A common sequence-associated physicochemical feature for proteins of beta-trefoil family

Ruizhen Xu, Yi Xiao∗

Biomolecular Physics and Modeling Group, Department of Physics, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China

Received 12 November 2004; received in revised form 10 December 2004; accepted 10 December 2004

Abstract

Different amino acid sequences can fold into similar tertiary structures but the reasons for it are not very clear. It has been suggested in the literature that these sequences may have some common features associated with them but the exact nature of such shared properties remains largely unknown. We studied a representative sample of proteins from the beta-trefoil family and observed that their amino acid sequences, despite being considerably divergent from each other, can be accounted for by matching to a repetition of three physicochemically similar segments. This observation in turn is consistent with the three-fold pseudo-symmetry in tertiary structures of these proteins.

© 2004 Elsevier Ltd. All rights reserved.

Keywords: Beta-trefoil family; Sequence repetition; Tertiary structure symmetry

1. Introduction

The β-trefoil family of proteins represents a diverse group comprising a structural "hyperfamily" (Orengo et al., 1994). According to SCOP (Murzin et al., 1995), proteins with this structural fold (or that contain this structural fold) include the fibroblast growth factors, interleukin-1, IP3 receptor type I binding core, mannos receptor, agglutinin, kunitz (STI) inhibitors, clostridium neurotoxins, fascin, histidine-rich actin binding proteins (Ago et al., 1991; Blaber et al., 1996; Graves et al., 1990; Zhu et al., 1991; Brych et al., 2001). The β-trefoil structure is composed of a six-stranded β-barrel closed off at one end by three β-hairpin structures and exhibits a characteristic pseudo-three-fold axis of symmetry when viewed down the β-barrel axis (Fig. 1). A detailed analysis of the geometry and architecture of the β-trefoil fold was done by Chothia and coworkers (Murzin et al., 1992). However, despite the retention of the symmetric tertiary architecture, the primary sequences of proteins of the β-trefoil family have diverged considerably.

How and why does the beta-trefoil family of proteins exhibit such a high degree of symmetry at the level of tertiary structure, and yet a low degree of repetition at the level of the primary sequence? Although this problem has been known for more than 10 years, we still do not know why. In the present paper, we solve this contradiction and show that the primary sequences of the proteins of the β-trefoil family, although diverged considerably, exhibit a novel common "hidden" repetition in an amino acid propensity, i.e., the frequencies of amino acids in β-strands (Chou and Fasman, 1978).

2. Methods

Symmetry means identity or repetition. For an example, the sequence, ADJGFADJGFADJGF, is composed of three identical parts and we say that it is exact three-fold repetition. But real protein sequences do not have such exact repetition and they in fact appear nearly random as indicated above. However, if primary sequences are composed of similar segments or repeats, they can form pseudo-symmetric structures. It is known from sequence alignment (Needleman and Wunsch, 1970; Smith and Waterman, 1981) that, if the...
primary sequences of two proteins have more than 25% identical amino acids, they likely have similar tertiary structures (Mount, 2001; Sweet and Eisenberg, 1983). One of the reasons may be that these two sequences are similar in physiochemical properties. This suggests that the primary sequences of proteins of the \( \beta \)-trefoil family may exhibit three-fold repetition if we see them from points of view of physiochemical properties.

To do this, we use the recurrence quantification analysis (Giuliani et al., 2002; Zhilut et al., 2004). 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; Wooston, 1997; Konopka and Smythers, 1987; Konopka and Chatterjee, 1988). Other variants of distance chart analysis are “coincidence (diversity) index method” (Shulman et al., 1981) and “periodicity analysis” (Michel, 1986).

To find repeats in sequences, we need to scan the segments of different lengths. The original form of the recurrence quantification analysis is not convenient to show how recurrence changes with the segment length. We use a modified version and it works as follows. Consider an arbitrary protein sequence \( S = \xi_1 \xi_2 \xi_3 \cdots \xi_N \), where \( \xi_i \) is one of the 20 amino acids and it can be denoted by one of the 2 symbols: P and N. An amino acid is denoted by P if its frequency in \( \beta \)-strands is no less than 1.3 (this includes F, I, L, V, W and Y) and by N if less than 1.3 (this includes A, C, D, G, H, K, M, N, P, Q, R, S, T and E) (Chou and Fasman, 1978), i.e., P represents amino acids that are strong \( \beta \)-strand former and N for others. Thus, \( S \) is a sequence with an alphabet of two symbols (P and N). For such a sequence, one constructs all possible segments of \( d \) consecutive symbols: \( X_i = \xi_{i+1} \xi_{i+2} \cdots \xi_{i+d-1} \) (\( 1 \leq i \leq N - d + 1 \)). The modified recurrence plot is built as follows: the horizontal axis denotes residue index in primary sequence and the vertical axis denotes segment length \( d \). A point is placed at \((i, d)\) if \( X_i \) and \( X_j \) is similar, i.e., the ratio of the number of identical symbols to the segment length \( d \) is larger than \( r \) (\( 0 \leq r \leq 1 \)). We can decrease the value of \( r \) gradually to detect repeats in primary sequences.

3. Results and discussions

We shall take hormone growth factor (PDB ID: 1jqz) as an example to show the hidden repetition of its primary sequence by using the modified recurrence plot. 1jqz is a member of the \( \beta \)-trefoil family and has a tertiary structure that contains a pseudo three-fold axis of symmetry (Fig. 1a), but its primary sequence appears irregular (Fig. 1b). Fig. 1c shows the modified recurrence plot of 1jqz. It shows that the modified recurrence plot of 1jqz does not show obvious repetition but only some similar short segments if we consider the sequences similar only when they have more than 85% identical P and N, e.g., \( r = 0.85 \). However, the primary sequence repetition emerges when \( r \) decreases. In particular, the primary sequence of 1jqz shows obvious repetition when \( r \) is about 0.70. In this case the modified recurrence plot shows three sharp boundary lines which divides the plot (also the sequence) into three almost the same parts beginning at the 1st, 50th and 90th amino acids. This means that the sequences of these three parts are similar. Thus, the primary sequence repetition is three-fold and this is the same as that of its tertiary structure.

However, it is less important if only one member of the \( \beta \)-trefoil family proteins shows the hidden repetition characteristic of its structure. Our aim is to detect the common features of all the members. So the primary sequences of 20 representative proteins of the \( \beta \)-trefoil family are selected and analyzed. It is noteworthy that almost all of them show the same hidden three-fold repetition (Fig. 2).
Our results demonstrate that all sequences of the representative proteins of the β-trefoil family considered here have one common feature, i.e., the same three-fold repetition as their tertiary structures. These protein sequences are typical ones and represent most of the protein sequences of the β-trefoil family. It is also noted that these sequences are also different and sequence alignment shows that their identical amino acids are less than 30%. So it may suggest that the formation of the symmetric tertiary structures of these protein domains is the result of their sequence repetition. In other words, the symmetries of these tertiary structures are encoded by their sequences. This result explains one aspect of the sequence degeneracy (Chothia and Lesk, 1986). It could be interpreted as a manifestation of correctness of the generally accepted assumption that the sequence of a protein determines its tertiary structure.

We believe it should be noted that another recent paper (Brych et al., 2004) draws essentially the same conclusion from completely different evidence. By increasing the primary sequence symmetry of human acidic fibroblast growth factor through loop deletion and point mutation, they found that the stability of the protein is increased but the activity of its functionality is reduced. This result also supports the hypothesis that the sequence repetition encodes and determines the tertiary structure symmetry (as opposed to the hypothesis that the tertiary structure symmetry is determined...
in some other way but imposes symmetric constraints on the sequence).

Finally, it could be interesting to see whether the kind of sequence–structure relation reported in this note (concordance of symmetries at two different levels of structure) can be found for other proteins with symmetric tertiary structures.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under grant no. 90103031 and 30470412 and the Foundation of the Ministry of Education.

References