OPEN ACCESS Freely available online

PLoS one

Symmetric Key Structural Residues in Symmetric Proteins with Beta-Trefoil Fold

Jianhui Feng[1☯], Mingfeng Li[1,2☯], Yanzhao Huang[1], Yi Xiao[1]*

1 Biophysics and Molecular Modeling Group, Department of Physics, Huazhong University of Science and Technology, Wuhan, China, **2** Department of Neurobiology and Kavli Institute for Neuroscience, Yale University School of Medicine, New Haven, Connecticut, United States of America

Abstract

To understand how symmetric structures of many proteins are formed from asymmetric sequences, the proteins with two repeated beta-trefoil domains in Plant Cytotoxin B-chain family and all presently known beta-trefoil proteins are analyzed by structure-based multi-sequence alignments. The results show that all these proteins have similar key structural residues that are distributed symmetrically in their structures. These symmetric key structural residues are further analyzed in terms of inter-residues interaction numbers and B-factors. It is found that they can be distinguished from other residues and have significant propensities for structural framework. This indicates that these key structural residues may conduct the formation of symmetric structures although the sequences are asymmetric.

Citation: Feng J, Li M, Huang Y, Xiao Y (2010) Symmetric Key Structural Residues in Symmetric Proteins with Beta-Trefoil Fold. PLoS ONE 5(11): e14138. doi:10.1371/journal.pone.0014138

Editor: Annalisa Pastore, National Institute for Medical Research, Medical Research Council, London, United Kingdom

Received July 24, 2010; **Accepted** November 4, 2010; **Published** November 30, 2010

Funding: This work is supported partly by the National Natural Science Foundation of China (www.nsfc.gov.cn) under Grant No.30870678, 11074084 and 30525037. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: yxiao@mail.hust.edu.cn

☯ These authors contributed equally to this work.

Introduction

Symmetric proteins [1] are ideal objects to investigate protein evolution and folding. It is generally accepted that symmetric proteins have been arisen from gene duplications and fusions [2,3]. However, these repetitive or symmetric signals were almost lost in their sequences during evolution but remain in their structures. Investigating how these proteins keep their symmetric structures by "asymmetric" sequences is a way to understand protein evolution and folding. On the other hand, understanding the building principle of symmetric proteins is also necessary for designing de novo proteins, because symmetric structures are relatively simple to be built from basic units. One solution to the problem above is that protein sequences may contain hidden symmetric signals that determine their symmetric structures [4–8]. Recently, we suggested that these hidden symmetric signals might be contributed by a small number (about 30%) of identical or key residues [9–15].

Multi-domain proteins provide ideal models to study the problem above since many of them consist of more than one domains evolved from the same ancestor and have similar structural symmetry but different sequence symmetry. For example, *Ricin Toxin B* (RTB, PDB id: 2aaib) is composed of two domains with the same beta-trefoil structure of three-fold symmetry [16–18]. It was speculated that RTB is the twice triplicate duplications of its ancestor, a galactose-binding peptide of about forty residues [18]. Rutenber *et al.* detected hidden three-fold sequence symmetry in both domains [18] but the degrees are very different. In its first domain the averaged sequence similarity index between the trefoil units equals 1.73 while in its second domain it is 2.63, i.e., one half larger than that of the first domain. This appears in contradiction with their almost identical structures. Since these two domains have evolved from the same ancestor, they are ideal model to understand sequence-structure relations of proteins. In fact, for RTB, Haze detected a three-fold repetitive QXW motif in both domains and regarded them as key structural residues [19]. Rutenber and Robertus also described a 12-residue hydrophobic core in both domains [20] and later Murzin *et al.* further showed that these residues are characteristic of the beta-trefoil fold [17]. It seems that these key residues may be the main factor to determine the symmetric structure. However, more evidences are needed to validate this conclusion. At least, we need to investigate other proteins in the same family.

According to *Structural Classification Of Proteins* (SCOP) databank [21], RTB belongs to *Plant Cytotoxin B-chain* (PCB) family and all proteins in this family contain two domains with beta-trefoil structure (see Materials and Methods). In this paper we shall analyze their sequence symmetries and identify their key structural residues by three different methods: structure-based multi-sequence alignments, residue interaction number and B-Factor analysis. We shall also extend our analysis to all presently known beta-trefoil proteins. Our results show that there exist similar key structural residues in all these proteins that may determine the symmetry of their structures.

Materials and Methods

Plant Cytotoxin B-chain Family

According to SCOP1.69, there are five species and sixteen protein chains in PCB family (Table 1). Among them, two species, *European mistletoe* and *Sambucus ebuLus*, have more than one protein chains. We select 1m2tb and 1hwmb as their representatives

Table 1. Characteristics of Plant Cytotoxin B-chain family.

Species	Protein Chain[a]	Resolution[b] (Å)	RMSD[c] (Å)
Castor bean	**2aaib**	2.50	1.50
Abrus precatorius	**1abrb**	2.14	1.24
MongoLian snake-gourd	**1ggpb**	2.70	1.77
European mistLetoe	**1m2tb**, 1pc8b, 1onkb, 1puub,		
1pumb,1oqlb, 1tfmb, 1ce7b, 2mllb	1.89	1.30	
Sambucus ebuLus	**1hwmb**,1hwob, 1hwnb, 1hwpb	2.80	1.50

[a]Bold entries indicate representative protein chains.
[b]Experiment resolution of crystal structure for representative protein chains.
[c]RMSD of structural superposition between domains for representative protein chains.
doi:10.1371/journal.pone.0014138.t001

because both have crystal structures of the highest experimental resolutions (Table 1) [22]. The atomic coordinates of the crystal structures (PDB file) and experimental resolutions are retrieved from Protein Data Bank (Table 1).

Detection and Quantification of Protein Sequence Symmetry

In a previous paper [12], we developed a modified recurrence plot (MRP) algorithm to detect protein sequence symmetry, and defined two parameters R and S to quantify the degree of the detected sequence symmetry. Here, we only introduce them briefly.

The MRP of a protein sequence $x_1\ x_2\ x_3 \ldots\ x_N$ is built as follows: the horizontal axis i denotes the location of the first residue of a segment in sequence and the vertical axis d denotes the length of the segment. For any segment $X_i = x_i\ x_{i+1}\ \ldots\ x_{i+d-1}$, if the number of its non-overlapping similar segments $X_j = x_j\ x_{j+1}\ \ldots\ x_{j+d-1}$ ($|j-i| \geq d$) is larger than the degree of symmetry you want to find, we plot a point at (i, d). The MRP is formed when this is done for all possible i and d. Two segments are similar if the percentage of their similar residues, obtained by using pair-wise global sequence alignment with PAM250 score matrix, is larger than a chosen number r and when p-value is lower than 0.05.

The parameter R is the Pearson's correlation coefficient between iMRP and rMRP, where iMRP denotes the ideal symmetric MRP corresponding to the real MRP (rMRP) of protein sequence. R reports the presence of non-overlapping repetitive patterns. Because the R value cannot definitely tell us the degrees of similarities of different patterns and so the degree of sequence symmetry, we introduce a parameter S to do this. S is the average value of the Pearson's correlation coefficients between all different patterns and describes the average similarity of different patterns. Therefore, the S value is a measure of the degree of sequence symmetry. For a sequence to be symmetric, both R and S should have large values. The details of this method can be found in ref. 12. It is noted that there existed other methods to find repeats of a protein sequence [4–8].

Evaluation of Residue Interactions

The residue interaction number (RIN) of a residue is the number of the interaction pairs between this residue and other residues that are more than four residues apart along sequence and their potential energies are lower than −0.5kcal/mol [23,24]. The potential energy is calculated with all-atom force field and implicit solvent model (GB/SA) [25,26]. It is the sum of three energy terms: Van der Waals energy, electrostatic energy and solvent polarized energy. The third term denotes electrostatic interactions ΔG_{pol} between the solute and solvent and is calculated by

$$\Delta G_{pol} = -166.0\left(1-\frac{1}{\varepsilon}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{\sqrt{r_{ij}^2+\alpha_i\alpha_j e^{-D_{ij}}}},$$

where $D_{ij} = r_{ij}^2/4\alpha_i\alpha_j$ and r_{ij} is the distance between atom i and atom j. q_i and q_j are the charges of atom i and atom j. ε is the dielectric constant of the solvent. α_i is the effective Born radius of atom i, which is related to the effective Born free energy of solvation. The molecular mechanics software we used is Tinker with Charmm27 force field [27,28]. Before formal calculations we optimize protein structure by conjugate-gradient method and the gradient tolerance is 0.1kcal/(Å mol).

Results and Discussions

Three-fold sequence symmetries of different degrees

Fig. 1 gives the MRPs of the two domains of the five representative protein chains ($r = 0.3$ as in the previous paper [12]). It shows that all MRPs contain three repetitive patterns. The R values of all domains are larger than 0.5, and all the S values are larger than 0.4 only with one exception (Table 2). In our previous work, $R \geq 0.5$ and $S \geq 0.4$ are set as the cutoff values to measure whether a MRP shows symmetry or not [12]. Thus, almost all domains show hidden three-fold sequence symmetries. However, the MRPs of all the second domains reveal a pattern of three approximately right-angled triangles and the pattern is much more distinguishable than those of the first domains (Fig. 1). This means the symmetry degree of the second domains is higher than that of the first domains. In agreement with this, the R and S values of the second domains are all larger than those of the first domains with only one exception (Table 2) and the differences of the S values are significant, equaling 0.18, 0.10, 0.30, 0.22 and 0.18, respectively, and being about 35.3%, 22.7%, 54.6%, 34.4% and 34.6% of their respective means. This is in agreement with the result of RTB [18].

For the five representative proteins, the first domains are superposed to their second domains with the aid of OPAAS [29] and the root-mean-square distances (RMSD) are all less than 2Å (Table 1), i.e., the first and second domains have similar structures. Therefore, the symmetry degrees of the first and second domains are the same at structural level but different at sequence level. This is also in agreement with the result for RTB [18].

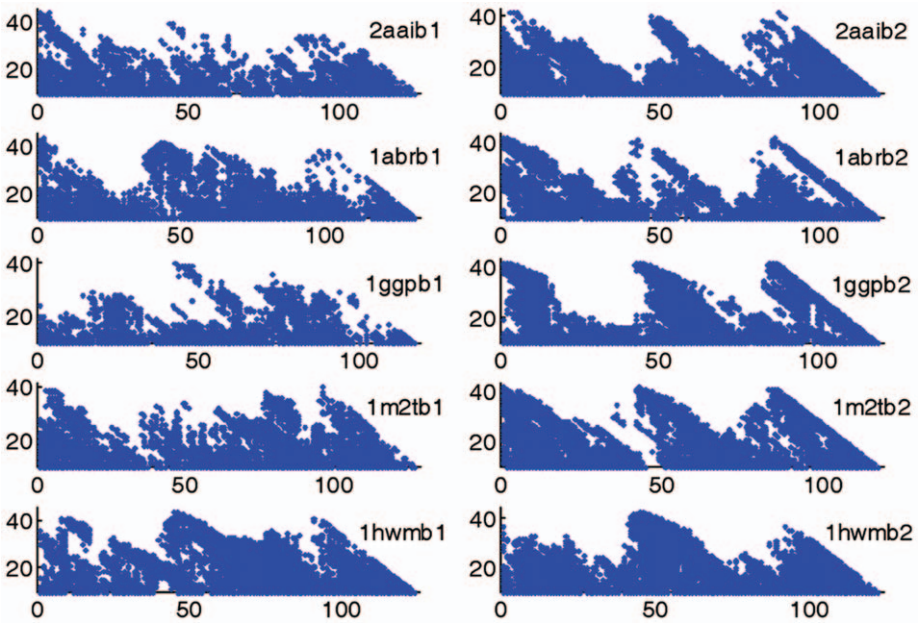

Figure 1. The MRPs of two domains in five representative protein chains. Column one is for the first domains and column two is for the second domains.
doi:10.1371/journal.pone.0014138.g001

Key structural residues of three-fold repetitions

Structure-based multi-sequence alignments. In the first and second domains of all the five representative protein chains of PCB family, we identified four repetitive motifs through structure-based multi-sequence alignments of trefoil units (Fig. 2) [30,31]. The repetitive motifs are $(I)_3$, $(L/M/V)_3$, $([I/L/V]X[I/L/M])_3$ and $(QXW)_3$, where X denotes any residue. They are totally composed of twenty-four residues and show three-fold repetitions (Fig. 3). The four different residues (I, L, M, V) are all large hydrophobic residues [32,33]. Generally, one residue is considered as buried if it has less than 25% solvent accessibility [34]. Using WHAT IF [35], we find that the four three-fold repetitive motifs are almost buried in the interior of their structures.

Consider RTB as an example to show the four three-fold repetitive (FTR) motifs in detail. The distribution of these motifs in the structure is illustrated in Fig. 3. It is shown that each beta strand has one motif and each trefoil unit has four motifs. Three-fold repetitions of the four motifs just correspond to the three-fold trefoil units in both domains. Moreover, these motifs are distributed symmetrically in the three-dimensional structures. The first motif is located at the top of the barrel structure, the fourth at the middle and the remaining two at the bottom. The FTR motifs seem to form the framework of the structures and act as key residues contributing to the formation of the symmetric structures, namely, the so-called key structural residues. Three previous works have reported some key structural residues in RTB [17,19,20]. Comparing them with the FTR motifs, we find they have a large overlap. Since other four representative protein chains show the same FTR motifs, they can be considered as the key structural residues of PCB family.

Inter-residue interactions. We use another approach to confirm the FTR motifs acting as key structural residues in PCB family. We calculate their inter-residue interactions. The key structural residues should have more interactions with others. RTB is selected as an example too. The average residue interaction number (RIN) of all residues, buried residues, and all residues in FTR motifs is 4.98, 6.31 and 8.50 respectively (Table 3). The average RIN of the FTR motifs is the largest among them (Table 4). The FTR motifs are mainly composed of buried residues. Generally, a buried residue likely has a large RIN.

Table 2. Sequence symmetries for five representative protein chains.

Protein chains	Domain I		Domain II		ΔR^a	$\Delta R/\langle R\rangle^b$ (%)	ΔS^a	$\Delta S/\langle S\rangle^b$ (%)
	R	S	R	S				
2aaib	0.80	0.42	0.70	0.60	−0.10	−13.3	0.18	35.3
1abrb	0.73	0.39	0.75	0.49	0.02	2.7	0.10	22.7
1ggpb	0.69	0.40	0.73	0.70	0.04	5.6	0.30	54.6
1m2tb	0.64	0.53	0.72	0.75	0.08	11.8	0.22	34.4
1hwmb	0.66	0.43	0.75	0.61	0.09	12.8	0.18	34.6

[a] $\Delta R = RII - RI$ and $\Delta S = SII - SI$;
[b] $\langle R\rangle = (RI+RII)$ and $\langle S\rangle = (SI+SII)$.
doi:10.1371/journal.pone.0014138.t002

	Barrel	Hairpin	Hairpin	Barrel
2aaib-1a :	PIVRIVGR	-CVDV--	GNAIQLWP--	-ANQLWTLKR
2aaib-1b :	-DNTIRSN	KCLTT--	GVYVMIYDC-	TDATRWQIWD
2aaib-1c :	-NGTIINP	LVL----	GTTLTVQTN-	-VSQGWLPTN
2aaib-2a :	FVTTIVG-	LCLQAN-	-GQVWIEDC-	-AEQQWALYA
2aaib-2b :	-DGSIRPQ	NCLTS--	ETVVKILSC-	--GQRWMFKN
2aaib-2c :	-DGTILNL	-VLDV--	--QIILYP--	-PNQIWLPLF
1abrb-1a :	--VRIGGR	-CVDV--	GNRIIMWK--	-ENQLWTL--
1abrb-1b :	--KTIR--	-CLTT--	GSYVMIYD--	AEATYWEIW-
1abrb-1c :	--GTIIN-	LVLSA--	GGTLTVQT--	-MRQGWRT--
1abrb-2a :	--TSISG-	LCMQAQ-	-SNVWMAD--	--EQQWAL--
1abrb-2b :	--GSIRS-	NCLTS--	GSTILLMGC-	-ASQRWVF--
1abrb-2c :	--GSIYS-	MVMDV--	--QIILWPY-	--NQIWLTLF
1ggpb-1a :	ATVRIAG-	FCADV--	-AAIILKKCA	-DNQLWTLKR
1ggpb-1b :	---TIRS-	--LTTAA	---AGIYDCT	--LSAWEIAD
1ggpb-1c :	---IINPA	--LSSG-	--DLGVQTNS	---QGWRTGN
1ggpb-2a :	--TQISGS	--MQAG-	--NLWMSEC-	-AEQQWALL-
1ggpb-2b :	--KSIRS-	NCLTS--	--TILLALC-	-ASQRWVFD-
1ggpb-2c :	---SILSL	-QMDSE-	---IILWWN-	--NQIWLALF
1m2tb-1a :	-IVRIVG-	MTVDV--	GNQIQLWPS-	DPNQLWTI--
1m2tb-1b :	--GTIR--	SCLTT--	GVYVMIFDC-	REATIWQIW-
1m2tb-1c :	--GTIIN-	LVLAA--	GTTLTVQTL-	-LGQGWLA--
1m2tb-2a :	RETTIYG-	LCMESA-	-GSVYVETC-	QENQRWAL--
1m2tb-2b :	--GSIRP-	QCLT---	-TVINIVSC-	-SGQRWVF--
1m2tb-2c :	--GAILN-	LAMDV--	-QRIIIYPA-	-PNQMWLPVP
1hwmb-1a :	FTRRIV--	-CVDV--	GTPIQLWP--	-RNQQWTFY-
1hwmb-1b :	--KTIRS-	KCMTA--	GSYIMITD--	-DATKWEVL-
1hwmb-1c :	--GSII--	LVMT---	RTTLLLENN-	-ASQGWTVS-
1hwmb-2a :	IATLIVG-	MCLQAN-	-NNVWMEDC-	-VQQQWALF-
1hwmb-2b :	--RTIRV-	LCVTS--	KDLIVIRKC-	-ATQRWFF--
1hwmb-2c :	---SVVN-	RVMDV--	-QEVIIFP--	-PNQQWRTQV

Figure 2. Structure based multiple sequence alignments of trefoil units in two domains of five representative protein chains. Conserved residues and most conserved residues are shaded gray and black respectively.
doi:10.1371/journal.pone.0014138.g002

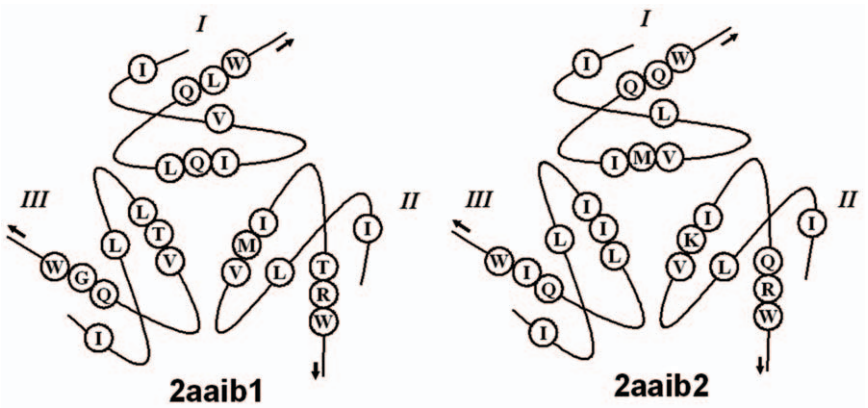

Figure 3. Schematic diagrams of four three-fold repetitive motifs (one-letter in circles) in two domains of RTB. The three trefoil units are shown in clockwise order. The arrows indicate the directions of beta strands.
doi:10.1371/journal.pone.0014138.g003

Table 3. The averaged residue interaction numbers and B-Factors.

Protein chains	Averaged RIN*			Averaged B-Factors*		
	A	B	R	A	B	R
2aaib	4.98	6.31	8.50	25.35	22.73	22.20
1abrb	5.08	6.33	8.92	23.12	18.00	17.26
1ggpb	4.82	6.18	8.33	19.32	14.61	11.68
1m2tb	4.81	5.95	8.79	40.55	37.03	36.51
1hwmb	5.10	6.03	8.92	20.88	16.52	16.37

*A-all residues, B-buried residues (eliminating buried residue in FTR motifs), R-FTR motifs.

doi:10.1371/journal.pone.0014138.t003

However, the average RIN of the FTR motifs are larger than that of other buried residues. This indicates that they may play the role of key structural residues. Furthermore, as shown in the plot of the RIN versus amino acids, the residues in the FTR motifs almost always have the locally largest RINs although they may not be the globally largest (Fig. 4A). As for other four representative protein chains, the results are similar (Table 3 and Fig. 4). Hence, it is a common feature that the residues of the FTR motifs have larger RIN and they play the role of hubs in the inter-residue interaction network.

Fig. 5 gives the interaction energies between the key structural residues of each representative protein chain (Fig. 5). In each plot there are six ''L''-like patterns along diagonal (each domain has three patterns), which denote the strong residue interactions. There are few interactions between different trefoil units. We compared these patterns with the positions of the key structural residues and found the six ''L''-like patterns are just corresponding to the six repetitions of the four motifs or the six trefoil units. Furthermore, the ''L''-like patterns indicate similar inter-

Table 4. The averaged residue interaction numbers (RINs) for FTR motifs in five representative protein chains. The superscript numbers are their indices in sequences.

Protein chains	Trefoil unit	Motif I	RIN	Motif II	RIN	Motif III	RIN	Motif IV	RIN
2aaib	2aaib-1a	I^{13}	7	V^{21}	7	$IQL^{34–36}$	9.33	$QLW^{47–49}$	8
	2aaib-1b	I^{57}	9	L^{64}	10	$VMI^{75–77}$	9.67	$TRW^{88–90}$	8.67
	2aaib-1c	I^{98}	7	L^{105}	9	$LTV^{118–120}$	8.33	$QGW^{129–131}$	9.33
	2aaib-2a	I^{144}	8	L^{152}	8	$VWI^{159–161}$	8	$QQW^{171–173}$	8.33
	2aaib-2b	I^{181}	8	L^{191}	8	$VKI^{202–204}$	7	$QRW^{214–216}$	11
	2aaib-2c	I^{224}	7	V^{233}	9	$IIL^{245–247}$	7.67	$QIW^{256–258}$	8.33
1abrb	1abrb-1a	$\mathbf{I}^{18}$	8	$\mathbf{V}^{26}$	9	$\mathbf{IIM}^{39–41}$	10	$\mathbf{QLW}^{52–54}$	8
	1abrb-1b	$\mathbf{I}^{62}$	9	$\mathbf{L}^{69}$	8	$\mathbf{VMI}^{80–82}$	10	$\mathbf{TYW}^{93–95}$	8.33
	1abrb-1c	$\mathbf{I}^{103}$	7	$\mathbf{L}^{110}$	8	$\mathbf{LTV}^{123–125}$	8.67	$\mathbf{QGW}^{134–136}$	10
	1abrb-2a	$\mathbf{I}^{149}$	8	$\mathbf{M}^{157}$	10	$\mathbf{VWM}^{164–166}$	7.67	$\mathbf{QQW}^{176–178}$	9.33
	1abrb-2b	$\mathbf{I}^{186}$	8	$\mathbf{L}^{196}$	8	$\mathbf{ILL}^{207–209}$	7.67	$\mathbf{QRW}^{219–221}$	11.67
	1abrb-2c	$\mathbf{I}^{229}$	7	$\mathbf{M}^{238}$	9	$\mathbf{IIL}^{250–252}$	9.67	$\mathbf{QIW}^{261–263}$	8.67
1ggpb	1ggpb-1a	$\mathbf{I}^{18}$	7	$\mathbf{A}^{26}$	6	$\mathbf{IIL}^{39–41}$	10	$\mathbf{QLW}^{52–54}$	8
	1ggpb-1b	$\mathbf{I}^{62}$	8	$\mathbf{L}^{69}$	9	$\mathbf{AGI}^{81–83}$	8	$\mathbf{SAW}^{93–95}$	8
	1ggpb-1c	$\mathbf{I}^{104}$	6	$\mathbf{L}^{112}$	8	$\mathbf{LGV}^{123–125}$	7	$\mathbf{QGW}^{134–136}$	9.33
	1ggpb-2a	$\mathbf{I}^{149}$	7	$\mathbf{M}^{157}$	11	$\mathbf{LWM}^{164–166}$	10	$\mathbf{QQW}^{176–178}$	9
	1ggpb-2b	$\mathbf{I}^{186}$	7	$\mathbf{L}^{196}$	9	$\mathbf{ILL}^{207–209}$	6.33	$\mathbf{QRW}^{219–221}$	11
	1ggpb-2c	$\mathbf{I}^{229}$	6	$\mathbf{M}^{238}$	9	$\mathbf{IIL}^{250–252}$	8.33	$\mathbf{QIW}^{261–263}$	7.33
1m2tb	1m2tb-1a	$\mathbf{I}^{262}$	7	$\mathbf{V}^{269}$	7	$\mathbf{IQL}^{282–284}$	9	$\mathbf{QLW}^{295–297}$	7.67
	1m2tb-1b	$\mathbf{I}^{305}$	8	$\mathbf{L}^{312}$	10	$\mathbf{VMI}^{323–325}$	10	$\mathbf{TIW}^{336–338}$	8.67
	1m2tb-1c	$\mathbf{I}^{346}$	8	$\mathbf{L}^{355}$	8	$\mathbf{LTV}^{366–368}$	7.67	$\mathbf{QGW}^{377–379}$	9.33
	1m2tb-2a	$\mathbf{I}^{392}$	9	$\mathbf{M}^{400}$	9	$\mathbf{VYV}^{407–409}$	8.33	$\mathbf{QGW}^{419–421}$	9.67
	1m2tb-2b	$\mathbf{I}^{429}$	8	$\mathbf{L}^{439}$	11	$\mathbf{INI}^{450–452}$	9	$\mathbf{QRW}^{462–464}$	10.67
	1m2tb-2c	$\mathbf{I}^{472}$	6	$\mathbf{M}^{481}$	10	$\mathbf{III}^{493–495}$	9	$\mathbf{QMW}^{504–506}$	8
1hwmb	1hwm-1a	$\mathbf{I}^{15}$	8	$\mathbf{V}^{23}$	7	$\mathbf{IQL}^{36–38}$	10.33	$\mathbf{QQW}^{47–49}$	8.33
	1hwm-1b	$\mathbf{I}^{57}$	8	$\mathbf{M}^{64}$	11	$\mathbf{IMI}^{75–77}$	10	$\mathbf{TKW}^{88–90}$	8.33
	1hwm-1c	$\mathbf{I}^{98}$	7	$\mathbf{M}^{107}$	9	$\mathbf{LLL}^{118–120}$	9	$\mathbf{QGW}^{129–131}$	10.67
	1hwm-2a	$\mathbf{I}^{144}$	6	$\mathbf{L}^{152}$	7	$\mathbf{VWM}^{161–163}$	8.33	$\mathbf{QQW}^{173–175}$	9.67
	1hwm-2b	$\mathbf{I}^{183}$	8	$\mathbf{V}^{193}$	9	$\mathbf{IVI}^{204–206}$	7.67	$\mathbf{QRW}^{215–217}$	11.67
	1hwm-2c	$\mathbf{I}^{226}$	6	$\mathbf{M}^{234}$	9	$\mathbf{VII}^{246–248}$	7.67	$\mathbf{QQW}^{257–259}$	9.33

doi:10.1371/journal.pone.0014138.t004

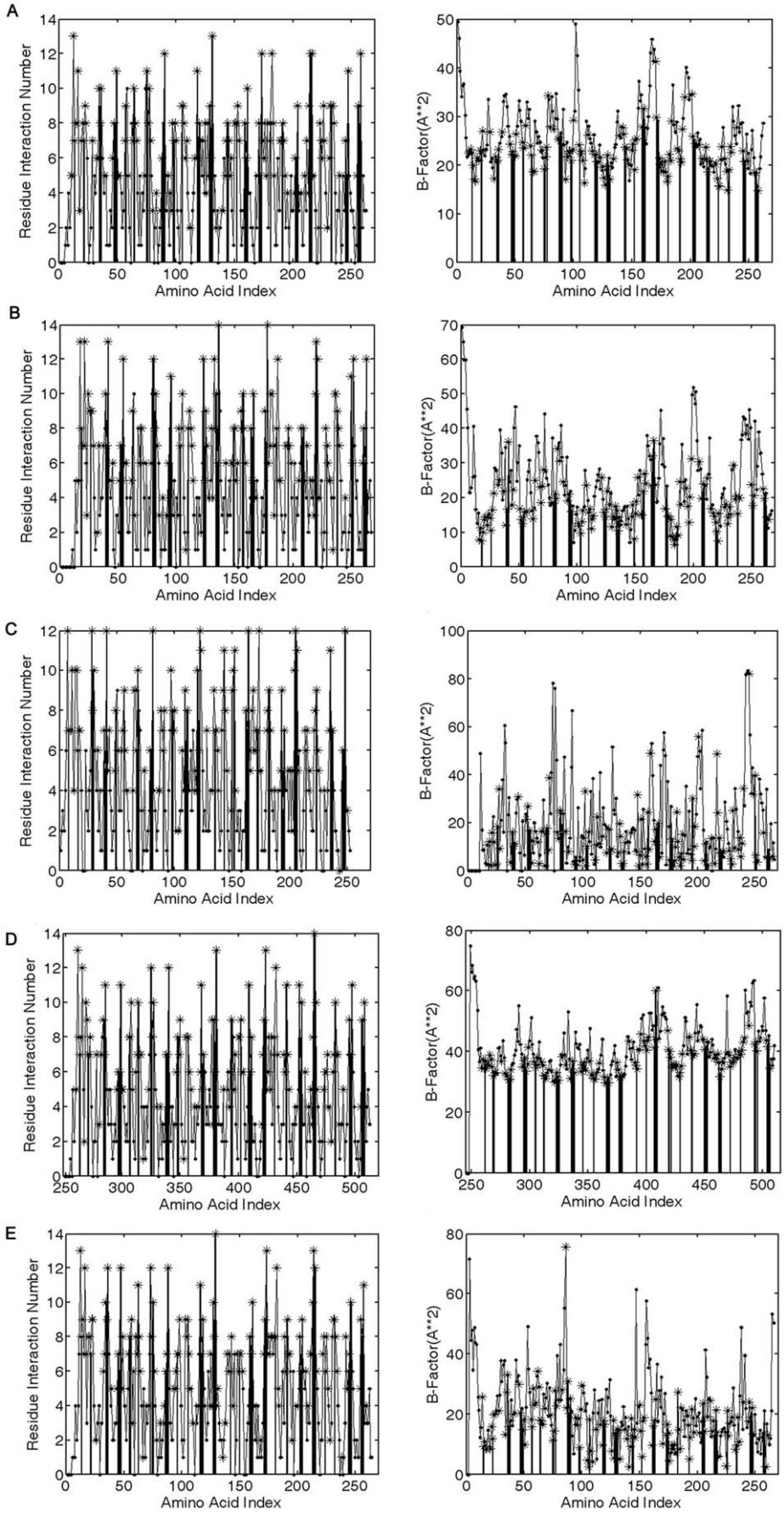
A
Residue Interaction Number
Amino Acid Index
B-Factor(A**2)
Amino Acid Index
B
Residue Interaction Number
Amino Acid Index
B-Factor(A**2)
Amino Acid Index
C
Residue Interaction Number
Amino Acid Index
B-Factor(A**2)
Amino Acid Index
D
Residue Interaction Number
Amino Acid Index
B-Factor(A**2)
Amino Acid Index
E
Residue Interaction Number
Amino Acid Index
B-Factor(A**2)
Amino Acid Index

Figure 4. The residue interaction numbers (column one) and B-Factors (column two) versus amino acid index for 2aaib(A), 1abrb(B), 1ggpb(C), 1m2tb(D) and 1hwmb(E). The symbols represent different type of residues: four three-fold repetitive motifs (bar), buried residues (star) and remaining residues (dot).
doi:10.1371/journal.pone.0014138.g004

residue interaction patterns in every trefoil unit. Therefore, every trefoil units not only have similar key structural residues but also similar strong residue interactions. This suggests that the repetitive key structural residues may determine the three-fold trefoil units. Finally, the "L"-like patterns show that the second motifs, $(L/M/V)_3$, have stronger interactions with other motifs. This may be that the second motifs are closer to other three motifs (Fig. 3).

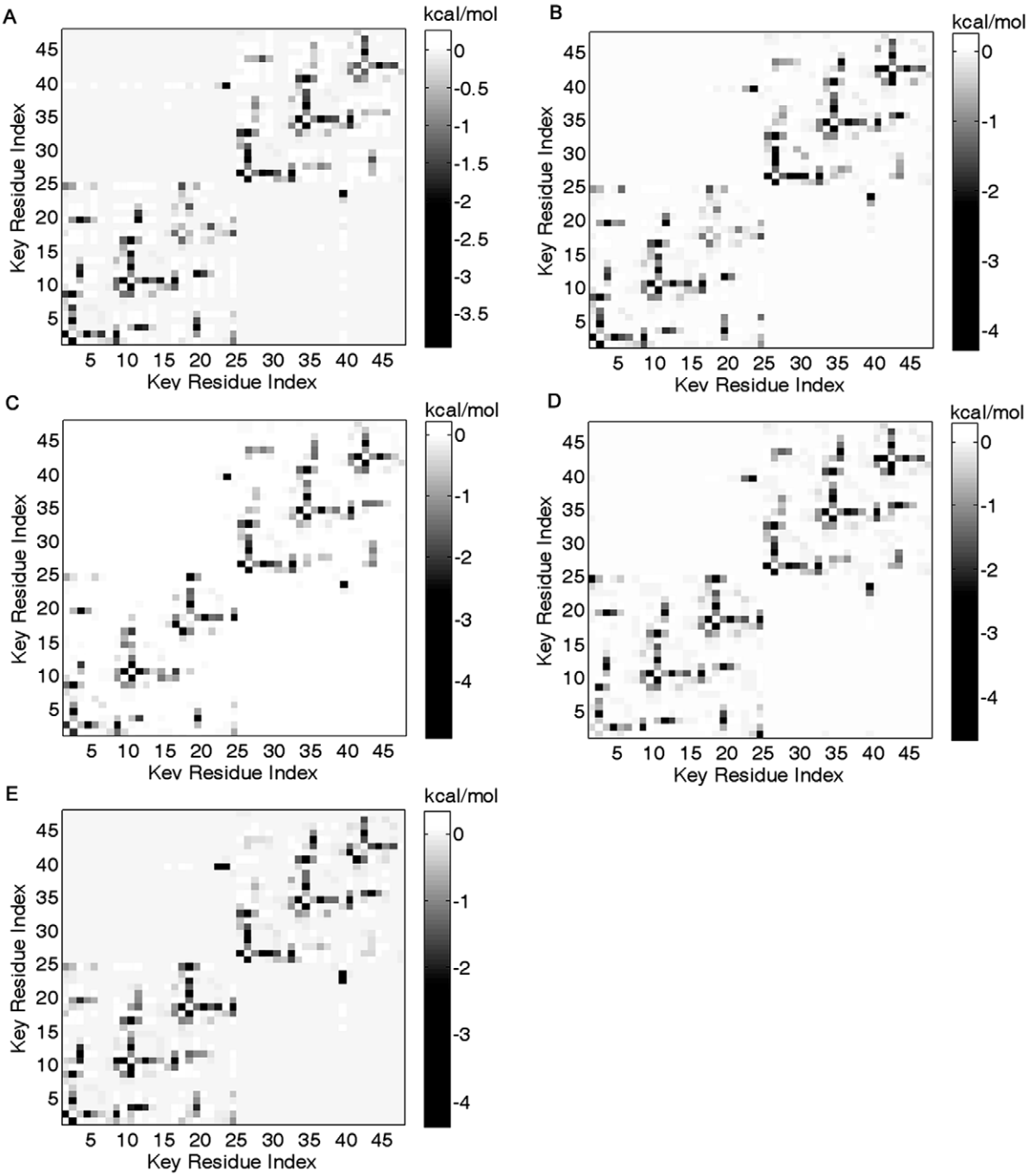

Figure 5. The potential energies of residue interactions between key structural residues for 2aaib(A), 1abrb(B), 1ggpb(C), 1m2tb(D) and 1hwmb(E). The key structural residues are arrayed along two axes according to their orders in the sequence. The magnitude of the interactions is indicated by the colorbar.
doi:10.1371/journal.pone.0014138.g005

B-factors. From an experimental point of view, since the key structural residues act as the skeleton of structures, they should be much more constrained than other residues. The B-factors retrieved from PDB file are generally characteristic of the degree of atomic constraint. We average the B-factors of all heavy atoms in one residue and designate the mean as the B-factor of this residue. For RTB, the average B-factor of all residues, buried residues, and all residues in the FTR motifs is 25.35, 22.73 and 22.20 respectively (Table 3). Clearly, the FTR motifs have the smallest average B-factor. Furthermore, as shown in the plot of the B-factors versus amino acids, the residues in the FTR motifs always have the locally smallest B-factors (Fig. 4A). As for other four representative protein chains, we gain the same results as RTB (Table 3 and Fig. 4). Therefore, the FTR motifs seem to be most strongly constrained. In summary, both the inter-residue interactions and B-factors also suggest that the FTR motifs may be key structural residues in PCB family.

Extension to all beta-trefoil folds

Are the three-fold repetitive key structural residues special for beta-trefoil proteins in PCB family or common for all proteins sharing beta-trefoil fold? In our recently published paper [12], thirty protein chains/domains were selected as the representatives of the presently known proteins with beta-trefoil fold. Because the two domains of 1vcla are homologous and also because only the atomic coordinates of alpha carbon atoms can be retrieved from PDB database for 2ila-, twenty-eight protein chains/domains are set as the representatives (Table S1 in Supporting file S1). Two algorithms, CE and TM-align integrated in STRAP [36–38], are used to do their structure-based multiple sequence alignments. Interestingly, both alignment methods detected similar twelve conserved motifs (Figure S1 and Figure S2 in Supporting file S1). We compare them with the FTR motifs and find they are similar. The twelve conserved motifs also show three-fold repetitions. In addition, we notice the twelve conserved residues as well as the FTR motifs are mainly composed of large hydrophobic residues (I, L, V, F, W), which is in agreement with the previous prediction by Murzin *et al.* that the large hydrophobic residues stabilize the beta-trefoil fold [17]. Recently, Chaudhuri *et al.* [39] pointed out that at least 80% propellers across families are similar at a level indicative of homology. To support their conclusion, one evidence is that all propellers share similar key sequence motifs across families. We [23,24] also studied the key residues in the protein domain G from transducin (PDB id: 1tbg), which is a propellerlike protein composed of seven similar blades or called WD-repeats and has a high structural symmetry. From a structure-based sequence alignment, it can be observed that there are five residues that are almost totally invariant in each repeat of the protein. These structurally conserved residues connect the outer strand of each blade to the inner three strands of the next blade, and are certainly considered as key residues critical for the structural stability of the G protein. We calculated the contact energies by all-atom force field and found that the residues with lowest contact energies (or strong inter-residue interactions) are in good agreement with the structurally conserved residues identified previously. Here, the proteins with beta-trefoil fold show the similar situation. All evidences suggest that the three-fold repetition of key structural residues should dominate the three-fold symmetric structures. Thus, the contradiction of different degrees of structure and sequence symmetries of the two domains of PCB family proteins can be interpreted in terms of similar key structural residues.

In conclusion, we analyzed the proteins with two repeated beta-trefoil domains in Plant Cytotoxin B-chain family and all presently known beta-trefoil proteins by three different methods and show that some key structural residues may play important roles in the formation of the three-fold symmetric structure of beta-trefoil fold. These key structural residues are (i) buried residues, (ii) symmetrically located in the structure, and (iii) have large residue interaction numbers and small B-Factors. This result may be helpful to design de novo proteins.

Supporting Information

Supporting File S1 Supplementary data (Table S1; Figures S1, S2)

Found at: doi:10.1371/journal.pone.0014138.s001 (3.50 MB DOC)

Acknowledgments

We thanks Prof. Anna Tramontano and Dr. Changjun Chen for valuable suggestions.

Author Contributions

Conceived and designed the experiments: ML YX. Performed the experiments: JF ML YH. Analyzed the data: JF ML. Wrote the paper: ML YX.

References

1. Brych SR, Blaber SI, Logan TM, Blaber M (2001) Structure and stability effects of mutations designed to increase the primary sequence symmetry within the core region of a beta-trefoil. Protein Sci 10: 2587–2599.
2. Lang D, Thoma R, Henn-Sax M, Sterner R, Ilmanns M (2003) Structural evidence for evolution of the alpha/beta barrel scaffold by gene duplication and fusion. Science 289: 1546–1550.
3. McLachlan AD (1976) Evidence for gene duplication in collagen. J Mol Biol 107: 159–174.
4. Giuliani A, Benigni R, Zbilut JP, Webber JCL, Sirabella P, et al. (2002) Nonlinear signal analysis methods in the elucidation of protein sequence-structure relationships. Chem Rev 102: 1471–1491.
5. Laskin AA, Kudryashov NA, Skryabin KG, Korotkov EV (2005) Latent periodicity of serine-threonine and tyrosine protein kinases and other protein families. Comput Biol Chem 29: 229–243.
6. Rackovsky S (1998) "Hidden" sequence periodicities and protein architecture. Proc Natl Acad Sci USA 95: 8580–8584.
7. Soding J, Remmert M, Biegert A (2006) HHrep: de novo protein repeat detection and the origin of TIM barrels. Nucleic Acids Res 34: W137–W142.
8. Szklarczyk R, Heringa J (2004) Tracking repeats using significance and transitivity. Bioinformatics 20 Suppl 1: i311–317.
9. Huang YZ, Li MF, Xiao Y (2007) Nonlinear analysis of sequence repeats of multi-domain proteins. Chaos Solitons Fractals 34: 782–786.
10. Huang YZ, Xiao Y (2007) Detection of gene duplication signals of Ig folds from their amino acid sequences. Proteins 68: 267–272.
11. Ji XF, Chen HL, Xiao Y (2007) Hidden symmetries in the primary sequences of beta-barrel family. Comput Biol Chem 31: 61–63.
12. Li M, Huang Y, Xiao Y (2008) Effects of external interactions on protein sequence-structure relations of beta-trefoil fold. Proteins 72: 1161–1170.
13. Li MF, Huang YZ, Xu RZ, Xiao Y (2005) Nonlinear analysis of sequence symmetry of beta-trefoil family proteins. Chaos Solitons Fractals 25: 491–497.
14. Wang XC, Huang YZ, Xiao Y (2008) Structural-symmetry-related sequence patterns of the proteins of beta-propeller family. J Mol Graph Model 26: 829–837.
15. Xu RZ, Xiao Y (2005) A common sequence-associated physicochemical feature for proteins of beta-trefoil family. Comput Biol Chem 29: 79–82.
16. McLachlan AD (1979) Three-fold structural pattern in the soybean typsin inhibitor (Kunitz). J Mol Biol 133: 557–563.
17. Murzin AG, Lesk AM, Chothia C (1992) Beta-trefoil fold patterns of structure and sequence in the Kunitz inhibitors interleukins-1beta and 1alpha and Fibroblast growth factors. J Mol Biol 223: 531–543.
18. Rutenber E, Ready M, Robertus JD (1987) Structure and evolution of ricin B chain. Nature 326: 624–626.
19. Hazes B (1996) The (QxW)3 domain: a flexible lectin scaffold. Protein Sci 5: 1490–1501.

20. Rutenber E, Robertus JD (1991) Structure of ricin B-chain at 2.5 Å resolution. Proteins 10: 260–269.
21. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
22. Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG, et al. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
23. Chen CJ, Li L, Xiao Y (2007) All-atom contact potential approach to protein thermostablity analysis. Biopolymers 85: 28–37.
24. Chen CJ, Li L, Xiao Y (2006) Identification of key residues in proteins by using their physical characters. Phys Rev E 73: 041926.
25. Qiu D, Shenkin PS, Hollinger FP, Still WC (1997) The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii J Phys Chem A 101: 3005–3014.
26. Still VC, Tempezvk A, Hawley RC, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 112: 6127–6129.
27. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Eva seck JD, et al. (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102: 3586–3617.
28. Ren P, Ponder JW (2003) Polarizable atomic multipole water model for molecular mechanics simulation. J Phys Chem B 107: 5933–5947.
29. Shih ESC, Hwang MJ (2004) Alternative alignments from comparison of protein structures. Proteins 56: 519–527.
30. Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinformatics.
31. Nicholas KB, Nicholas HB, Deerfield DW (1997) GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEWNEWS 4: 14.
32. Li TP, Fan K, Wang J, Wang W (2003) Reduction of protein sequence complexity by residue grouping. Protein Eng Des Sel 16: 323–330.
33. Riddle DS, Santiago JV, Bray ST, Doshi N, Grantcharova VP, et al. (1997) Functional rapidly folding proteins from simplified amino acid sequences. Nat Struc Biol 4: 805–809.
34. Bloom JD, Drummond DA, Arnold FH, Wilke CO (2006) Structural determinants of the rate of protein evolution in yeast. Mol Biol Evol 23: 1751–1761.
35. Vriend G (1990) WHAT IF: A molecular modeling and drug design program. J Mol Graph 8: 52–56.
36. Gille C, Frömmel C (2001) STRAP: editor for STRuctural Alignments of Proteins. Bioinformatics 17: 377–378.
37. Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng Des Sel 11: 739–747.
38. Zhang Y, Skolnick J (2005) TM-align: A protein structure alignment algorithm based on TM-score. Nucleic Acids Res 33: 2302–2309.
39. Chaudhuri I, Soding J, Lupas AN (2008) Evolution of the beta-propeller fold. Proteins 71: 795–803.