Vol 16 No 3, March 2007
1009-1963/2007/16(03)/0868-08
Chinese Physics

The preferences of orientations between the pairs of amino acids*

Chen Ying(陈 颖), Wang Jun(王 骏), and Wang Wei(王 炜)†

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

(Received 17 February 2006; revised manuscript received 13 October 2006)

In this work, we make an investigation on the preferences of orientations between amino acids using the orientation defined based on the local geometry of the amino acids concerned. It is found that there are common preferences of orientations (70°, 30°, 140°) and (110°, 340°, 100°) for various pairs of amino acids. Different side chains may strengthen or weaken the common preferences, which is related to the effect of packing. Some amino acids having specific local flexibility may possess some preferences of orientations besides the common ones, such as (10°, 280°, 210°). Another analysis on the pairs of the amino acids with different secondary-structure preferences shows that the directional interaction may affect the distribution of orientation more effectively than the packing or local flexibility. All these results provide us some insight of the organization of amino acids in protein, and their relation with some related interactions.

Keywords: native structure of protein, preferences of orientation, preferences of secondary structures
PACC: 8715B

1. Introduction

As a kind of important biopolymer, proteins regularly fold to some specific structures (known as the native structures) to fulfil their biological functions. Various kinds of proteins have different native structures, which have some intrinsic regularities, such as α-helices, β-sheets, and further high-level architectures.[1] Physically, these regularities are the results of the interactions between the amino acid residues. As a consequence, to understand the interactions from the regularities of protein structures is an interesting inverse problem and a useful idea in protein studies,[2−7] which has become a fundamental formulism of knowledge-based protein analysis.

There are many kinds of interactions in protein systems. Some of them are isotropic, and some others are directional, such as hydrogen bonds. Many experiments and simulation studies suggest that the directional interactions do contribute to the formation and stabilization of the secondary structures and the complementary packing of side chains.[8−10] The structural statistics by Thornton's group[11] also suggests the importance of the directional interactions. To understand the effect of directional interactions in the construction of protein structures is an interesting task, which is under active development.[6] Presently, the studies mainly focus on the orientations of side chains of residues.[6,12] The orientational information between backbone bonds, which is related to the basic topology of proteins, has not attracted much attention except the finding of the Ramachandran plot for the local dihedral angles. To what extent the directional interactions contribute to the formation of the overall topology of a protein chain and how much information is included in the orientations of the bonds at backbone level are still open questions. The statistics on the orientational information of backbones is a good entry to understand the directional interactions. The answers to these questions would be instructive for building better coarse-grain models.

In our work, we setup local reference frames (LRFs) based on backbone bonds, and define the orientation based on the projection between the LRFs concerned. A systematic statistics on the orientations is made for all kinds of pairs of amino acids. The common features are analysed in detail. Some specific orientations are investigated. Furthermore, by comparing the orientations of residues concerning some

*Project supported by the National Natural Science Foundation of China (Grant Nos 10204013, 90103031, 10074030, 10474041, 90403120 and 10021001), and the Nonlinear Project (973) of the National Science Ministry, China.

†E-mail: wangwei@nju.edu.cn

secondary-structure features, we illustrate the effect of directional interactions on the orientations. In a word, our work give out some detailed information of orientational arrangement of backbones which may benefit the future molecular modelling.

2. Model and method

2.1. Contacts between residues

Contacts are defined between two spatially neighbouring amino acids. The contacts are often related to interactions in coarse-grain models. Practically, there are various kinds of definitions for contacts.[13−16] A heavy-atom-based definition for contacts is employed in this paper. That is, two residues with the sequence indices i and j could form a contact when they have at least one pair of neighbouring heavy atoms (namely non-hydrogen atoms) whose distance is smaller than 5 Å. This kind of contacts show a large similarity to those defined with sophisticated CSU software,[13] and have more information than the definition based on the C_α atoms only. Besides the threshold for spatial distance between atoms (5Å), the sequential separation of a contact, which is defined as the loop length $|j-i|$, is required not smaller than 4. This requirement is based on the fact that the possible contacts with a short sequential distance are not the results of direct interactions but the results of local chemical constraints. This kind of filtering is regularly considered in the literature.[14,15,17]

2.2. Local reference frames of residues and orientations between them

As building blocks of proteins, amino acids are not isotropically spherical objects. Therefore, not only the distance but also the orientation are both important for describing the relation between contacted residues. Here, we mainly concentrate on the features of the orientations between the concerned pairs of amino acids. Same as many intermediate-level models, the side chain of an amino acid could be simply represented. At this level, the protein chains can be rebuilt from the traces of backbone atoms. Therefore, to consider the orientational information of amino acids from their backbone atoms (namely the atoms C_α, N, C) is reasonable and meaningful.

To describe the placement of an amino acid in space, a LRF is defined (as shown in Fig.1(a)). Mathematically, a coordinate system is defined based on the vectors of C_α-related chemical bonds, as $\overrightarrow{\boldsymbol{C}_\alpha - \boldsymbol{N}}$ $(= \vec{\boldsymbol{N}} - \vec{\boldsymbol{C}}_\alpha)$ and $\overrightarrow{\boldsymbol{C}_\alpha - \boldsymbol{C}}$ $(= \vec{\boldsymbol{C}} - \vec{\boldsymbol{C}}_\alpha)$, in the amino acid concerned. Here, $\boldsymbol{X}$ represents the position vector of atom X, with $X = C_\alpha$, C, N. For simplicity, an orthogonal coordinate system is implemented practically. In details, the atom C_α is located at the origin of the frame, the bond $\overrightarrow{\boldsymbol{C}_\alpha - \boldsymbol{N}}$ is placed on X axis, and the normal vector of the plane $C - 4C_\alpha - N$ is taken as the Z axis. Considering the chirality of natural amino acids, we regularly pick up the direction of the cross product $\overrightarrow{\boldsymbol{C}_\alpha - \boldsymbol{N}} \times \overrightarrow{\boldsymbol{C}_\alpha - \boldsymbol{C}}$ as the positive direction of Z axis. Consequently, the Y axis is the cross product of the unit vector of Z axis and that of X axis as $\vec{\boldsymbol{Y}} = \vec{\boldsymbol{Z}} \times \vec{\boldsymbol{X}}$. The set of three local-geometry-related axes builds up the LRF. The relative positions of atoms of amino acids can be reproduced from this LRF.

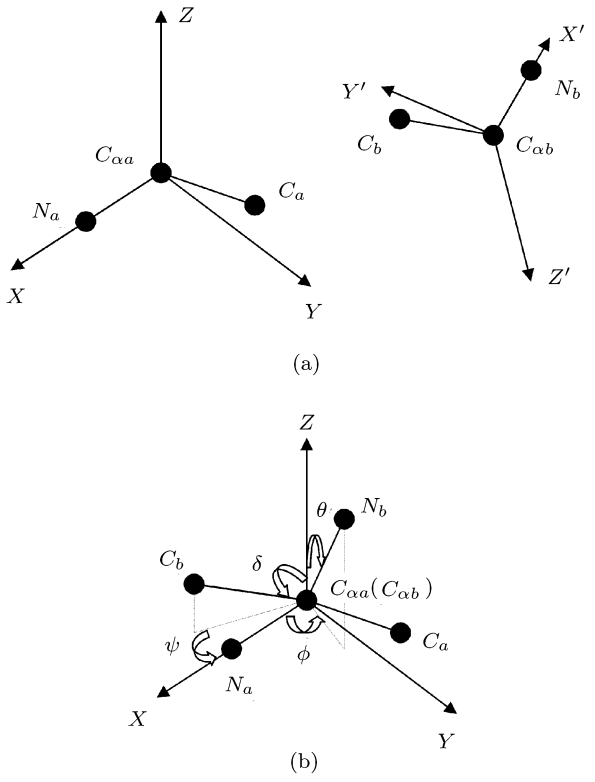

Fig.1. (a) The LRFs for residues a and b, respectively. Vector $\overrightarrow{\boldsymbol{C}_\alpha - \boldsymbol{N}}$ is set as X axis, then, Z axis is obtained by the cross product $\boldsymbol{Z} = \overrightarrow{\boldsymbol{C}_\alpha - \boldsymbol{N}} \times \overrightarrow{\boldsymbol{C}_\alpha - \boldsymbol{C}}$ finally, Y axis is determined by $\vec{\boldsymbol{Y}} = \vec{\boldsymbol{Z}} \times \vec{\boldsymbol{X}}$. (b) Residue b is projected onto residue a (i.e., the origins of the two LRFs are superimposed). Three orientational angles θ, ϕ and ψ are defined as shown in the graph.

With the definition of the local geometry of the

amino acids along the backbone, the orientation of one amino acid b with respect to the amino acid a could be described based on the projection of the LRF of b onto the LRF of a. Operationally, three angles (marked as θ, ϕ, and ψ shown in Fig.1(b)) are defined to describe the orientational relation between the two residues.

These angles satisfy the relations

$$\begin{aligned}\cos\theta &= (\vec{\boldsymbol{X}})_b \cdot (\vec{\boldsymbol{Z}})_a\,,\\ \cos\phi &= (\vec{\boldsymbol{X}})_b \cdot (\vec{\boldsymbol{X}})_a/\sin\theta\,,\\ \cos\psi &= (\overrightarrow{\boldsymbol{C}_\alpha - \boldsymbol{C}})_b \cdot (\vec{\boldsymbol{X}})_a/\sin\delta\,,\end{aligned} \tag{1}$$

with $\cos\delta = (\overrightarrow{\boldsymbol{C}_\alpha - \boldsymbol{C}})_b \cdot (\vec{\boldsymbol{Z}})_a$. The set of these three angles uniquely describes spatial placement of one amino acid relative to another. Indeed, the orientation may be affected by the distance between the two residues. Here, we omit the distance-dependent effect.

For two amino acids with indices i and j, it is easy to find out that the projection angles from the amino acid i to the amino acid j are different from the angles projecting amino acid j to the amino acid i, though both choices reflect the relation of the same pair of amino acids. Considering that there is a sequential order for protein chains (namely, the difference between C-terminus and N-terminus), the above two kinds of projections might both be meaningful. In our statistics, for the amino acid pair (i,j) with $i < j$, the projection from amino acid j to amino acid i is marked as positive projection, and the projection from amino acid i to j is marked as negative projection.

2.3. Data source and classification

In our study, 900 proteins are extracted from ASTRAL SCOP 1.69.[18] The set of sequences with less than 40% identity to each other is adopted. These proteins are randomly selected from the ASTRAL set. Different trials for the set of proteins were also carried out (data are not given here), which have consistent results. In our statistics, all the contacted pairs are classified based on the types of amino acids. Considering the projection order, there are totally 420 kinds of pairs. For a certain kind of pairs of amino acids, there are various kinds of configurations (thus orientations) for these two kinds of amino acids. These projection angles concerned are collected for further building up of the distribution of orientations in the space of projection angles. Practically, the space of projection angles is discretized. We split the whole ranges of the angles θ, ϕ, and ψ into 18, 36, and 36 bins, respectively. The distributions of orientations are represented by the normalized occurrence frequencies of projection angles in these bins. To avoid the difficulty to illustrate this 4-dimensional distribution, we re-arrange the data along a one-dimensional index AI as $\mathrm{AI} = [\phi/10]\times 36\times 18 + [\psi/10]\times 18 + [\theta/10]$, which keeps parts of the neighbouring relationship of bins. Here $[S]$ represents the integral part of the quantity S.

3. Results and discussion

Based on the basic geometrical knowledge, the quantities $\cos(\theta)$, $\cos(\phi)$, and $\cos(\psi)$ are uniformly distributed in the region $[-1, 1]$ for two amino acids with random orientations. However, the actual distribution of orientations is rather different from this distribution related to random orientations (see Fig.2(a)).

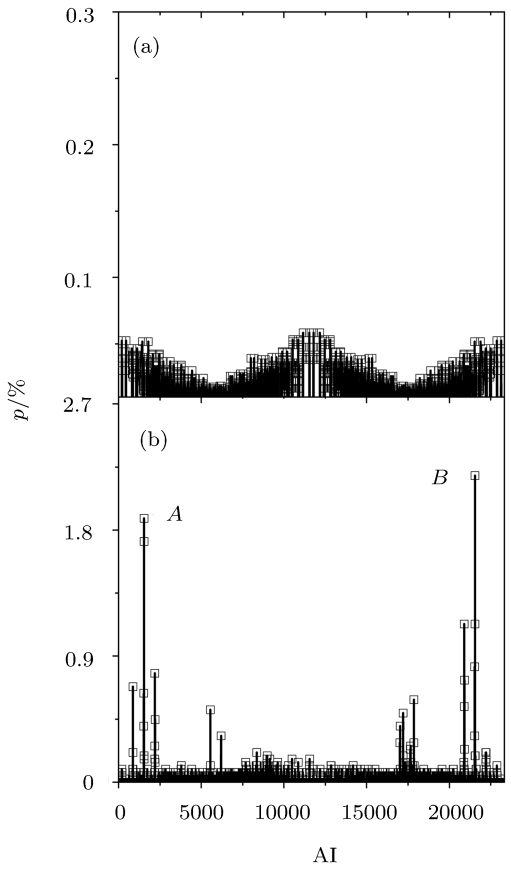

Fig.2. (a) The orientational distribution related to random orientations. (b) The orientational distribution for Ala–Arg pair. Two largest peaks are marked as A (70°, 30°, 140°) and B (110°, 340°, 100°).

Especially, the peaks for random orientations disappear in the actual distributions, while some new

peaks emerge in actual cases. This indicates that the orientations between amino acids in proteins are not random. Compared with isotropic orientations in the random case, the new peaks in actual cases suggest some preferences of orientations between some certain amino acids. The complex interactions introduce some orders into the packing of amino acids. More interestingly, there are some common features for the preferences between various kinds of pairs of amino acids, which are described in the following subsection. This may be concerned with the basic interactions of amino acids. Some preferences for specific kinds of amino acid pairs are also discussed, which demonstrates the speciality of various amino acids. Furthermore, the investigation related to the secondary-structure preferences is described in the final section, indicating the effect of directional interactions on the orientations.

3.1. Common preferences of the orientations for all pairs of amino acids

For all the pairs between amino acids, there are mainly two common preferences of orientations. That is, in the space of projection angles, there are two highly populated zones, which are marked as zone I and zone II, respectively (as shown in Fig.2(b)). The corresponding most probable projection angles (θ, ϕ, ψ) are (70°, 30°, 140°) and (110°, 340°, 100°). There may be some tiny variations (about 10°) of these peak positions for other types of amino acids. In each zone, the possible variation for any angles depends on the type of amino acids. There are some common features for these two cases. The angles θs are close to 90°, which suggests that the bond C_α-N of residue b is parallel to the plane C-C_α-N of residue a. At the same time, the difference between the angles ψ and ϕ is almost 110° which is about the bond angle of the conformation C-C_α-N. This observation indicates that the bond C_α-N of residue a is parallel to the plane C-C_α-N of residue b. Therefore, we could conclude that the two planes C-C_α-N of the two residues are parallel to each other. Based on these analysis, we could image the basic configurations of the organization of the two amino acids. As examples, two typical conformations corresponding to these two zones are shown in Fig.3 (with Fig.3(a) for zone I and Fig.3(b) for zone II).

The features discussed above are clearly illustrated in these figures.

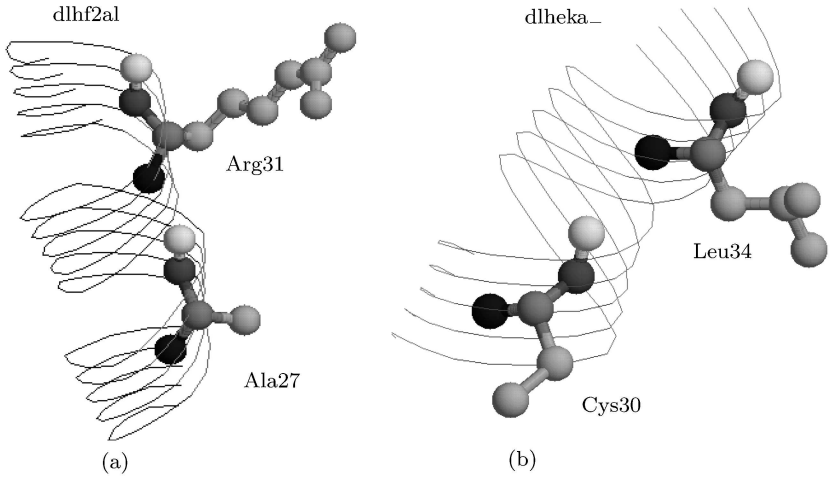

Fig.3. (a) The conformation for Arg to be positively projected to Ala with angles (70°, 30°, 140°) which belong to zone I. The three black balls represent backbone atoms, the light grey one represents atom O, and the other grey ones represent the side chain. (b) The conformation for Leu to be positively projected to Cys with angles (110°, 340°, 100°) which belong to zone II. The representations of the balls are same to those in Fig.3(a).

More interestingly, we find that there are different preferences for positive and negative projections. That is, positive projections prefer the zone II, and negative projections prefer the zone I. Especially, for the pairs constructed with two same-type amino acids, these kinds of preferences do exist (see Fig.4).

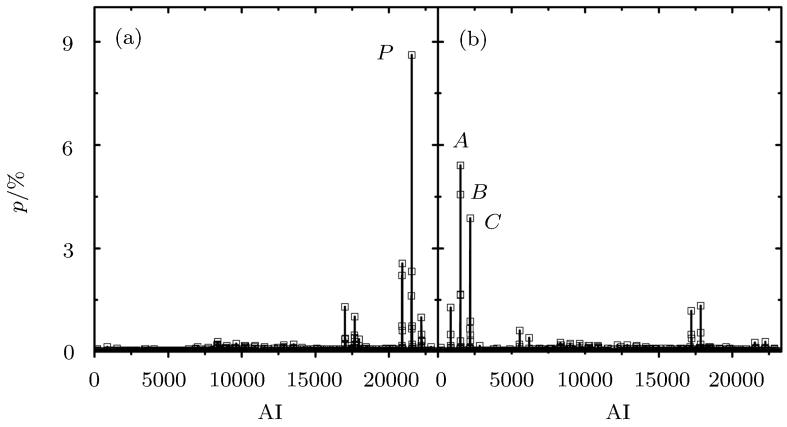

Fig.4. The orientational distribution for Ala–Ala pair. (a) corresponding to positive projections with the highest peak P at (110°, 340°, 100°) and (b) to negative ones with the three highest peaks A, B and C as (80°, 30°, 140°), (70°, 30°, 140°), and (70°, 40°, 140°), respectively.

Considering the facts that the angles θs for two zones are mutually supplementary and two ϕs make almost a round angle, it is easy to find out that different combinations of angles (θ, ϕ, ψ) in the two zones may be the different descriptions for a pair of amino acids. Consequently, we could conclude that the different appearance of two zones is the result of the different projections (positive or negative). As a check, we transform the distributions related to negative projections based on the above-mentioned transformation for the angles θ and ϕ; that is, we change the index AI to AI $= [(360-\phi)/10]\times 36\times 18+[\psi/10]\times 18+[(180-\theta)/10]$. The newly generated distributions are almost the same as the corresponding distributions related to positive projections. This further justifies our statements for the equivalence of the two zones. In this sense, the orientational preference between amino acids is rather unique, which reflects that there are some orders in the packing of amino acids in protein systems.

In detail, the shape of these zones is not all the same for various kinds of pairs. For some pairs of amino acids, these zones are very sharp with little bias towards other orientations, while the preferences for some pairs of amino acids are not prominent and even ambiguous (as shown in Fig.5(a)).

For example, for the contacts between amino acids Ala and Ala, the probability for the most probable orientation (80°, 30°, 140°) is about 50 times larger than the average probability for the corresponding distribution. Meanwhile, for the pair of amino acids Thr–Tyr, the ratio of maximal probability to the average probability is around 6. There are some more examples for these two kinds of pairs, with 'concentrated' preference or with dispersed distributions, such as, the pairs Cys–Gln, His–Asp whose preferences are rather 'concentrated', and the pairs Trp–Val, Tyr–Ser with relatively 'flat' distributions. Practically, we distinguish between the two classes of distributions by a threshold (10 in the present case) of the ratio of the maximal probability to the average probability. It is found that the pairs with small-size amino acids often have concentrated preferential zones. On the other hand, the orientations of the amino acids of large size seem to be more complex and, thus, have a weak preference towards the common zones. This kind of phenomenon is a reasonable result of packing in protein systems. The packing between small amino acids is relatively simple, while the packing between large amino acids is often affected by their large side chains. Consequently, the orientational preferences of backbones would be modulated to fit the local arrangement. Therefore, the orientations of large amino acids show more fluctuations. It is still interesting to note that the main preferences (zone I and zone II) are robust for various amino acids, which shows the importance of directional interactions in packing of proteins. This kind of analysis can be further supported by the examples of the pairs between a small amino acid and a large amino acid (see Fig.5(b)). Besides, for the pair Ala and Phe, the ratio of maximal probability to the average probability is still about 10, but there are much more peaks with their heights comparable to the highest one. In other words, the preferential zones have a broad distribution. Here, the large amino acid Phe also introduces some flexibility to the orientations.

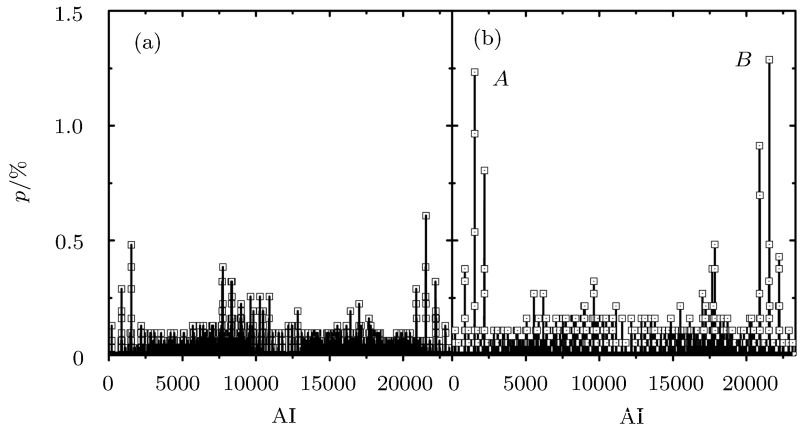

Fig.5. (a) The orientational distribution for Thr–Tyr pair. (b) The orientational distribution for Ala–His pair with the two highest peaks A and B as (70°, 30°, 140°) and (110°, 340°, 100°), respectively.

3.2. Specific preferences of orientations for some pairs of amino acids

Besides these major preferences, some pairs of amino acids may have some specific orientational preferences. Considering the effect of packing, these specific preferences are regularly related to the large amino acids. For example, the pair of Ser and Tyr has a preferential orientation with the projection angles (10°, 280°, 210°) as shown in Fig.6, which shows two almost perpendicular C-C_α-N planes.

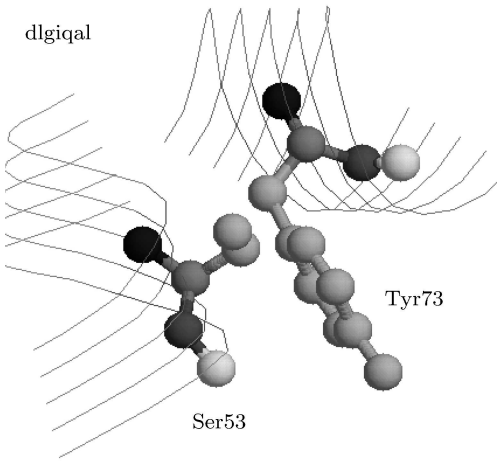

Fig.6. The conformation for Tyr to be projected onto Ser with angles (10°, 280°, 210°). The representations of the balls are same as those in Fig.3(a).

This is totally different from the parallel structure of common orientations for the above zones. From Fig.6, the effect of side-chain packing could be clearly visualized. This kind of perpendicular cases also happens for some other pairs of amino acids, such as Asn–Leu, Asp–Ile, Glu–Met etc. For some geometrical special amino acids, such as Pro, there are some else specific orientations different from any above cases. For example, the pair of amino acids Pro and Met has a preferential angle set of (100°, 170°, 290°).

The preferential angles for the pairs of amino acid Pro and other kinds of amino acids also show large variations. This may be attributed to the reduction of local flexibility of the protein chains caused by the special amino acid Pro. This suggests that the local flexibility of the chains could also affect the orientational organizations.

3.3. Orientational preferences related to secondary structures

Besides the exclusive volume effect of amino acids and flexibility of local structures, there are some other kinds of directional interactions, such as hydrogen bonds. How would these directional interactions affect the orientations between amino acids? It is well known that directional interactions are an important part of the origin of the secondary structures.[19] Therefore, the orientational preferences of amino acids with some certain secondary-structure preference would show features of directional interactions, which are given in the our following discussions.

Bearing this purpose, we divide the amino acids based on their preferences to secondary structures into three classes:[20] the ones preferring to α-helix structures, the ones preferring to β-strand structures, and the ones with weak preferences to any two kinds of secondary structures. For simplicity, these classes are named as α class, β class, and ordinary class, respec-

tively. Consequently, the orientations between α-class amino acids, between β-class amino acids, or between ordinary amino acids are analysed separately. Some representative cases are shown in Fig.7.

Clearly, different secondary structures produce different preferences of orientations. As shown in Fig.7(a), the orientations between two α-class amino acids (say Ala and Glu) show the preferences to the angles (70°, 30°, 140°) and (110°, 340°, 100°), which match nicely the geometry of α-helix. The large difference between the peaks and the background of the distribution is consistent with the fact that the α-class amino acids are likely to take the helix structures. Interestingly, for the other two kinds of pairs, the peaks related to α-helix structure still exist, though the strength of the preference is much weaker as compared with the case of α-class amino acids. The existence of the common peak illustrates the popularity of helical motif in protein structure, which suggests the special preference of helices in protein architectures, as indicated in many papers.[21,22] For the pair composed of two β-class amino acids (say Val and Ile), there is another preferential orientation with angles (120°, 150°, 320°) as shown in Fig.7(b). This kind of orientations are actually the same as the preferred orientations in ideal β-sheet structures. The height of this peak is comparable to those of the common peaks in this case. The population of this new peak indicates the emergence of a different kind of structure (β-sheet structures). For the case with ordinary amino acids (say Gly and Leu) (Fig.7(c)), the profile of the distribution is similar to that for β-class amino acids. Yet the positions of the peaks are different, and the whole distribution is somewhat flat. This is consistent with the feature of ordinary amino acids which have weak preferences to any structures.

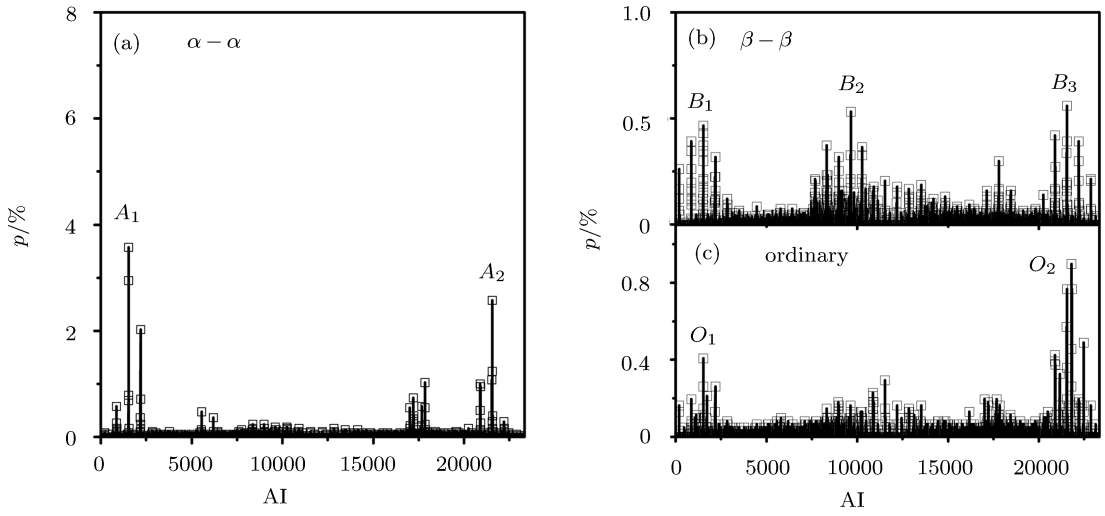

Fig.7. The distributions of different secondary-structure classes of amino acid pairs. (a) between α-type amino acids Ala–Glu with the peaks A_1 (70°, 30°, 140°) and A_2 (110°, 340°, 100°), (b) between β-type amino acids Val–Ile with the peaks B_1 (70°, 30°, 140°), B_2 (120°, 150°, 320°), and B_3 (100°, 340°, 100°), (c) between ordinary amino acids Gly–Leu with the peaks O_1 (70°,30°, 140°), and O_2 (110°, 340°, 100°).

Clearly, the secondary-structure-related directional interactions do produce some intrinsic patterns of orientational preferences. These directional interactions could affect the orientations between amino acids more globally compared with the packing or local flexibility. Another interesting thing is that we could find out the preferences of amino acids to secondary structures by inspecting the orientation-dependent distributions. This to some extent provides a statistical characterization for properties of amino acids.

4. Conclusion and outlook

In this paper, through the statistics on orientations of contacted pairs of amino acids, the preferences of orientations are analysed, and the corresponding regularities are presented. There are common preference of orientations (70°, 30°, 140°) and

(110°, 340°, 100°), which correspond to the configurations with parallel $C - C_\alpha - N$ planes. Different kinds of side chains may make the distributions flat, especially for the amino acids with large side chains, e.g. Trp–Val, Tyr–Ser. Yet, the common orientations do exists. Some special preferences of orientations may happen for some specific pairs of amino acids, such as (100°, 170°, 290°) for Pro–Met. The distributions between the amino acids with different secondary-structure preferences are also investigated, which shows a clear correlation between the distributions of orientations and the properties of amino acids. All these observations suggest that the packing, local flexibility, and the directional interactions may all affect the orientational properties of the pairs of amino acids. These investigations provide us the knowledge about the relationship between interactions and the orientational arrangement of amino acids in proteins, which may benefit for further modelling of proteins.

References

[1] Branden C I and Tooze J 1998 *Introduction to Protein Structure* 2nd ed (New York: Garland)

[2] Miyazawa S and Jernigan R L 1985 *Macromolecules* **18** 534

[3] Miyazawa S and Jernigan R L 1996 *J. Mol. Biol.* **256** 623

[4] Jernigan R L and Bahar I 1996 *Curr. Opin. Struct. Biol.* **6** 195

[5] Vajda S, Sippl M and Novotny J 1997 *Curr. Opin. Struct. Biol.* **7** 222

[6] Buchete N V, Straub J E and Thirumalai D 2004 *Curr. Opin. Struct. Biol.* **14** 225

[7] Sheng Y B, Li J, Ma B L and Wang W 2005 *Chin. Phys.* **14** 2365

[8] Kabsch W and Sander C 1983 *Biopolymer* **22** 2577

[9] Hunt N G, Gregoret L M and Cohen F E 1994 *J. Mol. Biol.* **241** 214

[10] Kellis J T Jr, Nyberg K and Fersht A R 1989 *Biochemistry* **28** 4914

[11] Mitchell J B O, Laskowski R A and Thornton J M 1997 *Proteins: Struct. Funct. Bioinfo.* **29** 370

[12] Yan A and Jernigan R L 2005 *Proteins: Struct. Funct. Bioinf.* **61** 513

[13] Sobolev V, Sorokine A, Prilusky J, Abola E E and Edelman M 1999 *Bioinformatics* **15** 327

[14] Chavez L L, Onuchic J N and Clementi C 2004 *J. Am. Chem. Soc.* **126** 8426

[15] Kaya H and Chan H S 2003 *J. Mol. Biol.* **326** 911

[16] Vendruscolo M, Paci E, Dobson C M and Karplus M 2001 *Nature* **409** 641

[17] Zuo G H, Wang J and Wang W 2006 *Proteins Struct. Genet. Bioinf.* **63** 165

[18] Chandonia J M, Hon G, Walker N S, Conte L L, Koehl P, Levitt M and Brenner S E 2004 *Nucl. Acids Res.* **32** D189

[19] Pauling L, Corey R B and Branson H R 1951 *Proc. Natl. Acad. Sci. USA* **37** 205

[20] Chou P Y and Fasman G D 1978 *Annu. Rev. Biochem.* **47** 251

[21] Emberly E G, Wingreen N S and Tang C 2002 *Proc. Natl. Acad. Sci. USA* **99** 11163

[22] Banavar J R and Maritan A 2003 *Rev. Mod. Phys.* **75** 23