Describing and Comparing Protein Structures Using Shape Strings

Nanjiang Shu, Sven Hovmöller* and Tuping Zhou

Structural Chemistry, Arrhenius Laboratory, Stockholm University, SE-106 91 Stockholm, Sweden

Abstract: Different methods for describing and comparing the structures of the tens of thousands of proteins that have been determined by X-ray crystallography are reviewed. Such comparisons are important for understanding the structures and functions of proteins and facilitating structure prediction, as well as assessing structure prediction methods. We summarize methods in this field emphasizing ways of representing protein structures as one-dimensional geometrical strings. Such strings are based on the shape symbols of clustered regions of φ/ψ dihedral angle pairs of the polypeptide backbones as described by the Ramachandran plot. These one-dimensional expressions are as compact as secondary structure description but contain more information in loop regions. They can be used for fast searching for similar structures in databases and for comparing similarities between proteins and between the predicted and native structures.

Keywords: Protein structure, secondary structure, structure comparison, Ramachandran plot, dihedral angle, shape strings.

1. INTRODUCTION

Protein structures can be described numerically or graphically. In either case there is a trade-off between completeness and perspicuity. The only way to comprehensively describe a protein structure is by listing the xyz coordinates of all atoms as found in the Protein Data Bank (PDB) [1] (http://www.pdb.org/). Most protein molecules contain hundreds and even thousands of atoms. It means that thousands of real numbers are needed to describe a protein structure by listing xyz coordinates of all atoms as is done in PDB. Not only is it difficult for the human brain to grasp so much information, it is also not easy to compare the structures of different proteins by computers.

With more than 42,000 (March, 2007) protein structures in the rapidly growing PDB, structure comparison techniques have become increasingly important. It is widely accepted that protein structures are more conserved than their amino acid sequences [2]. Thus, protein structure comparison can be used to detect distant homologues whose sequences have diverged so much that no obvious sequence similarity can be detected. In fact, analyses of structural families have shown that homologous proteins frequently share less than 20% sequence identity [3]. For example, the sequences of triosephosphate isomerase from *Escherichia coli* (PDB code 1TRE_A, 255 amino acids) and that from *Pyrococcus woesei* (1HG3_A, 225 amino acids) share only 18% sequence identity although they both belong to the triosephosphate isomerase (TIM) family and have very similar 3D structures (2.5Å root mean square deviation (RMSD) for 212 aligned Cα atoms, see Fig. **1**). The construction of structural families by comparing available 3D protein structures, in turn, supplies better knowledge about the variation of sequences within protein families [4].

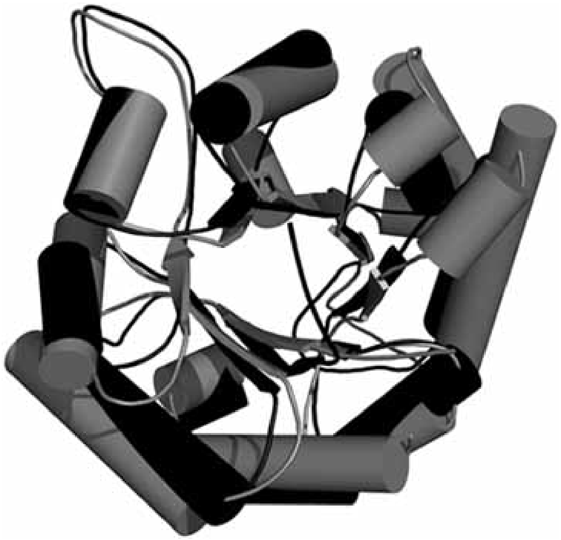


Fig. (1). The protein triosephosphate isomerase from *Pyrococcus woesei* (1HG3_A; black) is aligned with that from *Escherichia coli* (1TRE_A; gray) by rigid body superposition. The RMSD of the alignment based on Cα atoms is 2.5Å, although the sequence identity is only 18%. Structural alignment was made by CE [8]. Cartoon representation generated by PyMOL [9].

SCOP [5] and CATH [6], the two most widely used databases which classify protein structural domains hierarchically, provide a detailed and comprehensive description of the structural and evolutionary relationships between all domains of protein structures deposited in the PDB. SCOP is manually curated while CATH employs both automated computation and manual inspection.

The evaluation of the quality of predicted structure models requires a quantitative measure of the similarities between model structures and real structures, as is done in the Critical Assessment of Structure Prediction (CASP) [7]. In

*Address correspondence to this author at the Structural Chemistry, Arrhenius Laboratory, Stockholm University, SE-106 91 Stockholm, Sweden; E-mail: svenh@struc.su.se

1389-2037/08 $55.00+.00

CASP, hundreds or thousands of model structures are predicted for each target. These predicted model structures are compared with the released structures determined by X-ray crystallography or NMR and then ranked according to the similarity between a model structure and its real structure. Both global and local similarities are taken into account.

Comparing protein structures by superposing all atoms (often simplified by superposing Cα atoms only) of one protein onto the other as rigid bodies is especially powerful in detecting subtle structural changes, including two forms of the same protein under different conditions such as pH or temperature. Nevertheless, it is computationally expensive and cannot detect the similarities between protein structures with large motions (e.g. hinge-bending) and extensive insertions/deletions. This problem becomes crucial as more and more structures are available in PDB. Efficiency and simplicity in structure description and comparison becomes essential. Secondary structure based comparison methods [10-17] were introduced to handle these limitations. These methods compare secondary structure elements (SSEs), which are usually defined by DSSP [18], of protein structures first and then carry out a more careful Cα alignment between pairs of protein molecules (for reviews see Gibrat *et al.* [19], Carugo and Pongor [20] and Carugo [21, 22]). They are fast and can detect distantly related protein structures. However, on average, nearly half the amino acids in protein structures are in so-called random coil or loop regions and their conformations are not defined in the secondary structure description. Methods based on the alignment of SSEs are limited as a consequence.

Can we develop other methods to represent and compare protein structures which are as simple as secondary structure alignment based methods, yet not limited to SSEs but also taking the loop regions into account? Recently, several groups have endeavoured on developing such methods by representing the backbone structures of proteins as discrete torsion angles [23] or 1D strings of path in 3D space [24] or shape symbols [25]. These shape symbols represent clustered regions of φ/ψ torsion angle pairs in the Ramachandran plot [26] which has been revisited recently [27-30]. In this review, these newly developed methods are summarized and compared with methods based on the alignment of Cα atoms and SSEs. The advantages and disadvantages of the various methods are discussed and illustrated by several examples.

2. DESCRIBING PROTEIN STRUCTURES

2.1. Graphical Representation

Visual inspection is one of the most important and usually a first step in studying a protein structure, because this is the best way for the human brain to grasp the information. SCOP [5], which is widely used as the gold standard for protein structure comparisons [31-34], is manually created based on visual inspection. Even in CASP, although various methods have been used to facilitate evaluating the model structure, such as RMSD [35], Global Distance Test Total Score (GDT_TS) [36] (see Equation 1) and MaxSub score (a scalar in the range of 0 and 1, normalized from the size of the largest 'well-predicted' subset) [37], visual inspection is still used as a final decision [38, 39].

The standard way to visualize protein structures in 3D in a simplified way using helices and arrows was pioneered by Jane Richardson [40] who made beautiful drawings by hand (Fig. **2**). Today such drawings are generated by computers, using many macro-molecular visualization tools such as RasMol [41]. Other molecular representations, e.g. ball and stick, space-filling and solid surface are used for various purposes.

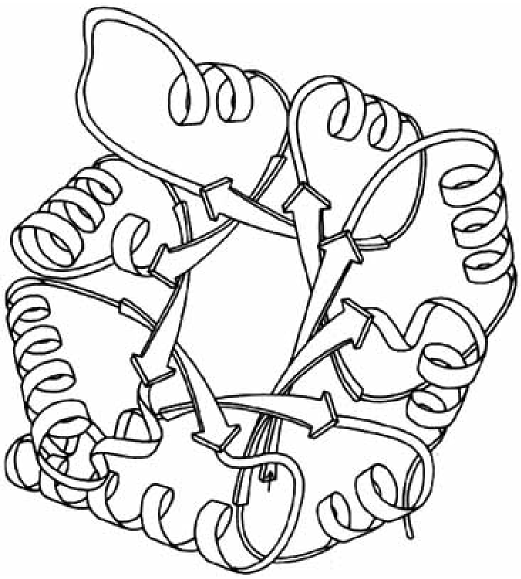
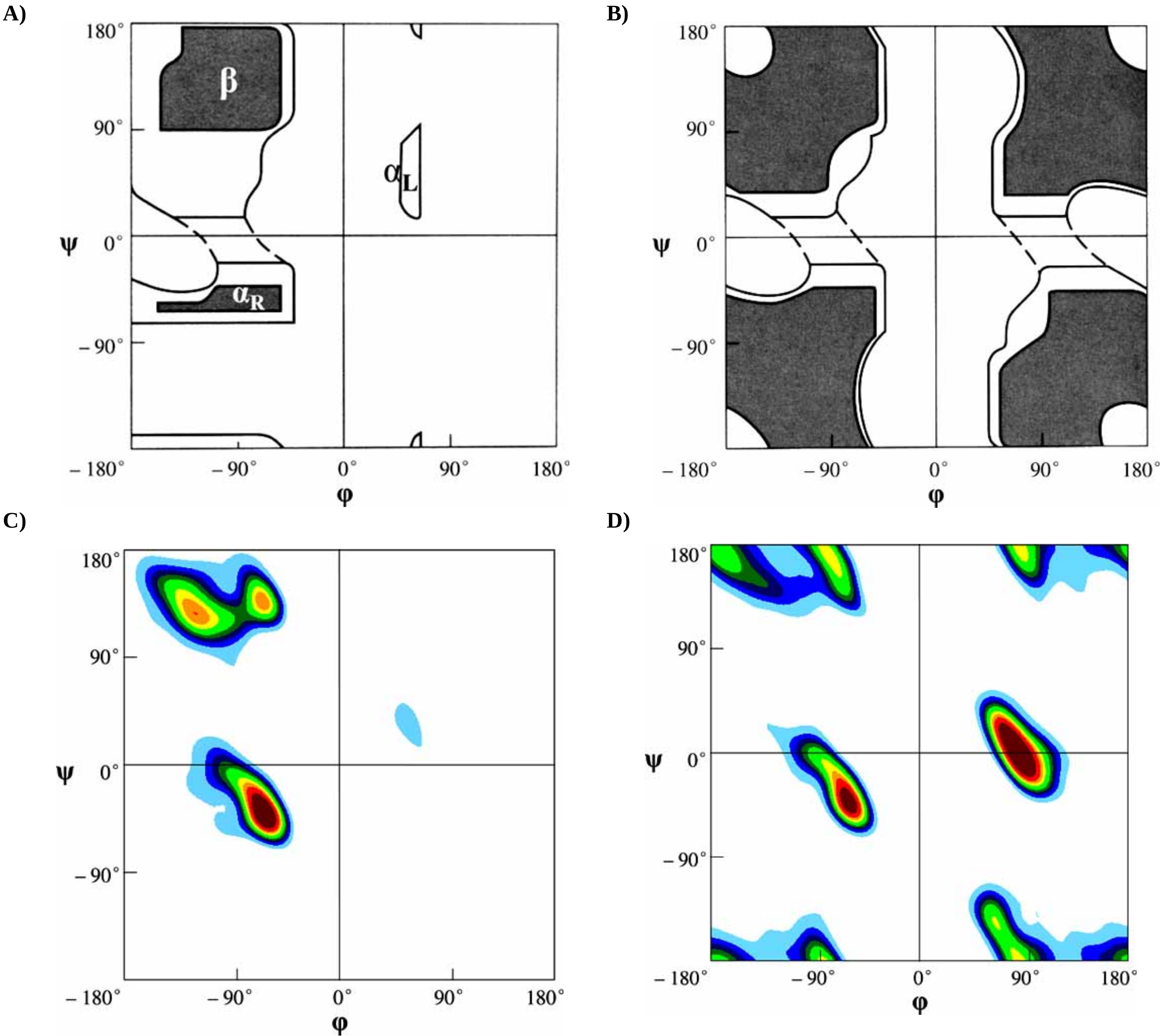


Fig. (2). Illustration of the triosephosphate isomerase (TIM) barrel domain represented by a set of helices and arrows, drawn by Jane Richardson. [Reproduced from web version of Richardson [40] with permission]

2.2. Shape Strings

The peptide bond in proteins is planar [42]. As a result of this, the backbone conformation in a polypeptide chain can be described by a pair of torsion angles (φ and ψ), per residue. Thus, the most compact, yet complete, description of the polypeptide chain needs just two numbers per amino acid. Ramachandran *et al.* [26] noted that only a few combinations of these torsion angles are possible in proteins, as seen in the Ramachandran plot. In their plot (Fig. **3a, b**), Ramachandran *et al.* predicted the commonly allowed regions: α_R, α_L and β-region, for φ/ψ-angle pairs in the Ramachandran plot based on the analysis of steric hindrances of short peptides. Recent studies on the Ramachandran plot in real protein structures, using high-resolution X-ray crystallography results in PDB, show that the allowed regions of φ/ψ-angle pairs in the observed plot differ from the original Ramachandran plot [27-30]. The first main difference is that α_R, α_L and β-sheet regions are diagonal in the observed Ramachandran plot (Fig. **3c, d**) while in the original Ramachandran plot (Fig. **3a, b**) the edges of these regions are mostly parallel to one or both of the φ or ψ axes. The second is that the β-region (Fig **3a**) is split into two diagonal lobes: the β-sheet region (left) and the polyproline II region (right) [28, 29] (Fig. **3c**). The third is that the two most populated regions for glycine (Fig. **3d**) are in regions predicted to be only permissible in the standard Ramachandran plot (Fig. **3b**). These discrepancies were explained partly in terms of local electrostatic interaction by Ho *et al.* [43].



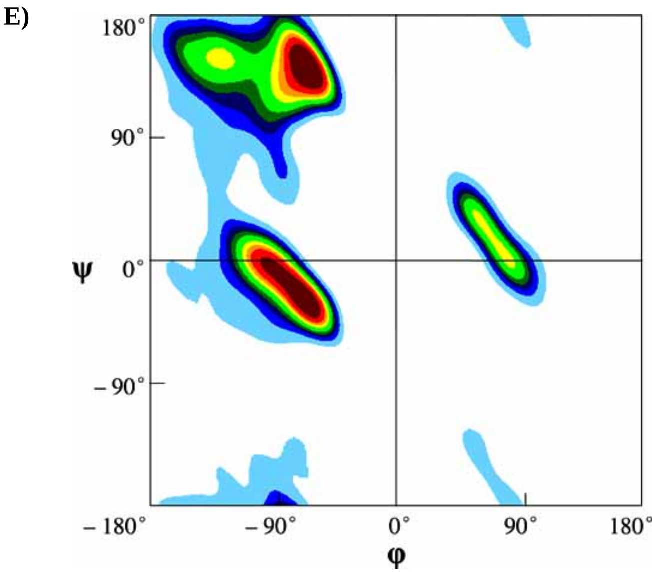


Fig. (3). The classical version of the Ramachandran plot for (a) alanine (but often taken as typical for all non-glycines except proline) with three commonly allowed regions: α_R, α_L and β-region and (b) glycine according to Ramachandran and Saisekharan [26]. The fully allowed regions are shaded and the partially allowed regions are enclosed by solid lines. The connecting regions enclosed by the dashed lines are permissible with slight flexibility of bond angles. These plots were calculated by modelling tripeptides based on steric clashes. Although the overall features of these plots are correct, the details differ from the experimentally observed Ramachandran plots for (c) all 19 non-glycines and (d) glycine. (e) All 20 amino acids in random coils. [Reproduced from Hovmöller *et al.* [28] with permission]

Knowing that the allowed combinations of φ/ψ angles in the Ramachandran plot are highly clustered, we can assign a symbol to each cluster in the Ramachandran plot (see Fig. **6c**). The backbone structure of any protein molecule can then be expressed as a one-dimensional (1D) sequence of such symbols, one symbol for each amino acid; a shape string

[25]. Each shape symbol in the shape string corresponds to a certain region of backbone dihedral angles in the Ramachandran plot. The shape string of an entire protein carries a description of the entire 3D backbone structure. In contrast, the common secondary structure description with only 3 symbols, H (α-helix), E (extended β-strand) and C (coil), can describe the α-helices and β-sheets accurately, but carries no information about the structure of the other half of all residues that are in the coils.

The most abundant shape symbol is the A shape which takes up about 45% of all residues (Table **1** and Fig. **4**). That is because almost all residues in α-helices are of the A shape and some residues in turn regions also have A shape. The second most abundant shape symbol is S which accounts for nearly 25% of all residues. Most of residues in β-sheets are S shape. R shape corresponds to polyproline II regions, but it exists also in some slightly distorted β-strands. The rest of shape symbols are less abundant, but they are also very important since they contain the extra information in the loop regions which is lacking in the standard secondary structure description. The distribution of shape symbols changes dramatically given the shape symbol of the preceding amino acid (Table **1** and Fig. **4**). For example, in total, the A shape accounts for nearly 45% of all residues. However, the probability for a residue with A shape following a residue with A shape is 79% but after an amino acid with S shape, the probability for A is only 11%.

In Fig. (**5**) we investigate the property for a shape to extend itself. The probability for a residue to be A shape following residues with non-A shapes is quite low (only 17%). However, the probability increases quickly to 53% when one previous residue is A shape. This probability becomes even higher when two previous residues are both A shape (Fig. **5a**, See also supplementary data for detailed values). The distribution of 8 shape symbols for the residues following more than two residues all with A shapes becomes nearly constant, keeping the percentage for A shape at a very high (~91%) level. A similar phenomenon exits for the S shape, although the percentages of S shape are not as prominent as that of the A shape.

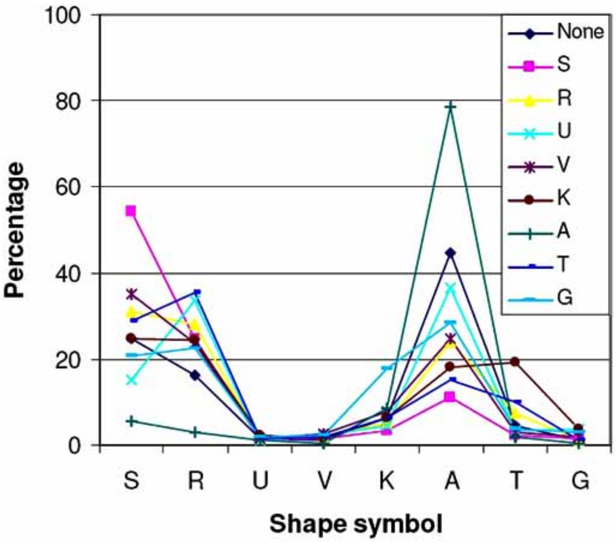


Fig. (4). Distribution of 8 shape symbols for all residues (represented by All in the legend, which gives out the background composition of shape symbols) and those following a residue with each of 8 shape symbols (S, R, U, V, K, A, T and G in the legend). See also Table 1 for detailed percentage values. The result were constructed from a non-redundant set of PDB (version 2006 April) culling at 30% sequence identity containing 4274 unique chains (created by PISCES server [44]).. The same dataset was used in the following text, unless specially mentioned.

2.3. Describing Loop Regions by Shape Strings

The φ/ψ torsion angle pairs of backbone structures in loop regions are scattered over the entire allowed regions in the Ramachandran plot (Fig. **3e**). This is contrary to α-helices (centred at φ = -61°, ψ = -41°) and β-sheets (centred

Table 1. Percentages for 8 Shape Symbols for All Residues and for Those Following a Residue with each of 8 Shape Symbols. Complementary to Fig. (4). For Definition of Shape Symbols, See (Fig. 6C).

Second a.a. First a.a.	S (%)	R (%)	U (%)	V (%)	K (%)	A (%)	T (%)	G (%)
All	24.7	16.2	1.3	1.1	6.4	44.7	4.5	1.2
S	54.2	24.7	1.4	1.4	3.2	11.2	2.3	1.6
R	31.2	27.9	1.5	1.8	5.1	24.0	7.4	1.0
U	15.2	33.6	1.3	2.1	4.4	36.5	3.3	3.7
V	35.0	23.7	1.6	2.5	7.9	24.6	2.8	1.9
K	24.8	24.5	2.3	1.4	6.2	18.2	19.2	3.5
A	5.7	3.1	1.0	0.5	8.4	78.7	2.0	0.6
T	28.9	35.6	1.6	1.8	6.1	15.0	10.1	1.0
G	20.8	22.5	1.8	2.2	17.7	28.4	3.6	3.1

at φ = -116°, ψ = 128° for parallel and φ = -122°, ψ = 135° for anti-parallel) [28]. Despite this irregularity of φ/ψ angle pairs, loops can be classified into distinct classes [40, 45, 46]. Moreover, Panchenko and Madej [47] showed a linear correlation between sequence similarity and average loop structural similarity, which indicates that the loop regions are rather systematic. Loops play an important role as structural determinants connecting the SSEs [40]. Residues in many loop regions are essential in stabilizing the local conformation [48] and are involved in enzymatic activities [49] and protein-protein interactions [50]. The accuracy of loop conformations often determines the usefulness of computational or experimental models, but this remains the most difficult part of comparative homology modelling [51, 52].

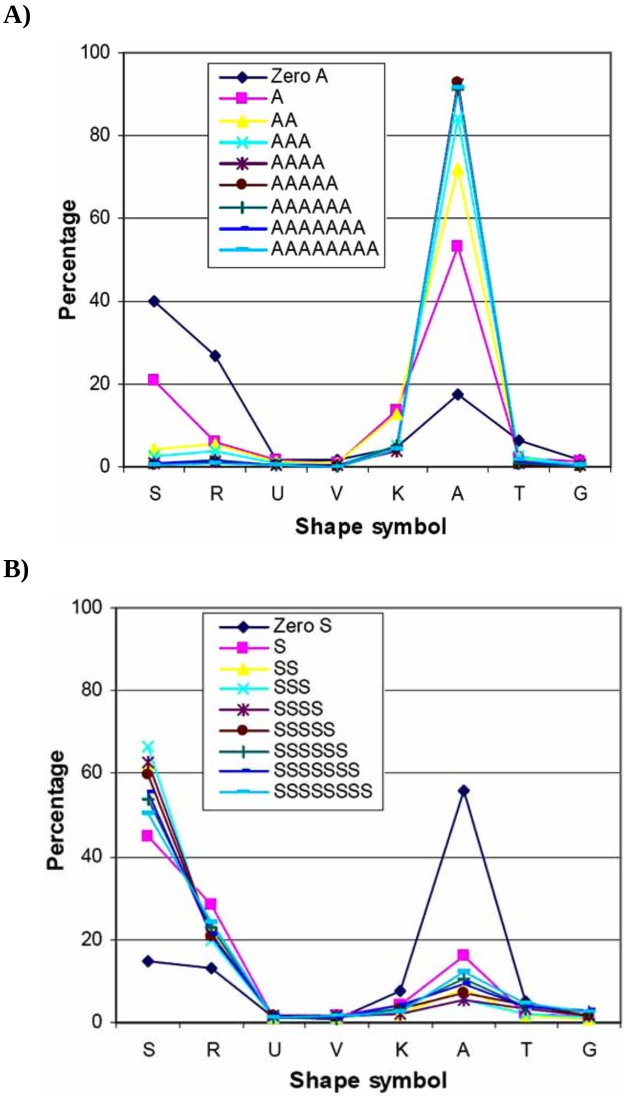


Fig. (5). Distribution of 8 shape symbols for residues following (a) 0 to 8 residues with A shapes and (b) 0 to 8 residues with S shapes. Note that the distribution following zero A is different from the background shape string composition as shown in Fig. **4**. When counting the background composition, residues with A shape are following any shapes, including A, but here zero A means following non-A shapes. Once an α-helix is formed, it has a strong tendency to continue. Most α-helices are terminated by an amino acid with K shape (71%). Most β-strands are terminated by R.

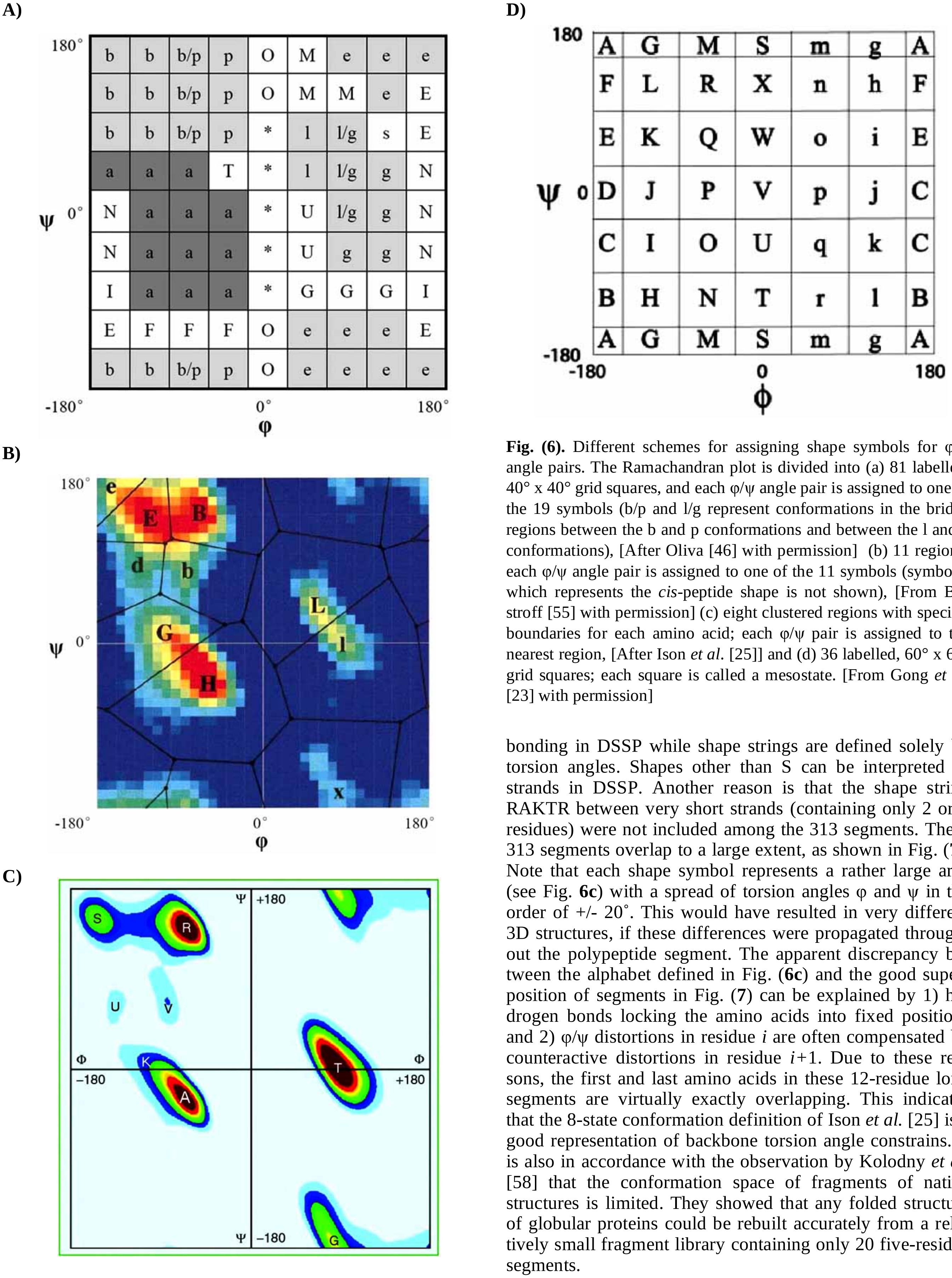
Accurate classification of loop conformation benefits structure predictions by comparative homology modelling. Such classification was originally carried out by visual inspection of protein structures [40, 53]. With the rapidly increasing number of available protein structures, automatic classification methods are required. Oliva *et al.* [46] developed a semi-automated method for classification of structures of short loop regions in proteins by assigning torsion angles into 81 labelled, 40° x 40° grid squares in the Ramachandran plot (Fig. **6a**). Loop conformations were denoted as strings of these shape symbols based on both the main-chain dihedral angles and the geometry of the bracing secondary structures. Their method can reproduce most features of classifications done by manual analysis and identify several novel motifs for loops. Espadaler *et al.* [54] extended this work and developed a program called ArchDB which can classify longer loops than in previous work and also yields a more stable classification. ArchDB carries out clustering based on a density search on the (φ,ψ) space of the loop conformation, which allows a second check by RMSD.

Fitzkee *et al.* [56] have recently built a Protein Coil Library (PCL) which classifies protein structures in loop regions by dividing backbone torsion angles into a coarse-grained 30° x 30° φ, ψ-grid following the work of Srinivasan and Rose [57]. The PCL library stores molecular coordinates, dihedral angles and sequence information for each segment in loop regions. The database also supplies searching and analysis tools, which make the database useful also for prediction and design of loop regions.

An accurate description of the conformation of short turns that connect two secondary structure elements is very important. In Table **2** we list the five most common shape strings of all short turns (from 2 to 5 residues) that connect two helices, a helix and a strand, a strand and a helix, or two strands. The shape strings in the turn regions are rather scattered, showing the rich conformation in these regions.

2.4. Describing Protein Structures by Shape Strings

As mentioned above, backbone structures of proteins can be expressed as 1D strings of shape symbols. These shapes are as compact as 1D strings of secondary structures but contain information also in loop regions. Ison *et al.* [25] assigned torsion angles into only 8 highly clustered regions (Fig. **6c**) in the Ramachandran plot, with specific boundary specification for each amino acid. A database with shape strings following this definition of all proteins in the PDB has been created by Zhou and Hovmöller (www.fos.su.se~/pdbdna). A computer program Frags [25] was developed for exploiting this database for various purposes. A surprising property of shape strings is that they are highly converged in 3D space, i.e., the backbone structures of all fragments with the same shape string are usually very similar in 3D. For example, the shape string segment SSSSRAKTRSSS (see Fig. (**6c**) for symbol assignment) is found 313 times in the non-redundant subset of PDB containing 4274 protein chains mentioned above. Note that the number of shape string SSSSRAKTRSSS retrieved by Frags is only 313, which is considerably less than 492, the number of shape string RAKTR found between two strands (Table **2**). One reason is that secondary structures are defined by the hydrogen



A)

b	b	b/p	p	O	M	e	e	e
b	b	b/p	p	O	M	M	e	E
b	b	b/p	p	*	l	l/g	s	E
a	a	a	T	*	l	l/g	g	N
N	a	a	a	*	U	l/g	g	N
N	a	a	a	*	U	g	g	N
I	a	a	a	*	G	G	G	I
E	F	F	F	O	e	e	e	E
b	b	b/p	p	O	e	e	e	e

D)

A	G	M	S	m	g	A
F	L	R	X	n	h	F
E	K	Q	W	o	i	E
D	J	P	V	p	j	C
C	I	O	U	q	k	C
B	H	N	T	r	l	B
A	G	M	S	m	g	A

Fig. (6). Different schemes for assigning shape symbols for φ/ψ angle pairs. The Ramachandran plot is divided into (a) 81 labelled, 40° x 40° grid squares, and each φ/ψ angle pair is assigned to one of the 19 symbols (b/p and l/g represent conformations in the bridge regions between the b and p conformations and between the l and g conformations), [After Oliva [46] with permission] (b) 11 regions; each φ/ψ angle pair is assigned to one of the 11 symbols (symbol c which represents the *cis*-peptide shape is not shown), [From Bystroff [55] with permission] (c) eight clustered regions with specific boundaries for each amino acid; each φ/ψ pair is assigned to the nearest region, [After Ison *et al*. [25]] and (d) 36 labelled, 60° x 60° grid squares; each square is called a mesostate. [From Gong *et al*. [23] with permission]

bonding in DSSP while shape strings are defined solely by torsion angles. Shapes other than S can be interpreted as strands in DSSP. Another reason is that the shape string RAKTR between very short strands (containing only 2 or 3 residues) were not included among the 313 segments. These 313 segments overlap to a large extent, as shown in Fig. (**7**). Note that each shape symbol represents a rather large area (see Fig. **6c**) with a spread of torsion angles φ and ψ in the order of +/- 20˚. This would have resulted in very different 3D structures, if these differences were propagated throughout the polypeptide segment. The apparent discrepancy between the alphabet defined in Fig. (**6c**) and the good superposition of segments in Fig. (**7**) can be explained by 1) hydrogen bonds locking the amino acids into fixed positions and 2) φ/ψ distortions in residue *i* are often compensated by counteractive distortions in residue *i*+1. Due to these reasons, the first and last amino acids in these 12-residue long segments are virtually exactly overlapping. This indicates that the 8-state conformation definition of Ison *et al.* [25] is a good representation of backbone torsion angle constrains. It is also in accordance with the observation by Kolodny *et al.* [58] that the conformation space of fragments of native structures is limited. They showed that any folded structure of globular proteins could be rebuilt accurately from a relatively small fragment library containing only 20 five-residue segments.

Table 2. Frequencies of the shape strings of short turns or loops (2 to 5 amino acids long), connecting two helices (H*H), a helix and a strand (H*S), a strand and a helix (S*H), and two strands (S*S), respectively. For each case the five most frequent shape strings are listed. As the loops get longer, there are of course more possible shape strings, making each individual shape string less abundant, as seen by low percentages. Note, however, that the shape string RAKTR is very common between two strands. See also Fig. (7) for structural alignment of two strands connected by the five-long turn with shape string RAKTR. The secondary structure is defined by DSSP [18] while the shape strings are defined according to Fig. (6c) [25]. The existence of A shape following a helix, e.g. the AS shape string between two helices, are caused by differences in definitions of DSSP and shape strings. The statistics is based on a non-redundant set of PDB containing 4274 protein chains as mentioned before.

H*H				H*S			S*H			S*S		
Size of turn	**Shapes**	**Count**	**%**	**Shapes**	**Count**	**%**	**Shapes**	**Count**	**%**	**Shapes**	**Count**	**%**
2	RR	551	18.6	RA	316	14.9	AS	438	17.5	TT	1600	38.2
	SR	260	8.8	TR	289	13.7	SR	339	13.6	GK	697	16.7
	KR	221	7.4	SA	215	10.2	RR	289	11.6	AK	251	6.0
	AS	170	5.7	TS	209	9.9	KS	226	9.1	GA	184	4.4
	RS	162	5.5	KT	132	6.2	SS	135	5.4	RT	161	3.9
3	TSR	248	8.4	KTR	410	13.3	SAS	132	6.8	SAK	115	5.6
	KRR	165	5.6	TRA	277	9.0	SKS	67	3.4	RRR	102	4.9
	ARR	156	5.3	KTS	264	8.5	RRR	66	3.4	AKG	91	4.4
	TRR	146	4.9	TSR	190	6.1	RAS	56	2.9	SAA	79	3.8
	KSR	112	3.8	ATR	148	4.8	ASR	54	2.8	ASR	74	3.6
4	KTRR	198	8.4	KTRA	203	7.2	AKRR	37	2.2	AAKT	286	9.1
	KTSR	118	5.0	ATRA	120	4.3	RRSR	34	2.0	AAAT	282	9.0
	ATRR	95	4.0	KTSR	110	3.9	SRRR	22	1.3	ASAK	144	4.6
	KTSS	56	2.4	KTRK	93	3.3	SAAR	21	1.3	RRTR	137	4.4
	KTRS	49	2.1	KTRR	88	3.1	RRRR	20	1.2	AKTR	131	4.2
5	KTASR	60	3.5	RRRTR	38	1.9	RAKRR	33	2.1	RAKTR	492	18.3
	ATASR	33	1.9	KTASA	38	1.9	RAARR	26	1.6	RAATR	153	5.7
	TSRRR	28	1.6	ATASA	32	1.6	RRTRR	23	1.4	RAKTS	97	3.6
	UAARR	18	1.1	RSRTR	30	1.5	SSAAS	15	0.9	SAKTR	76	2.8
	KTKSR	18	1.1	KTRAS	19	1.0	RRTRS	11	0.7	AAATR	67	2.5

Gong *et al.* [23] proved that it is possible to rebuild native protein conformation from highly approximated torsion angles grouped into 36 labelled, 60° x 60° grid squares, each called a mesostate (Fig. **6d**). The 3D backbone protein structure is hence represented by a linear string of mesostates. The rebuilding procedure was carried out first by replacing each five-residue fragment within a target mesostate sequence with candidates selected from the pre-built fragment library. Then the fragments were assembled by Monte Carlo simulation with simulated annealing by using an energy function with three simple terms: (i) steric exclusion, (ii) hydrogen bonding and (iii) global compaction. Despite the large range of each mesostate (60° x 60°) and the crude energy function, near native protein conformations could be rebuilt from such very approximate mesostate strings. This discovery is important because approximate torsion angles can either be obtained directly from NMR spectroscopy [59] or predicted from torsion angle prediction methods. The latter means that the prediction of discrete torsion angles can be a good starting point for *ab initio* 3D structure prediction. Bystroff *et al.* [55] pioneered the prediction of torsion angles with a Hidden Markov Model in their HMMSTR package. The Ramachandran plot was divided into 11 conformational states (Fig. **6b**). Kuang *et al.* [60] predicted backbone torsion angles based on machine learning methods. The torsion angles were grouped into three or four conformation states which were derived from the definition of the conformation state by Oliva *et al.* [46].

2.5. Other 1D Expressions of Backbone Structures

Several groups have endeavoured on 1D expression of backbone structures different from those introduced above.

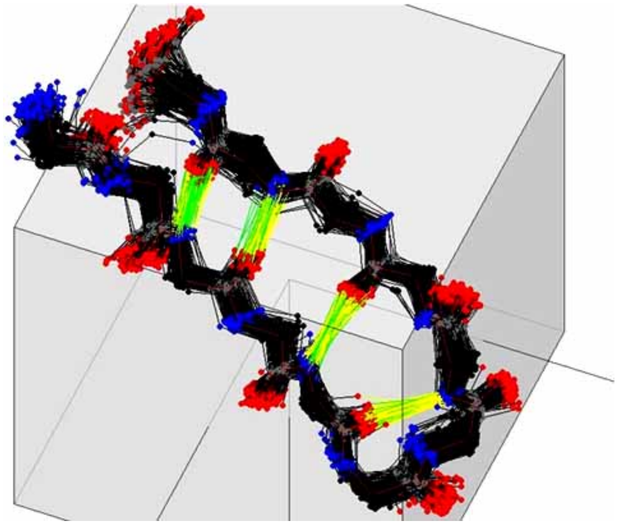


Fig. (7). All 313 backbone fragments with the same shape string segment SSSSRAKTRSSS (see Fig. **6c** for symbol assignment) found in 4274 unique protein chains (the non-redundant PDB mentioned above). They are very similar in the 3D conformation. Yellow/green lines are hydrogen bonds. Retrieved by Frags © Roger Ison [25].Colouring scheme: blue = nitrogen N, red = oxygen O, black = carbon Cα, grey = carbon C, green = strong (i.e. short) hydrogen bonds between N and O, yellow = weak (long) hydrogen bonds.

Instead of assigning to each residue a shape symbol based only on the torsion angles of itself, Zhi *et al.* [24] developed a highly simplified description of protein structure that minimized local structural information by smoothing the protein backbone. The protein backbone was smoothed by averaging positions of Cα atoms in a seven-residue window. It abstracted the protein structure as whether it is locally straight or curved. Even at this highly simplified level of protein structure description, the program can still classify structural domains successfully as compared with the SCOP [5] classification and other existing structural alignment programs. A potential reason for the success was explained by Zhi *et al.* [24] that natural proteins were constrained into a compact shape and there were only a limited number of ways to arrange a given turning angle series into a realistic compact shape. This explanation is consistent with the finding of Ison *et al.* [25] that the backbone structures of fragments with the same shape string are highly overlapping.

Friedberg *et al.* [61] represented protein structures as 1D strings called KL-string as developed by Kolodny and Levitt [58]. The KL-string was built based on a library of 20 fragments of protein backbone. Only the Cα atoms were used for each amino acid. Fragments were chosen randomly but with secondary structure constrain. The library of fragments was generated by grouping 7133 five-residue long non-overlapping fragments from 200 protein domains into 20 clusters based on their RMSD from one another as described in Kolodny and Levitt [62]. These 20 elements served as building blocks to encode protein structures as KL-strings. Kolodny and Levitt [62] showed that small all-α proteins could be adequately recreated by these 20 fragments, indicating that KL-strings are satisfactory 1D representations of the 3D protein structure. Given that these fragments have only five residues, loop regions much longer than five residues cannot be rebuilt correctly, since the selection of the fragments is constrained only by the SSEs, but there are no SSEs in loop regions.

X-ray structures deposited in PDB are all rigid. However, we should always keep in mind that protein structures are dynamical *in vivo*. Describing protein structure dynamics is outside the scope of this paper. Readers interested in protein dynamics can refer to Urbanc *et al.* [63], Krebs *et al.* [64] and Ming *et al.* [65].

3. COMPARING PROTEIN STRUCTURES

Comparing protein structures includes identifying similarities and dissimilarities among protein structures. This is often referred to as protein structure alignment. Although significant progress has been made over the past decades, a unique, reliable, fast and convergent method for 3D protein structure alignment is still lacking. Defining the similarity between two protein structures still remains a major problem [66]. When aligning two rather similar protein structures, rigid body superposition can overlay one structure onto the other successfully. The similarity between them can be easily measured by the overall RMSD of for example Cα positions. On the other hand, when comparing distantly related protein structures, identifying the equivalent core residues between compared protein structures remains difficult. This results in considerable ambiguity in describing the similarity between protein structures [67]. Many structural alignment methods yield different alignments for remotely related protein structures [32, 68].

3.1. Rigid Body Superposition

The earliest approach to compare two protein structures is to superpose all atoms of one protein onto the other as rigid bodies, as pioneered by Rao and Rossman [35]. The structural similarity is measured as the RMSD of the Euclidean distances between atoms (usually just the Cα) of equivalent residues [35]. Many algorithms have been proposed for comparing protein structures based on rigid body superposition (for a review see Holm and Sander [69]). These methods vary in procedures for identifying the equivalent core residues and the process for finding the optimal rotation. Comparisons between protein structures under different conditions are quite common in a variety of fields, such as structural biology [70] and drug design [71]. In these fields, protein structures under different physiological conditions, e.g. temperature, pH, ionic concentration or ligand binding/unbinding, are studied extensively. In those cases, usually only subtle conformational changes on a few residues or even side chain atoms take place.

Many studies have proven that two protein structures must be rather similar if their RMSD is small [72, 73]. However, it should also be noticed that the RMSD is affected not only by the conformational similarity, but also by the overall size [74] and the accuracy of the experimentally determined structures [75] of the proteins being compared. Due to these limitations, RMSD can not be used directly as a scale to measure the similarity between two protein structures, that is, one can not assure that two structures with a smaller

RMSD are more similar than two protein structures with a larger RMSD. As a consequence, Carugo and Pongor [76] suggested using $rmsd_{100}$, which normalized the RMSD of two protein structures as if they had 100 equivalent Cα atoms, to measure the similarity between two aligned structures. This method was then extended by Carugo [77] by introducing the P values to access the statistical significance between two $rmsd_{100}$ values. Comparing protein structures only by superposing the whole protein structure as rigid bodies is simple and understandable and can detect subtle conformation changes between similar structures, but it can not detect all similarities between two proteins with large conformation changes such as hinge-bending. Hinges are common in proteins in different conditions of ligand binding. An example is the calcium-binding protein calmodulin which undergoes a significant conformational change via binding of a 25-residue peptide (Fig. **9**). The rigid body superposition method is further challenged in comparing distantly related protein structures where extensive insertions/deletions exist. Another drawback of this method is that it is much more computationally expensive compared to sequence alignments.

Improvements of the rigid body superposition method have been proposed for comparing protein structures with flexibilities by introducing twists at positions with hinge-bending movements. This approach was originated by Wriggers and Schulten [78] and Verbitsky *et al.* [79] by assuming prior knowledge of the location of potential hinges. Shatsky *et al.* [80] presented an algorithm, FlexProt, which automatically detected hinge regions in protein structures. In their algorithm, the protein molecules to be compared were divided into a minimal number of separate fragments with maximal size. The equal-size fragment pairs were then detected based on RMSD and later the rigid fragment pairs were linked together and finally consecutive fragment pairs with a similar 3D transformation were clustered. This algorithm is as efficient as rigid body structure alignment algorithms despite the fact that FlexProt can detect hinge regions automatically. Ye and Godzik [81] developed another algorithm, FATCAT, for flexible structural alignment using dynamic programming to connect aligned fragment pairs (AFP). FATCAT introduces fewer twists with similar RMSDs compared to FlexProt.

3.2. Structural Alignment Based on SSEs

With the introduction of twists, protein structures with large motions (e.g. hinge-bending) can be aligned successfully, as was done in FlexProt and FATCAT. Even so, rigid body superposition still fails to align distantly related protein structures where extensive insertions/deletions exist. Since insertions/deletions occur more often in loop regions, one approach is to first discard the more variable loop regions and compare only SSEs between proteins. A more careful Cα alignment is performed later when the equivalence among SSEs of compared structures are found [10-17] (refer to Carugo and Pongor [20] for more details). Yang and Honig [17] introduced the protein structural distance (PSD) to measure similarities between compared protein structures, taking both high level secondary structure alignment and Cα alignment into account.

Another advantage of structure alignment based on SSEs is its efficiency. The average number of SSEs for a protein with a length of 300 amino acids is 20 to 25, i.e., more than one order of magnitude less than its number of Cα atoms. This makes structure comparisons based on SSEs much faster than comparisons based on rigid body superposition of Cα atoms. Therefore, structure comparisons based on SSEs are suitable for fast database searching of similar structures. However, when comparing two different protein structures with similar SSE orders, they might fail to recognize the difference in global conformation, since they neglect the conformations of loop regions connecting SSEs (see Fig. **9** for an example of two proteins with the same secondary structure orders while having significantly different topology). This will result in many false positives in fast database searching for similar structures. In addition, the determination of SSEs from 3D protein structures is not without ambiguities, especially at the beginning and end of SSEs. Andersen and Rost [82] recently reviewed the principles of the most popular assignment methods DSSP [18], STRIDE [83], DEFINE [84] and P-Curve [85]. They found that DEFINE and P-curve and likewise DEFINE and DSSP agreed in 74% of all residues, P-Curve and DSSP agreed in 79% of all residues, and DSSP and STRIDE agreed in 96% of all residues. For the assignment of β-turns, the agreement is lower, for example, DSSP and STRIDE agreed in 89% of all β-turn residues, and DSSP and DEFINE agreed in 60% of all β-residues [86].

3.3. Comparison of Different Methods

Different descriptions of protein structures are used for different purposes of structure comparisons. In studying subtle changes in side chain conformations of the active site residues upon ligand binding, methods superposing all atoms are most useful. When comparing only the polypeptide backbone, pair-wise rigid body superposition of Cα atoms may be sufficient. However, methods based on rigid body superposition do not satisfy the need for fast structure database searches. They are computationally too expensive and also cannot align remotely related structures. Methods based on the alignment of SSEs are fast and can handle extensive insertions/deletions but might miss the global similarity of compared protein structures. The 1D expression of backbone structures based on torsion angles [25, 87], averaged Cα positions [24] or basic fragments [61], as described in previous sections, are as compact as the traditional secondary structure description. Since the structure is reduced to a 1D sequence, standard dynamic programming can be used. Thus, these 1D strings of shape symbols are as efficient as secondary structure sequences for fast database searching for similar structures. In addition, they are more accurate in identifying structural similarities because of their extra information in loop regions compared to the three-state secondary structure description.

We benchmarked FATCAT, CE, KL string alignment, shape string alignment and BLAST pairwise sequence alignment in finding the similar protein structures based on the FSB dataset used in Friedberg *et al.* [61]. This dataset was constructed on the basis of SCOP 1.61 cutting at 40% sequence identity. The benchmark has 6233 pairs of similar structures (within the same SCOP fold) and 8769 pairs of dissimilar structures (not in the same SCOP fold). The results for FATCAT and CE were obtained from the FATCAT website (http://fatcat.burnham.org/fatcatbench/) and the results for KL string alignment and BLAST pairwise sequence alignment were obtained from Friedberg *et al.* [61]. The shape string alignment was carried out using the standard Smith and Waterman dynamic programming with -12 and –2 as penalties for gap opening and extension, respectively. The raw alignment scores were used to represent the structural similarity. The substitution matrix was derived from structural alignment of 2430 pairs of SCOP domains from SCOP 1.38 cutting at 40% sequence identity (the structural alignments were obtained from Levitts website: http://csb.stanford.edu/levitt/, see also the supplemental data for the substitution matrix used by the shape string alignment). The ROC curves for different methods are shown in Fig. **8**. According to the ROC-curves, FATCAT is the most successful method to identify the similarities in protein structures, followed by CE. Among three 1D alignment based methods, namely BLAST, KL string alignment and shape string alignment, shape string alignment and KL string alignment performed significantly better than BLAST, the amino acid sequence based alignment. Shape string alignment performed slightly better than the KL-string alignment. As to the computational efficiency, BLAST is the fastest method, KL string alignment and shape string are at the same level and these two are both about three orders of magnitude faster than CE and FATCAT.

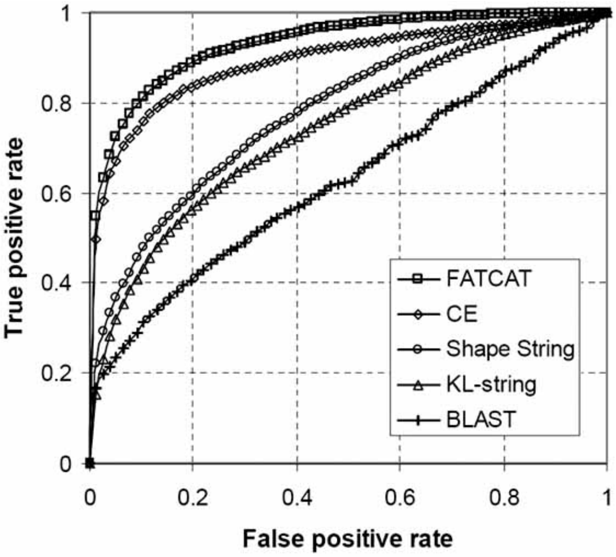


Fig. (8). ROC curves for five methods, FATCAT, CE, Shape string alignment, KL-string alignment and BLAST pairwise sequence alignment in identifying similar protein structures benchmarked on a dataset constructed from SCOP 1.61 cutting at 40% sequence identity.

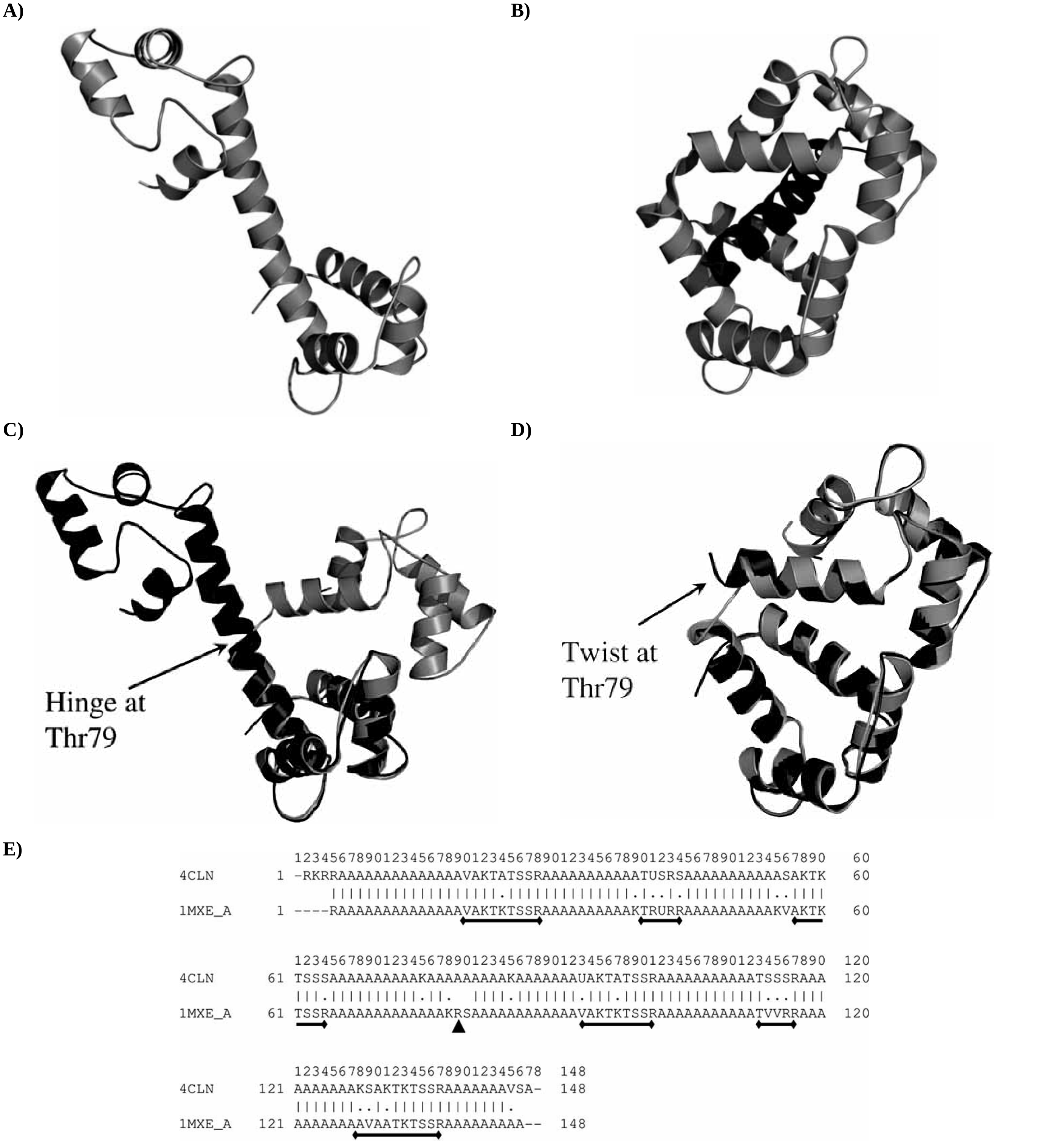
Two examples are given below to demonstrate the efficiency of shape strings in comparing protein structures.

3.3.1. Comparing Protein Structures with Hinges

Calmodulin from *Drosophila melanogaster* is a calcium binding protein. It has quite different 3D conformations in two PDB entries, 4CLN [88] (Fig. **9a**) and 1MXE_A [89] (Fig. **9b**). Note that both 4CLN and 1MXE_A are calcium-loaded. The structure of 1MXE_A differs greatly from that of 4CLN because of the binding of an alpha helical peptide (Fig. **9b**). A standard pair-wise Cα structural alignment considering the whole protein as a rigid body does not superpose one protein onto the other. Only half of the residues are close in 3D space (Fig. **9c**). In contrast, shape string alignment clearly shows the structural similarity of the two molecules throughout the entire protein (Fig. **9e**). The shape symbols are all matched perfectly, except for the significant mismatches at positions 79 and 80 at the location of the hinge. Other minor mismatches, e.g. position 128 and 129 and at the N- and C-terminals, correspond to smaller deviations in structure at these regions. Note that loop regions as marked in Fig. (**9e**) contain complex and valuable strings of shape symbols; while three-state secondary structure description will ignore this rich conformational information by setting them all as C (coil). Despite the complexity of shape symbols in loop regions, shape string alignment can align them satis-factorily, indicating its advantage in aligning loop regions. FATCAT [81] can also detect the similarity of the whole structure successfully by allowing one twist at Thr79 of 4CLN (Fig. **9d**), but it is about three orders of magnitude more computational expensive than shape string alignment.

3.3.2. Comparing Protein Structures with the Same SSEs

The A chain of the protein tagatose-1,6-bisphosphate (1GVF [91]) and the A chain of human mitochondrial 2,4-dienoyl-CoA reductase (1W6U [92]) have the same order of secondary structure elements HEHEHEHEHEHEHEHEH (H: α-helix, E: β-strand). While 1GVF_A has a TIM beta/ alpha-barrel fold, 1W6U_A has an NAD(P)-binding Rossmann-fold domain (Fig. **10a, b**) according to SCOP [5]. Rigid body superposition as is done by CE reveals the great differences in 3D structure between 1GVG_A and 1W6U_A (Fig. **10c**). However, a simple dynamic programming based shape string alignment (Fig. **10e**) reveals not only the similarity in secondary structures (by the well aligned A shape and R/S shape) but also the massive global conformational difference (by the extensive gaps and mismatches of these two proteins, since those gaps and unmatched regions are all representing different 3D conformations). Note that the mismatches between A and T, R and T and S and T (see Fig. **6c** for symbol assignment) in some loop regions as marked in Fig. (**10e**) indicate as much as 180° difference in the φ angles. In addition, the equally efficient structural alignment based simply on SSEs (note that here we mean the pure one-dimensional SSE alignment) might not distinguish the difference between these two structures due to the same orders of secondary structure elements for 1GVF_A and 1W6U_A. Interestingly, FATCAT superposes 1GVG_A quite well (significantly similar annotated by FATCAT server with P-value of 1.31e-2) on 1W6U_A, which falsely gives an impression that these two protein structures are quite similar in 3D. This is because FATCAT has introduced too many twists (5 twists) in the alignment. The introduction of each twist will disregard the structural dissimilarity in a short loop region. The introduction of too many twists tends to result in an overoptimistic structure alignment, especially for proteins with similar SSEs but different overall topologies, such as the proteins in this example.

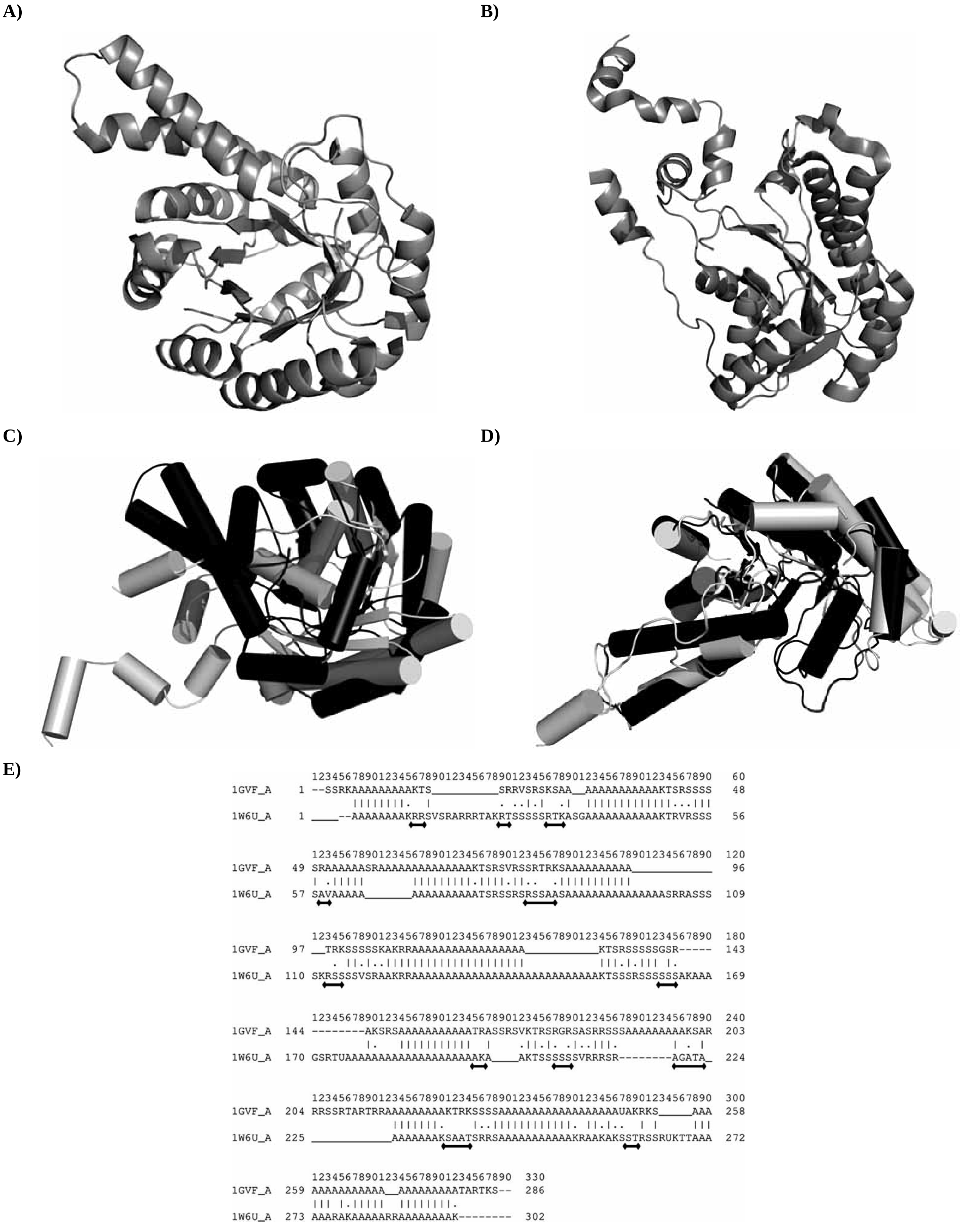


```
              123456789012345678901234567890123456789012345678901234567890   60
4CLN        1 -RKRRAAAAAAAAAAAAAAVAKTATSSRAAAAAAAAAAAATUSRSAAAAAAAAAASAKTK   60
                  |||||||||||||||||||.|||||||||||||||.|..|.|||||||||..||||
1MXE_A      1 ----RAAAAAAAAAAAAAAVAKTKTSSRAAAAAAAAAAAKTRURRAAAAAAAAAKVAKTK   60

              123456789012345678901234567890123456789012345678901234567890  120
4CLN       61 TSSSAAAAAAAAAAKAAAAAAAAAKAAAAAAAUAKTATSSRAAAAAAAAAAATSSSRAAA  120
              |||.||||||||||.||.  ||||.|||||||.|||.||||||||||||||||...||||
1MXE_A     61 TSSRAAAAAAAAAAAAAKRSAAAAAAAAAAAAVAKTKTSSRAAAAAAAAAAATVVRRAAA  120
                                ▲

              1234567890123456789012345678  148
4CLN      121 AAAAAAAKSAKTKTSSRAAAAAAAVSA-  148
              |||||||..|.|||||||||||||.
1MXE_A    121 AAAAAAAAVAATKTSSRAAAAAAAAA--  148
```

Fig. (9). Different 3D conformation of calmodulin in PDB structures (a) 4CLN [88] and (b) 1MXE_A[89] as well as the structural alignment of 4CLN and 1MXE_A made by (c) rigid body superposition (by program CE [8]), (d) FATCAT [81] and (e) shape string alignment (using standard dynamic programming of Needleman and Wunsch [90]). The 3D structure of calmodulin in the PDB structure 1MXE_A undergoes a dramatic conformational change upon binding of a helical peptide (shown in black b). Note that this conformational change is different from the conformation change of calmodulin upon calcium binding. In the structural alignment (c) and (d), 4CLN is shown in black while 1MXE_A is in grey. In the shape string alignment, small conformational changes, i.e. between two regions adjacent in the Ramachandran plot (Fig. **6c**) are marked by '.'. Cartoon representations were made by PyMOL [9].
'- - -' (hyphen): Positions without shape string definition at C- and N-terminals as well as residues without atomic coordinates in PDB or with too high temperature factor, the same in Fig. **10e**. ▲: The position of hinge.⟷: Loop regions.



```
             123456789012345678901234567890123456789012345678901234567890   60
1GVF_A     1 --SSRKAAAAAAAAAKTS__________SRRVSRSKSAA__AAAAAAAAAAAKTSRSSSS   48
                   ||||||||.  |          . ..|.|  .|  |||||||||||||...|||
1W6U_A     1 ____--AAAAAAAAKRRSVSRARRRTAKRTSSSSSRTKASGAAAAAAAAAAAKTRVRSSS   56

             123456789012345678901234567890123456789012345678901234567890  120
1GVF_A    49 SRAAAAAASRAAAAAAAAAAAAAKTSRSVRSSRTRKSAAAAAAAAAAA____________   96
             | .|||||       |||||||||.||||.||..  .|||||||||||
1W6U_A    57 SAVAAAAA_______AAAAAAAAAATSRSSRSRSSAASAAAAAAAAAAAAAAASRRASSS  109

             123456789012345678901234567890123456789012345678901234567890  180
1GVF_A    97 __TRKSSSSSKAKRRAAAAAAAAAAAAAAAAA___________KTSRSSSSSGSR-----  143
                . ||.|..|||||||||||||||||||||           |||.|.||| |.
1W6U_A   110 SKRSSSSVSRAAKRRAAAAAAAAAAAAAAAAAAAAAAAAAAAAKTSSSRSSSSSSAKAAA  169

             123456789012345678901234567890123456789012345678901234567890  240
1GVF_A   144 ---------AKSRSAAAAAAAAAATRASSRSVKTRSRGRSASRRSSSAAAAAAAAAKSAR  203
                     |.   |||||||||||  |    .||.|. .|..|||.        | . |
1W6U_A   170 GSRTUAAAAAAAAAAAAAAAAAAAAKA____AKTSSSSSVRRRSR---------AGATA_  224

             123456789012345678901234567890123456789012345678901234567890  300
1GVF_A   204 RRSSRTARTRRAAAAAAAAAKTRKSSSAAAAAAAAAAAAAAAAAAAUAKRKS_____AAA  258
                         |||||||.    |..||||||||||||. ||.|..  | |     |||
1W6U_A   225 ____________AAAAAAAKSAATSRRSAAAAAAAAAAAAKRAAKAKSSTRSSRUKTTAAA  272

             123456789012345678901234567890  330
1GVF_A   259 AAAAAAAAAAA__AAAAAAAAATARTKS--  286
             ||| |.|||||  ||||||||.
1W6U_A   273 AAARAKAAAAARRAAAAAAAAAK--------  302
```

Fig. (10). Cartoon representations of the different proteins 1GVF_A (a) and 1W6U_A (b), and the structural alignments by (c) CE, (d) FATCAT and (e) shape strings of these two proteins. In (c) and (d), 1GVF_ A is shown in black while 1W6U_A is in light grey. Rigid body

(Legend Fig. 10) contd….

superposition as was done by CE (c) reveals the dissimilarity of these two structures in 3D. However, FATCAT aligns these two structures quite well by introducing five twists, which might result in an overoptimistic impression about the similarity between these two structures. On the other hand, the shape strings match the secondary structure elements while also revealing the many conformation differences in loop regions by the extensive gaps and mismatches in the loop regions as marked by the fat horizontal double-arrows.
'____' (underscore): Gap. '- - -' (hyphen): Positions without shape string definition. ⟷: Loop regions with significant mismatch in shape string.

Shape string alignment still has its limitations. Alignment between all alpha proteins (or all beta proteins) tends to give out a high score for alignment even though the overall topologies are different. Therefore, the shape string identity is not sufficient for proving structural similarity. In contrast to the well studied amino acids sequence alignment methods [93-95], no study has yet been done to assess the statistical significance of shape string alignment.

3.4. Comparing Model Structures to Real Structures

In protein structure prediction, a common question is to assess the quality of the predicted structures. The judgement of the prediction accuracy of a model structure is often carried out by comparing the model structure to its real structure as is done in CASP.

Comparing the similarity between model structures predicted by various 3D structure prediction methods and their real structures is essential for evaluating structure prediction methods. Generally speaking, the comparison between a model structure and its real structure is relatively easy since they have the same amino acid sequence, which means the equivalent residues are already determined. Hence, the major difficulty lies in the measurement of the similarity between protein structures that are compared. Model structures generated by comparative homology modelling based on close homologues are usually quite close to their real structures. Thus the similarity can be easily measured by the overall RMSD or normalized RMSD (e.g. rmsd_{100}) after optimal superposition of one protein onto the other as rigid bodies. However, for model structures predicted based on fold recognition or even *de novo* structure prediction when no obvious homology can be detected from the available structure database, rigid body superposition is not sufficient any more, since these model structures are often quite different from their real structures. In CASP, one of the standard evaluation methods is the GDT_TS (Global Distance Test Total Score) [36, 96],

$$\text{GDT\_TS} = (\text{GDT\_P1} + \text{GDT\_P2} + \text{GDT\_P4} + \text{GDT\_P8})/4 \qquad (1)$$

where GDT_Pn denotes percent of residues under distance cutoff ≤ nÅ. The GDT_TS score can identify multiple maximum substructures of the model structure that can be superimposed over the real structure cutting at different distance thresholds (e.g. 1, 2, 4 and 8 Å as represented in Equation 1). It was used by all three assessors (comparative modelling, fold recognition and new fold prediction) in both CASP5 [39] and CASP6 [38]. Another important method for evaluating model structures is MaxSub [37], which is based on similar principles as GDT (Global Distance Test). MaxSub computes a normalized score ranging from 0 (for a completely wrong model) to 1 (for a perfect model) and thus it is especially suitable for evaluating the model structures predicted by fully automatic servers. MaxSub has been applied in CAFASP2 [97] and CAFASP3 [98] successfully.

Both GDT_TS and MaxSub focus only on the size of the substructure, while the spatial information of the templates outside of the substructure is partially neglected. In addition, the scores of GDT_TS and MaxSub are power-law dependent on the size of the protein [99]. Zhang and Skolnick [99] developed a new scoring function, TM-score (template modelling score), which overcomes the above two limitations of GDT_TS and MaxSub by introducing a protein size-dependent scale to eliminate the inherent protein size dependency, and by evaluating all residue pairs in alignment/modelling in the proposed score. TM-score shows a significant outperformance to both GDT_TS and MaxSub when benchmarked in CASP5 targets. Based on TM-score, Zhang and Skolnick have also developed a protein structural alignment program, TM-align [100], which shows advantage in both speed and accuracy compared to conventional structural alignment methods, e.g. CE [8] and DALI [101].

Sims and Kom [102] compared the HOPPscore with GDT_TS score by analyzing CASP6 models. Although no significant correlation exists between HOPPscore and GDT_TS score, they found that in general, model structures with low GDT_TS scores had also large differences in HOPPscores while models with fairly good GDT_TS scores had good HOPPscores as well. Actually, the vectors used in Sims and Kom [102] to denote fragments of model structures are quite similar to shape strings. They are both 1D strings of discrete φ/ψ angle pairs, while the former are rounded down to a certain bin size and the latter are mapped to highly clustered regions representing the natural conformation constrains in the Ramachandran plot. It means that shape strings might be more reliable in representing the natural conformation space of protein fragments and a useful tool for evaluating the quality of model structures. The program Frags by Ison *et al.* [25] supplies a function for returning the frequency of occurrence for any shape string fragment in the library, but no implementation has been done yet with respect to evaluating the quality of entire model structures.

4. CONCLUSIONS

The increasing number of structures in PDB gives us great opportunities to dig out information about the structures of proteins. On the other hand, it requires efficient and accurate methods for describing and comparing protein structures. The secondary structure expression of protein structures reduces the complexity in describing protein structure enormously as opposed to listing xyz coordinates of all atoms in proteins. The widely used protein structure classification databases such as SCOP and CATH are all based on secondary structure elements. As to comparing the protein structures, all atom rigid body superposition is the most useful methods for quantifying detailed structural differences of proteins. However, it is computationally demanding and cannot align distantly related protein structures with many indels. Structure comparison methods based on alignment of secondary structure elements have been introduced to handle these limitations.

Given that secondary structure description does not distinguish the conformations in loop regions (which account for on average nearly half of all amino acids), comparisons based on secondary structure elements might not classify global similarity correctly. Methods based on one-dimensional geometrical representation of protein backbone structures are becoming important. These one-dimensional representations, especially shape strings, encode loop regions as well as secondary structure elements in a rather accurate way. In this review, we have shown some advantages of shape strings in structure comparison and homology identification. However, the current applications of shape strings are rather preliminary. For structure comparison, more accurate alignment that best make use of the properties of shape strings as well as the analysis of the statistical significance of shape string alignment are required. Moreover, prediction of shape strings instead of the secondary structures of proteins might be an alternative way to start 3D structure prediction. Both the prediction of shape strings and the building of 3D structures from shape strings need further research.

Shape strings facilitate fast database searching for similar structures, classification of loop regions and evaluation of model structures. We can expect more widely use of such methods in the near future.

ACKNOWLEDGEMENT

We thank Dr. Iddo Friedberg for supplying the benchmark results for FATCAT, CE, BLAST and KL-string alignment. Calidris, Sollentuna is thanked for financial support.

REFERENCES

[1] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. **(2000)** *Nucleic Acids Res., 28*, 235-242.

[2] Rost, B. **(1997)** *Fold. Des., 2*, S19-24.

[3] Sander, C. and Schneider, R. **(1991)** *Proteins, 9*, 56-68.

[4] Holm, L. and Sander, C. **(1998)** *Nucleic Acids Res., 26*, 316-319.

[5] Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia, C. **(1995)** *J. Mol. Biol., 247*, 536-540.

[6] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. and Thornton, J. M. **(1997)** *Structure, 5*, 1093-1108.

[7] Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T. and Tramontano, A. **(2007)** *Proteins, 69*, 3-9.

[8] Shindyalov, I. N. and Bourne, P. E. **(1998)** *Protein Eng., 11*, 739-747.

[9] DeLano, W. L. **(2002)** *The PyMOL Molecular Graphics System* on World Wide Web http://www.pymol.org.

[10] Kawabata, T. and Nishikawa, K. **(2000)** *Proteins, 41*, 108-122.

[11] Madej, T., Gibrat, J. F. and Bryant, S. H. **(1995)** *Proteins, 23*, 356-369.

[12] Krissinel, E. and Henrick, K. **(2004)** *Acta Crystallogr. D. Biol. Crystallogr., 60*, 2256-2268.

[13] Harrison, A., Pearl, F., Sillitoe, I., Slidel, T., Mott, R., Thornton, J. and Orengo, C. **(2003)** *Bioinformatics, 19*, 1748-1759.

[14] Lu, G. G. **(2000)** *J. Appl. Crystall., 33*, 176-183.

[15] Vesterstrom, J. and Taylor, W. R. **(2006)** *J. Comput. Biol., 13*, 43-63.

[16] Orengo, C. A., Brown, N. P. and Taylor, W. R. **(1992)** *Proteins, 14*, 139-167.

[17] Yang, A. S. and Honig, B. **(2000)** *J. Mol. Biol., 301*, 665-678.

[18] Kabsch, W. and Sander, C. **(1983)** *Biopoly., 22*, 2577-2637.

[19] Gibrat, J. F., Madej, T. and Bryant, S. H. **(1996)** *Curr. Opin. Struct. Biol., 6*, 377-385.

[20] Carugo, O. and Pongor, S. **(2002)** *Curr. Protein. Pept. Sci., 3*, 441-449.

[21] Carugo, O. **(2006)** *Curr. Bioinform., 1*, 75-83.

[22] Carugo, O. **(2007)** *Curr. Protein. Pept. Sci., 8*, 219-241.

[23] Gong, H., Fleming, P. J. and Rose, G. D. **(2005)** *Proc. Natl. Acad. Sci. USA, 102*, 16227-16232.

[24] Zhi, D., Krishna, S. S., Cao, H., Pevzner, P. and Godzik, A. **(2006)** *BMC. Bioinformatics, 7*, 460.

[25] Ison, R. E., Hovmöller, S. and Kretsinger, R. H. **(2005)** *IEEE Eng. Med. Biol. Mag., 24*, 41-49.

[26] Ramachandran, G. N. and Sasisekharan, V. **(1968)** *Adv. Protein Chem., 23*, 283-438.

[27] Chakrabarti, P. and Pal, D. **(2001)** *Prog. Biophys. Mol. Biol., 76*, 1-102.

[28] Hovmöller, S., Zhou, T. and Ohlson, T. **(2002)** *Acta Crystallogr. D. Biol. Crystallogr., 58*, 768-776.

[29] Kleywegt, G. J. and Jones, T. A. **(1996)** *Structure, 4*, 1395-1400.

[30] Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. and Richardson, D. C. **(2003)** *Proteins, 50*, 437-450.

[31] Sam, V., Tai, C. H., Garnier, J., Gibrat, J. F., Lee, B. and Munson, P. J. **(2006)** *BMC Bioinformatics, 7*, 206.

[32] Gerstein, M. and Levitt, M. **(1998)** *Protein Sci., 7*, 445-456.

[33] Brenner, S. E., Chothia, C. and Hubbard, T. J. **(1998)** *Proc. Natl. Acad. Sci. USA, 95*, 6073-6078.

[34] Leplae, R. and Hubbard, T. J. **(2002)** *Bioinformatics, 18*, 494-495.

[35] Rao, S. T. and Rossmann, M. G. **(1973)** *J. Mol. Biol., 76*, 241-256.

[36] Zemla, A. **(2003)** *Nucleic Acids Res., 31*, 3370-3374.

[37] Siew, N., Elofsson, A., Rychlewski, L. and Fischer, D. **(2000)** *Bioinformatics, 16*, 776-785.

[38] Moult, J., Fidelis, K., Rost, B., Hubbard, T. and Tramontano, A. **(2005)** *Proteins, 61*. (Suppl. 7), 3-7.

[39] Moult, J., Fidelis, K., Zemla, A. and Hubbard, T. **(2003)** *Proteins, 53*. (Suppl. 6), 334-339.

[40] Richardson, J. S. **(1981)** *Adv. Protein. Chem., 34*, 167-339.

[41] Sayle, R. A. and Milner-White, E. J. **(1995)** *Trends Biochem. Sci., 20*, 374.

[42] Pauling, L., Corey, R. B. and Branson, H. R. **(1951)** *Proc. Natl. Acad. Sci. USA, 37*, 205-211.

[43] Ho, B. K., Thomas, A. and Brasseur, R. **(2003)** *Protein Sci., 12*, 2508-2522.

[44] Wang, G. and Dunbrack, R. L., Jr. **(2003)** *Bioinformatics, 19*, 1589-1591.

[45] Kwasigroch, J. M., Chomilier, J. and Mornon, J. P. **(1996)** *J. Mol. Biol., 259*, 855-872.

[46] Oliva, B., Bates, P. A., Querol, E., Aviles, F. X. and Sternberg, M. J. **(1997)** *J. Mol. Biol., 266*, 814-830.

[47] Panchenko, A. R. and Madej, T. **(2005)** *BMC Evol. Biol., 5*, 10.

[48] Parker, M. H. and Hefford, M. A. **(1997)** *Protein Eng., 10*, 487-496.

[49] Funhoff, E. G., Ljusberg, J., Wang, Y., Andersson, G. and Averill, B. A. **(2001)** *Biochemistry, 40*, 11614-11622.

[50] Park, Y. Y., Kim, H. J., Kim, J. Y., Kim, M. Y., Song, K. H., Cheol Park, K., Yu, K. Y., Shong, M., Kim, K. H. and Choi, H. S. **(2004)** *Mol. Endocrinol., 18*, 1082-1095.

[51] Fiser, A., Feig, M., Brooks, C. L., 3rd and Sali, A. **(2002)** *Acc. Chem. Res., 35*, 413-421.

[52] Fiser, A., Do, R. K. and Sali, A. **(2000)** *Protein Sci., 9*, 1753-1773.

[53] Milner-White, E. J. and Poet, R. **(1986)** *Biochem. J., 240*, 289-292.

[54] Espadaler, J., Fernandez-Fuentes, N., Hermoso, A., Querol, E., Aviles, F. X., Sternberg, M. J. and Oliva, B. **(2004)** *Nucleic Acids Res., 32*, D185-188.

[55] Bystroff, C., Thorsson, V. and Baker, D. **(2000)** *J. Mol. Biol., 301*, 173-190.

[56] Fitzkee, N. C., Fleming, P. J. and Rose, G. D. **(2005)** *Proteins, 58*, 852-854.

[57] Srinivasan, R. and Rose, G. D. **(1999)** *Proc. Natl. Acad. Sci. USA, 96*, 14258-14263.

[58] Kolodny, R., Koehl, P., Guibas, L. and Levitt, M. **(2002)** *J. Mol. Biol., 323*, 297-307.

[59] Chou, J. J., Li, S. and Bax, A. **(2000)** *J. Biomol. NMR, 18*, 217-227.

[60] Kuang, R., Leslie, C. S. and Yang, A. S. **(2004)** *Bioinformatics, 20*, 1612-1621.

[61] Friedberg, I., Harder, T., Kolodny, R., Sitbon, E., Li, Z. and Godzik, A. **(2007)** *Bioinformatics, 23*, e219-224.

[62] Kolodny, R. and Levitt, M. **(2003)** *Biopolymers, 68*, 278-285.

[63] Urbanc, B., Borreguero, J. M., Cruz, L. and Stanley, H. E. **(2006)** *Methods Enzymol., 412*, 314-338.

[64] Krebs, W. G. and Gerstein, M. **(2000)** *Nucleic Acids Res., 28*, 1665-1675.

[65] Ming, D., Kong, Y., Lambert, M. A., Huang, Z. and Ma, J. **(2002)** *Proc. Natl. Acad. Sci. USA, 99*, 8620-8625.

[66] Koehl, P. **(2001)** *Curr. Opin. Struct. Biol., 11*, 348-353.

[67] Feng, Z. K. and Sippl, M. J. **(1996)** *Fold. Des., 1*, 123-132.

[68] Kolodny, R., Koehl, P. and Levitt, M. **(2005)** *J. Mol. Biol., 346*, 1173-1188.

[69] Holm, L. and Sander, C. **(1994)** *Proteins, 19*, 165-173.

[70] Anderson, A. C. **(2005)** *Acta Crystallograph Sect. F. Struct. Biol. Cryst. Commun., 61*, 258-262.

[71] Powers, R., Copeland, J. C., Germer, K., Mercier, K. A., Ramanathan, V. and Revesz, P. **(2006)** *Proteins, 65*, 124-135.

[72] Maiorov, V. N. and Crippen, G. M. **(1994)** *J. Mol. Biol., 235*, 625-634.

[73] Cohen, F. E. and Sternberg, M. J. **(1980)** *J. Mol. Biol., 138*, 321-333.

[74] Maiorov, V. N. and Crippen, G. M. **(1995)** *Proteins, 22*, 273-283.

[75] Carugo, O. **(2003)** *J. Appl. Crystal., 36*, 125-128.

[76] Carugo, O. and Pongor, S. **(2001)** *Protein Sci., 10*, 1470-1473.

[77] Carugo, O. **(2007)** *Protein Eng. Des. Sel., 20*, 33-37.

[78] Wriggers, W. and Schulten, K. **(1997)** *Proteins, 29*, 1-14.

[79] Verbitsky, G., Nussinov, R. and Wolfson, H. **(1999)** *Proteins, 34*, 232-254.

[80] Shatsky, M., Nussinov, R. and Wolfson, H. J. **(2002)** *Proteins, 48*, 242-256.

[81] Ye, Y. and Godzik, A. **(2003)** *Bioinformatics, 19.* (Suppl. 2), II246-II255.

[82] Andersen, C. A. and Rost, B. **(2003)** *Methods Biochem. Anal., 44*, 341-363.

[83] Frishman, D. and Argos, P. **(1995)** *Proteins, 23*, 566-579.

[84] Richards, F. M. and Kundrot, C. E. **(1988)** *Proteins, 3*, 71-84.

[85] Sklenar, H., Etchebest, C. and Lavery, R. **(1989)** *Proteins, 6*, 46-60.

[86] Bornot, A. and de Brevern, A. G. **(2006)** *Bioinformation, 1*, 153-155.

[87] Gong, H. and Rose, G. D. **(2005)** *Proteins, 61*, 338-343.

[88] Taylor, D. A., Sack, J. S., Maune, J. F., Beckingham, K. and Quiocho, F. A. **(1991)** *J. Biol. Chem., 266*, 21375-21380.

[89] Clapperton, J. A., Martin, S. R., Smerdon, S. J., Gamblin, S. J. and Bayley, P. M. **(2002)** *Biochemistry, 41*, 14669-14679.

[90] Needleman, S. B. and Wunsch, C. D. **(1970)** *J. Mol. Biol., 48*, 443-453.

[91] Hall, D. R., Bond, C. S., Leonard, G. A., Watt, C. I., Berry, A. and Hunter, W. N. **(2002)** *J. Biol. Chem., 277*, 22018-22024.

[92] Alphey, M. S., Yu, W., Byres, E., Li, D. and Hunter, W. N. **(2005)** *J. Biol. Chem., 280*, 3068-3077.

[93] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. **(1990)** *J. Mol. Biol., 215*, 403-410.

[94] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. **(1997)** *Nucleic Acids Res., 25*, 3389-3402.

[95] Altschul, S. F. and Gish, W. **(1996)** *Methods Enzymol., 266*, 460-480.

[96] Zemla, A., Venclovas, C., Moult, J. and Fidelis, K. **(1999)** *Proteins,* (Suppl. 3), 22-29.

[97] Fischer, D., Elofsson, A., Rychlewski, L., Pazos, F., Valencia, A., Rost, B., Ortiz, A. R. and Dunbrack, R. L., Jr. **(2001)** *Proteins,* (Suppl. 5), 171-183.

[98] Fischer, D., Rychlewski, L., Dunbrack, R. L., Jr., Ortiz, A. R. and Elofsson, A. **(2003)** *Proteins, 53.* (Suppl. 6), 503-516.

[99] Zhang, Y. and Skolnick, J. **(2004)** *Proteins, 57*, 702-710.

[100] Zhang, Y. and Skolnick, J. **(2005)** *Nucleic Acids Res., 33*, 2302-2309.

[101] Holm, L. and Sander, C. **(1993)** *J. Mol. Biol., 233*, 123-138.

[102] Sims, G. E. and Kim, S. H. **(2006)** *Proc. Natl. Acad. Sci. USA, 103*, 4428-4432.

Received: April 09, 2007 Revised: January 07, 2008 Accepted: February 19, 2008