Journal of Molecular Graphics and Modelling 28 (2010) 788–791

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Identification of sequence repetitions in immunoglobulin folds

Xiaofeng Ji, Haiying Wang, Jianhua Hao, Yuan Zheng, Wei Wang, Mi Sun *

Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao 266071, Shandong, China

ARTICLE INFO

Article history:
Received 17 August 2009
Received in revised form 20 January 2010
Accepted 5 February 2010
Available online 1 March 2010

Keywords:
Repetitions
Protein sequence
Structure
Pearson's correlation coefficient
Ig folds

ABSTRACT

A lot of evidence suggests that many proteins with the symmetric structures have evolved by internal duplication and fusion. Meanwhile many internal sequence repeats correspond to functional and structural units. These proteins, which have internal structural symmetry, this means that their sequences should be made up of identical repeats. However, many of these repeat signals can only be seen at the structural level yet. We have developed a de novo algorithm, modified recurrence correlation analysis, to detect the symmetries in the primary sequences of immunoglobulin folds (Ig folds), which adopt highly symmetrical tertiary structures while their sequences appear nearly random. Using this method, we show that the internal repetitions of the immunoglobulin folds could be identified directly at the sequence level. These results may give us some help to study the hypotheses about the origin of Ig folds by duplication of simpler fragments and it may also give us some helps to understand the relationship between the sequences and their tertiary structures.

1. Introduction

Six out of the ten most popular folds (superfolds) possess an approximate structural symmetry [1,2]. But in many proteins that adopt one of these folds have no detectable symmetry in their sequences and their sequences are even random. If the chain conformations of protein are mainly determined by the information contained in its amino acid sequence, there must be signals which indicate their structural symmetry in the sequences of these proteins. The detection of the repeats in sequences would be helpful to understand the protein evolution mechanisms.

Ig folds comprise two basic types: the constant domain (C) and the variable domain that can be found in both heavy and light chains of immunoglobulin [3]. Classical Ig-like domain is composed of 7–10 β strands distributed between two sheets with typical topology and connectivity. Constant domains are made of 7 β strands: 4 form one sheet and 3 form another. The structures have Greek key barrel topology. Similar to constant domains, variable domains have 9 β strands that arranged in two sheets through 4 and 5 strands, respectively. The 5-stranded sheet is structurally homologous to the 3-stranded sheet of constant domains, but contains two extra strands. The remainder strands have the same topology and similar structure as their counterparts in the constant domain of immunoglobulin folds. A disulfide bond links the strands in opposite sheets, as in constant domains [1]. Previous investigations have shown that one of the common properties of this family is the pseudo 2-fold symmetry, which is observed in all known Ig structures [4]. Furthermore, Yanzhao Huang and Yi Xiao showed that the internal repetitions of the immunoglobulin folds could be identified directly at the sequence level by calculating the Pearson's correlation coefficients between every sub-matrices of column size for the similarity matrix [5].

In order to detect repeats or periods in protein sequences, different methods have been proposed [6–15]. Among them, there are also two popular web servers of repeat detection: RADAR [7] and TRUST [10]. They identify repeats based on suboptimal self-sequence alignment. These two methods are proposed for general repeats detection but not especially for symmetric sequence repeats. What is worth mentioning in particular is that Xiao's group used the method of modified recurrence plot to detect periods in the sequences of beta-trefoil [11], beta-barrel [14], beta-propeller [15] and Ig fold [5]. In this study, we present a fast and sensitive method to detect the latent periodicities in proteins based on the protein sequences analysis.

2. Methods

In developing recurrence correlation analysis, we were guided by the idea of recurrence quantification analysis [16] to find the correlation between two segments in the aspect of polarity distribution.

The detail of recurrence correlation analysis method can be found in our previous paper [17]. This method includes three steps: (i) replacement of residues. Consider the arbitrary sequence

* Corresponding author. Tel.: +86 53285819525; fax: +86 53285819525.
E-mail address: sunmi@ysfri.ac.cn (M. Sun).

1093-3263/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2010.02.003

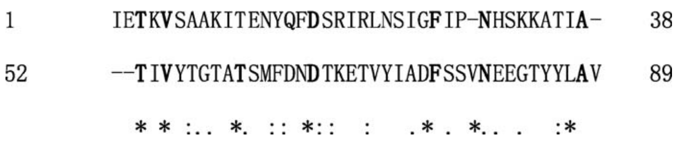
(S = x1,x2,x3,. . .,xN) to be analyzed, where N is the sequence length and xi is one of the 20 amino acids. Using the Grantham polarity value [18] to denote corresponding amino acid, a vector representation of the protein sequence, as A = a1,a2,a3,. . .,aN, is achieved. (ii) Calculation of the Pearson's correlation coefficients. We select a certain length segment Ai = aiai + 1,. . .,ai + d − 1 from the sequence for the target segment and let this segment slide along the sequence, then calculate the correlation coefficients between this target segment and the remaining segments along the sequence. The correlation coefficients of segment Ai and Aj (Aj = ajaj + 1,. . .,aj + d − 1 ($j \neq i$) is of the same length with Ai in sequence S) is saved in the correlation matrix r as r(i,j). r is formed when this is done for all possible i and d. (iii) Determine the threshold. We generated 1000 pairs of sequence and 1000 pairs of random one-dimensional array for six length groups, respectively. The similarity of each pair of sequences that we have generated is not less than 0.25. Then we calculated the Pearson's correlation coefficient for each pair of sequences and arrays. The results showed that the distribution of the correlation coefficient for random one-dimensional array is similar to selected sequences and the correlation coefficient of the two statistical results is 1. According to the opinion of statistics, if the correlation coefficient of two arrays is greater than 0.5, the two arrays are related. So we choose 0.5 as the threshold of our program. In order to further validate our results, we also made a statistics for the similarity between the pairs of sequences that the Pearson's correlation coefficient is not less than 0.5. We can find that the percentage of segments, which the similarity between the pair of segments is not less than 0.25, is greater than 98% when the length of the segments is not less than 20. The percentage even reached 99.58% when the length of the segments is greater than 30. Based on the above statistics and related articles, we choose 0.5 as the threshold of our program. (iv) Construction of result plot. If r(i,j) is not less than the threshold and when p-value is lower than 0.01, we think segment Ai and Aj are similar. Then we plot a point at (i,j) and (j,i) in the result plot.

3. Results and discussion

To test our method for detecting the internal structure-related sequence repetition, we apply it to a sugar binding protein (PDB ID: 1TL2) (Fig. 1). This protein is the typical example of repeated sequence of the propeller fold, where five nearly identical segments form a highly symmetrical five-bladed domain (Fig. 1a) [19]. From the result plot of Fig. 1c, we can easily find that the leading diagonal divides the whole plane equally into the same two parts. The whole upper triangular zone (or lower triangular zone) was partitioned into five parts. The five segments are 47 residues length, their beginning residues are 2, 29, 96, 143 and 192 and their Pearson's correlation coefficients between each other are shown in Fig. 1d. It demonstrates the latent 5-fold periodicity in this amino acid sequence, i.e. the protein with 5-fold symmetry. It is worth noting that the locations of these five parts and the locations of the symmetric parts in the structure have a very good match.

The example above suggests that our method maybe effective to detect the internal repeats in protein sequences. Furthermore, it is interesting to see whether it is possible to identify repeats in the sequences of Ig folds. The tertiary structure and the result of eight typical proteins are shown in Fig. 2. We can see the result plots can be divided into 3 groups. Group I includes Glycosyl hydrolase (1c1c), Immune system (1kgc) and Transferase (1wwc). Immune system (1mju) and hydrolase (1qho) are the members of group II and the remaining three proteins, Immune system (1seq), Immune system (1x9q) and complex (1a14), belong to group III.

For group I, from the result plot we can get that the slanting straight line, which parallels the main diagonal, is continuous and it intersect the horizontal axis. It demonstrates that there is a continuous segment, which begins with the first amino acid, is correlated with another segment, e.g. Glycosyl hydrolase (PDB ID: 1clc), its two repeats can be easily seen from the result plot. The segments I1-A38 and T52-V89 are correlated and its coefficient r is 0.5220.

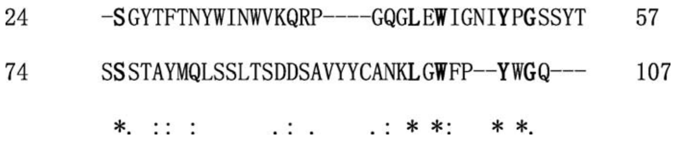


```
1     IETKVSAAKITENYQFDSRIRLNSIGFIP-NHSKKATIA-    38

52    --TIVYTGTATSMFDNDTKETVYIADFSSVNEEGTYYLAV    89

        * * :.. *. :: *::  :   .* . *.. .  :*
```

Immune system (PDB ID: 1mju) is the typical protein of group II. We can see two sequence repeats clearly from the plot, with a length of about 33 amino acids. Although the slanting straight line is continuous, it is different from the members of group I, for it does not intersect the horizontal axis. The two repeats, S24-T57 and S74-Q107, are correlated and the coefficient is 0.5173.

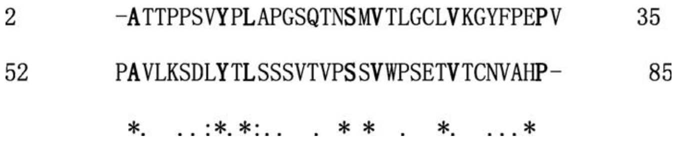


```
24    -SGYTFTNYWINWVKQRP----GQGLEWIGNIYPGSSYT    57

74    SSSTAYMQLSSLTSDDSAVYYCANKLGWFP--YWGQ---    107

       *. :: :     . : .    .:  * *:   * *.
```

For group III, the slanting straight line intersects the horizontal axis, but it is not continuous. Immune system (PDB ID: 1mju) is the member of this group. The 2-fold sequence repeats can be directly seen from the result plot and the segments S24-T57 and S74-Q107 are correlated. The Pearson's correlation coefficient is 0.5173.



```
2     -ATTPPSVYPLAPGSQTNSMVTLGCLVKGYFPEPV    35

52    PAVLKSDLYTLSSSVTVPSSVWPSETVTCNVAHP-    85

       *.  ..:*.*:..  . * * .   *.  ...*
```

It is easy to extend the analysis above to the amino acid sequences of all other proteins in Ig folds. Table 1 gives the calculated results of repeat number at sequence level for all the 19 representative primary sequences, which are selected from CATH [20]. Furthermore, among them the identical amino acids between any two sequences are less than 30%. Therefore, these proteins can be taken as the models of this fold. In Table 1, we also list the correlation coefficients of the repeat segments that show apparent symmetric patterns.

From Table 1, we can easily see that the method we present here can find the 2-fold repeats signals, as their tertiary structures, in almost all of the protein sequences. It is noted that in the proteins of Ig fold mentioned above, Radar and Trust methods cannot find any repeat signals. The reasons may be that these methods use the standard sequence alignment and depend on sequence homologues. Therefore, they can only detect repeats with higher similarity, like those in 1tl2.

The results we showed here may suggest that protein sequences are not random and the formation of the symmetric tertiary structures of these protein domains is the result of their sequence repetition. More importantly, these results are in agreement with

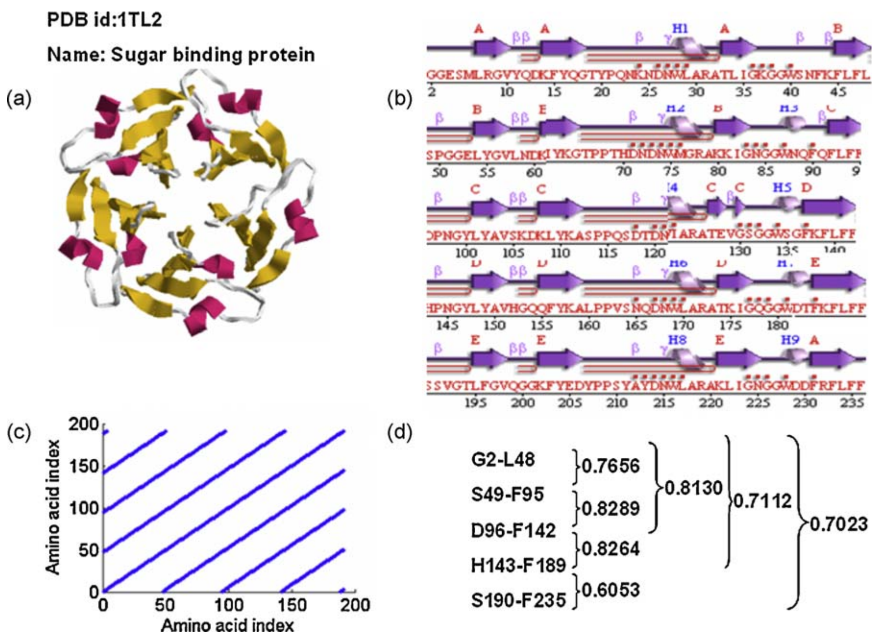

PDB id:1TL2
Name: Sugar binding protein
(a)
(b)
GGESMLRGVYQDKFYQGTYPQNKNDNWLARATLIGKGGWSNFKFLFL
SPGGELYGVLNDKIYKGTPPTHDNDNWMGRAKKIGNGGWNQFQFLFF
DPNGYLYAVSKDKLYKASPPQSDTDNWIARATEVGSGGWSGFKFLFF
HPNGYLYAVHGQQFYKALPPVSNQDNWLARATKIGQGGWDTFKFLFF
SSVGTLFGVQGGKFYEDYPPSYAYDNWLARAKLIGNGGWDDFRFLFF
(c)
Amino acid index
Amino acid index
0
50
100
150
200
(d)
G2-L48
S49-F95
D96-F142
H143-F189
S190-F235
0.7656
0.8289
0.8264
0.6053
0.8130
0.7112
0.7023

Fig. 1. The sugar binding protein (PDB ID: 1TL2). (a) The tertiary sequence; (b) the primary sequence and secondary structure; (c) the result plot. Both the horizontal and vertical axes denote the amino acid index in the sequence; (d) the Pearson's correlation coefficients between any two segments.

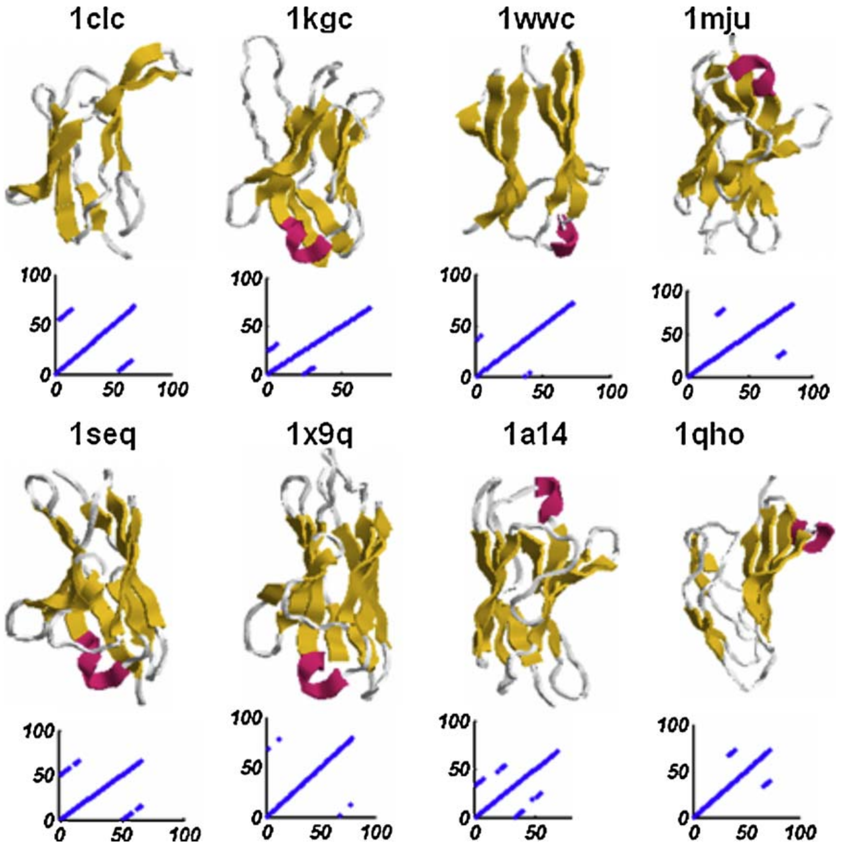


Fig. 2. The structures and recurrence plots of representative proteins of the Ig fold: (a) PDB ID; (b) the tertiary structures; (c) the recurrence plots: based on the polarity distribution of amino acids. The horizontal axis is the residue index in primary sequences and the vertical axis is the repeat length *d*.

Table 1
The calculation results of repeat number at sequence level, the repeat segments and their Pearson's correlation coefficients for proteins in the Ig folds.

PDB ID	Repeat number		Repeat segment	Pearson's correlation coefficients
	Structure level	Sequence level		
1a14	2	2	D1-N39; N34-S72	0.5253
1mju	2	2	S24-T57; S74-Q107	0.5173
1seq	2	2	A2-V35; P52-P85	−0.5264
1k3i	2	2	T1-S31; Y61-A91	0.5191
1vca	2	2	L39-V79; T62-K102	−0.5115
1q0x	2	2	T30-K64; K64-T98	0.5682
1qho	2	2	G33-Q71; A66-T104	0.5385
1mfa	2	2	S25-K67; W47-E89	0.5000
1eaj	2	2	L1-L39; V51-T89	0.5138
1ncw	2	2	R1-V24; F57-F80	0.5499
1oga	2	2	N1-L24; D41-A64	0.6500
1wwc	2	2	T1-L34; L37-T70	−0.5148
1clc	2	2	I1-A38; T52-V89	0.5220
1x9q	2	2	D2-N34; F68-V100	−0.5147
1k5n	2	2	M1-H32; S56-T87	0.5358
1mqk	2	2	E17-L47; S67-L97	0.5216
1mex	2	2	Q5-K38; F62-K95	0.5039
1edq	2	2	A23-T53; D38-T69	0.5083
1kgc	2	2	G1-L44; S25-P68	−0.5116

the theory that modern proteins evolved by gene duplications and fusions [21]. We hope that our method shall be helpful to understand the sequence–structure relationship of proteins.

Acknowledgements

This work is supported by the National Nature Science Foundation of China under Grant no. 30700613 and 40706051 and the National High Technology Research and Development Program ("863"Program) of China (2007AA091602).

References

[1] G.M. Salem, E.G. Hutchinson, C.A. Orengo, J.M. Thornton, Correlation of observed fold frequency with the occurrence of local structural motifs, J. Mol. Biol. 287 (1999) 969–981.

[2] J. Söding, A.N. Lupas, More than the sum of their parts: on the evolution of proteins from peptides, Bioessays 25 (2003) 837–846.

[3] D.M. Halaby, J.P.E. Mornon, The immunoglobulin superfamily: an insight on its tissular, species, and functional diversity, J. Mol. Evol. 46 (1998) 389–400.

[4] E. Padlan, Anatomy of the antibody molecule, Mol. Immunol. 31 (1994) 169.

[5] Y.Z. Huang, Y. Xiao, Detection of gene duplication signals of Ig folds from their amino acid sequences, Proteins: Struct. Funct. Bioinform. 68 (2007) 267–272.

[6] S. Rackovsky, "Hidden" sequence periodicities and protein architecture, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 8580–8584.

[7] A. Heger, L. Holm, Rapid automatic detection and alignment of repeats in protein sequences, Proteins: Struct. Funct. Genet. 41 (2000) 224–237.

[8] A. Heger, L. Holm, Many large proteins have evolved by internal duplication and many internal sequence repeats, Proteins 41 (2000) 224–237. available at: www.ebi.ac.uk/Radar/.

[9] A. Giuliani, R. Benigni, J.P. Zbilut, C.L. Weber, P. Sirabella, A. Colosimo, Nonlinear signal analysis methods in the elucidation of protein sequence–structure relationships, Chem. Rev. 102 (2002) 1471–1491.

[10] R. Szklarczyk, J. Heringa, Tracking repeats using significance and transitivity, Bioinformatics (Part 26) 20 (2004) 1311–1317. available at: http://ibivu.cs.vu.nl/programs/trustwww.

[11] R. Xu, Y. Xiao, A common sequence-associated physicochemical feature for proteins of beta-trefoil family, Comput. Biol. Chem. 29 (2005) 79–82.

[12] J. Söding, M.R. Remmert, A. Biegert, HHrep: de novo protein repeat detection and the origin of TIM barrels, Nucleic Acids Res. 34 (2006) W137–W142.

[13] V.P. Turutina, A.A. Laskin, N.A. Kudryashov, K.G. Skryabin, E.V. Korotkov, Identification of amino acid latent periodicity within 94 protein families, J. Comput. Biol. 13 (2006) 946–964.

[14] X.F. Ji, H.L. Chen, Y. Xiao, Hidden symmetries in the primary sequences of beta-barrel family, Comput. Biol. Chem. 31 (2007) 61–63.

[15] X.C. Wang, Y.Z. Huang, Y. Xiao, Structural-symmetry-related sequence patterns of the proteins of beta-propeller family, J. Mol. Graphics Model. 26 (2008) 829–833.

[16] A.K. Konopka, Sequence complexity and composition, in: D.N. Cooper (Ed.), Nature Encyclopedia of the Human Genome, vol. 5, Nature Publishing Group Reference, London, 2003, pp. 217–224.

[17] X.F. Ji, Y.J. Wang, H.Y. Wang, M. Sun, Identification of protein latent periodicities using recucorrelation analysis, J. Theor. Biol. 255 (2008) 316–319.

[18] Y. Tsuneyuki, M. Takeo, Evidence for the neutral hypothesis of protein polymorphism, Science 178 (1972) 56–58.

[19] H.G. Beisel, S. Kawabata, S. Iwanaga, R. Huber, W. Bode, Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab *Tachypleus tridentatus*, EMBO J. 18 (1999) 2313–2322.

[20] A.L. Cuff, I. Sillitoe, T. Lewis, O.C. Redfern, R. Garratt, J. Thorntonm, C.A. Orengo, The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies, Nucleic Acids Res. 37 (2009) D310–D314, available at: http://www.cathdb.info/..

[21] C. Chothia, J. Gough, C. Vogel, S.A. Teichmann, Evolution of the protein repertoire, Science 300 (2003) 1701–1703.