Protein Engineering vol. 16 no. 12 pp. 913–923, 2003
DOI: 10.1093/protein/gzg116

Multiple structural alignment for distantly related all β structures using TOPS pattern discovery and simulated annealing

A.Williams[1], D.R.Gilbert[2] and D.R.Westhead[1,3]

[1]School of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT and [2]Bioinformatics Research Centre, Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK

[3]To whom correspondence should be addressed.
E-mail: westhead@bmb.leeds.ac.uk

Topsalign is a method that will structurally align diverse protein structures, for example, structural alignment of protein superfolds. All proteins within a superfold share the same fold but often have very low sequence identity and different biological and biochemical functions. There is often significant structural diversity around the common scaffold of secondary structure elements of the fold. Topsalign uses topological descriptions of proteins. A pattern discovery algorithm identifies equivalent secondary structure elements between a set of proteins and these are used to produce an initial multiple structure alignment. Simulated annealing is used to optimize the alignment. The output of Topsalign is a multiple structure-based sequence alignment and a 3D superposition of the structures. This method has been tested on three superfolds: the β jelly roll, TIM (α/β) barrel and the OB fold. Topsalign outperforms established methods on very diverse structures. Despite the pattern discovery working only on β strand secondary structure elements, Topsalign is shown to align TIM (α/β) barrel superfamilies, which contain both α helices and β strands.

Keywords: pattern discovery/protein topology/simulated annealing/structural alignment/superfolds

Introduction

There are currently over 19 000 structures in the Protein Data Bank (Berman *et al*., 2000) and many more protein structures being experimentally determined in structural genomic initiatives. The ultimate aim is to know the function of all proteins and structural comparison will play a part in this process. Comparison of protein structures will also help our understanding of evolutionary relationships and physico-chemical constraints on protein folds. An essential part of structure comparison is superposition of protein structures and generation of a corresponding structure-based sequence alignment.

Various tools have been developed for protein structure comparison. Dynamic programming methods that had previously been applied to sequence comparison methods (Needleman and Wunsch, 1970) have been used in protein 3D-structure comparison. SSAP (Taylor and Orengo, 1989a,b; Orengo and Taylor, 1990) uses a double dynamic programming method that takes into account several different features of protein structure including phi/psi angles, accessibility and inter-residue vectors to align two protein structures. COMPARER (Sali and Blundell, 1990) also makes use of many protein features but combines dynamic programming with simulated annealing to produce multiple structural alignments. MNYFIT (Sutcliffe *et al*., 1987) performs superposition of two or more structures using the least squares-fitting algorithm. However, this method is sensitive to concerted shifts in secondary structure elements (Sali and Blundell, 1990). STAMP (Russell and Barton, 1992) also produces multiple structural alignments using the least squares-fitting method, dynamic programming and the structure comparison algorithm of Argos and Rossmann (1976). DALI (Holm and Sander, 1993) is a pairwise structural alignment method used to produce fold classification based on structure–structure alignment of proteins (FSSP) (Holm and Sander, 1996). SARF2 (Alexandrov and Go, 1994) is a structural alignment program that aims to detect 3D similarity of the backbone fragments of a protein without topological restrictions. More recently a Combinatorial Extension method using Monte Carlo optimization has been reported for pairwise comparisons (Shindyalov and Bourne, 1998) and for multiple structure alignments that uses pairwise alignments as a starting point (Guda *et al*., 2001). Prosup (Lackner *et al*., 2000) is a method based on rigid body superposition that keeps the RMSD (root mean square deviation) below a threshold value by applying a distance cut-off for structurally equivalent residues so the main measure of similarity is the number of structurally equivalent residues. K2 (Szustakowski and Weng, 2000) is a method that aligns two structures (based on previous program KENOBI). It implements a fast vector-based technique using a genetic algorithm.

Other structural comparison methods are based on comparing secondary structure elements and their relationships between proteins. For example, GRATH (Harrison *et al*., 2002) is a graph-based algorithm that compares the axial vectors of α helices and β strands of two proteins, together with the distances, angles and chirality between these vectors. It is based on a method by Grindley and co-workers (Grindley *et al*., 1993), but has been developed to include a statistical approach for assessing the significance of any similarities detected. Earlier work by Koch *et al*. (1996) uses a graph method to find maximal common secondary structure elements in a pair of proteins.

Here we present a new method for multiple protein structural alignment that uses topological descriptions of proteins and pattern discovery to build an initial structural alignment that is then optimized using simulated annealing. It has been developed for the construction of multiple structural alignments of very distantly related structures, sharing the same fold, but where substantial differences in loops, the lengths of shared secondary structure elements (SSEs) and non-shared SSEs make alignment difficult for existing approaches.

Simulated annealing has been used previously to solve biological optimization problems. It has been used on numerous occasions for multiple sequence alignment (Hirosawa *et al*., 1993; Ishikawa *et al*., 1993; Kim *et al*., 1994) and in

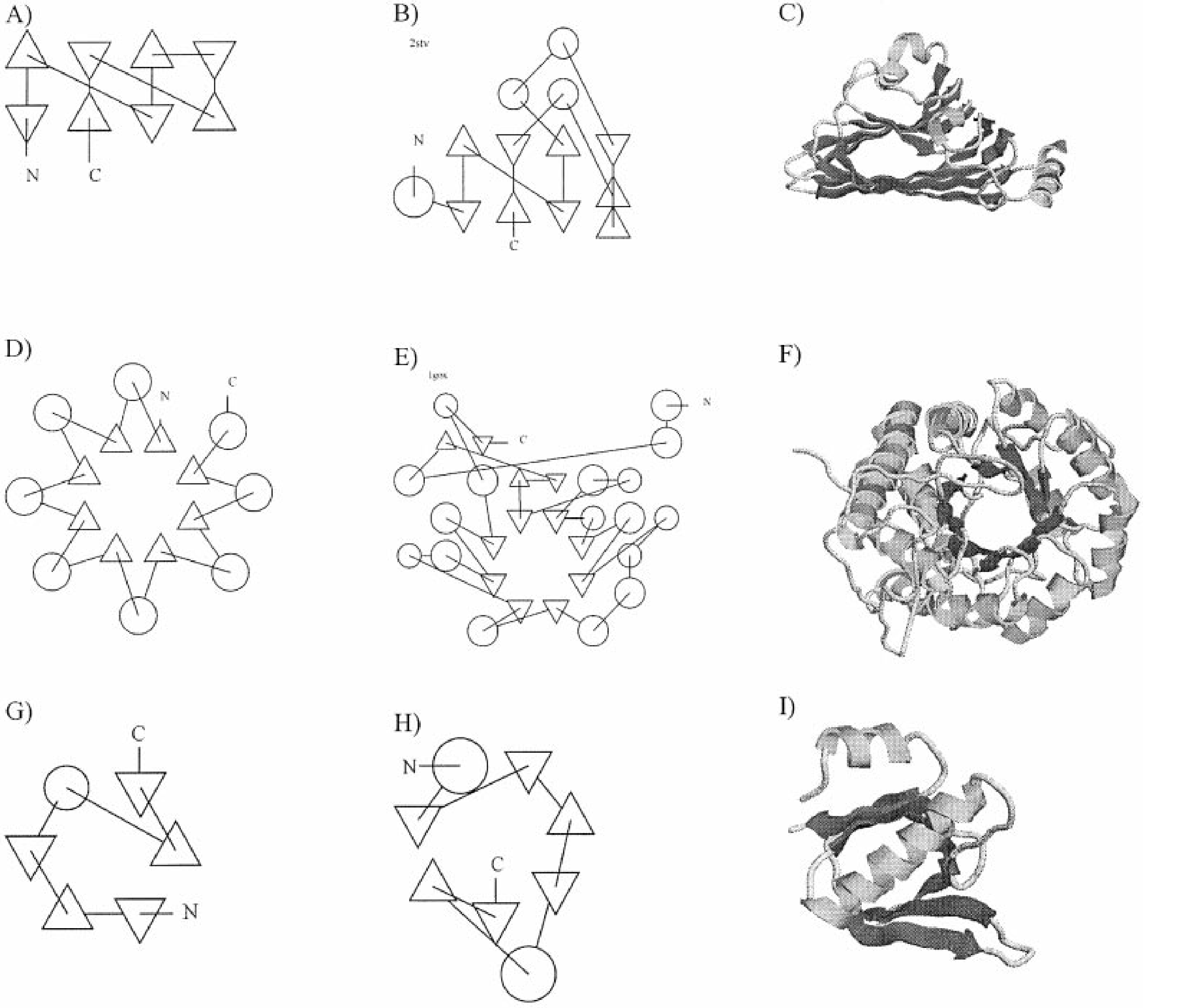


Fig. 1. (**A**) A TOPS cartoon of a β jelly roll. (**B**) A TOPS cartoon of 2stv (a protein with β jelly roll fold). (**C**) A RasMol picture of 2stv. (**D**) A TOPS cartoon of a TIM α/β barrel. (**E**) A TOPS cartoon of 1gox (a protein with a TIM barrel fold). (**F**) A RasMol picture of 1gox. (**G**) A TOPS cartoon of an OB fold. (**H**) A TOPS cartoon of 1tiiD (a domain with an OB fold). (**I**) A RasMol picture of 1tiiD. In the TOPS cartoons, each SSE has a direction (N to C), which is either 'up' (out of the plane of the diagram) or 'down' (into the plane of the diagram). The direction of elements can be deduced from the connecting lines. If the N terminal connection is drawn to the edge of the symbol and the C terminal one to the centre of the symbol, then the direction is up; otherwise, the N terminal connection is drawn to the centre and the C terminal one to the edge and the direction is down. The direction information is duplicated for strands. 'Up' strands are indicated by upward pointing triangles and 'down' strands by downward pointing triangles.

combination with dynamic programming for the COMPARER multiple structure alignment method. It has been suggested that simulated annealing only works well to improve an alignment, i.e. when the method is given an alignment that is already close to optimal and is not trapped in a global minimum (Notredame and Higgins, 1996). Here we use simulated annealing to improve an initial structural alignment, derived by topological pattern discovery.

The multiple structural alignment program has been tested on three protein folds: β jelly roll, TIM α/β barrel and oligomer binding (OB) fold. These three folds are all superfolds (Orengo *et al.*, 1994). Superfolds are unusual in that they support a wide variety of different biological and biochemical functions. The defining characteristics of a superfold are its appearance in many protein superfamilies with no detectable similarity in sequence and diverse functions. They are also very diverse in structure around the basic fold, so are a good test for this method.

The β jelly roll consists of two Greek key motifs that adopt an eight-stranded β sandwich structure (Richardson, 1981). The hydrogen-bonding pattern between adjacent strands is broken in two places and as a consequence the structure comprises of two four-stranded β sheets. Both sheets are purely anti-parallel, with strands adjacent in sequence appearing in different sheets with the exception of the fourth and fifth strands, which are in the same sheet. This leads to a structure with only one hairpin, all other β–β connections being arches. Figure 1A is a Topology Of Protein Structure (TOPS) cartoon of the β jelly roll topology; the two four-stranded anti-parallel β sheets are clearly shown. There are a wide range of biological functions associated with the β jelly roll including viral coat and capsid proteins, tumour necrosis factor proteins, isomerases involved in the glycosylation pathway, lectins and glucanases. Figure 1B and C are a TOPS cartoon of 2stv and a RasMol picture of 2stv, respectively. 2stv is a protein with β jelly roll topology. It is a satellite tobacco necrosis virus coat protein and a member of the viral coat and capsid protein superfamily.

The TIM α/β barrel fold contains eight β–α units in which the strands form a sheet wrapped around into a closed barrel. The helices are on the outside of the sheet. Eight parallel strands form the sheet and the helices are approximately parallel to the strands. TIM α/β barrels mostly function as enzymes and adopt a wide range of functions (Pajados and Palau, 1999). Approximately 10% of enzymes may contain such a fold (Gerlt, 2000). Notwithstanding the diversity of their catalytic reactions, the active site is always found at the C-terminal end of the barrel sheets (Brändon and Tooze, 1991). TIM α/β barrel enzymes are most often involved in molecular or energy metabolism within the cell (Nagano *et al.*, 2002). They include triosephosphate isomerases, (*trans*)glycosidases, aldolases and cobalamin (vitamin B12)-dependent enzymes. Figure 1D shows a TOPS cartoon of a TIM α/β barrel, with the parallel β barrel and surrounding helices clearly shown. Figures 1E and F show a TOPS cartoon and RasMol picture of 1gox, respectively. 1gox is a glycolate oxidase, a member of the FMN-linked oxidoreductase superfamily.

The OB fold comprises of a five-stranded anti-parallel β sheet coiled to form a closed β barrel. This barrel is capped by an α helix located between the third and fourth strands (Murzin, 1993). Figure 1G shows a TOPS cartoon of the OB fold. It is observed in non-homologous proteins that bind oligonucleotides and oligosaccharides. There are three variable loops that contribute residues in the oligomer binding site. Examples of proteins with this fold include staphylococcal nuclease, bacterial enterotoxins, anticodon-binding domain of asp-tRNA synthetase, inorganic pyrophosphatase and tissue inhibitor of metalloproteinases (TIMP). Figure 1H and I show a TOPS cartoon and RasMol picture of 1tiiD, a domain with an OB fold. 1tiiD is a heat-labile toxin from the bacterial enterotoxins superfamily. The three loops at the top of the structure in the RasMol picture are the variable loops that contribute residues in the oligomer binding site.

Materials and methods

Datasets

The structural alignment program (Topsalign) has been tested on three protein folds: β jelly roll, OB fold and TIM α/β barrel.

The β jelly roll is recognized by CATH (Orengo *et al.*, 1997), where most examples can be found under the jelly roll topology level (2.6.120). In SCOP (version 1.53) (Murzin *et al.*, 1995), there is no unique fold level classification for β jelly rolls, but under the all β protein class, folds 9, 12, 13, 17, 18, 21, 22 and 80 are annotated as containing a β jelly roll topology. The β jelly roll dataset used here contained all the proteins with β jelly roll structure from CATH and SCOP. The OB fold is classified in SCOP (version 1.59) as a fold group under the all β protein class. All proteins in this fold group were used in the OB fold dataset. The TIM barrel is classified in SCOP (version 1.59) as a fold group under the α and β proteins (α/β) class. All proteins in this fold group were used. In all three datasets, redundancy was removed at the 90% sequence identity level. The datasets were split into SCOP superfamilies. There were 19 β jelly roll superfamilies, 25 TIM barrel superfamilies and 8 OB fold superfamilies. Structural alignments within superfamilies were performed on superfamilies containing two or more structures. Two β jelly roll superfamilies (the viral coat and capsid proteins and Con A-like lectins/glucanases) are large and very diverse and so were divided into four subgroups. One TIM barrel superfamily [(*trans*)glycosidases] was split into two subgroups for the same reason.

TOPS cartoons and diagrams

Protein structure can be represented in two dimensions using TOPS cartoons (Flores *et al.*, 1994; Westhead *et al.*, 1998). TOPS cartoons represent the structure as a sequence of SSEs—strands (depicted as triangles) and helices (depicted as circles), how they are connected in a sequence from amino to carboxyl terminus, and their relative spatial positions and orientations. See Figure 1 for a more detailed description. TOPS diagrams (Gilbert *et al.*, 1999) are a more formal description of protein structure topology, based on the underlying information used to construct the TOPS cartoon. In particular, they encode the sequence of secondary structure elements and their relative spatial relationships, including hydrogen bonds and strand order in sheets and supersecondary structure chiralities. Figure 2A shows a TOPS diagram for the β jelly roll; the corresponding cartoon is shown in Figure 1A. The topological description is precise and complete for assemblies of β strands, where relative directions (parallel/antiparallel) and order within the sheet can be defined unambiguously. This is not possible for assemblies of α helices. Our methods are currently only applicable to mainly β structures (see Discussion for more details). The strength of these descriptions is their ability to capture the essential topology of a protein fold, ignoring details such as loop and SSE lengths that differ between diverse structures of the same fold.

TOPS pattern matching, pattern discovery and alignments

A TOPS pattern (or motif) (Gilbert *et al.*, 1999) is similar to a TOPS diagram, but is a generalization describing several diagrams that conform to some common topological characteristics. This generalization is achieved by specifying the insertion of SSEs into the sequence of SSEs; indeed a diagram is just a pattern where no inserts are permitted. An insert is indicated by the length of its sequence. Figure 2B shows a TOPS pattern for the β jelly roll. For example, Figure 1B, a TOPS cartoon of viral coat protein 2stv, matches the β jelly roll pattern with 8 β strands. The helices are treated as inserts between strands 4 and 5 and between 6 and 7.

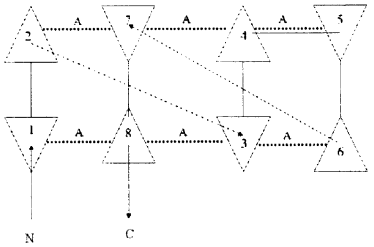
We have previously developed two fast constraint-based methods for matching TOPS patterns to TOPS diagrams (Viksna and Gilbert, 2001). These methods exploit the fact that TOPS graphs (diagrams and patterns) are vertex ordered, a property which reflects the underlying biology whereby an amino acid sequence gives rise to a sequence of SSEs. A TOPS pattern matches a TOPS diagram if, and only if, the pattern is a subgraph of the corresponding TOPS diagram graph, preserving the order of the vertices. Although such matching corresponds to subgraph isomorphism and is thus a NP-complete problem, our methods perform well on TOPS representations of protein structures. The result of a match is a 'correspondence list' that identifies each SSE in the pattern with a corresponding SSE in the matched diagram. Some SSEs in the diagram may not correspond to a SSE in the pattern,

A) β Jelly Roll = (E, H, C), where

$E = (\beta_{-1}, \beta_{+2}, \beta_{-3}, \beta_{+4}, \beta_{-5}, \beta_{+6}, \beta_{-7}, \beta_{+8})$

$H = \{(\beta_{-1}, A, \beta_{+8}), (\beta_{+8}, A, \beta_{-3}), (\beta_{-3}, A, \beta_{+6}), (\beta_{+2}, A, \beta_{-7}), (\beta_{-7}, A, \beta_{+4}), (\beta_{+4}, A, \beta_{-5})\}$

$C = \{\}$



B) β Jelly Roll = (T, H, C), where

$T = (\beta_{\ominus 1} - (0,N) - \beta_{\oplus 2} - (0,N) - \beta_{\ominus 3} - (0,N) - \beta_{\oplus 4} - (0,N) - \beta_{\ominus 5} - (0,N) - \beta_{\oplus 6} - (0,N) - \beta_{\ominus 7} - (0,N) - \beta_{\oplus 8})$

$H = \{(\beta_{\ominus 1}, A, \beta_{\oplus 8}), (\beta_{\oplus 8}, A, \beta_{\ominus 3}), (\beta_{\ominus 3}, A, \beta_{\oplus 6}), (\beta_{\oplus 2}, A, \beta_{\ominus 7}), (\beta_{\ominus 7}, A, \beta_{\oplus 4}), (\beta_{\oplus 4}, A, \beta_{\ominus 5})\}$

$C = \{\}$

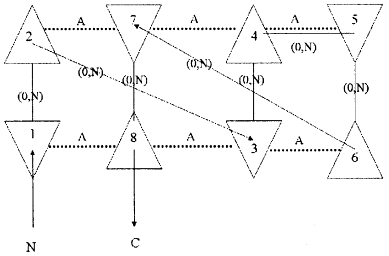


Fig. 2. Formal definitions of (**A**) a β jelly roll diagram and (**B**) a β jelly roll pattern. A TOPS diagram is a triple (E,H,C) where $E = S_1,\ldots,S_k$ is a sequence of length k of SSEs, and H and C are relations over the SSEs (H-bonds and chiralities). An SSE S is either {α,β} standing for helix or strand and ± refers to the direction of the SSE, up/down. Only β strands are involved in H-bonds, and each bond either parallel or anti-parallel {P,A}. Chiralities are associated with handedness, but are not considered in this example. A TOPS pattern is a triple (T,H,C) where T is a sequence V_1-(n,m)-V_2-…-(n,m)-V_k comprising SSEs indicated by V_k and between each of these an insert description. Each insert description is a pair (n,m), where n stands for the minimum and m for the maximum number of SSEs which can be inserted at that position. N is the largest number of SSEs in any TOPS diagram. H and C are the same as in the diagrams. However, since TOPS diagrams exhibit rotational invariance of 180° about the *x* and *y* axes, a direction variable, ⊕ or ⊖, is associated with each SSE in a pattern.

indicating that they are insert positions. Also, there may be more than one way in which a pattern matches a diagram and hence more than one possible correspondence. In practice, we take the correspondence list from the first successful match encountered.

We have also developed a method to automatically discover TOPS patterns from sets of TOPS diagrams (Gilbert *et al.*, 2001). Pattern discovery for sequences is a well-established technique (Brazma *et al.*, 1998); we have adapted a ‘pattern-driven’ approach and applied it to TOPS. Given a set of TOPS diagrams, our algorithm works by finding the smallest pattern that matches all diagrams in the set (using the fast matching algorithm described above) and then repeatedly expanding the pattern and matching it to the members of the set until the largest such pattern is found.

In more detail, our algorithm discovers patterns of H-bonds (and supersecondary structure chiralities) based on the properties of sheets for TOPS diagrams as well as patterns of the associated sequences of SSEs and inserts. Briefly, the algorithm attempts to discover a new sheet by finding, common to all the target set of diagrams, a (fresh) pair of strands sharing an H-bond with a particular direction. Then it attempts to extend the sheet by repeatedly inserting a fresh strand that is H-bonded to one of the existing strands in the (current) sheet. The algorithm then finds all further H-bonds between all the members of the current sheet. The entire process is repeated until no more sheets can be discovered; any supersecondary chiralities arcs between the SSEs in the pattern are then discovered by a similar process. The ranges of the size of the gaps in the corresponding insert positions in the SSE sequence of the pattern are then found. The result is the least general common TOPS pattern characterizing all the members of the target set of protein descriptions.

When such a pattern has been identified, it is used to build an initial alignment. The alignment operation exploits the fact that the pattern must match each structure in the target set, and utilizes the ‘correspondence list’ for each matched diagram. Effectively, the SSEs from each diagram which correspond to (match to) the first SSE in the pattern are said to align, and so on for each SSE in the pattern. We thus generate a structure-based multiple alignment of the sequences of SSEs in the input set of diagrams.

Structural alignments

Topsalign works by taking equivalent SSEs between a set of structures and producing a residue level multiple structural alignment from these. The equivalent SSEs are produced by the TOPS pattern discovery algorithm described above. An initial multiple structure-based sequence alignment is produced from these SSE equivalencies simply by aligning the central residues of equivalent SSEs. Initially, any residues not in SSEs are treated as unaligned positions. A multiple superposition of the structures is built up from pairwise superpositions. The structure with highest sum of residues in SSEs is chosen as a 'model' structure and a pairwise superposition is performed between this structure and all others. Equivalent residues between the 'model' structure and other structures are taken from the structure-based sequence alignment. The algorithm used to superimpose two protein structures is based on an algorithm originally written by McLachlan (McLachlan, 1972; Kabsch, 1976). The method determines the best left rotation matrix to minimize the least-squares differences in the positions of corresponding atoms. The corresponding atoms are Cα atoms taken from the structurally equivalent residues between the model structure and each other structure in the multiple structure-based sequence alignment. Each pairwise alignment is evaluated by a scoring function (Sc):

$$\mathrm{Sc} = N/(1 + \mathrm{RMSD})$$

where N is the number of structurally equivalent residues in each pairwise superposition, and RMSD is calculated from the Cα atom of each structurally equivalent residue. An overall score for the multiple alignment is calculated by taking an average of all pairwise superposition scores. We have found this simple scoring function to be effective in producing good alignments by empirical testing. It is designed to produce alignments with large numbers of aligned residues and small average RMSDs. Where given elsewhere in this paper, RMSDs and number of equivalent residues for a multiple alignment are averages of the pairwise alignment values.

Optimization of the alignment

Once the initial alignment has been produced it is optimized using a simulated annealing algorithm (Metropolis *et al.*, 1953; Kirkpatrick *et al.*, 1983). Simulated annealing is a metaheuristic based on the idea of annealing in physics. It can be used to solve combinatorial optimization problems, especially to avoid local minima that cause problems when using simpler local search methods. Figure 3 shows the simulated annealing procedure used for optimization of the alignment. It shows that if the new configuration is better than the old alignment (calculated by comparing the scores), it is always accepted, or if the new configuration is worse than the old one, the new configuration will be accepted with a probability based on the Boltzmann distribution ($e^{-\Delta E/T}$). The annealing cooling schedule used was as follows: starting temperature = 10, finishing temperature = $10E^{-9}$, temperature decrease = 0.7. At each temperature step 100 random mutations are made to the alignment. Mutations are made by changing information held in a data structure about the position of a SSE in a protein sequence. The data structures hold the sequence start and end positions of each SSE in the structures and each mutation changes one or both (depending on the type of mutation) of these residue numbers. There are three different types of mutation and Figure 4 illustrates each type. The 'move' mutation will change the position of a SSE in one protein. The 'grow' mutation increases the number of structurally equivalent residues in one protein and so may extend the alignment. The alignment may be extended into loop regions or to include SSEs that were not identified by the pattern discovery process. The 'shrink' mutation decreases the number of structurally equivalent residues in one protein. The protein to be mutated, which SSE to be mutated, the type of mutation and in which direction (left or right) are all chosen randomly. The alignment is evaluated by comparing scores after each mutation to see whether to use the new configuration or keep the old one. After optimization, a final alignment is produced. A multiple structure-based sequence alignment is produced and a multiple superposition of the structures.

```
Initialise T
Generate random configuration, $X_{old}$
    WHILE $T>T_{min}$ DO
    FOR u=1 to $N_c$ DO
        Generate new configuration, $X_{new}$
        Calculate new energy, $E_{new}$
        Calculate $\Delta E=E_{new} - E_{old}$
        IF $\Delta E<0$ or random<prob=$e^{-\Delta E/T}$ THEN
            $X_{old} <= X_{new}$
            $E_{old} <= E_{new}$
        ENDIF
    END FOR
    reduce T (T*0.7)
ENDWHILE
```

Where N_c is the number of random changes in configuration at each temperature, T. The variable random is a randomly generated number in the range [0,1].

Fig. 3. The simulated annealing procedure for optimization of the alignment.

```
Original alignment:           123456789
                              yLVWRRIen  2-7
                              mIYVWRPlr  2-7

Mutation 1: Move
                                    or
          YLVWRRien   1-6            ylVWRRIEn  3-8
         mIYVWRPlr    2-7             mIYVWRPlr 2-7

Mutation 2: Grow
                                    or
          YLVWRRIen   1-7            yLVWRRIEn   2-8
          MIYVWRPlr   1-7            mIYVWRPLr   2-8

Mutation 3: Shrink
                                    or
          ylVWRRIen   3-7            yLVWRRien   2-6
          miYVWRPlr   3-7            mIYVWRplr   2-6
```

Fig. 4. A diagram showing the three possible mutations that can be made to an alignment during optimization. The structurally equivalent positions are shown in bold uppercase letters. Lowercase letters represent unaligned positions. The protein to be mutated, which SSE, type of mutation and the direction to make the mutation in are all chosen randomly. The numbers to the right of the sequences show the position of the structurally equivalent residues in the sequence.

Comparisons

Topsalign has been compared with the existing structural alignment packages: STAMP and DALI. STAMP produces multiple structural alignments, whereas DALI can be used to perform a database search where structural neighbours of the query are returned or to perform a pairwise comparison. STAMP was used to align the β jelly roll, OB fold and TIM barrel superfamilies and the results compared with Topsalign results. The STAMP (version 4.2) SCAN method was used. This requires a domain with which to scan the other domains to be superimposed. To select this domain for each superfamily, the domain with highest average sequence identity (i.e. the sequence most similar to all others) was chosen. The SCAN method was chosen because it works particularly well with structures that are very diverse. One of the β jelly roll folds, the double-stranded β helix fold, contains six superfamilies (SCOP version 1.59): RmlC-like, clavaminate synthase-like, cAMP-binding domain-like, regulatory protein AraC, thiamin pyrophosphokinase substrate-binding domain and Trp RNA-binding attenuation protein (TRAP). To align structures from different superfamilies, a representative from each of the six superfamilies was chosen (one with highest average sequence identity). All versus all pairwise alignments were performed for these representatives using DALI and Topsalign. An interactive DALI pairwise comparison was performed at the DALI website (http://www.ebi.ac.uk/dali). STAMP and DALI alignments were evaluated using the Topsalign scoring function to make comparisons possible, and also by comparing RMSD and numbers of aligned residues.

Results

Optimization of the alignment

To investigate the performance of the optimization by simulated annealing, we chose a pair of structures (Figure 5) where the initial pattern was limited to the equivalence of 2 SSEs. Figure 5 shows the anti-freeze protein type III superfamily multiple superposition before and after optimization. It is clear that from the two equivalent SSEs of the initial pattern that simulated annealing optimization has extended the alignment over most of the residues in each structure to produce a high quality alignment.

β jelly roll

Within superfamilies—comparison with STAMP. There were 19 groups of β jelly roll proteins aligned. The 19 groups consisted of 13 different superfamilies with two split into four sub-groups (viral coat and capsid protein and con A-like lectins/glucanases). Figure 6 is a graph comparing the results of Topsalign and STAMP using the Topsalign scoring function. From Figure 6, it can be seen that Topsalign produces alignment scores better than or equal to STAMP for 13 of the 19 groups of β jelly roll proteins. For example, for TNF-like superfamily STAMP scored 35.25 (RMSD = 1.44 Å, number of equivalent residues = 86) and Topsalign scored 44.38 (RMSD = 1.35 Å, number of equivalent residues = 98), corresponding to a better alignment. For the PHM/PNGase F superfamily, STAMP scored 11.49 (RMSD = 3.44 Å, number of equivalent residues = 51) and Topsalign scored 21.31 (RMSD = 3.17 Å, number of equivalent residues = 86). For three superfamilies, segmented RNA genome viruses, viral coat and capsid proteins (all four subgroups), and viral proteins, STAMP did not align all the structures and aligned only a subset. Topsalign aligned all structures within these superfamilies. Figure 7 shows a RasMol picture of the Topsalign multiple superposition of the segmented RNA genome viruses superfamily. STAMP aligned three of the four structures in this superfamily using 1ahsA as the seed structure for the SCAN method. It aligned the domains 2btvS1, 1ahsA and 2hmgC, but not 1flcE1. STAMP aligned these three proteins with an RMSD of 2.99 Å over 61 residues. Topsalign achieved a score of 14.67 for the alignment of all four structures with an RMSD of 2.46 Å over 51 structurally equivalent residues. Table I shows the Topsalign score, RMSD and number of structurally equivalent residues that Topsalign achieved for the six alignments that STAMP did not produce. Table I shows that Topsalign successfully managed to align superfamilies with RMSD values ranging from 2.99 Å over 71 residues for the viral coat and capsid 4 group to 3.24 Å RMSD over 86 residues for the viral coat and capsid 3 group. A less convincing result Topsalign achieved was an RMSD of 4.91 Å RMSD over 91 residues for the viral coat and capsid 1 group. It has not been possible to produce a good quality alignment of the viral protein superfamily with either STAMP or Topsalign (RMSD = 11.93 over 104 equivalent residues).

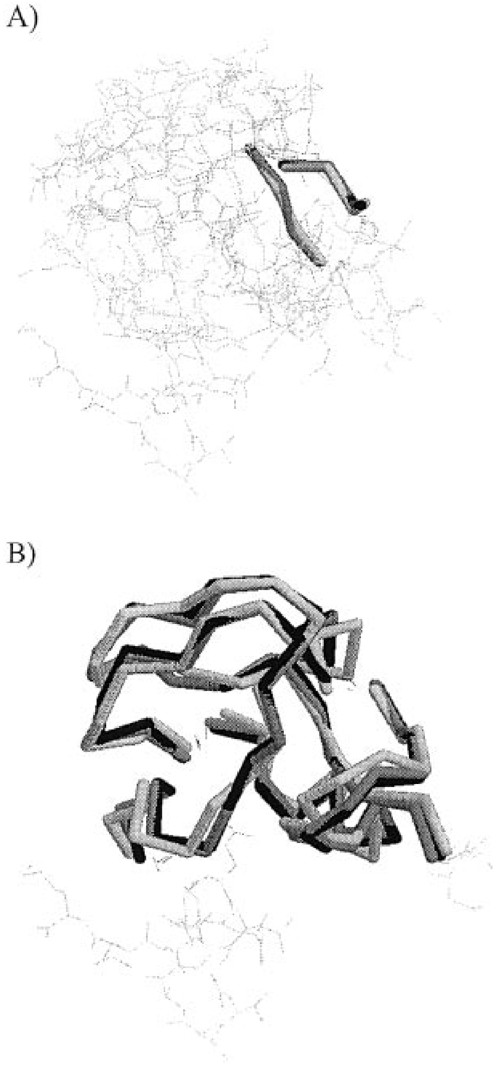


Fig. 5. Multiple superposition of 3rdn (light grey), 6msi (grey) and 1ops (black), all members of the anti-freeze protein type III superfamily. This superfamily belongs to the SCOP β clip fold, which is annotated as having β jelly roll topology. Wireframe display indicates unaligned positions and backbone display indicates aligned positions. (**A**) Before optimization. (**B**) After optimization.

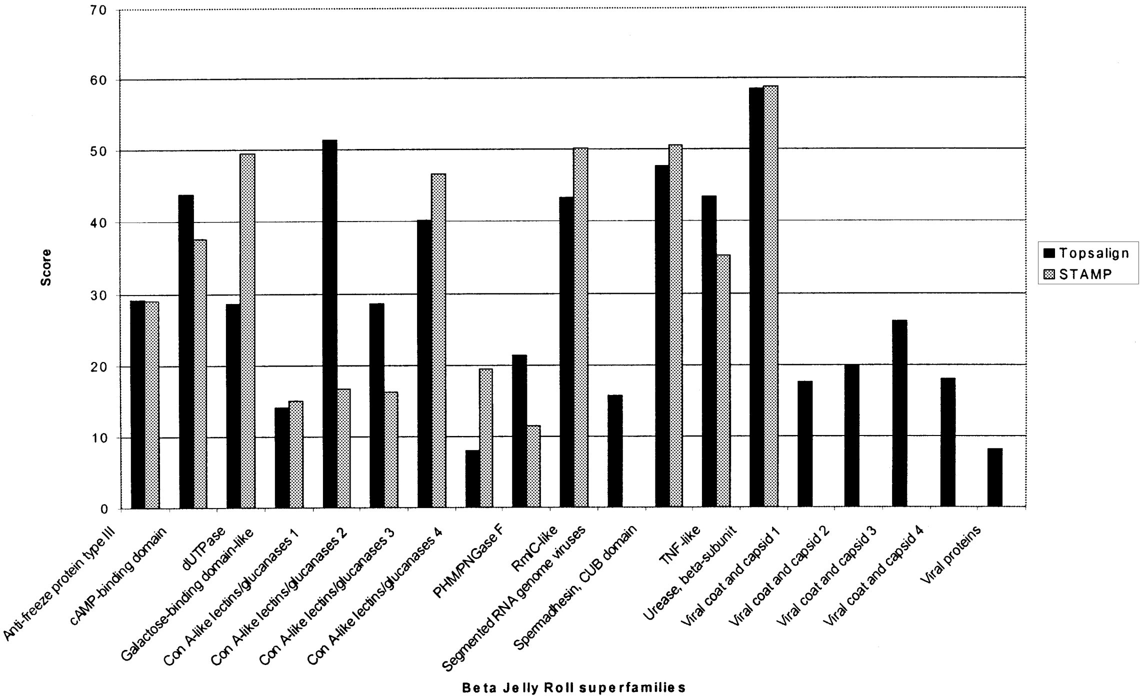


Fig. 6. A graph comparing the multiple structural alignments of β jelly roll superfamilies by STAMP and Topsalign.

Table I. Shows the Topsalign score, RMSD and number of structurally equivalent residues that Topsalign achieved for the alignments of superfamilies that STAMP did not align

Superfamily	Topsalign score	RMSD (Å)	Number of equivalent residues	Length of shortest protein in alignment
Segmented RNA genome viruses	14.67	2.46	51	104 (1flcE1)
Viral coat and capsid 1	17.48	4.91	91	147 (1a34A)
Viral coat and capsid 2	17.97	3.85	80	190 (1f15B)
Viral coat and capsid 3	25.09	3.24	86	141 (1stmA)
Viral coat and capsid 4	17.98	2.99	71	184 (2stv)
Viral proteins	8.04	11.93	104	367 (1cjdC)

The length of the shortest protein in the alignment is also shown to put the number of equivalent residues into context, the PDB code of the shortest protein is given in parentheses.

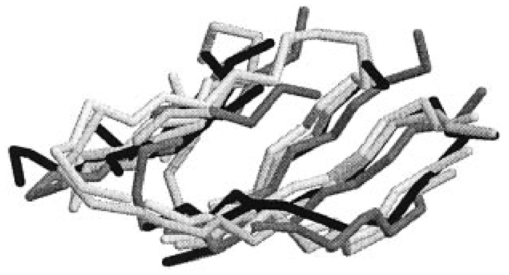


Fig. 7. Topsalign alignment of the β jelly roll superfamily segmented RNA genome viruses. Shows 2btvS1 (white), 1ahsA (light grey), 2hmgC (grey) and 1flcE1 (black). Only structurally equivalent residues are shown.

Between superfamilies—comparison with DALI. The double stranded β helix fold (annotated as having β jelly roll topology in SCOP) was used to test pairwise structural alignments between different superfamilies. It contains six superfamilies. Figure 8 is a graph showing the results of pairwise structural alignments of a representative of each double-stranded β helix superfamily against a representative from all other double-stranded β helix superfamilies. Results are shown for DALI and Topsalign. Topsalign produces alignment scores better than or equal to DALI for 13 out of 15 alignments. On seven occasions, DALI does not produce an alignment, presumably because of the large structural diversity of the proteins. Topsalign produced an alignment for all 15 pairwise alignments. An example of a pairwise alignment that DALI did not produce is between a thiamin pyrophosphokinase, substrate-binding domain protein (1ig3A from the SCOP version 1.59 thiamin pyrophosphokinase, substrate binding domain superfamily) and a TRAP protein (1wapA from the SCOP version 1.59 TRAP superfamily). Figure 9 shows a RasMol picture of the Topsalign superposition of 1ig3A and 1wapA. Topsalign achieved a score of 13.35 for this alignment, with a RMSD of 2.59 Å over 48 structurally equivalent residues. Table II shows the Topsalign score, RMSD and number of structurally equivalent residues that Topsalign achieved for the 7 alignments that DALI did not produce. Successful Topsalign results in Table II include the alignment of 1rgs02 and 1ig3A, which has an RMSD of 3.16 Å over 62 equivalent residues and the alignment of 1cauB and 1wapA, which has an RMSD of 3.28 Å over 56 equivalent residues. Other less impressive results in the table include the alignment of 1ipsA and 1rgs02 with an RMSD of 4.61 over 90 equivalent residues and the alignment of 1ipsA and 1wapA with an RMSD of 7.88 over 63 equivalent residues.

OB fold

Within superfamilies—comparison with STAMP. There are six superfamilies in the SCOP OB fold group. These were aligned using STAMP and Topsalign. Figure 10 is a graph showing the score achieved by both STAMP and Topsalign for each OB fold superfamily. Topsalign produces alignment scores better than or equal to STAMP for 4 out of 6 alignments. STAMP

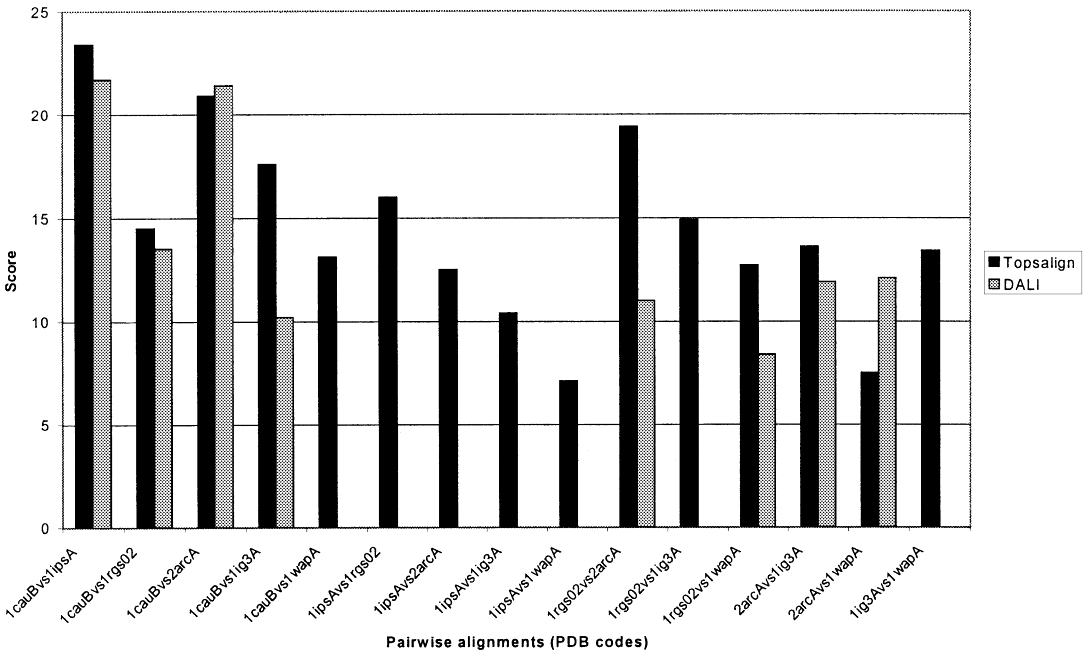


Fig. 8. A graph comparing the pairwise alignments of representative structures from the double stranded β helix six superfamilies by DALI and Topsalign.

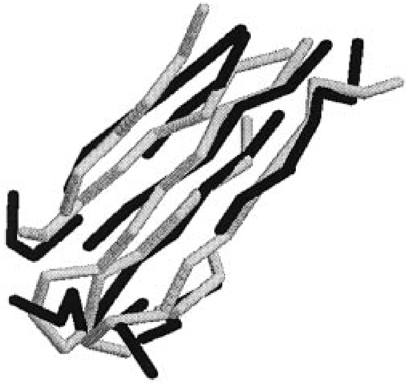


Fig. 9. A pairwise alignment of 1wapA (light grey) and 1ig3A (black) produced by Topsalign. Only structurally equivalent residues are shown.

Table II. The Topsalign score, RMSD and number of structurally equivalent residues that Topsalign achieved for the pairwise alignments that DALI did not align

Pairwise alignment	Topsalign score	RMSD (Å)	Number of equivalent residues	Length of shortest protein in alignment
1cauB – 1wapA	13.08	3.28	56	68 (1wapA)
1ipsA – 1rgs02	16.03	4.61	90	132 (1rgs02)
1ipsA – 2arcA	12.51	5.99	87	161 (2arcA)
1ipsA – 1ig3A	10.40	5.83	71	85 (1ig3A)
1ipsA – 1wapA	7.09	7.88	63	68 (1wapA)
1rgs02 – 1ig3A	14.92	3.16	62	85 (1ig3A)
1ig3A – 1wapA	13.35	2.59	48	68 (1wapA)

Proteins are identified by PDB code. The chain and SCOP domain number given if relevant. The length of the shortest protein in the alignment is also shown to put the number of equivalent residues into context, the PDB code of the shortest protein is given in parentheses.

does not align all the structures for two superfamilies (nucleic acid binding proteins and bacterial enterotoxins), whereas Topsalign aligns all structures for all six superfamilies. For example, STAMP aligns only 14 of the 17 proteins in the bacterial enterotoxins superfamily (using 1se401 as the seed sequence) with an RMSD of 2.20 Å over 39 equivalent residues. Topsalign aligns all 17 structures with an RMSD of 3.33 Å over 48 residues.

TIM α/β barrel (an example containing some helices)

Within superfamilies—comparison with STAMP. There were 19 SCOP TIM barrel superfamilies aligned by Topsalign and STAMP. The (*trans*)glycosidases superfamily was split into two subgroups. Topsalign only produces alignment scores better than or equal to STAMP for 4 out of 20 alignments. Figure 11 is a graph showing the Topsalign and STAMP scores for these alignments. In certain cases STAMP produces much higher alignment scores than Topsalign. For example, for the metallo-dependent hydrolases superfamily, STAMP scored 44.37 (RMSD = 2.11 Å, number of equivalent residues = 138) and Topsalign only scored 8.68 (RMSD = 4.35 Å, number of equivalent residues = 44). In other cases, the alignment scores are more equal. For example, for the quinolinic acid phosphoribosyltransferase superfamily STAMP scored 58.52 (RMSD = 1.70 Å, number of equivalent residues = 158) and Topsalign scored 58.56 (RMSD = 1.44 Å, number of equivalent residues = 143). For the ribulose-phosphate binding barrel superfamily, STAMP scored 19.94 (RMSD = 2.41 Å, number of equivalent residues = 68) and Topsalign scored 19.97 (RMSD = 2.06 Å, number of equivalent residues = 56). It also shows that STAMP failed to align all structures within the cobalamin-dependent enzymes and (*trans*)glycosidases group 2 superfamilies. Topsalign did not produce alignments for the aldolase and (*trans*)glycosidases group 2 superfamilies. Topsalign did not produce alignments for these two superfamilies because the TOPS pattern discovery could not identify any equivalent SSEs between the structures.

Discussion

We have shown that our multiple structure-based sequence alignment method is able to match two well-established

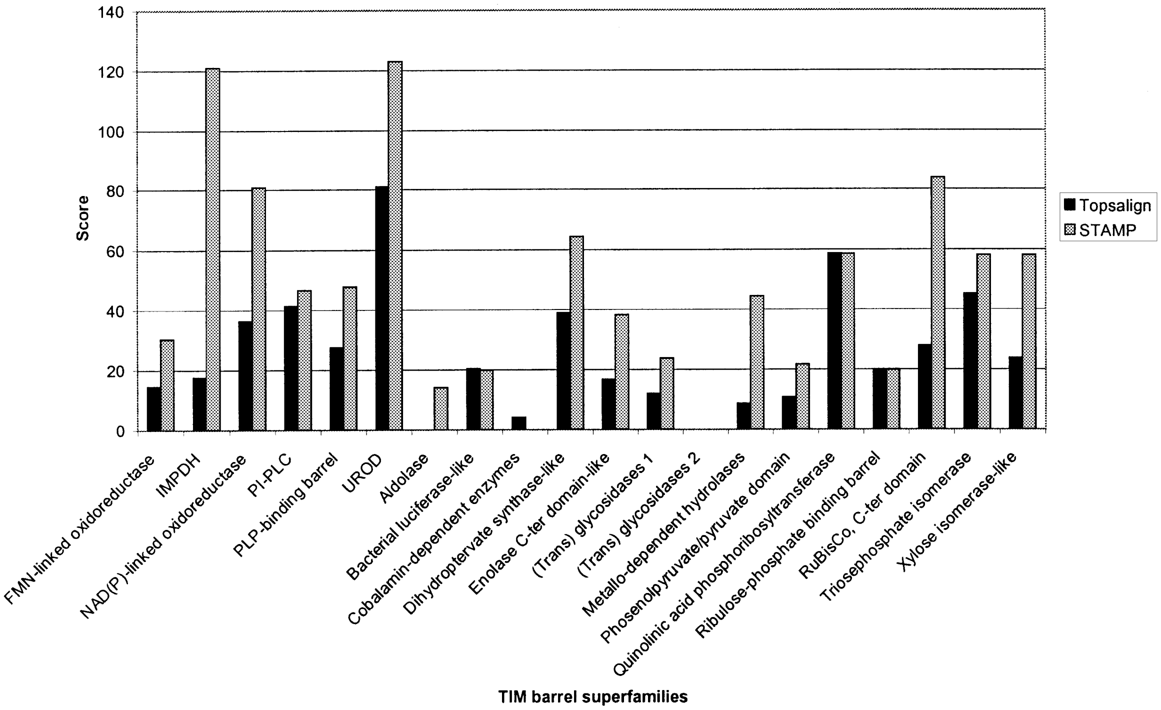


Fig. 10. A graph comparing the multiple structural alignments of OB fold superfamilies by STAMP and Topsalign.

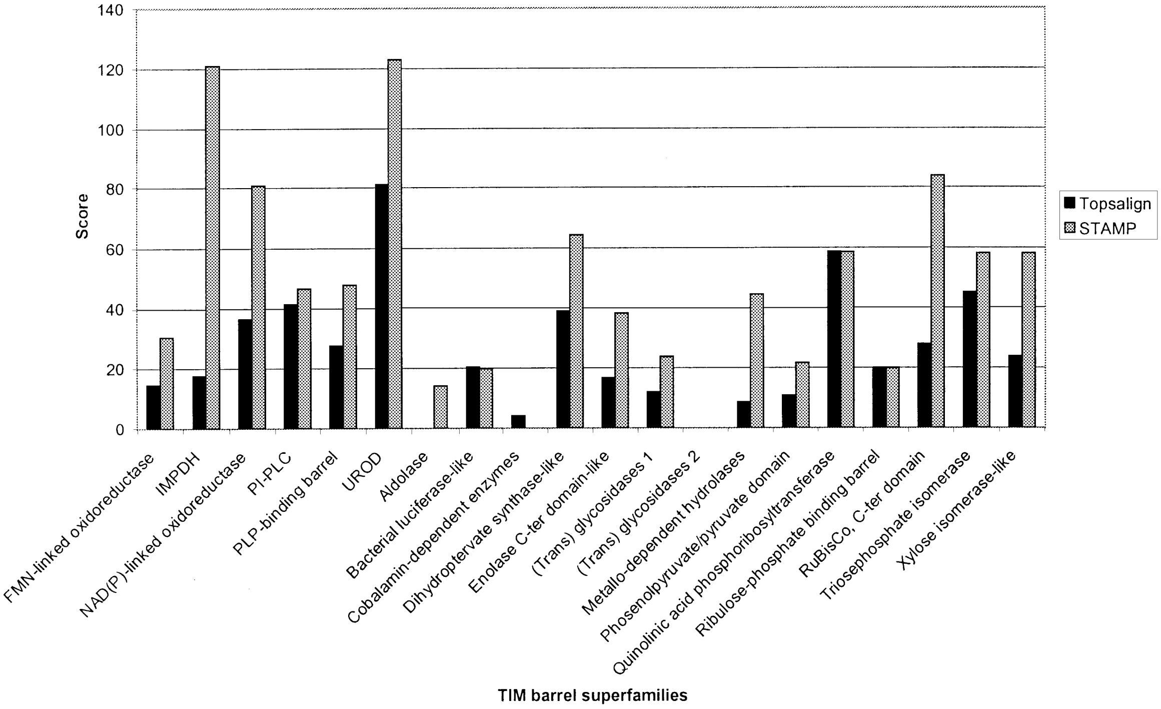


Fig. 11. A graph comparing the multiple structural alignments of TIM α/β barrel superfamilies by STAMP and Topsalign.

methods for alignment within SCOP superfamilies and between superfamilies with the same fold. In some cases, it produces better alignments, with more equivalent residues and lower RMSDs. For some proteins, the established methods were unable to produce alignments, despite the fact that the proteins concerned share the same fold or even the same superfamily. In several of these cases, Topsalign was successful in producing alignments with reasonable RMSD values and numbers of equivalent residues.

β jelly roll and OB fold

Overall, within superfamilies Topsalign seems to perform better than STAMP at aligning β jelly roll superfamilies and achieves similar results for OB fold superfamilies. These

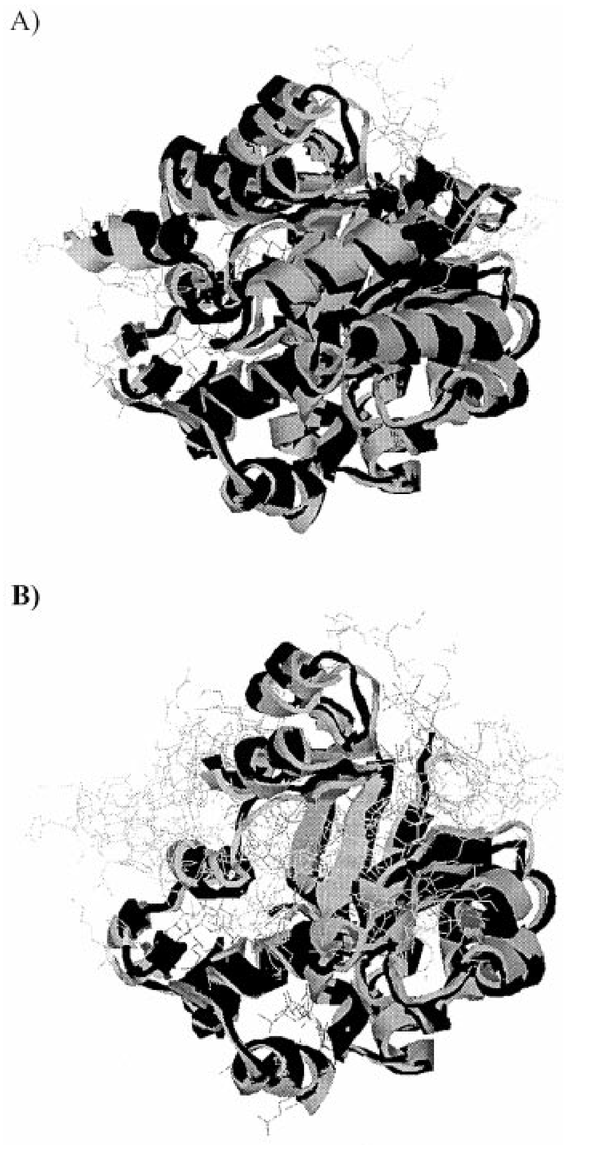


Fig. 12. Superposition of 1uroA (black) and 1j93A (light grey), members of the TIM α/β barrel fold UROD superfamily. Structurally equivalent residues are shown in cartoon display and unaligned residues in wireframe. (**A**) STAMP superposition. (**B**) Topsalign superposition.

superfamilies are all classified in the all β protein class. Proteins with β jelly roll structure are very diverse, sometimes even within superfamilies. Between superfamilies with the same fold, overall Topsalign performs better than DALI does for the pairwise structural alignments of the double-stranded β helix superfamilies. These structures are more diverse than aligning within superfamilies and Topsalign performs well.

TIM α/β barrel

Topsalign performs badly in comparison with STAMP. The reason for this is that the TOPS pattern discovery algorithm is limited to patterns in β sheet topology and does not include helices. This is because the current TOPS data structures only contain a minimum amount of information regarding α helices. We are currently enhancing the TOPS data structures with more structural information, including more helix packing relationships, to extend the applicability of our pattern discovery algorithm to structures with substantial helix content. The TIM α/β barrel fold contains both β strands and α helices. It is classified by SCOP in the α and β protein class. This means that the α helical content of the fold has been ignored by the TOPS pattern discovery and hence has not been included in equivalent SSEs and this will affect the resulting structural alignment. Optimization of the alignment has succeeded in extending the initial strand equivalencies into some α helical regions. For example, the UROD superfamily structural alignment scores for STAMP and Topsalign are 122.93 and 80.97, respectively. The RMSD values for STAMP and Topsalign for the UROD superfamily are comparable (RMSD for alignment by STAMP is 1.66 Å and by Topsalign is 1.47 Å), but the number of structurally equivalent residues is a lot lower for Topsalign (200) than STAMP (327). Figure 12 shows the UROD multiple superpositions in RasMol by STAMP and Topsalign. Comparing the two superpositions it can be seen that Topsalign has aligned some α helical structure, but STAMP has aligned more.

Conclusion

We have produced a multiple structural alignment program that performs well, mainly on β structures. However, despite the pattern discovery working only on β strand SSEs, Topsalign is shown to align TIM (α/β) barrel superfamilies. We see the main use of our program as alignment of very diverse β structures with the same fold, where core secondary structures and topological relationships are present, but with substantial structural diversity in other parts of the fold.

Acknowledgement

We thank the BBSRC for sponsorship.

References

Alexandrov,N.N. and Go,N. (1994) *Protein Sci.*, **3**, 866–875.

Argos,P. and Rossmann,M. (1976) *J. Mol. Biol.*, **105**, 75–95.

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

Brändon,C.I and Tooze,J. (1991) In Brändon,C.I. and Tooze,J. (eds), *Introduction to Protein Structure*. Garland Publishing, New York, pp. 43–57.

Brazma,A., Jonassen,I., Eidhammer I. and Gilbert D.R. (1998) *J. Comput. Biol.*, **5**, 277–303.

Flores,T.P.J., Moss,D.M. and Thornton,J.M. (1994) *Protein Eng.*, **7**, 31–37.

Gerlt,J.A. (2000) *Nat. Struct. Biol.*, **7**, 171–173.

Gilbert,D.R., Westhead,D.R., Nagano,N. and Thornton,J.M. (1999) *Bioinformatics*, **15**, 317–326.

Gilbert,D.R., Westhead,D.R, Viksna,J. and Thornton,J.M. (2001) *J. Comput. Chem.*, **26**, 23–30.

Grindley,H.M., Artymiuk,P.J., Rice,D.W. and Willet,P. (1993) *J. Mol. Biol.*, **229**, 707–721.

Guda,C., Scheef,E.D., Bourne,P.E. and Shindyalov,I.N. (2001) *Pac. Symp. Biocomput.*, **6**, 275–286.

Harrison,A., Pearl,F., Mott,R., Thornton,J. and Orengo,C. (2002) *J. Mol. Biol.*, **323**, 909–926.

Hirosawa,M., Hoshida,M., Ishikawa,M. and Toya,T. (1993) *Comput. Appl. Biosci.*, **9**, 161–167.

Holm,L. and Sander,C. (1993) *J. Mol. Biol.*, **233**, 123–138 (see also: http://www.ebi.ac.uk/dali/domain).

Holm,L. and Sander,C. (1996) *Science*, **273**, 595–602.

Ishikawa,M., Toya,T., Hoshida,M., Nitta,K., Ogiwara,A. and Kanehisa,M. (1993) *Comput. Appl. Biosci.*, **9**, 267–273.

Kabsch,W. (1976) *Acta Crystallogr. A*, **32**, 922–923.

Kim,J., Paramanik,S. and Chung,M.J. (1994) *Comput. Appl. Biosci.*, **10**, 419–426.

Kirkpatrick,S., Gerlatt,C.D. and Vecchi,M.P. (1983) *Science*, **220**, 671–680.

Koch,I., Lengauer,T. and Wanke,E. (1996) *J. Comput. Biol.*, **3**, 289–306.

Lackner,P., Koppensteiner,W.A., Sippl,M.J. and Domingues,F.S. (2000) *Protein Eng.*, **13**, 745–752.

McLachlan,A.D. (1972) *Acta Crystallogr. A*, **28**, 656–657.

Metropolis,N., Rosenbluth,A.W., Rosenbluth,M.N., Teller,A.H. and Teller,E. (1953) *J. Chem. Phys.*, **21**, 1087–1092.

Murzin,A. (1993) *EMBO J.*, **12**, 861–867.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) *J. Mol. Biol.*, **247**, 536–540 (see also: http://scop.mrc-lmb.cam.ac.uk/scop).

Nagano,N., Orengo,C.A. and Thornton,J.M. (2002) *J. Mol. Biol.*, **321**, 741–765.

Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol.*, **48**, 443–453.

Notredame,C. and Higgins,D.G. (1996) *Nucleic Acids Res.*, **24**, 1515–1524.

Orengo,C.A. and Taylor,W.R. (1990) *J. Theor. Biol.*, **147**, 517–551.

Orengo,C.A., Jones,D.T. and Thornton,J.M. (1994) *Nature*, **372**, 631–634.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure*, **5**, 1093–1108 (see also: http://www.biochem.ucl.ac.uk/bsm/cath).

Pajados,G. and Palau,J. (1999) *Biologia*, **54**, 231–254.

Richardson,J.S. (1981) *Adv. Protein Chem.*, **34**, 167–339.

Russell,R.B. and Barton,G.J. (1992) *Proteins Struct. Funct. Genet.*, **14**, 309–323.

Sali,A. and Blundell,T.L. (1990) *J. Mol. Biol.*, **212**, 403–428.

Shindyalov,I.N and Bourne,P.E. (1998) *Protein Eng.*, **11**, 739–747.

Sutcliffe,M.J., Haneef,I., Carney,D. and Blundell,T.L. (1987) *Protein Eng.*, **1**, 377–384.

Szustakowski,J.D. and Weng,Z. (2000) *Proteins Struct. Funct. Genet.*, **38**, 428–440.

Taylor,W.R. and Orengo,C.A. (1989a) *Protein Eng.*, **2**, 505–519.

Taylor,W.R. and Orengo,C.A. (1989b) *J. Mol. Biol.*, **208**, 1–21.

Viksna,J. and Gilbert,D.R. (2001) *WABI 2001: 1st Workshop on Algorithms in BioInformatics*, LNCS **2149**, pp. 98–111.

Westhead,D.R., Hatton,D.C. and Thornton,J.M. (1998) *Trends Biochem. Sci.*, **23**, 35–36 (see also: http://www3.ebi.ac.uk/tops).

Received June 3, 2003; revised September 25, 2003; accepted October 21, 2003