Pro	0.139	0.115	0.095	0.065	0.097	0.088	0.047	0.053	0.133	0.107	0.096	0.068	0.108	0.095	0.095	0.086	0.105	0.114	0.095	0.037	
	Leu	Val	Ile	Met	Phe	Tyr	Cys	Trp	Ala	Gly	Thr	His	Glu	Gln	Asp	Asn	Lys	Ser	Arg	Pro	

The shape distribution of knotted proteins and the contrast in the case of unknotted proteins are also shown in Fig.3. The abscissa denotes the interval of the shape factor δ^*, and the vertical ordinate presents the proportion of the interval. An conspicuous variance between the knotted proteins and the unknotted proteins is that about 75% knotted proteins are located in the shape factor interval of 0–0.05, while only 15% proteins are located in this interval for unknotted proteins. We also calculate the average shape factor $\langle\delta^*\rangle$, which is defined as

$$\langle\delta^*\rangle = (1/N)\sum_{i=1}^{N}\delta_i^*, \tag{5}$$

where N is the total number of proteins, i.e. N=273

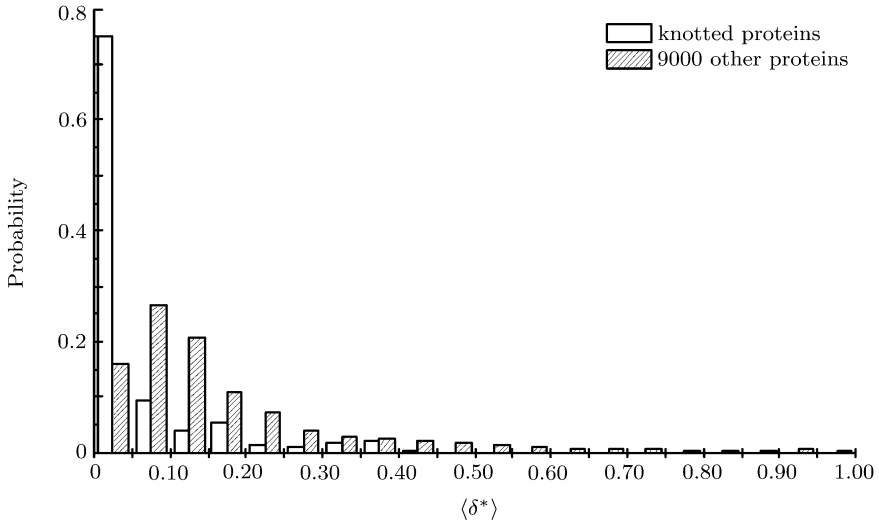


Fig.3. Comparison of the probability of shape factor δ^* in different intervals between knotted proteins (all 273 entries in the PDB, and not just a subset) and 9000 other proteins in the PDB.

for knotted proteins and $N = 9000$ for random proteins. Our results are $\langle\delta^*\rangle = 0.059$ for knotted proteins and $\langle\delta^*\rangle = 0.182$ for random proteins. The value of δ^* varying from 0 to 1 implies that the shape changes from sphere to rod, while the value of δ^* close to zero means that the shape of protein is globular. We can conclude that most of the knotted proteins are absolutely globular because the average shape factor is 0.059 for the knotted proteins, which is only about one-third of the average shape factor for the unknotted proteins. Additionally, the shape factor δ^* of all knotted proteins is less than 0.45, while for unknotted proteins, there remain 0.68% of unknotted proteins with a large value of $\delta^*(\approx 1.0)$. Those facts show that knots may make structures more globular.

4. Conclusions

The main purpose of the present study is to gain an understanding of structural statistical properties of all 273 knotted proteins in the PDB. Moreover, we also count up to 9000 other entries to make an accurate comparison, and we do find some significant differences. We observe a higher probability of long-range contacts in the knotted proteins than those in other ones, implying their important role in achieving knots. In addition, the shape distribution of the knotted proteins and the unknotted proteins are also presented. The comparison indicates that knots may make structures more globular.

References

[1] Virnau P, Kantor Y and Kardar M 2005 *J. Am. Chem. Soc.* **127** 15102
[2] Mansfield M L 1994 *Macromolecules* **27** 5924
[3] Lua R C, Borovinskiy A L and Grosberg A Y 2004 *Polymer* **45** 717
[4] Mansfield M L 1994 *Nat. Struct. Mol. Biol.* **1** 213
[5] Mansfield M L 1997 *Nat. Struct. Mol. Biol.* **4** 166
[6] Taylor W R 2000 *Nature* **406** 916
[7] Taylo W R and Lin K 2003 *Nature* **421** 25
[8] Lua R C and Grosberg A Y 2006 *PLOS Comp. Biol.* **2** 350
[9] Liang C and Mislow K 1994 *J. Am. Chem. Soc.* **116** 3588
[10] Crippen G M 1974 *Theor. Biol.* **45** 327
[11] Crippen G M 1975 *Theor. Biol.* **51** 495
[12] Connolly M L, Kuntz I D and Crippen G M 1980 *Biopolymers* **19** 1167
[13] Klapper M H and Klapper I Z 1980 *Biochim. Biophys. Acta* **626** 97
[14] Kikuchi T, Némethy G and Scheraga H A 1986 *J. Comput. Chem.* **7** 67
[15] Kikuchi T, Némethy G and Scheraga H A 1989 *J. Comput. Chem.* **10** 287
[16] Mao B 1989 *J. Am. Chem. Soc.* **111** 6132
[17] Benham C J and Jafri M S 1993 *Protein Sci.* **2** 41
[18] Mao B 1993 *Protein Sci.* **2** 1057
[19] Murzin A G, Brenner S E, Hubbard T and Chothia C 1995 *J. Mol. Biol.* **247** 536
[20] Tyagi R, Duquerroy S, Navaza J, Guddat L W and Duggleby R G 2005 *Protein Sci.* **14** 3089
[21] Wagner J R, Brunzelle J S, Forest K T and Vierstra R D 2005 *Nature* **438** 325
[22] Virnau P, Mirny L A and Kardar M 2006 *PLOS Comp. Biol.* **2** 1074
[23] Ji G F, Xue B and Wang W 2004 *Chin. Phys.* **13** 1574
[24] Jiang F and Li N 2007 *Chin. Phys.* **16** 392
[25] Chen Y, Wang J and Wang W 2007 *Chin. Phys.* **16** 868
[26] Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 *Nucleic Acids Research* **28** 235
[27] Miyazawa S and Jernigan R L 1996 *J. Mol. Biol.* **256** 623
[28] Miyazawa S and Jernigan R L 1985 *Macromolecules* **18** 543
[29] Crippen G M 1977 *Biopolymers* **16** 2189
[30] Miyazawa S and Jernigan R L 1982 *Biopolymers* **21** 1333
[31] Jernigan R L and Miyazawa S 1983 *Biopolymers* **22** 79
[32] Gromiha M M 2001 *Biophys. Chem.* **91** 71
[33] Gromiha M M and Selvaraj S 1999 *Biophys. Chem.* **77** 49
[34] Gromiha M M and Selvaraj S 2001 *Int. J. Biol. Macromol.* **29** 25
[35] Jiang Z T, Zhang L X, Chen J, Xia A G and Zhao D L 2002 *Polymer* **43** 6037
[36] Chen J, Zhang L X, Li J, Wang X H, Jiang Z T and Zhao D L 2003 *Biophys. Chem.* **105** 11
[37] Zhang L X, Li J, Jiang Z T and Zhao D L 2003 *Polymer* **44** 1751
[38] Solc K and Stockmayer W H 1971 *J. Chem. Phys.* **54** 2756
[39] Solc K 1971 *J. Chem. Phys.* **55** 335
[40] Zifferer G and Preusser W 2001 *Macromol. Theory Simul.* **10** 397
[41] Jagodzinski O, Eisenriegler E and Kremer K 1992 *J. Phys.* **2** 2243
[42] Fasman G D 1989 *Prediction of Protein Structure and Principles of Protein Conformation* (New York: Plenum)