

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900

Open access books available

186,000

International authors and editors

200M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



# Protein Structure Alphabetic Alignment

Jiaan Yang<sup>1,\*</sup> and Wei-Hua Lee<sup>2</sup>

<sup>1</sup>MicrotechNano LLC

<sup>2</sup>Department of Pediatrics and Anatomy,  
Indiana University School of Medicine, Indianapolis,  
USA

## 1. Introduction

This study presents a fast approach to compare protein 3D structures with protein structure alphabetic alignment method. First, the folding shape of 5 consecutive residues is represented by protein folding shape code (PFSC) (Yang, 2008) and thus protein folding conformation can be completely described by PFSC. With complete description for folding shape along the backbone, any protein with given 3D structure can be converted into an alphabetic string and aligned for comparison. Consequently, this approach is able to provide a unique score to assess the global similarity in structure while it supplies an alignment table for analysis of local structure. Several sets of proteins with diverse homology or different degrees in complexity are compared. The results demonstrate that this approach provides an efficient method for protein structure alignment which is significant for protein structure search with high throughput screening of protein database.

Comparison of protein structures is a challenging task because of complication of 3D structure which involves ambiguous procedure in analysis. First, protein structure obviously is not a simple geometric subject. It is not easily to superimpose two proteins together because the specific emphasis of one portion of structures may cause other parts with similar structures to orient toward different directions in geometric space. In practice, an individual turning point in protein may overshadow entire similarity between two structures. Second, it is hard to develop a uniform process to compare the proteins with different homologies. For protein structures with identical amino acid sequence or with mutation in sequence, the comparison often requires sensitivity to distinguish the conformers with higher similarity in structure. However, for proteins with drastic difference in structural conformation, the good comparison expects a consistent procedure to evaluate the similarity in variant cases. Significant variation of protein conformation is primarily determinate by sequence difference, which affects the formation of hydrogen bond, van der Waals force interaction and disulfide bridge. Also, the protein conformation may be changed by other factors, such as solvent effect, protein-protein interaction, ligand docking and so on. From view of topological order of secondary structure, if two structures belong to different categories in protein classification, such as under different families, superfamilies,

---

\* Corresponding Author

folds and class, the structural comparison becomes more difficult. An ideal method should have a consistent process to assess the similarity for proteins with various homologies in structures.

Many established methods for protein structure comparison were developed and evaluated (Kolodny et al., 2005). DALI method (Holm & Sander, 1993; Holm & Park, 2000) is frequently used in protein structure comparison based on the alignment of distance matrices. LGA method (Zemla, 2003) generates the different local superposition to detect the regions where the structures are more similar. CE method (Shindyalov & Bourne, 1998) is for calculating pairwise structure alignments. Two proteins are aligned by using characteristics of local geometry between C-alpha positions. Heuristics are used in defining a set of optimal paths joining aligned fragment pairs with gaps. The path with the best RMSD is subject to dynamic programming to achieve an optimal alignment. 3D-BLAST method (Mavridis & Ritchie, 2010) is developed to align the protein structures using 3D spherical polar Fourier for protein shape. There are many of well known methods, including DAL (Kryshtafovych et al., 2005; Hvidsten et al., 2003), MAMMOTH (Ortiz et al., 2002), ProSup (Lackner et al., 2000), VAST (Madej et al., 1995; Gibrat et al., 1996), SSAP (Taylor & Orengo, 1989), STRUCTAL (Subbiah et al., 1993), LSQMAN (Kleywegt & Jones, 1994), SSM (Krissinel & Henrick, 2004), FlexProt (Shatsky et al., 2002), FATCAT (Yuzhen & Adam, (2003) and TM-align/score (Zhang Y & Skolnick J. 2005).

For optimistic solution, most of methods attempt to find out higher number of equivalent residues while obtain lower value of root-mean-square deviation (RMSD) through superimposition of protein 3D structures or alignment of structural fragments. Unfortunately, it is tough to optimize these two parameters simultaneously because the intention of higher number of equivalent residues leans higher RMSD, or the favor of lower RMSD leads less number of equivalent residues. In protein structural superimposition two factors, the cutoff distance for RMSD and the initiative focusing location, may be artificially adjusted. These artificial factors are not unique for various methods and they may be changed on case-by-case basis with using same method. Apparently it directly affects the outcome of protein structural comparison. So, it is not surprised that with different methods or even same method, it may produce different values of RMSD and different numbers of equivalent residues. Consequently, different methods may generate unlike rank of similarity in assessment of proteins structures.

The structural alignment is a popular approach for protein comparison which has been developed by different strategies. First strategy is the rigid body alignment, which directly superimposes two proteins with possible best fitting to obtain the lowest RMSD and higher number of equivalent residues. Second strategy is the non-rigid body alignment, which allows smaller structural fragments of proteins with certain flexibility to orient or shift for better fittings, and then adopts various algorithms of measurement for similarity. However, no matter how the protein structure is partitioned, the acquisition of optimum result still involves obtaining the lowest RMSD and highest number of equivalent residues, which are two of contradictory adjustments. The attempt of direct alignment of geometric objects is difficult because no unique resolution is able to handle a geometric object of more than three points with no double superposition. In order to avoid direct alignment of geometric objects, the structural alphabetic alignment is a solution.

The earliest application of structural alphabets was the reorganization of the secondary structure in protein, and then adopted letter "A" for  $\alpha$ -helix, "E" for  $\beta$ -strand and "C" for coil. Furthermore, the structural alphabetic methods (Brevern et al., 2000; Kolodny et al., 2002; Micheletti et al., 2000; Rooman et al., 1990; Schuchhardt et al., 1996; Unger et al., 1989; Sander et al., 2006; Tung et al., 2007; Ku & Hu, 2008; Karplus et al., 2003; Murphy et al., 2000) have been developed for more detail assignment for representative folding shapes. Different approaches in structure alphabets defined different length of peptide and adopted different number of prototypes for folding shapes. With pentapeptide motif, Protein Blocks (PBs) method determined 16 of folding shapes and use alphabets represent these primary prototypes (Kolodny et al., 2002). Thus, it was applied to protein structural alignment (Brevern, 2005; Joseph et al., 2011). Based on different designs in structural alphabets, a variety of methods of structural alphabetic alignment have been developed (Ku & Hu, 2008; Karplus et al., 2003; Tyagi et al., 2006; Melo & Marti-Renom, 2006; Friedberg et al., 2007; Tyagi et al., 2006; Guyon et al., 2004; Sacan et al., 2008; Wang & Zheng, 2008). The performance of structural alphabetic alignment approaches are significantly faster than the methods based on 3D structural comparison, and the unambiguousness is avoided during structural superimposition. However, to date the prototypes of folding shapes in structural alphabetic methods are obtained by observations from training database, and then the primary motifs for folding patterns are determined by statistics judgment. With training database, the experimental observations may collect most of folding patterns with higher frequency of appearance in protein, but may leave out certain folding shapes as leak because of its rare appearance in proteins. Also, each prototype of folding pattern or alphabet is isolated without association meaning. A recently developed structural alphabets approach, protein folding shape code (PFSC) (Yang, 2008), overcomes the shortcomings, which is comprised by complete folding patterns for motif of five residues, and all folding patterns have the meaningful interrelated relationship.

In this study, a set of 27 PFSC vectors is used to describe the folding shapes of protein structure, and to apply to structural alignment. The 27 PFSC vectors are rigorously obtained by mathematical derivation to cover an enclosed space, and represent all possible folding shapes for any five of successive  $C\alpha$  atoms (Yang, 2008). The 27 PFSC vectors are symbolized by 26 alphabetic letters plus \$ symbol, which are capable completely to describe the change of protein folding shapes along protein backbone from N-terminus to C-terminus without gap. With complete description of folding shape for any given protein 3D structure, a consistent method for alignment of protein structures is developed, which is able to assess the structural similarity with various homologies.

## 2. Method

### 2.1 Conversion of alphabet description

The protein 3D structure is first converted into alphabetic description with protein folding shape code (PFSC) (Yang, 2008). With PFSC approach, a set of 27 PFSC vectors represent all possible folding shapes for each five successive  $C\alpha$  atoms. The 27 PFSC vectors, prototypes of folding shapes and alphabets are shown on top of Fig.1. The 27 PFSC vectors are able to map all possible folding shapes, including the regular secondary structure and irregular coil and loop. The 27 PFSC alphabetic codes are able to describe the change of protein folding shapes along based on five successive  $C\alpha$  atoms. It provides a complete alphabetic

description of protein structural conformation from N-terminus to C-terminus without gap. To take protein structure of 8DFR (PDB ID) as sample, the folding shape of each of each five successive C $\alpha$  atoms is converted into one of 27 PFSC alphabetic letters along protein backbone. Consequently, the structural folding conformation is expressed by the PFSC alphabetic description and is demonstrated on bottom of Fig. 1.

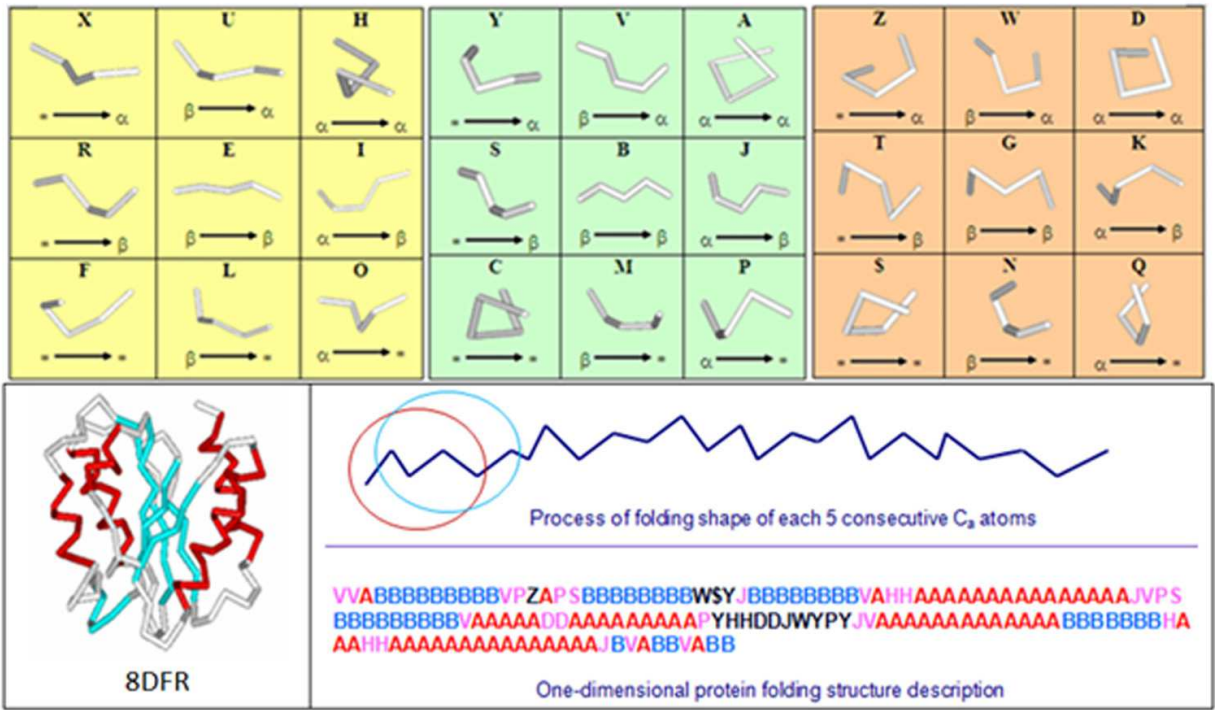


Fig. 1. The 27 protein folding shape code and the conversion of protein alphabetic description. Top: Three blocks represent three regions of pitch distance of motif for five residues; the nine vectors in each block represent the nine folding shape patterns determined by two torsion angles; each vector is simultaneously represented by a letter, a folding shape pattern and an arrow. The vector characteristic is represented by an arrow line. The “ $\alpha$ ”, “ $\beta$ ” or “\*” at each end of vector indicates the folding features similar to  $\alpha$ -helix,  $\beta$ -strand or random coil respectively. Bottom: 8DFR (PDB ID) is a sample to illustrate how protein backbone conformation is converted into PFSC alphabetic description. The folding shape of each five successive C-alpha atoms in a protein backbone from N-terminal to C-terminal is converted into alphabetic description. “A” represents a typical alpha helix with red color and “B” beta strand with blue. The folding shape is derived from secondary structure in pink color, and shape for loop or coil in black.

2.2 Protein Folding Shape Alignment (PFSA)

With one-dimensional PFSC alphabetic description, the protein conformation structures are able to be compared by protein folding shape alignment (PFSA) approach (Yang, 2011). Similarly as sequence alignment, the PFSC alphabetic strings for proteins are aligned to match the similarity. The Needleman-Wunsch algorithm of dynamic programming technique (Needleman SB & Wunsch, 1970) is used in the PFSA for structural alignment. Therefore, the structural similarity of two proteins is able to be discovered by structural alphabetic alignment with PFSA approach.

In PFSA approach, a substitution matrix for 27 PFSC vectors is defined according relationship of vector similarity. Within substitution matrix  $S$ , each element of similarity matrix  $S[i, j]$  is determined by the similarity between  $PFSC[i]$  and  $PFSC[j]$ , which is determined by the integrated relationship of 27 PFSC vectors (Yang, 2008). For identical folding shape, the value  $S[i, i] = 2$ ; for analogous folding shape, the value  $S[i, j] = 1$  and for different folding shape, the value  $S[i, j] = 0$ . The substitution matrix  $S$  is displayed in Table 1. In next step, a similarity matrix for two proteins is constructed. According substitution matrix  $S$ , all elements of similarity matrix  $M$  are able to be determined. It assumes that  $m$  and  $n$  are the lengths of amino acid sequence for protein A and B respectively. Thus the lengths PFSC strings for protein A and B are  $m-4$  and  $n-4$ . With the protein folding shape strings of protein A[3... $m-2$ ] and protein B[3... $n-2$ ], a similarity matrix  $M$  with  $(m-4) \times (n-4)$  dimension is constructed for a pair proteins of A and B in structural alignment. The third step is to obtain a sum matrix by computing the elements of the similarity matrix according the Needleman-Wunsch algorithm. With the sum matrix, an optimized structural alignment is obtained based on tracing elements from the largest value to smaller value. When the track shifts from diagonal in the sum matrix, it actually tries to reduce the mismatch by insertion of gap for match of identical or analogous folding shape.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	\$
A	2			1				1	1							1						1			1		
B		2			1		1			1			1						1			1					
C			2			1							1			1			1						1		1
D	1			2							1						1						1			1	
E		1			2				1			1						1			1						
F			1			2						1			1			1							1		
G		1					2				1			1						1		1					
H	1							2	1						1						1			1			
I					1			1	2	1					1			1									
J	1	1							1	2	1					1			1								
K				1			1			1	2						1			1							
L					1	1						2	1		1						1						
M		1	1									1	2	1		1						1					
N							1						1	2			1						1				1
O						1		1	1			1			2	1											
P	1		1							1			1		1	2	1										
Q				1							1			1		1	2										1
R					1	1			1									2	1						1		
S		1	1							1								1	2	1					1		
T							1				1								1	2						1	1
U					1			1				1									2	1		1			
V	1	1											1								1	2	1		1		
W				1			1							1								1	2			1	
X						1		1										1			1				2	1	
Y	1		1																1			1			1	2	1
Z				1																1			1		1	2	1
\$			1											1		1			1							1	2

Table 1. The substitution matrix of 27 PFSC vectors. The top row and the left column list the 27 PFSC letter. The value of element in substitution matrix is 2 for identical folding shape code; 1 for analogues folding shape code; empty means zero for different folding shape code.

## 2.3 Similarity score

With optimized alignment, the protein structural similarity score is calculated. Each match of identical folding shape is assigned by 2; analogous folding shape 1; different folding shape 0; penalty of open a gap -2 and penalty of extended a gap -0.25. The value of protein folding structure alignment score (PFSA-S) is determined by the total contribution of identical folding shapes, analogous folding shapes and gaps. The score is normalized with below function.

$$PFSA-S = \frac{2 \times ID_{FS} + 1 \times AN_{FS} - (2 \times GPO + 0.25 \times GPE)}{2 \times TSQ}$$

Here  $ID_{FS}$  is the number of identical folding shapes,  $AN_{FS}$  the number of analogous folding shape,  $GPO$  the number of open gaps,  $GPE$  the number of extended gaps and  $TSQ$  is the length of PFSC of protein. The denominator in formula,  $2 \times TSQ$ , assures the value of PFSA-S to equal numeral one for comparison of two identical structures. When similarity between two protein structures decreases, the value of PFSA-S will decrease. When two proteins have less similarity, the structural alignment produces larger number of gaps, which may give negative value for PFSA-S and signify no noteworthy similarity existing. For normalization, the value of PFSA-S is limited to larger or equal to zero, so any negative value of PFSA-S is converted as zero. Therefore, the PFSA approach provides a normalized score between one and zero to evaluate the protein structural similarity.

## 2.4 Alignment table

With comparison of one-dimensional alphabetic strings for protein folding conformation, the PFSA alignment table is generated. There are two types of alignment tables, i.e. sequence-dependence mode and sequence-independence mode. For same protein or proteins with mutation, the structural alignment for conformation analysis may prefer the sequence-dependent mode because gap insertion is not necessary. For proteins with different sequence and size, the structural alignment takes the advantage of the sequence-independent mode, which allows inserting gaps to obtain the best match in local structural similarity.

The PFSA alignment table possesses several features. First, the alignment table is able explicitly to reveal the similarity and dissimilarity for local structure. Second, the alignment table exhibits how all similar fragments are matched or shifted with insertion of gaps. Third, it intuitively display how the structural folding shape associates with the corresponding residue of five consecutive amino acids, which is able to assist the analysis of relationship between amino acid sequence-structure-function in protein.

## 3. Results

### 3.1 Conformation analysis

Protein structure 1M2F (PDB ID) has 25 conformers obtained by NMR spectroscopy and show in Fig. 2(A). 1M2E (PDB ID) in Fig. 2(B) is the average structural models of 25 conformers of 1M2F (Williams et al., 2002). All of these structures apparently have identical

sequence and similar 3D structural conformations. To differentiate the structures with higher similarity requests a tool with higher sensitivity to distinguish each conformer in global and local structure. With PFSA approach, each conformer of protein 1M2F and the structure of 1M2E are converted into one-dimensional PFSC alphabetic description, and then are aligned for comparison. The PFSA alignment table is displayed in Table 2.

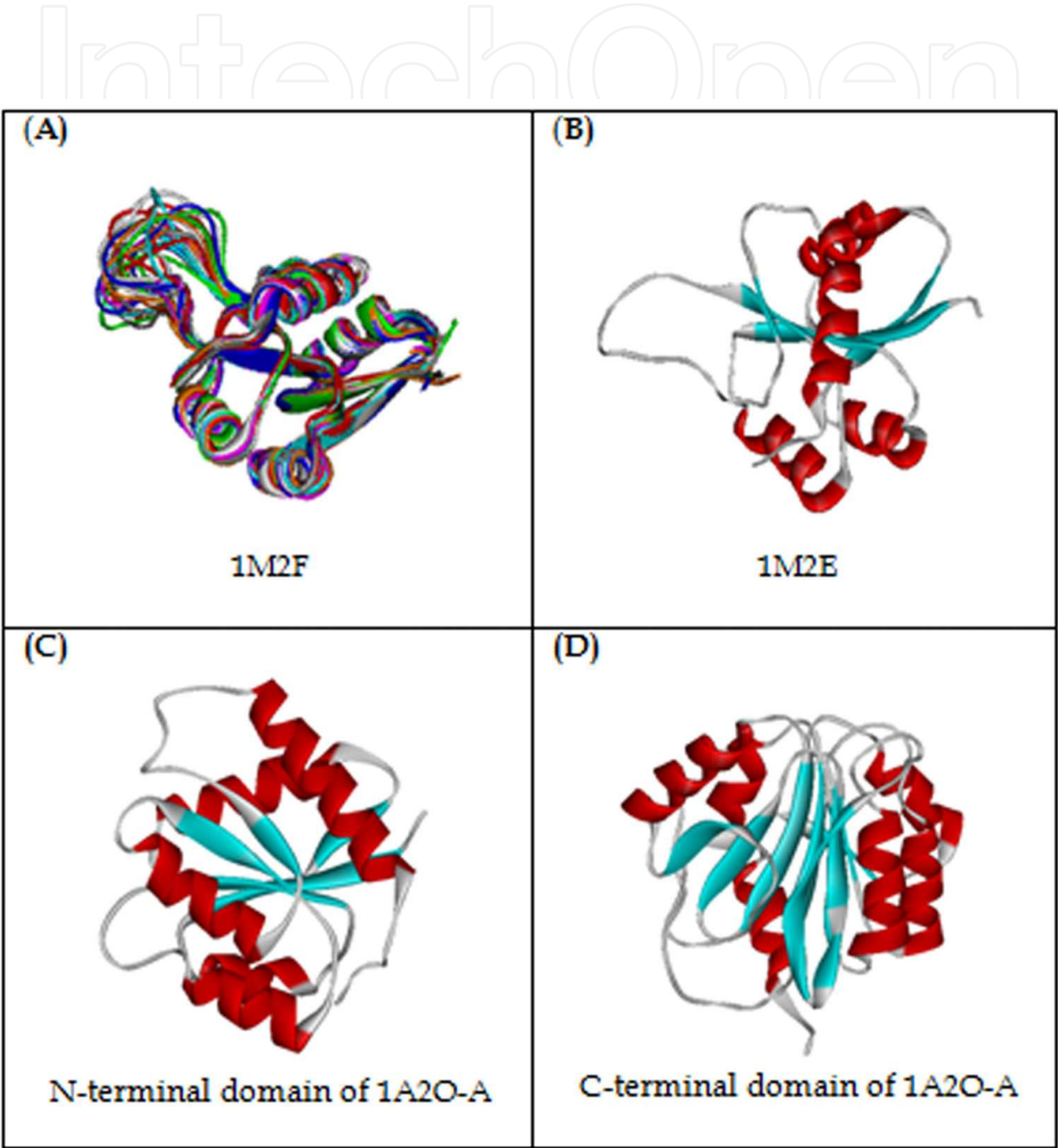


Fig. 2. Images of protein for structural comparisons. (A): 25 conformers of protein 1M2F. (B): protein structure of 1M2E. (C) and (D): N-terminal domain and C- terminal domain in chain A of 1A2O.

	0	1	2	3	4	5	6	7	8	9	100															
Rule	12345678901234567890123456789012345678901234567890123456789012345678901234567890123456	12345678901234567890123456789012345678901234567890123456789012345678901234567890123456	12345678901234567890123456789012345678901234567890123456789012345678901234567890123456	12345678901234567890123456789012345678901234567890123456789012345678901234567890123456	12345678901234567890123456789012345678901234567890123456789012345678901234567890123456	12345678901234567890123456789012345678901234567890123456789012345678901234567890123456	12345678901234567890123456789012345678901234567890123456789012345678901234567890123456	12345678901234567890123456789012345678901234567890123456789012345678901234567890123456	12345678901234567890123456789012345678901234567890123456789012345678901234567890123456	12345678901234567890123456789012345678901234567890123456789012345678901234567890123456	01234567890123456															
Seq.	MLSQIAICIWVESTAILQDCQRALSADRYQLQVCESGEMLLEYAQTHRQIDCLILVAANPSFRAVVQQLCFEGVVVPAIVVGDRDSEDPDEPAKEQLYHSAELHL																									
1M2E-A	--WY	BBBBBBBVJ	AAAAADAAAAB	VAAJ	BBBBBV	AAAAAAAAAAD	JVAAJ	VJBBBBV	HJ	AAAAAAAAAAD	JBBWS	SBBBVJ	BBBBBV	JVAJ	BBWY	AP	CYAA	BBBB								
1M2F-1	--WY	IBBBBBBBVJ	AAAAADAAAAB	VAAJ	BBBBBV	AAAAAAAAAAD	JVAAJ	VJBBBBV	HJ	AAAAAAAAAAD	JBBWS	SBBBVJ	BBBV	JBBVAJ	VJ	BV	HP	CYAA	BBBB							
1M2F-7	--WY	JBBBBBBBVJ	AAAAADAAAAB	VAA	BBBBBBBV	AAAAAAAADAAD	JVAAJ	VJBBBBV	HJ	AAAAAAAAAAD	JBBWS	SBBBVJ	BBBBBBV	VAJ	JB	VP	YAP	CYAA	JBBB							
1M2F-2	--WY	BBBBBBBVJ	AAAAADAAAAB	VAAJ	BBBBBV	AAAAAAAADAAD	JVAAJ	VJBBBBV	HJ	AAAAAAAAAAD	JBBWS	SBBBVJ	BBB	PSBVAJ	JB	VP	YAP	QC	YAA	BBBB						
1M2F-18	--PY	BBBBBBBVJ	AAAAADAAAAB	VAAJ	BBBBBV	AAAAAAAAAAD	JVAAJ	VJBBBBV	HJ	AAAAAAAAAAD	JBBWS	SBBBVJ	BBBBBH	BVAJ	JB	VAA	P	G	YAA	JBBB						
1M2F-25	--WY	IBBBBBBBVJ	AAAAADAAAB	VAAJ	BBBBBV	AAAAAAAAAAD	JVAAJ	WS	SBBBV	HJ	AAAAAAAAAAD	JBBWS	SBBBVJ	VJ	VJBBVAJ	JB	VP	YAP	CYAA	HBBB						
1M2F-8	--WY	IBBBBBBBVJ	ADAAAAADAB	VAAJ	BBBBBV	AAAAAAAAAAD	JVAAJ	WS	SBBBV	HJ	AAAAAAAAAAD	JBBWS	SBBBVJ	BBH	JBBVAJ	JB	VP	YAP	QC	YAA	JBBB					
1M2F-20	--PY	BBBBBBBVJ	AAAAADAAAB	VAAJ	BBBBBV	AAAAAAAAAAD	JVAAJ	WS	SBBBV	HJ	AAAAAAAAAAD	JBBWS	SBBBV	VAJ	BBHBBVAJ	VJ	W	HP	CYAA	BBBB						
1M2F-13	--PS	BBBBBBBVJ	AAAAADAAAAB	VAAJ	BBBBBV	AAAAAAAADAAD	JVAAJ	VJBBBBV	HJ	AAAAAAAAAAD	JBBWS	SBBBVJ	BBH	JBBAAJ	BB	V	HP	CYAA	HBBB							
1M2F-15	--PY	BBBBBBBVJ	AAAAAAAADA	JVA	ABBBBBBV	AAAAAAAADAAD	JVAAJ	VJBBBBV	VH	AAAAAAAAAAD	JBBWS	SBBBVJ	BBBBBBV	VAJ	JB	VAA	P	CYAA	JBBV							
1M2F-14	--WY	BBBBBBBVJ	AAAAADAAAAB	VAAJ	BBBBBV	AAAAAAAAAD	JVAAJ	WS	SBBBV	HJ	AAAAAAAAAAD	JBBWS	SBBBV	AB	HJB	V	QY	BBV	PYAP	QC	YAA	BBBV				
1M2F-10	--AJ	JBBBBBBBVJ	AAAAADAAAAB	VAA	BBBBBBBV	AAAAAAAADAAD	JVAAJ	VJBBBBV	HJ	AAAAADAAAD	JBBWS	SBBBV	VAJ	WS	BV	JVAAJ	BB	VHP	G	YAA	BBBB					
1M2F-12	--PY	BBBBBBBVJ	AAAAADAAAB	VAAJ	BBBBBV	AAAAADAAAD	JVAAJ	WS	SBBBV	HJ	AAAAADAAAD	JBBWS	SBBBVJ	BH	AJ	JB	VAAJ	JB	VAA	P	CYAA	BBBB				
1M2F-3	--WY	IBBBBBBBVJ	AAAAADAAAAB	VAA	BBBBBBBV	AAAAAAAADAAD	JVAAJ	WS	SBBBV	HJ	AAAAADAAAD	JBBWS	SBBBVJ	BBB	WY	JV	HJB	V	VAA	P	CYAA	HBBB				
1M2F-16	--PS	BBBBBBBVJ	AAAAAAAADAB	VAAJ	BBBBBV	AAAAAAAADAAD	JVAAJ	WS	SBBBV	VH	AAAAADAAAD	JBBWS	SBBBVJ	BBV	VAJ	JB	VAD	J	B	V	HP	CYAA	BBBB			
1M2F-24	--PS	BBBBBBBVJ	AAAAAAAADA	JVA	ABBBBBBV	AAAAAAAADAAD	JVAAJ	VJBBBBV	VA	HJ	AAAAADAAAD	JBBWS	SBBBVJ	BBBBBBV	DJ	JB	VP	YHP	G	YAA	BBBB					
1M2F-9	--WY	BBBBBBBVJ	AAAAADAAAB	VAAJ	BBBBBV	AAAAADAAAD	JVAAJ	WS	SBBBV	HJ	AAAAADAAAD	JBBWS	SBBBVJ	BV	HJB	JB	VAJ	V	HP	YHP	G	YAA	BBBV			
1M2F-21	--PY	BBBBBBBVJ	AAAAAAAADAB	VAAJ	BBBBBV	AAAAAAAADAAD	JVAAJ	WS	SBBBV	HJ	AAAAADAAAD	JBBWS	SBBBVJ	BBBBBBV	A	H	H	P	YAP	QC	G	YAA	BBBB			
1M2F-11	--PY	BBBBBBBVJ	AAAAADAAAAB	VAA	BBBBBBBV	AAAAAAAADAAD	JVAAJ	WS	SBBBV	VH	AAAAADAAAD	JBBWS	SBBBWY	JW	SB	VJ	VAJ	BB	B	VHP	CYAA	JBBB				
1M2F-17	--WY	IBBBBBBBVJ	AAAAADAAAAB	VAAJ	BBBBBV	AAAAAAAADAAD	JVAAJ	VJ	SBBBV	VHJ	AAAAADAAAD	JBBWS	SBBBV	AB	W	SB	VJ	VAA	H	HP	YAP	G	YAA	HBBB		
1M2F-23	--PS	BBBBBBBVJ	AAAAADAAAAB	VAAJ	BBBBBV	AAAAAAAADAAD	JVAAJ	VJ	SBBBV	VAH	J	AAAAADAAAD	JBBWS	SBBBV	AB	HPY	B	VAA	B	VAA	P	CYAA	BBBU			
1M2F-19	--WS	BBBBBBBVJ	AAAAADAAAAB	VAAJ	BBBBBV	AAAAAAAADAAD	JVAAJ	VJ	SBBBV	VH	AAAAADAAAD	JBBWS	SBBBVJ	BBH	BW	YP	Z	J	B	P	YHP	G	YAA	HBBB		
1M2F-5	--WY	JBBBBBBBVJ	AAAAADAAAB	VAAJ	BBBBBV	AAAAAAAADAAD	JVAAJ	VJ	SBBBV	VHJ	AAAAADAAAD	JBBWS	SBBBV	VA	HAJ	BB	VAA	H	HP	YHP	G	YAA	BBBB			
1M2F-22	--WY	IBBBBBBBVJ	AAAAADAAAB	VAA	BBBBBBBV	AAAAAAAADAAD	JVAAJ	VJ	SBBBV	VH	AAAAADAAAD	JBBWS	SBBBV	A	H	H	J	BB	VAA	H	B	B	H	HP	CYAA	BBBB
1M2F-6	--WY	YBBBBBBBVJ	AAAAADAAAB	VAAJ	BBBBBV	AAAAAAAADAAD	JVAAJ	WS	SBBBV	VH	AAAAADAAAD	JBBWS	SBBBVJ	BBB	V	H	B	V	JWY	J	V	HAP	G	YAA	HBBV	
1M2F-4	--WC	SBBBBBBBVJ	AAAAAAAADA	JWZ	ABBBBBBV	AAAAAAAADAAD	JVAAJ	WS	SBBBV	VH	AAAAADAAAD	JBBWS	SBBBVJ	BBBBH	JV	J	VAJ	B	AA	P	CYAA	BBBB				

Table 2. PFSA Alignment Table: Conformation alignment of 1M2E-A and 25 conformers of 1M2F. 1M2F and 1M2E are listed on left column. The amino acid sequence and rule for number of residue protein folding shape code (PFSC) for each conformer is listed following the structure name. The with red color, the  $\beta$ -strands with blue color and the tertian fragments with black. Also, the analog structure are remarked with pink color.The PFSA alignment table has the capability for analysis of alignment table does not only align the secondary structure (font with red and blue colors), but it tertiary structure (font with black color). Second, the alignment table exhibits the detail element alignment within each of fragment of secondary structure. The font with pink color indicates the alteration in or the flexible terminal of secondary structure. Third, the alignment table is able intuitively to reveal structural stability or flexibility. For example, in regions of fragment of residues (50-54) and fragment conformations show the fluctuation in 25 conformers, which indicates these two regions with more protein segments.

The PFSA approach has capability for evaluation of global similarity. It provides PFSA-S as score to assess the global structural similarity. The 1M2E in Fig. 2(B), as average structure, is compared with each of 25 conformers of 1M2F in Fig. 2(A). The similarity scores are listed with descending order of PFSA-S in Table 3, including the number of identical and number of analogous folding shapes. Also, the results are compared with LGA method (Zemla, 2003). Both of PFSA-S and PFSA alignment table explicitly display the structural difference in protein conformation analysis. Apparently, the PFSA approach has ability to differentiate each conformer with its appropriate sensitivity.

Name	PFSA				LGA		
	PFSA-S	ID_FS	AN_FS	Gaps	N	RMSD	GDT_TS
1M2F-1	0.964	118	11	0	133	0.93	96.111
1M2F-7	0.962	117	12	0	131	0.85	95.741
1M2F-2	0.956	120	7	0	135	0.91	96.296
1M2F-18	0.956	120	7	0	134	0.89	95.370
1M2F-25	0.954	116	12	0	134	0.94	95.000
1M2F-8	0.952	115	13	0	135	0.79	97.037
1M2F-20	0.952	112	17	0	133	0.94	95.926
1M2F-13	0.950	117	10	0	131	0.95	95.370
1M2F-15	0.948	113	15	0	130	0.80	95.185
1M2F-14	0.943	113	14	0	133	0.91	95.962
1M2F-10	0.939	114	12	0	135	1.09	95.556
1M2F-12	0.939	114	12	0	133	0.92	95.370
1M2F-3	0.937	116	9	0	129	0.74	94.074
1M2F-16	0.937	113	13	0	133	0.70	96.296
1M2F-24	0.937	110	17	0	133	0.89	95.185
1M2F-9	0.931	113	12	0	132	1.14	95.000
1M2F-21	0.931	113	12	0	130	0.80	95.556
1M2F-11	0.929	109	17	0	134	0.84	95.926
1M2F-17	0.927	114	10	0	133	0.80	96.296
1M2F-23	0.926	113	11	0	132	1.00	93.704
1M2F-19	0.920	116	6	0	134	0.95	96.111
1M2F-5	0.918	112	11	0	134	1.04	95.556
1M2F-22	0.914	110	13	0	131	0.85	95.000
1M2F-6	0.906	106	17	0	130	1.05	92.963
1M2F-4	0.893	106	20	0	130	1.01	94.444

Table 3. Comparison of 25 conformers of 1M2F to average structural model of 1M2E with PFSA approach and LGA method. All data are sorted by values of PFSA-S. Left column lists the names of 25 conformers of 1M2F. PFSA: approach of protein folding structure alignment; PFSA-S: PFSA score for structural similarity; ID\_FS: number of identical folding shapes; AN\_FS: number of analogous folding shapes and Gaps: number of insertion gaps. LGA: LGA method (Zemla, 2003); GDT\_TS: an estimation of the percent of residues (largest set) that can fit under the distance cutoff of 1, 2, 4 and 8 Å. N: number of superimposed residues under a cutoff distance and RMSD: root mean square deviation of all corresponding C-alpha atoms.

3.2 Domain-domain comparison

The proteins belong to different categories in the structural classification of protein (SCOP) (Murzin et al., 1995) are compared. The structures 1M2E in Fig. 2(B) is compared with N-terminal domain of chain A of 1A2O (1A2O-A) in Fig. 2(C) and then its C-terminal domain in Fig. 2(D) respectively. Although, all of three structures are classified as the class of alpha and beta proteins ( $\alpha/\beta$ ), they belong to two of different folds in SCOP. Both structures of 1M2E and N-terminal domain of 1A2O-A belong to Flavodoxin-like fold, but the C-terminal domain of 1A2O-A belongs to Methylesterase CheB fold. The summary of structural classification of 1M2E, N-terminal domain and C-terminal domain of 1A2O-A is listed in Table 4.

First, the alignment table provides the detail information of alignment for local structural fragments. Table 5 shows the comparison of 1M2E and N-terminal domain 1A2O-A while Table 6 shows the comparison of 1M2E and C-terminal domain of 1A2O-A. The alignment tables in Table 5 and Table 6 display how the fragments with similar local folding shapes are matched up with insertion of gaps. In alignment table, the aligned identical protein folding shape code is marketed with “|”, the analogue with “\*”, the different with “^” and the insertion with “+”. Actually, the alignment table shows the optimized structural alignment with matching all local structural fragments between two proteins. Second, the PFSA-S provides the quantitative assessment of similarity for global structural comparisons. The PFSC-S values are listed in Table 4, including the numbers of identity and analog of folding shapes, and the number of insertion gaps. In contrast to C-terminal domain, the comparison of N-terminal domain of 1A2O-A and 1M2E have higher PFSA-S similarity score (0.7214 vs. 0.2109), larger number of identical and analogous folding shapes and less number of gaps. The results reflect the homologous difference of these two pairs of proteins in structure classification.

Name	1M2E	1A2O-A	
		N- terminal domain	C- terminal domain
Class	Alpha and beta proteins ( $\alpha/\beta$ )	Alpha and beta proteins ( $\alpha/\beta$ )	Alpha and beta proteins ( $\alpha/\beta$ )
Fold	Flavodoxin-like	Flavodoxin-like	Methylesterase CheB
Superfamily	CheY-like	CheY-like	Methylesterase CheB
Family	Circadian clock protein KaiA	CheY-related	Methylesterase CheB
Protein	Circadian clock protein KaiA	Methylesterase CheB,	Methylesterase CheB
PFSA-S	1.000	0.7214	0.2109
Number of Identity	131	69	59
Number of Analog	0	38	43
Number of Gaps	0	3	25

Table 4. Comparison of structure of 1M2E with N-terminal domain and C-terminal domain of Chain A of 1A2O. Top rows show the structural classification of 1M2E, N-terminal domain and N-terminal domain of 1A2O-A. The bottom four rows list the values of structural similarity of comparison of 1M2E with N-terminal domain and N-terminal domain of 1A2O-A respectively with using PFSA approach

Seq	1:	-MLSQIAICIWVESTAILQDCQRALSADRYQLQVCESGEMLLEYAQTHRDQIDCLILVAANPSFRAVVQQLCFE
1M2E-A		---WYBBBBBBBVJVAAAAADAAAABVAAJB BBBBVA AAAAAAAAAADJVAJVVJB BBVBHVJVA AAAAAAAAAAD + ^ ^         * *   * *         *       *   * *   ^ ^                         * ^ *   * *   * ^ * ^ ^ ^ *   * *           *
1A2O-A_N_DoM		VJB BBBBVBVJV AAAAAAAAAAAAAA JVAJB BWYBBV AAAAAAAAAAADJBBVJBBBVBHVQSWYADDA AAAAAA
Seq	1:	MSKIRVLSVDDSA LMRQIMTEI INSHSDMEMVATAPDPLVARDLIK FKNPDVLTLDVEMPRMDGLDFLEKLMRI

---

Seq	100:	HSAELHLGIHQLEQ-LPYQVDAALAEFLRLAPVETMA---
1M2E-A		Y A A B B B B H A A J V A A - A D D A A A A A A A A A A J V J B B B ----- ^ ^ *       ^ ^ * *       + * * *                   * * * ^   * + + + + +
1A2O-A_N_DoM		BWYBBB B F Z P S V A A I V A A A A A A A A A A A A A A H B V A A J --
Seq	100:	AIDFVTKPQLGIREGMLAYSEMI AEKVRTAARARIAAHKPF

Table 5. PFSA Alignment Table: Protein Folding Structure Alignment of 1M2E and N-Terminus Domain of 1A2O are listed on left column. Amino acid sequence is listed on right column. The protein folding shape code (PFSC) for each structure is listed following the structural elements. The  $\alpha$ -helices are remarked with red color, the  $\beta$ -strands with blue color and the tertiary fragments with black. Amino acids involved in secondary structure are remarked with pink color. The “|” indicates the alignment with identical folding shapes; “^” different folding shapes. The “+” represents the insertion of gaps.

Seq	1:	-----MLSQIAICIWV-ESTAILQDCQRAL--SADRYQLQVCESGEMLLLEYAQTHRDQI--DCLIL-
1M2E-A		-----WYBBBBBBB-VJVAAAAADAAAA--BVAAJBBBBBVAAAAAAAAAADJVAAJ--VJBBB-
		+^#####^*   +^**   **  +^***   **   **  **+*+* +*+
1A2O-A_C_DoM		HJBAJVVAPCRWYAAJBBBBBVDAAAAAAAAADAAHHAHPBBBBBVVPYAAAAAAAAAAAAAJVJJB-VH
Seq	141:	MAAPTTLKAGPLLSSEKLIAGASTGGTEAIRHVLQPLPLSSPAVIITQHMPPGFTRSFARLNKLCQISV-KR
Seq	74:	GVVV-P-AI-VVGDRDSEDPDEP-A---KEQ-----LY-HSAELHL---GIHQLEQLPYQVDAALAEFLRLAE
1M2E-A		JBBW-S-BB-BVJBBBBBJVJJB-B---WYA-----PC-YAABBBB---HAAJVAAADDAAAAAAAAAAJVC
		* + +*+ *+ ** ++++^#####+ ++++* +++++^+*  *  + + + **   **   +++++^**
1A2O-A_C_DoM		BB-WGYHBVIBBBB-----WC SVQYBVJBVAAAAAAAAADDJVAAJBBBBWYHAPSVAAAAAAAAA-----QSB
Seq	237:	SG-ANYQIKIHDGP-----PVNRHRPSVDVLFHFSVAKHAGRNAVGVILTGMGNDGAAGMLAMYQAGAWTIL
		AR-SGANYQIKIHD----GPPVNRHRPSVDVLFHFSVAKHAGRNAVGVILTGMGNDGAAGMLAMYQ-----AGAW
Seq		-----
1M2E-A		-----
		+++++
1A2O-A_C_DoM		BVHAAAAAAAAAAAAAPYA--
Seq	327:	VVDLSQVSQQMLAKISAGQAI

Table 6. PFSA Alignment Table: Protein Folding Structure Alignment of 1M2E and C-Terminus Do of 1M2E and N-terminus domain of 1A2O are listed on left column. Amino acid sequence is listed on right column. The protein folding shape code (PFSC) for each structure is listed following the structural remarks. The  $\alpha$ -helices are remarked with red color, the  $\beta$ -strands with blue color and the tertiary fragments with black. Also, the loops and turns with secondary structure are remarked with pink color. The “|” indicates the alignment with identical residues; “^” analogous folding shapes; “^” different folding shapes. The “+” represents the insertion of gaps.

### 3.3 Protein comparison

Proteins may be comprised by single domain or multiple domains in the chain structure. To take protein chain-chain in alignment will related to multiple domain comparison. For example, insulin-like growth factor 1 receptor (IGF1R) and insulin receptor (INSR), transmembrane proteins belonging to the tyrosine kinase super-family, have multiple domains in structure. Over the past two decades, rich structural data of IGF1R/INSR has been accumulated, and the sequence alignment was applied in comparison (Werner et al., 2008; McKern, 2006; Garrett, 1998; Pautsch, 1997; Hubbard, 1997; Lou, 2006; Garza-Garcia, 2007). In this study, instead, the folding conformations of IGF1R and INSR are directly aligned for structural comparison. The crystal structures of first three domains of L1-CR-L2 structures of IGF1R (PDB ID: 1IGR) (Hubbard, 1997) and INSR (PDB ID: 2HR7) (Murzin, 1995) are available in PDB. The images of first three domains for IGF1R (1IGR) and INSR (chain A of 2HR7) are displayed in Fig.3. Both L1 and L2 domains consist of a right-handed  $\alpha$ -helix conformation. The CR domain is composed of seven modules with eight disulphide-bond connectivity.

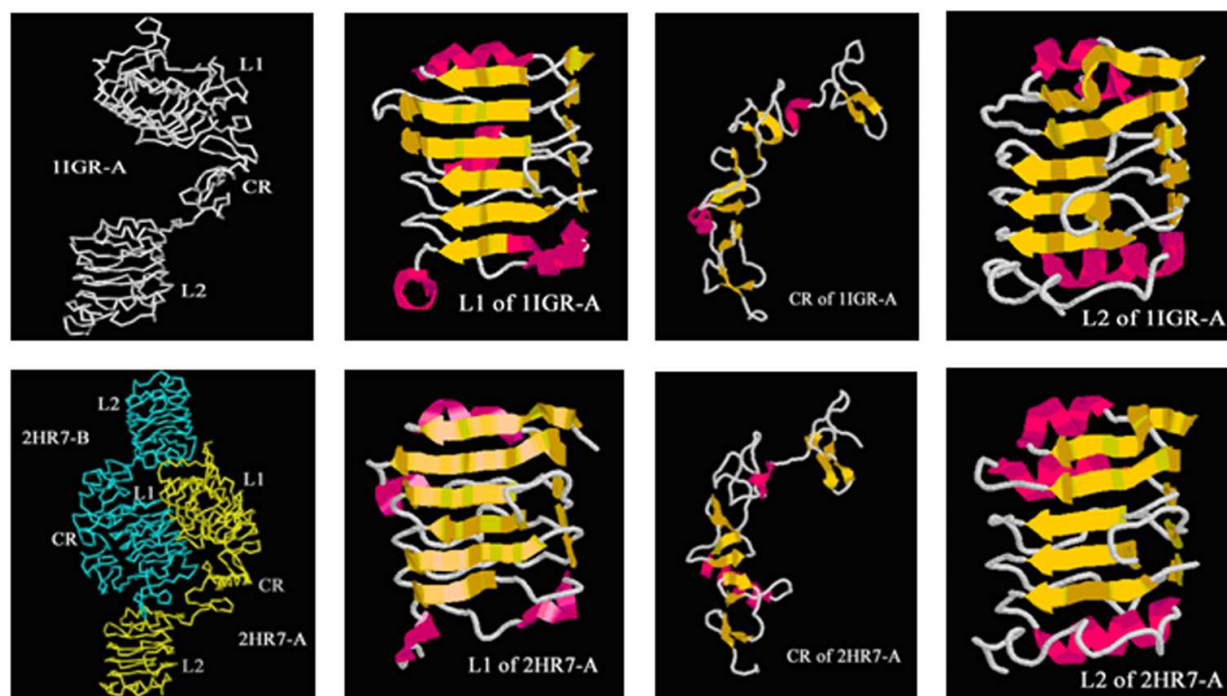


Fig. 3. Structural images of IGF1R (1IGR) and INSR (2HR7). The IGF1R and its L1, CR and L2 domains are shown in the top row; INSR and its L1, CR and L2 domains are shown in the bottom row. The atomic geometric coordinates of 1IGR and 2HR7 are obtained from the protein databank (PDB). The image was drawn with RasWin Molecular Graph

ics V. 2.6. The structural images are displayed by  $\alpha$ -carbon backbone or cartoon views.



Table 7. PFSA Alignment Table: Insulin-like growth factor 1 receptor (IIGR-A) and insulin receptor represents the chain A of IGF1R; 2HR7-A for chain A of INSR. Seq: sequence of amino acid. PFSC labeled  $\alpha$ -helix; blue  $\beta$ -strand; pink analogous  $\alpha$ -helix and  $\beta$ -strand. Numeric rules are displayed at top. The rules indicate the hundredth, tenth and digit numbers. PhyChe: The physicochemical properties of physicochemical properties are labeled with nonpolar side chain for hydrophobicity: P, acidic side chain for hydrophilicity: A, basic side chain for hydrophilicity: B, hydroxyl group for hydrophilicity: O, glycine with unchanged hydrogen as side chain: H. The Polar group of amino acids is divided into three properties. Amino acid with hydroxyl group -OH: O, with carbonyl and amide group O=CNH2: N, with amine group -NH2: M. Geom: geometric size and characteristic of amino acid. Tiny: T, small: S, large: L and hardness: H.

The structural is assessed. The sequence similarity is evaluated by the percentage of identical residues. The structural similarity is quantitatively assessed by PFSA score. The similarity of three domains of L1, CR and L2 for IGF1R and INSR are summarized in Table 8. Overall, two protein structures have 60% of sequence in identity with structural similarity score at 0.860. Furthermore, each pair of domains is compared. The L1 domain has 67% of sequence in identity with structural similarity score at 0.909, the L2 domain 64% of sequence in identity with structural similarity score at 0.929 and CR domain 49% of sequence in identity with structural similarity score at 0.749. The PFSA scores specified that the L1 and L2 domains have higher structural similarity than the CR domains. Also, L2 domains have a higher degree of structural homology than L1, even though L1 has a higher degree of identity of sequence. With PFSA approach, the quantitative assessment of similarity between IGF1R and INSR agrees with previous quality specifications by sequence alignment. However, detail structural features are exposed for comparison.

Domain	Residues	Sequence Identity	Structural Homology				
			PFSA Score	Identical	Analog	GapO	GapE
L1 + CR + L2	1IGR-A (1-459) 2HR7-A (4-469)	60%	0.860	353	68	10	6
L1	1IGR-A (1-147) 2HR7-A (4-154)	67%	0.909	119	20	4	0
CR	1IGR-A (148-299) 2HR7-A (155-309)	49%	0.749	96	30	5	6
L2	1IGR-A (300-459) 2HR7-A (310-469)	64%	0.929	133	16	1	0

Table 8. Quantitative assessment of similarity for domain structures between IGF1R and INSR. PFSA Score: protein folding structural alignment score – a value for structural similarity; Identical: number of identical folding shapes; Analog: number of analogous folding shapes; GapO: number of opened gaps for structural alignment and GapE: number of extended gaps for structural alignment. Sequence identity is obtained with running J Aligner at <http://jaligner.sourceforge.net/>

4. Discussion

4.1 Feature of PFSA approach

4.1.1 Consistent procedure

The PFSA approach provides an unambiguous procedure for protein comparison based on structural alphabetic alignment. First, the PFSA approach relies on complete assignment of protein conformation. The PFSC provides a complete assignment of protein conformation for any protein with given 3D structure. Without usage of training database, all 27 PFSC are obtained by restrict mathematical derivation. Each PFSC vector or alphabetic letter represents a special folding shape of five successive C $\alpha$  atoms in protein backbone. The folding shape of each of five successive C $\alpha$  atoms in protein backbone is assuredly assigned

by one among 27 PFSC vectors. Therefore, the protein backbone from N-terminal to C-terminal gets complete alphabetic assignment for folding conformation without gap. Second, the PFSA alignment of alphabetic strings is a consistent process. The PFSA approach is able to avoid the artificial choice of geometric parameters in structural comparison, such as the adjustment of initiative focusing location, cutoff distance for RMSD and the length of segment. Similarly as sequence alignment, the structural alphabetic alignment provides a fast and steady procedure for protein structure comparison. Third, the PFSA approach is able to handle protein comparison in various homologies, i.e. in wider scope of structure difference. This feature is well demonstrated by results of comparison of conformers in Table 2, comparison of different proteins in Table 5 and Table 6, and comparison of protein with complicated structures in Table 7. Furthermore, the PFSA approach is able to categorize the protein structures according structural classification in homology. With structure classification of protein SCOP (Murzin, 1995; Andreeva, 2008) as gold standard, the PFSA assessed the homologous degree for a set of protein structures, and the distribution of similarity scores, PFSA-S, was overall agreed with the categories in SCOP (Yang, 2011).

#### 4.1.2 Normalized score and unique measurement

With normalization of PFSA-S score, the structural similarity of various proteins is easily assessed. If two structural data are an identical protein structure, the PFSA-S equals one. If the structural similarity decreases, the value of PFSA-S decreases. When the value of PFSA-S is near zero or less than zero, two proteins have large difference in conformation shape. The PFSA-S score is normalized by size of protein. In PFSA approach, the length of protein folding shape string is used as the denominator in formula for normalization when the PFSA-S is calculated. If a pair of two proteins is compared, anyone of proteins may be taken as the referent protein. If a set of proteins are compared with a reference protein, the similarity scores are normalized according the length of referent protein. The PFSA approach provides a unique quantitative measurement to evaluate the similarity in protein structural comparison.

#### 4.1.3 Local structural comparison

The PFSA alignment table is able to compare protein structures in detail. The one-dimensional alphabetic string expresses the change of protein folding conformation along backbone. A letter of PFSC represents the folding shape of fragment for five successive amino acids. In alignment table, the protein folding conformations are aligned with similarity. The PFSA alignment table is comprised with the amino acid residues by adhesive to the associated folding shape code. Furthermore, the PFSA alignment table includes physicochemical properties of amino acid residue which are expresses by alphabetic letters as seen in Table 7. Therefore, the PFSA alignment table may become a good tool to study the relationship between sequence-structure-function. The PFSA alignment table has capability to exam the structural similarity as well as dissimilarity. In alignment table, if the local structures match with identical or analogous folding shapes, it reveals the structural similarity; if local structures align with different folding shapes, it exhibits the dissimilarity. Also, some of unmatched local structures are shifted with insertion of gaps to display the dissimilarity. In general, it is hard to straightforwardly expose both of similarity and dissimilarity with protein 3D structural image or computer modeling animation. Protein

modeling provides visualization for view of 3D structure, but PFSA alignment table provides digit description for conformation. The combination of application of protein 3D modeling with PFSA alignment table is helpfully to inspect both of similarity and dissimilarity in protein structures.

## 4.2 Comparison with other methods

Different methods adopt various strategies to study specific geometric parameters for protein structural comparison. With different parameters and approaches, all methods have a common goal trying to evaluate the similarity of protein structures. As complexity, it not surprise there is no unique outcome for protein comparison. In this study, the results from PFSA approach are compared with other methods.

### 4.2.1 PFSA vs. LGA

LGA method (Zemla, 2003) is an important approach for protein structure comparison. Specially, it is extensively applied for assessment of similarity for protein prediction in Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Kryshtafovych et al., 2007; Moult et al., 2009). The 25 conformer of 1M2F and its average model of 1M2E are compared by both of PFSA approach and LGA method respectively. The results are listed in Table 3, where all structures are ranked by the order of PFSA-S. LGA method and PFSA approach adopt different strategies to assess the structural similarity. LGA method is designed to evaluate the longest continuous segments (LCS) searching for the largest set of 'equivalent' residues that deviate by no more than a specified distance cutoff. GDT\_TS is an estimation of the percent of largest set of residues that can fit under selected cutoff distances. A scoring function (LGA\_S) was defined as a combination of these values and can be used to evaluate the level of structure similarity of selected regions. However, PFSA takes the fixed length of segment of five successive  $C_\alpha$  atoms to determine the folding shape, and then directly makes the alignment with structural alphabets. It is not surprised that PFSA and LGA methods present different ranks in structural comparisons. Due to higher similarity, the comparison of 25 conformer of 1M2F requires a tool with sensitivity to distinguish structural perturbation. The PFSA approach provides finer description for folding conformation. Each PFSC code steadily represents the folding shape of five successive residues and each of PFSC vector can be transformed from one to another. 27 PFSC vectors cover all possible folding shapes. Therefore, each conformer of 1M2F acquires a complete assignment along protein backbone, so the alignment is performed with full length of structure from N-terminal to C-terminal. Furthermore, with structural alphabets, the PFSA adopts an unambiguous process in alignment for protein comparison. Except similarity score PFSA-S, with folding shape for each five residues, the PFSA approach provides explicit comparison in alignment table. Therefore, the PFSA approach offers a complementary tool in analysis of protein conformation.

### 4.2.2 PFSA vs. CE

The combinatorial extension (CE) method (Shindyalov & Bourne, 1998) breaks each structure in the query set into a series of fragments that it then attempts to reassemble into a complete alignment. A series of pairwise combinations of fragments are used to define a similarity matrix through which an optimal path is generated to identify the final alignment.

The size of each aligned fragment pairs is usually set to empirically determined values of 8 and 30 respectively. One group of 20 structures, the quaternary complex of cAMP dependent protein kinase, has certain structural similarities and is compared with the structure of 1ATP-E by CE method and PFSA approach respectively. The results of comparisons between 1ATP-E and 20 of cAMP dependent protein kinases are listed in Table 9 which is sorted in the order by Z Score of CE. Two conclusions are observed from results. First, the ranks of similarity are overall agreed between CE and PFSA, except the structures with number 5, 8, 9, 10 and 18. With PFSA approach, the assessment of similarity is an aggregate of matched folding shape, structural topological distribution, gap and size of protein. Thus, the rank for structural similarity may be adjusted by relative size of compared proteins. 1ATP-E, as reference protein, has sequence length of 335 and other structures, as target proteins, with the matched size may have higher PFSA-S. Referring the length of 1ATP-E, for example, the similarity of structures 5 and 8 is assigned with lower values of PFSA-S because of having large difference in length of 298 and 438; the similarity of structures 18 with higher value of PFSA-S because of having matched length of 366.

No.	PDB ID	CE					PFSA			
		Size	N <sup>A</sup>	N <sup>G</sup>	RMSD	Z Score	ID_FS	AN_FS	Gaps	PFSA-S
1	1APM-E	350	336	0	0.3	7.9	282	48	5	0.9145
2	1CDK-A	350	336	0	0.4	7.9	271	57	3	0.8968
3	1YDR-E	350	336	0	0.5	7.9	274	43	0	0.9092
4	1CTP-E	350	303	0	1.5	7.4	249	52	19	0.8074
5	1PHK	298	255	28	2.5	7.2	161	83	59	0.5034
6	1KOA	491	258	20	2.7	7.1	205	88	21	0.6122
7	1KOB-A	387	260	20	2.8	7.1	182	99	24	0.6382
8	1AD5-A	438	237	31	2.5	7.0	117	133	102	0.3825
9	1CKI-A	317	260	47	2.8	6.9	128	113	46	0.4725
10	1CSN	298	249	37	2.4	6.8	117	117	43	0.4439
11	1ERK	364	254	55	2.6	6.8	171	106	43	0.5768
12	1FIN-A	298	253	69	2.2	6.8	168	88	52	0.5557
13	1GOL	364	254	55	2.6	6.8	174	99	43	0.5535
14	1JST-A	298	253	69	2.4	6.7	161	96	50	0.5553
15	1IRK	306	244	69	3.3	6.5	136	103	53	0.4552
16	1FGK-A	310	251	54	3.5	6.2	141	81	66	0.4322
17	1FMK	452	245	19	2.8	6.2	122	129	104	0.4040
18	1WFC	366	240	72	3.1	5.6	157	94	54	0.5294
19	1KNY-A	253	112	79	4.3	3.9	114	89	83	0.3592
20	1TIG	94	54	3	4.2	3.9	71	17	228	0.0565

Table 9. Comparison of 20 of quaternary complex of cAMP with strucrure of 1ATP-E (PDB ID: 1ATP, chain E, sequence length 335). N<sup>A</sup>: number of aligned position; N<sup>G</sup>: number of non-aligned position; RMSD: results based on C alpha atoms over the length of the alignment; Z Score: measure of the statistical significance of the result relative to an alignment of random structures. ID\_FS: number of identical folding shapes; AN\_FS: number of analogous folding shapes and Gaps: number of insertion gaps; PFSA-S: PFSA score for structural similarity.

Second, the CE method indicates that 20 protein structures have similar fold as structure of 1ATP-E. However, the PFSA has capability further to distinguish the dissimilarity between 20 structures of cAMP dependent protein kinases. According to CE method, if Z Score is larger than 3.5, the compared proteins have similar fold in structure. The values of Z Score of 20 structures of cAMP are from 3.9 to 7.9, so they all have similar fold structure as 1ATP-E. The values of PFSA-S for 20 structures are distributed in the wide range of 0.9145 - 0.0565. According to PFSA approach, the value of PFSA-S is near one when two structures have high similarity, and on the contrary, the value of PFSA-S is near zero when two structures with less similarity. The PFSA-S value 0.0565 is for comparison between 1ATP-E and No. 20 of structure. PFSA-S near zero indicates that the pair of structures is dissimilar. It is noted that the sequence length of structures 20 is 94. To compare with 1ATP-E, two structures have big difference in length and the alignment generates 228 gaps which give the lower value of PFSA-S. Therefore, the PFSA has ability to distinguish the structural deference in more detail.

#### 4.2.3 PFSA vs. other methods

A set of 10 pairs of proteins with lower structural similarity was recognized as difficult structures for comparison, and was evaluated by VAST (Madej et al., 1995; Gibrat et al., 1996), DALI (Holm & Sander, 1993; Holm & Park, 2000), CE (Shindyalov & Bourne, 1998), Prosup (Lackner et al., 2000) and LGA (Zemla, 2003) methods respectively. The structural similarity was evaluated by two optimistic parameters, i.e. lower RMSD and larger number of equivalent residues. It is apparently, in Table 10, that various methods gave comparative results for each pair of proteins. The results from various methods provide complementary information for protein structural comparison. Overall, the ProSup and LGA methods provided consistent results with restriction of RMSD less than 3.0. The PFSA, however, offers new observation for assessment of similarity of protein structures. First, the similarity is able to be evaluated by a single value of the PFSA-S. In order to compare with other methods, information of (sum of number of identical and analogous shapes) / (number of gaps) / (PFSA-S) is listed in Table 10. The similarity score of PFSA-S is determined by number of identical shapes, number of analogous shapes and number of gaps. Second, the value of score PFSA-S may judge the similarity crossing isolated comparisons, i.e. the values of score PFSA-S from unrelated comparisons can be used to assess the protein homologous degree. Each pair of proteins in Table 10 is a lonely comparison without common reference structure, but the value of PFSA-S may indicate which pair of structures has higher similarity. In Table 10, the results of each pair of comparison are sorted according the values of PFSA-S descendingly. For example, the pair of comparison of 1CEW-I and 1MOL-A has the PFSA-S = 0.564 and the pair of comparison of 1CID and 2RHE has the PFSA-S = 0.384. A conclusion may be obtained that the pair of 1CEW-I and 1MOL-A has higher structural similarity than the pair of 1CID and 2RHE. Third, relative size of compared proteins makes the contribution to structural similarity in PFSA approach. With various methods, the value of RMSD is often used to make judgment of structural similarity. For example, the pair of comparison of 1CEW-I and 1MOL-A have the RMSD (VAST: 2.0, DALI: 2.3, CE: 2.3, ProSup: 1.9 and LGA: 2.0); the pair of comparison of 1TEN and 3HHR-B have the RMSD (VAST: 1.6, DALI: 1.9, CE: 1.9, ProSup: 1.7 and LGA: 1.9). Both pairs have lower RMSD than other

remaining pairs and have overall agreement with various methods. However, PFSA approach distinguishes these two pairs by PFSA-S. With PFSA-S = 0.456, the pair of 1TEN and 3HHR-B is ranked below other five pairs, including the pair of 1CEW-I and 1MOL-A with PFSA-A = 0.564. The separation is explained by a factor that the pair of 1CEW-I and 1MOL-A has comparable length of sequence (108 : 94), but the pair of 1TEN and 3HHR-B has larger different in length (99 : 195). The contribution of relative difference of size is counted in PFSA approach. Therefore, with normalization of PFSA-S, for separated comparisons, the similarity degree still can be evaluated without common reference protein.

Proteins	N1	Proteins	N2	VAST	DALI	CE	ProSup	LGA	PFSA
1CEW-I	108	1MOL-A	94	75/2.0	81/2.3	81/2.3	76/1.9	79/2.0	69/24/0.564
1FXI-A	96	1UBQ	76	48/2.1	52/2.5	64/3.8	54/2.6	61/2.6	59/28/0.538
1BGE-B	159	2GMF-A	121	71/2.3	94/3.3	107/3.9	87/2.4	91/2.5	92/47/0.526
1CRL	534	1EDE	310	186/3.7	212/3.6	219/3.8	161/2.6	182/2.6	264/237/0.475
3HLA-B	99	2RHE	114	58/2.3	74/3.0	83/3.3	71/2.7	74/2.5	69/25/0.473
1TEN	99	3HHR-B	195	76/1.5	86/1.9	87/1.9	85/1.7	87/1.9	72/71/0.456
1TIE	166	4FGF	124	76/1.6	114/3.1	116/2.9	101/2.4	104/2.3	88/35/0.456
2AZA-A	129	1PAZ	110	70/2.1	82/3.0	84/2.9	82/2.6	80/2.2	81/36/0.453
2SIM	381	1NSB-A	390	299/4.2	289/3.2	275/3.0	248/2.6	269/2.6	249/129/0.428
1CID	177	2RHE	114	78/2.0	96/3.1	97/2.9	84/2.3	93/2.3	75/61/0.384

Table 10. Comparison of 10 pairs of proteins with lower structural similarity using various methods. N1 and N2 are the lengths of proteins. For methods of VAST, DALL, CE, ProSup and LGA, the results are presented as (Number of identical residues) / (RMSD). For PFSA approach, results are presented as (sum of number of identical and analogous shapes) / (number of gaps) / (PFSA-S). The results of each pair of comparison are sorted according values of PFSA-S descendingly.

5. Conclusion

The PFSA approach adopts the vector of folding shape of five residues as element, and the geometric feature of folding shape is embedded by alphabets as representation. With application of alphabets, the alignment of protein structures is straightforward and steady. This study demonstrates two advantages in PFSA approach. First, 27 PFSC vectors are able to cover all possible folding shapes of five successive Ca atoms in protein. This is fundamental important because it offers a complete description of folding conformation for any protein with given 3D structure. Second, with consistent procedure, the PFSA approach generates unique score for similarity and detail information in alignment table, which provides new observation for the protein structure comparison.

6. Acknowledgments

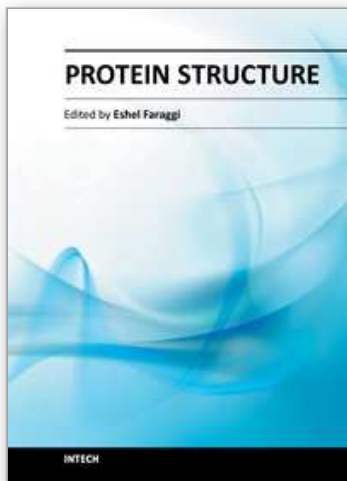
This work was supported in part by a grant from the Indiana Spinal Cord and Brain Injury Research Fund (2009-2011). The algorithms of PFSC and PFSA have been coded with Java (J2SE v.1.5.0\_07) computer language. Requests for additional information will be accepted via e-mail to info@proteinshape.com, jiaan@microtechnano.com, jiaanyang@comcast.net or via Website: http://www.proteinshape.com.

## 7. References

- Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C & Murzin AG. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* 36 (Database issue), 419–425.
- Brevern AG, (2005). New assessment of a structural alphabet. *Silico Biol.*, 5, 283-289
- Brevern AG, Etchebest C & Hazout S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41, 271-287
- Friedberg I, Harder T, Kolodny R, Sitbon E, Li Z, Godzik A. (2007). Using an alignment of fragment strings for comparing protein structures, *Bioinformatics*, 23, 219-224
- Garrett TPJ, McKern NM, Lou M, Frenkel MJ, Bentley JD, Lovrecz GO, Elleman TC, Cosgrove LJ & Ward CW. (1998). Crystal structure of the first three domains of the type-1 insulin-like growth factor receptor. *Nature*, 394, 395-399
- Garza-Garcia A, Patel DS, Gems D & Driscoll PC. (2007). RILM: a web-based resource to aid comparative and functional analysis of the insulin and IGF-1 receptor family. *Hum Mutat.*, 28, (7), 660-668
- Gibrat JF, Madej T & Bryant SH. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, 6, 377–385
- Guyon F, Camproux AC, Hochez J & Tuffery P. (2004). SA-Search: a web tool for protein structure mining based on a structural alphabet. *Nucleic Acids Res.*, 32, W545-W548
- Holm L & Sander C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, 233, 123-138
- Holm, L. & Park, J. (2000). DaliLite workbench for protein structure comparison. *Bioinformatics*, 16, 566-567
- Hubbard SR. (1997). Crystal structure of the activated insulin Cysreceptor tyrosine kinase in complex with peptide substrate and ATP analog. *EMBO J*, 16, 5572–5581
- Hvidsten TR, Kryshtafovych A, Komorowski J & Fidelis K. (2003) A novel approach to fold recognition using sequence-derived properties from sets of structurally similar local fragments of proteins. *Bioinformatics*, 19 (Suppl 2):II81–II91
- Joseph AP, Srinivasan N, Brevern AG. (2011). Improvement of protein structure comparison using a structural alphabet, *Biochimie.*, 93,(9),1434-1445
- Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M & Hughey R. (2003). Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. *Proteins*, 53, Suppl 6:491-6
- Kleywegt GJ & Jones TA. (1994). A super position. *ESF/CCP4 Newsletter*, 31, 9-14
- Kolodny R, Koehl P & Michael Levitt M. (2005). Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures. *J. Mol. Biol.*, 346, 1173-1188
- Kolodny R, Koehl P, Guibas L., & Levitt M. (2002). Small libraries of protein fragments model native protein structures accurately, *J. Mol. Biol.*, 323, 297-307
- Krissinel E & Henrick K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr.*, 60, 2256-2268
- Kryshtafovych A, Milostan M, Szajkowski L, Daniluk P & Fidelis K. (2005). CASP6 data processing and automatic evaluation at the protein structure prediction center. *Proteins*, 61, (Suppl 7):19–23
- Kryshtafovych A, Prlic A, Dmytriv Z, Daniluk P, Milostan M, Eyrich V, Hubbard T & Fidelis K. (2007). New tools and expanded data analysis capabilities at the protein structure prediction center. *Proteins*, 69, S8:19- 26

- Ku SY & Hu YJ, (2008). Protein structure search and local structure characterization, *BMC Bioinformatics*, 9, 349
- Lackner P, Koppensteiner WA, Sippl MJ & Domingues FS. (2000). ProSup: a refined tool for protein structure alignment. *Protein Eng.*, 13, 745-752
- Lou M, Garrett TP, McKern NM, Hoyne PA, Epa VC, Bentley JD, Lovrecz GO, Cosgrove LJ, Frenkel MJ & Ward CW. (2006). The first three domains of the insulin receptor differ structurally from the insulin-like growth factor 1 receptor in the regions governing ligand specificity. *Proc. Natl. Acad. Sci. USA.*, 103, 12429-12434
- Madej T, Gibrat JF & Bryant SH. (1995). Threading a database of protein cores. *Proteins*, 23, 356-369
- Mavridis L & Ritchie DW, (2010). 3D-blast: 3d protein structure alignment, comparison, and classification using spherical polar fourier correlations, *Pac Symp Biocomput*, 15, 281-292
- McKern NM, Lawrence MC, Streltsov VA, Lou MZ, Adams TE, Lovrecz GO, Elleman TC, Richards KM, Bentley JD, Pilling PA, Hoyne PA, Cartledge KA, Pham TM, Lewis JL, Sankovich SE, Stoichevska V, Da Silva E, Robinson CP, Frenkel MJ, Sparrow LG, Fernley RT, Epa VC & Ward CW. (2006). Structure of the insulin receptor ectodomain reveals a folded-over conformation. *Nature*, 443, 218-221
- Melo F & Marti-Renom MA. (2006). Accuracy of Sequence Alignment and Fold Assessment Using Reduced Amino Acid Alphabets, *PROTEINS: Structure, Function, and Bioinformatics*, 63, 986-995
- Micheletti C, Seno F, Maritan A. (2000). Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. *Proteins*, 40, 662-674.
- Moult, J., et al. (2009). Critical assessment of methods of protein structure prediction – Round VIII. *Proteins*, 77, (Suppl 9):1-4
- Murphy LR, Wallqvist A & Levy RM. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding, *Protein Eng.*, 13, (3): 149-152
- 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
- Needleman SB & Wunsch CD. (1970). A general method applicable to search for similarities in the amino acid sequences of two proteins. *J Mol Biol*, 48, 442-453
- Ortiz AR, Strauss CE, Olmea O. (2002). MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, 11:2606-2621
- Pautsch A, Zoephel A, Ahorn H, Spevak W, Hauptmann R & Nar H. (2001). Crystal structure of bisphosphorylated IGF-1 receptor kinase: insight into domain movements upon kinase activation. *Structure*. 9, 955-965
- Rooman MJ, Rodriguez J, Wodak SJ. (1990). Automatic definition of recurrent local structure motifs in proteins. *J. Mol. Biol.*, 213, 327-336
- Sacan A., Toroslu IH & Ferhatosmanoglu H. (2008). Integrated search and alignment of protein structures. *Bioinformatics*, 24, 2872-2879
- Sander O, Sommer I & Lengauer T. (2006). Local protein structure prediction using discriminative models, *BMC Bioinformatics*, 7, 14
- Schuchhardt J, Schneider G, Reichelt J, Schomburg D & Wrede P. (1996). Local structural motifs of protein backbones are classified by self-organizing neural networks. *Protein Eng.*, 9, 833-842

- Shatsky M, Nussinov R, & Wolfson HJ. (2002). Flexible protein alignment and hinge detection, *Proteins*, 48, 242-256
- Shindyalov IN & Bourne PE. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, 11, 739-747
- Subbiah S, Laurents DV & Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.*, 3, 141-148
- Taylor WR & Orengo CA. (1989). Protein structure alignment. *J. Mol. Biol.*, 208, 1-22
- Tung CH, Huang JW & Yang JM. (2007). Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. *Genome Biol.*, 8 R31
- Tyagi M, Gowri VS, Srinivasan N, Brevern AG & Offmann B. (2006). A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications, *Proteins*, 65, (1):32-9
- Tyagi M, Sharma P, Swamy C S, Cadet F, Srinivasan N, Brevern A.G and Offmann B. (2006). Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. *Nucleic Acids Research*, 34, Web Server issue W119-W123.
- Unger R, Harel D, Wherland S & Sussman JL. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins*, 5, 355-373
- Wang S & Zheng WM. (2008). CLePAPS: fast pair alignment of protein structures based on conformational letters. *J. Bioinform. Comput. Biol.*, 6, 347-366
- Werner H, Weinstein D & Bentov I. (2008) Similarities and differences between insulin and IGF-I: structures, receptors, and signaling pathways. *Arch Physiol Biochem.* 114,(1),17-22
- Williams SB, Vakonakis I, Golden SS & LiWang AC. (2002). Structure and function from the circadian clock protein KaiA of *Synechococcus elongatus*: a potential clock input mechanism. *Proc Natl Acad Sci USA.* 26, 99(24), 15357-62
- Yang J. (2008). Comprehensive description of protein structures using protein folding shape code. *Proteins*, 71, 3, 1497-1518
- Yang, J, (2011). Complete Description of Protein Folding Shapes for Structural Comparison. In: *Series: Protein Biochemistry, Synthesis, Structure and Cellular Functions: Protein Folding*. Edited by Walters EC. New York, Nova Science Publishers, 421-442 (ISBN: 978-1-61761-259-6),
- Yuzhen Y, & Adam G. (2003). Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19(suppl.2), ii246-ii255
- Zemla A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res*, 31, 3370-3374
- Zhang Y & Skolnick J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33, 7, 2302-2309



## **Protein Structure**

Edited by Dr. Eshel Faraggi

ISBN 978-953-51-0555-8

Hard cover, 396 pages

**Publisher** InTech

**Published online** 20, April, 2012

**Published in print edition** April, 2012

Since the dawn of recorded history, and probably even before, men and women have been grasping at the mechanisms by which they themselves exist. Only relatively recently, did this grasp yield anything of substance, and only within the last several decades did the proteins play a pivotal role in this existence. In this expose on the topic of protein structure some of the current issues in this scientific field are discussed. The aim is that a non-expert can gain some appreciation for the intricacies involved, and in the current state of affairs. The expert meanwhile, we hope, can gain a deeper understanding of the topic.

### **How to reference**

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Jiaan Yang and Wei-Hua Lee (2012). Protein Structure Alphabetic Alignment, Protein Structure, Dr. Eshel Faraggi (Ed.), ISBN: 978-953-51-0555-8, InTech, Available from: <http://www.intechopen.com/books/protein-structure/protein-structure-alphabetic-alignment>

**INTeCH**  
open science | open minds

### **InTech Europe**

University Campus STeP Ri  
Slavka Krautzeka 83/A  
51000 Rijeka, Croatia  
Phone: +385 (51) 770 447  
Fax: +385 (51) 686 166  
[www.intechopen.com](http://www.intechopen.com)

### **InTech China**

Unit 405, Office Block, Hotel Equatorial Shanghai  
No.65, Yan An Road (West), Shanghai, 200040, China  
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元  
Phone: +86-21-62489820  
Fax: +86-21-62489821

© 2012 The Author(s). Licensee IntechOpen. This is an open access article distributed under the terms of the [Creative Commons Attribution 3.0 License](https://creativecommons.org/licenses/by/3.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

IntechOpen

IntechOpen