PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

CLEMAPS: Multiple alignment of protein structures based on conformational letters

Xin Liu,[1] Ya-Pu Zhao,[1] and Wei-Mou Zheng[2]*

[1] Institute of Mechanics, Chinese Academy of Sciences, Beijing 100080, China
[2] Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100080, China

ABSTRACT

CLEMAPS is a tool for multiple alignment of protein structures. It distinguishes itself from other existing algorithms for multiple structure alignment by the use of conformational letters, which are discretized states of 3D segmental structural states. A letter corresponds to a cluster of combinations of three angles formed by C_α pseudobonds of four contiguous residues. A substitution matrix called CLESUM is available to measure the similarity between any two such letters. The input 3D structures are first converted to sequences of conformational letters. Each string of a fixed length is then taken as the center seed to search other sequences for neighbors of the seed, which are strings similar to the seed. A seed and its neighbors form a center-star, which corresponds to a fragment set of local structural similarity shared by many proteins. The detection of center-stars using CLESUM is extremely efficient. Local similarity is a necessary, but insufficient, condition for structural alignment. Once center-stars are found, the spatial consistency between any two stars are examined to find consistent star duads using atomic coordinates. Consistent duads are later joined to create a core for multiple alignment, which is further polished to produce the final alignment. The utility of CLEMAPS is tested on various protein structure ensembles.

Proteins 2008; 71:728–736.

Key words: protein structure; multiple structural alignment; protein conformational alphabet.

INTRODUCTION

Protein folds into specific three-dimensional structure to fulfill its function.[1] The comparison of protein structures plays a significant role in our understanding of the organization of life. The detection of local or global structural similarity between a new protein and a protein with a known function allows the prediction of the new protein's function. Since protein structures are better conserved than amino acid sequences, remote homology is detectable more reliably by comparing structures. Structural comparison methods are useful for organizing and classifying known structures, and for discovering structure patterns and their correlation with sequences.

Protein structure comparison is most often performed by a protein structure alignment program, and many tools have been developed for this. Despite the existence of various pairwise structural alignment algorithms and several methods of multiple alignment, efficient and reliable algorithms for multiple alignment are in ever increasing demand for analyzing the rapidly growing data of protein structure. 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 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.[2,3] Such pairwise-based methods have the limitation that alignments that are optimal for the whole input set might be missed. There are a handful of truly multiple methods.[4–6] Many multiple alignment tools often start with sets of structurally common fragments extracted from as many as possible input proteins, and then combine them into a global common substructures. For example, in doing this, MASS implements a two-level alignment, using both secondary structure and atomic representation.

Local similarity is a necessary condition for the global structural alignment, but insufficient. Structurally similar fragments first found in different proteins by seed matches form the basis objects for further examination of their consistency in the spacial arrangement required by the global alignment. Consistent pieces then may be joined to obtain the global alignment. Different methods use various criteria and strategies for seed matching, consistency checking and

The Supplematerial Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/
Grant sponsor: National Natural Science Foundation, China; Grant sponsor: National Basic Research Program of China; Grant number: 973 program, 2007CB310504.
*Correspondence to: Wei-Mou Zheng, Institute of Theoretical Physics, Academia Sinica, Beijing 100080, China. E-mail: zheng@itp.ac.cn
Received 26 June 2006; Revised 18 June 2007; Accepted 21 July 2007
Published online 2 November 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21739

pieces merging. Generally, a stringent criterion for local similarity would create fewer seed matched objects, and hence speed up the merging process. However, it would miss some substructures constituting the final global alignment. On the other hand, because of the insufficiency of the local similarity to the global alignment, a loose condition of local similarity would overload the later filtering task. One has to balance sensitivity with specificity, and compromise efficiency with accuracy.

Protein structural alignment involves the geometric representation of structures. In most cases, only the backbone of pseudobonds formed by C_α atoms are considered. Coordinates of C_α atoms, which change under translation and rotation in 3D space, are not geometric invariants. Distances used by DALI are the intrinsic property of a geometric object.[7,8] The bending and torsion angles of pseudobonds, as the chain counterparts of curvature and torsion of a smooth curve, are also geometric invariants. VAST and MASS replace secondary structure elements by the vectors of their axes,[9,10] but this representation is not very accurate for structural elements.

Another way to represent structures is to use conformational alphabets, which are discretized conformational states of certain fragment units of protein backbones.[11–15] We use the bending and torsion angles of pseudobonds to describe protein backbones. The smallest unit possessing one-to-one correspondence between angles and coordinates is the quadrupeptide unit, which admits two bending angles and one torsion angle. Our conformational alphabet of 17 letters is obtained by clustering based on the probability distribution in the phase space spanned by the three angles. The description by conformational letters provides a good balances between accuracy and simplicity, and converts a 3D structure to a 1D sequence of letters. Substitution matrices such as the popular PAM and BLOSUM are essential to amino acid sequence alignment algorithms. Without a conformational substitution matrix the use of a conformational alphabet is very limited. To implement fast structural comparison in terms of conformational alphabets, we have derived a substitution matrix of conformational letters called CLESUM from a representative pairwise aligned structure set of the FSSP (families of structurally similar proteins) database of Holm and Sander.[7,8] It has been verified that CLESUM aptly measures the similarity between the conformational letter states.[15]

Here we report a tool called CLEMAPS developed for fast multiple alignment of protein structures by fully using our conformational alphabet and its substitution matrix CLESUM. We demonstrate its utility with several types of protein ensembles.

METHODS

The input of the algorithm is m protein structures: P_1, P_2, …, P_m. First, the coordinates $\{\mathbf{r}_{\kappa i}\}$ of C_α atoms of each protein P_κ are converted to its sequence S_κ of conformational letters. Since each letter corresponds to a quadrupeptide unit, the length of S_κ is shorter than that of P_κ by three. By convention, we assign the first letter to the third residue, the second to the fourth and so on, until finally the last letter is assigned to the last residue but one.

In our algorithm CLEMAPS, a scaffold comprising aligned fragments of a fixed width or blocks is first built for multiple alignment. The scaffold is then refined into the core of final multiple alignment. The members of a block share common local similarity, each from a distinct protein, and the total number of members, that is, the block size, may be smaller than m, the size of the input set. We assume that the multiple alignment contains at least two blocks. It is obvious that any two blocks are consistent in their spatial arrangement. That is, if fragments a_1, b_1 are in block 1, and a_2, b_2 in block 2 with a_1, a_2 being from protein A, and b_1, b_2 from B, one can superpose a_1, a_2 on b_1, b_2 to make them ε-congruent.[4] A fragment set of common local similarity from distinct proteins without considering the spatial consistency with other fragments may be used to initiate the search for a block. Such a set will be defined as a *star*. The algorithm starts with finding stars consisting of structurally similar fragments from different proteins. We illustrate some main concepts of CLEMAPS in Figure 1, and explain its main steps in the following subsections.

Finding stars by the center-star approach

For a given block it is natural to assume that there exists a member in the block, which shares the greatest similarity with all the other members. Using this heuristic, we iteratively choose a string of width l from every sequence S_κ as a seed to search in all the other sequences for the string most similar to the seed. We use CLESUM to score the pair similarity. A threshold T_0 is used to filter out insignificant fragmental similarity. We define the score for seed $s_{\kappa,i}\ s_{\kappa,i+1}\ \ldots\ s_{\kappa,i+l-1}$ of protein κ as

$$\Sigma = \sum_{\nu=1}^{m'} \Theta\left(\max_{1\le j\le n_\nu - l + 1} \sum_{k=0}^{l-1} C\left(s_{\kappa,i+k}, s_{\nu,j+k}\right) - T_0 \right), \quad (1)$$

where the prime at the summation sign denotes that the summation excludes sequence κ, $C(x, y)$ is the xy entry of CLESUM, and $\Theta(x)$ is the step function with $\Theta(x) = x + T_0$ for $x \ge 0$ and $\Theta(x) = 0$ otherwise.

By thinking in graph theory, the seed from sequence S_κ is the center node. If the most similar string in sequence S_ν is found with a similarity to the seed above T_0, the string is a neighbor node of the seed, and an edge is linked between it and the seed node or the center. The center and its neighboring nodes form a center-star, which may be simply called a star, and is scored by Σ, the sum of pairwise similarity with respect to the center

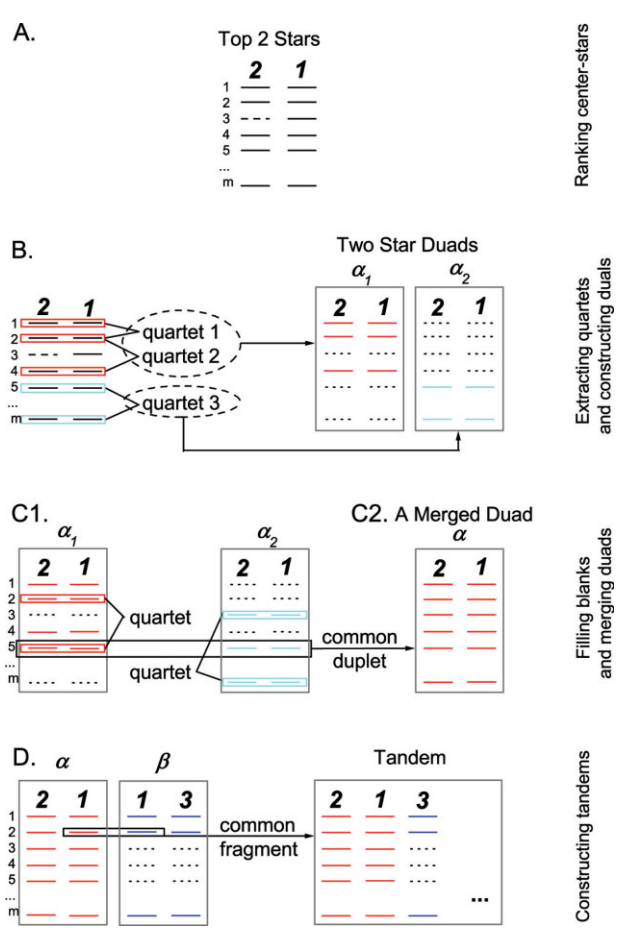

Figure 1

Illustration of CLEMAPS. Non-overlapping center-stars are found in m *proteins, and ranked according to their CLESUM scores in descending order. The top two stars (**A**) are indexed by their ranks. Each line represents a protein. The dashed segment in the line representing protein 3 indicates the absence of a neighbor of star 2 in protein 3. By examining the consistence conditions, quartets are found for center-stars 1 and 2. Three quartets consisting of five duplets are grouped into two star duads* α_1 *and* α_2 *(**B**). Under less stringent conditions, a member of star 2 is found in protein 3, and two quartets are found, one between proteins 2 and 5, and the other between 3 and* m *(**C1**). (Unrelevant fregments are represented by dotted segments.) The duplet of protein 5 is now shared by both duads* α_1 *and* α_2*, and hence leads to the mergence of the two duads (**C2**). Star duads may be further joined into tandems. Joining duad* α *of stars 1 and 2 with duad* β *of stars 1 and 3 through the common fregment of protein 2 in star 1 results in a part of a tandem (**D**).*

seed. The center-star approach has been used to search for motifs in amino acid sequences and in MultiProt.[4,16]

After finding all center-stars, we sort them according to their scores in descending order. We retain the top N stars for further processing.

Finding consistent star duads

Consider two stars found with the center-star approach. They contain fragments a_1, a_2 of protein A and b_1, b_2 of B, where 1 and 2 are indices of the two stars. We examine the spatial consistency between (a_1, a_2) and (b_1, b_2). If the following three conditions are all satisfied, we say that (a_1, a_2) and (b_1, b_2) are consistent duplets, and form a quartet (a_1, a_2; b_1, b_2).

1. Fragments a_1 and a_2 do not overlap, and neither do b_1 and b_2.
2. Local distance similarity $D(a_1, b_1) < D_0$, and $D(a_2, b_2) < D_0$. Distance similarity of fragments x and y is defined as

$$D(x,y) = \frac{1}{l+3}\sum_{i=1}^{l+3}\left|d_{ic}^{x} - d_{ic}^{y}\right|, \qquad (2)$$

   where d_{ic}^{x} denotes the distance between residue x_i of fragment x and the mass center x_c of x; similarly for d_{ic}^{y}. Notice that the number of residues is greater than the string width by three.
3. Distance consistency $R(a_1, a_2; b_1, b_2) < R_0$, where $R(x, y; u, v)$ is defined as

$$R(x,y;u,v) = \frac{1}{(l+3)^2}\sum_{i=1}^{l+3}\sum_{j=1}^{l+3}\left|d_{x_i,y_j} - d_{u_i,v_j}\right|. \qquad (3)$$

Here the distance similarity condition of D is a refinement of the pair similarity of CLESUM. A rough criterion for R is obtained by looking at fewer atoms, say one in every four. In this way, the total number of distance difference terms for summation reduces to about just one-sixteenth.

The members of a quartet are its two duplets. A star *duad* is defined as a union of quartets. A member of a star duad is a fragment duplet (a_1, a_2) from the single protein A. Each protein contributes to a duad at most one duplet. Two duplets (x, y) and (u, v) of the quartet (x, y; u, v) must belong to the same duad. For any member (x, y) of a duad, there must exist at least one member (u, v) in the duad, which together with (x, y) forms the quartet (x, y; u, v). According to this definition of star duad, starting with the top-1 star as the first, we iteratively pick the next star in succession from the descending list of stars to examine the quartet conditions. When the first quartet is met, the construction of the first duad is initiated. Members of a single star need not appear in a single duad. From two given stars, which admit at least one quartet, we often obtain several duads after examining the consistency conditions.

Taking stars one by one according to the star order, we examine the quartet conditions between the fragments in a newly chosen star and those in the earlier found duads to find new quartets and then new duads. Assume that we have examined up to the top-K star, and obtained k duads. We next examine the quartet conditions between the fragments in the K + 1st star and those in the k duads for new duads. It may happen that a newly found

duad shares a common duplet with an earlier duad. In this case we should merge the two incomplete duads into one (by taking the union). If a duplet in the newly found duad conflicts with a duplet in an earlier duad during merging, only the earlier one is kept.

We continue with this procedure for finding new duads until all top N stars have been examined for consistency. After finishing the search, we sort the duads in descending order according to their depth, that is, the number of duplets.

Filling blanks in duads

The approach for constructing a center-star is greedy since only the string in a sequence with the greatest similarity to the center seed is kept. The greedy approach and stringent parameters of T_0, D_0, and R_0 help to accelerate the finding of high quality quartets, which play the role of a scaffold for further global alignment.

In the ideal case, a duad should be a subset of a block pair in the final multiple alignment. Compared with the block pair, the duad could fail to spot some duplets under stringent conditions. Such duplets are called missing blanks. Another case is when the block pair does not contain any duplets from some proteins of the input set. These are genuine blanks. With the aid of the already found quartets, we may fill in the missing blanks in duads according to their depth order in a less greedy way.

To fill in the blanks in a duad, we take each duplet of the duad as a seed. In each sequence which so far has no duplet in the duad, we find all the string pairs which are neighbors of the seed (with fragmental similarity above threshold T_0), and sort them according to the similarity sum in descending order. We examine the quartet conditions between the seed and its neighbors one at a time. Once a quartet is found, we may fill the blank using the quartet and continue. If no quartet is found, we cannot fill in the blank as yet. Blank filling may result in duad merging.

After we have completed filling in the blanks for all the duads, we replace parameters T_0, D_0, and R_0 with the less stringent T_1, D_1, and R_1, and perform a second run of blank filling. If necessary, a third run can be further taken using even looser parameters.

Global extension and final refinement

If two duads share any common fragments we may join them to form a tandem. More duads can be further added to an already existing tandem with the aid of shared fragments to make the tandem grow. The final duad tandem forms an approximation to the core for alignment.

Each duad defines a multiple correspondence among residues of different proteins. Fragments of two given proteins extracted from a scaffold determine a correspondence between selected atoms of the two proteins, from which a rigid transformation can be determined to minimize the root mean square deviation (RMSD) distance of the correspondence.[17] The RMSD of two corresponding point sets $\{\mathbf{r}_i\}_1^n$ and $\{\mathbf{t}_i\}_1^n$ is defined as

$$\mathrm{RMSD}(\{\mathbf{r}_i\},\{\mathbf{t}_i\}) = \left(\frac{1}{n}\sum|\mathbf{r}_i-\mathbf{t}_i|^2\right)^{1/2} = \left(\frac{1}{n}\sum d_i^2\right)^{1/2}. \quad (4)$$

We introduce a distance threshold δ for the deviation between a pair of corresponding points. If the deviation of a point pair is greater than δ we remove the pair from the correspondence list. For a pair of corresponding fragments from two proteins, we may also calculate the deviation between their flanking site pairs one by one on both ends. If the deviation of a point pair is smaller than δ, we add the pair to the correspondence list to extend the aligned fragment pair.

Collecting all duads of the same greatest depth and forming a subcore from them, we may determine a structural templet as follows. Taking the first protein in the subcore as the initial templet, we determine the rigid transformation for each of the remaining proteins that minimizes the RMSD of the atom sets of correspondence between the protein and the templet. The averages of the transformed coordinates define the first update version of the templet. The convergent templet obtained by iteration of this procedure is the final templet. We may elongate the fragments of the templet by calculating the average of the flanking site positions with the identified rigid transformations. Taking the templet as the pivot and using the δ criterion, we update the list of global correspondence. We may then calculate the RMSD of each structure with respect to the templet, as well as the average RMSD for the multiple alignment.

A refined templet may also be constructed for a particular subset of the input proteins. Such a templet serves as a guide for finding the optimal alignment specific to the subset.

RESULTS

We have applied CLEMAPS to three protein ensembles. The PDB codes of the three sets are listed in Table I. All the experiments were performed on a personal computer (Pentium IV 1600 MHz processor with 512MB RAM) with a SuSE Linux 7.3 operating system. The ensembles cover various challenging cases of structural alignment. Ensemble 1 contains structural homologies at different levels, ensemble 2 exhibits different topologies, while ensemble 3 contains a large number of proteins. The empirically recommended parameters for CLEMAPS are $l = 8$, $T_0 = 19$, $D_0 = 0.6$ Å, $R_0 = 0.8$ Å, $T_1 = 11$, $D_1 = 2.3$ Å, $R_1 = 2.1$ Å, and $\delta = 6$ Å. Parameters D, R, and δ are almost independent of l while T is l-dependent.

Table I
Three Ensembles Used to Test CLEMAPS

Ensemble name	Proteins
CL-GL	1ak6, 1cfyB, 1cnuA, 1f7s, 1svy, 2vik, 1a0nA 27-152,1a0nA 153-262, 1a0nA 263-383,1a0nA 384-532, 1a0nA 533-628, 1a0nA 629-755
C2 domain	1a25A, 1bdyA, 1d5rA, 1dsyA, 1e8yA, 1gmiA, 1qasA, 1rlw, 1rsy, 3rpbA
Serine Proteinase	1agjA, 1agjB, 1arb, 1arc, 1boqA, 1csoE, 1ct0E, 1ct2E, 1ct4E, 1ds2E, 1exfA, 1gbaA, 1gbbA, 1gbcA, 1gbdA, 1gbeA, 1gbfA, 1gbhA, 1gbiA, 1gbjA, 1gbkA, 1gblA, 1gbmA, 1hpgA, 1ky9A, 1ky9B, 1lcyA, 1p01A, 1p02A, 1p03A, 1p04A, 1p05A, 1p06A, 1p09A, 1p10A, 1p11E, 1p12E, 1qq4A, 1qrwA, 1qrxA, 1qtfA, 1sgc, 1sgpE, 1sgqE, 1sgrE, 1sgt, 1tal, 2alp, 2lprA, 2sfa, 2sga, 2sgpE, 2ull, 3lprA, 3proA, 3proB, 3sgaE, 3sgbE, 4proA, 4proB, 4sgaE, 4sgbE, 5lprA, 5sgaE, 6lprA, 7lprA, 8lprA, 9lprA

The first four letters of an entry are a protein name in the PDB code, followed by optional items: chain id, and the sites of the first and the last residue.

Generally, to fit gaps in alignment a shorter l is preferable, while a longer l is advantageous in reducing search space. Normally, a higher T should be used for a longer l.

Subset alignment detection

Ensemble 1 consists of 12 sequentially nonredundant structures belonging to the fold "Actin depolymerizing proteins". According to the SCOP database,[18] this fold contains four structures in the Cofilin-like (CL) family (PDBid: 1ak6, 1cfyB, 1cnuA, and 1f7s) and eight structures in the Gelsolin-like (GL) family (PDBid(chain sites): 1d0nA 27–152, 1d0nA 153–262, 1d0nA 263–383, 1d0nA 384–532, 1d0nA 533–628, 1d0nA 629–755, 1svy, and 2vik). The ith segments of the six segments of 1d0nA will be denoted by adding an extra subscript as $1d0nA_i$ ($i = 1, 2, \ldots, 6$). The two families share a central five-stranded β-sheet substructure of the form BACDE that is flanked between two α-helices: long helix (α_1) between strands D and E and short helix (α_2) in the C terminus. There are two additional α-helices in the CL family: an N terminus helix, and a short helix between strands B and C. The two families are related structurally but not sequentially.[19,20]

The common alignment of the MASS structural alignment of all 12 proteins consists of 28 residues with an RMSD of 1.9 Å. Strands A, C, D, E, and helix α_1 are structurally conserved. Strand B is only partially conserved due to a slight twist. The CLEMAPS alignment generally agrees with that of MASS. At a weaker RMSD of 2.2 Å, the common alignment consists of 71 residues. This is very close to the alignment of CE-MC.[21] Compared with MAMMOTH-mult, at the similar RMSD, the core of MAMMOTH-mult alignment is a little longer. There is no clearly defined unique way to evaluate the quality of protein structure alignments. Since different criteria are used no simple direct comparison exists. For example, a high RMSD would lead to a large number of equivalent residues. Methods without an extra restriction in the minimal size for aligned segment usually have a larger number of residues in the core of alignment than others. To make a direct comparison, we calculate the numbers of aligned residues at different cutoffs δ. Both MAMMOTH-mult and CE-MC alignments provide the list of correspondence among residues of various proteins. Using the procedure described in the earlier sections we generate a structure templet from the correspondence list. After aligning every structure against the templet, we count the numbers of aligned residues which have their distances from the templet below different cutoffs after superposition. A comparison of CLEMAPS with MAMMOTH-mult and CE-MC by these counts of aligned residues is given in Table II. (In a more careful way of counting residues we have to discard the residues with large deviations from the list of correspondence, rebuild the templet, and then re-count the number.)

For this ensemble the top-1 star (of score sum 375 at $l = 8$) is found with seed AIGCAHII starting at site 686

Table II
Comparison of CLEMAPS with MAMMOTH-mult and CE-MC: Counts of aligned residues as a function of the deviations from the templet of alignment after superposition

	Ensemble CL-GL			Ensemble C2		
Deviation (Å)	CLEMAPS	CE-MC	MAMMOTH-mult	CLEMAPS	CE-MC	MAMMOTH-mult
<1	274	291	327	519	477	215
1–2	384	469	432	252	326	176
2–4	288	376	334	149	233	448
<4	946	1136	1093	920	1036	839
4–6	68	80	154	34	75	209
>6	12	24	240	7	24	1207

```
aa vELSKKVTGKLDKttPGIQIWRIENMEMVPVPTKSY--------GNFYEGDCYVLLSTrktgsgfs
ss cccCCCCCCCCCcccceeEEEECHHHCcceeccccc--------ccccCCCCEEEEEeeeccccee
cl   OGNIKNJNPLEngELDEEDEPGNOGCECBEAJKG--------KMGCPKLDEEEEEEeceailqd
    FCCAJJJJIKJK  LBDDEEECCQNKCGDECBLCG LCGAJJ JMGCAKLDFDEEEA

aa YNIHYWLGKNSSQDEQG AaAIYTTQMDeylgsvAVQHREVQG HESETFRAYFkqgliykqggvasgmk
ss eeeeeeecCCCCHHHHH HHHHHHHHHhhhcccCCCEEEEECc CCHHHHHHHcccccccccccccccc
cl EEEEEEBPOMLCAHHHH IiHHHHHHIJkijoomGFDEDEEAK MKGNJIIJIMplqdeeeccnpjmml
   CLEEBEBEAIGCAIIIHH  IHIHIHII      GCFEEEDCNM GCAJIIJJM
```

Figure 2

The alignment of 2vik to the templet of alignment. Markers "aa", "ss" and "cl" indicate the sequences of amino acids, secondary structures and conformational letters, respectively. An additional line shows the consensus of the alignment, where the total CLESUM pair scores of each letter from the alphabet of 17 letters to all conformational letters of a given aligned residue column is calculated, and the one with the highest sum is taken as the representative letter for the column consensus. Lowercase letters of amino acids or gaps ("–") indicate structural nonequivalence. Uppercase letters of secondary structures indicate the fragments before refinement. When two such fragments or aligned elements are concatenated one to the other, a space is added to separate them.

of protein 1a0nA (residue 58 of $1a0nA_6$). The star has its members in all proteins, and all positions of the members are correct with respect to the final alignment. Furthermore, stars which are a shift of the top-1 star as a whole also have high scores. This implies that the aligned segment could be longer. Indeed, the final width elongated from the star is 15. Width $l = 9$ was used for this ensemble. In fact, in a wide range of $l \geq 8$ top-1 stars share the same kernel, which is the top-1 star of $l = 8$.

CLEMAPS is able to find subpatterns not shared by the whole ensemble. The final refined templet for the multiple alignment of the whole ensemble comprises eight elements, which are indexed from 1 to 8 according to the order from the N to C terminus. All the four CL proteins contain the first seven elements with element 8 missing, (having the common pattern 1234567,) while the GL family is characterized by the common pattern of 245678. In the GL family, element 1 is found only in $1d0nA_1$, 2vik, and element 3 only in $1d0nA_3$. As an example, the alignment of 2vik to the templet is shown in Figure 2.

To have a close comparison among CLEMAPS, CE-MC, and MAMMOTH-mult, we examine the overlaps among their alignments. Three (elements 4, 5, and 8) out of the above eight structural elements are common to all the three methods; they cover 456 residues. Elements 2,6, and 7 are easily identifiable within a shift up to two residues, and the proportions of shifted structures are 3/12, 1/12, and 1/12, respectively.

Pattern detection in proteins with different topology

It may happen that even though some proteins are quite different in their sequences and spatial arrangement of structural elements, their 3D structures as a whole are surprisingly similar. They could exhibit nontopological similarities in structure alignment, where the order of polypeptide fragments does not follow their linear order in sequences.[22]

Ensemble 2 consists of 10 proteins, four "Synaptotagmin-like" proteins and six "PLC-like" proteins, taken from two families of the "C2 domain" superfamily. The two families are related by a circular permutation while each forms a topological group.

Width $l = 9$ was used for this ensemble. The top-1 star is obtained by using seed PGLDFNGCC of protein 1rsy. Among the 10 members of the star, only seven remain in the final alignment, and the other three (of proteins 1bdyA, 1d5rA, 1e8yA) are from a wrong position. These wrong members are removed later by examination of the quartet conditions. Such situations happen often when the substructure of a seed is dominanted by regular secondary structures. The wrong members are removed by examination of the consistency conditions. The blanks left by removing the inconsistent members are later filled by a less greedy search.

The CLEMAPS alignment consists of nine elements indexed from 1 to 9. The model alignment for the four "Synaptotagmin-like" proteins is 123456789 in the element indices while that for the six "PLC-like" proteins is 234567891. That is, element 1 is located in the N terminus for the former, but in the C terminus for the latter. Furthermore, in the "PLC-like" family, 1bdyA has elements 4, 5, 7 missing, 1d5rA has 4, 5, 9 missing, and 1e8yA has 7 missing. As an example from the "Synaptotagmin-like" family, the alignment of 1a25 to the templet is shown in Figure 3 together with the consensus for the nine elements of the alignment.

MASS conducts no dynamic programming. The nontopological alignment of this ensemble was detected by MASS.[5] The core of the MASS alignment consists of 58 residues within an RMSD of 1.9 Å and forms a sandwich of eight β-stands. The alignments of MASS and CLEMAPS

```
aa erRGRIYIQAHIDR EVLIVVVRDAKNLVPMDpngLSDPYVKLKLIPDpkseSKQKTKTIKC
ss ccccEEEEEEEECc ceeeEEEEEEECCccccccccCCEEEEEEEeccccccCEECCCCCC
cl   KCQDDEEEEDEN NGEEEEBBLDDNGCCBLajoGDDFECEEEDQNGpiomJGEFDCCECN
     LQDDEEDEEDBN CBDEEEBGLDFNGCCPL   GDBFEDEEEDQNG    MGEFDCCECN

aa SLNPEWNETFRFqlkesdkDRRLSVEIWD WDLTSRnDF MGSLSFGISElqkAGVDGWFKLLSqeegeyfnv
ss ccCEEEEEEEeeccchhhhCEEEEEEEE CCCCCCCEE EEEEEEEHHhhccccEEEEEECCChhhhccccc
cl GBKEEBPGDEEEefeahjjHGDDEEEEED CEBKLCcCC BLDEFDCAHHihhLDFELQECCFCaihjjkld
   GPKEEBPGEEDE       KGFDEEEEE DCEPKLC  CCPLDDDECAJJ   LQEEDDECCFC
```

Figure 3

The alignment of 1a25 to the templet. Markers "aa", "ss" and "cl" indicate the sequences of amino acids, secondary structures and conformational letters, respectively. An additional line shows the consensus of the alignment. Lowercase letters of amino acids indicate structural nonequivalence. Uppercase letters of secondary structures indicate the fragments before refinement.

agree each other on the whole, but the CLEMAPS core is longer (63) within a smaller RMSD (1.7 Å). Both CE-MC and MAMMOTH-mult are designed for detecting a preserved topology. The core of CE-MC alignment is roughly 123456789. The element 1 of the six "PLC-like" proteins, which is nontopological with respect to the core, is missing from the alignment. Except for this, the alignments of CLEMAPS and CE-MC are rather close. The core of MAMMOTH-mult alignment is roughly 1234567891. The same element 1 of the six "PLC-like" proteins and the four "Synaptotagmin-like" proteins is treated as two distinct elements. The core of the MAMMOTH-mult alignment is shorter than that of CLEMAPS or CE-MC for this ensemble. The alignment of 1e8yA by MAMMOTH-mult shows no overlap with that by either CLEMAPS, or CE-MC, or MASS. A detailed comparison of CLEMAPS with CE-MC and MAMMOTH-mult is also given in Table II.

Large-scale structural alignment

Ensemble 3 is taken from Ref. 5. It comprises 68 molecules of the SCOP family "Prokaryotic trypsin-like serine protease". The discretization of continuous 3D conformational states into a few letters reduces the computational cost for locating local similarity enormously. CLEMAPS won an easy success in fast aligning the large ensemble (27 s vs. 85 min of MASS). Aligning this ensemble was not acceptable to the MAMMOTH-mult web server due to its large size.

Using width $l = 11$, the CLEMAPS alignment for this ensemble is found to consist of 11 elements, which are indexed with digits from 0 to 9, and an extra letter *a*. The templet of the alignment comprises 138 residues with an RMSD of 0.79 Å. Elements 1, 5 and *a* are shared by all members. They form the common core of 49 residues with an RMSD of 1.1 Å. The least common elements are 4 and 8, each of which is shared by only 46 proteins. In the 68 proteins, 39 share the full pattern of the multiple alignment, and only two proteins contain no elements other than 1, 5, and *a*. As an example, the alignment of 1boqA to the templet is shown in Figure 4 together with the consensus for the eleven elements of the alignment.

Figures of structural alignment for these three ensembles are given in the supplementary material.

DISCUSSION

CLEMAPS distinguishes itself from 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 compromises 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. Furthermore, CLESUM extracted from the database FSSP of structure alignments 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.

CLEMAPS in one or the other respect resembles some other algorithms, such as the center-star approach used by MultiProt, the ordered pairs of secondary structure elements used by MASS. The use of conformational letters for a fast local similarity search can be integrated in many existing tools to improve their efficiency.

Another greedy strategy of CLEMAPS is to take into account only the binary consistency when constructing global alignment. Considering any two proteins in the input set and regarding a duplet of a duad as a node of a bipartite graph, the alignment between the two proteins

```
aa ..slCSVGFSVTRGatKGFVTAGHCGTVNA TARIGGAVvGT FAARVFPGNDRAwvsltsaq
ss ..eeeECCEEEEEECCEeEEEECHHHCCCcc EEEECCEEEEE EEEEECCCCCEeeeeecccc
cl ..qfEDCELDEDBNngEEDCECAJMLDCNG CEEENOGCbLD CPLEDEQKLELEdeefcajg
      EDCBLDEDBN  FEDDEQAJMLDCNGC EEENOGC  LDCPLEDEQNLELE

aa tllPRVANGSSFVTV RGSTEAAVGAAVCRSGRTtgYQCGTITAKNI TANYAEGAVRGltqg
ss eeeeEEEECCEEEEC CCCCCCCCCCEeeeccccccCEEEEEEEEEE EEEECCEEEEEeeee
cl eeeAGDCBCNGDDCC PLCBCCCCNGCEDDDEBKkfQEFEDCBLDNG EEEEAILQEFNgcfe
      AGDCBCNGDD CCBLCBCCCCKGCEDEEEAK  QEFEDCPLDNGE EBEAILQEFN

aa nacMGRGDSGGSWITSAgQAQGVMSGGNVQsngnncgipasqRSSLFERLQPILSQyglslvtg
ss cccccCCCCCCEEECCcccEEEEEEEECCCccccccccchhhcCEEEEEHHHHHhhccceeccc
cl bmcEDCKMGNGCEEDAJoGCBLDCBLFAGFajolbmjlcajiGBBFEDCPMIHHIIkogecce
      EDCKMGNGCBEDAJ GCBLDCBLFAQF            GBBFEDCPMIHHII
```

Figure 4

The alignment of 1boqA to the templet for the alignment of 68 proteins. Markers "aa", "ss" and "cl" indicate the sequences of amino acids, secondary structures and conformational letters, respectively. An additional line shows the consensus of the alignment. Lowercase letters of amino acids indicate structural nonequivalence. Uppercase letters of secondary structures indicate the fragments before refinement.

corresponds to a maximal clique in the bipartite graph.[9] In the worst case, the binary constrains would correspond only to a chain (the so-called "sausage effect"). For multiple alignment the situations where the sausage effect happens simultaneously at each pair of proteins should be rare. Wrong assignment of correspondence can be detected and then removed or corrected in a later stage after a rigid transformation.

The tuning of parameters used by CLEMAPS is crucial to its optimal performance. A large value of basis width l or similarity threshold T would reduce search times, but at the price of sensitivity. Our strategy is to use stringent parameters first for building reliable scaffold of the alignment core, and then fill in the missing blanks for later compensation of the sensitivity loss at relaxed parameters.

There is scope for further improvement in our approach. For example, a dynamic width l may be used by joining nearby high scored local alignments as DALI does. An alternative is to use first a large l, and then a small l.

CLESUM only considers information of conformation. However, the FSSP alignments from which CLESUM was derived also contain the amino acid information. The use of a modified CLESUM that also includes such information would illuminate the biochemical role in alignment.[15]

REFERENCES

1. Branden C, Tooze J. Introduction to protein structure, 2nd ed. New York: Garland; 1999.
2. Ye J, Janardan R. Approximate multiple protein structure alignment using the sum-of-pairs distance. J Comput Biol 2004;11:986–1000.
3. Lupyan D, Leo-Macias A, Ortiz AR. A new progressive-iterative algorithm for multiple structure alignment. Bioinformatics 2005;21: 3255–3263.
4. Shatsky M, Nussinov R, Wolfson H. MultiProt—a multiple protein structural alignment algorithm. In: Guigo R, Gusfield D, editors. Lecture notes in computer science 2452. Rome: Springer Verlag; 2002. pp 235–250.
5. Dror O, Benyamini H, Nussinov R, Wolfson H. MASS: multiple structural alignment by secondary structures. Bioinformatics 2003; 19 (Suppl 1):i95–i104.
6. Dror O, Benyamini H, Nussinov R, Wolfson H. Multiple structural alignment by secondary structures: algorithm and applications. Protein Sci 2003;12:2492–2507.
7. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acid Res 1994;22:3600–3609.
8. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acid Res 1997;25:231–234.
9. Madej T, Gibrat JF, Bryant SH. Threading a database of protein cores. Proteins 1995;23:356–369.
10. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
11. Rooman MJ, Kocher J-PA, Wodak SJ. Prediction of protein backbone conformation based on seven structure assignments: influence of local interactions. J Mol Biol 1991;221:961–979.
12. Park BH, Levitt M. The complexity and accuracy of discrete state models of protein structure. J Mol Biol 1995;249:493–507.
13. Edgoose T, Allison L, Dowe DL. An MML classification of protein structure that knows about angles and sequences. In: Proceedings of third Pacific Symposium on Biocomputing (PSB-98). Hawaii, USA; 1998. pp 585–596.
14. Camproux AC, Tuffery P, Chevrolat JP, Boisvieux JF, Hazout S. Hidden Markov model approach for identifying the modular framework of the protein backbone. Protein Eng 1999;12:1063–1073.
15. Zheng WM, Liu X. A protein structural alphabet and its substitution matrix CLESUM. In: Priami C, Zelikovsky A, editors. Lecture notes in Bioinformatics 3680. Berlin: Springer Verlag; 2005. pp 59–67.

16. Zheng WM. Relation between weight matrix and substitution matrix: motif search by similarity. Bioinformatics 2005;21:938–943.
17. Kabsch W. A discussion of the solution for the best rotation to related two sets of vectors. Acta Crystal 1978;34A:827–828.
18. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
19. Hatanaka H, Ogura K, Moriyama M, Ichikawa S, Yahara I, Inagaki F. Tertiary structure of destrin and structural similarity between two actin-regulating protein families. Cell 1996;85:1047–1055.
20. Benyamini H, Gunasekaran K, Wolfson H, Nussinov R. Conservation and amyloid formation: a study of the gelsolin-like family. Proteins 2003;51:266–281.
21. Guda C, Lu S, Sheeff ED, Bourne PE, Shindyalov IN. CE-MC: a multiple protein structure alignment server. Nucleic Acids Res 2004; 32:W100–W103.
22. Alexandrov NN, Fischer D. Analysis of topological and nontopological structural similarities in the PDB: new examples with old structures. Proteins 1996;25:354–365.