Motif–based searching in TOPS protein topology databases

David Gilbert[1,3]* David Westhead[2,3] Nozomi Nagano[4] and Janet Thornton[3,4,5]

[1] Department of Computer Science, City University, London EC1V 0HB, UK
[2] Current address: School of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK
[3] European Institute of Bioinformatics, Hinxton, Cambridge CB10 1SD, UK
[4] Department of Biochemistry, University College, London WC1E 6BT, UK
[5] Crystallography Dept., Birkbeck College, London WC1E 7HX, UK

Abstract

Motivation: TOPS cartoons are a schematic abstraction of protein 3-dimensional structures in two dimensions, and are used for understanding and manual comparison of protein folds. Recently an algorithm that produces the cartoons automatically from protein structures has been devised and cartoons have been generated to represent all the structures in the structural databank. There is now a need to be able to define target topological patterns and to search the database for matching domains.
Results: We have devised a formal language for describing TOPS diagrams and patterns, and have designed an efficient algorithm to match a pattern to a set of diagrams. A pattern-matching system has been implemented, and tested on a database derived from all the current entries in the Protein Data Bank (15,000 domains). Users can search on patterns selected from a library of motifs, or alternatively they can define their own search patterns.
Availability: The system is accessible over the Web at `http://tops.ebi.ac.uk/tops`
Contact: drg@cs.city.ac.uk

Introduction

Protein structure comparison and searching An understanding of the similarities and differences between protein structures is very important for the study of the relationship between sequence, structure and function, and for the analysis of possible evolutionary relationships. This has lead to the need for computational methods of structure comparison, and algorithms for searching structural databases. Furthermore, the rapid increase in the size of structural databases means that search and comparison techniques should be both fast and accurate. This area of research is too large to survey here, but see (Orengo 1994) and (Gibrat *et. al.* 1996) for recent reviews.

Protein structures can be described at various levels of detail, ranging from atomic coordinates, through vector approximations to secondary structures elements (SSEs), to 'topological' models. These latter abstractions typically consider a sequence of SSEs, i.e. helices or strands, together with relationships like spatial adjacency within the fold and approximate orientation, neglecting details like lengths and structures of loops, and the lengths of the secondary structure elements themselves.

The topological description has the advantage of simplicity, which makes it possible to implement very fast search and comparison algorithms. Further, by neglecting many of the details which typically vary between related structures, like lengths and structures of loops, and exact lengths, spatial positions and orientations of SSEs, it has the potential to detect more distant structural relationships than could be found by methods based on more geometrical descriptions. On the other hand, its disadvantages are that there may be structures which, although related at the topological level, are very different from a geometric point of view, and have no meaningful biological relationship.

* To whom correspondence should be addressed

TOPS cartoons and diagrams The research described here is concerned with topological descriptions of protein structure based on TOPS diagrams ('cartoons' in their graphical form). TOPS cartoons are used for understanding and manual comparison of protein folds.

In TOPS cartoons, strands are represented by triangles and helices by circles, connected in a sequence from the amino (N) terminus to the carboxy (C) terminus. An example cartoon is shown on the right in Figure 1. This is a pseudo two-dimensional representation of the protein structure, in which a third dimension is implied, and SSEs should be considered to have an approximate direction of 'up' or 'down'. This direction is implied in the way the connecting lines to the symbols are drawn: connections drawn to the edge of a symbol imply connection to the base and those drawn to the centre imply connection to the top, and the direction is that taken by the protein chain from N to C terminus. The direction information is duplicated for strands: upward pointing triangles have the direction 'up' and downward pointing ones the direction 'down'. Although, strictly speaking, many hydrogen bonds exist between the atoms in a pair of strands, at the abstract level of TOPS descriptions each adjacent strand pair is considered to be connected by a single H-bond, having the property of being parallel or anti-parallel, according to the pattern of underlying atomic H-bonds. These H-bonds are not explicitly drawn in TOPS cartoons, but their existence is implied between adjacent strands in such cartoons. Another important aspect of protein structure is that of chirality or handedness. The path traced by the protein chain in three-dimensions is, in general, chiral, that is, it is not super-imposable on its own mirror image. Naturally occurring proteins show preference for certain chiralities, the best known of which is the preference for right–handed connection between two parallel strands within the same sheet (Richardson 1977), (Sternberg and Thornton 1977). The chiralities of these connections, and connections between long parallel helices, are implicit in TOPS cartoons.

We consider in more detail the illustrative example shown in Figure 1, the DNA-binding protein 2bop (Hegde *et. al.* 1992). On the left the structure is illustrated in Molscript style (Kraulis 1991) where extended (strand) type SSEs are represented by light grey ribbons, helices by dark grey helical ribbons and connecting loops by thin cords. It can be seen from the TOPS cartoon, and less clearly from the three-dimensional representation that the structure of 2bop comprises five strands and three helices. The five strands are arranged so that they make a 'sheet' in which the right hand edge is the C terminal strand (the strand nearest the C terminus of the chain). Numbering the SSEs in sequential order with the N terminal strand as number 1 the numbers of the strands in the sheet from the right hand edge are (8,1,6,[4,5]) where strands 4 and 5 have the same position in the sheet, on the left hand edge next to strand 6. The sheet is 'anti-parallel', which means that all pairs of adjacent strands have opposite directions. The helices all lie on one side of the sheet, reflecting the right-handed nature of the connections between strands 1 and 4, and 6 and 8. It is obvious that this type of topological information relating sequential order and relative spatial position is much easier to deduce from the TOPS cartoon than the three-dimensional structure.

TOPS cartoons were originally drawn manually (Sternberg and Thornton 1977); recently an algorithm that produces the cartoons automatically from protein structures has been devised (Flores *et. al.* 1994), (Westhead *et. al.* 1998b), (Westhead *et. al.* 1998a) and implemented as a computer program. Although the cartoons do not explicitly display the information about hydrogen bonding between strands or chirality connections between SSEs, the program outputs a data structure which contains, among other things, the information about these bonds and connections. We refer to this richer representation as a TOPS *diagram*; it is from this form that the cartoons are produced, which can then stored in graphical format (for example postscript or gif files).

Atlas of cartoons A representative set of protein domains has been constructed (Westhead *et. al.* 1998a), based on clustering structures in the structural databank (Bernstein *et al.* 1977), (Abola *et al.* 1987) using the standard single linkage clustering algorithm at 95% sequence similarity, and contains to date over 3000 members. TOPS diagrams and their accompanying cartoons have been generated for this set and are stored in an Oracle database; there is a simple Web interface to the database whereby a user can search for a named member of the Atlas and view the corresponding cartoon.

Motif-based searches We have implemented a topology–based search facility which, given a motif description, searches over a database of TOPS diagrams, and returns the names of the matching diagrams and additional information about the matches. In addition to the Atlas, we have created a database which contains TOPS diagrams for all the domains in the PDB, with over 15,000 entries as of April 1998. A search on a pattern will find all the diagrams which match the pattern, and all the ways in which the pattern can match the diagram.

An important consideration was the efficiency of the matching algorithm, since several TOPS cartoons (i.e. *graphical representations*) can represent the same protein structure. For example, the cartoon in Figure 2 could be drawn in four different equivalent orientations, corresponding to rotations of 180° around the x and y axes. Equally, the exact positions of the circular symbols is not constrained, except to lie on the side of the sheet required by chirality. Because the cartoons are flexible in this way, we decided not to implement a database search which seeks similar *cartoons*, but rather to seek *diagrams* with the same underlying topological constraints, based on H-bonds, chiralities and the types and relative orientations of SSEs.

The system which we have constructed permits users to search on patterns selected from a library of motifs which we have defined; alternatively users can define their own search patterns. There is a Web interface, and the system is accessible over the Web at `http://tops.ebi.ac.uk/tops`

Contents of the paper Until now there has been no formal definition of the diagrams and no definition either informal or formal, of TOPS patterns; in this paper we give definitions for both. We then give an efficient algorithm to match a pattern to a diagram, describe briefly the implementation, and evaluate it on a set of TIM barrels.

Methods

TOPS diagrams A TOPS diagram is a formalisation of a cartoon, based on the underlying topological information from which the cartoon is generated – secondary structure elements (SSEs), H-bonds and chiralities. Such a diagram is a sequence of secondary structures and two associated sets defining relations, or constraints, over elements in the sequence; relations can be of the form of H-bonds or chiralities.

More formally, a TOPS diagram is a triple (E, H, C) where $E = S_1, \ldots, S_k$ is a sequence of length k of secondary structure elements (SSEs) and H and C are relations over the SSEs, called respectively H-bonds and chiralities. In this description an H-bond constraint refers to a ladder of individual hydrogen bonds between adjacent strands in a sheet.

We characterise a TOPS diagram as a *string-graph*, since the sequence E can be regarded as either a string, with two relations H and C over elements of the string, or as a path through a graph whose nodes are $\{S_1, \ldots, S_k\}$, with directed arcs $\{S_1 \to S_2, S_2 \to S_3, \ldots, S_{k-1} \to S_k\}$, and two sets of non-directed arcs H and C over a subset of the nodes. We will later refer to the *length* of a diagram as the length of the sequence E.

In our formalism an SSE S is a character from the alphabet $\{\alpha, \beta\}$ standing for helix and strand respectively. Since each SSE in a TOPS diagram is associated with a direction *up* or *down* we associate a direction symbol, $+$ or $-$, with each letter of our alphabet, giving $\{\alpha_+, \alpha_-, \beta_+, \beta_-\}$, called α-helix-up, α-helix-down, β-strand-up, and β-strand-down, respectively.

Both H-bonds and chiralities are symmetric relations (non-directed arcs in the graph). An H-bond constrains the types of the two SSE's involved to be strands, and each bond is associated with a relative direction $\delta \in \{P, A\}$, indicating whether the bond is between parallel or anti-parallel strands. Chiralities are associated with handedness $\chi \in \{L, R\}$ (left and right respectively), and only occur between pairs of SSEs of the same type. We denote the H-bond relationship between two SSEs S_i and S_j by (S_i, δ, S_j) and a chirality relationship by (S_i, χ, S_j).

As an example, we give a TOPS diagram for 2bop in Figure 2; we can 'stretch out' this diagram to give a linear form, as shown in Figure 4, and represent this TOPS diagram in our notation as follows:

2bop $= (E, H, C)$, where
$E = (\beta_{+_1}, \alpha_{-_2}, \alpha_{-_3}, \beta_{+_4}, \beta_{+_5}, \beta_{-_6}, \alpha_{+_7}, \beta_{-_8})$
$H = \{(\beta_{+_1}, A, \beta_{-_6}), (\beta_{+_1}, A, \beta_{-_8}), (\beta_{+_4}, A, \beta_{-_6}), (\beta_{+_5}, A, \beta_{-_6})\}$
$C = \{(\beta_{+_1}, R, \beta_{+_4}), (\beta_{-_6}, R, \beta_{-_8})\}$

TOPS patterns A TOPS *pattern* (or *motif*) is similar to a TOPS diagram, but is a generalisation which describes several diagrams conforming to some common topological characteristics. This generalisation is achieved by permitting 'gaps', standing for the insertion of SSEs (and any associated H-bond and chiralities), in the sequence of secondary structure elements; indeed a diagram is just a pattern where no inserts are permitted. A gap is described by the number of SSEs inserted.

Formally a TOPS pattern is a triple (T, H, C) where T (referred to as a T-pattern) is a sequence $V_1 - (n_1, m_1) - V_2 - \ldots - (n_{k-1}, m_{k-1}) - V_k$ comprising secondary structure elements indicated by V_i 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. The range of n and m is from zero to the largest number of SSE's in any TOPS diagram (approximately 60); in the following we let $\mathbf{N}$ stand for this large number. This notation for T is very similar to that used by Prosite (Bairoch *et al.* 1996) to describe protein sequences. H are H-bonds and C are chiralities, just as in the diagrams.

In principle, just as for TOPS diagrams, each SSE in a TOPS pattern is associated with a direction *up* or *down* (+ or − respectively), and is a character from the alphabet $\{\alpha, \beta\}$. However, since TOPS diagrams exhibit rotational invariances of 180° about the x and y-axes, therefore associate a *direction variable*, $\oplus$ or $\ominus$ with each SSE in a pattern P s.t. they satisfy the constraint

$$\forall \oplus, \ominus \in P : opp(\oplus, \ominus) \leftrightarrow (\oplus = + \wedge \ominus = -) \vee (\oplus = - \wedge \ominus = +)$$

For example a TOPS pattern which describes plaits (2bop is an instance of a plait) is illustrated in non-linear and linear form in Figures 3 and 5 respectively; arrows between SSEs in the sequence have been annotated with pairs of integers standing for (n_i, m_j), in this case $(0, \mathbf{N})$. The formal definition of this TOPS pattern is below:

Plait $= (V, H, C)$, where
$V = ((\beta_{\oplus_1} - (0, \mathbf{N}) - \alpha_{\ominus_2} - (0, \mathbf{N}) - \beta_{\oplus_3} - (0, \mathbf{N}) - \beta_{\ominus_4} - (0, \mathbf{N}) - \alpha_{\oplus_5} - (0, \mathbf{N}) - \beta_{\ominus_6})$
$H = \{(\beta_{\oplus_1}, A, \beta_{\ominus_4}), (\beta_{\oplus_1}, A, \beta_{\ominus_6}), (\beta_{\oplus_3}, A, \beta_{\ominus_4})\}$
$C = \{(\beta_{\oplus_1}, R, \beta_{\oplus_3}), (\beta_{\ominus_4}, R, \beta_{\ominus_6})\})$

The Matching Algorithm

We have designed a backtracking algorithm which is guaranteed to find all the ways in which a TOPS pattern matches a TOPS diagram; for each match it returns the set of pairs of corresponding nodes between the pattern and the diagram, and the set of corresponding insert sizes in the pattern.

Constraints are used in the algorithm in order to prune the search space, and we assume the existence of a constraint solver for equations and inequations over integers. In general, a variable x over some finite domain can be thought of standing for a set $S_x = \{v_1, v_2, \ldots, v_n\}$ of possible values for x, i.e. $x \in S_x$. During a constrained computation, values are removed from the set, so that $x \in S'_x$, $S'_x \subseteq S_x$. If S'_x is a singleton, then we can write $x = v_j$ instead of $x \in \{v_j\}$, and we say that v_j is a *valuation* for x. If the final set for x is empty, then there is no solution, and the computation fails. Although constraint solving may result in a set of more than one value for x, we have constructed our algorithm so that in the case of a successful computation all constraint variables are given a valuation, i.e. are instances.

Essentially, the matching algorithm initially *imposes constraints* on the SSE numbers in the diagram that can match the SSE numbers in the pattern, and also on the number of inserts between matching SSE numbers in the diagram. This is done in the **Initialise** stage by setting up (i) a correspondence between each node number in the pattern and a range of possible node numbers in the diagram that it may match, and (ii) a list of corresponding insert sizes. The values that the matching node numbers and inserts can take are then *constrained further* by first matching on the H-bonds (**H-match**) and chiralities (**C-match**). Finally sequence matching (**S-match**) *generates* instances of the matching SSE numbers in the diagram, and their associated insert values.

This approach of 'constrain the data, then generate solutions' is much more efficient than a pure backtracking which would attempt to generate all possible sequence matches, and then reject those which do not conform to the constraints imposed by the hydrogen bonds and chriralities.

The algorithm is:

Given :

(1) a pattern $P = (T, H_p, C_p)$
$\Sigma = \{\alpha_\oplus, \alpha_\ominus, \beta_\oplus, \beta_\ominus\}$
$T = V_1 - (n_1, m_1) - V_2 - \ldots - (n_{k-1}, m_{k-1}) - V_k,\ V_j \in \Sigma,\ n_j \leq m_j$
$H_p = \{(S_i, \delta, S_j) | S_i, S_j \in \{\beta_\oplus, \beta_\ominus\},\ \delta = P \leftrightarrow S_i = S_j,\ \delta = A \leftrightarrow S_i \neq S_j\}$
$C_p = \{(S_i, \chi, S_j) | \chi \in \{R, L,\},\ S_i, S_j \in \Sigma\}$
and an associated constraint *opp* (defined above) over the direction variables $\oplus, \ominus$ in P, and

(2) a diagram $Diag = (S, H_d, C_d)$
$\Sigma = \{\alpha_+, \alpha_-, \beta_+, \beta_-\}$,
$E = (S_1, \ldots, S_k),\ S_i \in \Sigma$
$H_d = \{(S_i, \delta, S_j) | S_i, S_j \in \{\beta_+, \beta_-\},\ \delta = P \leftrightarrow S_i = S_j,\ \delta = A \leftrightarrow S_i \neq S_j\}$
$C_d = \{(S_i, \chi, S_j) | \chi \in \{R, L,\},\ S_i, S_j \in \Sigma\}$

Initialise :

(1) $Corr$, the correspondence between node numbers in the diagram and node numbers in the pattern s.t. $Corr := (d_1, d_2, \ldots, d_k)$, where d_i $(i \in 1 \ldots k)$ is a constraint variable representing the node number of the SSE in the diagram matching an SSE with node number i in the pattern,

(2) Ins, the sequence of insert sizes. s.t. $Ins := (I_1, I_2, \ldots, I_{k-1})$, where I_i $(i \in 1 \ldots k-1)$ is a constraint variable representing the number of inserts between nodes i and $i+1$ in the pattern,

with the constraints for $i \in 1..k$,

C1 $0 < d_i \leq \mathbf{N}$
C2 $n_i \leq I_i \leq m_i$,
C3 $d_i + I_i + 1 = d_{i+1}$

Note: Constraint C1 gives the range of d_i; C2 and C3 ensure that the sequence is preserved and that the constraints on the gap size given by I_i are taken into account. **N** represents the size of the largest insert that can be made, in practice less than 60. Thus for example at this stage $1 \leq d_1 \leq d_k - I_1 - I_2 - \ldots - I_{k-1} - k - 1$ and $d_1 + 1 \leq d_2 \leq d_k - I_2 - \ldots - I_{k-1} - k - 2$.

A note on character matching: We assume that each character is represented as a pair (C, U) where C is an SSE type $(C \in \{\alpha, \beta\})$ and U is a direction ($U \in \{+, -\}$ for diagrams, and $U \in \{\oplus, \ominus\}$ for patterns). Character matching between a character in the pattern and a character in the diagram is defined by matching the corresponding SSE types and directions (respecting the the global constraint *opp*).

Perform in order : (omit 1b and 2b for weak matching[6]):

1. **H-match**:
   (a) for each H-bond (V_i, δ, V_j) in H_p, find (U_k, δ, U_l) in H_d such that
      i. V_i character matches U_k, V_j character matches U_l, and
      ii. $d_i = k$ and $d_j = l$ for d_i, d_j in $Corr$, respecting the constraints C1, C2 and C3 on k and l.

   (b) for each H-bond (U_k, δ, U_l) in H_d such that $d_k = i$ and $d_l = j$ are in $Corr$, find (V_i, δ, V_j) in H_p such that V_i character matches U_k, and V_j character matches U_l.
2. **C-match**:
   (a) for each chirality (V_i, χ, V_j) in C_p, find (U_k, χ, U_l) in C_d such that
      i. V_i character matches U_k, and V_j character matches U_l, and
      ii. $d_i = k$ and $d_j = l$ for d_i, d_j in $Corr$, respecting the constraints C1, C2 and C3 on k and l.

   (b) for each chirality (U_k, χ, U_l) in C_d such that $d_k = i$ and $d_l = j$ are in $Corr$, find (V_i, χ, V_j) in C_p, such that V_i character matches U_k, V_j character matches U_l.
3. **S-match**: For each V_i in T find U_k in S such that $d_k = i$ is in $Corr$

Return :

in the case of success, all the possible pairs ($Corr$,Ins), where $Corr$ is the set of corresponding nodes, and Ins is the set of corresponding insert sizes for that match.

Complexity of the pattern matching algorithm: The pattern matching algorithm presented here can be regarded as an algorithm for solving the subgraph isomorphism problem for a special type of ordered graphs. Although the problem for ordered graphs seems to be easier than in the general case, it still can be shown to be NP complete. If we assume a sequence with x hbonds and y chiralities, in the worst case our algorithm will find all possible matches in time $O(2^n/\sqrt{n})$ (where $n = x + y$) which would appear to be too large for sequences from our database.

Running times on real examples are considerably smaller, due to the nature of the diagrams in the database and the patterns sought. For example, in the TOPS Atlas of 3000 diagrams the maximal length of a sequence is 59, the maximal number of H-bonds is 38 and the maximal number of chiralities is 16. In addition, the maximal number of H-bonds from one strand is 5, and the maximal number of chiralities from a node is 3. In addition, typical examples from the database contain very few cycles, which largely appears to be the reason why algorithm works in very reasonable time on practical examples. However, so far we have not been able to obtain a lower theoretical estimate of the upper bound of complexity for types of graphs within our databese.

Constraint pruning by H-bonds and chiralities is quite efficient, and it is an open question which to perform first. H-bonds constrain the SSEs involved to be strands and to have a relative direction. This prunes the search space much faster than trying to match a T-pattern onto a sequence, since T-patterns are much shorter than diagram sequences, and the inserts permit the T-pattern to be matched in many ways. If all SSEs in the pattern are members of the H-bond set (i.e. are H-bonds), then matching is significantly pruned. Our technique is most efficient for all-β or $\alpha\beta$ domains, because all β strands are constrained by one or more H-bonds, and may in addition be constrained by chirality relations; the only possible constraints on α helices are chiralities, and not all helices are thus constrained.

[6] In terms of graph theory, strong matching, i.e. performing 1(a), 1(b), 2(a) and 2(b), corresponds to an induced subgraph isomorphism, whereas weak matching corresponds to a (non-induced) subgraph isomorphism.

Example

We now sketch the matching of the plait pattern to the domain 2bop, using the algorithm above. Given:

Pattern plait : $P = (T, H, C)$, where
$T = ((\beta_{\oplus_1}–(0, \mathbf{N})–\alpha_{\ominus_2}–(0, \mathbf{N})–\beta_{\oplus_3}–(0, \mathbf{N})–\beta_{\ominus_4}–(0, \mathbf{N})–\alpha_{\oplus_5}–(0, \mathbf{N})–\beta_{\ominus_6})$,
$H = \{(\beta_{\oplus_1}, A, \beta_{\ominus_4}), (\beta_{\oplus_1}, A, \beta_{\ominus_6}), (\beta_{\oplus_3}, A, \beta_{\ominus_4})\}$,
$C = \{(\beta_{\oplus_1}, R, \beta_{\oplus_3}), (\beta_{\ominus_4}, R, \beta_{\ominus_6})\})$

Diagram 2bop : $D = (E, H', C')$, where
$E = (\beta_{+_1}, \alpha_{-_2}, \alpha_{-_3}, \beta_{+_4}, \beta_{+_5}, \beta_{-_6}, \alpha_{+_7}, \beta_{-_8})$,
$H' = \{(\beta_{+_1}, A, \beta_{-_6}), (\beta_{+_1}, A, \beta_{-_8}), (\beta_{+_4}, A, \beta_{-_6}), (\beta_{+_5}, A, \beta_{-_6})\}$,
$C' = \{(\beta_{+_1}, R, \beta_{+_4}), (\beta_{-_6}, R, \beta_{-_8})\}$

Initialise $Corr = (d_1, d_2, d_3, d_4, d_5, d_6)$
and $Ins = (I_1, I_2, I_3, I_4, I_5, I_6)$ and for $i \in 1..6$ let $0 < d_i \leq \mathbf{N}$ and $d_i + I_i + 1 = d_{i+1}$, then

(1) H-matching produces the correspondence
$Corr = (1, d_2, d_3, 6, 7, 8)$, $2 \leq d_2 \leq 3, 4 \leq d_3 \leq 5$,
$Ins = (I_1, I_2, I_3, 0, 0)$, $0 \leq I_1 \leq 1, 0 \leq I_2 \leq 2, 0 \leq I_3 \leq 1$

(2) C-matching then constrains the correspondence and insert values further since the first chirality in the pattern set $(\beta_{+_1}, R, \beta_{+_3})$ can only match with the first in the diagram set $(\beta_{+_1}, R, \beta_{+_4})$, and $\oplus$ in the pattern has been bound to $+$ by the H-match:
$Corr = (1, d_2, 4, 6, 7, 8)$, $2 \leq d_2 \leq 3$, $Ins = (I_1, I_2, 1, 0, 0)$, $0 \leq I_1 \leq 1, 0 \leq I_2 \leq 1$

(3) S-matching results in two instances
Match$_1$: $Corr_1 = (1, 2, 4, 6, 7, 8)$, $Ins_1 = (0, 1, 1, 0, 0)$
Match$_2$: $Corr_2 = (1, 3, 4, 6, 7, 8)$, $Ins_2 = (1, 0, 1, 0, 0)$

Implementation

Generating TOPS diagrams TOPS diagrams were originally generated by the TOPS program (Flores *et. al.* 1994). We have recently made extensive modifications to this program in order to improve its robustness and reliability over a much larger set of protein structures (Westhead *et. al.* 1998b). The new program has been used to generate an atlas of TOPS cartoons available over the WWW (http://tops.ebi.ac.uk/tops) (Westhead *et. al.* 1998a). In addition a Java program has been developed which permits display and editing of these cartoons.

A PDB entry is first analysed by the program DSSP (Kabsch and Sander 1983), which locates SSEs and atomic hydrogen bonds, and the TOPS diagrams which are later produced are dependent on the functioning of the DSSP program. TOPS uses this information in a topological analysis which includes the location of sheets, barrels and sandwiches and the positioning of strands within them; analysis of sheet curvature; analysis of connection chirality; and, analysis of spatial neighbour relationships. The results of this analysis are then used, along with a cartoon optimization procedure, to produce a valid TOPS cartoon (Westhead *et. al.* 1998a), (Westhead *et. al.* 1998b). Finally TOPS writes all topological information, along with the coordinate details necessary to produce the cartoon to an ASCII file (a TOPS file).

We have constructed a program to compile TOPS files to a string–graph database; the 6747 TOPS files (total 50MB) describing all the protein structures in the current PDB are compiled into 15400 domain descriptions in graph format (total 11MB) in 11 minutes elapsed time (5 sec CPU time) on a Dec-Alpha.

Searching the database We have implemented a search engine based on our matching algorithm. High-level descriptions have been written for some well-known motifs: left-handed beta-alpha-beta, Greek key, jelly-rolls, plaits, Rossmann type NAD-binding domains, immuno-globulins, barrels and Tim-barrels (perfect and distorted), trefoils, and propellers. For example, a type 1 jelly-roll contains two sheets, each anti-parallel, with hydrogen bonds 1-8-3-6 and 2-7-4-5 and a type 2 jelly-roll contains two sheets, each anti-parallel, with hydrogen bonds 8-1-6-3 and 7-2-5-4. Users can search on these or define their own search motifs.

The entire system has been written in the finite domain constraint logic programming system clp(FD) (Codognet and Diaz 1996), which incorporates a finite domain solver over integers. This system compiles source programs into stand-alone executable code whose efficiency approaches that of programs written in an imperative language such as 'C'. Constraint operations over integers can be encoded quite naturally, and the constraint solver permits pruning of the search space when a match is being sought; moreover the backtracking mechanism of the logic programming engine permits us to return all solutions to our algorithm, i.e. all matches, in an elegant manner.

Our work has been implemented and tested on a realistic data-set, namely all 15000 descriptions of protein domains currently in the Brookhaven Protein Data Bank (PDB) (Bernstein *et al.* 1977), (Abola *et al.* 1987) (April 1998). A database paging mechanism has been developed in order to overcome memory restrictions when the database based on the entire PDB is to be queried; loading overheads are approximately 3sec/MB, i.e. 33 sec for the PDB database. We have evaluated the system using the CATH database (Orengo *et. al.* 1997) at University College, London (see 'Results' below), and the structural domain definitions for our database are those as defined by that database.

The match algorithm can output parameters such as the number of inserts in the matched domain – this measure increases as proteins diverge as a result of evolution. Results can be sorted on a these parameters and the node numbers of the SSEs matching the description can also be output. A Web interface to the system has also been constructed using the PiLLoW routines (Cabeza *et. al.* 1996) which we have ported to clp(FD) (Gilbert *et. al.* 1997).

The system is fast, with an average time per attempted match of 3ms on a Dec-Alpha. Typical search times over the TOPS Atlas are, for example, 8.4 seconds to find all matches to the jelly-roll type 1 description (181 hits out of 2900 domains) on a Dec-Alpha. Some queries take advantage of precompiled topological information generated by the TOPS program; for example a search for Tim barrels takes 1.2 seconds to find 50 matches in the database. A domain can be tested against all motif descriptions to see which it matches and by how much. To run this battery test over all entries in the Atlas takes 68 seconds. Query times scale linearly with the size of the database.

Results In order to evaluate our system we chose some simple examples of topological searches which might be carried out by a typical user. The first such search was for protein domains with the so-called TIM barrel topology. This protein fold comprises a central eight-stranded parallel

β barrel in which the strands are positioned in sequential order around the barrel. Within the connections between each adjacent pair of strands are zero or more helices. In order to find such domains we set up a query for eight strands in sequence with a total of eight parallel hydrogen bonds between the following eight strand pairs: (1,2), (2,3), (3,4), (4,5), (5,6), (6,7), (7,8), (8,1), The helices were not specified because knowledge of protein structure indicated that they would not add specificity to the query. This query is referred to as TIM8PAR.

Secondly, we constructed a query such that of the 8 strands in the barrel, one was in a different direction to the rest; we refer to this query as ENOL, since it should identify domains in the enolase family. Finally we constructed a query called TIM which found all those domains identified either by TIM8PAR or ENOL.

We ran our queries over a reference list (subset of the PDB) of 263 domains which had been previously identified as TIM barrels by using an algorithm based on shear-number analysis (Nagano *et. al* 1998) and which also uses SSTRUC (Smith 1989), an implementation of the DSSP (Kabsch and Sander 1983) algorithm. The query TIM identified 235 of these domains, of which 226 were identified by TIM8PAR and 9 by ENOL; 28 were not identified by TIM. We then developed a test for distorted TIM barrels where the barrel is not complete due to one or more missing H-bonds, and called this test TIMVAR. This test was run on the test set, and identified all remaining 28 members as distorted barrels.

Finally, we ran our queries over the entire structural databank (15361 domains). We identified 260 domains using TIM, of which 250 were identified using using TIM8PAR and 10 using ENOL.

The search TIMVAR identified 279 domains. The results of running the tests over the Atlas and the PDB are presented in Table 1.

Of the 260 domains in the PDB identified as complete TIM barrels by our method TIM, 10 were not identified as complete TIM barrels by the method based on shear numbers. In 9 of these the shear number is not 8, thus preventing their automatic identification as complete barrels (there are some exceptional TIM barrels with shear numbers of 7 or 9 (Nagano *et. al* 1998)). In the remaining case (1oneA2) bulge residues on the barrel sheet prevented detection using shear numbers.

Discussion

We have developed a reduced representation of protein structure, based on a sequence of secondary structure elements, hydrogen bonds between strand type elements, and connection chiralities. This representation enables us to describe in a formal way topological representations of protein structures and motifs.

This language constitutes a new formal way of describing protein structural topology, which extends previous approaches (Richardson 1977), (Rawlings *et. at.* 1985), (Koch *et. al.* 1992), (Flower 1994a), (Flower 1994b), (Flower 1995), (Koch *et. al.* 1996), (Koch and Lengauer 1997), (Tuskamoto *et. at.* 1997). Our language is more expressive than Richardson formulae (Richardson 1977), which can only describe single sheets, and also than that of Flower (Flower 1994a), (Flower 1994b). As with Richardson formulae, the latter does not permit the explicit description of sandwiches, nor of α helices and chiralities. To our knowledge there are no other practical systems which have implemented searches of protein structural databases using motifs based on a formal language. The system reported by (Rawlings *et. at.* 1985) is perhaps most similar to ours, but relied on some hand-compilation of the very small Prolog database, did not employ constraints to prune the search space, and was relatively slow.

Existing motif search systems are based on developments of structure comparison programs, where a 'probe' (representative or core structure) is compared with a set of target structures. The essential difference between searching based on matching and searching based on comparison is that the former only returns a result for those structures in the target set which exactly match the probe pattern, whereas comparison outputs a result (distance) for each structure compared with the probe. In work reported separately, we have designed a system to compare TOPS diagrams, based on making a structural alignment using a common discovered pattern.

The simplicity of the representation has allowed us to develop a very fast algorithm to match a motif to a protein description, exploiting constraints generated by H-bonds and chiralities to prune the search space. Using this algorithm, searches over the entire structural databank are possible interactively over the WWW. The principal disadvantage of the reduced representation is that real protein structures do not always contain all the hydrogen bonds that would be expected in the ideal case. This is illustrated by the need to use the TIMVAR query in order to find distorted TIM barrels not containing the ideal eight hydrogen bonds. However, speed is a great advantage, and our experiments show that carefully constructed queries can find all related domains.

Acknowledgments

The authors wish to thank Alvis Brazma of the EBI for his invaluable suggestions, Juris Viksna of the University of Latvia for his analysis of the complexity of the algorithm, Dan Hatton for his painstaking testing, and the CATH group at UCL for their help in using CATH.

David Gilbert has been partially supported by an EPSRC Visiting Fellowship at the European Bioinformatics Institute whilst on sabbatical from City University.

References

E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle, and J. Weng (1987): Protein Data Bank. In F. H. Allen, G. Bergerhoff, and R. Sievers, editors, *Crystallographic Databases-Information Content,* *Software Systems, Scientific Applications*, pages 107–132. Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

A. Bairoch, P. Bucher, and K. Hofman (1996): The PROSITE database, its status in 1995. *Nucleic Acids Research*, 24(1):189–196.

F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi (1977): The Protein Data Bank: a Computer–based Archival File for Macromolecular Structures. *J. Mol. Biol.*, 112:535–542.

D. Cabeza, M. Hermenegildo, and S. Varma (1996): The PiLLoW/CIAO Library for INTERNET/WWW Programming using Computational Logic Systems. In *Proceedings of the 1st Workshop on Logic Programming Tools for INTERNET Applications*, pages 72–90, JICSLP'96, Bonn. Text and code available from http://www.clip.dia.fi.upm.es/miscdocs/pillow/pillow.html.

P. Codognet and D. Diaz (1996): Compiling constraints in clp(FD). *Journal of Logic Programming*, 27(3):185–226

T.P. Flores, D.M. Moss, and Thornton J.M (1994): An algorithm for automatically generating protein topology cartoons. *Protein Engineering*, 7(1):31–37, 1994.

D.R. Flower (1994): Automating the identification and analysis of protein beta barrels. *Protein Engineering*, 7(11):1305–1310.

D.R. Flower (1994): β-Sheet toplogy: A new system of nomenclature. *FEBS Letters*, 344(14007):247–250.

D.R. Flower (1995): FOLD: Integrated analysis and display of protein secondary secondary structure. *Journal of Molecular Graphics*, 13:377–384.

J-F. Gibrat, T. Madej, and S. H. Bryant (1996): Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, 6:377–385.

D. Gilbert, I. Eidhammer, and I. Jonassen (1997): StructWeb: Biosequence structure searching on the Web using clp(FD). Technical Report 1997/05, Department of Computer Science, City University, London, UK ISSN 1364-4009.

R. S. Hegde, S. R. Grossman, L. A. Laimins, and P. B. Sigler (1992): Crystal structure at 1.7 A of the bovine papillomavirus-1 E2 DNA–binding domain bound to its DNA target. *Nature*, 359(6395):505–512, Oct 1992.

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

I. Koch, F. Kaden, and J. Selbig (1992): Analysis of protein sheet topologies by graph theoretical methods. *Proteins: Structure, function and genetics*, 12:314–323.

I. Koch and T. Lengaue (1997): Detection of distant structural similarities in a set of proteins using a fast graph-based method. In T. Gaasterland et al, editor, *Proceedings of the 5th International Conference on Intelligent Systems in Molecular Biology*, pages 167–178. AAAI Press.

I. Koch, T. Lengauer, and Wanke E. (1996): An algorithm for finding maximal common subtopologies in a set of protein structures. *Journal of Computational Biology*, 3(2):289–306.

P.J. Kraulis (1991): MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *Journal of Applied Crystallography*, 24:946–950, 1991.

N. Nagano, G. Hutchinson, and J.M. Thornton (1998): Barrel structures in proteins - Automatic identification and classification including a sequence analysis of TIM barrels *Protein Science*, 1998. to be submitted.

C. Orengo (1994): Classification of protein folds. *Current Opinion in Structural Biology*, 4:429–440.

C. A. Orengo, A. D. Michie, D. T. Jones, M. B. Swindells, and J. M. Thornton (1997): CATH – a hirearchic classification of protein domain structures. *Structure*, 5(8):1093–1108.

C.J. Rawlings, W.R. Taylor, J. Nyakairu, J. Fox, and M.J.E. Sternberg (1985): Reasoning about protein topology using the logic programming language PROLOG. *Journal of Molecular Graphics*, 3(4):151–157.

J.S Richardson (1977): Beta-sheet topology and the relatedness of proteins. *Nature*, 268:495–500.

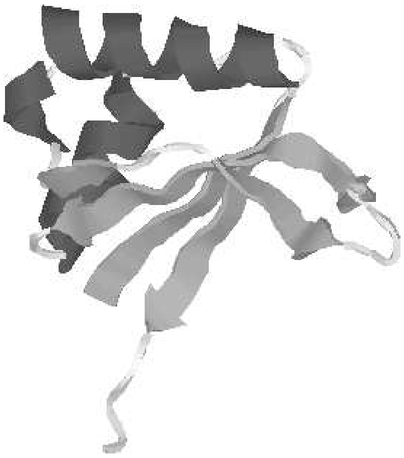
D Smith (1989): Secondary structure calculation program. personal communication.

M.J.E. Sternberg and J.M. Thornton (1977): On the conformation of proteins: The handedness of the connection between parallel beta strands. *Journal of Molecular Biology*, 110:269–283, 1977.

Y. Tsukamoto, K. Takiguchi, K. Satou, T. Furuichi, E. Takagi, and S. Kuhara (1997): Application of a deductive database system to search for topological and similar three-dimensional structures in protein. *CABIOS*, 13(2):183–190, 1997.

D.R. Westhead, T.W.F. Slidel, T.P.J. Flores, and Janet M. Thornton (1998): Protein structural topology: automated analysis, diagrammatic representation and database searching. *Protein Science*, submitted.

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



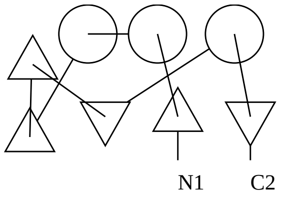


Fig. 1. Rasmol diagram (left) and TOPS cartoon (right) for 2bop

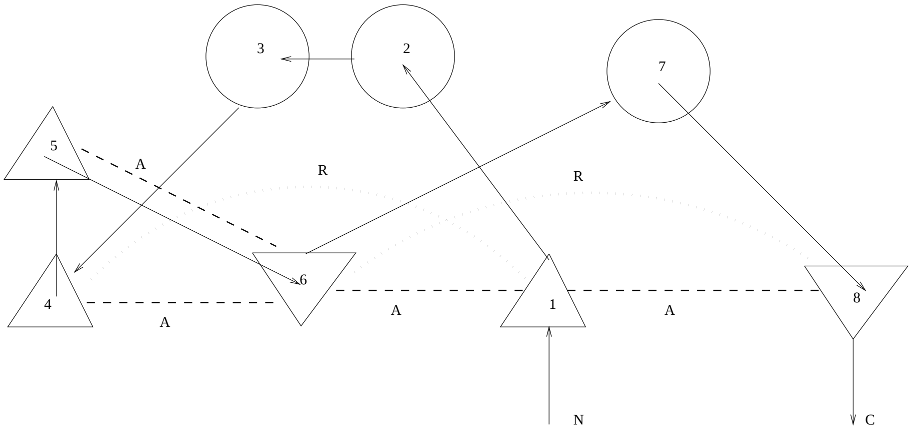


Fig. 2. TOPS diagram for 2bop

This article was processed using the LaTeX macro package with LLNCS style

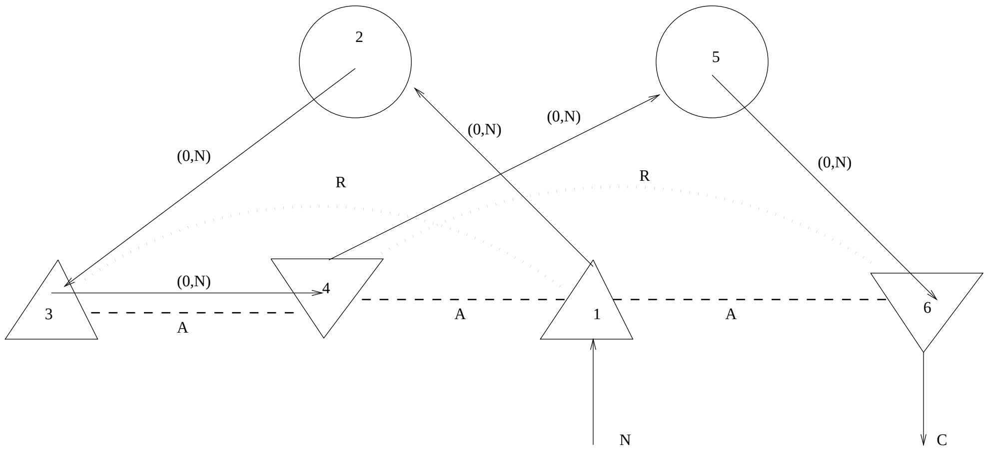


Fig. 3. TOPS diagram for the plait motif

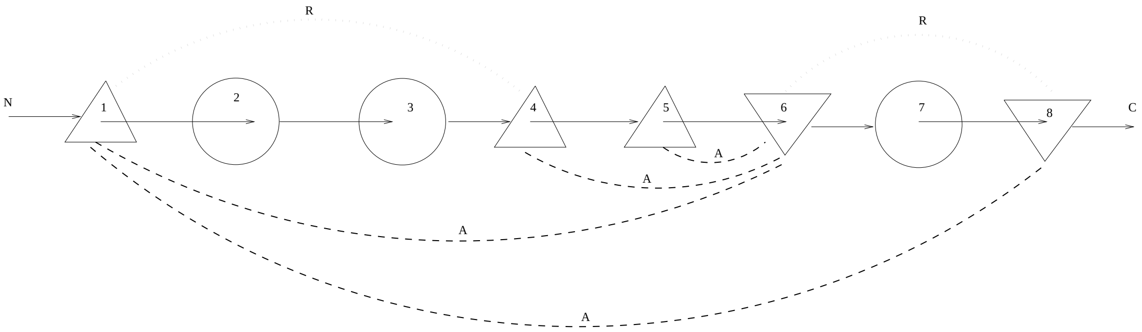


Fig. 4. Linearised TOPS diagram for 2bop

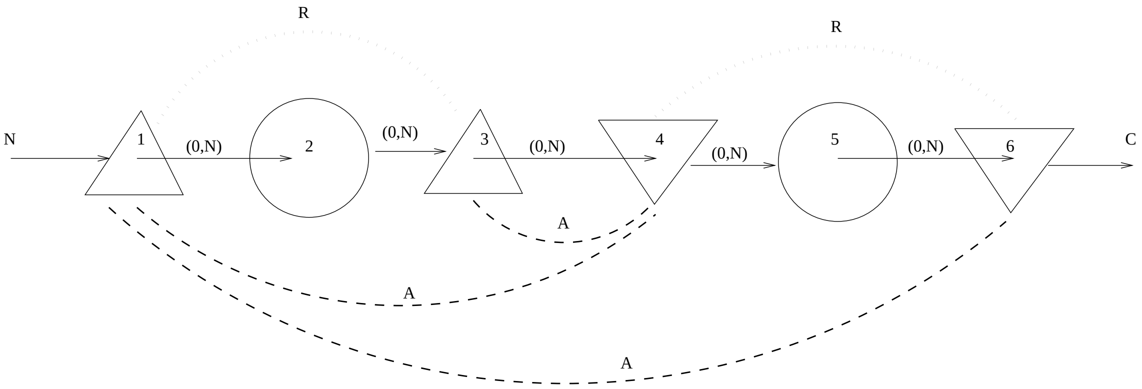


Fig. 5. Linearised TOPS diagram for the plait motif

Query	Database	N_{hits}	Search Time (ms)
TIM	Atlas	50	1196
TIM8PAR	Atlas	49	723
ENOL	Atlas	1	739
TIMVAR	Atlas	44	1178

Query	Database	N_{hits}	Search Time (ms)
TIM	PDB	260	7785
TIM8PAR	PDB	250	4497
ENOL	PDB	10	4528
TIMVAR	PDB	279	6873

Table 1. Results of topological searches for TIM barrels over the Atlas (3000 domains, left) and the the entire structural databank (15361 domains, right). Query names are explained in the main text. N_{hits} gives the total number of domains found containing the query topological pattern. Times exclude database load overhead.