Available online at www.sciencedirect.com

SCIENCE @ DIRECT®

Polymer 45 (2004) 1045–1053

polymer

www.elsevier.com/locate/polymer

Analysis of structural statistical properties of proteases and nonproteases

Tingting Sun, Linxi Zhang*, Jin Chen

Department of Physics, Zhejiang University, Hangzhou 310027, People's Republic of China

Received 21 October 2003; received in revised form 30 November 2003; accepted 30 November 2003

Abstract

The different structural properties of proteases and nonproteases are well investigated in this paper. The average percentage $\overline{P_L}$ of residues having a number of long-range contacts greater than or equal to $N_L(\geq N_L)$ for the proteases is larger than that for the nonproteases. The average number of long-range contacts per residue $\overline{P_{L,\mu}}$ in four secondary structure μ ($\mu =$ H, E, T, and N) for the proteases is also larger than that for the nonproteases. We calculate the average contact order (CO) per protein, the average long-range order (LRO) per protein, and the average total contact distance (TCD) per protein, and find that the average value of LRO for the nonproteases is smaller than that for the proteases. However, both proteases and nonproteases have the same average values of CO and TCD. The average number of long-range contacts per residue C_L for the proteases is larger than that for the nonproteases, however, the average number of short-range contacts per residue C_S for the proteases is smaller than that for the nonproteases. It is also shown that the square of radius of gyration for the proteases is relatively smaller than for the nonproteases. This finding implies that proteases are more compact than nonproteases. In protein molecule, each residue has a different ability to form contacts, and in general the number of residues having a small number of contacts is greater than that having a large number of contacts. Here we have concluded that the probability $P(n)$ of amino acid residues having n pairs of contacts in all residues fits a good Gaussian distribution, and there has the same form of Gaussian distribution for 20 amino acid residues. The most probable number of contacts, n_C, for the proteases is greater than that for the nonproteases for 20 amino acid residues and one has a good correlation with the Fauchere-Pliska hydrophobicity (FPH) scale. Finally we discuss the relative contribution of amino acid residues involved in cation–π interactions. The higher fraction of cation–π interactions observed in the proteases is found to be reflection of the more general, more frequent occurrence of these interactions in these proteins. All these findings would be helpful for us to understand structural differences between the proteases and other proteins.

Keywords: Contact; Cation–π interaction; Protease and nonproteases

1. Introduction

Protease refers to a group of enzymes whose catalytic function is to hydrolyze (breakdown) peptide bonds of proteins. They are also called proteolytic enzymes or proteinases. Protease is a very well studied class of protein. Before the development of recombinant methods for protein expression, digestive enzymes were the subjects of many early structural and mechanistic studies, because they were easy to obtain in large quantities from natural sources [1]. Today, the database of protease structures has grown to include a variety of molecules that play critical roles in many biological processes ranging from viral replication to the development and growth of an organism [2].

Protease is one of the most important proteins in organism. Regulation is particularly important for protease, because all proteins are their natural substrates. There are different mechanisms in protease regulation. These include inhibition by specific protease inhibitors and synthesis as zymogens with covalently attached, inhibitory prosegments [3]. Proteases may also be restricted to some parts of the cell or function only under specific environmental conditions.

Stawiski et al. [2] has found that proteases have smaller than average surface areas, smaller radii of gyration, and higher C^{α} densities by comparing with other proteins of similar size. All these imply that proteases are, as a group, more tightly packed than other proteins. Furthermore, there are also notable differences in secondary structure content between two groups of proteins: proteases have fewer

* Corresponding author. Tel./fax: +86-571-87951328.
E-mail address: Zhanglx@mail.hz.zj.cn (L. Zhang).

0032-3861/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.polymer.2003.11.041

helices and more loops. Besides, they also have successfully trained a neural network to use global structural characteristics to predict protease function.

In this paper, we will discuss structural difference between proteases and nonproteases in more detail. First, we will investigate the degree of structure compactness for the proteases and nonproteases through calculating the percentage of residues having a large number of long-range contacts in all the residues of proteins. In the meantime, we also consider the average ability of forming long-range contacts in secondary structure μ ($\mu =$ H, E, T, and N) for the proteases and nonproteases. Three parameters of contact order (CO), long-range order (LRO), and total contact distance (TCD), and the average number of long-range and short-range contacts per residue for the proteases and nonproteases are also discussed. Furthermore, the probability distribution of amino acid residues according to the different number of contacts for each residue, and cation–π interactions for the proteases and nonproteases are also considered. Some discussions about the reason why there exists the difference between the proteases and nonproteases are made.

2. Method of calculation

2.1. Database

There is considerable redundancy in the databases of proteases and nonproteases, as many proteins are identical or very similar in sequence. However, statistical analyses of protein sequence-structure relation require nonredundant data [4]. To reduce redundancy in the Protein Data Bank of 3D protein structures, which is caused by many homologous proteins in the data bank, we have selected a representative set of structures. The selection algorithm was designed to: (1) select as many nonhomologous structures as possible, (2) select structures of good quality. This representative set may reduce time and effort in statistical analyses [5]. The criteria for the structure selection with no more than 25% sequence identity and crystallographic resolution better than 0.25 nm was given by Hobohm and Sander [6,7]. Only structures of biologically active, monomeric proteins were used. The structure selection criteria were the same for the proteases except that the sequence identity cutoff for proteases was 35% so as to include more examples. In both cases, the molecular mass range was 14–54 kDa. The final sets only contained 36 protease and 154 nonprotease structures. The 36 proteases are 1ac5, 1arb, 1bqb, 1bxo, 1cgh, 1cj1, 1cv8, 1dan, 1eag, 1elt, 1exf, 1ezm, 1gci, 1hfc, 1hne, 1htr, 1iab, 1iag, 1igb, 1kuh, 1lam, 1lay, 1lml, 1mat, 1obr, 1sgp, 1smp, 1try, 1xjo, 2alp, 2asi, 2ctc, 2prk, 3pbh, 5gds, and 8pch. At the same time, the PDB identifiers of the 154 nonproteases are also listed here, and they are 153l, 16pk, 1a0p, 1a17, 1a26, 1a34, 1aog, 1a8e, 1a9s, 1ad6, 1ads, 1ah7, 1ak0, 1akl, 1ako, 1akz, 1alu, 1am7, 1amm, 1amx, 1anf, 1aoh, 1aol, 1aqb, 1arv, 1ash, 1asw, 1at0, 1atg, 1aua, 1auk, 1ax8, 1axn, 1azo, 1bc5, 1bd8, 1bg0, 1bgc, 1bgf, 1bjk, 1bkb, 1bol, 1bpl, 1brt, 1bvl, 1bxw, 1byb, 1byq, 1c25, 1c3d, 1ceo, 1cex, 1cfb, 1cfr, 1chd, 1ckn, 1cnv, 1cpo, 1csh, 1csn, 1dad, 1dhr, 1dhs, 1dhy, 1edg, 1fdr, 1fmt, 1fts, 1g3p, 1gky, 1gpl, 1grj, 1gso, 1ha1, 1hxn, 1idk, 1ihp, 1inp, 1ips, 1ixh, 1juk, 1lba, 1lcl, 1lit, 1lki, 1maz, 1mml, 1mpg, 1mrj, 1mrp, 1msk, 1mup, 1nar, 1nif, 1nkr, 1np4, 1npk, 1ois, 1opr, 1oyc, 1pda, 1pgs, 1phc, 1phm, 1pmi, 1pne, 1poc, 1pot, 1pta, 1pty, 1pud, 1qnf, 1qtq, 1ra9, 1rcf, 1rec, 1rhs, 1rss, 1rsy, 1sbp, 1tca, 1tde, 1tfr, 1thv, 1tib, 1tml, 1uae, 1uxy, 1v39, 1vhh, 1vid, 1vjs, 1wab, 1zin, 2abk, 2baa, 2cba, 2cyp, 2dri, 2end, 2gar, 2hft, 2ilb, 2liv, 2pia, 2plc, 2pth, 2sns, 2thi, 3nll, 3seb, 3sil, 4xis, and 6cel. Now we will discuss the statistical properties of two kinds of proteins above, especially the differences between them. Deducing the functions of proteins from their structures would be beneficial.

2.2. Computation for short- and long-range contacts

Each residue in a protein molecule is represented by C^{α} atom. Residues whose C^{α} atoms are closer than R_{C} are defined to form a contact. This kind of simple method to evaluate the number of residue–residue contacts in proteins has often been used in many articles [4–11]. In addition, we choose the value $R_{\mathrm{C}} = 0.60$ nm.

For a given residue, the composition of surrounding residues is discussed in terms of the location at the sequence level and the contributions from $\leq \pm 4$ residues are treated as a short-range contact, and $> \pm 4$ residues as a long-range contact [12–17].

2.3. Percentage of residue with a large number of long-range contacts for the proteases and nonproteases

The number of long-range residue–residue contacts can be calculated easily and effectively. In order to know the structures of proteases and nonproteases in more detail, here we discuss the average percentage $\overline{P_{\mathrm{L}}}$ of residues having a number of long-range contacts greater than or equal to N_{L} ($\geq N_{\mathrm{L}}$) per protein molecule for the proteases and nonproteases. First we introduce the percentage P_{L} of residues having a number of long-range contacts greater than or equal to N_{L} ($\geq N_{\mathrm{L}}$) [17]

$$P_{\mathrm{L}} = \frac{N_{P_{\mathrm{L}}}}{N'} \tag{1}$$

here N' is the total number of amino acid residues in a protein molecule, and $N_{P_{\mathrm{L}}}$ is the number of residues whose number of long-range contacts is greater than or equal to N_{L} ($\geq N_{\mathrm{L}}$) in a protein. Therefore, we can consider the average percentage $\overline{P_{\mathrm{L}}}$ of proteases and nonproteases

$$\overline{P_{\mathrm{L}}} = \frac{\sum_{i=1}^{M} P_{\mathrm{L},i}}{M} \tag{2}$$

Here M represents the total number of proteins in the proteases and nonproteases, and $M = 36$ and 154, respectively. $P_{L,i}$ is the percentage of residues having a number of long-range contacts greater than or equal to N_L ($\geq N_L$) for i-th protein.

It is important for us to know different statistical properties in the secondary structure of proteins. Here the secondary structure definition follows the convention of Kabsch and Sander [18], except that H includes all residues marked H, G, I and P in the program DSSP, E strands for both E and B, T for S and T, and N represents the residues with no characteristic secondary structure. We first introduce the number of long-range contacts per residue $S_{L,\mu}$ in secondary structure μ ($\mu =$ H, E, T, and N), and define it as

$$S_{L,\mu} = \frac{N_{L,\mu}}{N_\mu} \quad (\mu = \text{H, E, T, and N}) \tag{3}$$

Here N_μ is the total number of residues in secondary structure μ, and $N_{L,\mu}$ is the total number of long-range contacts in secondary structure μ for a protein. We only consider the total number of residues in the same secondary structure, and different types of amino acid residues in the same secondary structure are not distinguished. Therefore, there include different types of amino acid residues in the same secondary structure μ ($\mu =$ H, E, T, and N) in proteins.

We are eager to know whether $S_{L,\mu}$ is the same or not for the proteases and nonproteases. So we consider the average number of long-range contacts per residue in secondary structure μ, and it is

$$\overline{S_{L,\mu}} = \frac{\sum_{i=1}^{M} S_{L,\mu,i}}{M} \tag{4}$$

Here M represents the total number of proteins in the proteases and nonproteases, and $M = 36$ and 154, respectively.

2.4. The calculation of three contact parameters (CO, LRO, and TCD) for the proteases and nonproteases

Previous studies found that CO, LRO, and TCD have significant correlations with folding rate of protein [19–24]. Here, we also calculate the values of three parameters to see whether they are different or not for the proteases and nonproteases. In fact, those three parameters represent the total statistical properties of contacts in proteins.

The logarithms of folding rates ($\ln k_f$) of proteins that fold with two- or weakly three-state kinetics has a surprisingly simple and statistically significant correlation with a single parameter called contact order (CO) [22], and CO is defined as

$$\text{CO} = \frac{1}{n_c n_r} \sum_{|j-i|>l_{\text{cut}}}^{n_c} |j - i| \tag{5}$$

where n_r is the number of amino acid residues of a protein, and n_c is number of nonlocal residue–residue contacts, i and j represent the positions of two residues. A nonlocal contact is defined as two heavy atoms (excluding hydrogen atoms) within a cutoff distance R_C and separated by at least a residue separation cutoff value l_{cut}. Here $l_{\text{cut}} = 2$. This parameter reflects the relative importance of nonlocal contacts in protein structures. In fact, nonlocal contacts here include short- and long-range contacts.

Another parameter is found to correlate better with $\ln k_f$ than CO, which is called long-range order (LRO) for a protein from the knowledge of long-range contacts (contacts between two residues that are close in space and far in the sequence) in protein structure [21]. It is defined as

$$\text{LRO} = \frac{\sum n_{ij}}{n_r} \qquad n_{ij} = \begin{cases} 1 & |j - i| > 12 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

here i and j represent two residues for which the $C^\alpha - C^\alpha$ distance is ≤ 0.60 nm and n_r is the number of amino acid residues of a protein [21]. The new results suggest the importance of the long-range contacts in folding kinetics. The difference between the two parameters is that LRO only considers long-range contacts, while CO also discusses all the contacts of proteins.

Third parameter called total contact distance (TCD) was brought forward by Zhou and Zhou [23]. It is shown to be the best in correlation with the logarithms of folding rates, and the expression for TCD is

$$\text{TCD} = \frac{1}{n_r^2} \sum_{|j-i|>l_{\text{cut}}}^{n_c} |j - i| \tag{7}$$

Comparing with Eq. (5), we can easily find that CO is a quantity per contact whereas TCD is the summation over all the contacts per residue.

2.5. Average number of short- and long-range contacts per residue

Average number of contacts per residue indicates the ability to form contacts. Here we define the average number of long-range contacts per residue $C_{\alpha,L}$ and the average number of short-range contacts per residue $C_{\alpha,S}$ as

$$C_{\alpha,\eta} = \frac{\sum_{\beta=\text{Ala,Asp,Cys,Glu,...,Tyr}} N_{\alpha-\beta,\eta}}{N_\alpha} \quad (\eta = \text{S, or, L; } \alpha = \text{Ala, Asp, ..., Tyr}) \tag{8}$$

where N_α is the number of residue α in all proteins, and $N_{\alpha-\beta,\eta}$ is the number of short-range or long-range contacts between residues α and β. So $C_{\alpha,\eta}$ indicates the relative ability to form contacts. If residue α has a large value of $C_{\alpha,\mathrm{L}}$, it means that residue α has a high tendency of forming long-range contacts.

2.6. Probability $P(n)$ of residue with forming n pairs of contacts

In protein molecule, each residue has a different ability of forming contacts, and in general, residues in the interior of proteins have a large number of contacts. Here we introduce the probability $P(n)$ of residues with forming n pairs of contacts in all residues, and it is defined as

$$P(n) = \frac{N_n}{N} \tag{9}$$

here N_n is the total number of amino acid residues with forming n pairs of contacts (including short-range and long-range contacts), and N is the total number of residues in all proteases (or nonproteases).

In the meantime, we also consider the probability $P_\alpha(n)$ for special type α of residue, and it is written as

$$P_\alpha(n) = \frac{N_{\alpha,n}}{N_\alpha} \qquad (\alpha = \mathrm{Ala, Asp, \ldots, Tyr}) \tag{10}$$

here N_α is the total number of residue α in all proteins, and $N_{\alpha,n}$ is the total number of residue α with forming n pairs of contacts in all proteases (or nonproteases). So $P_\alpha(n)$ is the probability distribution of residue α with forming n pairs of contacts in all residues α for the proteases or nonproteases.

2.7. Estimations of cation–π interactions

The cation–π interaction is an important force for molecular recognition in biological receptors [25–30]. Within a protein, cation–π interactions can occur between the cationic sidechains of either Lys, or Arg and the aromatic sidechains of Phe, Tyr, or Trp. The cation–π interactions in each protein can be calculated using the program, CAPTURE, developed by Gallivan and Dougherty [28] available at http://capture.caltech.edu. The percentage composition of a specific amino acid residue contributing to cation–π interactions can be obtained by

$$C_{\mathrm{cat}-\pi,\alpha} = \frac{n_{\mathrm{cat}-\pi,\alpha}}{n_\alpha} \tag{11}$$

here α represents one of Lys, Arg, Phe, Tyr, and Trp, $n_{\mathrm{cat}-\pi,\alpha}$ is the total number of residue α involved in cation–π interaction and n_α is the total number of residue α in all proteases or nonproteases.

In order to analyze the cation–π interactions in more detail, we introduce another parameter $P_{\mathrm{cat}-\pi}$, and it is defined as

$$P_{\mathrm{cat}-\pi} = \frac{N_{\mathrm{cat}-\pi}}{\sum\limits_{\alpha=\mathrm{Lys,Arg,Phe,Tyr,Trp}} n'_\alpha} \tag{12}$$

here $\sum_{\alpha=\mathrm{Lys,Arg,Phe,Tyr,Trp}} n'_\alpha$ is the total number of residues involved in cation–π interactions and $N_{\mathrm{cat}-\pi}$ is the number of total cation–π interactions in a protein.

3. Results and discussions

3.1. The average percentage of residues having a large number of long-range contacts for the proteases and nonproteases

We calculate the average percentage $\overline{P_\mathrm{L}}$ of residues having a number of long-range contacts greater than or equal to N_L ($\geq N_\mathrm{L}$) for the proteases and nonproteases according to Eq. (2). Here N_L ranges from 1 to 6, and $R_\mathrm{C} = 0.60$ nm. The results are given in Fig. 1. From Fig. 1 we can easily find when N_L increases, $\overline{P_\mathrm{L}}$ decreases both for the proteases and nonproteases. The values of $\overline{P_\mathrm{L}}$ for the proteases are relatively larger than that for the nonproteases. Why can the tendency be occurred for the proteases and nonproteases? This may have something with the content of four different structural classes: i.e. all-α, all-β, $\alpha + \beta$, and α/β for the proteases and nonproteases. Here we calculate the percentage of all-α protein, all-β protein, $\alpha + \beta$ protein, and α/β protein in 36 proteases and 154 nonproteases, respectively. The percentage of all-α, all-β, $\alpha + \beta$, and α/β in the proteases is 0, 45.5, 31.4 and 22.9%, respectively. However, it is 18.4, 25.4, 17.5 and 38.6% in the nonproteases, respectively. There is no any all-α protein (0%), and all-β protein has the largest percentage of 45.5% in the proteases. The percentage of all-α protein in the nonproteases is larger than that in the proteases, while the percentage of all-β protein in the nonproteases is smaller than that in the proteases. Gromiha et al. [15] have analyzed

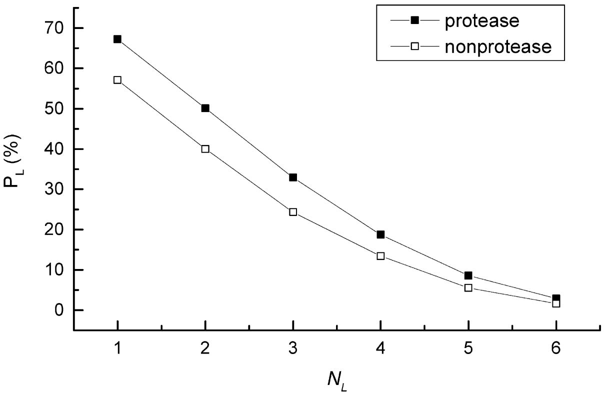


Fig. 1. The average percentage $\overline{P_\mathrm{L}}$ of residues with a number of long-range contacts greater than or equal to N_L ($\geqq N_\mathrm{L}$) versus N_L. (■) Represents proteases and (□) nonproteases, and $R_\mathrm{C} = 0.60$ nm.

that the residues in all-β class of proteins have more long-range contacts than that in all-α proteins, and in our earlier work [17], we have also concluded that the average percentage $\overline{P_L}$ for all-α proteins is greatly lower than that for all-β proteins. Therefore, the reason that the average percentage $\overline{P_L}$ for the proteases is larger than for the nonproteases can be explained clearly.

3.2. The average number of long-range contacts per residue in four secondary structures of H, E, T, and N for the proteases and nonproteases

The average number of long-range contacts per residue in four secondary structures of H, E, T, and N for the proteases and nonproteases has been calculated. Here the secondary structure definition of H, E, T, and N is according to Kabsch and Sander [18]. We find that the value of $\overline{S_{L,\mu}}$ (μ = H, E, T, and N) for the proteases is larger than that for the nonproteases. The average number $\overline{S_{L,\mu}}$ in secondary structure of H, E, T and N is 0.750, 3.23, 1.30, and 1.79, respectively for the proteases, and 0.598, 3.10, 1.00, and 1.52, respectively for the nonproteases. We also observe that the average number of long-range contacts in secondary structure E is much larger than the other three, and 4–5 times as many as in secondary structure H. Therefore, the residues in secondary structure E can form long-range contacts easily.

3.3. The average values of three parameters (CO, LRO, and TCD) for the proteases and nonproteases

According to Eqs. (5)–(7), we calculate the average values of CO, LRO, and TCD for the proteases and nonproteases, and the results are given in Table 1. Those values are averaged over all 36 proteases and 154 nonproteases, respectively. From Table 1, we can draw two conclusions: (1) as there is no difference of the average values of CO and TCD between proteases and nonproteases, they are almost the same in the average statistical properties of all contacts (including short- and long-range contacts) for the proteases and nonproteases. (2) The average value of LRO for the proteases is obviously larger than that for the nonproteases, so we can know that the average statistical properties of long-range contacts are different between the proteases and nonproteases because LRO only considers long-range contacts, therefore LRO can be used to distinguish proteases from nonproteases.

Table 1
Average values of three parameters (CO, LRO, and TCD) for the proteases and nonproteases with $R_C = 0.60$ nm

	$R_C = 0.60$ nm		
	CO	LRO	TCD
Proteases	0.134	1.38	0.340
Nonproteases	0.129	1.10	0.311

3.4. The average number of contacts per residue for the proteases and nonproteases

We calculate the average number of long-range contacts per residue C_L and the average number of short-range contacts per residue C_S for 20 amino acid residues according to Eq. (8), and the results are given in Table 2. Here we discuss the average number of contacts per residue for 20 amino acid residues with $R_C = 0.60$ nm. The amino acids of Leu, Val, Ile, Met, Phe, Tyr, Cys, Trp, Ala, and Gly have a large value of C_L, and the amino acids of Thr, His, Glu, Gln, Asp, Asn, Lys, Ser, Arg, and Pro have a small value of C_L. The results agree with our previous calculations basically [11]. The reason why there exists small deviation between our previous calculation and this work may be that the distributions of all-α, all-β, $\alpha + \beta$, and α/β proteins for the proteases and nonproteases are asymmetry here. Comparing with the value of C_L for the proteases and nonproteases, we find that the value of C_L for the proteases is larger than that for the nonproteases. In Table 2, we also find that the average number C_S of short-range contacts per residue is almost the same for different amino acid residue both proteases and nonproteases, and this means that residue plays an equally important role in forming short-range contacts, however, it is different in forming long-range

Table 2
Average number of contacts per residue for the proteases and nonproteases. C_L (C_S) is the average number of long-range (short-range) contacts per residue. Here $R_C = 0.60$ nm

20 amino acid residues	$R_C = 0.60$ nm			
	Proteases		Nonproteases	
	C_L	C_S	C_L	C_S
Leu	3.88	1.42	2.61	2.16
Val	4.49	1.17	3.62	1.53
Ile	4.13	1.40	3.42	1.76
Met	3.60	1.71	2.87	2.23
Phe	4.00	1.51	3.37	1.94
Tyr	4.03	1.34	3.43	1.83
Cys	4.25	1.35	4.22	1.76
Trp	4.08	1.64	3.29	2.00
Ala	3.86	1.69	2.95	2.25
Gly	4.16	1.16	3.34	1.60
Thr	3.53	1.30	2.95	1.63
His	3.14	1.64	2.92	1.79
Glu	2.25	1.68	1.72	2.07
Gln	2.60	1.70	1.87	2.36
Asp	2.63	1.63	2.02	1.81
Asn	2.55	1.43	2.30	1.76
Lys	2.62	1.55	2.04	2.00
Ser	3.18	1.25	2.76	1.72
Arg	3.00	1.46	2.41	2.10
Pro	2.76	0.88	2.21	1.04

contacts. Opposite to C_L, the value of C_S for the proteases is smaller than that for the nonproteases.

3.5. The distribution of the square of radius of gyration for the proteases and nonproteases

In order to assess the shapes of proteases and nonproteases, we compute the square of radius of gyration for the proteases and nonproteases, and plot the square of radius of gyration as a function of total number of residues with a number of long-range contacts greater than or equal to N_L ($\geq N_L$). In fact, the total number of residues with a large number of long-range contacts means the total number of residues in the interior of proteins, and this measure is sensitive to the mass distribution of proteins. Here we consider two cases of $N_L = 2$ in Fig. 2(a) and $N_L = 3$ Fig. 2(b), respectively. The fit lines of nonproteases are both given in Fig. 2, and the slope is 0.0113 for $N_L = 2$ and 0.0156 for $N_L = 3$, respectively. The slope of $N_L = 2$ is smaller than $N_L = 3$. In Fig. 2(a), most of the proteases, especially for the total number of residues with a number of long-range contacts greater than or equal to N_L are smaller than 350, fall along the lower edge of the line. In Fig. 2(b), most of the proteases fall along the lower edge of the line, especially for those smaller than 200. It suggests that proteases are more compact than nonproteases, and it also judges that the proteases are globally better packed.

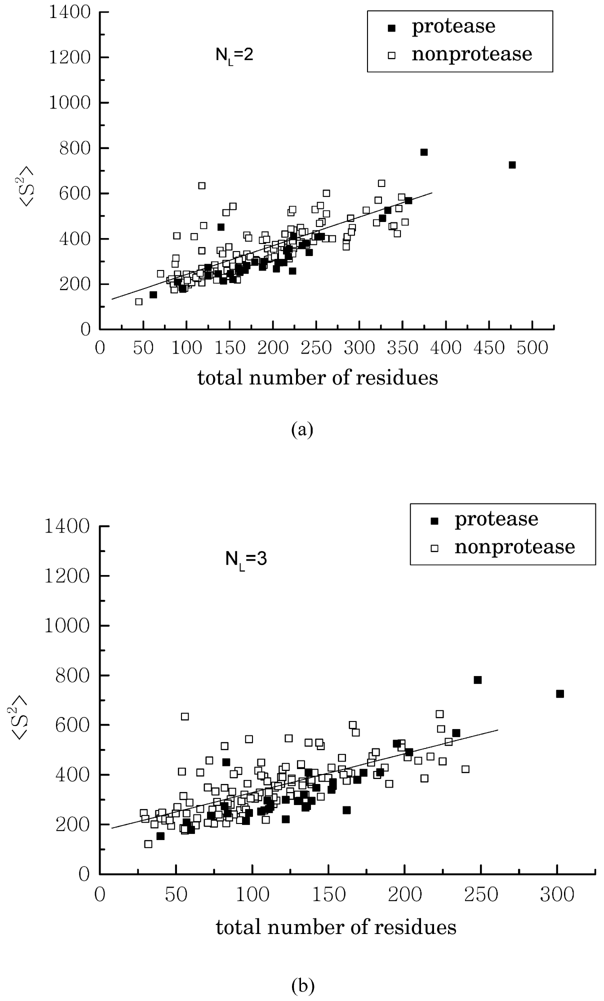


Fig. 2. The square of radius of gyration $\langle S^2 \rangle$ versus total number of residues with a number of long-range contacts greater than or equal to N_L ($\geqq N_L$). (■) Represents proteases and (□) nonproteases. (a) For $N_L = 2$ and (b) for $N_L = 3$, and the slope is 0.0113 in (a) and 0.0156 in (b), respectively. Here $R_C = 0.60$ nm.

3.6. The probability distribution P(n) of amino acid residues with forming n pairs of contacts in all residues

First, we use Eq. (9) to calculate the probability of amino acid residues with forming *n* pairs of contacts in all residues. In Fig. 3, *n* ranges from 0 to 14, and contacts include short- and long-range contacts here. This means some residues in

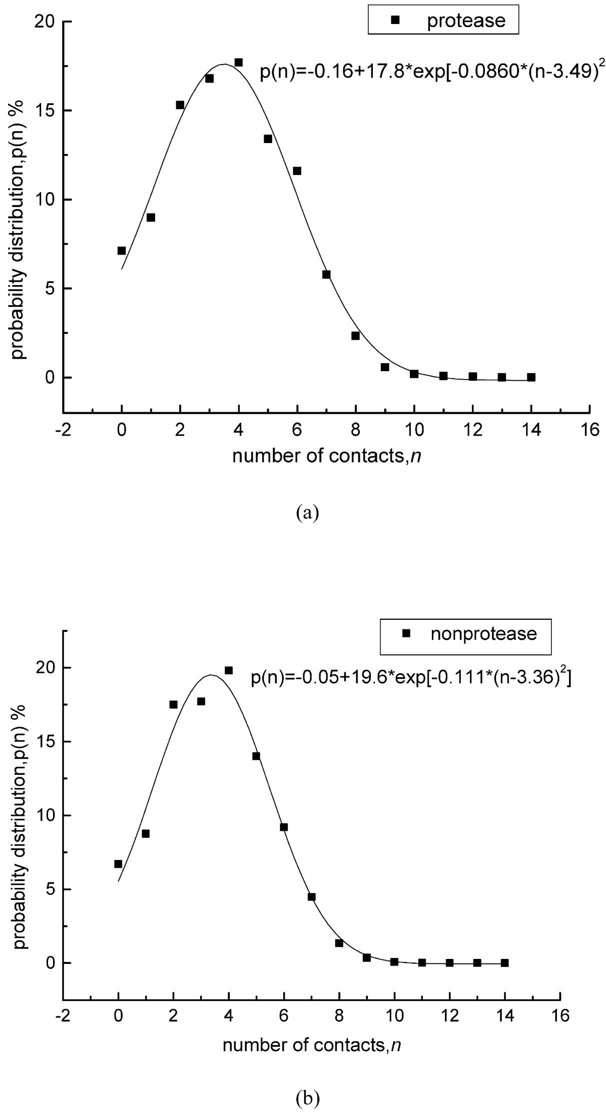


Fig. 3. The probability distribution $P(n)$ of amino acid residues having *n* pairs of contacts (including short- and long-range contacts) versus number of contacts *n*. Here (a) is for the proteases and (b) for the nonproteases, and $R_C = 0.60$ nm.

proteins can form many contacts. Of course, the number of residues with forming many contacts is very few. In the meantime, the maximum of $P(n)$ occurs in the region of $n = 2-4$ when $R_C = 0.60$ nm in Fig. 3. Comparing with Fig. 3(a) and (b), the position of n at the maximum of $P(n)$ moves toward left as $n_C = 3.49$ for the proteases in Fig. 3(a) and $n_C = 3.36$ for the nonproteases in Fig. 3(b). We also find there exists the similar relationship between the probability distribution $P(n)$ and n for both the proteases and nonproteases in Fig. 3(a) and (b). The important thing in Fig. 3 is that they both fit Gaussian distribution. In Fig. 3, the probability distribution $P(n)$ of amino acid residues with forming n pairs of contacts in all residues can be expressed as Gaussian distribution

$$P(n) = P_0 + a\exp[-b(n - n_C)^2] \tag{13}$$

here $P_0 = -0.16, -0.05$, $a = 17.8, 19.6$, $b = 0.086, 0.111$, and $n_C = 3.49, 3.36$ in Fig. 3(a) and (b), respectively.

In this paper, we also calculate the probability distribution $P(n)$ of residue with having n pairs contacts for 20 amino acid residues, and there are different probability distributions for 20 amino acid residues. For example, the probability distribution $P(n)$ for residue Leu is given in Fig. 4. Here the range of n is much smaller, and from 0 to 6 for the proteases. However, the range of n becomes from 0 to 9 for the nonproteases. Comparing with proteases and nonproteases, we find that the maximum of $P(n)$ for the nonproteases (Fig. 4(b)) is larger than that for the proteases (Fig. 4(a)). However, the value of $P(n)$ for $n \geq 3$ in Fig. 4(b) is smaller than that in Fig. 4(a). A good fit for Gaussian distribution function is also found in Fig. 4, and the expression of $P(n)$ for Leu is

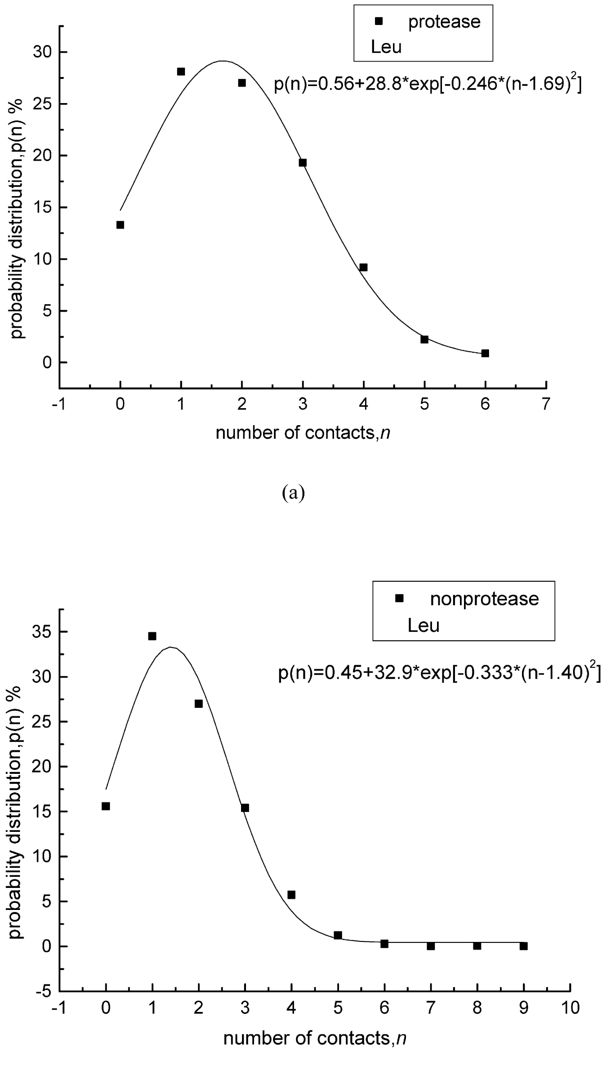
$$P_{\text{Leu}}(n) = P_{0,\text{Leu}} + a_{\text{Leu}}\exp[-b_{\text{Leu}}(n - n_{\text{C,Leu}})^2] \tag{14}$$

Here $P_{0,\text{Leu}} = 0.56, 0.45$, $a_{\text{Leu}} = 28.8, 32.9$, $b_{\text{Leu}} = 0.246, 0.333$, and $n_{\text{C,Leu}} = 1.69, 1.40$, for the proteases and nonproteases, respectively.

An important parameter in Gaussian distribution function is n_C, the most probable number of contacts. We obtain different values of n_C of 20 amino acid residues for the proteases and nonproteases. To a certain extent, n_C, the most probable number of contacts, represents its Gaussian distribution function. We think that n_C may have some correlations with the hydrophobicity scales of 20 amino acid residues. In Fig. 5, we plot n_C as a function of the Fauchere-Pliska hydrophobicity (FPH) scale for both the proteases and nonproteases, and find that the value of n_C increases with FPH value. The relationship between n_C and the FPH scale is expressed approximately as

$$n_C = a + b \times \text{FPH} \tag{15}$$

$a = 1.16$ and $b = 0.34$ for the proteases, and $a = 1.10$ and $b = 0.28$ for the nonproteases.



(a)

(b)

Fig. 4. The probability distribution $P(n)$ of residue Leu having n pairs of contacts (including short- and long-range contacts) versus number of contacts n. Here (a) is for the proteases and (b) for the nonproteases, and $R_C = 0.60$ nm.

3.7. Relative contribution to cation–π interactions for the proteases and non-proteases

The cation–π interaction is increasing recognized as an important noncovalent binding interaction relevant to structure biology. We calculate the percentage composition of a specific amino acid residue contributing to cation–π interactions according to Eq. (11) for the proteases and nonproteases. The relative contributions to cation–π interaction for five amino acid residues in the proteases and nonproteases are shown in Fig. 6. We found that the tendency to forming cation–π interaction for the positively charged residues (Lys, Arg) in the proteases is higher than that in the nonproteases, and the percentage of the residue involving cation–π interactions is 9.4 and 20.2% for Lys and Arg, respectively in the proteases, however, it is 6.9 and

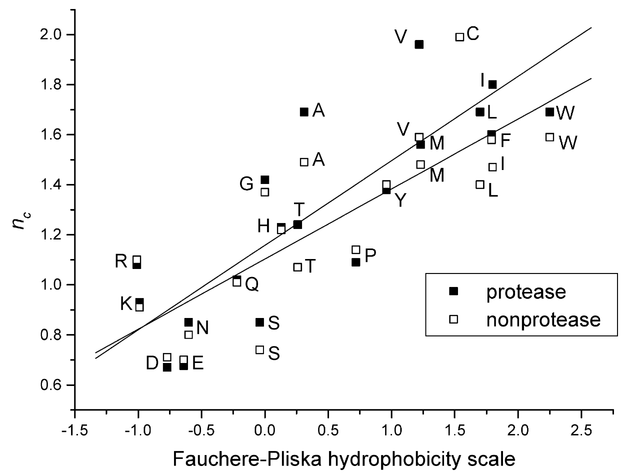


Fig. 5. n_C versus Fauchere-Pliska hydrophobicity scale (FPH) of 20 amino acid residues for the proteases (■) and nonproteases (□), and $R_C = 0.60$ nm.

15.8%, respectively in the nonproteases. For the aromatic residues (Phe, Tyr, Trp), the relative contribution to cation–π interaction in the nonproteases is larger than that in the proteases. The corresponding percentage of the aromatic residues contributing towards cation–π interactions is 8.0, 10.9, and 26.8% in the proteases, respectively, and 7.9, 12.2, and 33.0% in the nonproteases, respectively.

Eq. (12) gives another way to estimate cation–π interactions. Using this equation, we calculate $P_{\text{cat}-\pi}$ in the proteases and nonproteases, and we divide the values of $P_{\text{cat}-\pi}$ into nine regions and the interval is 2%. The relative distribution of $P_{\text{cat}-\pi}$ of the nine regions is shown in Fig. 7. In Fig. 7, we find that in the region of small $P_{\text{cat}-\pi}$, the distribution percentage for the nonproteases is larger than that for the proteases. $P_{\text{cat}-\pi}$ becomes the largest percentage of probability in the region of [10%, 12%]. However, its largest is in the region of [4%, 6%] for the nonproteases.

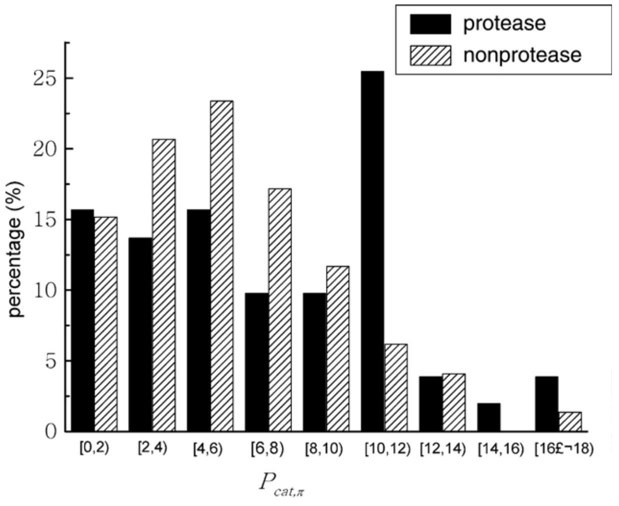


Fig. 7. The percentage distribution of $P_{\text{cat}-\pi}$ in nine different regions for the proteases and nonproteases.

We have analyzed the structural properties of proteases and nonproteases using some methods and parameters. The proteases have more compact structure than nonproteases. We also give some explanations why there exist some structure differences between proteases and nonproteases. Some new characteristics such as Gaussian distribution $P(n)$ of residues with forming n pairs of contacts are also found. This investigation can help us to know the different statistical properties in proteases and nonproteases in more detail, and also help us to distinguish other classes.

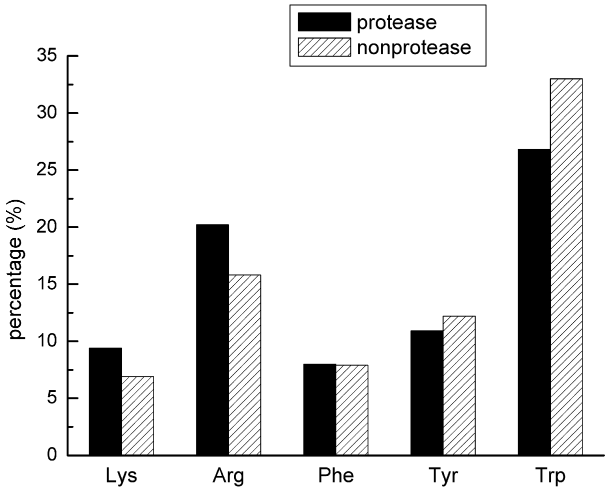


Fig. 6. The percentage of five amino acids (Lys, Arg, Phe, Tyr, and Trp) contributing towards cation–π interactions in the proteases and nonproteases.

Acknowledgements

The research was financially supported by National Natural Science Foundation of China (Nos. 29874012, 20174036, 20274040), and Natural Science Foundation of Zhejiang Province(No. 10102). We also thank the referees for their critical reading of the manuscript and their good ideas.

References

[1] Neurath H. Science 1984;224:350–7.
[2] Stawiski EW, Baucom AE, Lohr SC, Gregoret LM. Proc Natl Acad Sci USA 2000;97:3954–8.
[3] Kham AR, James MNG. Protein Sci 1998;7:815–36.
[4] Crippen GM. Biopolymers 1977;16:2189–96.
[5] Miyazawa S, Jernigan RL. Biopolymers 1982;21:1333–63.
[6] Hobohm U, Scharf M, Schneider R, Sander C. Protein Sci 1992;1:409–17.
[7] Hobohm U, Sander C. Protein Sci 1994;3:522–4.
[8] Miyazawa S, Jernigan RL. J Mol Biol 1996;256:623–4.
[9] Miyazawa S, Jernigan RL. Macromolecules 1985;18:543–52.
[10] Jernigan RL, Miyazawa S. Biopolymers 1983;22:79–85.
[11] Zhouting J, Linxi Z, Jin C, Agen X, Delu Z. Polymer 2002;43:6037–47.
[12] Tanaka S, Scheraga HA. Proc Natl Acad Sci 1975;72:3802–5.
[13] Bahar I, Kaplan M, Jernigan RL. Proteins 1997;29:292–308.
[14] Gromiha MM, Selvaraj S. J Biol Phys 1997;23:209–17.

[15] Gromiha MM, Selvaraj S. Biophys Chem 1999;77:49–68.
[16] Gromiha MM, Selvaraj S. J Biol Phys 1997;23:151–62.
[17] Jin C, Linxi Z, Jing L, Youxing W, Zhouting J, Delu Z. Biophys Chem 2003;105:11–21.
[18] Kabsch W, Sander C. Biopolymers 1983;22:2577–637.
[19] Alm E, Baker D. Proc Natl Acad Sci USA 1999;96:11305–10.
[20] Dinner AR, Kaplus M. Nat Struct Biol 2001;8:21–2.
[21] Gromiha MM, Selvaraj S. J Mol Biol 2001;310:27–32.
[22] Plaxco KW, Simons KT, Baker D. J Mol Biol 1998;277:985–94.
[23] Zhou H, Zhou Y. Biophys J 2002;82:458–63.
[24] Linxi Z, Jing L, Zhouting J, Agen X. Polymer 2003;44:1751–5.
[25] Dougherty DA. Science 1996;271:163–8.
[26] Sussman JL, Harel M, Frolow F, Oefner C, Goldman A, Toker L, Silman I. Science 1991;253:872–9.
[27] Zhong W, Gallivan JP, Zhang Y, Li L, Lester HA, Dougherty DA. Proc Natl Acad Sci USA 1998;95:12088–93.
[28] Gallivan JP, Dougherty DA. Proc Natl Acad Sci USA 1999;96: 9459–64.
[29] Gromiha MM. Biophys Chem 2003;103:251–8.
[30] Wintjens R, Lievin J, Rooman M, Buisine E. J Mol Biol 2002;302: 393–410.