Topology–based protein structure comparison using a pattern discovery technique

David Gilbert[1,2] ; David Westhead[3] ; Juris Viksna[1,4] ; Janet Thornton[2,5,6]
[1] Department of Computer Science; City University, Northampton Square, London EC1V 0HB, UK
[2] European Institute of Bioinformatics; Hinxton, Cambridge CB10 1SD, UK
[3] School of Biochemistry and Molecular Biology; University of Leeds, Leeds LS2
[4] Institute of Mathematics and Computer Science; University of Latvia, Riga LV-1459, Latvia
[5] Department of Biochemistry; University College, London WC1E 6BT, UK
[6] Crystallography Dept.; Birkbeck College, London WC1E 7HX, UK
[1]drg@soi.city.ac.uk ; [3]westhead@bmb.leeds.ac.uk ; [4]jviksna@cclu.lv ; [5]thornton@biochem.ucl.ac.uk

Abstract

We describe the design and implementation of a fast topology–based method for protein structure comparison. The approach uses the TOPS topological representation of protein structure, aligning two structures using a common discovered pattern and generating measure of distance derived from an insert score. Heavy use is made of a constraint-based pattern matching algorithm for TOPS diagrams that we have designed. The system is maintained at the European Bioinformatics Institute and is available over the Web via the at tops.ebi.ac.uk/tops. Users submit a structure description in Protein Data Bank (PDB) format and can compare it with structures in the entire PDB or a representative subset of protein domains, receiving the results by email.

Keywords: structure comparison, constraints, pattern matching, pattern discovery, protein motifs, protein topology. [1]

1 Introduction

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; furthermore, the rapid increase in the size of structural databases means that techniques to compare a given structure with member of such a database should be fast.

Various structure comparison methods have emerged, ranging from those which make detailed geometrical comparisons of backbone coordinates Taylor and Orengo (1989), through methods using vector approximations to secondary structure elements, or SSEs, Mitchell et al. (1989); Grindley et al. (1993); Artymuik et al. (1994), and finishing with methods based on highly simplified models of structure Koch et al. (1996); Koch and Lengauer (1997); Tsukamoto et al. (1997). These latter methods typically consider a sequence of SSEs, along 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. This type of description of a protein structure is commonly known as a 'topological' description.

The topological description has the advantage of simplicity, which makes it possible to implement very fast 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

[1]Proceedings of the AISB-00 Symposium on AI in Bioinformatics , ISBN 1 902956 12 X, pp 11-17, published by The Society for the Study of Artificial Intelligence and the Simulation of Behaviour, 19-20 April 2000

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.

2 TOPS diagrams and patterns

TOPS cartoons were originally drawn manually Sternberg and Thornton (1977) and comprise graphical representations of secondary structure elements (SSEs), their relative orientations and some indication of spatial adjacency. Subsequently a richer representation of the topological structure has been devised Flores et al. (1994); Westhead et al. (1999, 1998), termed a TOPS *diagram*, which includes information about hydrogen bonding between strands and chirality connections between SSEs; this representation is used to automatically produce graphical cartoons.

We have previously described in detail our formal representation of TOPS diagrams and patterns as graphs, and the design of a fast pattern matching program Gilbert et al. (1999). In this paper we describe a pattern discovery algorithm for TOPS diagrams and show how we use it to structurally align diagrams and compute a comparison measure.

TOPS diagrams In TOPS diagrams (for example the diagram for 2bop in Figure 1), strands are represented by triangles and helices by circles, connected in a sequence from the amino (N) terminus to the carboxy (C) terminus. SSEs are considered to have a direction of 'up' or 'down', 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'. The existence of hydrogen bond ladders between a pair of strands is indicated by a single H-bond in the TOPS representation, labelled as being parallel or anti-parallel, according to the relative directions of the two strands that it joins. In addition, TOPS diagrams also represent the chiralities of connections between connections between two parallel strands within the same sheet and connections between long parallel helices. A more detailed description of TOPS diagrams can be found in Gilbert et al. (1999).

More formally, a TOPS diagram is a triple (S, H, C) where $S = S_1, \ldots, S_k$ 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 will later refer to the *length* of a diagram as the length of the sequence S.

In our formalism an SSE 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 $\{\alpha_+, \alpha_-, \beta_+, \beta_-\}$.

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).

The formal definition of a TOPS diagram $D = (S, H_d, C_d)$, given $\Sigma = \{\alpha_+, \alpha_-, \beta_+, \beta_-\}$, is
$S = (S_1, \ldots, S_k),\ S_i \in \Sigma$
$H_d = \{(S_i, \delta, S_j) | S_i, S_j \in \{\beta_+, \beta_-\},\ \delta = P \leftrightarrow S_i = S_j,\ \delta = A \leftrightarrow S_i \neq S_j\}$
$C_d = \{(S_i, \chi, S_j) | S_i, S_j \in \Sigma,\ \chi \in \{R, L,\}\}$

As an example, consider the TOPS diagram for 2bop in Figure 1; we can 'stretch out' this diagram to give a linear form, as shown in Figure 3, and represent it formally as 2bop = (S, H, C), where
$S = (\beta_{+1}, \alpha_{-2}, \alpha_{-3}, \beta_{+4}, \beta_{+5}, \beta_{-6}, \alpha_{+7}, \beta_{-8})$
$H = \{(\beta_{+1}, A, \beta_{-6}), (\beta_{+1}, A, \beta_{-8}), (\beta_{+4}, A, \beta_{-6}), (\beta_{+5}, A, \beta_{-6})\}$
$C = \{(\beta_{+1}, R, \beta_{+4}), (\beta_{-6}, R, \beta_{-8})\}$

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 specifying the insertion of SSEs (and any associated H-bond and chiralities) into the sequence of secondary structure elements; indeed a diagram is just a pattern where no inserts are permitted. The length of an insert is constrained to be within the range of the lengths of the sequences that can be inserted. A TOPS pattern is thus a triple, similar to that of a TOPS diagram; in this case, however, we refer to the sequence of SSEs with inserts permitted as *T-pattern*. The inserts are similar to wild cards with length constraints; we extend the definition of TOPS patterns given in Gilbert et al. (1999) to permit such wild cards before the beginning of, and after the end of the sequence of SSEs.

Formally a TOPS pattern is a triple (T, H, C) where T (referred to as a T-pattern) is a sequence $(n_0, m_0) - V_1 - (n_1, m_1) - V_2 - \ldots - (n_{k-1}, m_{k-1}) - V_k - (n_k, m_k)$ comprising secondary structure elements indicated by V_i and between each of these an insert description, as well as an insert description (n_0, m_0) before V_1 and also an insert (n_k, m_k) after V_k. 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). H are H-bonds and C are chiralities, just as in the diagrams. Since TOPS diagrams exhibit rotational invariances of 180° about the x and y-axes, we associate a *direction variable*, $\oplus$ or $\ominus$ with each SSE in a pattern P s.t. they satisfy the constraint
$\forall \oplus, \ominus \in P : opp(\oplus, \ominus) \leftrightarrow (\oplus = + \wedge \ominus = -) \vee (\oplus = - \wedge \ominus = +)$

The formal definition of a TOPS diagram pattern $P = (T, H_p, C_p)$, $\forall \oplus, \ominus \in P : opp(\oplus, \ominus)$, given $\Sigma = \{\alpha_\oplus, \alpha_\ominus, \beta_\oplus, \beta_\ominus\}$ is:
$T = (n_0, m_0) - V_1 - (n_1, m_1) - V_2 - \ldots - (n_{k-1}, m_{k-1}) - V_k - (n_k, m_k)$, $V_j \in \Sigma$, $n_j \leq m_j$
$H_p = \{(S_i, \delta, S_j) | S_i, S_j \in \{\beta_\oplus, \beta_\ominus\},\ \delta = P \leftrightarrow S_i = S_j,\ \delta = A \leftrightarrow S_i \neq S_j\}$
$C_p = \{(S_i, \chi, S_j) | \chi \in \{R, L, \},\ S_i, S_j \in \Sigma\}$

For example a TOPS pattern which describes plaits, of which 2bop is an instance, is given by Plait = (V, H, C), where
$V = ((0, \mathbf{N}) - \beta_{\oplus_1} - (0, \mathbf{N}) - \alpha_{\ominus_2} - (0, \mathbf{N}) - \beta_{\oplus_3} - (0, \mathbf{N}) - \beta_{\ominus_4} - (0, \mathbf{N}) - \alpha_{\oplus_5} - (0, \mathbf{N}) - \beta_{\ominus_6} - (0, \mathbf{N}))$
$H = \{(\beta_{\oplus_1}, A, \beta_{\ominus_4}), (\beta_{\oplus_1}, A, \beta_{\ominus_6}), (\beta_{\oplus_3}, A, \beta_{\ominus_4})\}$
$C = \{(\beta_{\oplus_1}, R, \beta_{\oplus_3}), (\beta_{\ominus_4}, R, \beta_{\ominus_6})\})$

Figures 2 and 4 illustrate this in non-linear and linear form respectively.

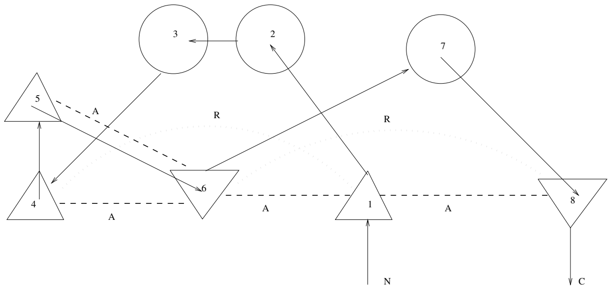

Figure 1: TOPS diagram for 2bop

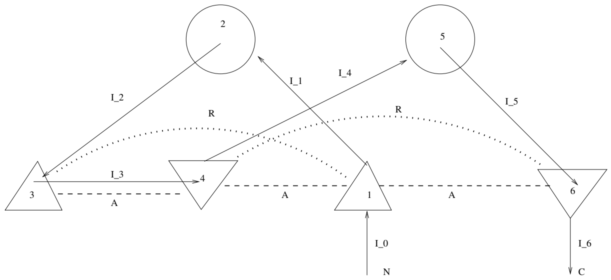

Figure 2: TOPS diagram for the plait motif

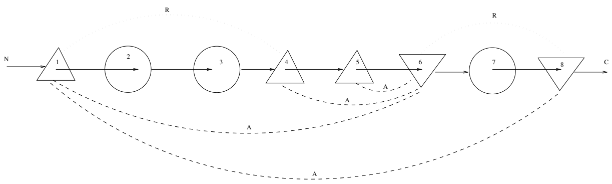

Figure 3: Linearised TOPS diagram for 2bop

3 Methods

We have designed a measure to compare the similarity between two TOPS diagrams, in order to be able to perform structure comparison at the topological level. Our method works by performing a structural alignment of the SSEs of the diagrams and computing a score based on an edit distance over aligned blocks of SSEs plus contributions from the H-bond and chirality sets of the diagrams. In order to perform the alignment we use a least general common pattern generated by a pattern discovery technique which we have designed; this in turn makes heavy use of our constraint-based pattern matching method for TOPS diagrams.

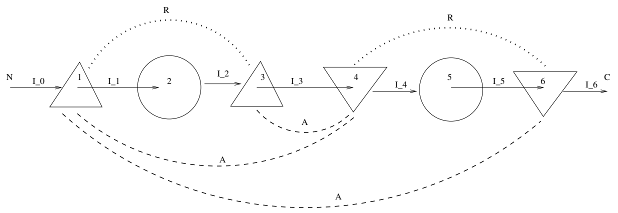

Figure 4: Linearised TOPS diagram for the plait motif

3.1 Pattern discovery for TOPS diagrams

Pattern discovery for sequences is a well-established technique Brazma et al. (1998) which could be applied to TOPS diagrams and patterns as follows. The first, "pattern driven" (PD) is based on enumerating candidate patterns in a given solution space and picking out the ones with high fitness; the second, "diagram driven" (DD) comprises algorithms that try to find patterns by comparing given diagrams and looking for local similarities between them. In the equivalent of DD for sequences, an algorithm may be based on constructing a local multiple alignment of given sequences and then extracting the patterns from the alignment by combining the segments common to most of the sequences.

Essentially the difference between pattern discovery for sequences and TOPS diagrams is that techniques for the former assume that the grammar of the former is regular whilst that of the latter is context–sensitive due to the fact that H-bond and chirality arcs may cross (i.e. they describe a "copy language"). Thus in a naive version of a PD approach for TOPS diagrams not only would we have to enumerate an exponentially large number of patterns comprising not only all the possible combinations of the SSEs (and their orientations) in a pattern of length k, but also all the possible H-bond and chirality connections over them.

Our algorithm discovers patterns of H-bonds (and chiralities) based on the properties of sheets for TOPS diagrams; we also derive T-patterns, i.e. the associated sequences of SSEs and insert sizes. 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 which 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 chirality arcs between the H-bonds in the pattern are then discovered by a similar process. The numbers of inserts between each strand in the pattern are then computed for all the patterns in the learning set, and the minimum and maximum size of the gaps in the corresponding insert positions in the pattern are thus found, and combined with the SSE sequence to give the T-pattern. The result is the least general common TOPS pattern characterising the target set of protein descriptions.

Naive insertion of a new SSE into an existing sequence of SSEs is expensive: consider the case when the existing sequence is of length 2. The new H-bond can be inserted at the beginning of the sequence, at the end of the sequence or between the existing two SSEs. Moreover, a new H-bond must be discovered between the new SSE and one of the existing SSEs in the sequence. We use a 'seed' derived from one of the target set of diagrams in order to give the insertion point: the H-bond pattern is extended in one diagram first by selecting one of the remaining bonds from the diagram H-bond set; if this fails to give a pattern which matches the other diagram, then an alternative bond is selected.

An alternative approach would be to adapt that of Koch et al Koch et al. (1996), which constructs an edge product graph for two graphs and then employs Bron and Kerbosch's algorithm Bron and Kerbosch (1973) which enumerates all the maximal cliques in the graph. Although Koch et al improve Bron and Kerbosch's algorithm by restricting the search process to cliques representing connected substructures, they determine common substructures in more than two topology graphs by forming the intersections between all substructures of all cliques resulting from a pairwise comparison.

The worst-time complexity for the learning algorithm based on repeated matching is approximately $O(k * n^n)$, where k is the number of sequences, and n the number of secondary structures (helices and strands) in a sequence. The maximal clique method has complexity $O((n^k/c_k)!)$ (with little information about c_k, except $c_k \geq 1$) for the same n and k. These are approximations assuming that number of nodes is approximately the same as the number of edges — this is more or less true in TOPS. In terms of implementation, the clique algorithm (for $k = 2$) tends to be slower (up to 10 times) in comparison with the repeated matching algorithm, although it sometimes

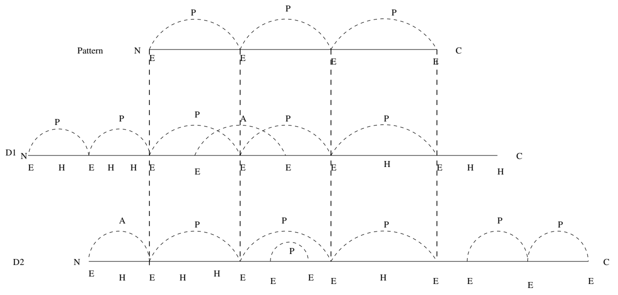

Figure 5: Making an alignment

produces better results.

We use a variant of the repeated matching algorithm to discover common patterns in all-α domains, where patterns of chirality arcs are discovered only between discovered strands.

Distance measure

Given two TOPS diagrams $D1 = (S1, H1, C1)$, $D2 = (S2, H2, C2)$ and a least general common pattern $P = (SP, HP, CP)$, we can make a structural alignment of S1 and S2 by matching P with $D1$ and $D2$. If $length(SP) = N$, then there are $N + 1$ insert positions in the pattern, corresponding to $N + 1$ blocks of unaligned SSEs in $D1$ and $S2$. An example is illustrated in Figure 5, where aligned blocks in $S1$ and $S2$ are indicated by $S1_1 \ldots S1_5$ and $S2_1 \ldots S2_5$ respectively.

The distance measure M between $D1$ and $D2$ is given by the normalised sum of the edit distances (Levenshtein (1965)) of all the blocks plus a contribution from the extra (when compared with the pattern) H-bonds and chiralities in the diagrams.

We have evaluated our method by performing a pairwise comparison of 1396 domains from the SCOP PDB40d database Murzin et al. (1995) and computed the error versus coverage data using the SCOP numbers as an indication of structural homology. Two domains are defined as homologous if at least their first three SCOP numbers are identical; the domains are non-homologous if only their first SCOP numbers are identical. Matches between domains with with only the first two SCOP numbers identical are ignored (not performed) since the SCOP hierarchy does not differentiate homologous and non-homologous pairs at this level. Coverage versus error results are given in Figure 6.

Times per comparison pair are typically 30–400ms on average (DEC Alpha).

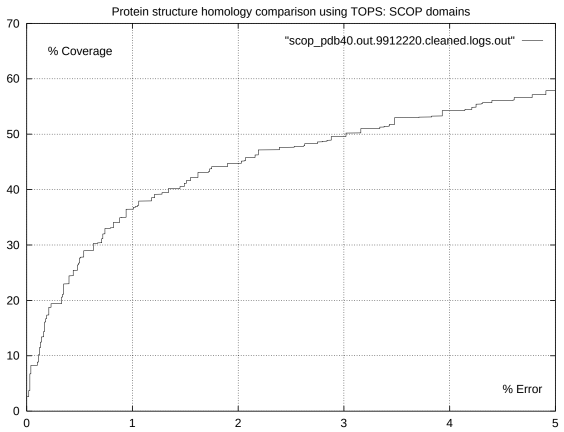

Figure 6: Coverage vs error

System availability: structure comparison server

The comparison system can be used via the Web at tops.ebi.ac.uk/tops. Target structures can be compared against either a database of TOPS diagrams corresponding to all the domains currently in the PDB (currently over 24000 domains) or with a representative subset (the TOPS Atlas Westhead et al. (1998)), 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 containing to date over 3000 members.

Users upload a target structure description in PDB format, select a database against which to compare, and enter their email address in order to receive the result. The target description is first analysed using the DSSP program Kabsch and Sander (1983) which locates SSEs and atomic hydrogen bonds. The TOPS program Flores et al. (1994); Westhead et al. (1999) uses this information in a topological analysis which includes analysis of connection chirality; the resulting file is then translated into a TOPS diagram in logic programming format by a compiler we have written in clp(FD) Codognet and Diaz (1996). The comparison is then performed off-line, the result of each comparison comprising the distance measure, the name of the domain compared, and its hierarchic classification according to the CATH system developed at UCL Orengo et al. (1997). The output is sorted by distance from the target protein, and returned to the user by email. Users may also request the output for each comparison to be annotated with the numbers of the corresponding residues and also the common discovered pattern.

The system is fast; a comparison of one structure against the entire PDB (15000 domains) takes from under 10 minutes to 1 hour or more on a DEC Alpha, depending on the complexity of the structure submitted.

4 Conclusions

Although our pattern discovery algorithm produces the richest patterns over α–β domains, when both H-bond and chirality connections can be discovered, it also discovers patterns of H-bonds for all-β domains and patterns of chiralities for all-α domains. However, the null pattern will be discovered when comparing two all-α domains with no chirality information, and thus in this case neither an alignment no a meaningful comparison measure can be computed. The accuracy of the system as measured by coverage against error falls in between those for a well-performing atom-coordinate approach (ranging from 60% coverage at 1% error to 78% coverage at 5% error) and sequence-based approaches (ranging from 16% coverage at 1% error to 18% coverage at 5% error).

A disadvantage of the topological approach is that no RMSD output can be made - the best that can be done is to return the numbers of the matching residues of the matching SSEs, which is not a one to one relationship between residues, but rather between between SSEs which are potentially of different lengths. However, an advantage of our pattern–based declarative approach is that the patterns can be returned to the user - these contain more information than is conveyed by the comparison score alone, for example that both pattern contained a complete barrel.

Finally, our pattern discovery algorithm can be used to make multiple alignments of TOPS structures, since it is is linear time in the number of members of the target set.

Acknowledgements

The authors wish to thank Alvis Brazma of the EBI for his invaluable suggestions, 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. 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, 1987.

P.J. Artymuik, H.M. Grindley, A.R. Poirette, D.W. Rice, E.C. Ujah, and P. Willett. Identification of β-Sheet motifs, of ψ-loops, and of patters of amino acid residues in three dimensional protein structures using a subgraph isomorphism algorithm. *J. Chem. Inf. Comput. Sci.*, 34:54–62, 1994.

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. The Protein Data Bank: a Computer–based Archival File for Macromolecular Structures. *Journal of Molecular Biology*, 112:535–542, 1977.

A. Brazma, I. Jonassen, I Eidhammer, and D. R. Gilbert. Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, 5(2):277–303, 1998.

C. Bron and J. Kerbosch. Algorithm 457: Finding All Cliques of an Uundirected Graph. *CACM*, 16 (9):575–577, 1973.

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

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

D. R. Gilbert, D. R. Westhead, N. Nagano, and J. M. Thornton. Motif–based searching in tops protein topology databases. *Bioinformatics*, 15(4):317–326, 1999.

H.M. Grindley, P.J. Artymuik, D.W. Rice, and P. Willett. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *Journal of Molecular Biology*, 229(707–721), 1993.

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

I. Koch and T. Lengauer. 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, 1997.

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

V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii nauk SSSR (in Russian)*, 163(4):845–848, 1965. Also in Cybernetics and Control Theory, vol 10, no. 8, pp 707–710, 1996.

E.M. Mitchell, P.J. Artymuik, D.W. Rice, and P. Willett. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *Journal of Molecular Biology*, 212:151–166, 1989.

A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chotia. **scop**: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247: 536–540, 1995.

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

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

W.R. Taylor and C.A. Orengo. Protein structure alignment. *Journal of Molecular Biology*, 208:1–22, 1989.

Y. Tsukamoto, K. Takiguchi, K. Satou, T. Furuichi, E. Takagi, and S. Kuhara. 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, D. C. Hutton, and J. M. Thornton. An atlas of protein toplogy cartoons available on the World Wide Web. *Trends in Biochemical Sciences*, 23, 1998.

D. R. Westhead, T. W. F. Slidel, T. P. J. Flores, and J. M. Thornton. Protein structural topology: automated analysis and diagrammatic representation. *Protein Science*, 8(4):897–904, 1999.