BIOINFORMATICS

Vol. 19 Suppl. 2 2003, pages ii246–ii255
DOI: 10.1093/bioinformatics/btg1086

Flexible structure alignment by chaining aligned fragment pairs allowing twists

*Yuzhen Ye and Adam Godzik**

Program in Bioinformatics and Systems Biology, The Burnham Institute, La Jolla, CA 92037, USA

Received on March 17, 2003; accepted on June 9, 2003

ABSTRACT

Motivation: Protein structures are flexible and undergo structural rearrangements as part of their function, and yet most existing protein structure comparison methods treat them as rigid bodies, which may lead to incorrect alignment.

Results: We have developed the Flexible structure AlignmenT by Chaining AFPs (Aligned Fragment Pairs) with Twists (FATCAT), a new method for structural alignment of proteins. The FATCAT approach simultaneously addresses the two major goals of flexible structure alignment; optimizing the alignment and minimizing the number of rigid-body movements (twists) around pivot points (hinges) introduced in the reference protein. In contrast, currently existing flexible structure alignment programs treat the hinge detection as a post-process of a standard rigid body alignment. We illustrate the advantages of the FATCAT approach by several examples of comparison between proteins known to adopt different conformations, where the FATCAT algorithm achieves more accurate structure alignments than current methods, while at the same time introducing fewer hinges.

Contacts: adam@burnham.org

INTRODUCTION

Protein structure comparison has been a classic challenge in computational molecular biology for more than two decades. Many programs addressing this challenge have been developed (Holm and Sander, 1993; Boutonnet *et al.*, 1995; Madej *et al.*, 1995; Shindyalov and Bourne, 1998; Eidhammer *et al.*, 2001), but they all share similar limitations stemming from treating proteins as rigid bodies. We know, however, that proteins are flexible molecules that undergo significant structural changes as part of their normal function (Wuthrich and Wagner, 1978; Schulz and Schirmer, 1979; Bennett and Huber, 1984; Jacobs *et al.*, 2001). When flexible molecules in different conformations are compared to each other as rigid bodies, even strong structural similarities can be missed and significant errors in alignments can occur because such algorithm compensate global rearrangements with local alignment shifts.

Figure 1 illustrates the problem using an example of homologous spectrin repeats from *Drosophila sp.* (PDB code 2spc, chain A) and from *Gallus gallus* (PDB code 1aj3). 1aj3 has three helices (denoted as helix I, II and III) with similar length (Fig. 1b), and 2spcA has two helices (I and II) with helix II being much longer. The structural comparison of these two (or any other) pair can be visualized as a dot matrix of AFPs (Vriend and Sander, 1991; Shindyalov and Bourne, 1998), each AFP being represented by a diagonal line (Fig. 1a). Any structural alignment can be viewed as a chain of non-overlapping AFPs. Structural comparison programs such as CE (Shindyalov and Bourne, 1998) or DALI (Holm and Sander, 1993), can not find an alignment between the above two proteins that spans their entire lengths; instead, they find local alignments: CE aligns helices II and III in 1aj3 with helix I and the N-terminal half of helix II in 2spcA ((DALI aligns helices I and II with the same two helices in 2spcA) (Fig. 1b). Both alignments show significant structural similarity between the two proteins (CE's alignment has 56 aligned positions with RMSD 2.02 Åand DALI's alignment has 67 aligned positions with RMSD 3.24 Å), but they don't capture the actual homology that exists in strong sequence similarity along the entire protein (Fig. 1c). The analysis based on both structures and sequences shows that there are two structural rearrangements (one major and one minor) between the two proteins, and the major one causes their overall structural dissimilarity (three-helix bundle versus two-helix structure). Such large structural rearrangements make 'rigid-body'structure alignments unable to detect the real structural similarity even in homologous proteins.

To address these issues we developed a flexible protein structure alignment algorithm (FATCAT), which naturally incorporates conformational flexibility in structure comparison. The problem is formulated as follows: given two protein structures A and B, find the optimal structure alignment between them with the least number

*To whom correspondence should be addressed.

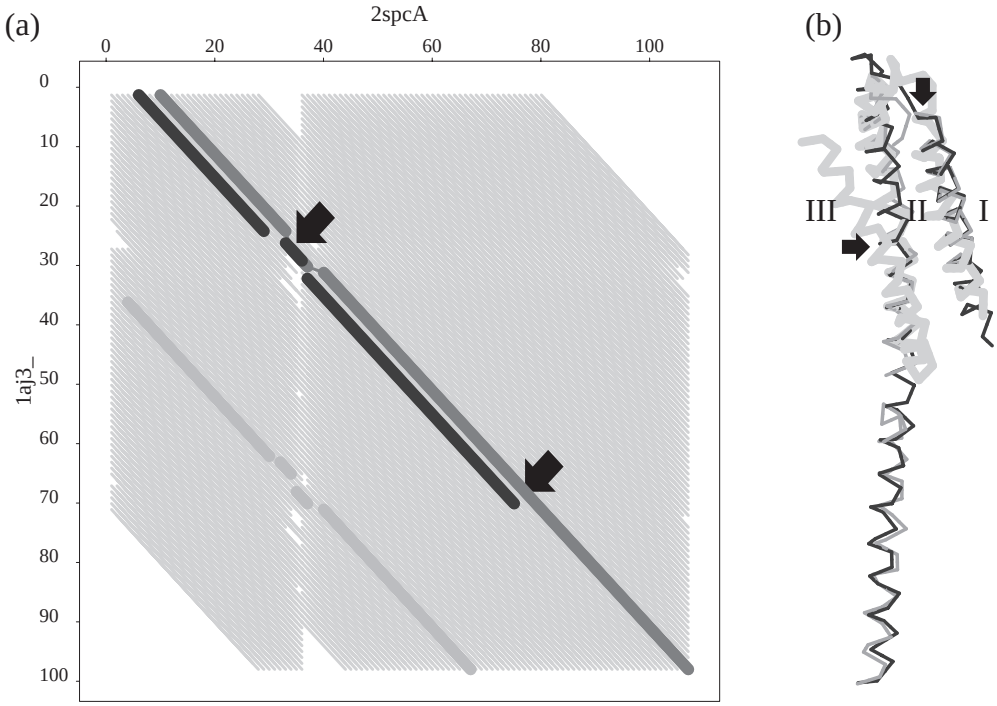

Fig. 1. Comparison between 2spcA and 1aj3. (a) AFP chains derived by FATCAT (with two twists pointed by arrows), DALI and CE are shown as the dark gray, black, and light gray bold lines in the dot matrix of AFPs, respectively. (b) The superposition of 2spcA (black lines) and 1aj3. 1aj3, superimposed onto 2spcA according to DALI alignment, is shown in gray bold lines, and the twisted 1aj3, superimposed onto 2spcA according to FATCAT alignment, is shown in gray thin lines. (c) The sequence alignment between the two proteins by BLAST.

of rearrangements (twists) in one of the structures. Several implicit approaches to solve this problem have been described in the literature. For example, Wriggers and Schulten (1997) partition a protein into rigid domains that are extracted by an adaptive selection procedure using least-squares fitting, followed by characterization of relative movements of the domains as happening at specific hinges. Boutonnet *et al.* (1995); Ochagavia *et al.* (2002) solve the problem using a multiple linkage clustering algorithm to identify segment combinations which yields optimal global structure alignments. The clustering trees are further analyzed to detect the rigid-body movements between structure elements. In a more recent work, Shatsky *et al.* (2002) search for the largest set of congruent AFPs, and then look for a subset that describes a possible alignment of two structures with flexibility by clustering consecutive AFPs. These approaches search for rotation points (hinges) by analysis of the initial rigid body alignments. However, the initial alignments of two structures that differ by a subdomain movement is often so wrong that it precludes finding the correct hinges by post-alignment analysis.

In this paper, we propose a dynamic programming algorithm to connect Aligned Fragment Pairs (AFPs) by combining gaps and twists between consecutive AFPs, each with its own score penalty. Therefore, the minimization algorithm compares solutions involving twists and simple extensions and in this way it performs the alignment and hinge detection simultaneously. The entire algorithm has been implemented in a fast and efficient computer program FATCAT and systematically tested on a large alignment benchmark.

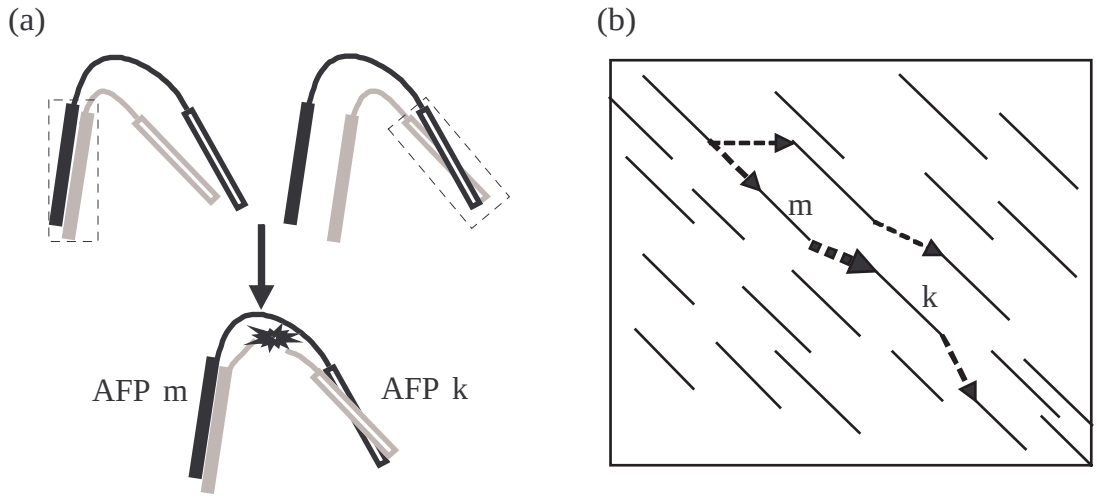

Fig. 2. Structure alignment by AFPs chaining. (a) A twist is introduced in one structure to connect AFP *m* and *k*. (b) Dot matrix of AFPs. Each AFP is shown as a line. A chain linking AFPs corresponds to an alignment between two structures (for clarity, only two chains are shown in the graph).

METHODS

In this section, we provide a detailed description of the new algorithm and its implementation.

Definitions

Given two protein structures, denote a match of two fragments, one from each protein, as an Aligned Fragment Pair (AFP), the starting positions of an AFP k in the two proteins as $b^1(k)$ and $b^2(k)$, and its ending positions in the two proteins as $e^1(k)$ and $e^2(k)$, respectively. Each AFP describes one way of superimposing one protein on the other. We say that two consecutive AFPs are compatible if they result in the same (or very similar) superposition of the proteins; otherwise, they are not compatible unless one structure is modified. If two AFPs are compatible, they can be simply connected to each other, if they are not, they can still be connected if a twist is introduced in the connection of two AFPs (Fig. 2a), and there is a hinge in the structures at the twist position.

We focus on the consecutive or sequence-dependent structural comparison, where fragments conserve their relative position along the sequence; thus a structure alignment is viewed as a chain of AFPs (Fig. 2b). A rigid structure alignment is a chain of compatible AFPs connected to each other. Dynamic programming can be used to identify the optimal set of connected AFPs. In contrast, a flexible structure alignment is a chain of AFPs in which some connections between AFPs are achieved by introducing twists. We define a block of AFPs as a set of consecutive compatible AFPs; in other words, n twists in an AFP chain divide the chain into $n + 1$ blocks. We say that there is a structural distortion between the two structures when one structure has to be twisted around the hinges to be aligned by a rigid body superposition. Overall RMSD is the structural similarity of two such structures, defined as the root mean square deviation of all of their aligned C_α atoms based on the rigid body superposition after one structure is modified.

AFP detection

Two fragments of fixed size L (e.g. 8) form an AFP if the RMSD (root mean square deviation after optimal superposition) of their C_α atoms is less than a certain threshold (C_t, e.g. 3.0 Å) (Shindyalov and Bourne, 1998).

The test for AFPs compatibility

The compatibility of consecutive AFPs pairs is measured by the root mean square deviation between distance matrices of residues in the fragments from each protein forming connected AFPs. It is denoted as D_{mk} for AFP m and k (see Equation 1). The high similarity between two distance matrices means that these two AFPs are compatible; otherwise a twist has to be introduced for connecting the AFPs pair, as shown in Fig. 2a.

$$D_{mk} = \sqrt{\sum_{s=1}^{L} (d^1_{b^1(m)+s,b^1(k)+s} - d^2_{b^2(m)+s,b^2(k)+s})^2} \quad (1)$$

where $d^1_{i,j}$, $d^2_{i,j}$ is the distance between residue i and j in protein 1 and protein 2, respectively, $b^1(m)$, $b^1(k)$, $b^2(m)$

and $b^2(k)$ are the starting positions of AFP m and k, in proteins 1 and 2 respectively, as defined earlier, and L is the length of each AFP.

Flexible structure alignment

Author: Gusfield 1999 not in reference list

Flexible structure alignment can be formulated as the AFPs chaining process (Gusfield 1999) allowing at most t twists, and the flexible structure alignment is transformed into a rigid structure alignment when t is 0. Dynamic programming is used to find the optimal chaining. If we denote $S(k)$ as the best score ending at AFP k, it can be calculated from the best ending at previous AFPs that can be connected with AFP k subject to the following constraints (Fig. 2b),

$$S(k) = a(k) + \max\left\{ \max_{\substack{e^1(m) < b^1(k) \\ e^2(m) < b^2(k)}} \left[(S(m) + c(m \rightarrow k)\right], 0 \right\} \quad s.t. T(k) \leq t \qquad (2)$$

where $a(k)$ is the score of AFP k itself; $c(m \rightarrow k)$ is the score of introducing a connection between AFP m and AFP k; $T(k)$ is the number of twists required for connecting the chain of AFPs leading up to $S(k)$, which is calculated by,

$$T(k) = T(m) + t(m \rightarrow k) \qquad (3)$$

where $t(m \rightarrow k)$ is 1 if a twist is required to connect AFP m and k and 0 if no twist is required.

The score of an AFP k is determined by its RMSD (d_k) and length (L); long AFPs are rewarded and large RMSDs are penalized,

$$a(k) = R_s \times L \times F(d_k) \qquad (4)$$

where R_s is the rewarding score associated with a good aligned position and $F(d_k)$ is the function of d_k.

The score for connecting AFP m and k is the function of the compatibility of the AFPs and the mis-matched regions (p) and/or gaps (q) created by the connection of the two AFPs,

$$c(m \rightarrow k) = W(D_{mk}) \times P_c + F(p, q) \qquad (5)$$

$$W(D_{mk}) = \begin{cases} 1 & \text{if } D_{mk} > D_c \\ \left(\frac{D_{mk} - D_0}{D_c - D_0}\right)^2 & \text{elsif } D_0 < D_{mk} \leq D_c \\ 0 & \text{else} \end{cases} \qquad (6)$$

$$F(p, q) = M_c \times p + M_s \times q \qquad (7)$$

where D_{mk} is the root mean square of the distance matrix between AFP m and k, as defined above; D_c is the threshold for defining a twist; D_0 is the threshold for penalizing a connection; P_c is the maximum penalty for connecting two AFPs; M_c is the penalty involved with mis-matching two positions; M_g is the penalty for a gap.

Post-processing of AFP chaining

Several post processing steps are applied after deriving the best AFP chain defined by the scoring system presented above. Additional twists are introduced into the AFP chains if its overall RMSD is larger than a fixed threshold. Unnecessary twists that do not lower the overall RMSD are removed. Finally, we apply iterative refinement of structure alignments by dynamic programming performed on the distance matrix calculated from the two superimposed structures as described in previous studies (Feng and Sippl, 1996; Shindyalov and Bourne, 1998; Lackner *et al.*, 2000).

RESULTS

We implemented the FATCAT approach in C++ on a Linux platform. The running time of FATCAT comparing a pair of protein structures on a 1.8GHz Pentium varies from seconds to a few minutes, depending on the number of AFPs the two structures have. For instance, 42 060 AFPs were detected in comparing protein 1fmk (with 438 residues) and 1tki (with 321 residues) (alignment result is shown below), and the whole process took 76 seconds.

We first applied FATCAT to several alignments described as 'difficult'in the literature (Fischer *et al.*, 1996) and compared its performance with three rigid alignment programs, DALI (Holm and Sander, 1993), VAST (Madej *et al.*, 1995) and CE (Shindyalov and Bourne, 1998). We then compared FATCAT's performance with the results of the only other readily available flexible alignment program, FlexProt (Shatsky *et al.*, 2002). Finally, to obtain a broader overview we applied FATCAT to a large set of similar structures extracted from the non-redundant SCOP database (proteins are clustered at 40% sequential identity) (Murzin *et al.*, 1995). To avoid bias from large families, we retained only one pair per family, leading to 6437 pairs of structurally similar proteins, including 854 family-level protein pairs, 3200 superfamily-level protein pairs (one representative structure per family), and 2383 fold-level protein pairs (one representative structure per superfamily). The same parameters ($t = 5$; $L = 8$; $C_t = 3.0$; $D_c = 5.0$; $D_0 = 1.0$; $R_s = 3.0$; $P_c = -25$; $M_s = -0.5$ and $M_g = -0.5$) were used in all the calculations.

Comparison with rigid structure alignment programs

FATCAT works well in aligning distantly similar protein structures, comparable to the performances of the rigid structure alignment programs, DALI, VAST and CE. In the test of 10 'difficult'examples (Fischer *et al.*, 1996), FATCAT produced good alignments (no twists needed, similar alignment length and similar RMSD) in 8 out of 10 examples, except that 2 and 5 twists are introduced in

Table 1. Comparison of structure alignments of 10 'difficult' pairs of structures from (Fischer *et al.*, 1996) by different methods

Pro1	Pro2	VAST		DALI		CE		FATCAT		
		Size	RMSD	Size	RMSD	Size	RMSD	Size	RMSD	Twist
1fxiA	1ubq_	48	2.1	-	-	-	-	63	3.01	0
1ten_	3hhrB	78	1.6	86	1.9	87	1.9	87	1.9	0
3hlaB	2rhe_	-	-	63	2.5	85	3.5	79	2.81	2
2azaA	1paz_	74	2.2	-	-	85	2.9	87	3.01	0
1cewI	1molA	71	1.9	81	2.3	69	1.9	83	2.44	0
1cid_	2rhe_	85	2.2	95	3.3	94	2.7	100	3.11	0
1crl_	1ede_	-	-	211	3.4	187	3.2	269	3.55	5
2sim_	1nsbA	284	3.8	286	3.8	264	3.0	286	3.07	0
1bgeB	2gmfA	74	2.5	98	3.5	94	4.1	100	3.19	0
1tie_	4fgf_	82	1.7	108	2.0	116	2.9	117	3.05	0

The data for VAST, DALI and CE are from (Shindyalov and Bourne, 1998). Descriptions for the items are: Size, the number of aligned positions; Twist, the number of twists introduced in FATCAT. The RMSD value in FATCAT is the overall RMSD.

comparing (3hlaB, 2rhe_) and (1crl_, 1ede_), respectively (Table 1). In both cases, however, the FATCAT alignment is arguably better, with either a lower RMSD or a longer alignment. This result shows that FATCAT is not specifically biased to detect hinges.

FATCAT obviously outperforms rigid structure alignment programs with respect to its capability to detect hinges in protein structures. For example, in the comparison between 2spcA and 1aj3 discussed in the Introduction section, FATCAT identified a structure alignment spanning the entire length of both proteins by introducing two twists (Fig. 1), a result which is consistent with their evolutionary relationship. In contrast, the rigid structure alignment programs, such as CE and DALI were only able to identify short local alignments, either stopping around the hinge position (DALI) or aligning non-homologous regions (CE).

Comparison with FlexProt

As mentioned earlier, the main features of FATCAT are its ability to optimize the structure alignment and introduce the fewest number of twists at the same time. Its advantage over the FlexProt (Shatsky *et al.*, 2002) is demonstrated by the examples listed in Table 2. Overall, FATCAT alignments have a smaller number of twists but similar RMSDs and lengths as FlexProt alignments, suggesting that the strategy of separating the hinge detection and the chaining process introduces unnecessary twists into the alignments. For instance, FATCAT created an alignment of 238 aligned positions with overall RMSD of 3.08 Åbetween the human tyrosine-protein kinase C-SRC (PDB code 1fmk) and the titian protein (PDB code 1tki), whereas FlexProt was forced to introduce two hinges to get an even shorter alignment (231 aligned positions) and a higher RMSD (3.28 Å).

In the second example, the tissue factor (PDB code 1a21, chain A) was compared to the growth hormone-binding protein (PDB code 1hwg, chain C). Four hinges are detected by FlexProt which results in a structure alignment of 163 aligned positions with an RMSD of 2.75 Åbetween these two proteins (Shatsky *et al.*, 2002). On the other hand, FATCAT created a slightly shorter structure alignment of 153 aligned positions with an RMSD of 3.16 Å, but introduced only one twist into the alignment. Although the difference between the aligned positions in the two alignments is modest, its influence on the overall alignment is significant (4 hinges in FlexProt versus 1 hinge in FATCAT).

The third example is the comparison between the histocompatibility antigen (PDB code 2clr, chain A) and the neonatal FC receptor (PDB code 3fru, chain A) (Fig. 3). FlexProt alignment has 253 aligned positions with an RMSD of 2.71 Åand it introduced two hinges. In contrast, FATCAT gave a structure alignment with similar quality (245 aligned positions with an RMSD of 3.06 Å) without introducing any twists, see Figure 3.

Overall structural distortions in similar structures

We applied FATCAT to the 6437 pairs of structurally similar proteins on SCOP fold, superfamily and family levels, the results are summarized in Table 3. Overall, structural distortions are significant between many pairs: twists are detected in aligning about half of the protein pairs, and more twists are found in aligning protein pairs on the fold- and superfamily-level than in the protein pairs on the family-level. This was expected because functions of proteins change among different families or superfamilies and they are often accompanied by changes in the structures.

Except for those rare cases where structure distortions are caused by the structure determination process itself, the distortions can be grouped into two types. The

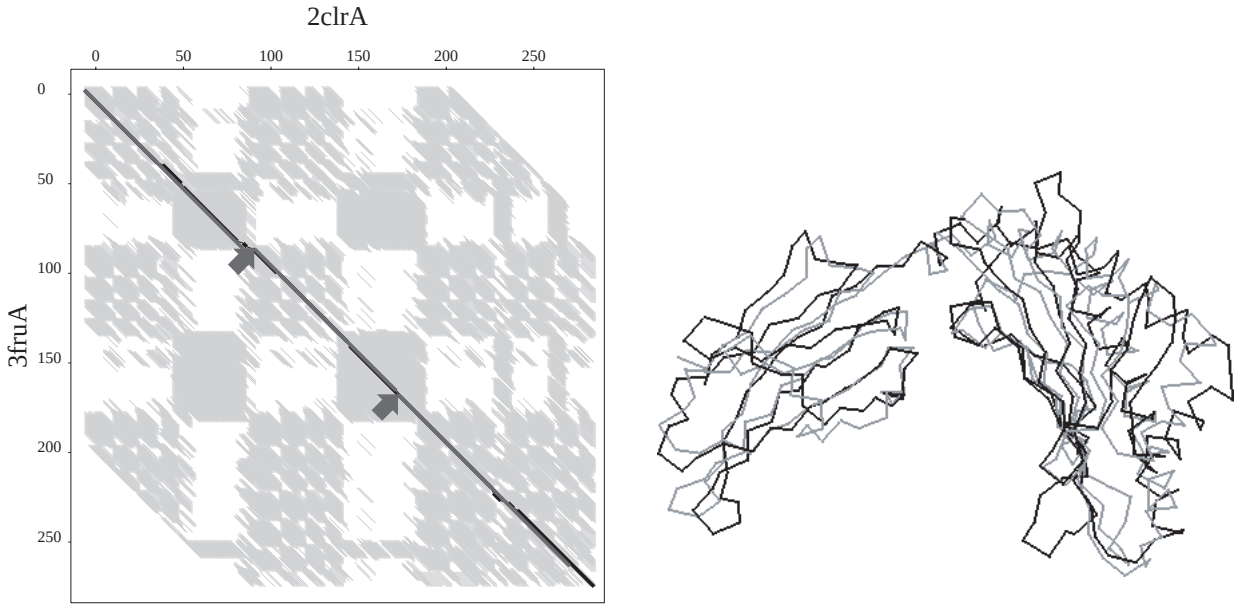

Fig. 3. The left graph shows the comparison between the histocompatibility antigen (2clr, chain A) and the neonatal FC receptor (3fru, chain A), in which FlexProt detected 2 hinges (shown by the arrows) but no twists are introduced by FATCAT. The right graph shows the superposition of the two proteins according to the FATCAT alignment, in which the histocompatibility antigen is shown in black lines and the neonatal FC receptor is shown in gray lines.

Table 2. Comparison of FlexProt and FATCAT

		FlexProt			FATCAT		
Pro1	Pro2	Size	RMSD	Twist	Size	RMSD	Twist
1wdnA	1gggA	218	0.94	2	220	1.01	2
1hpbP	1gggA	220	2.34	2	213	1.59	2
2bbmA	1cll_	139	2.22	1	144	2.28	1
2bbmA	1top_	147	2.40	3	145	2.28	3
1akeA	2ak3A	200	2.44	2	202	1.54	2
2ak3A	1uke_	182	2.90	2	188	2.97	0
1mcpL	4fabL	218	1.93	1	217	1.40	1
1mcpL	1tcrB	212	2.33	1	213	2.20	1
1lfh_	1lfg_	691	1.41	2	686	0.89	2
1tfd_	1lfh_	291	1.98	2	290	1.37	2
1b9wA	1danL	75	2.78	1	80	2.39	2
1qf6A	1adjA	323	4.43	1	351	2.68	1
2clrA	3fruA	253	2.71	2	245	3.06	0
1fmk_	1qcfA	424	1.25	2	433	2.27	0
1fmk_	1tkiA	231	3.28	2	238	3.07	0
1a21A	1hwgC	163	2.75	4	153	3.16	1

The data for FlexProt are taken from (Shatsky *et al.*, 2002), in which 'Twist' is equal to the 'Number of flexible regions'and 'RMSD' is equal to the 'Total RMSD' Refer Table 1 for other descriptions.

Table 3. Comparison of the overall structural distortions between structures from different levels

Items	Family	Superfamily	Fold	All
Total pairs	854	3200	2383	6437
Pairs with twists	111 (13.0%)	1796 (56.1%)	1284 (53.9%)	3191 (49.6%)
Average twists	2.0	2.4	2.1	2.3

first type is caused by the conformational flexibility of structures—which is seen as a distortion between structures of the same (or homologous) proteins in different functional states (e.g. with and without the ligand). This type of structural distortions has been studied in protein dynamic analysis (Gerstein *et al.*, 1994; Echols *et al.*, 2003) as well as by flexible structure comparison (Wriggers and Schulten, 1997; Shatsky *et al.*, 2002). We found that many protein pairs on the family level belong to this type.

There are, however, many cases that do not fit into this type and we group them into a different set, calling them structural distortions caused by evolution. It includes the structural differences between very distant homologies. We suspect that such large structural changes are related to the development of new functions, interactions with new partners or significant changes of existing functions (Kinch and Grishin, 2002). Although such cases have been studied with respect to improvement in the comparative modeling of proteins (Reddy and Blundell, 1993; Reddy

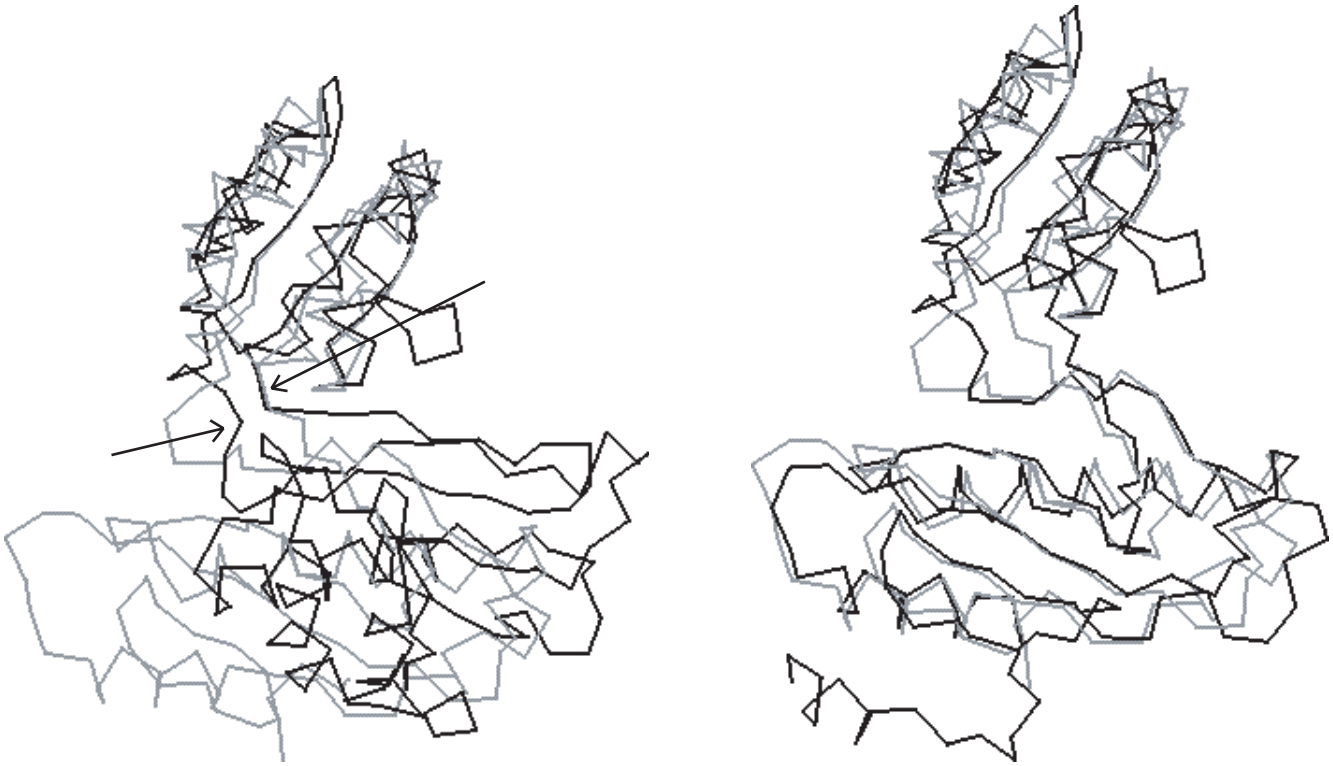

Fig. 4. Comparisons between the 50S ribosomal protein L1P from Methanococcus jannaschii (SCOP code d1cjsa_, shown in gray lines) and the ribosomal protein L1 mutant S179C (SCOP code d1ad2__, shown in black lines). Left graph shows the superposition of these two proteins according to a rigid comparison. Obviously, only one domain from each protein is superimposed in this way. However, after FATCAT introduced two hinges into d1ad2__ (pointed by two arrows), the twisted d1ad2__ is superimposed to entire d1cjsa_ with an RMSD of 2.96 Å, as shown in the right graph.

et al., 1999), they are rarely discussed in the structure comparison. In the following sections, we will discuss some of the examples from both types.

Structural distortions caused by conformational flexibility

A typical case of structural distortions caused by flexibility is illustrated by comparing the ribosomal protein L1 mutant S179C from *Thermus thermophilus* (SCOP code d1ad2__) with the 50S ribosomal protein L1P from *Methanococcus jannaschii* (SCOP code d1cjsa_). Both proteins have two domains that are not linked by a single hinge, as found in many known cases, but instead linked by two hinges (N-terminal and C- terminal form one domain while the middle part of the protein forms the other one). Two twists are introduced in aligning these two proteins by FATCAT, as shown in Figure 4, resulting in a good superposition between the two proteins spanning their entire structures. Indeed, Unge *et al.* (1997) have studied the conformational flexibility of ribosomal protein L1 and showed that this protein has a small but significant opening of the cavity between its two domains, which is suspected to be necessary to accommodate the larger conformational change needed for an induced fit mechanism upon binding RNA.

More complicated structural distortions are found in comparing the apo-dethiobiotin synthase (SCOP code d1byi__) and the adenylosuccinate synthetase from *E.coli* (SCOP code d1qf5a_) (Fig. 5). Both proteins belong to the nitrogenase iron protein like family from the P-loop containing nucleotide triphosphate hydrolases fold, based on the SCOP classification. Except for the many loops in these proteins that can not be aligned, they are well superimposed along the entire proteins when one structure is modified along twists introduced in the FATCAT alignment of 175 aligned positions with an RMSD of 2.96 Å. We see that d1qf5a_ has longer loops than d1byi__, and the hinges detected by FATCAT are distributed in these loop regions. Moreover, the important P-loops from each protein are well superimposed according to the FATCAT alignment, but this is missed by the CE program (Fig. 5b). In fact, the conformational flexibility of adenylosuccinate

synthetase has been reported in comparing hydantocidin complex and the unligated synthetase, which involves the collapse of some structural elements toward the active site crevice (Poland *et al.*, 1996). Our flexible structural comparison provides a simple description of such a collapse process involving movements of many secondary elements by rearranging the structure at several 'twists'positions. We further compare the d1byi__ with other structures that belong to the same family and show the result in Figure 5c. It is clear that the twist positions are generally conserved among these structures, strongly suggesting that our analysis goes beyond an introduction of artificial parameters to improve alignment quality.

Structural distortions caused by evolution

Hinges are often detected in comparing proteins from different families but the same superfamily. These proteins have similar but distinct functions, such as two enzymes having the same catalytic mechanism but different substrate specificity, which often is accompanied by a significant structure distortion. For instance, hinges are found in comparing the L-2-haloacid dehalogenase (PDB code d1zrn__) and the β-phosphoglucomutase (d1lvha_) (Fig. 6). Both proteins belong to the HAD-like superfamily but form different families based on SCOP classification. In contrast, comparing d1zrn__ with other structures in the same family no twists are found. For example, L-2 haloacid dehalogenase complexed with ACY (d1zrn__) and L-2-haloacid dehalogenase (d1jud__) are well superimposed with no twist, resulting in an alignment of 220 aligned positions with an RMSD of 0.27 Å. Although it is reported that Asp10-Ser20, Tyr91-Asp102 and Leu117-Arg135 regions move to the active site in the L-2 haloacid dehalogenase complexed with its reaction intermediates (Li *et al.*, 1998), our calculation shows that this movement cannot be described by a rigid body movement. We further compared d1zrn__ with other structures of the same superfamily and twists are found around the same positions in most of the cases, as shown in Figure 6c. All of these data suggest that the structural distortions between these structures are not caused by the structural flexibility itself, but instead are an evolutionary result in developing enzymes with different specificities. This hypothesis is also supported by the inspection of the structures. Overall, HAD-like proteins are composed of two domains, a helical cap domain and an α/β core domain (Fig. 6a), and active sites are located in the domain interface (Lahiri *et al.*, 2002). This arrangement enables the binding pockets to be changed efficiently during the evolution by a few simple rigid-movements in the cap domain while the structure of α/β core domain remains unchanged, as shown in the observation that the detected twists are all distributed in the cap domain (Fig. 6c).

Hinges are also found in comparisons between domains from the same protein. For instance, FATCAT produced an alignment of 81 aligned positions with RMSD 2.51 Åbetween restriction endonuclease FokI N-terminal (recognition) domain 1 (SCOP code d2foka1) and domain 2 (SCOP code d2foka2) by introducing 2 twists; in contrast, CE produced an alignment of only 48 aligned positions with a high RMSD (6.92 Å). From the evolutionary point of view, the domains in a protein are probably the result of duplication, followed by mutations and the accompanying structural changes required for structural stability or new functions. Flexible structure alignment programs are more suitable for comparing such cases than the rigid ones.

DISCUSSION

Flexible structure alignments are essential in comparing proteins. Proteins are designed to be flexible, and this flexibility is often part of their function. X-ray crystallography sometimes captures this flexibility by solving different structures in various functional states corresponding to different global conformations. In such cases, rigid body alignments can introduce errors to compensate for global structural changes and often miss the structural similarity altogether. The analysis of local and global conformational changes between proteins provides important information for the evolutionary study of structures, the study of structure and function relationship and the homology modeling considering structure changes.

A new flexible structure alignment method has been developed and implemented in a program FATCAT. It provides good alignments both between flexible and rigid structures, in the former case compared to the existing rigid alignment programs and in the latter to the flexible program FlexProt. The major feature of FATCAT is that it optimizes the alignment while it minimizes the number of twists introduced. Therefore, it avoids the problem of existing flexible structure alignment programs that separate these two goals, such as introducing too many hinges into the alignment or missing the optimal alignment because of significant errors in the initial rigid-body alignment.

The results presented here represent the first, exploratory application of the FATCAT algorithm. The scoring functions and associated parameters (Equations 1–7) used in this paper need to be optimized further. The scoring functions used here were chosen based on intuition, rather then extensive optimization, and the parameters were selected on the basis of a small number of examples listed in Table 1 and Table 2. For instance, the penalty of introducing a twist into the alignment is determined solely by the impact of the twist on the RMSD of the two structures (see Equation 5). We expect

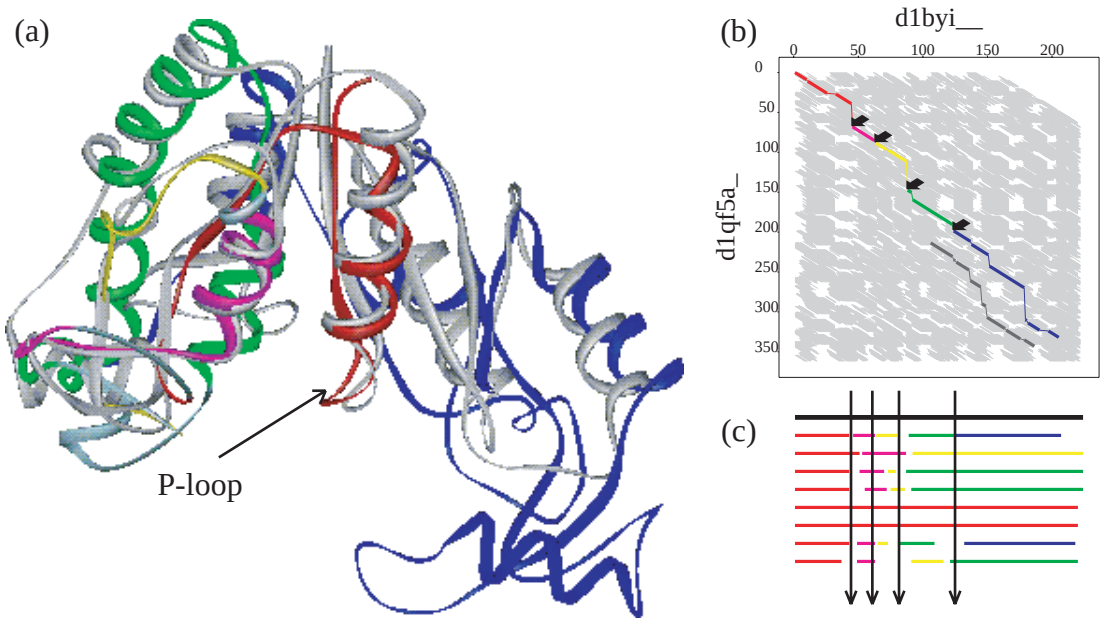

Fig. 5. Comparisons between the nitrogenase (SCOP code d1byi__) and its structural homologues. (a) The superposition of d1byi__ (gray ribbons) and twisted nitrogenase (SCOP code d1qf5a_) according to the FATCAT alignment. Different parts of d1qf5a_ separated by the twists are shown in different colors. (b) AFP chains between d1byi__ and d1qf5a_ by FATCAT and the CE alignment (the short path in black lines) are shown in the dot matrix of AFPs. (c) The schematic representation of the comparisons between d1byi__ (the top line) and its structural homologues (lines below), i.e. d1qf5a_, d1eg7a_, d1cp2a_, d1g3qa_, d1jpna2, d1fts_2, d1ihua1 and d1ihua2. The twists found in comparing d1qf5a_ to d1byi__ are marked by 4 vertical arrows. Different blocks are shown in different colors.

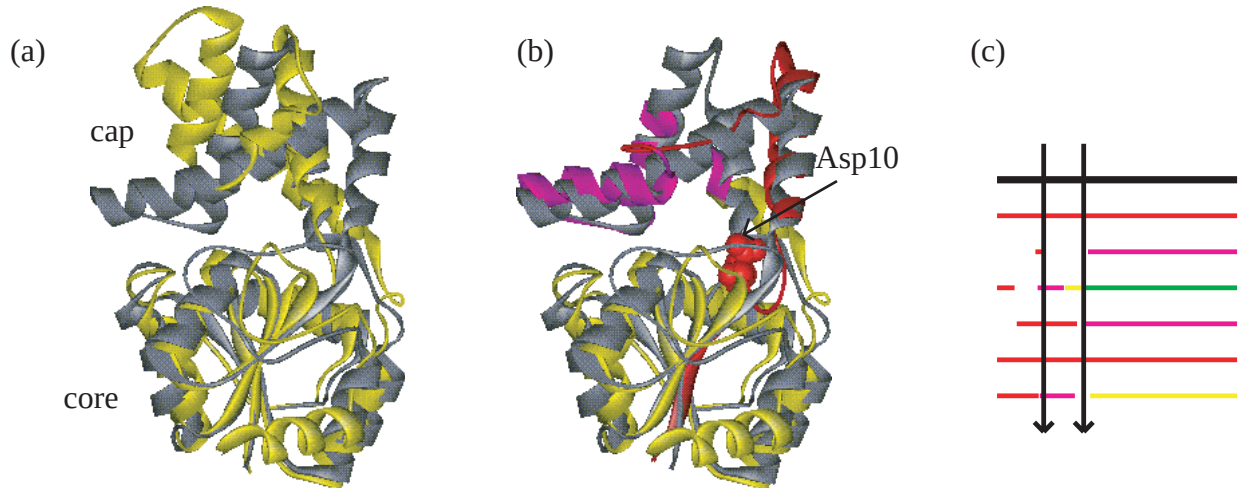

Fig. 6. Comparisons of proteins from HAD-like superfamily. (a) The rigid superposition of the L-2-haloacid dehalogenase (SCOP code d1zrn__, gray ribbons) and the β-phosphoglucomutase (SCOP code d1lvha_, yellow ribbons). The core domains are superimposed but the cap domains are not. (b) The superposition of d1zrn__ and twisted d1lvha_ (with different parts separated by twists shown in different colors) according to the FATCAT alignment. The nucleophile Asp10 of d1lvha_ is shown in CPK. (c) The schematic representation of the comparisons between d1zrn__(the top line) and the other HAD-like proteins from the SCOP 40% database (lines below), i.e. d1qq5a_ (L-2-haloacid dehalogenase), d1ek1a1 (epoxide hydrolase), d1feza_ (phosphonoacetaldehyde hydrolase), d1k1ea_ (probable phosphatase YrbI) and d1lvha_ (β-phosphoglucomutase). The twists found in comparing d1lvha_ to d1zrn__ are marked by 2 vertical arrows. Different blocks are shown in different colors.

to derive more realistic flexible alignments by using a range of penalties for different types of twists. We are also developing a benchmark for flexible alignments to systematically improve and evaluate FATCAT.

ACKNOWLEDGMENTS

We thank Bruce Worcester for help in editing. This work was supported by NIH grant GM63208.

REFERENCES

Bennett,W. and Huber,R. (1984) Structural and functional aspects of domain motions in proteins. *Crit. Rev. Biochem.*, **15**, 291–384.

Boutonnet,N.S., Rooman,M.J., Ochagavia,M.E., Richelle,J. and Wodak,S.J. (1995) Optimal protein structure alignments by multiple linkage clustering: application to distantly related proteins. *Protein Engng*, **8**, 647–662.

Echols,N., Milburn,D. and Gerstein,M. (2003) Molmovdb: analysis and visualization of conformational change and structural flexibility. *Nucleic Acids Res.*, **31**, 478–482.

Eidhammer,I., Jonassen,I. and Taylor,W.R. (2001) Structure comparison and structure pattern. *J. Comput. Biol.*, **7**, 685–716.

Feng,Z.K. and Sippl,M.J. (1996) Optimum superimposition of protein structures: ambiguities and implications. *Fold Des.*, **1**, 123–132.

Fischer,D., Elofsson,A., Rice,D. and Eisenberg,D. (1996) Assessing the performance of fold recognition methods by means of a comprehensive benchmark. In *Pacific Symposium on Biocomputing*. pp. 300–318.

Gerstein,M., Lesk,A.M. and Chothia,C. (1994) Structural mechanisms for domain movements in proteins. *Biochemistry*, **33**, 6739–6749.

Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.

Jacobs,D.J., Rader,A.J., Kuhn,L.A. and Thorpe,M.F. (2001) Protein flexibility predictions using graph theory. *Proteins*, **44**, 150–165.

Kinch,L.N. and Grishin,N.V. (2002) Evolution of protein structures and functions. *Curr. Opin. Struct. Biol.*, **12**, 400–408.

Lackner,P., Koppensteiner,W.A., Sippl,M.J. and Domingues,F.S. (2000) Prosup: a refined tool for protein structure alignment. *Protein Engng*, **13**, 745–752.

Lahiri,S.D., Zhang,G., Dunaway-Mariano,D. and Allen,K.N. (2002) Caught in the act: the structure of phosphorylated beta-phosphoglucomutase from *Lactococcus Lactis*. *Biochemistry*, **41**, 8351–8359.

Li,Y.F., Hata,Y., Fujii,T., Hisano,T., Nishihara,M., Kurihara,T. and Esaki,N. (1998) Crystal structures of reaction intermediates of l-2-haloacid dehalogenase and implications for the reaction mechanism. *J. Biol. Chem.*, **273**, 15 035–15 044.

Madej,T., Gibrat,J.F. and Bryant,S.H. (1995) Threading a database of protein cores. *Proteins*, **23**, 356–369.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Ochagavia,M.E., Richelle,J. and Wodak,S.J. (2002) Advanced pairwise structure alignments of proteins and analysis of conformational changes. *Bioinformatics*, **18**, 637–640.

Poland,B.W., Fromm,H.J. and Honzatko,R.B. (1996) Crystal structures of adenylosuccinate synthetase from *Escherichia Coli* complexed with GDP, IMP hadacidin, NO3-, and Mg2 +. *J. Mol. Biol.*, **264**, 1013–1027.

Reddy,B.V. and Blundell,T.L. (1993) Packing of secondary structural elements in proteins. Analysis and prediction of inter-helix distances. *J. Mol. Biol.*, **233**, 464–479.

Reddy,B.V., Nagarajaram,H.A. and Blundell,T.L. (1999) Analysis of interactive packing of secondary structural elements in alpha/beta units in proteins. *Protein Sci.*, **8**, 573–586.

Schulz,G.E. and Schirmer,R.H. (1979) *Principles of Protein Structure*. Springer, New York.

Shatsky,M., Nussinov,R. and Wolfson,H.J. (2002) Flexible protein alignment and hinge detection. *Proteins*, **48**, 242–256.

Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engng*, **11**, 739–747.

Unge,J., Al-Karadaghi,S., Liljas,A., Jonsson,B.H., Eliseikina,I., Ossina,N., Nevskaya,N., Fomenkova,N., Garber,M. and Nikonov,S. (1997) A mutant form of the ribosomal protein L1 reveals conformational flexibility. *FEBS Lett.*, **411**, 53–59.

Vriend,G. and Sander,C. (1991) Detection of common three-dimensional substructures in proteins. *Proteins*, **11**, 52–58.

Wriggers,W. and Schulten,K. (1997) Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. *Proteins*, **29**, 1–14.

Wuthrich,K. and Wagner,G. (1978) Internal motion in globular proteins. *Trends Biochem. Sci.*, **3**, 227–230.