Fast Multiple Alignment of Protein Structures Using Conformational Letter Blocks

Sheng Wang and Wei-Mou Zheng*

Institute of Theoretical Physics, Academia Sinica, Beijing 100190, China

Abstract: Most approaches for protein structure alignment start from a search for similar fragments since this local similarity is necessary to the alignment even though is insufficient. In contrary to the sequence alignment, any insignificant trial alignment for structures can be detected by structure superposition and then excluded. It is then practicable to select from locally similar fragments those responsible for alignment and build up it. An efficient way for local similarity search is to use a conformational alphabet, which is a discretized description of protein chain local geometry. Using our conformational alphabet and its substitution matrix CLESUM, we propose a tool called BLOMAPS for fast multiple structure alignment.

By means of the conformational alphabet, a structural fragment is mapped to a string, and two strings with their CLESUM score being higher than a preset threshold form a similar fragment pair (SFP). A string from one protein as a seed and its highly similar fragments from other proteins form a similar fragment block. Taking one protein as the pivot, BLOMAPS uses the rigid transformation for SFPs in a block to superimpose proteins and initiate an anchor-based alignment. BLOMAPS is greedy in nature, guided by CLESUM similarity scores. It consists of several steps including finding similar fragment blocks based on a pivot protein, removing block redundancy, constructing scaffold by checking consistency in spatial arrangement among fragments from different blocks, dealing with unanchored structures, and the final step of refinement where the average template for alignment is obtained and motifs missing from the pivot protein are found and added. The utility of BLOMAPS is tested on various protein structure ensembles including large scale ones, and compared with several other tools including MATT.

BLOMAPS is available at: www.itp.ac.cn/zheng/blomaps.rar

Keywords: Protein structure, Multiple structural alignment, Protein conformational alphabet.

1. INTRODUCTION

Protein structures are better conserved than amino acid sequences. Remote homology is detectable more reliably by comparing structures. Multiple alignment carries significantly more information than pairwise alignment, and hence is a much more powerful tool for classifying proteins, detecting evolutionary relationship and common structural motifs, and assisting structure/function prediction. Most recent reviews on protein structure alignment include Refs. [1, 2]. Being able to provide very useful information and insights for proteomics, structural bioinformatics and drug development, the comparison and alignment of protein structures has come to be a fundamental and widely used task in computational structure biology. However, developing accurate and fast methods for multiple structure alignment is still regarded as an open challenge. Here, we propose a tool called BLOMAPS for fast multiple structure alignment.

The common goal of all multiple structure alignment methods is to identify a set of residue 'columns' from each 'row' protein that are structurally similar, or to find an optimal correspondence among the atoms of these molecular structures. For a given multiple alignment, several aligned 'blocks', which correspond to contiguous columns, usually can be identified. Each block is again composed of locally similar fragments. This local similarity within a block may be phrased as 'vertical equivalency'. The local similarity is necessary to the alignment, but is insufficient. For any two structures in the multiple alignment, the transformation to superimpose a fragment duad in an aligned block should also bring the fragment duads in other blocks spatially close. This is the 'horizontal consistency'. In contrary to the sequence alignment, any insignificant trial alignment for structures can be detected by structure superposition and then excluded. Thus, it is practicable to select from locally similar fragments those responsible for the global alignment and build up it.

Structure alignment involves the geometric representation for structures. In most cases, only the backbone of pseudobonds formed by C_α atoms are considered. Atom coordinates, which change under translation and rotation in 3D space, are not geometric invariants. Distances used by DALI [3, 4] or CE [5] are the intrinsic property of a geometric object. The peptide chain dihedral angles or the bending and torsion angles of pseudobonds [6] are also geometric invariants. The unit-vector used by MAMMOTH [7] implies pseudobond angles, but is not invariant. MASS [8, 9] replace secondary structure elements by the vectors of their axes. This vector

*Address correspondence to this author at the Institute of Theoretical Physics, Academia Sinica, Beijing 100190, China; E-mail: zheng@itp.ac.cn

1875-0362/09

representation speeds up the computation, but has a low precision for structural elements. BLOMAPS uses pseudobond angles, but goes one step further. The three pseudobond angles formed by four contiguous residues have been coarse-grained into discrete 'conformational letters', and a substitution matrix called CLESUM has been constructed for these letters based on a database of aligned protein structures [10, 11]. By means of this conformational alphabet, a structural fragment is mapped to a string. BLOMAPS measures the similarity of two fragments with the CLESUM score of their strings.

The horizontal consistency described above is based on superposition. In the language of distances, the consistency is expressed as the distance constraint that any corresponding distances from two structures in alignment should be nearly equal. A few methods like DALI or CE construct an alignment by joining fragment duads satisfying the distance constraint without superposition. Many methods start with an initial trial correspondence, and then iteratively update the optimal transformation and correspondence in turn until the best correspondence is finally found. Often a dynamic programming algorithm is employed in piecing up a global alignment; it enforces the alignment to be collinear. MATT (Multiple Alignment with Translations and Twists), including the difference between transformations for superposing fragment duads as a part of penalty for dynamic programming, is able to deal with local flexibility [12].

The pairwise structure alignment forms the basis for the multiple structure alignment. Most existing methods of multiple structural alignment combine a pairwise alignment and some heuristic with a progressive-type layout to merge pairwise alignments into a multiple alignment [6, 7, 13-16]. Such pairwise-based methods have the limitation that alignments which are optimal for the whole input set might be missed. Another limitation is in speed. There are a handful of truly multiple methods [8, 9, 17, 18]. BLOMAPS also belongs to the category as well as CLEMAPS, another tool developed in our group [19].

BLOMAPS is developed from our CLEPAPS, a fast tool for pairwise protein structure alignment [20]. CLEPAPS searches for similar fragment pairs (SFPs) by string comparison based on CLESUM scores. Like ProSup [21], CLEPAPS is anchor-based. It takes a single good SFP as an initial correspondence. The optimal transformation for this seed SFP is used to superimpose the pair proteins and then to update the correspondence. The procedure of progressively building up larger correspondence is iterated until the best correspondence is finally found. CLEPAPS adopts a greedy strategy guided by CLESUM scores. In fact, the usage of conformational letters may be integrated into any approaches engaging SFPs. However, the number of operations involved in consistency checking for a CE-type method is, roughly speaking, quadratic in the number of relevant residues, while for a ProSup-type it is linear. Our CLEMAPS, being developed before CLEPAPS, is a multiple version of the CE-type, which is less superior in speed than the ProSup-type. Encouraged by the performance of CLEPAPS, we decided to extend it to BLOMAPS, a multiple version. The design of BLOMAPS' architecture has been briefly reported in [22] and [11]. Recently, appeared a new tool MATT, which outperforms other programs in alignment quality on distant structures, and sets a high standard for other tools to reach. We present here a full version of BLOMAPS, which includes not only a detailed description of the method and careful analysis of its implementation, but also some tests on large scale ensembles and case study. BLOMAPS is compared with several other tools including MATT.

2. METHODS

BLOMAPS is greedy in nature. Its several steps include finding similar fragment blocks based on a pivot protein, removing block redundancy, constructing scaffold by checking consistency in spatial arrangement among fragments from different blocks, dealing with unanchored structures, and the final step of refinement where the average template for alignment is obtained and motifs missing from the pivot protein are found and added.

2.1. Abbreviations

Five main abbreviations SFP, AFP, SFB, HSFB and MAB are frequently used and listed as follows. Their more precise definitions will be given in the text.

SFP = Similar fragment pair. As the name suggests, it is a pair of segments, each from each of two given proteins. SFPs are based only on local geometry.

AFP = Aligned fragment pair. We save word 'aligned' only for global features such as orientation of secondary structure elements and overall topology. An AFP must be an SFP, not the other way around. AFPs are those SFPs which finally appear in the global alignment of two whole structures.

SFB = Similar fragment block. SFB is the counterpart of SFP for multiple proteins. We use a simplified version of SFB, for which there is a seed fragment member, and all other members are compared only with this seed. SFBs are based only on local geometry.

HSFB = Highly similar fragment block. For a given fragment as the seed, there are usually many SFBs. The HSFB has its members most similar to the seed. That is, among these SFBs, for any protein in comparison, the member in the HSFB is more similar to the seed than those in any other SFBs.

MAB = Multi-aligned block. MAB is the counterpart of AFP for multiple proteins. MABs form the core of a multiple structure alignment, and then can be extracted from the alignment. Thus, an MAB must be an SFB, not the other way around.

2.2. Conformational Alphabet

To represent protein structures with conformational alphabets, which are discretized conformational states of

certain fragment units of backbones, is an old idea [23-29]. BLOMAPS uses a 3D structure coding of protein backbones consisting of C_α pseudobonds. Three contiguous C_α atoms determine two pseudobonds and a bending angle between them. Four contiguous C_α atoms, say a, b, c and d, determine two such bending angles (θ,θ') and a torsion angle (τ) which is the dihedral angle between the two planes of triangles abc and bcd. Since the length of pseudobones in the dominant trans configuration is almost constant these bending and torsion angles, as the chain counterparts of curvature and torsion of a smooth curve, maintain the three dimensional information. The smallest unit possessing one-to-one correspondence between angles and coordinates is the quadrupeptide unit. By using a mixture model for the density distribution of the three angles, the local structural states have been clustered as 17 discrete states or letters (A to Q) of a protein conformational alphabet. When using structural codes for the structural comparison, a score matrix similar to the BLOSUM for amino acids is desired. Based on the alignments for representative structures in the database FSSP of Holm and Sander, we have constructed a substitution matrix called CLESUM for the conformational letters. The matrix (in an updated version) is shown in Table **1**, where a scaling factor of 20 instead of 2 is used to show more details [10]. To the best of our knowledge, CLESUM is the first substitution matrix directly derived from structure alignments for a conformational alphabet. Among the 17 letters, H represents the prototype of helices, while E represents that of extended strands. In Table **1** similar conformational letters have been grouped closely. CLESUM reflects not only a geometrical similarity, but also an 'evolutionary' similarity. For example, the diagonal entry for the frequent H is relatively small despite the high geometrical similarity between two helices.

An essential parameter of BLOSUM is its valid evolutionary distance, which is controlled by the identity rate of sequences for training. Similarly, we use the structure family indices of FSSP to carefully control the similarity between structures in our training set. A training set of too high similarity will make most non-diagonal entries of the substitution matrix negative.

2.3. Finding Similar Fragment Blocks

Suppose that protein P is one from the structures to be aligned. The coordinates $\{r_i\}$ of C_α atoms of the protein are converted to a sequence S of conformational letters. Since each letter corresponds to a quadrupeptide unit, the length of S is shorter than that of P by three. The first letter is assigned to the third residue by convention, the second to the fourth and so on, until finally the last letter is assigned to the last residue but one. (This assignment is supported by an analysis on the position dependence of the mutual information between the letters and amino acids.)

For given two fragments of the same length l, one starts at residue i of P and the other at j of another protein P' (with conformational sequence S'), their local structural similarity may be measured by

$$\sigma = \sum_{k=0}^{l-1} M(s_{i+k}, s'_{j+k}), \tag{1}$$

where $M(a,b)$ is the (a,b)-entry of the CLESUM, and s_{i+k} and s'_{j+k} are the conformational letters of the

Table 1. CLESUM: The Conformation Letter Substitution Matrix (in Units of 0.05 Bit)

J	37																
H	13	23															
I	16	18	23														
K	13	5	21	49													
N	-2	-34	-11	28	90												
Q	-44	-87	-62	-24	32	90											
L	-32	-62	-41	-1	8	26	74										
G	-21	-51	-34	-13	-8	8	29	69									
M	16	-4	1	12	7	-7	5	21	61								
B	-57	-96	-74	-50	-11	12	-12	13	-13	51							
P	-34	-60	-49	-36	-3	7	-12	5	8	42	66						
A	-23	-45	-31	-19	10	16	-11	-6	-2	20	35	73					
O	-24	-55	-34	5	15	-13	-4	-1	5	-12	4	25	104				
C	-43	-77	-56	-33	-5	29	0	-4	-12	7	4	13	3	53			
E	-93	-127	-108	-84	-43	-6	-21	-22	-47	15	-5	-25	-48	3	36		
F	-73	-107	-88	-69	-32	3	-16	-5	-33	7	0	-20	-30	20	26	50	
D	-88	-124	-105	-81	-44	14	-22	-31	-49	13	-10	-17	-42	21	22	21	52
	J	H	I	K	N	Q	L	G	M	B	P	A	O	C	E	F	D

corresponding residues. Here the same index is kept for a residue and its conformational letter. If the pair score σ is greater than a preset threshold T, the pair is called a similar fragment pair (SFP), which defines a correspondence of l residue duads. Comparing each string of S with length l against S', we can find all SFPs. Usually, a small l and a low T will result in a long list of SFPs.

A rigid transformation can be found to superimpose the two members of a given long enough SFP and make the spatial deviation of its duad C_α atoms very small [30, 31]. Since an SFP is determined only by local similarity, a superposition valid for one SFP need not be valid for another. We define the spatial 'separation' between two members of a certain SFP under a given transformation by

$$\delta = \max_{(r_i, r'_j) \in SFP} \{|x_i - x'_j|, |y_i - y'_j|, |z_i - z'_j|\}, \quad (2)$$

where (r_i, r'_j) is a residue duad of the SFP after transformation, and (x, y, z) denotes the 3D coordinates of r. A small separation δ implies a good superposition of the two SFP members.

Selecting a 'pivot' protein from a set of structures to be aligned and taking an l long fragment of the pivot as a seed, we search the structures other than the pivot for fragments similar to the seed. A fragment which forms an SFP with the seed is called a *neighbor* of the seed. For a given seed, several neighbors may be found in a single structure. A block of similar fragments (SFB) may be formed by taking one neighbor from each structure which has neighbors of the seed. There is a block which consists of the seed and those neighbors which score the highest amongst neighbors in each structure possessing neighbors. Such a block is called a 'highly similar fragment block (HSFB)'. Of course, a seed might have no neighbor in some structures. The total number of fragments in an HSFB will be called the depth of the HSFB. The total score Σ of an HSFB is defined as the sum of σ scores between the seed and each of its neighbors in the HSFB. Another characteristic of an HSFB is its consensus, which is defined as follows. For a given set of conformational letters, their consensus letter is defined as the letter which belongs to the set and has the highest sum of CLESUM scores between itself and all letters in the set. The consensus of an HSFB is then defined as the string which consists of the consensus letters of columns when aligning the 'row' strings of the HSFB. Thus, besides the positions of its member fragments, an HSFB has a depth, score and consensus.

To examine each l long string of every conformational sequence and find all possible SFBs is inefficient. BLOMAPS simply takes the shortest protein as the pivot to create all HSFBs. That is, all seeds are extracted from this protein. (When seeking SFBs, CLEMAPS conducts an all-against-all search for the best centers and finding SFBs as so-called center-stars. The center of such a block is always the consensus. However, besides the high computation cost, it is hard to efficiently remove redundancy from such blocks.) Bearing only local similarity, an SFB need not correspond exactly to a multi-aligned block (MAB), which appears in the multiple structural alignment. Obviously, an MAB must be an SFB in the sense of the CLESUM score. For a set of closely related structures, we expect that there is a good chance of finding certain members of MABs in some HSFBs.

For an HSFB with a large depth and score, shifts of its seed would plausibly also generate HSFBs with a large depth and score. To remove this redundancy, the HSFBs are sorted first in descending order of depth and then in that of score. A 2D grid of atom indices is created with its rows corresponding to individual proteins. The atom indices of the first HSFB in the grid are marked, and then the second HSFB is examined. If the overlapping positions between the two HSFBs are less than $\Gamma = \gamma l m$, where l is the width of the block, m the depth of the second HSFB and γ a parameter for redundancy, we fill in the grid the second HSFB. Otherwise, we skip it, and examine the third. When examining a new HSFB, the number of its positions that overlap with the marked indices of the grid is counted. Only when its overlapping proportion is less than γ do we fill in its place in the grid. This procedure is continued until the last HSFB is examined.

2.4. Building up a Scaffold

Multiple structure alignment requires both the vertical equivalency and horizontal consistency. This increases the difficulty of alignment, but also reduces the chance of making a wrong alignment. That is, by superposition the wrong alignment can be detected and then discarded. The multiple alignment algorithms which progressively merge pairwise alignments may be classified as horizontal-first. Our BLOMAPS may be regarded as vertical-first. To speed up, all structures are aligned against some template, and at the beginning the shortest protein is taken as the template to create HSFBs. BLOMAPS then starts with an HSFB taken from the top K HSFBs as an 'anchor'. To choose the most representative structure, the template is updated to the protein whose member in the anchor HSFB has the highest similarity score with the consensus. The updating of template may result in that the new pivot protein does not have a fragment in an HSFB. Since structures are aligned against the template, it is desired for the template to have as many HSFB members as possible. When an HSFB lacks a fragment of the pivot protein, the consensus of the HSFB is used to search the pivot for possible neighboring fragments with a lower threshold T_-, and add the optimal neighbor to the HSFB.

The transformations to superimpose the fragments of the anchor HSFB are also used to superimpose the whole structures where the fragments are. After having superimposed all the structures which possess a fragment in the anchor HSFB, the horizontal consistency is examined as follows (see a schematic in Fig. **1**). We mark all the HSFBs and their fragments 'uncolored', and set up a counter for each HSFB. If an HSFB contains a fragment of the pivot protein, the fragment will be regarded as the center of the HSFB, and separations between the center and other

members of the HSFB are calculated. When a separation is found below a preset cutoff d_1 the count of the counter for the HSFB is advanced by one, the HSFB is marked colored and both the fragment and the center colored. At the first time when such an HSFB is found, the anchor HSFB, and its two fragments taken for creating the transform for superposition are also colored. The total number of colored HSFBs and that of colored fragments are assigned to the anchor HSFB. Having calculated the total counts for all the top K anchor HSFBs, we select the 'optimal' anchor HSFB to go on to the next step by inspecting first the total number of colored HSFBs, and then the total number of colored fragments if necessary.

For an optimal anchor HSFB, its colored HSFBs are spatially consistent with the anchor. Thus, the colored fragments of the pivot protein are supported both horizontally and vertically, and form a scaffold for multiple alignment. It may happen that the anchor HSFB has no neighbors in some structure (even after an extra search with the block consensus), which is then regarded as an unanchored protein. No transformation can be found based on the anchor HSFB to superimpose such an unanchored protein on the pivot protein.

If an protein has a colored fragment in an anchor HSFB, but has no fragment in a colored HSFB, we search the protein for fragments which do not overlap with any colored fragments and are similar to the center either by checking the similarity score first, or directly examining the separation with respect to the center fragment. If a separation is found being smaller than the cutoff d_1 the corresponding fragment is colored and added to the HSFB. For a structure which has a fragment in the anchor HSFB, if the number of its colored fragments is smaller than a lower bound n_h the protein is also regarded as unanchored, and otherwise as anchored (see Fig. **1**).

2.5. Improving the Scaffold

Improving of the scaffold greatly follows the pairwise CLEPAPS [20]. By using the colored fragments, the transformation to superimpose two structures, which so far is based only on a single pair of fragments, can be updated based on more fragment pairs satisfying consistency. That is, for an anchored protein, the transformation optimal for superimposing all its colored fragments on the pivot protein is the updated transformation for the two proteins. With the transformation updated, we use a width l' smaller than l and a more stringent cutoff d_2 to examine the consistency of colored fragments. If an l long colored fragment has at least l' residues whose coordinates deviate from their counterparts on the pivot protein within d_2, the fragment remains colored, and otherwise it is changed to uncolored.

Having examined all l long colored fragments, we 'recruit aligned fragment pairs (AFPs)' for every anchored protein as follows. For a given anchored protein, the colored fragments and their partner fragments on the pivot define the primary AFPs. They are masked from the anchored and pivot proteins. With each fragment of length l' from the unmasked region of the pivot, the unmasked region of the anchored protein is searched for SFPs at a lenient threshold T'. After sorting the found SFPs in descending order of scores, we examine the separations of the SFPs in succession. Whenever a separation is found to be smaller than d_2 and the positions of the corresponding SFP do not conflict with the existing AFPs, the SFP is recorded as an AFP, and its fragment on the pivot is assigned as a part of the scaffold. An extended scaffold is then obtained. The AFPs map residues of proteins other than the pivot to those of the pivot protein, and define columns of residue correspondence. We construct the first average template by averaging transformed coordinates of atoms over individual columns, and use it for dealing with unanchored structures.

2.6. Dealing with Unanchored Structures

The optimal anchor HSFB does not provide any guidance for aligning an unanchored protein. However, the protein may still have members in other colored HSFBs. Any of such members can be used to generate a transformation for superimposing the protein on the template for examining the consistency of other fragments. For efficiency, we may sort the members of colored HSFBs according to their depths or similarity scores, and then examine only the top K. If an unanchored protein does not have enough number of members which can be associated with colored HSFBs, a pairwise alignment is necessary.

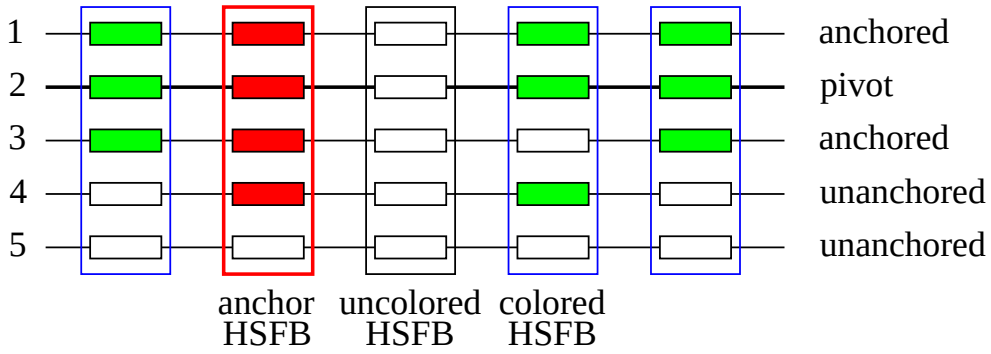

Fig. (1). Schematic of horizontal consistency examination. Selected by the consensus of the anchor HSFB, structure 2 is the pivot, on which all other structures are superimposed. The anchor HSFB has 4 colored HSFBs and 13 colored fragments. For $n_h = 3$ structures 1 and 3 are anchored.

There is a difference between the pairwise alignment here and the ordinary one. The scaffold on the pivot protein now provides a guidance. Each fragment of length l is taken from the scaffold as a seed, and the unanchored protein is searched for SFPs at threshold T. After sorting the found SFPs in descending order of scores, the top K SFPs are used to generate transformations and the top $J(\geq K)$, say $10K$, are used for consistency checking. (The algorithm speed is insensitive to the choice of J). The transformation derived from the fragment pair of an SFP in the top K is used to superimpose the unanchored protein on the template. Separations of the top J SFPs are then successively examined. We count the total number n_p of non-overlapping SFPs whose separations are less than d_2. The SFP which has the largest n_p among the top K and its consistent SFPs are used to update the transformation for superimposing the unanchored protein. The procedure of recruiting AFPs is then applied to extend the portion aligned to the scaffold. Every unanchored protein can be treated in this way.

2.7. Refinement

2.7.1. Updating the Average Template

The average template can be updated once unanchored proteins are aligned on the template and missing motifs are discovered. A cutoff d_3, even more stringent than d_2, is applied to examine the deviation between residue duads of AFPs and their flanking sites. Fragments are elongated or shrunken according to the deviation cutoff d_3, and hence the AFPs updated. The modified AFPs lead to an updated average template. This is an iteration, and its convergence is usually rather fast. At the final iteration step, the distance cutoff d_0 is used to control residue duad distances (instead of separations which are just maximal coordinate differences).

2.7.2. Seeking for Missing Motifs

The patterns or motifs considered here are ungapped fragments which are structurally similar locally, and are arranged in space very alike globally. It is doomed by the greedy nature of the above scheme that only patterns shared by the pivot protein have a chance to be discovered. Some patterns could be shared by a subset of structures, but be absent from the pivot protein. They are 'missing motifs' to the pivot protein. Their information has to be extracted from the structures sharing them.

A motif in a set of structures must also be a motif of their conformational sequences. Although the reverse need not be true, the latter provides a candidate for the former. There are many methods for discovering motifs in a set of sequences, e.g. a simple center-star approach in the sense of strings instead of sequences [32]. An exceptionally large protein will have a large proportion of 'blank' regions, most of which have no contribution to missing motifs, and hence a procedure to directly examine all fragments would be rather inefficient. The inefficiency occurs also when the number of structures is large.

Considering the fact that the structures have been superposed in the space, we propose a way to rescue missing motifs by 'cell registration'. The space occupied by the structures after superimposition is divided into uniform cubic cells of a finite size, say 6Å . The number of different proteins which have their residues falling in a given cell is the depth of the cell. The cells with their depths below a preset cutoff are discarded. The remaining cells are sorted in descending order of depth. Picking a cell of a large depth as a base, in each dimension we select from its two nearest neighbor cells the one with the higher depth to expand the base cell and double its size into an 'octad'. The residues falling into the octad are marked. In this octad, if a protein which has at least three marked residues with their indices within l' is found, we take the segment of the protein which covers the most marked residues as a seed, and examine all the segments which are from other proteins and have some marked residues by checking their separation from the seed. If AFBs are found, we discard the eight cells of the octad after finding the AFBs, and continue with the cell of the next highest depth until all cells are examined.

So far it is implicitly assumed that there is a common core shared by the whole set of structures. In the case the structure set is actually divided into subsets according to common cores, the above scheme extracts a subset based on the pivot. The algorithm first accomplishes the alignment for this subset, and then treats the remainders as a new input set.

2.8. Evaluation

The final alignment may be described with the involved rigid transformations and viewed visually. Another convenient way is to give the complete residue correspondence. A full column of the correspondence has residues from every protein of the structure set. Usually, the common core of alignment is rigorously defined by all full columns. A 'partial core' may also be defined by introducing a parameter of proportion. For example, core-60 is given by columns covering sixty percent or more proteins of the structure set. Assume that a new motif is found on a protein other than the pivot. With the help of the common core, we can map the protein together with the missing motif on the template, and update the template by averaging coordinates over residues in individual new columns. After having superimposed structures on the final average template, we can calculate the total squared deviation of aligned residues with respect to the template for full or partial core of the alignment, and then the RMSD (root mean square deviation) for evaluation.

The default parameters of BLOMAPS are listed in Table **2**. Finally, we summarize the overall algorithm of BLOMAPS: 1) creating HSFBs using the shortest protein as a template, sorting HSFBs, and deriving redundancy-removed HFSBs; 2) for each HSFB in top K, selecting the pivot protein based on the HSFB consensus, superimposing other proteins on the pivot, finding consistent HSFBs, and selecting the best HSFB according to the number of consistent HSFBs; 3)

Table 2. Default Parameters of BLOMAPS

Symbol	Value	Meaning
l	12	width of wide SFBs
T	200	similarity threshold for wide HSFBs
K	5	number of wide HSFBs tested as anchors
l'	8	width of narrow SFPs for scaffold improvement
T'	100	similarity threshold for narrow SFPs
γ	0.5	overlap proportion for removing redundancy of HSFBs
ρ	4	minimum length of aligned fragments
d_0	5Å	distance cutoff for evaluating overall alignment
d_1	10Å	separation threshold for consistency
d_2	7.5Å	separation bound for recruiting AFPs based on the pivot
d_3	5Å	separation cutoff for recruiting AFPs based on the template
n_h	3	minimal number of colored fragments for an anchored protein

building a primary scaffold from the consistent HSFBs, updating the transformation using the consistent HSFBs, recruiting aligned fragments, and creating an average template; 4) dealing with unanchored proteins, finding missing motifs by cell registration, and refining the alignment; and finally 5) evaluating the alignment.

3. RESULTS

BLOMAPS has been tested on 16 protein structure ensembles as well as some large scale ensembles. The 16 ensembles, covering various challenging cases of structural alignment, are taken from several references: five from Ref. [17], three from [8], four from [9], and four from [6]. Some ensembles contain structural homologies at different levels, some exhibit submotifs not shared by all members or different topologies, while others contain a large number of proteins, or exhibit symmetry or repetition. The ensembles are briefly summarized in Table **3**. The meaning of the abbreviated names and original references for the ensembles listed in the table are as follows. MicrRib: Microbial ribonucleases, Subtil: Subtilisins, TIM61: a set of 61 TIM barrels, Serin5: a set of 5 Serine proteinases, Serpin: Serpins, Thior: Thioredoxins, Beta: All beta immunoglobulins, Glob10: a set of 10 Globins, Glob16: another set of 16 Globins, Serin68: another set of 68 Serine proteinases, CaBind: Calcium-binding proteins, CL-GL: Cofilin-like/Gelsolin-like proteins, PLP: PLP-dependent transferases, C2: C2-domains, TIM7: another set of 7 TIM barrels, and HelBun: Helix-Bundles (with the number of proteins less by one due to the fail in tracing PDB-ID in the updated PDB version). DBNWa: MASS [8], DBNWb: MASS [9], SNW: MultiProt [17], YJ: Ye & Janardan [6], CK: Chew & Kedem [13].

To conduct large scale comparisons between BLOMAPS and other tools, they should be downloadable to run locally, and provide readable alignments easy for further handling. Fortunately, MAMMOTH-mult (abbreviated as MAMMOTH later on) [7], Mustang [16] and MATT [12] are such softwares available from the web or authors. They all use dynamic programming, and build multi-alignment progressively from pairwise alignments. Occasionally, we also manually make comparisons with some other tools. (As mentioned before, CLEMAPS is of the CE-type, involving many pair comparisons. The use of conformational letters makes it still fast, but its alignment quality is less satisfactory at least for the published version. The paper of CLEMAPS reported that the running time for Serin68 was 27s. The same set takes BLOMAPS 1.84s. Since CLEMAPS has been compared with MAMMOTH, we shall not make further comparison with CLEMAPS.)

3.1. Implementation of BLOMAPS on 16 Test Sets

BLOMAPS uses several greedy strategies. It starts by taking the shortest protein as a pivot for finding HSFBs. It could happen that the shortest one poorly represents the set. The found HSFBs will then have a low quality, and even cannot pass the examination of vertical equivalency and horizontal consistency. In this case, BLOMAPS will get warned at the very beginning, and a second protein has to be taken as a new pivot. However, taking the shortest as a pivot works for all the 16 ensembles. Furthermore, for five ensembles (MicrRib, Subtil, TIM61, Serin5, Serpin), three of which have members over 60, the similarity between members in a set is so high that the shortest always well represents the set, and the optimal anchor HSFBs, which always have members from all proteins in the set, directly correspond to a MAB (except for one member in 63 of MicrRib). They are then 'trivial cases'. The top few HSFBs of local similarity also agree with the global alignment. Mere string comparison for conformational codes with the CLESUM matrix can lead to the right answer.

Although these trivial cases are not the best examples to show the power of the conformational letter description we mention a pair of structures, 1gci: 1gnsA, in the set Subtil (see Table **3**), whose optimal anchor HSFB is supported by 12 other colored HSFBs. The amino acid identity rate for the pair in these 13 HSFBs as the aligned proportion is 88/156 while 120/156 of their conformational letters are identical. A simple elongation of colored HSFBs with duads of positive CLESUM scores with respect to the pivot protein leads to total of 216 aligned positions for the set.

The first notable ensemble is the set of Thioredoxins ('Thior' in Table **3**). The shortest protein 1fo5A is of length 85, rather shorter than all the rest nine (between 105 and 112). The HSFB of rank 1 is selected as the optimal anchor HSFB, having six colored HSFBs. (Here, 'rank 1' means that the HSFB is ordered first according to its depth and then to its total score among HSFBs.) The seed of the HSFB has codes FEEENOGCEDEQ from 1fo5A, while the consensus of the HSFB selects EEEENOGCPLDE of 2tir as a representative, then 2tir becomes the pivot protein. The shift of center member results in a score change of the HSFB from 439 to 688. This HSFB happens to be an MAB. Interestingly, being supported by only two colored fragments, 1fo5A turns to be unanchored. Its member in the anchor HSFB may be regarded as a 'weak member'. That is, although it finally appears in the global alignment, it is not strong enough for inferring the alignment. However, SFPs of other colored HSFBs can still be used to superimpose 1fo5A against the pivot for checking horizontal consistency, and then align it. If the members of a protein in all the colored HSFB are wrong or week, which should be rare, we have to conduct a direct pairwise alignment to align the protein against the scaffold. However, this does not happen here.

The next notable ensemble is the set of 16 globins (Glob16), which was first studied by Chew and Kedem [6, 13]. The shortest protein is 1eca. The optimal anchor HSFB, with 1bdbA being the pivot selected by the consensus of the HSFB, has 6 colored HSFBs, all with depth 16. The total number of colored fragments is 38. Since an HSFB reflects only the local similarity it need not be an MAB. For this set the anchor HSFB has two 'wrong' members from 1eca and 2hbg, i.e., it is not the fragment appearing in the final alignment. No supported SFPs are found for these two wrong fragments in the examination of horizontal consistency, and they are then rejected. In fact, there are two more members of the HSFB which fail in the horizontal consistency examination. The scaffold after recruiting aligned fragments consists of 17 pieces. By using the colored HSFBs, the four structures unanchored so far are easily aligned against this scaffold.

Ensemble C2 consists of ten proteins, four 'Synaptotagmin-like' proteins and six 'PLC-like' proteins, taken from two families of the 'C2 domain' superfamily [8]. The two families are related by a circular permutation while each forms a topological group. The lengths spread over a wide range from 123 (1bdyA) to 841 (1qasA). No dynamic programming is conducted in MASS, so the non-topological alignment of this ensemble was detected by MASS [8]. The core of the MASS alignment forms a sandwich of eight β-stands, which will be indexed from a to h. The model alignment for the four 'Synaptotagmin-like' proteins is

Table 3. The 16 Ensembles Used to Test BLOMAPS. The Last Five Columns are the Rank, Depth, Number of Wrong Members of the 'Optimal' Anchor HSFB and Numbers of its Colored HSFBs and Fragments, respectively

Name	Ref	Size	L	Rank	Depth	n_{wrong}	N_b	N_f
MicrRib	DBNWb	63	100:104	5	63	1	5	309
Subtil	DBNWb	60	263:281	2	60	0	13	770
TIM61	DBNWb	61	385:443	4	61	0	22	1217
Serin68	DBNWa	68	181:396	1	68	8	9	410
Serin5	SNW	5	274:279	2	5	0	14	58
Serpin	SNW	13	337:420	1	13	0	16	150
Thior	YJ	10	85:112	1	10	0	6	38
Beta	YJ,CK	6	95:115	1	6	0	4	16
Glob10	YJ	10	136:158	3	10	3	7	32
Glob16	YJ,CK	16	136:158	2	16	2	6	38
CL-GL	DBNWb	12	96:174	2	12	0	5	32
PLP	DBNWa	11	361:730	1	11	4	15	54
C2	DBNWa	10	123:841	1	10	4	6	26
CaBind	SNW	6	75:185	3	6	4	3	7
TIM7	SNW	7	247:491	5	7	5	4	10
HelBun	SNW	9	106:159	3	9	5	4	11

Size: Ensemble size, L: Length range.

abcdefgh in the element indices while that for the six 'PLC-like' proteins is *bcdefgha*. In the MASS alignment element *d* was absent in 3rpbA and 1bdyA. Any multiple alignment tool which uses dynamic programming is able to discover only collinear alignment. For example, in the MATT alignment element *a* of 'PLC-like' (near *C*-terminus) was missing, but element *d* exists in all the 10 structures. The missing element *a* counts 8 columns, and makes the MATT alignment shorter than BLOMAPS. According to PDB files the segments corresponding to element *d* of 3rpbA and 1bdyA are not annotated as sheets, which explains the missing of the element in the alignment of MASS, a secondary structure based tool. In fact, the conformational codes of element *d* for the whole set are typical of strands. BLOMAPS is able to detect all the eight elements.

Compared with the other members in ensemble C2, protein 1qasA is tremendously large (about seven times larger than the smallest). This ensemble is a good example to demonstrate the cell registration technique for missing motifs. Besides the above mentioned core of eight β-strands, there are five submotifs, which are not shared by all proteins. After the 10 structures have been superimposed together, they occupy a volume of $12\times15\times11=1980$ (in units of 6.0Å for sides). Among the 1980 cells, only 101 cells contain points from at least three proteins. By sorting these cells in descending order of depth, The five subpatterns are discovered using the first, second, fifth, sixth and eighth cells. The third, fourth and seventh cells are removed during the octa formation. (A cell size of 5.0Å has been tested, and also works.)

Another ensemble showing subset alignment is CL-GL, which consists of 12 structures belonging to the fold 'Actin depolymerizing proteins', four from the Cofilin-like(CL) family and eight from the Gelsolin-like (GL) family. The two families share five β-strands (indexed from *a* to *e*) and two α-helices (indexed as 3 and 4), and family CL has two additional helices (indexed as 1 and 2). Written in these indices, the structurally conserved common core is *acd*3*e*; the CL family is characterized by 1*ab*2*cd*3*e* while GL family by *abcd*3*e*4 [9]. The shortest structure, being 1d0n, and the pivot protein selected by the consensus of the optimal anchor HSFB are the same one. It belongs to the GL family. The anchor HSFB is supported by 5 colored HSFBs and 32 colored fragments. Although the HSFB is an MAB, *i.e.*, all its members appear in the final alignment, there are still three members from CL family which fail in the consistency examination. The first scaffold is then built based on the rest nine structures. After recruiting aligned fragment pairs the improved scaffold consists of 10 pieces. There is no difficulty to align the three unanchored structures against the scaffold by using colored HSFBs. Since the pivot structure belongs to the GL family, the two additional helices specific to the CL family have to be detected as missing motifs. By cell registration we find not only these two CL helices, but also many other subpatterns. For example, the helix 4 is split into two submotifs according to CL and GL, and it is missing in CL protein 1f7sA. However, 1f7sA shares with the two other CL members some submotifs which are missing in CL member 1ak6. If looking only at the core of full aligned columns, the BLOMAPS alignment agrees with those of other tools, but is a little longer.

An example of large ensemble is Serin68, which comprises 68 molecules of the SCOP family 'Prokaryotic trypsin-like serine protease' [33]. Aligning this ensemble is not acceptable to the MAMMOTH-mult web server due to its large size (although it is not an actual limit of MAMMOTH). The optimal anchor HSFB has 68 members, and is top-1 HSFB in the sorted list, with 8 members being wrong. Protein 1csoE is selected as the pivot by the consensus of this HSFB. It is supported by eight other HSFBs, seven of which have their depths 68 (full). The transformations based on single SFPs of the anchor HSFB are able to align 60 structures against all the nine fragments of the primary scaffold. There is no difficulty in aligning the 8 unanchored proteins on the scaffold by using colored HSFBs other than the anchor. The lengths of proteins vary over a large range in this ensemble, which requires a step for missing motifs. It should be pointed out that this ensemble is highly redundant. At least 40 members have lengths between 195 and 198; their members in the top five HSFBs are highly identical in both amino acids and conformational letters. The core of alignment for these 40 structures has 195 residues with RMSD 0.139Å.

Ensemble CaBind has six proteins of EF hand-like superfamily extracted from three families [17]. According to the pairwise relation, the ensemble can be split into two groups: non-repetitive (3icb, and 4cpv) and repetitive (2scpA, 1scmB, 1top and 2sas). For the latter, a pairwise alignment of a single structure against itself admits multiple solutions. This set is rather tricky. With the shortest 3icb, 10 HSFBs are found. The consensus of the optimal HSFB picks up 2sas as the pivot. However, due to a lack of enough colored fragments no structures are anchored, and a pairwise alignment becomes necessary. After aligning the five structures against 2sas, the common core of BLOMAPS finally has 56 residues. The BLOMAPS alignment is rather close to that of MATT.

Since various criteria are used it is difficult to define a general comparison between different aligning methods. A common goal for structural alignment is to minimize the deviation of the conserved core while maximizing the size of the core. We set up a cutoff 3.0Å to standardize the deviation criterion for the core: only when the root mean squared deviation (RMSD) of the residues in a given aligned column with respect to their corresponding site in the average template is smaller than 3.0Å will the column be kept in the core. Note that the RMSD here is in the sense of a single aligned column. The core size obtained with this cutoff is denoted by N_c. Such N_c can be compared directly. Besides, an important measure for comparison is some identity rate between the compatible cores of alignments of two methods. Here we take the BLOMAPS alignment as a reference. For example, when we compare BLOMAPS with

Table 4. Comparison of BLOMAPS Alignments with those of MUSTANG, MAMMOTH and MATT

			BLOMAPS	MUSTANG			MAMMOTH			MATT		
Name	**Size**	L_{mean}	N_c^-	N_c	N_c^-	$N_0 + N_\pm$	N_c	N_c^-	$N_0 + N_\pm$	N_c	N_c^-	$N_0 + N_\pm$
Serin5	5	277	238	236	230	224+ 6	241	239	227+ 7	243	232	224+11
Serpin	13	369	305	301	299	289+ 9	303	297	285+14	307	301	293+10
Thior	10	105	75	74	67	66+ 4	80	80	71+ 2	76	73	66+ 4
Beta	6	107	76	77	76	71+ 2	79	76	70+ 1	78	69	69+ 3
Glob10	10	147	116	116	112	110+ 2	117	116	110+ 1	118	118	111+ 3
Glob16	16	146	99	98	93	89+ 3	98	95	90+ 2	105	98	87+10
CL-GL	12	126	66	62	62	49+11	62	59	45+12	64	58	49+12
PLP	11	443	198	143	119	118+15	167	154	124+19	194	131	124+23
C2	10	258	84	4	3	0+ 0	32	30	26+ 1	73	62	67+ 3
CaBind	6	140	56	41	41	28+ 3	0	0	0+ 0	59	52	50+ 5
TIM7	7	390	91	1	0	0+ 0	0	0	0+ 0	69	50	11+ 4
HelBun	9	131	60	0	0	0+ 0	18	18	4+ 1	33	29	13+ 1

L_{mean}: the mean length. N_c: the number of fully aligned columns at cutoff 3.0Å without the restriction on minimum length of aligned fragments; N_c^-: the number of the full columns with the restriction set to 4 (applied before cutoff 3.0Å). N_0 (or $N_\pm$): effective numbers of fully aligned columns (in N_c) which coincide with those of BLOMAPS (or shift at most four sites).

MATT, usually corresponding columns from two cores of alignment can be easily recognized by counting the identical site indices, which should not be less than the half of the ensemble size. The identical indices are summed, after divided by the ensemble size, to give the effective number of 'identical columns' N_0. To include also small shifts, we count nonidentical site indices which deviate at most four sites (one turn of a helix) in the related columns. The number of site indices is converted to another effective number of columns $N_\pm$. The number $N_0 + N_\pm$ may be then taken as a measure of accordance between two alignments. The comparison of BLOMAPS with MUSTANG, MAMMOTH and MATT on 12 of the 16 sets is summarized in Table **4**. For the 4 sets with set size over 60, either the size is beyond the limit of a tool, or the running time is too long except for BLOMAPS, so these sets are not included in the table. BLOMAPS's N_c for these four large sets MicrRib, Subtil, Tim61 and Serin68 are 99, 257, 364 and 120, respectively. A running time comparison conducted on subsets of various sizes extracted from Tim61 will be presented later. One case of discrepancy between BLOMAPS and other tools mentioned above is ensemble CaBind due to repetition. Ensembles CaBind, TIM7 and HelBun contain members from different superfamilies or even different folds, besides the symmetry and repetition. Thus, observing discrepancy among different methods in these ensembles is not so surprising. It is seen that, in absence of symmetry and repetition, alignments of BLOMAPS usually agree with those of other three tools; BLOMAPS is most close to MATT. As a rule, the aligned length obtained by a tool using dynamic programming is longer and scrappier than by a tool without it. BLOMAPS superadds a minimum length of aligned fragments which is set to four as default. To make a close comparison, the aligned lengths N_c^- with this same restriction (applied before the cutoff 3.0Å for fairness) are also given in the table.

3.2. Large Scale Comparison of BLOMAPS with other Tools

Large scale tests are conducted on two sets: ASTRAL40-fam and SABmark-sup. ASTRAL40 is extracted from %40 percentage identity filtered ASTRAL SCOP genetic domain sequence subsets [34]. By excluding families with members less than 3 or over 25, and three more families containing member proteins which cannot be traced by their PDB indices in the updated version of PDB, ASTRAL40-fam covers 852 SCOP families. Taking the BLOMAPS alignment as reference, we divide the set into two subsets: alike and unlike. For a family in subset 'alike', N_c of BLOMAPS is never less than 85% of the largest ('best') among the three N_c values from MUSTANG, MAMMOTH and MATT. At the same time, $N_0 + N_\pm$ between BLOMAPS and the best tool is never less than 85% of the smaller in the two N_c values. (When a tool reports no alignment for a family, the N_c is assigned as zero.) This subset 'alike' has 783 families, and the rest 69 families form subset 'unlike'. The comparison of BLOMAPS, MUSTANG, MAMMOTH and MATT is summarized in Table **5**.

The original SABmark, designed as a sequence alignment benchmark, provides structurally alignable groups of SCOP superfamily level, and consists of two subsets: Twilight Zone (-twi) and Superfamilies (-sup) [35]. The

Table 5. Comparison Between BLOMAPS, MUSTANG, MAMMOTH and MATT on Test Set ASTRAL40-fam

Alike(783)	BLOMAPS	MUSTANG	MAMMOTH	MATT
BLOMAPS	**123.3**	5.3	5.7	7.2
MUSTANG	103.7	**114.0**(109.8)	4.9	5.4
MAMMOTH	106.8	101.2	**117.4**(115.3)	6.1
MATT	110.2	103.8	106.1	**124.2**(116.5)
Unlike(69)	**BLOMAPS**	**MUSTANG**	**MAMMOTH**	**MATT**
BLOMAPS	**57.7**	1.9	2.9	3.2
MUSTANG	23.2	**43.4**	2.9	2.2
MAMMOTH	21.4	22.6	**44.6**	2.8
MATT	29.5	29.4	28.5	**65.0**

Average $N_c(N_c^-)$ (diagonal), N_0 (lower-left) and $N_\pm$ (upper-right) are listed for two subsets of 'alike' and 'unlike'.

former is more divergent in sequences than the latter. Thus, SABmark-twi might be better for the case study, and for the large scale comparison we use only SABmark-sup of 425 groups, each of which contains 3 to 25 members. (Note that the 'false positive' sequence counterparts for the purpose of discriminant analysis are never used here.) Ten groups contain member proteins which cannot be traced by their PDB indices. The remain 415 groups are then used for comparison. The set can also be divided into 'alike' and 'unlike' subsets with the same criteria for ASTRAL40-fam. The averaged values of N_c, N_0 and $N_\pm$ for the two subsets are summarized in Table **6**.

Finally, we give three examples of case study in more details. The first example, taken from SABmark-sup (Group 323), is the SCOP superfamily Glutamine synthetase/guanido kinase of three domains: d1m15a2, d1qh4a2 of family Guanido kinase and d1f52a2 of family Glutamine synthetase with a fold description of the common core being two beta-alpha-beta2-alpha repeats. Domain d1f52a2 is longer than the other two by about 100. Written in helical elements (indexed from *1* to *5*) and beta strands (indexed from a to g), the BLOMAPS alignment is *1a2bc3de4fg5* (from the N- to C-terminus), which well represents the fold. Values of N_c (at cutoff 3.0Å) are 147, 21, 34 and 74 for BLOMAPS, MUSTANG, MAMMOTH and MATT, respectively. MATT aligns *1a2bc3* of domain d1f52a2 to *3e4fg5* of the other two. Alignments reported by MUSTANG and MAMMOTH are not better than that of MATT. Comparative alignments are illustrated in Fig. (**2**).

The second example, taken from ASTRAL40-fam, consists of 12 proteins of SCOP family Thiolase-related. This is an example of submotifs. The 12 proteins form three groups: I=(d1m3ka2, d2bywa2, d1ox0a2, d1tqyb2, d1tqya2), II=(d1tqyb1, d1tqya1, d2bywa1, d1ox0a1), and III=(d1wdkc1, d1afwa1, d1m3ka1). With helical elements being indexed by arabic digits and strands by letters, the common core for the whole family can be characterized as $a1b2c3d$, while group I as $a1b2c3c''d4'$, group II as $a1'a'a''1''1b2c2'2''3d3'd'4'$, and group III as $a1b2c3d45fg6h7$. The segment $c'c''$, consisting of a long loop intervened between two very short strands c' and c'', is missing in d1m3ka2 of I; elements d' and $4'$ are missing in d1tqyb1 of II. The BLOMAPS alignment of the family is shown in Fig. (**3**). The aligned cores of BLOMAPS,

Table 6. Comparison Between BLOMAPS, MUSTANG, MAMMOTH and MATT on Test Set SABmark-Sup

Alike(345)	BLOMAPS	MUSTANG	MAMMOTH	MATT
BLOMAPS	**105.0**	5.6	6.0	7.8
MUSTANG	80.6	**91.4**(87.6)	5.0	5.7
MAMMOTH	82.5	74.9	**94.0**(92.0)	6.8
MATT	89.0	79.8	80.7	**104.6**(96.4)
Unlike(70)	**BLOMAPS**	**MUSTANG**	**MAMMOTH**	**MATT**
BLOMAPS	**53.0**	2.5	3.4	3.5
MUSTANG	16.1	**34.3**	2.3	2.6
MAMMOTH	15.5	13.4	**36.1**	4.1
MATT	21.5	18.7	20.5	**57.1**

Average $N_c(N_c^-)$ (diagonal), N_0 (lower-left) and $N_\pm$ (upper-right) are listed for two subsets of 'alike' and 'unlike'.

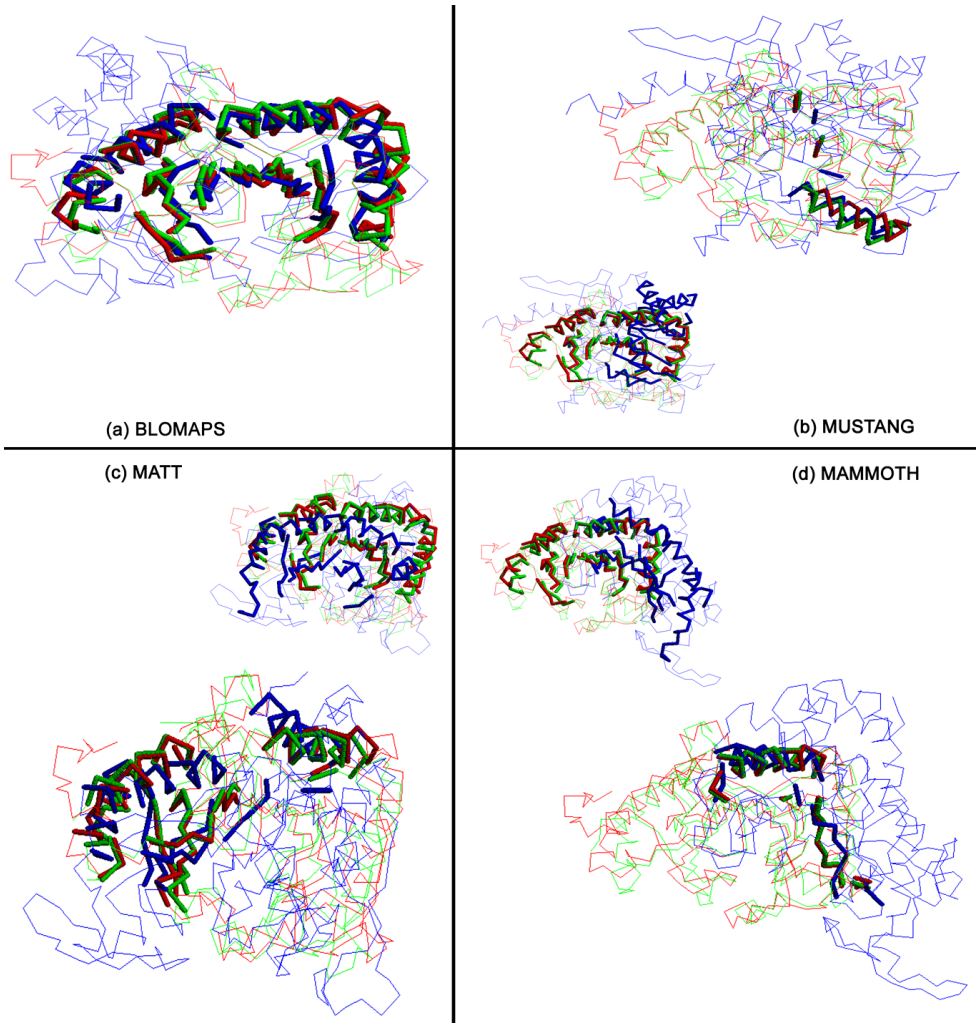

Fig. (2). Comparative alignments of SABmark-sup Group 323. (**a**) The core of BLOMAPS alignment is shown with bold lines for d1f52a2 (in blue), d1m15a2 (in green) and d1qh4a2 (in red). A more stringent cutoff 2.5Å is used for a perceptible visualization. The fragments of this core will be also shown in the following alignments of MUSTANG, MAMMOTH and MATT as insets. (**b**) The core of MUSTANG alignment, (**c**) The core of MATT alignment, and (**d**) The core of MAMMOTH alignment. (Pictures were drawn with RasMol [36]).

MUSTANG and MATT are very similar, while that of MAMMOTH is much shorter.

The third example, taken also from ASTRAL40-fam, is the SCOP family Legume lectins of 4 proteins in class All beta. Domains d1gzca represents itself and other two structures (d1v6ia and d1dhkb), while d1nlsa forms another group of the family. Each structure consists of two segments A and B of eight strands. They can also be identified in the aligned core of BLOMAPS; the lengths corresponding to A and B are of 102 and 80 residues. Structure d1gzca is characterized as AB while d1nlsa as BA. The two segments cannot be related by a simple symmetry of rotation. The BLOMAPS alignment of the family is shown in Fig. (**4**). An aligning tool using dynamic programming can at most find the longer segment A as the core. Indeed, while N_c of BLOMAPS is 182, those of MUSTANG, MAMMOTH and MATT are 102, 101 and 103, respectively.

A comparison on the running time is summarized in Table **7**. BLOMAPS is the fastest. Furthermore, the running

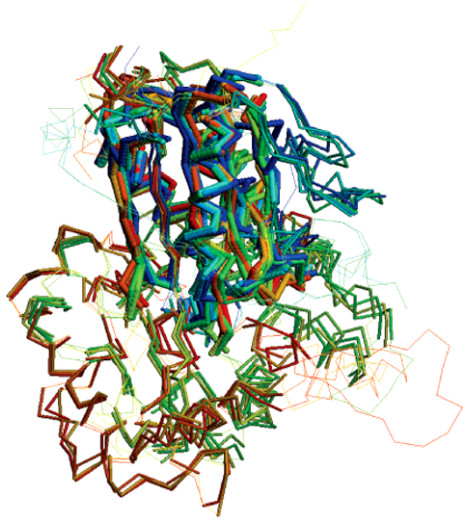

Fig. (3). BLOMAPS alignment of SCOP family Thiolase-related. The full core is shown with bold lines, and partial core with thick lines. Structures in groups I, II and III are colored in blue, green and red tones, respectively. (Pictures were drawn with RasMol [36]).

time for BLOMAPS grows linearly with the set size in contrast to the quadratic growth for the other three tools.

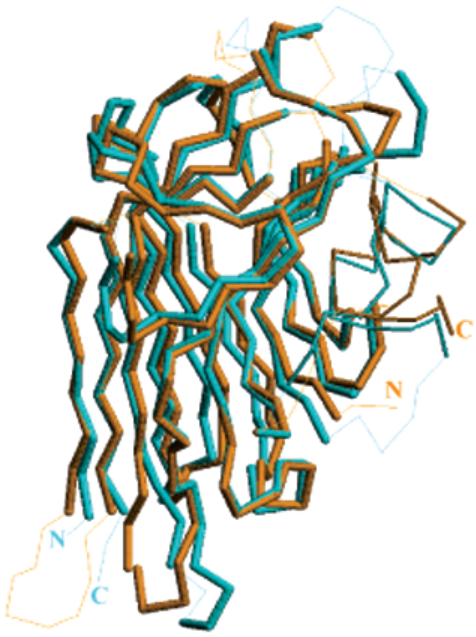

Fig. (4). BLOMAPS alignment of two representative structures 1gzca (brown) and 1nlsa (cyan) in SCOP family Legume lectins. The common core is shown with bold lines. The N- and C-termini of the two structures are indicated, showing a topological shift. (Pictures were drawn with RasMol [36]).

4. DISCUSSION

BLOMAPS distinguishes itself from most other existing algorithms for multiple structure alignment by its use of conformational letters. The description of 3D segmental structural states by a few discrete conformational letters gives a compromise between precision and simplicity. The substitution matrix CLESUM provides us with a proper measure of the similarity between these discrete states or letters. Such a description fits ε-congruent problems very well [17]. Furthermore, extracted from the database FSSP of structure alignments, CLESUM contains information of the structure database statistics. For example, scores between two frequent helical states are relatively low, which reduces the chance of accidental matching of two irrelevant helices. The conversion of coordinates of a 3D structure to its conformational codes requires little computation. Once we have transformed the 3D structures to 1D sequences of letters, tools for analyzing ordinary sequences can be directly applied. The use of conformational letters for a fast local similarity search can be integrated in many existing tools to improve their efficiency.

BLOMAPS is developed from our CLEPAPS, a tool for pairwise structure alignment, which is based on similar fragment pairs (SFPs) defined by CLESUM scores from string comparison. CLEPAPS takes a single good SFP as an initial correspondence, and iteratively builds up larger correspondence with ever stringent thresholds of 'zoom-in'. CLEPAPS adopts a greedy strategy guided by CLESUM scores. When dealing with the multiple alignment, we inevitably encounter the problem of combinatorial explosion. The concept of similar fragment pairs is extended to that of aligned fragment blocks. A multiple alignment has to satisfy the equivalency among the fragments of differen proteins inside an aligned fragment block, and, at the same time, the consistency among different aligned fragment blocks of the alignment. The strong restriction of both equivalency and consistency reduces the chance of making an insignificant alignment. Our heuristic to avoid the combinatorial explosion includes the using of a single pivot protein and HSFBs (instead of SFBs). Wrong assignment of correspondence can be detected, then removed, and replaced by a correct one in a later stage. In contrast with MATT, BLOMAPS uses a 'zoom-in' technique to go from local to global alignment. Thanks to the operation reduction to merely string comparison, greedy strategy guided by CLESUM scores and the cell-registration technique, the implement of BLOMAPS is considerably fast. At the same time, its alignment quality is competitive with other programs.

Contrary to CLEPAPS, for multi-alignment BLOMAPS excludes multiple solutions. Thus, it is encouraged to have a survey of the symmetry and repetition on the structures to be aligned, for example, by running a pairwise alignment on a randomly picked structure with itself. An inspection on conformational sequences can also be informative. In existence of a symmetry, the possibility of a cycle permutation should be examined. For repetition, we may

Table 7. Comparison of Running Time (in Units of s) Among BLOMAPS, MUSTANG, MAMMOTH and MATT on Different Sets

Test Set	BLOMAPS	MUSTANG	MAMMOTH	MATT
C2 (10 structures)	0.16	508.53	6.50	151.58
ASTRAL40-fam (averaged over 852 groups)	0.15	73.77	2.70	63.68
SABmark-sup (averaged over 415 groups)	0.26	167.87	3.90	97.88
First 5 structures of TIM61	0.20	46.20	4.31	137.42
First 10 structures of TIM61	0.40	181.19	16.97	718.05
First 15 structures of TIM61	0.61	450.25	36.97	1744.16
First 20 structures of TIM61	0.81	807.83	67.06	3193.61

mask the already aligned portion of a structure, and then have a second run of alignment. This technique of masking is also useful for detecting components with a domain move or translations and twists of MATT.

It should be emphasized that all the results presented above are derived with the same set of the default parameters listed in Table **2**. The tuning of parameters is generally crucial to an optimal performance of an algorithm. For example, for BLOMAPS a large value of fragment width l or similarity threshold T would reduce search times, but at the price of sensitivity. A too long l would leads to too few HSFBs. Similarly, a too high T would result in too sparse HSFBs. Without an enough number of dense HSFBs an efficient checking of vertical equivalency and horizontal consistency becomes impossible. Our strategy is to use moderately stringent parameters first for building a reliable primary scaffold for alignment, and then fill in the missing blanks for later compensation of the sensitivity loss with relaxed parameters. Under the assumption that structures to be aligned are independent of each other, a relatively low threshold T, which might be weak for a pairwise alignment, can still be significant for a multiple alignment.

The large ensembles we have tested are taken from the literature and all highly redundant. Extremely close structures can be detected rather reliably by merely aligning their conformational letter sequences. To conduct an alignment only for representative structures is more efficient and less biased.

CLESUM only considers information of conformation. However, the FSSP alignments from which CLESUM was derived also contain the amino acid information. The use of modified CLESUM matrices that also include such information would illuminate the biochemical role in alignment [10]. Recently, we have clustered amino acids into two groups *AVCFIWLMY* (type h) and *DEGHKNPQRST* (type p), and then obtained CLESUM matrices of types p-p, h-h and h-p [11]. We expect that such matrices would further improve the efficiency of our tools for structure alignments.

ACKNOWLEDGEMENTS

We are grateful to the authors of MAMMOTH-mult for providing their source code. This work is supported by National Natural Science Foundation of China and National Basic Research Program of China (2007CB814800).

REFERENCES

[1] H. Hasegawa and L. Holm, "Advances and pitfalls of protein structural alignment", *Curr. Opin. Struct. Biol.*, vol. 19, pp. 341-348, 2009.

[2] M.A. Marti-Renom, E. Capriotti, I.N. Shindyalov and P.E. Bourne, "Structure Comparison and Alignment", In: *Structural Bioinformatics*, 2nd ed. J. Gu and P.E. Bourne, Eds., John Wiley & Sons: USA 2009, pp. 397-417.

[3] L. Holm and C. Sander, "The FSSP database of structurally aligned protein fold families", *Nucleic Acid Res.*, vol. 22, pp. 3600-3609, 1994.

[4] L. Holm and C. Sander, "Dali/FSSP classification of three-dimensional protein folds", *Nucleic Acid Res.*, vol. 25, pp. 231-234, 1997.

[5] I.N. Shindyalov and P.E. Bourne, "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path", *Protein Eng.*, vol. 11, pp. 739-747, 1998.

[6] J. Ye and R. Janardan, "Approximate multiple protein structure alignment using the sum-of-pairs distance", *J. Comput. Biol.*, vol. 11, pp. 986-1000, 2004.

[7] D. Lupyan, A. Leo-Macias and A. R. Ortiz, "A new progressive-iterative algorithm for multiple structure alignment", *Bioinformatics*, vol. 21, pp. 3255-3263, 2005.

[8] O. Dror, H. Benyamini, R. Nussinov and H. Wolfson, "MASS: Multiple structural alignment by secondary structures", *Bioinformatics*, vol. 19, pp. i95-i104, 2003.

[9] O. Dror, H. Benyamini, R. Nussinov and H. Wolfson, "Multiple structural alignment by secondary structures: Algorithm and applications", *Protein Sci.*, vol. 12, pp. 2492-2507, 2003.

[10] W.M. Zheng and X. Liu, "A protein structural alphabet and its substitution matrix CLESUM", In: *Lecture Notes in Bioinformatics*, vol. 3680, C. Priami and A. Zelikovsky, Eds. Berlin: Springer Verlag, 2005; pp. 59-67.

[11] W.M. Zheng, "Protein Conformational Alphabets", In: *Protein Conformations: New Research*, L.B. Roswell, Ed. Nova Science Publishers, 2008, pp. 1-49.

[12] M. Menke, B. Berger and L. Cowen, "Matt: Local flexibility aids protein multiple structure alignment", *PLoS Comput. Biol.*, vol. 4, no. e10, pp. 88-99, 2008.

[13] L.P. Chew and K. Kedem, "Finding the consensus shape of a protein family", In: *18th Annual ACM Symposium on Computational Geometry*, ACM, 2002, pp. 64-73.

[14] C. Guda, S. Lu, E.D. Sheeff, P. E. Bourne and I. N. Shindyalov, "CE-MC: A multiple protein structure alignment server", *Nucleic Acids Res.*, vol. 32, pp. W100-W103, 2004.

[15] Y. Ye and A. Godzik, "Multiple flexible structure alignment using partial order graphs", *Bioinformatics*, vol. 21, pp. 2362-2369, 2005.

[16] A. Konagurthu, J. Whisstock, P. Stuckey and A. Lesk, "MUSTANG: A multiple structural alignment algorithm", *Proteins*, vol. 64, pp. 559-574, 2006.

[17] M. Shatsky, R. Nussinov and H. Wolfson, "MultiProt - A multiple protein structural alignment algorithm", In: *Lecture Notes in Computer Science*, vol. 2452, R. Guigo and D. Gusfield, Eds., Springer Verlag: Rome 2002, pp. 235-250.

[18] J. Ebert and D. Brutlag, "Development and validation of a consistency based multiple structure alignment algorithm", *Bioinformatics*, vol. 22, pp. 1080-1087, 2006.

[19] X. Liu, Y. Zhao and W.M. Zheng, "CLEMAPS: Multiple alignment of protein structures based on conformational letters", *Proteins*, vol. 71, pp. 728-736, 2008.

[20] S. Wang and W.M. Zheng, "CLEPAPS: Fast pair alignment of protein structures based on conformational letters", *J. Bioinform. Comput. Biol.*, vol. 6, pp. 347-366, 2008.

[21] P. Lackner, W.A. Koppensteiner, M.J. Sippl and F. S. Domingues, "ProSup: A refined tool for protein structure alignment", *Protein Eng.*, vol. 13, pp. 745-752, 2000.

[22] W.M. Zheng, "The use of a conformational alphabet for fast alignment of protein structures", In: *Lecture Notes in Computer Science*, Springer: USA, 2008, vol. 4983, pp. 331-342.

[23] M.J. Rooman, J.P. Kocher and S.J. Wodak, "Prediction of protein backbone conformation based on seven structure assignments: Influnce of local interactions", *J. Mol. Biol.*, vol. 221, pp. 961-979, 1991.

[24] B.H. Park and M. Levitt, "The complexity and accuracy of discrete state models of protein structure", *J. Mol. Biol.*, vol. 249, pp. 493-507, 1995.

[25] T. Edgoose, L. Allison and D.L. Dowe, "An MML classification of protein structure that knows about angles and sequences", In: *3rd Pacific Symposium on Biocomputing*, 1998, pp. 585-596.

[26] A.C. Camproux, P. Tuffery, J.P. Chevrolat, J.F. Boisvieux and S. Hazout, "Hidden markov model approach for identifying the modular framework of the protein backbone", *Protein Eng.*, vol. 12, pp. 1063-1073, 1999.

[27] B. Offmann, M. Tyagi and A.G. de Brevern, "Local protein structures", *Curr. Bioinform.*, vol. 2, pp. 165-202, 2007.

[28] R.W. Montalvao, R.E. Smith, S.C. Lovell and T.L. Blundell, "CHORAL: A differential geometry approach to the prediction of the cores of protein structures", *Bioinformatics*, vol. 21, pp. 3719-3725, 2005.

[29] P.L. Chang, A.W. Rinne and T.G. Dewey, "Structure alignment based on coding of local geometric measures", *BMC Bioinform.*, vol. 7, pp. 346-356, 2006.

[30] W. Kabsch, "A discussion of the solution for the best rotation to related two sets of vectors", *Acta Crystallogr.*, vol. 34A, pp. 827-828, 1978.

[31] S. Umeyama, "Least-squares estimation of transformation parameters between 2-point patterns", *IEEE Trans. Pattern Anal. Machine Intel.*, vol. 13, pp. 376-380, 1991.

[32] W.M. Zheng, "Relation between weight matrix and substitution matrix: motif search by similarity", *Bioinformatics*, vol. 21, pp. 938-943, 2005.

[33] A.G. Murzin, S.E. Brenner, T. Hubbard and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures", *J. Mol. Biol.*, vol. 247, pp. 536-540, 1995.

[34] J.M. Chandonia, G. Hon, N.S. Walker, L.L. Conte, P. Koehl, M. Levitt and S.E. Brenner, "The ASTRAL compendium in 2004", *Nucleic Acids Res.*, vol. 32, pp. D189-D192, 2004.

[35] I.V. Walle, I. Lasters and L. Wyns, "SABmark --- a benchmark for sequence alignment that covers the entire known fold space", *Bioinformatics*, vol. 21, pp. 1267-1268, 2005.

[36] R.A. Sayle and E.J. Milner-White, "RasMol: Biomolecular graphics for all", *Trends Biochem. Sci.*, vol. 20, pp. 374-376, 1995.

Received: September 09, 2009 Revised: October 09, 2009 Accepted: October 09, 2009