J. Biomedical Science and Engineering, 2016, 9, 307-314
Published Online May 2016 in SciRes. http://www.scirp.org/journal/jbise
http://dx.doi.org/10.4236/jbise.2016.96025

Hidden Sequence Repeats: Additional Evidence for the Origin of TIM-Barrel Family

Xiaofeng Ji, Yuan Zheng, Zhipeng Wang, Jun Sheng*

Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, China
Email: *shengjun@ysfri.ac.cn

Received 19 April 2016; accepted 28 May 2016; published 31 May 2016

Abstract

Most proteins adopt an approximate structural symmetry. However, they have no symmetry detectable in their sequences and it is unclear for most of these proteins whether their structural symmetry originates from duplication. As one of the six popular folds (super-folds) possessing an approximate structural symmetry, the triosephosphate isomerase barrel (TIM-barrel) domain has been widely studied. Using modified recurrent quantification analysis of primary sequences, we identified the same 2-, 3-, and 4-fold symmetry pattern as their tertiary structures. This result indicates that the symmetry in tertiary structure is coded by symmetry in the primary sequence and that the TIM-barrel adopts a 2-, 3-, or 4-fold repeat pattern during evolution. This discovery will be useful for understanding the evolutionary mechanisms of this protein family and the symmetry pattern that may be a clue into the ancient origin of duplication of half-barrels or the β a unit.

Keywords

TIM-Barrel, Hidden Symmetry, Primary Sequences, Repeat Pattern, Recurrence Quantification Analysis

1. Introduction

Proteins are amino acid polymers that can adopt a wide range of structures uniquely determined by sequence. It is well-known that the information regarding structure formation is contained within their amino acid sequences [1]. Nevertheless, many proteins exhibit obvious symmetry at the level of tertiary structures and yet seldom show periodicity in their primary sequences [2] [3]. A detailed analysis of the repeats in protein sequences may

*Corresponding author.

How to cite this paper: Ji, X.F., Zheng, Y., Wang, Z.P. and Sheng, J. (2016) Hidden Sequence Repeats: Additional Evidence for the Origin of TIM-Barrel Family. *J. Biomedical Science and Engineering*, **9**, 307-314.
http://dx.doi.org/10.4236/jbise.2016.96025

help us to better understand the evolutionary mechanisms proteins used to adapt their structure and function under evolutionary pressure.

The eight-stranded β/α barrel (triosephosphate isomerase [TIM] barrel) is by far the most common tertiary fold observed in high-resolution protein crystal structures and it mediates diverse function maintaining overall structure. It is estimated that 10% of all known enzymes have this fold [4]. By itself, the TIM-barrel fold has typically approximately 250 residues, with a minimum of approximately 200 residues required to form its structure; branched hydrophobic side chains dominate the core of β/α barrels [5]. The closed parallel β-domain structure of the $(\beta/\alpha)_8$-barrel is formed from eight parallel (β/α)-units linked by hydrogen bonds (**Figure 1**). Based on structural [6] and sequence [7] analysis of HisA and HisF, the $(\beta/\alpha)_8$ barrel domain of both of these enzymes appears to be the result of a gene duplication and fusion. Richter and colleagues suggested a two-step evolutionary pathway in which a HisF-N1-like predecessor was duplicated and fused twice to yield HisF [8]. Despite many experimental studies showing that the $(\beta/\alpha)_8$-barrel may evolve from an ancestral half or quarter-barrel [6] [9] [10] and structures of this family are approximately symmetrical, evidence for an origin of this common ancestor by 4-fold duplication is lacking.

Internal repeats in protein sequences have wide-ranging implications for the structure and function of proteins. The ability to detect repeated structures based only on sequence analysis would support the evolutionary hypotheses that a large fraction of modern-day enzymes evolved from a basic structural unit. In order to detect latent symmetries in protein sequences, some effort has been made. Different methods [11]-[19] have been proposed to detect periods in the sequences of beta-trefoil [20], beta-barrel [21], beta-propeller [22] [23], Ig fold [24] [25], and left-handed beta-helix fold [26], among others. Notably, there are popular web tools available that detect repeats: RADAR [18], TRUST [17], HHrep [16], REPETITA [14], and FAIR [13]. These tools identify repeats in protein and DNA sequences based on suboptimal self-sequence alignment. These tools are useful for general repeats detection, but are less useful for symmetric sequence repeats. In our previous paper [27], a modified recurrence plot was used to detect latent periodicities in proteins with an Ig fold. At that time, the amino acids were denoted by their corresponding Grantham polarity values [28] and Pearson's correlation coefficients were used to characterize similarity. If the two segments showed a higher correlation, they were considered to be more similar. In order to understand the evolution of the $(\beta/\alpha)_8$-barrel family, here we propose a fast and sensitive modified quantification analysis method to detect the hidden symmetries in the primary sequence of non-homologous sequences with CATH [29] Code 3.20.20. In this study, hydrophilic and hydrophobic features were used to denote the corresponding amino acids. Additionally, the percentages of their identical symbols were used to characterize similarity. Our result showed that nearly all numbers of this family were 2-, 3-, and 4-fold symmetric. This result may increase the understanding of the evolutionary mechanisms of $(\beta/\alpha)_8$-barrel family.

2. Methods

The method of modified recurrence plot, which was guided by the idea of recurrence quantification analysis [30] was used to identify internal repeats in the TIM-barrel family. The flow chart of this method is shown in **Figure 2**.

Consider an arbitrary sequence $S = x_1 x_2 x_3 \cdots x_N$, where N is the length of the sequence and x_i denotes one of the 20 amino acids. First, the complexity of the protein sequence should be reduced. From the Introduction, we can easily find that the $(\beta/\alpha)_8$-barrel is mainly characterized by α-helix and β-strand, and their structural features are mainly determined based on their hydrophilic and hydrophobic regions. Hence, we reduce protein sequence complexity by grouping the 20 amino acids into four groups based on their individual hydrophobicity according to the ranges of the hydropathy scale (**Table 1**) [31]. After this step, a vector representation of the protein sequence, as $A = a_1 a_2 a_3 \cdots a_N$, is achieved. Next, sets of possible segments, as described in our previous paper

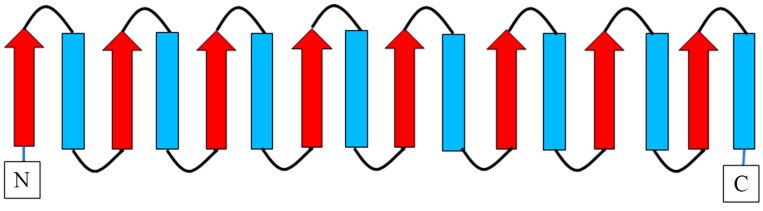

Figure 1. The topological structure diagram of the eight-stranded β/α barrel.

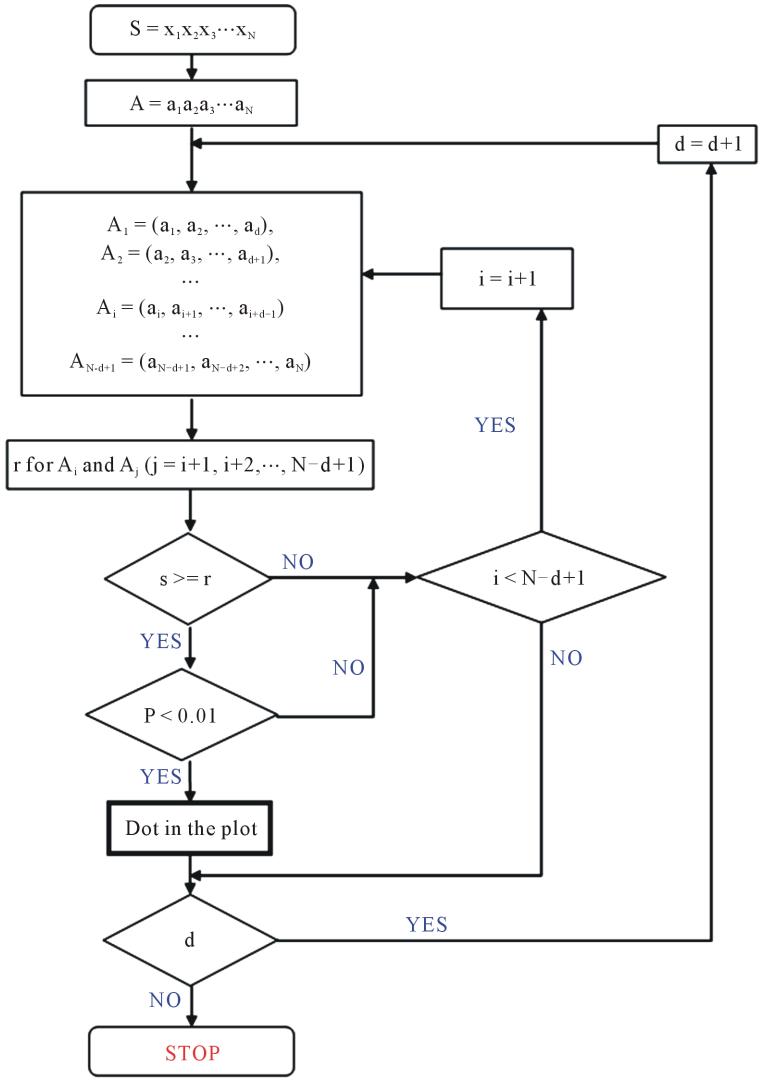

Figure 2. The flow chart of the method.

Table 1. Hydropathy characteristics.

Hydropathy characteristic	Abbreviation	Amino acids
Strongly hydrophilic (polar)	POL	R, D, E, N, Q, K, H
Strongly hydrophobic	HPO	L, I, V, A, M, F
Weakly hydrophilic or weakly hydrophobic (ambiguous)	AMBI	S, T, Y, W
Undefined	UND	C, G, P

[27], were constructed. For any segment $X_i = x_i x_{i+1} \cdots x_{i+d-1}$ $(1 \le i \le N - d + 1)$, if we can identify another segment $X_j = x_j x_{j+1} \cdots x_{j+d-1}$ $(j \ne i)$ of the same length in the sequence S and at the same time the two segment are similar, we plot a point at (i, d) and (j, d) in the modified recurrence plot. Two segments are similar if the percentage (s) of their identical symbols is larger than a chosen number r ($0 < r < 1$) and when P-value is lower than 0.01. When this was completed for all the possible i and d, the modified recurrence plot was formed. We decreased the value of r gradually to detect symmetries in primary sequences.

In order to assess the performance of our method for repeat detection, our results were compared with those

obtained using the web tools discussed in the Introduction section. Among these tools, HHrep and REPETITA are based on existing knowledge and they use information from sequence profiles. Moreover, FAIR can only identify short segments. Hence, only the de novorepeat detection methods REPRO, RADAR, and TRUST were used for the accession procedure. Compared with these three methods, our method showed high accuracy for all selected proteins (**Table 2**) for repeats and residues. Our method also showed a higher sensitivity for repeat prediction, although the sensitivity was lower than that of REPRO if repeat residues were counted.

3. Results and Discussion

We used typical proteins of eight-stranded β/α barrel family as examples to demonstrate the effectiveness of our methods for detecting symmetries in protein sequence. The TIM-barrel is an ancient fold with considerable sequence diversity. It evolved from the half- or quarter-barrel. Particularly, the prototypical $(\beta/\alpha)8$-barrel proteins HisA (PDB id: 1QO2) and HisF (PDB id 1THF) provided evidence that this fold evolved from a $(\beta/\alpha)4$-half or $(\beta/\alpha)4$ quarter-barrel ancestor. If the chain conformations of protein are primarily determined by the information contained in its amino acid sequence, there must be signals which indicate the structural symmetry in the sequences of these proteins. Here, we used HisA and HisF as examples.

Figure 3(c) shows that the entire zone was partitioned into two main parts. This demonstrates the latent 2-fold periodicity in both of these sequences. For HisF, the recurrence plot shows that at position 122, the sharp boundary line divides the plot into two parts. This means that segments 1 - 122 and 123 - 253 are symmetric. Similarly to HisF, the sharp boundary line divides the recurrence plot of HisA into two parts in $x_i = 118$. This result agrees with the experimental findings that the TIM-barrel family evolved from repeated duplication of simpler units.

It is easy to extend the analysis above to the amino acid sequences of all other proteins in this family. Sixteen proteins were selected from the fold of TIM-barrel in CATH, among them the identical amino acids between any two sequences are less than 30%. Furthermore, among these, identical amino acids between any two sequences

Table 2. Sensitivity and accuracy for different selected proteins from PROPEAT.

Sensitivity:

Folds	Repeats				Residues			
	Radar	Trust	Repro	Our	Radar	Trust	Repro	Our
β-trefoill	28.79	28.79	65.15	96.31	30.74	27.81	99.26	77.44
Jelly-roll	22.50	13.85	96.92	97.65	-------	-------	-------	-------
Ig like	9.38	15.63	90.63	94.11	8.64	17.69	110.23	98.82
TIM-barrel	23.75	22.50	50.00	91.63	19.54	57.72	107.94	97.36
Ferredoxin-like	0	0	100	100	0	0	62.79	88.49
Total	20.18	19.01	73.01	93.47	16.65	24.00	93.12	90.65

Accuracy:

Folds	Repeats				Residues			
	Radar	Trust	Repro	Our	Radar	Trust	Repro	Our
β-trefoill	63.16	68.42	42.50	72.52	76.16	76.88	63.20	80.76
Jelly-roll	-------	-------	-------	-------	-------	-------	-------	-------
Ig like	50.00	70.00	44.83	78.96	8.33	52.34	48.29	61.03
TIM-barrel	47.05	56.25	43.74	82.47	29.64	39.88	29.57	63.11
Ferredoxin-like	0	0	50.00	100.00	0	0	92.00	97.87
Total	50.00	59.57	47.37	85.45	48.15	54.58	45.06	76.62

were less than 30%. Therefore, these proteins can be considered representatives of the TIM-barrel family. We showed that the modified recurrence plot clearly revealed 2-fold, 4-fold, and even 3-fold symmetry in the primary sequence. First, we found the 2-fold symmetry in all members of this family had a similarity degree of r = 0.4 for the alignment, supporting the hypothesis of the origin of protein domains by duplication and recombination of simpler peptides. **Figure 4** shows the modified recurrence plot of typical proteins of the TIM-barrel family, and all of the results are listed in **Table 2**. Based on the partitioned mode of the plot, the modes of origin can be classified into three main categories (**Table 3**).

Categories 1 (e.g., **Figure 4**, S1) clearly contained a nearly 4-fold repeat structure with all three sub-optimal alignments visible; 4 + 4 indicates that the proteins evolved from an ancestral half-barrel. However, when we restricted the threshold, the multi-fold symmetry of the primary sequence emerged. This result supports that the

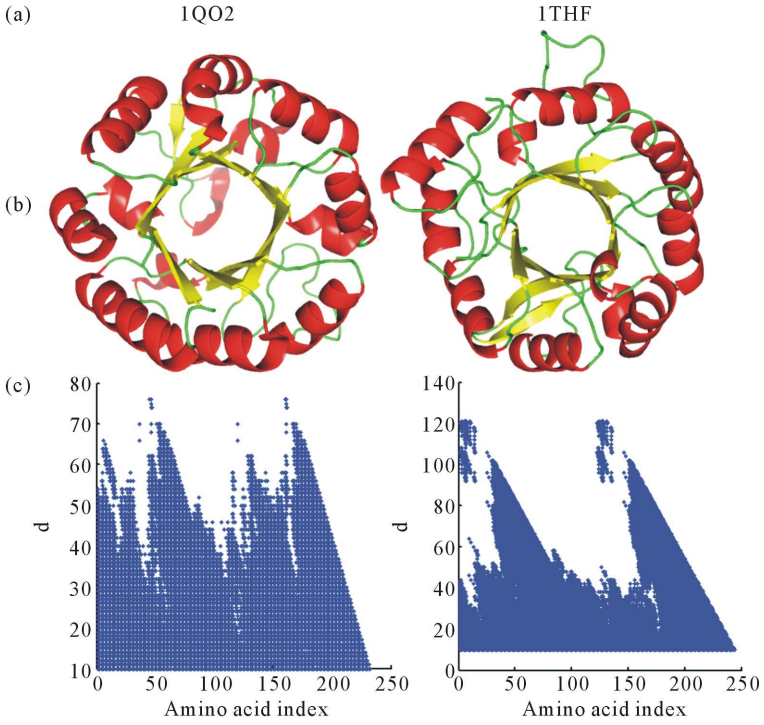

Figure 3. The tertiary structures and recurrence plot of imidazoleglycerol phosphate (PDBid:1thf) and Isomerase (PDBid:1qo2). (a) PDBid of the protein; (b) the tertiary structure. This figure was generated by Pymoland it was shown in rainbow cartoon; (c) the recurrence plot.

Table 3. Result of all the proteins is classified into three categories[#].

Categories	R	Format	PDB id
S1 (e.g. **Figure 4**: S1)	0.4 and 0.5	4 + 4	1eex, 1gk8, 1hzy, 1s2w, 1bd0, 1eye, 1i1w, 1v93
	0.6	2 + 2 + 2 + 2	
S2 (e.g. **Figure 4**: S2)	0.4	4 + 4	1o1z
	0.5 and 0.6	2 + 3 + 3	
S3 (e.g. **Figure 4**: S3)	0.4 and 0.5	4 + 4	1luc, 1kko, 1ex1, 1muw, 1req, 1ccw, 1oc7
	0.6	5 + 3	

[#]Here, we regard the βa domain as the basic unit to form the tertiary structure. We use a formula $N_1 + N_2 + \cdots + N_i + \cdots + N_n$ to express "format". In the formula N_i (i = 1, 2, 3, ···, n) means the number of βa domain to form a beta-domain; n means the number of beta-domain to form the whole structure. (e.g. Format 4 + 4 means 4 βa domains form a beta-domain, and the whole structure is grouped by the two domains.)

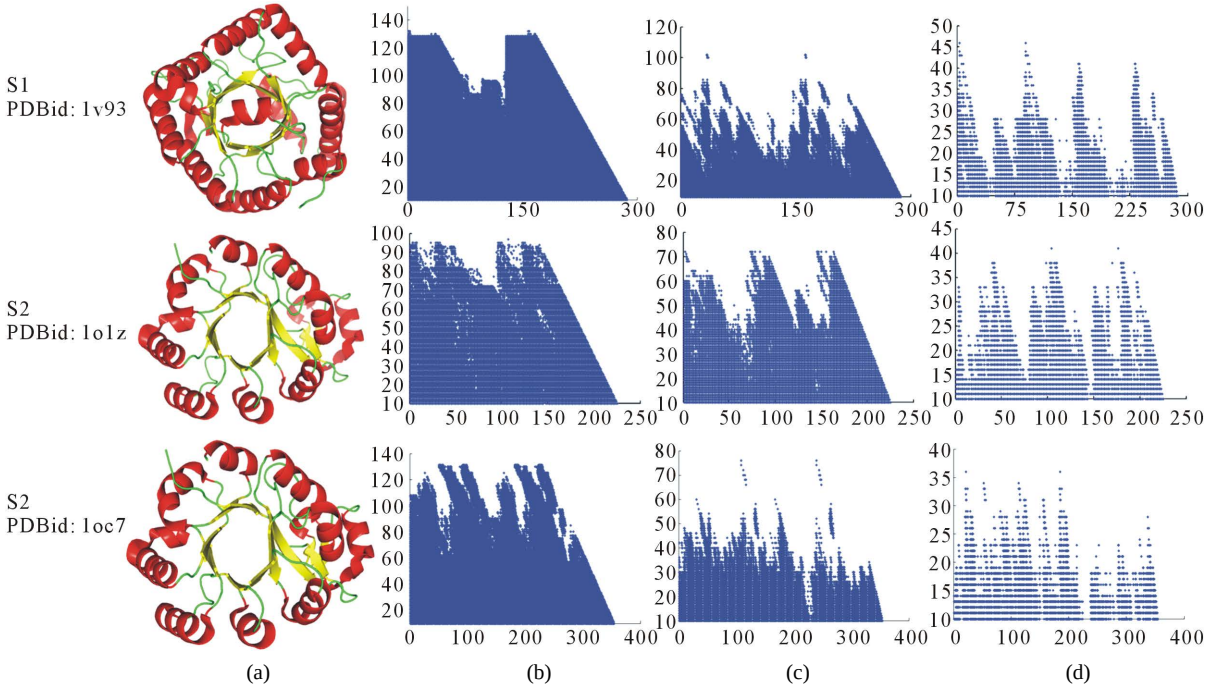

Figure 4. Structures and recurrence plots of the representative proteins. (a) The tertiary structures of proteins. (b)-(d) Modified recurrence plot with the values of r = 0.40, 0.50, 0.60 respectively. S means "categories".

ancient module may have arisen by 2-fold duplication of an aβ precursor, which would have given rise to the 8-fold symmetry. The same is true for other representative numbers (1EEX, 1GK8, 1HZY, 1S2W, 1BD0, 1EYE, 1I1W) of this family (not shown here).

Categories 2 (e.g., **Figure 4**, S2), the 3-fold symmetry emerged as the similarity degree increased. The protein may have had three ancestral segments, but the structure alignment showed that the latter two domains (3 + 3) were similar (rmsd = 3.77). One can speculate that the ancient $\beta\alpha$ domain may have duplicated to form the $\beta\alpha\beta\alpha\beta\alpha$ domain, and the other domain evolved by tandem duplication and fusion from the formed domain.

Categories 3 (e.g., **Figure 4**, S3), with the format of 5 + 3, the former domain (fi = 5) may have contained an βa domain as the ancestral segment and the latter domain (fi =3) contained another; therefore, we speculated that these proteins evolved by gene duplication from two ancestral segments, which formed the domain by duplication respectively during the early stage of evolution.

4. Conclusion

An internal repeat is a character that proteins use to adapt their structures and functions under evolutionary pressure. A detailed analysis of internal repeats within protein sequences may have wide-ranging implications for protein evolutionary trends. In this study, we used modified recurrence analysis method to detect hidden symmetries within proteins from the TIM-barrel family which accounted clearly for the 2-, 3-, and 4-fold symmetry. This result was consistent with the idea that TIM-barrels evolved from repeated duplication of simpler units. These findings support the hypothesis that protein evolution typically occurs by duplication, mutation, and shuffling from existing protein domains. Occasionally, the domains themselves are produced de novo, but they primarily belong to an established set. This result suggests that the symmetries at the structure level are due to those at sequence level. We hope that our results are useful for the development of structural prediction methods and understanding the mechanisms of protein evolution.

Acknowledgements

This work is supported by the Special Scientific Research Funds for Central Non-profit Institute, Yellow Sea Fisheries Research Institutes (Grant no. 20603022015012 and 20603022013016).

References

[1] Anfinsen, C.B. (1973) Principles That Govern the Folding of Protein Chains. *Science*, **181**, 223-230. http://dx.doi.org/10.1126/science.181.4096.223

[2] Soding, J. and Lupas, A.N. (2003) More than the Sum of Their Parts: On the Evolution of Proteins from Peptides. *Bioessays*, **25**, 837-846. http://dx.doi.org/10.1002/bies.10321

[3] Lupas, A.N., Ponting, C.P. and Russell, R.B. (2001) On the Evolution of Protein Folds: Are Similar Motifs in Different Protein Folds the Result of Convergence, Insertion, or Relics of an Ancient Peptide World? *Journal of Structural Biology*, **134**, 191-203. http://dx.doi.org/10.1006/jsbi.2001.4393

[4] Nagano, N., Orengo, C.A. and Thornton, J.M. (2002) One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on Their Sequences, Structures and Functions. *Journal of Molecular Biology*, **321**, 741-765. http://dx.doi.org/10.1016/S0022-2836(02)00649-6

[5] Branden, C. and Tooze, J. (1991) Introduction to Protein Structure. Garland, New York.

[6] Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. and Wilmanns, M. (2000) Structural Evidence for Evolution of the Beta/Alpha Barrel Scaffold by Gene Duplication and Fusion. *Science*, **289**, 1546-1550. http://dx.doi.org/10.1126/science.289.5484.1546

[7] Fani, R., Lio, P., Chiarelli, I. and Bazzicalupo, M. (1994) The Evolution of the Histidine Biosynthetic Genes in Prokaryotes: A Common Ancestor for the hisA and hisF Genes. *Journal of Molecular Evolution*, **38**, 489-495. http://dx.doi.org/10.1007/BF00178849

[8] Lee, J. and Blaber, M. (2011) Experimental Support for the Evolution of Symmetric Protein Architecture from a Simple Peptide Motif. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 126-130. http://dx.doi.org/10.1073/pnas.1015032108

[9] List, F., Sterner, R. and Wilmanns, M. (2011) Related (Betaalpha)8-barrel Proteins in Histidine and Tryptophan Biosynthesis: A Paradigm to Study Enzyme Evolution. *ChemBioChem*, **12**, 1487-1494. http://dx.doi.org/10.1002/cbic.201100082

[10] Richter, M., Bosnali, M., Carstensen, L., Seitz, T., Durchschlag, H., Blanquart, S., Merkl, R. and Sterner, R. (2010) Computational and Experimental Evidence for the Evolution of a ($\beta\alpha$)$_{< sub> 8}$-Barrel Protein from an Ancestral Quarter-Barrel Stabilised by Disulfide Bonds. *Journal of Molecular Biology*, **398**, 763-773. http://dx.doi.org/10.1016/j.jmb.2010.03.057

[11] Pellegrini, M., Renda, M.E. and Vecchio, A. (2012) *Ab Initio* Detection of Fuzzy Amino Acid Tandem Repeats in Protein Sequences. *BMC Bioinformatics*, **13**, S8. http://dx.doi.org/10.1186/1471-2105-13-S3-S8

[12] Luo, H., Lin, K., David, A., Nijveen, H. and Leunissen, J.A. (2012) ProRepeat: An Integrated Repository for Studying Amino Acid Tandem Repeats in Proteins. *Nucleic Acids Research*, **40**, D394-D399. http://dx.doi.org/10.1093/nar/gkr1019

[13] Senthilkumar, R., Sabarinathan, R., Hameed, B.S., Banerjee, N., Chidambarathanu, N., Karthik, R. and Sekar, K. (2010) FAIR: A Server for Internal Sequence Repeats. *Bioinformation*, **4**, 271-275. http://dx.doi.org/10.6026/97320630004271

[14] Marsella, L., Sirocco, F., Trovato, A., Seno, F. and Tosatto, S.C. (2009) REPETITA: Detection and Discrimination of the Periodicity of Protein Solenoid Repeats by Discrete Fourier Transform. *Bioinformatics*, **25**, i289-i295. http://dx.doi.org/10.1093/bioinformatics/btp232

[15] Nirjhar Banerjee, N.C.D.M. (2008) An Algorithm to Find All Identical Internal Sequence Repeats. *Current Science India*, **95**, 188-195.

[16] Soding, J., Remmert, M. and Biegert, A. (2006) HHrep: *De Novo* Protein Repeat Detection and the Origin of TIM Barrels. *Nucleic Acids Research*, **34**, W137-W142. http://dx.doi.org/10.1093/nar/gkl130

[17] Szklarczyk, R. and Heringa, J. (2004) Tracking Repeats Using Significance and Transitivity. *Bioinformatics*, **20**, i311-i317. http://dx.doi.org/10.1093/bioinformatics/bth911

[18] Heger, A. and Holm, L. (2000) Rapid Automatic Detection and Alignment of Repeats in Protein Sequences. *Proteins*: *Structure, Function, and Bioinformatics*, **41**, 224-237. http://dx.doi.org/10.1002/1097-0134(20001101)41:2<224::AID-PROT70>3.0.CO;2-Z

[19] Rackovsky, S. (1998) "Hidden" Sequence Periodicities and Protein Architecture. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 8580-8584. http://dx.doi.org/10.1073/pnas.95.15.8580

[20] Xu, R. and Xiao, Y. (2005) A Common Sequence-Associated Physicochemical Feature for Proteins of Beta-Trefoil Family. *Computational Biology and Chemistry*, **29**, 79-82. http://dx.doi.org/10.1016/j.compbiolchem.2004.12.003

[21] Ji, X., Chen, H. and Xiao, Y. (2007) Hidden Symmetries in the Primary Sequences of Beta-Barrel Family. *Computa-*

tional Biology and Chemistry, **31**, 61-63. http://dx.doi.org/10.1016/j.compbiolchem.2007.01.002

[22] Yadid, I. and Tawfik, D.S. (2011) Functional Beta-Propeller Lectins by Tandem Duplications of Repetitive Units. *Protein Engineering, Design and Selection*, **24**, 185-195. http://dx.doi.org/10.1093/protein/gzq053

[23] Wang, X., Huang, Y. and Xiao, Y. (2008) Structural-Symmetry-Related Sequence Patterns of the Proteins of Beta-Propeller Family. *Journal of Molecular Graphics and Modelling*, **26**, 829-833. http://dx.doi.org/10.1016/j.jmgm.2007.04.014

[24] Ji, X., Wang, H., Hao, J., Zheng, Y., Wang, W. and Sun, M. (2010) Identification of Sequence Repetitions in Immunoglobulin Folds. *Journal of Molecular Graphics and Modelling*, **28**, 788-791. http://dx.doi.org/10.1016/j.jmgm.2010.02.003

[25] Huang, Y. and Xiao, Y. (2007) Detection of Gene Duplication Signals of Ig Folds from Their Amino Acid Sequences. *Proteins*: *Structure, Function, and Bioinformatics*, **68**, 267-272. http://dx.doi.org/10.1002/prot.21330

[26] Shen, X. (2011) Conformation and Sequence Evidence for Two-Fold Symmetry in Left-Handed Beta-Helix Fold. *Journal of Theoretical Biology*, **285**, 77-83. http://dx.doi.org/10.1016/j.jtbi.2011.06.011

[27] Ji, X., Sheng, J., Wang, F., Zhang, S., Hao, J., Wang, H. and Sun, M. (2011) Identification of Latent Periodicity in Domains of Alkaline Proteases. *Biochemistry* (*Moscow*), **76**, 1037-1042. http://dx.doi.org/10.1134/S0006297911090082

[28] Yamazaki, T. and Maruyama, T. (1972) Evidence for the Neutral Hypothesis of Protein Polymorphism. *Science*, **178**, 56-58. http://dx.doi.org/10.1126/science.178.4056.56

[29] Sillitoe, I., Cuff, A.L., Dessailly, B.H., Dawson, N.L., Furnham, N., Lee, D., Lees, J.G., Lewis, T.E., Studer, R.A., Rentzsch, R., Yeats, C., Thornton, J.M. and Orengo, C.A. (2013) New Functional Families (FunFams) in CATH to Improve the Mapping of Conserved Functional Sites to 3D Structures. *Nucleic Acids Research*, **41**, D490-D498. http://dx.doi.org/10.1093/nar/gks1211

[30] Konopka, A.K. (2005) Sequence Complexity and Composition. eLS.

[31] Panek, J., Eidhammer, I. and Aasland, R. (2005) A New Method for Identification of Protein (Sub) Families in a Set of Proteins Based on Hydropathy Distribution in Proteins. *Proteins*: *Structure*, *Function*, *and Bioinformatics*, **58**, 923-934. http://dx.doi.org/10.1002/prot.20356