Abstract

In this paper, we analyze the symmetries of beta-barrel proteins at both structure and sequence levels by using a modified recurrent quantification analysis. It shows that the structures and sequences have the same two-fold symmetry, although the later diverged considerably. This result may be helpful to understand the mechanism of protein evolution.

Keywords: Beta-barrel family; Sequence; Tertiary structure; Symmetry

1. Introduction

Many protein domains are combinations of recurring substructures (Edward and Hwang, 2004). However, these domains usually have almost random primary sequences (Stephen et al., 2001; Taylor et al., 2002). How and why do the proteins exhibit obvious symmetry at the level of tertiary structures, and yet seldom symmetry in their primary sequences? Many progresses have been made in this problem (Zbilut et al., 2002; Heger and Holm, 2000). In a previous paper, we have revealed the hidden three-fold repetitions in the sequences of proteins from the beta-trefoil family taking account of physicochemical properties of amino acids (Xu and Xiao, 2005). In this paper we shall extend our method to the repetition analysis of both sequence and structure and take the beta-barrel folds as an example.

The beta-barrel domains (Fig. 1) are built of beta strands that can vary in number from 4 or 5 to over 10. These beta strands form two beta sheets that have the usual twist. Two such twisted beta sheets form a barrel-like structure when they are packed against each other. According to their topology structures, this family can be divided into up-and-down beta-barrel and jelly-roll beta-barrel. The up-and-down beta-barrel is formed by an array of beta-strands arranged in an antiparallel manner with each strand hydrogen-bonded to neighboring strands nearly always adjacent in the amino acid sequence (LaLonde et al., 1994). The jelly-roll beta-barrel is usually formed by two Greek-keys, and these proteins usually form two sheets with few if any hydrogen bonds between strands that belong to the different beta sheets. From direct observation of the tertiary structures we can easily find that all of the proteins in this family have two-fold quasi-symmetric structures. This shall be further confirmed in the following by a quantitative method. We shall also try to find the hidden repetitions of the primary sequences corresponding to their tertiary structures.

2. Methods

We shall investigate the two-fold symmetries of the primary sequences and tertiary structures by using a modified recurrence quantification analysis. The recurrence quantification analysis is a QSAR-related equivalent of a known sequence analysis tool that has originally been called “distance chart analysis” (Konopka, 1994, 1997, 2003; Wootton, 1997; Konopka and Smythers, 1987; Konopka and Chatterjee, 1988). The detail of the modified recurrence plot can be found in the previous paper (Xu and Xiao, 2005). Here, we only introduce it briefly. The modified recurrence plot of a protein sequence \( S = x_1 x_2 x_3 \ldots x_N \) is built as follows: the horizontal axis \( i \) denotes the location of the first residue of a segment in primary sequence and the vertical axis \( d \) denotes the length of the segment. For any segment \( X_i = x_{i} \ldots x_{i+d-1} \), if we can find another segment \( X_j = x_{j} \ldots x_{j+d-1} (j \neq i) \) of the same length in the sequence \( S \), and the sequences are similar if the percentage of their
identical symbols is larger than a chosen number $r$ ($0 < r < 1$) and when $p$-value is lower than 0.01.

We extend our method to analyze the repetitive local structures in the tertiary structure of a protein. In this case, $x_i$ represents the coordinate of the atom $C_{\alpha}$ of the $i$th amino acid. The two local structures are similar if the value of the distance root-mean-square deviation (dRMSD) (Wallin et al., 2002) is less than 2.0 Å.

3. Results and discussions

We shall take Atpase (PDB ID: 1E32) (Fig. 1) as an example to show the hidden symmetry of its primary sequence by using the modified recurrence plot. Fig. 1c is the modified structure recurrence plot of 1E32 and it is clear that the tertiary structure has a pseudo two-fold axis of symmetry. The three-dimensional structures of subsequences 1–45 and 46–89 are similar with each other.

Fig. 1. The structure and recurrence plots of Atpase (PDB ID: 1E32). (a) The primary sequence and secondary structure; (b) the tertiary structure; (c) the modified recurrence plot of the structure with (dRMSD ≤ 2.0 Å) and (d) the modified recurrence plot with the value of $r = 0.70$.

Fig. 2. The two similar parts (1–45, 46–89) of the tertiary structure of the protein Atpase (PDB ID: 1E32). (a) The primary and secondary structure of the two parts; (b) the structure of two parts: the green one is for 1–45 and the red one for 46–89; (c) the two parts from the view perpendicular to the barrel axis and (d) along the barrel axis.
Fig. 3. The Pearson correlation coefficients between the structure based and sequence based plots for all the representative proteins.

other and with the dRMSD being 2.0898 (Fig. 2). The secondary structures of the two parts are also very similar (Fig. 2a). However, this kind of similarity cannot be seen directly in its primary sequence (Fig. 2a). We analyzed the sequence by the modified recurrence plot with $r = 0.7$, the modified recurrence plot shows obvious two-fold symmetry, with the first part containing amino acid from 1 to 44 and the second part from 45 to 89 (Figs. 1d and 2a). It means that the amino acid sequence of 1E32 has a two-fold repetitive pattern corresponding to the tertiary structure. The Pearson correlation coefficient between the structure based plot and the sequence based plot is 0.9330.

Due to the great variation of the tertiary structures of the beta-barrel family, not all of them can show significant two-fold symmetries by the modified structure recurrence analysis with a strict condition $dRMSD \leq 2.0 \text{Å}$. However, the tertiary structures really have two-fold quasi-symmetries to some degrees. The analysis of the primary sequences of 21 representative proteins of the beta-barrel family shows that they also have the same hidden two-fold quasi-symmetry. It is noted that these sequences are different and sequence alignment shows that their identical amino acids are less than 30%.

To give a quantitative description of the relationship between the structure based and sequence based plots, we calculated the Pearson correlation coefficients between them for all the representative proteins (Fig. 3). It shows that most of them show strong correlations.

Thus, almost all of proteins studied here show the same two-fold quasi-symmetry at both sequence and structure levels. This suggests that the symmetries at structure level are due to those at sequence level. We hope that our results shall be useful in the development of structural prediction methods and the understanding of mechanisms of protein evolution.

Acknowledgements

This work is supported by the NSFC under Grant Nos. 30525037 and 30470412 and the Foundation of the Ministry of Education of China.

References

Konopka, A.K., 1994. Sequences and codes: fundamentals of biomolecular cryp-