MSAT – a Multiple Sequence Alignment tool based on TOPS

Te Ren, Mallika Veeramalai, Aik Choon Tan and David Gilbert

Bioinformatics Research Centre
Department of Computer Science
University of Glasgow
Glasgow, G12 8QQ United Kingdom
Tel:+44 141 330 2421, Fax:+44 141 330 3690

{rent, mallika, actan, drg}@brc.dcs.gla.ac.uk

Keywords

TOPS, multiple sequence alignment, multiple structure alignment, T-Coffee, ClustalW, Fold classification, Fold assignment, HMM profiles

Abstract

This paper describes the development of a new method for multiple sequence alignment based on fold-level protein structure alignments, which provides an improvement in accuracy compared to the most commonly used sequence only based techniques. This method integrates the widely used progressive multiple sequence alignment approach ClustalW with the TOPS topology based alignment algorithm. The TOPS approach produces a structure alignment for the input protein set by using a topology-based pattern discovery program, providing a set of matched sequence regions that can be used to guide a sequence alignment using ClustalW. The resulting alignments are more reliable than a sequence-only alignment, as determined by 20-fold cross validation with a set of 106 protein examples from the CATH database, distributed in 7 superfold families. The method is particularly effective for sets of proteins which have similar structures at the fold level, but low sequence identity. The aim of this research is to contribute towards bridging the gap between protein sequence and structure analysis, in the hope that this can be used to assist the understanding of the relationship between sequence, structure and function. The tool is available at http://balabio.dcs.gla.ac.uk/msat/

INTRODUCTION

The number of known structures in the Protein Data Bank (PDB) is increasing rapidly, due in particular to structural genomics initiatives (Goldsmith-Fischman and Honig, 2003; Kim, 1998) that aim to populate protein fold space using high-throughput experimental technologies. Most of these projects focus on proteins whose fold cannot be easily recognised by simple sequence comparison with proteins of known structure (Benner and Levitt, 2000). Recent research in structural biology has contributed to our understanding of the relationships between amino acid sequences and protein structures, and between different protein structures (Goldsmith-Fischman and Honig, 2003). One well-accepted observation is that the structure of a protein is more conserved than its underlying amino acid sequence. Hence, learning the similarities (or differences) between protein structures is very important in understanding the relationship between protein sequence, structure and function, and for the analysis of possible evolutionary relationships.

Several protein structure knowledge bases such as SCOP (Murzin et al, 1995), CATH (Orengo et al, 1997) and DALI (Holm and Sander, 1997), ranging from being manually curated to fully automated, have been created in order to further our understanding about the relationships between protein sequence and structure. Most of these databases classify protein structures in a hierarchical manner: several studies have shown that these databases can differ from one another due to differing domain definitions rather than by the assignment of the same domain to different folds (Day et al 2003; Hadley and Jones, 1999). We have employed the CATH classification scheme and domain assignments as our primary data source, but our approach can be adapted to use other structural classification schemes.

Experimental protein structure determination is expensive and time consuming; therefore sophisticated computational methods have been developed and applied to detect, search for and compare remote protein homology at the sequence level in the hope that annotations of proteins of known function can be transferred to a protein of unknown function. Thus methods that are capable of modelling protein families are important and useful for protein fold recognition studies as well as for discovering relationships between sequences and folds. There are sufficient protein families at the fold level in the CATH database to allow sophisticated analysis to compute over them. Important information for homology modelling is provided by identifying those positions within a sequence which are tolerant to mutation. Moreover sequence segments, based on conserved structural features, can be used to predict the structures of more distant members of a family (Orengo, 1993). The general steps for computational methods to model and detect protein families are as follows (Mian and Dubchak, 2000):

(i) Define a protein family where the members usually share sequence and/or structural features based on the analysis of multiple sequence alignments

(ii) Construct a model (pattern) that characterises the family

(iii) Design a scoring function scheme such that, given a model and a set of positive and negative examples, a score is returned for the examples covered by the model.

Most computational tools developed for protein fold prediction are primarily based on sequence similarity. If a new protein sequence with an unknown structure has high sequence similarity to a protein of known structure, then the new protein may share a similar fold with this structure. Closely related proteins can be detected by comparing their sequences, using standard bioinformatics tools such as BLAST (Altschul et al, 1990), PSI-BLAST (Altschul et al, 1997) and FASTA (Pearson and Lipman, 1988).

In order to obtain the maximum benefit from the wealth of known protein structures, fast and sensitive methods should be used to classify the Protein Databank into fold families. Unfortunately, structure comparison based on 3D coordinates is expensive in terms of computational power and time. In order to reduce the computational time, several heuristic methods have been proposed in the literature. TOPS is one such approach that employs machine learning and heuristic algorithms to discover common structural patterns (or motifs), and enables these patterns to be matched to a set of TOPS descriptions (Gilbert et al 1999; Gilbert et al 2001; Viksna and Gilbert 2001). The advantage of this system is the simplicity representation of the protein structure in topological model, where in this level of abstraction only the sequence of secondary structure elements, SSEs (i.e. strands and helices) and the spatial relationships between SSEs (i.e. hydrogen bonds and chirality) are considered. The performance of the TOPS system has been evaluated over the SCOP database entries and has proved to be much faster than the usual atom-coordinates driven algorithms (Gilbert et al, 2001). Also, the accuracy of the TOPS system approaches that of pure 3D structure driven methods, and much better than sequence-based approaches (Gilbert et al, 2001). A principal disadvantage of this system is the lack of detailed explanatory power (i.e. amino acid residues) required for the understanding of the relationship between protein sequence, structure and function. The addition of sequence information to TOPS descriptions, for example as in the work described in this paper, will greatly improve the effectiveness of the TOPS system.

The objective of the research reported in this paper is to generate TOPS based multiple sequence alignments for sequences that have low sequence similarity (< 10%) but share common topology or fold. In order to make good use of the TOPS system and to increase its usefulness for biologists, this research combines the properties of sequences and structure alignments in the MSAT program (Multiple Sequence Alignment tool based on TOPS). The system can generate structure alignments of a set of proteins along with their underlying sequences; evolutionary relationships among the sequences in each identified "structure segment" can then be revealed by a sequence-driven multiple alignment tool. MSAT multiple sequence alignments can then be used, for example, to make HMM profiles. As work in progress, not reported here, we are constructing a database of such profiles for fold families and developing a system to search a query protein sequence against these profiles in order to perform fold assignment.

BACKGROUND

TOPS

Protein structure can be described at a highly simplified 'topological' level using TOPS cartoons (Sternberg and Thornton, 1977). These are schematic abstractions of protein three-dimensional structures in two dimensions. A sample cartoon for 2bopA0 is shown in Figure 1(b) (for comparison a Rasmol cartoon is given in Figure 1(a)). The TOPS cartoon shows the secondary structure elements (SSEs) - β-strands (depicted by triangles) and α-helices (depicted by circles). TOPS cartoons were originally drawn manually; subsequently an algorithm that automatically produces cartoons from protein structures has been devised and implemented (Flores et al., 1994; Westhead et al, 1998, 1999).

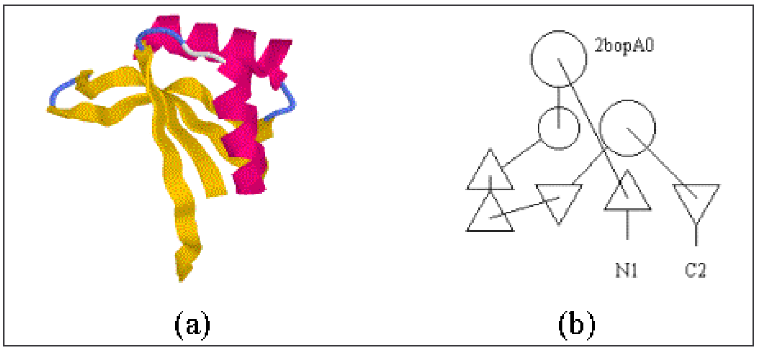
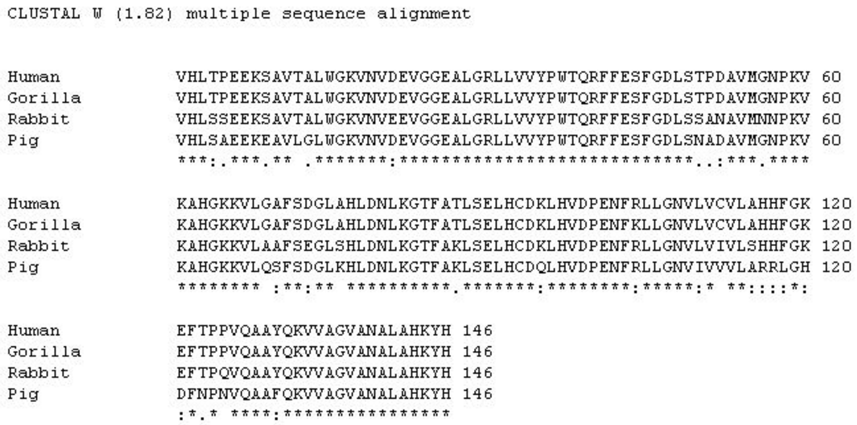


Figure 1: (a) Rasmol (Sayle and Milner-White, 1995) cartoon view of 2bop (b) TOPS cartoon

Although the cartoons do not explicitly display the information about hydrogen bonding between strands or chirality connections between SSEs, the cartoon generation program outputs a data structure containing, among other things, the information about these bonds and connections. We refer to this richer representation as a TOPS diagram (Figure 4); it is from this form that the cartoons are produced, which can then be stored in graphical format, for example as postscript or gif files.

Multiple sequence alignment

Multiple sequence alignments are usually inferred from amino acid sequences alone. Biologists produce high quality multiple sequence alignments by hand using their expert knowledge of protein sequence evolution. They must consider many issues to generate a good alignment, such as highly conserved residues, the influence of secondary and tertiary structure, etc. Obviously, the manual construction of multiple alignments is laborious and the development of automatic multiple sequence alignment methods has become an active topic of research for the Bioinformatics community. The most widely used method in molecular biology to align sets of nucleotide or amino acid sequences, is to build up a multiple alignment progressively. In this iterative process, the most closely related sequences are aligned together, keeping the early alignments fixed. There are two main strategies when considering the alignments between two protein sequences: global alignment (Needleman and Wunsch, 1970) and local alignment (Smith and Waterman, 1981). Figure 2 shows a multiple sequence alignment of four haemoglobin sequences generated by ClustalW (Thompson et al, 1994) which is based on progressive pairwise global alignments.



```
CLUSTAL W (1.82) multiple sequence alignment


Human           VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV 60
Gorilla         VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV 60
Rabbit          VHLSSEEKSAVTALWGKVNVEEVGGEALGRLLVVYPWTQRFFESFGDLSSANAVMNNPKV 60
Pig             VHLSAEEKEAVLGLWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSNADAVMGNPKV 60
                ***:.***.** .*******:****************************..:***.****

Human           KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK 120
Gorilla         KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFKLLGNVLVCVLAHHFGK 120
Rabbit          KAHGKKVLAAFSEGLSHLDNLKGTFAKLSELHCDKLHVDPENFRLLGNVLVIVLSHHFGK 120
Pig             KAHGKKVLQSFSDGLKHLDNLKGTFAKLSELHCDQLHVDPENFRLLGNVIVVVLARRLGH 120
                ******** :**:** **********.*******:********:*****:* **::::*:

Human           EFTPPVQAAYQKVVAGVANALAHKYH 146
Gorilla         EFTPPVQAAYQKVVAGVANALAHKYH 146
Rabbit          EFTPQVQAAYQKVVAGVANALAHKYH 146
Pig             DFNPNVQAAFQKVVAGVANALAHKYH 146
                :*.* ****:****************
```

Figure 2: An example of ClustalW multiple sequence alignment

A multiple alignment of sequences S_1, S_2, … S_k is a series of strings S_1’, S_2’, …S_k’ such that

1. |S_1’| = |S_2’| = … = |S_k’|
2. S_j’ is an extension of S_j, obtained by insertion of gaps Def .1

METHODS

We have devised MSAT (Multiple Sequence Alignment tool based on TOPS) that integrates TOPS generated alignments of secondary structures with multiple sequence alignments from tools such as ClustalW. The MSAT package incorporates two existing programs: the TOPS alignment program (Gilbert et al, 2001) and ClustalW v1.83 (Thompson et al, 1994), and requires as input the amino-acid sequence and the 3D protein structure description for each member of the protein family under consideration. MSAT generates the model by the following steps:

(i) Generate TOPS descriptions from 3D coordinate data, using the DSSP program and the TOPS cartoon generation program.
(ii) Generate a structural alignment for all the members of a family using the TOPS alignment program.
(iii) Divide the aligned sequences of the family members into several segments corresponding to the common secondary structure elements indicated by the TOPS alignments;
(iv) Generate a multiple sequence alignment for those segments using ClustalW;
(v) Concatenate the segments to produce the full length alignment;

This section is divided into two parts. Firstly we describe the representation of TOPS diagrams, TOPS patterns and the theory behind the production of a TOPS alignment. Secondly we describe the sequence segmentation stage and the usage of ClustalW. We briefly describe the underlying formal basis of each program and direct interested readers to the relevant references for a detailed explanation of each technique.

MSAT

Figure 3 illustrates the general architecture of the MSAT system. Protein domains can be selected from the CATH hierarchy. TOPS descriptions of the input set of proteins are retrieved from a relational database and passed to the TOPS pattern discovery program to produce a TOPS alignment of secondary structure elements. Next the amino-acid sequences of input set are divided into several sequence segments according to the aligned SSEs. These segments are then aligned by ClustalW and concatenated together to make the full-length sequence alignments. We have also used T-Coffee (Notredame et al, 2000) to produce the multiple sequence alignments, but found that it performed less well then ClustalW in our evaluations.

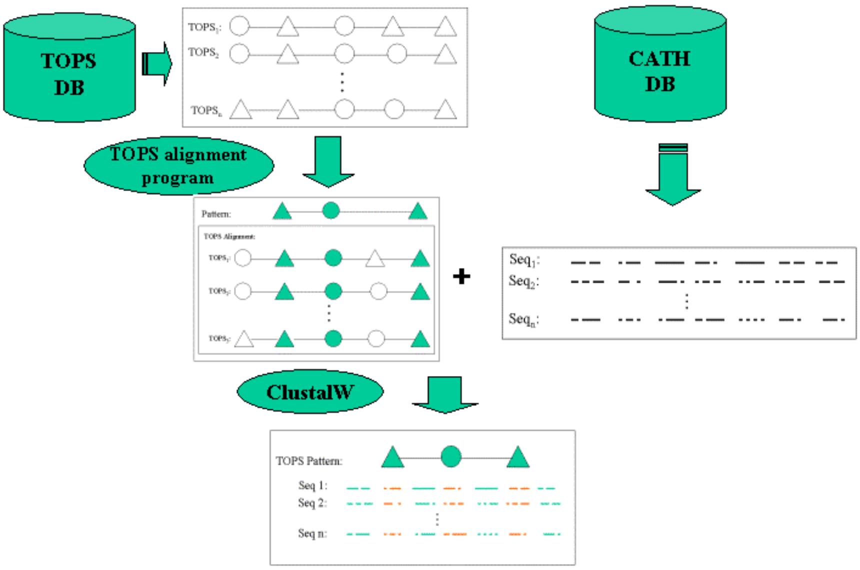


Figure 3: Architecture of MSAT

TOPS diagrams and pattern discovery

Formally, a TOPS diagram is a triple $T = (S, H, C)$ where $S = S_1, S_2, \ldots, S_n$ is a sequence length n of secondary structure elements (SSEs), H is a set of topological representations of hydrogen bonds, and C is a set of chirality connections. The loop regions between each SSE are implicitly represented in the sequence, e.g. there is a possible loop region between any S_i and S_{i+1}, as well as before S_1 and after S_n. We keep a record of which amino acid residue is at the start and end of each SSE, and hence we also have a record of the residues corresponding to the start and end of each loop region. In this TOPS description an “H-bond” refers to a ladder of individual hydrogen bonds between adjacent strands in a sheet, and chiralities are a subset of those generated by Slidel’s algorithm (Slidel and Thornton, 1996). TOPS descriptions are generated from atom coordinate files via DSSP (Kabsch and Sander, 1983) and the TOPS cartoon generation program (Westhead et. al., 1999).

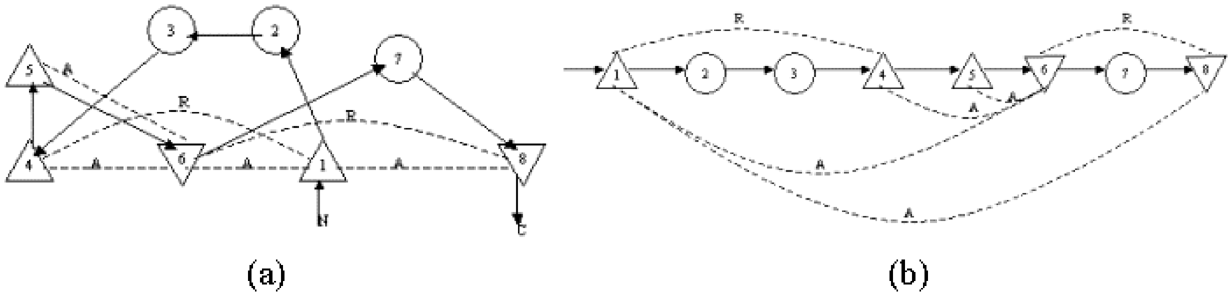


Figure 4: (a) TOPS diagram for 2bopA0. (b) Linearised TOPS diagram for 2bopA0.

A TOPS pattern is a generalisation that describes a set of TOPS diagrams which conform to some common topological characteristics. The form of a TOPS pattern *P= (SP,HP,CP)* is similar to a TOPS diagram, except that $SP = I_0\text{-}S_1\text{-}I_1\text{-}S_2\text{-}I_2\text{-}\ldots\text{-}I_{n-1}\text{-}S_n\text{-}I_n$ is a TOPS sequence pattern of length $2n+1$, where S_i is a pattern SSE and I_i is an insert. An insert describes those SSEs that are common to some but not all of the diagrams in the input set, as well as representing the loop regions between the SSEs. There is an insert in the pattern between each pair of SSEs in the pattern, as well as at each end of the SSE sequence (i.e. at both the N-terminus and the C-terminus). We say that the length of a TOPS pattern is the number of SSEs in the sequence pattern *SP*. The TOPS pattern discovery algorithm operates by discovering patterns of H-bonds and chiralities based on the properties of sheets for TOPS diagrams; it also derives TOPS sequence patterns, *SP*, i.e. the associated sequences of SSEs. For more details of the pattern discovery algorithm, see (Viksna and Gilbert, 2001).

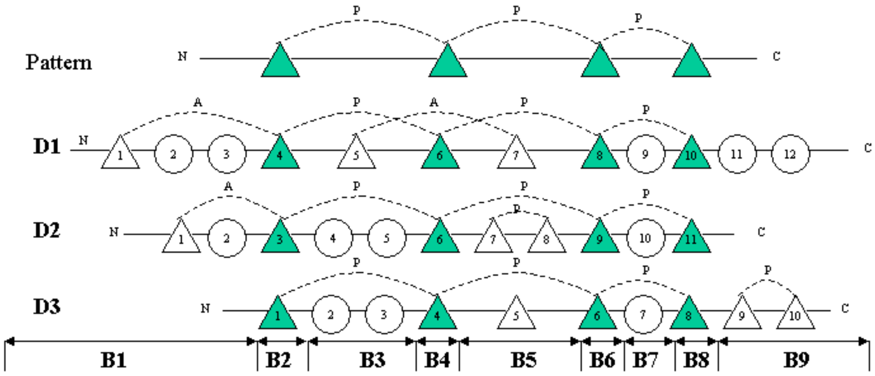


Figure 5: Making a TOPS alignment

For example, given three TOPS diagrams $D1 = (S1, H1, C1)$, $D2 = (S2, H2, C2)$, $D3 = (S3, H3, C3)$ and a least general common pattern $P = (SP, HP, CP)$ (Figure 5), we can make a structural alignment of these domains by matching P with D1, D2 and D3 respectively. If there are *n* SSEs and hence *n*+1 insert positions in the pattern P, then there are 2*n*+1 corresponding blocks in each of D1, D2 and D3. For example in Figure 5 the common pattern for the three domains contains 4 strands, and hence there are 4 aligned SSE blocks and 5 unaligned blocks in D1, D2 and D3, giving a total of 9 blocks (B1,…,B9) where the underlying amino acid sequences in each block will be subsequently be aligned.

The compression of a pattern with respect to a set of structures is computed in a standard way by reference to the size of the pattern and the total size of the components of the TOPS structures which are not included in the pattern. This value is normalised to the range 0 (worst) to 1 (best). Intuitively, the compression is a measure of how much a set of structures has in common. A good value indicates that they all share most of their elements. A poor value is, in turn, an indicator of a diverse group that shares little common structure. A set of structures that are all identical will naturally have a common motif that is identical to all the members of the group, and hence the compression of this motif with respect to the group will be exactly 1. Sets with no amount of common structure (a set comprising an all-alpha and an all-beta protein, for example) will have an empty motif which will have a compression of 0.

Multiple sequence alignment strategy

We apply a "divide-and-conquer" strategy when generating sequence alignments and then concatenate the underlying amino acid sequence of the aligned SSE regions (represented as A in Def.2) and other regions (represented as L) respectively. For the underlying amino-acid sequence S associated with a TOPS diagram:

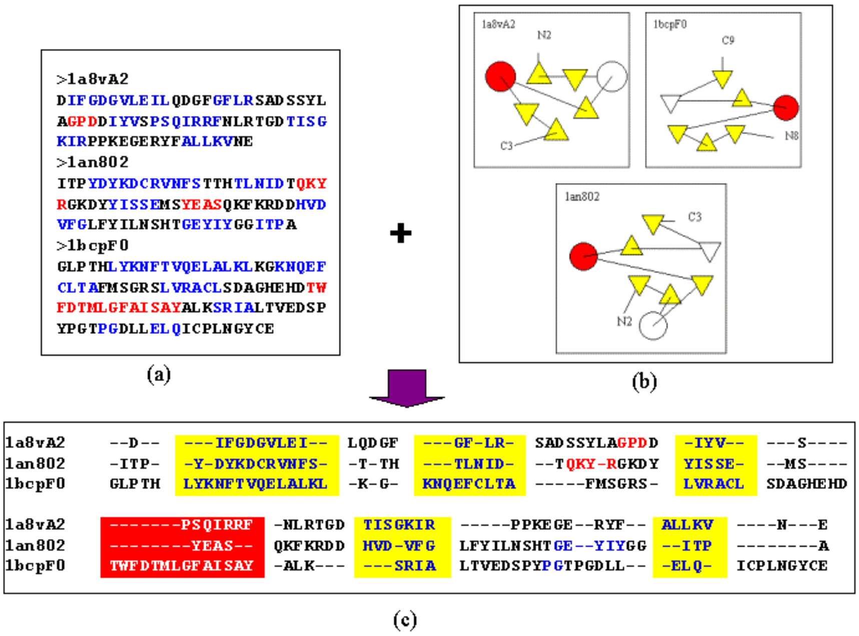
$S = L_1 \wedge A_1 \wedge L_2 \wedge A_2 \wedge \ldots \wedge L_{m-1} \wedge A_{m-1} \wedge L_m$ where

A_i is an amino-acid sequence of a corresponding aligned SSE

L_i is an amino-acid sequence for a region between two aligned SSEs S_{i-1} and S_i

^ represents sequence concatenation

$m=2*|SP|+1$ Def. 2



```
>1a8vA2
DIFGDGVLEILQDGFGFLRSADSSYL
AGPDDIYVSPSQIRRFNLRTGDTISG
KIRPPKEGERYFALLKVNE
>1an802
ITPYDYKDCRVNFSTTHTLNIDTQKY
RGKDYYISSEMSYEASQKFKRDDHVD
VFGLFYILNSHTGEYIYGGITPA
>1bcpF0
GLPTHLYKNFTVQELALKLKGKNQEF
CLTAFMSGRSLVRACLSDAGHEHDTW
FDTMLGFAISAYALKSRIALTVEDSP
YPGTPGDLLELQICPLNGYCE
```

(a)

(b)

```
1a8vA2    --D--  ---IFGDGVLEI--  LQDGF  ---GF-LR-  SADSSYLAGPDD  -IYV--  ---S----
1an802    -ITP-  -Y-DYKDCRVNFS-  -T-TH  ---TLNID-  --TQKY-RGKDY  YISSE-  --MS----
1bcpF0    GLPTH  LYKNFTVQELALKL  -K-G-  KNQEFCLTA  -----FMSGRS-  LVRACL  SDAGHEHD

1a8vA2    -------PSQIRRF  -NLRTGD  TISGKIR  -----PPKEGE--RYF--  ALLKV  ----N---E
1an802    --------YEAS--  QKFKRDD  HVD-VFG  LFYILNSHTGE--YIYGG  --ITP  --------A
1bcpF0    TWFDTMLGFAISAY  -ALK---  ---SRIA  LTVEDSPYPGTPGDLL--  -ELQ-  ICPLNGYCE
```

(c)

Figure 6: MSAT sequence alignment strategy. *Figure 6(a) illustrates the input amino-acid sequences in FASTA format. Figure 6(b) shows the corresponding TOPS diagrams, where light triangles and dark circles represent matched SSEs discovered by the TOPS alignment program. Figure 6(c) is a concatenated multiple sequence alignment, where the light and dark coloured regions correspond to the matched SSE regions in 6(b).*

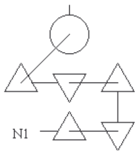
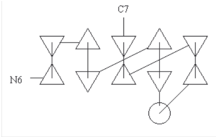
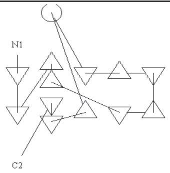
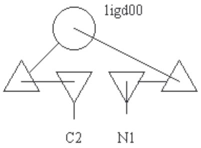
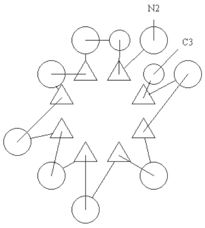
ClustalW v1.83 (Thompson et al, 1994) is used to generate the multiple sequence alignment for each TOPS aligned or unaligned region. For example, given three CATH domains '1a8vA2', '1an802' and '1bcpF0', the TOPS structure alignment program will firstly produce a structure alignment, the individual segments are then aligned using ClustalW, and finally all of the segments are concatenated to produce a full length multiple sequence alignment. This process is illustrated in Figure 6.

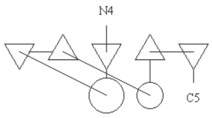
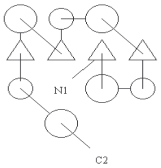
EVALUATION

Data sources

To demonstrate the usefulness of MSAT we have performed training and testing on examples of different folds compiled from the CATH knowledge base. CATH is a hierarchical classification of protein domain structures, which clusters proteins into four major levels, Class (C), Architecture (A), Topology (T) and Homologous superfamily (H) (Orengo et al, 1997). A CATH Homologous superfamily comprises sequences that might have low sequence similarity, but whose structure and function suggest a common evolutionary origin. The Topology (T) level in CATH represents the fold of a protein; fold families contain sequences that have structural similarities. We have selected 106 protein domains distributed in 7 common β-rich super folds from the CATH database as our test cases (Table 1), namely OB folds, Jelly Rolls, IG-like, UB rolls, TIM barrels, alpha-beta plaits and Rossman folds. We have selected domains so that each family has an average pair-wise sequence similarity of less than 10% and a TOPS compression value greater than 0.5.

Table 1: Fold families used in this research

Fold	CATH code	Nr	3D Structure	TOPS Cartoon	Average Structure Compression	Average Pair-wise Sequence Similarity
OB fold	2.40.50	15			0.71	7.12%
Jelly Rolls	2.60.120	15			0.63	7.40%
Immunoglobulin -like (Ig-like)	2.60.40	14			0.87	8.15%
UB rolls	3.10.20	14			0.62	7.14%
TIM Barrel	3.20.20	15			0.82	7.27%

Fold	CATH code	Nr	3D Structure	TOPS Cartoon	Average Structure Compression	Average Pair-wise Sequence Similarity
Alpha-Beta Plaits	3.30.70	18			0.97	8.12%
Rossmann fold	3.40.50	15			0.96	8.92%

Evaluation method

Hidden Markov models (HMMs) (Eddy, 1998; Durbin et al, 1998) have been shown to be effective in classifying protein families and recognising protein motifs from amino acid sequences. These models consist of a set of states, each with a probability of generating a particular residue, and a set of transition probabilities for moving from one state to the next. For protein sequences, there is usually a single state for each residue position in the model, and the transition possibilities are restricted to three: transition to the next residue position, insertion, or deletion of the next position. Thus HMMs are constrained to represent only correlations that are local in the amino acid sequence - the amino acid observed at a particular position only depends on the position immediately preceding it. In addition, profile-based HMMs have been widely applied in bioinformatics to identify remote homologs and to generate multiple sequence alignments for protein families (Mian and Dubchak, 2000; Durbin et al, 1998). We have employed the HMMER package (Eddy, 1998) in the MSAT system to generate a representative profile-based HMM for each of the seven superfold families using the multiple sequence alignment from MSAT. We also separately generated two more profiles for each superfold families using multiple sequence alignments generated by ClustalW, and T-Coffee respectively without prior alignment via TOPS. These pure sequence alignments were used to evaluate those generated by MSAT. We further generated MSAT profiles using T-Coffee (Notredame et al. 2000) for the multiple sequence alignment stage, and evaluated them using the method described below; they performed less well than the MSAT profiles using ClustalW and are not illustrated here.

In order to evaluate the discriminative power of a multiple sequence alignment, we hypothesise that:

S belongs to the family f described by alignment A, if the e-value e(A,S) generated by the hmmsearch program is below some threshold e_0.

		Predicted Class	
		True	False
Actual Class	Positive examples	True Positives (TP)	False Negatives (FN)
	Negative examples	False Positives (FP)	True Negatives (TN)

Figure 7: Contingency table for a two-class classification/prediction

We have evaluated this hypothesis by comparison to the ‘fold’ classification of the CATH database (Orengo et al, 1997) which we take as the ‘gold standard’. The CATH fold classification can be used to separate protein domains into those which are truly related (Positive examples, Np) and those which are not truly related (Negative examples, Nn) at the fold level. Our approach can then be used to classify CATH domains deemed similar by the *hmmsearch* program (i.e. entries whose sequence e-value $e(A,S) < e_0$) and constructed a contingency table for each fold families as shown in Figure 7. The positive examples Np comprise True Positives (TP), protein entries correctly predicted as members of the true class by *hmmsearch*, and False Negatives (FN), those wrongly predicted as members of the false class. The negative examples Nn comprise False Positives (FP), those incorrectly predicted as belonging to the true class, and True Negatives (TN), those negative examples correctly predicted as not belonging to the true class. Obviously, the numbers TP (e_0) and FP (e_0) of true and false positives respectively depend on the choice of the threshold e_0. Following Gribskov and Robinson (1996), we characterise the algorithm by its coverage rate c (e_0)= TP (e_0)/Np and its false positive rate f(e_0) = FP(e_0)/Nn. The plot of c(e_0) against f(e_0) is known as a Receiver Operating Characteristic (ROC) curve.

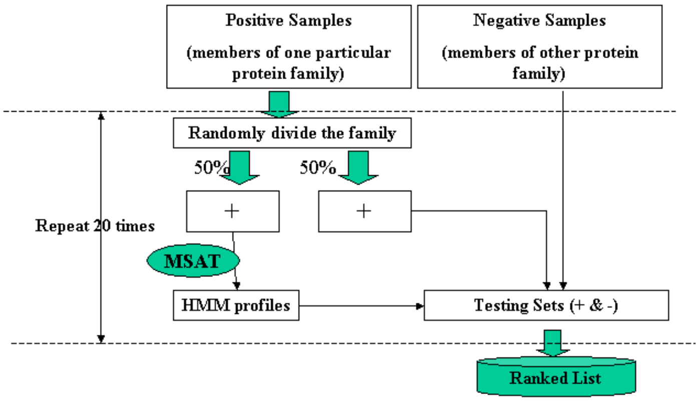


Figure 8: Evaluation schema

In order to carry out a form of cross validation over a fold family of N protein domains, half of the domains were removed from the set, multiple sequence alignment and profiles generated for the remaining N/2 domains by ClustalW and *hmmbuild*, and the family membership of the removed domains predicted by *hmmsearch* and its accuracy measured (Figure 8). This process was repeated by randomly dividing the N domains 20 times into two sets. This variation of cross validation is used because the size of N is less than 20 for each family in our case. As mentioned above, the number of true positives and false positives is largely dependent on the threshold e_0. In our case, a large e-value (9999) is artificially selected as an input parameter to *hmmsearch* so that all domain entries (including positive sample and negative samples) are listed in ascending order of e-value. Then the e-value e_0 for each positive entry is recorded as well as the number of true positives and false positives above it in the ordered list of matches. Finally using this data, the *coverage rate* and *false positive rate* is computed for that family.

RESULTS AND DISCUSSION

We have performed three different alignment methods and compared their performance based on the CATH folds listed in Table 1. ClustalW v1.83 (Thompson et al, 1994a) and T-Coffee v1.37 (Notredame et al. 2000) are sequence-based, and MSAT is our method. All these methods were used with the default parameters for the corresponding package.

Figure 9 shows the ROC curves of every method for each fold. In ROC curve space, the points (0,0) and (1,1) represent the HMM profile which always predicts negative class and positive class respectively, the point (0,1) represents the ideal HMM profile, and (1,0) represents the HMM profile that gets it all wrong. Therefore, given two curves in ROC space, the upper curve performs better than the lower one. We can see from Figure 9 that the MSAT method is generally better than the other two approaches. In order to evaluate the significance of differences found between these methods, we have computed several different statistical measurements: Positive Predicted Value (PPV), F-measure and the Matthew's Coefficient Correlation value, using the cut-off point at 80% coverage. Positive Predicted Value (Eq. 1) evaluates the positive predicted reliability of the method; the values range from 0 to 1, with the higher a value representing a better prediction. F-measure (Eq. 2, van Rijsbergen, 1979) is a measurement that has been widely used in the Information Retrieval community to balance the trade-off between the coverage and the PPV of a classifier (HMM profile in our case); values range from 0 (worst) to 1 (best). Matthew's Coefficient Correlation (Eq. 3) computes the correlation between positive and negative classifications, and is 1.0 when there are no false positives or negatives, 0 where the classifier is a random classifier, and -1.0 when there are only false positives and false negatives (Brazma et al, 1998). Table 2 indicates the PPV, CC and F-measure of each method at 80% coverage.

$$PPV = \frac{TP}{TP + FP} \qquad \text{Eq. 1}$$

$$F - Measure = \frac{2 \times TP}{2 \times TP + FN + FP} \qquad \text{Eq. 2}$$

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + TN)(TN + FN)(TP + FN)(TN + FP)}} \qquad \text{Eq. 3}$$

Overall, from our results (Table 2), we observed that MSAT alignment do improve the multiple sequence alignment in five out of seven families when compared to ClustalW and T-Coffee.

Table 2: Different statistical measurements at 80% coverage

Fold		MSAT	ClustalW	T-Coffee
OB fold	PPV	0.068	0.066	0.067
	F-Measure	0.126	0.123	-0.098
	CC	-0.08	-0.189	-0.098
Jelly Rolls	PPV	0.545	0.314	0.293
	F-Measure	0.631	0.449	0.423
	CC	0.624	0.450	0.422
Ig-like	PPV	0.160	0.130	0.106
	F-Measure	0.256	0.215	0.188
	CC	0.225	0.171	0.133
UB roll	PPV	0.142	0.105	0.139
	F-Measure	0.238	0.186	0.228
	CC	0.214	0.139	0.186
TIM Barrel	PPV	0.945	0.951	0.845
	F-Measure	0.896	0.899	0.842
	CC	0.891	0.894	0.833
Alpha-Beta Plaits	PPV	0.170	0.142	0.133
	F-Measure	0.282	0.242	0.231
	CC	0.229	0.165	0.160
Rossmann fold	PPV	0.298	0.326	0.284
	F-Measure	0.438	0.463	0.420
	CC	0.441	0.464	0.423

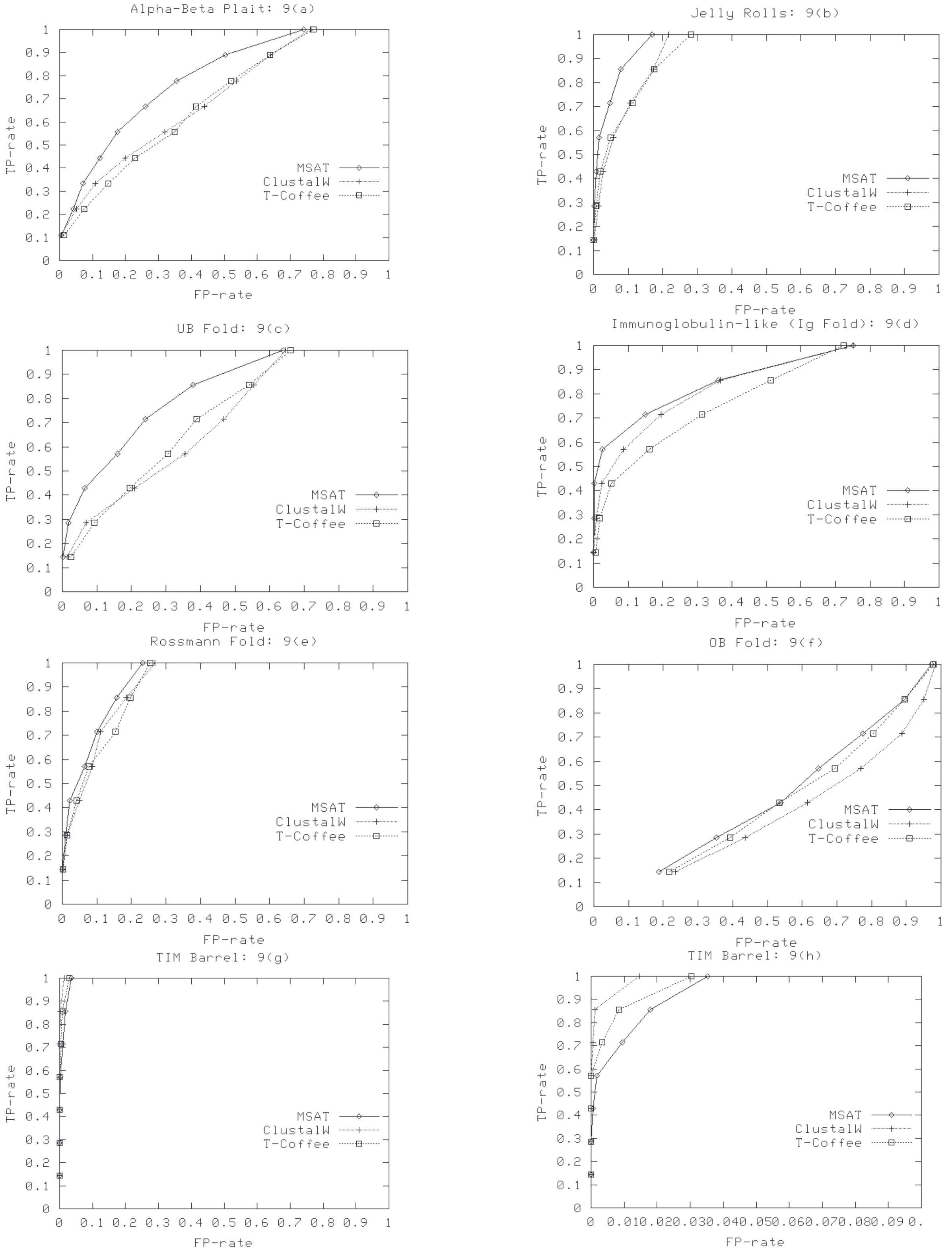
Table 3: TOPS pattern properties

Fold	# SSEs	#β strands	#H-bonds	#Chirality arcs	Avg. Structure Compression
OB Fold	8	5	3	0	0.71
Jelly Rolls	11	11	9	0	0.63
Ig – like	7	7	5	0	0.87
UB Roll	8	5	4	2	0.62
TIM Barrel	16	8	7	7	0.82
Alpha-Beta Plaints	6	4	3	2	0.97
Rossmann Fold	10	5	4	4	0.96

Furthermore, we have analysed the topological properties of the patterns discovered by the TOPS alignment program, in terms of numbers of SSEs, number of beta strands, number of H-bonds and chiralities, as tabulated in Table 3. We have found that MSAT significantly improves sequence alignments when the average structure compression for the family is less than 0.65 and the TOPS patterns are sufficiently large with many constraints (i.e. H-bonds and chiralities). In the case of Jelly rolls and UB rolls, the alignments generated from MSAT outperform their sequence-driven counterparts (ClustalW and T-Coffee). This observation suggests that the TOPS program can discover distant related protein structures that share a common fold, and hence can generate good structure-driven multiple sequence alignments. This additional (fold) information will be valuable knowledge for biologists interested in detecting protein members with low sequence similarity but that preserve a core common fold.

Generally, MSAT generates a better multiple sequence alignment for folds which are β strand rich (e.g. Class 2 of CATH). This is especially true for Jelly rolls and immunoglobulin-like folds where both have long β sheets in their TOPS patterns. The performance of MSAT for OB folds is not very good compared to ClustalW or T-Coffee alone due to the irregular amino acid length of this fold. Furthermore, the TOPS pattern for OB fold is very small and simple (less constrained) and hence did not contribute much to the alignment. Although alpha-beta plaints are another small fold, the TOPS pattern for this fold is more constrained (with two extra chirality arcs) thus making MSAT extremely sensitive, resulting in a large improvement in the alignment.

One interesting finding from our results is that MSAT only made a modest improvement in the case of TIM barrels and Rossmann folds, although both of these folds have large TOPS patterns as well as significant H-bonds and chirality constraints. This observation does not imply that MSAT is inconsistent in its alignments, but suggests that MSAT is returning biologically relevant results. Several studies have shown that the sequence of TIM barrels is more conserved in loop regions, rather than in the core secondary structure elements. In this context it is relevant that TOPS does not align loop regions, since the inserts in the TOPS patterns represent unaligned SSEs as well as loops, and these mixed blocks are passed on to ClustalW. The TIM barrel is a very common fold that is involved in various biological functions, and one of the conclusions from Nagano et al (2002) concerns patterns "*[…] whose sequences are so diverse that even the most powerful approaches find few relationships, yet whose active sites all cluster at one end of the barrel*". This is also the case for Rossmann fold, where most of the NAD binding sites are located at the C-terminal end of a parallel β sheet (Bell et al 1997). Although MSAT did not improve much on the sequence-based alignment of TIM barrels, this fold represents one of the best predicted fold by all three methods (Table 2); this is probably due to the regular βαβ arrangement of the secondary structure elements.



Alpha-Beta Plait: 9(a)

Jelly Rolls: 9(b)

UB Fold: 9(c)

Immunoglobulin-like (Ig Fold): 9(d)

Rossmann Fold: 9(e)

OB Fold: 9(f)

TIM Barrel: 9(g)

TIM Barrel: 9(h)

Figure 9: ROC Curves *Figure 9(h) shows a clearer representation of Figure 9(g). This has been made possible by widening the FP-rate scale of Figure 9(h)*

CONCLUSIONS

The multiple sequence alignment tool – MSAT, integrates the TOPS alignment program and ClustalW, in a useful framework facilitating the routine automated generation of structure driven sequence alignment. The tool is available as a web-service at http://balabio.dcs.gla.ac.uk/msat/.

Currently we are constructing a database of HMMER profiles and associated TOPS patterns for CATH fold families and developing a system to search a query protein sequence against these profiles in order to perform fold recognition. When a query sequence matches a profile, it can be associated with a corresponding TOPS pattern, as well as being assigned possible membership of a fold family. Thus the system can be used as an initial fold recognition tool, assisting the understanding of the relationship between sequence, structure and function. Our ultimate goal is to be able to automatically recognise the fold of a query protein sequence: the system described here is our first step towards this goal.

Finally, the TOPS alignment program does not at present perform very well on all α families due to the lack of constraint information in the descriptions for these structures. However, work is in progress to improve the TOPS algorithms by taking into account helix packing information which is now incorporated into the TOPS database (Dalton et al, 2003; Michaelopoulos et al, 2004).

ACKNOWLEDGEMENTS

MV and ACT were supported by studentships from University of Glasgow. The TOPS project was supported by a BBSRC-EPSRC grant (BIO14449).

REFERENCES

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25: 3389–3402.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215: 403-410.

Bell, C.E., Yeates, T.O. and Eisenberg, D. (1997) Unusual conformation of nicotinamide adenine dinucleotide (NAD) bound to diphtheria toxin: a comparison with NAD bound to the oxidoreductase enzymes. *Protein Science*, 6: 2084-2096.

Brazma, A., Jonassen, I., Eidhammer, I. and Gilbert, D.R. (1998) Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, 5: 277-303.

Brenner, S.E. and Levitt, M. (2000). Expectations from structural genomics. *Protein Science*, 9:197-200.

Dalton, J.A.R., Michalopoulos, I. and Westhead, D.R. (2003) Calculation of helix packing angles in protein structures. *Bioinformatics* 19: 1298-1299.

Day, R., Beck, D.A.C., Armen, R.S. and Daggett, V. (2003) A consensus view of fold space: combining SCOP, CATH and the DALI domain dictionary. *Protein Science*, 12: 2150-2160.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (1996) Biological Sequence Analysis, Cambridge University Press.

Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics*, 14: 755-763.

Flores, T.P., Moss, D.M. and Thornton, J.M. (1994) An algorithm for automatically generating protein topology cartoons. *Protein Engineering*, 7: 31-37.

Gilbert, D.R., Westhead, D.R., Nagano, N. and Thornton, J.M. (1999) Motif-based searching in TOPS protein topology databases. *Bioinformatics*, 15: 317-326.

Gilbert, D.R., Westhead, D.R., Viksna, J. and Thornton, J.M. (2001) A computer system to perform structure comparison using TOPS representations of protein structure. *Computer and Chemistry*, 26: 23-30.

Goldsmith-Fischman, S. and Honig, B. (2003). Structural genomics: computational methods for structure analysis. *Protein Science*, 12: 1813-1821.

Gribskov, M. and Robinson, N.L. (1996): Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computer and Chemistry*, 20: 25-33,

Hadley, C. and Jones, D.T. (1999) A systematic comparison of protein structure classifications: SCOP, CATH, and FSSP. *Structure,* 7: 1099–1112.

Holm, L. and Sander, C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.*, 25: 231-234.

Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, 22: 2577-2637.

Kim, S.-H. (1998). Shining a light on structural genomics. *Nature Structural Biology,* 5: 643-645.

Mian, I.S. and Dubchak, I. (2000) Representing and reasoning about protein families using generative and discriminative methods. *Journal of Computational Biology*, 7: 849-862.

Michaelopoulos, I., Torrance, G.M., Gilbert, D.R. and Westhead, D.R. (2004) TOPS - An enhanced database of protein structural topology. *Nucleic Acid Res.*, *in press*.

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

Nagano, N., Orengo, C.A. and Thornton, J.M. (2002) One fold with many functions: the evolutionary relationships between tim barrel families based on their sequences, structures and functions. *J. Mol. Biol*, 321: 741-765.

Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol.*, 48:443-453.

Notredame, C., Higgins, D. and Heringa, J. (2000) T-Coffee: A Novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol,* 302: 205-217.

Orengo, C.A. (1993) Classification of protein folds. *Curr. Opin. Struct. Biol.,* 4: 429-440.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. (1997) CATH- a hierarchic classification of protein domain structures. *Structure,* 5: 1093-1108.

Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc Natl Acad Sci,* 85: 2444-2448.

Sayle, R.A. and Milner-White, E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem Sci*, 20: 374.

Slidel, T.W.F and Thornton, J.M. (1996) Chirality in protein structure. In H. Bohr and S. Brunak (ed.), *Protein Folds: a distance-based approach*, p. 253-264. CRC Press.

Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol.,* 147:195-197.

Sternberg, M.J.E. and Thornton, J.M. (1977) On the conformation of proteins: the handedness of the connection between parallel beta-strands. *J. Mol. Biol.,* 110: 269-283.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res,* 22: 4673-4680.

van Rijsbergen, C, J. (1979) Information retrieval. London: Butterworths.

Viksna, J. and Gilbert, D.R., (2001) Pattern matching and pattern discovery algorithms for protein topologies, *Algorithms in Bioinformatics*, *LNCS*, 2149: 98-111.

Westhead, D.R., Hutton, D.C. and Thornton, J.M.(1998) An atlas of protein topology cartoons available on the World Wide Web. *Trends in Biochemical Science,* 23: 35-36.

Westhead, D.R., Slidel, T.W.F., Flores, T.P.J. and Thornton, J.M. (1999) Protein structural topology: automated analysis and diagrammatic representation. *Protein Science,* 8: 897-904.