Structural bioinformatics

A novel method for comparing topological models of protein structures enhanced with ligand information

Malika Veeramalai\(^1,2,\ast\) and David Gilbert\(^1,3\)

\(^1\)Bioinformatics Research Centre, Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK
\(^2\)Joint Center for Molecular Modeling, Burnham Institute for Medical Research, La Jolla, CA 92037, USA
\(^3\)School of Information Systems, Computing and Mathematics Brunel University, Uxbridge, Middlesex UB8 3PH, UK

Received on 31 March, 2008; revised on September 29, 2008; accepted on October 4, 2008

Advance Access publication October 7, 2008
Associate Editor: Anna Tramontano

ABSTRACT

We introduce TOPS+ strings, a highly abstract string-based model of protein topology that permits efficient computation of structure comparison, and can optionally represent ligand information. In this model, we consider loops as secondary structure elements (SSEs) as well as helices and strands; in addition we represent ligands as first class objects. Interactions between SSEs and between SSEs and ligands are described by incoming/outgoing arcs and ligand arcs, respectively; and SSEs are annotated with arc interaction direction and type. We are able to abstract away from the ligands themselves, to give a model characterized by a regular grammar rather than the context sensitive grammar of the original TOPS model. Our TOPS+ strings model is sufficiently descriptive to obtain biologically meaningful results and has the advantage of permitting fast string-based structure matching and comparison as well as avoiding issues of Non-deterministic Polynomial time (NP)-completeness associated with graph problems. Our structure comparison method is computationally more efficient in identifying distantly related proteins than BLAST, CLUSTALW, SSAP and TOPS because of the compact and abstract string-based representation of protein structure which records both topological and biochemical information including the functionally important loop regions of the protein structures. The accuracy of our comparison method is comparable with that of TOPS. Also, we have demonstrated that our TOPS+ strings method out-performs the TOPS method for the ligand-dependent protein structures and provides biologically meaningful results.

Availability: The TOPS+ strings comparison server is available from http://balabio.dcs.gla.ac.uk/mallika/WebTOPS/topsplusr.html.

Contact: mallikav@burnham.org

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Protein structure comparison is vital for analyzing evolutionary relationship and functional characterization. The number of known structures in the Protein Data Bank (PDB) (Berman et al., 2002) is increasing rapidly every year and as of spring 2008, it contains more than 49,500 3-dimensional (3D) structures. This highlights the importance of fast and reliable protein structure comparison methods. Structural comparison methods at the abstract level of topology are based on comparing secondary structure elements, and their relationships, between proteins. For example, GRATH (Harrison et al., 2003) is a graph-based algorithm that compares the axial vectors of \( \alpha \)-helices and \( \beta \)-strands of two proteins, together with the distances, angles and chirality between these vectors. Earlier work by Koch et al. (1996) uses a graph method based on the Bron/Kerbosch algorithm to find maximal common secondary structure elements for a given pair or set of proteins. Similarly, other abstract-level structure comparison approaches include VAST, which is a vector-based protein structure comparison method (Madej et al., 1995) and TOPS, a graph-based method applied to the topological representation of the protein structures (Viksna and Gilbert, 2001). Although these methods permit fast protein structure comparison, in most cases the abstract nature of their models precludes the exploitation of the biologically significant information in the underlying atomic-level model. Moreover, the problem of functional annotation at the fold level is hard, due to the fact that the number of protein fold is limited, while their range of functions is very diverse. For example, the TIM barrel proteins are associated with 15 distinct enzyme families (Gromiha et al., 2004; Nagano et al., 2002).

Thus, we have been motivated in our research to develop a novel topological model for protein structures, which is enhanced with structural and biochemical features, including ligand interaction information and the amino acid sequence length of the secondary structures, in order to improve the biological significance of comparisons. In this article, we introduce a novel protein structure comparison method, which is based on the enhanced TOPS+ strings model. This method computes an edit distance between two protein structures using a dynamic programming approach. Our method gives the distance between two proteins as a comparison score in the form of normalized compression. Our method can also compute the longest common substructures (LCS) between the TOPS+ strings models of two proteins.

2 TOPS MODEL AND COMPARISON METHOD—BACKGROUND

Protein structure can be described at a highly simplified ‘topological’ level using TOPS cartoons (Sternberg and Thornton, 1977). An automatic cartoon layout program was devised by Westhead et al. (1998, 1999) to produce...
A novel method for comparing topological models of protein structures enhanced with ligand information

3 A NOVEL ENHANCED TOPS MODEL

3.1 The TOPS+ and TOPS+ strings model

The TOPS model (Gilbert et al., 1999, 2000; Torrance et al., 2005; Viksna and Gilbert, 2001) represents protein structures at the fold level by a graph where the nodes stand for SSEs—(up or down) α-helices and β-strands—and (non-directed) edges represent right-or left-handed chirality and parallel or anti-parallel hbond relationships. In addition, there is a total ordering over the nodes, corresponding to the backbone of the protein. Our TOPS+ model enhances the original TOPS graph model with structural and biochemical features, such as ligand interaction information and amino acid sequence length of the secondary structures. We have added extra nodes for loops (represented as a first class object; i.e. SSE) and ligands as well as maintaining the existing nodes for α-helices and β-strands. We have compiled the protein–ligand interaction information for all PDB entries by running the InterCal program, which analyses interactions between proteins and ligands (Wallace, A. and Michalopoulos, I., personal communications). Interactions with ligands can be represented as residue–ligand with many-to-many relationship based on atomic-level interactions, which include interaction properties, such as hbond, covalent, van der Waals, electrostatic forces, etc. (see in Fig. 1a, for example, A342 and X13). In the current version, we have considered protein–ligand interactions at the SSE–ligand level with a many-to-one relationship. The residue–ligand atomic-level interactions are abstracted in two ways: (i) by considering residue–ligand interaction as a single entity at SSE level rather than the actual residues and (ii) likewise all the ligand atoms are considered as a single object (Fig. 1b). The earlier TOPS model did not consider loop regions; however, most protein–ligand interactions take place at the surface of the proteins, often in the loop regions. Moreover, the loops are not regular, and there are some cases where proteins have contiguous strands and helices without intervening loops, for example, the hydrolase protein Ippe. According to SCOP version 1.61, we found that 10.337 out of 28298 protein domains have some contiguous strand–helix connections without intervening loops within the protein structure.

Further, we have designed a string model based on our TOPS+ graph model where the long-range and short-range spatial interactions between the SSEs are converted into incoming and outgoing arcs for each SSE, which encode the direction of an interaction relative to the backbone order, and its type. This model can describe protein structures alone without ligands, or optionally with ligand interactions. In the latter case, all relevant SSE nodes are enhanced with SSE–ligand interaction information that includes loop–ligand interaction information. We abstract away from the ligands themselves, to give a linear model called TOPS+ strings, which preserves the essential biochemical information, whilst permitting more efficient and non-heuristic algorithms for comparison.

In detail, each SSE node is enriched with SSEType, SSE segment details are indicated by PS–SL where PS is the PDB start number and SL is the SSE segment length, total incoming arcs (InArc) and total outgoing arcs (OutArc), total number of ArcTypes and total number of ligand arcs (LigArc). The SSEType is given by a member of the set {E, e, H, h, U, u} where, ‘E’ and ‘e’ represent ‘up’ and ‘down’ oriented β-strands; ‘H’ and ‘h’ indicate ‘up’ and ‘down’ oriented α-helices; ‘U’ and ‘u’ represent ligand-bound and ligand-free loops. The InArcType is represented as a member of {R, L, P, A}, where ‘R’ and ‘L’ represent right and left chiralities; ‘P’ and ‘A’ represents parallel and anti-parallel hbonds, respectively. The OutArcType is represented in a similar manner by {R’, L’, P’, A’}. Ligand arcs are indicated by LT=AA where LT is the ligand type and AA is the PDB number. For example, Figure 2c shows the 3D cartoon of the protein domain Ifnb01 with FAD and SO4 ligands represented with sticks; Figure 2a and d provides the visual representation for Ifnb01 from the enhanced TOPS model or TOPS+ model and the reduced ‘TOPS+ strings model’, respectively. Here, triangles represent β-strands; a half circle represents the α-helix; an ellipse indicates loop regions and arcs indicate hbonds between two β-strands, called the anti-parallel β-sheet. Figure 2b represents the TOPS+ strings model of the protein domain Ifnb01 as a string notation. The length of a TOPS+ strings is given by the number of SSEs; thus the length of Ifnb01 is 19. We have generated the TOPS+ model and the TOPS+ strings representation of 28976 and 28298 protein domains including 16163 and 14887 ligand-bound protein domains corresponding to the CATH 2.4 and the SCOP 1.61, respectively.

We have chosen this strings representation because our linear representation of protein topology can be described by a regular...
The TOPS+ strings model permits pairwise structure comparison by TOPS graphs of Gilbert proteins sharing same fold have a similar function (Bray and evolution (Doolittle, 1995); moreover at the superfamily level, because these are the basic units of protein structure, function considers global alignment (the Needleman and Wunsch Algorithm). Computing the edit distance between two proteins using dynamic similarity, for example, multiple domains or repeats (Barton, 1996). SSE–ligand interactions at a local level across different folds. Also to find the local structural similarity or patterns, such as similar local alignment (Smith and Waterman Algorithm) can be applicable is appropriate for proteins that are known to share similarity at strings model records topological features, so that a global match is meaningful protein structure comparison which has the advantage of being very fast compared with the earlier TOPS comparison method. Our enhanced descriptions effectively double the number of nodes by introducing loops as SSEs; moreover the introduction of ligand nodes effectively destroys the linear ordering of the nodes in the graph. Our TOPS+ strings representation is given in Equation 1.

\[ M[i,j] = \min \{ M[i-1,j)+1; M[i-1,j]+1; M[i,j-1]+\text{SSEArcMatch}(t_i, s_j) \} \]  

where, SSEArcMatch is 0 if \( t_i = s_j \) (match); SSEArcMatch is 1 if \( t_i \neq s_j \) (mismatch).

At the end, \( M[n,m] \) will contain the edit distance \( ED \) value of the two TOPS+ strings models of the protein structures. Once all the values of the edit distance matrix have been computed, the LCS of the two models can be retrieved in the standard manner by tracing the shortest paths in the matrix from the last position to the first.

### 3.3 Modified version of the edit distance algorithm

We have added some additional functions in the basic edit distance algorithm in order to compare the TOPS+ strings models of the two protein structures. In our method, the construction of the edit distance matrix (the dynamic programming table) is the major component; this is achieved through the recursive definition of optimal matching between two TOPS+ strings SSE elements of the given target \( t_i \in T \) and source \( s_j \in S \) protein domains using the ‘SSEArcMatch’ function (see Section 0). When comparing two proteins of length \( n \) and \( m \), respectively, each position \( M[i,j] \) in the matrix is computed row–by-row based on the recurrence relation given in Equation 1.

In PDB structures, proteins having the same fold can interact with ligands of similar or different types. Similarly, a set of proteins with different folds can interact with similar ligands. Therefore, it can be useful to design the comparison process based either on local or global alignment according to the user’s requirements. In this TOPS+ strings comparison, we have considered only the abstract-level ligand arc information without any associated ligand name. However, in our case study, we have used a new version of the TOPS+ strings comparison method which integrates the ligand pattern discovery (LPD) within the comparison (see next section).

Our comparison method comprises five major steps:

1. Recursive definition of optimal match (alignment) score (based on SSEArcMatch Score).
2. Construction of a 2D edit distance matrix (Dynamic programming table).
3. Building a trace-back on the edit distance matrix (Dynamic programming table).
4. Obtaining the LCS.
5. Computation of the comparison score.

### 3.4 The SSEArcMatch function

In our method, the computation of the edit distance matrix \( M \) is the main process in which the SSEArcMatch function plays a key role in assigning scores for each match or mismatch between the TOPS+ strings of the two protein models. This function considers the SSEs elements of the query and targets TOPS+ strings models one by one. At the initial level, this function checks whether the SSEType in the source and the target is matched or not. If SSEType is matched then in the second stage an attempt is made to match all additional arc features, such as the total number of incoming, outgoing and ligand arcs as well as their actual ArcType properties.
If all the conditions are satisfied then a zero score is assigned, otherwise a mismatch score of ‘1’ is assigned when any one feature is not matched between the query and the target SSEs. Note that in this version we have not included the condition for matching the length of SSE segments, but have treated SSEs at a more abstract level without considering their segment length.

3.5 Incorporation of LPD within the TOPS+ strings comparison

3.5.1 Ligand compound classification We found that there are no proper ligand classification resources existing for all the PDB ligand molecules. According to the HET_Dictionary library from the PDB database (Berman et al., 2000) there are currently more than 5000 different ligand molecules recorded in the PDB which are bound to protein structures. We obtained the hierarchical ligand classification information from the macromolecular Crystallographic Information File (mmCIF) data dictionaries (Bourne et al., 1997). The dictionary file mmCIF contains the information on ligand classification at the hierarchical-level based on compound information given as follows: ‘Type’, ‘Class1’ and ‘Class2’. Type provides information on the higher level ligand groups, such as atom, DNA-Linking or Non-polymer, etc., while at the next level Class1 classifies those ligand types based on their compound properties. For example, if the ligand is of ‘non-polymer type’ then this can be further divided into any one of the following groups, such as [nucleotide, nucleoside, ion, simple group] and so on. Similarly, we can further divide Class1 into Class2 ligands based on the actual chemical property of the ligand, for example, if a ligand is of ‘non-polymer’ type and belongs to ‘nucleotide’ in Class1 then Class2 gives the information about type of nucleotide, such as ‘ADENOSINE-NUCLEOTIDE’ or ‘NICOTINAMIDE-ADENINE-DINUCLEOTIDE’. We have included the molecular weight and their chemical formula as an additional discriminative feature for each ligand molecule.

3.5.2 Ligand pattern discovery In general, several different approaches are available for LPD (von Grothus et al., 2003; Xue et al., 1999) and ligand-based analysis of protein structures (Chalk et al., 2004; Nobeli et al., 2005). In our LPD method, we utilize hierarchical ligand classifications for ligand molecules based on their compound properties. Our LPD method takes a set of protein domains \( D = \{d_1, d_2, \ldots, d_n\} \) where \( d_j = (k_j, k_2, \ldots, k_{m_j}) \), a set of HET ligands \( K = \{k_1, k_2, \ldots, k_m\} \), which contains ligand_id information and the unique list of ligand classification library \( C = (c_1, c_2, \ldots, c_r) \) where \( c_i = k_i \). Formula: Molwg:Class1:Class2. We have obtained this unique HET ligand set from the given protein domains in order to avoid an exhaustive search of the ligand space, because the original set contains more than 5000 unique ligands according to their ligand_id. This ligand classification library contains ligand classification information in a string format given as follows ‘ligID:Formula: Molwg: Class1: Class2’. The LPD method outputs ‘L’ a set of ligands common to \( D \), where \( L \subseteq K \). It compares each element of \( K \) against all proteins in \( D \) using the ‘MatchAllProLog’ function. We can perform the ligand match in two ways: (i) the ‘LigandNameMatch’ compares the ligand_id and reports the result as a Boolean and (ii) the ‘LigandCompoundClassMatch’ performs ligand match based on ligand classification information by extracting, parsing corresponding ligand classification information from \( C \), followed by comparison process and reports the Boolean result. The goodness of the ligand pattern is evaluated by a compression measure, which is adapted from Brazma et al. (1998).

3.6 Computation of comparison score

We have calculated the comparison scores based on ED values using the formula given in Equation (2). This score is only based on the number of SSEs without considering any arc information, where \( D_1 \) and \( D_2 \) correspond to the sets of SSEs from the target \( T \) and the source \( S \) protein domains, respectively.

\[
\text{Average} = \frac{2^* \text{ED}}{(|D_1| + |D_2|)}
\]  

(2)

4 EXPERIMENT

4.1 Datasets

We have considered two different datasets, namely the large alldp_PDB40 (containing both ligand-bound and ligand-free protein domains) and the small randomly selected ligandomonlydp dataset. We obtained the PDB40 dataset corresponding to the SCOP version 1.61 from the ASTRAL database (Brenner et al., 2000; Chandonia et al., 2004). The original PDB40 dataset contains 4774 protein domains from all classes; we considered the following SCOP structural classes namely, all alpha (\( \alpha \)), all beta (\( \beta \)), alphabeta (\( \alpha/\beta \)), alpha-beta (\( \alpha + \beta \)), which comprise 4220 protein domains. This dataset was subsequently reduced into 2620 domains with respect to the data that was available for the entries in the TOPS database (Michalopoulos et al., 2004) and in our enhanced TOPS+ database; we thus obtained 3430890 domain pairs \( n(n - 1)/2 \). Furthermore, we did not compare protein domain pairs for which only the first two levels of their SCOP numbers match since the SCOP classification does not differentiate between homologous and non-homologous pairs at this fold level. Thus, the final alldp_PDB40 dataset was reduced to 940383 protein domain pairs (Table 1). From this alldp_PDB40 dataset, we randomly selected the ligandomonlydp dataset; it contains 1362 domains pairs, in which all proteins are bound to one or many ligands. We have used this small dataset to test the biological significance of our novel TOPS+ strings model and comparison (Table 2).

4.2 Evaluation analysis procedures

We compared the TOPS+ strings comparison method against the TOPS comparison method which uses a Maximal Common Subgraph algorithm (Torrance et al., 2005; Viknsa and Gilbert, 2001) based on the alldp_PDB40 dataset. Based on the ligandomonlydp dataset.

<table>
<thead>
<tr>
<th>SCOP classes</th>
<th>Homolog Non-homolog Total Homolog (%) Non-homolog (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha (( \alpha ))</td>
<td>1550 85750 87300 2 98</td>
</tr>
<tr>
<td>All beta (( \beta ))</td>
<td>4528 99180 94708 5 95</td>
</tr>
<tr>
<td>Alphabeta (( \alpha/\beta ))</td>
<td>9720 379490 389210 2 98</td>
</tr>
<tr>
<td>Alpha-beta (( \alpha + \beta ))</td>
<td>2532 366633 369165 1 99</td>
</tr>
<tr>
<td>Total</td>
<td>18330 922053 940383 2 98</td>
</tr>
</tbody>
</table>

Table 1. SCOP superfamily-level homolog versus non-homolog statistics for the alldp_PDB40 dataset
dataset, we compared the biological significance of our TOPS+ strings method against the existing comparison methods including TOPS (Torrance et al., 2005) and detailed 3D protein structure comparisons, such as FATCAT—a flexible structure alignment by chaining aligned fragment pairs allowing twists (Ye and Godzik, 2003), (we have used the FATCAT with rigid option), SSAP—a sequential structure alignment program for protein structure comparison (Orengo and Taylor, 1996), DaliLite—a pairwise protein structure comparison method based on distance matrices (Holm and Park, 2000) and protein sequence comparison method BLAST (Altschul et al., 1997). We constructed the receiver operating characteristics (ROC) curve and computed the area under the ROC curve (AUC) value as a quantitative measure to evaluate the performance of all the methods based on the gold standard SCOP superfamily classification numbers as an indication of structural homology (Bradley, 1997). Tables 1 and 2 give homologous and non-homologous statistics for the alldp_PDB40 and ligandonlydp datasets with respect to the SCOP structural classes, respectively.

5 RESULTS

The performance on the alldp_PDB40 dataset computed using these AUC values shows that our TOPS+ strings method gives similar results compared with TOPS in the case of SCOP class alpha+beta (α + β) and all alpha (α) proteins and slightly worse results for all beta (β) and alpha/beta (αβ) (Fig. 3 and Table 3). The main reasons for false positives in the case of all beta and alpha/beta classes are due to the following: (i) at the superfamily level, we find variable number of SSEs between protein families due to the fact that insertion and deletion of amino acid residues during evolutionary process (Mizuguchi and Blundell, 2000). These variable numbers of SSEs and their long-range/short-range spatial relationships are recorded in the form of incoming/outgoing arc and ligand arc information in our TOPS+ strings model. (ii) In our general TOPS+ comparison, we have used the abstract-level ligand arc information, whether a SSE have a LigArc or not without using their explicit ligand details. This leads to false positives as the same SSE with similar topological property can interact with different kinds of ligands, across different proteins which perform entirely different functions. For example, the SCOP TIM barrel superfAMILY proteins belong to the alpha/beta class, and are associated with more than 30 different superfamilies due to distinct ligand interactions. This explains the poor result for the SCOP classes alpha/beta protein domains.

Our method is faster in identifying distantly related proteins than TOPS because our strings-based model has a lower degree of complexity than that of TOPS which is graph-based and highly sensitive to the number of arcs in a description. Moreover, our method exploits the biochemical information, which is recorded in our model including functionally important loop regions that are ignored by other protein comparison methods.

For the ligandonlydp dataset, our TOPS+ strings method gives better results compared with the BLAST method (Fig. 4 and Table 4). This method works well for close homologous proteins rather than the distantly related proteins for all the SCOP classes. This result confirms that the protein structures are much more conserved than the sequences, because we have used protein domains from the superfamily level and in most cases the sequence identity between any two member have very low sequence identity (normally <30%). Although BLAST has the same time complexity O(nm), our method is faster than BLAST in comparing protein domains. This is because BLAST considers amino acid sequences (average, size is ~300 amino acid); while the TOPS+ strings method consider the SSEs (average, size is ~50 nodes) of the protein.

In the case of alpha/beta protein domains on the ligandonlydp dataset, the TOPS+ strings method can identify more distantly related proteins compared with the SSAP method even at the 70–100% coverage regions (see SCOP alpha/beta proteins in Fig. 4). This illustrates the fact that geometrical structural features are not always conserved in distantly related proteins with a similar fold. Thus the more detailed 3D structure comparison method (SSAP) was not able match the geometrical structural features in more

### Table 2. SCOP superfamily-level homolog versus non-homolog statistics for the ligandonlydp dataset

<table>
<thead>
<tr>
<th>SCOP classes</th>
<th>Homolog</th>
<th>Non-homolog</th>
<th>Total</th>
<th>Homolog (%)</th>
<th>Non-homolog (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha (α)</td>
<td>96</td>
<td>16</td>
<td>112</td>
<td>86</td>
<td>14</td>
</tr>
<tr>
<td>All beta (β)</td>
<td>28</td>
<td>22</td>
<td>50</td>
<td>56</td>
<td>44</td>
</tr>
<tr>
<td>Alpha/beta (αβ)</td>
<td>228</td>
<td>247</td>
<td>475</td>
<td>48</td>
<td>52</td>
</tr>
<tr>
<td>Alpha/beta (α + β)</td>
<td>65</td>
<td>40</td>
<td>105</td>
<td>62</td>
<td>38</td>
</tr>
<tr>
<td>Total</td>
<td>443</td>
<td>919</td>
<td>1362</td>
<td>33</td>
<td>67</td>
</tr>
</tbody>
</table>

### Table 3. AUC values for the PDB40 dataset from TOPS and TOPS+ strings methods

<table>
<thead>
<tr>
<th>SCOP Classes</th>
<th>TOPS</th>
<th>TOPS+</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha (α)</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td>All beta (β)</td>
<td>0.90</td>
<td>0.84</td>
</tr>
<tr>
<td>Alpha/beta (αβ)</td>
<td>0.80</td>
<td>0.73</td>
</tr>
<tr>
<td>Alpha/beta (α + β)</td>
<td>0.87</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Fig. 3. ROC curves of TOPS and TOPS+ strings (TOPS+) method for PDB40 dataset.
distantly related proteins. Even though the FATCAT method has good overall performance, this method struggles at 90–100% of the coverage region, since it builds alignment by chaining aligned fragment pairs (AFPs) (Ye and Godzik, 2003). The same is true to some extent in the case of TOPS, even though this method does not depend on geometry, but it relies on topological relationships. So the TOPS method matches the core region of the protein domains very quickly, while the variable but functionally important surface regions, such as loops are ignored by this method (Fig. 4 and Table 4). Thus, our TOPS+ strings model demonstrates the key advantages of the abstract nature, while at the same time it efficiently records the topological features and importance of biochemical features including functionally important loop regions of the proteins, which are always ignored by other methods. The DaliLite and FATCAT methods give better performance compared with all the other methods for all SCOP classes of protein domains except the all beta class. The SSAP method provides the best results for all beta class proteins compared with all the other method. The different results yielded by the methods TOPS+, SSAP, FATCAT and BLAST are caused by the very different underlying assumptions and algorithms used. However, any protein comparison method is extremely sensitive to the set of chosen thresholds, weights, etc., simply because there is no commonly accepted measure for protein similarity measure.

We performed all the analyses using a RedHat 7.2 linux environment with an Intel Pentium IV 1.6 GHz processor. The TOPS+ strings and the TOPS comparison methods took 46 s and 160 s, respectively to complete 1362 comparisons. The FATCAT method with rigid option took 1965 s (i.e. 33 min), the DaliLite method completed the comparison task in 6128 s (i.e. 1 h and 42 min) and the SSAP method took 13892 s (i.e. 33 min), the DaliLite method completed the comparison task in 6128 s (i.e. 1 h and 42 min) and the SSAP method took 13892 s (i.e. 33 min). The overall results show that our TOPS+ strings comparison method has better performance compared with the protein sequence-based approach BLAST, and that the performance of our method is comparable with the graph-based TOPS comparisons. FATCAT and SSAP perform well for all protein classes. The TOPS+ strings comparison method is faster in identifying distantly related proteins than FATCAT, SSAP and TOPS because of the abstract TOPS+ strings model that records both topological and biochemical information including functionally important loop regions that are ignored by other methods.

6 FUNCTIONAL CLASSIFICATION OF THE NAD(P)-BINDING PROTEINS

We evaluated the biological significance of the TOPS+LPD comparison method incorporating LPD based on functional classification of 14 NAD(P)-binding proteins. We compared the performance of the TOPS+LPD method in the ClustalW (sequence-based), SSAP (3D atomic coordinates-based) and TOPS (Topology-based) comparison methods. We obtained the comparisons score from the all-against-all pairwise comparison for 14 protein domains and plotted the clusters using the ‘OC’ program (Barton, 1997). The clustering analysis results of the 14 NAD(P)-binding protein domains obtained by ClustalW (Higgins, 1994), SSAP (Orengo and Taylor, 1996), TOPS (Gilbert et al., 2001) and TOPS+LPD methods are given in the Supplementary Figure 1a–d. Supplementary Figure 1d shows the highlighted clusters of the NAD(P)-binding protein domains obtained from the TOPS+LPD method. Each subcluster is indicated by different colored boxes, and the protein domains are annotated with their function and organism name. The Supplementary Figure 2a provides the statistics of protein–ligand interaction ordered based on subclusters as shown in the Supplementary Figure 1d. For each domain, it gives total number of ligand interactions based on unique ligand_Id, ligand residue_num and SSEType–ligand interactions with respect to helix, strand and loop (Supplementary Fig. 2a). The conserved SSEs of Dihydropteridine reductase from rat in green (1dhr00) and human in red (1dhr00) identified by the TOPS+LPD method are shown in the Supplementary Figure 2h as a 3D superposed cartoon model and non-matched region indicated by a backbone trace.

Even the layout topology of the clusters from the different methods highlights the variation of the results between the different protein comparison approaches (Supplementary Fig. 1). The sequence-based approach clustalW is not able to cluster the protein domains correctly. In the case of the SSAP comparison approach, it distinguishes the proteins into three subclusters, while there are still four protein domains, which are not grouped properly. This indicates that the structural features are not always conserved in proteins with a similar fold (Russell and Barton, 1994). The detailed 3D atomic coordinate-based structure comparison method, for example, SSAP relies on the geometrical properties of the
structural features of the amino acids, which are not conserved in distantly related proteins even though the fold is preserved. In contrast, the TOPS topological comparison method clusters the six protein domains into a single group, because this method performs protein structure comparison only at the abstract topological level and does not depend explicitly on any detailed geometrical features. In other words, although we have used the basic geometrical properties to build the TOPS model, we represented structural features at an abstract level. For example, the β-sheet conformations based on the number of hydrogen bonds between two β-strands are oversimplified as a single interaction relationship. Therefore, this method is able to identify distantly related proteins more efficiently than the SSAP method. Our method gives six different clusters and the branch length between any two proteins, which indicates the dissimilarity (distance) between the protein domains. Our method divides the protein domains into several subclusters. This highlights the fact that our method is able to group the protein domains according to topological information and additionally according to the constraints from the protein–ligand interaction and their properties, which are exploited in our LPD method within our TOPS+ strings comparison.

This also confirms the hypothesis that detailed geometrical properties of the structural features are not always conserved in distantly related proteins with a similar fold (Russell and Barton, 1994). Therefore, a detailed 3D structure comparison method, such as SSAP may not be able to match those geometrical structural features in more distantly related protein structures. The same is true to some extent in the case of the TOPS method, even though this does not depend on geometry, but relies on topological relationships. Thus, the TOPS method can match the scaffold or core structure of the protein domains very quickly, while the other parts of the proteins are not recognized. Hence, our TOPS+ strings model demonstrates the key advantages of abstraction, while at the same time efficiently recording both topological and biochemical features including functionally important loop regions of the proteins, which are ignored by most of the structure comparison methods.

7 CONCLUSIONS

We have presented a new protein structure comparison method based on a novel topological model of protein structure enhanced with ligand information. Our TOPS+ strings model has many interesting features, such as the description of loop regions, and the abstraction of long-range interactions between SSEs into incoming and outgoing arc information. SSE–ligand interactions are described in a similar manner. This level of abstraction permits the characterization of our models using a regular grammar rather than the context-sensitive grammar of the existing graph-based TOPS model. This enables the use of fast exact string-matching algorithms, avoiding the disadvantages associated with graph-based techniques, which employ heuristics as an attempt to minimize the effect of the NP-complete nature of the subgraph isomorphism problem.

In general, there is an acceptable tradeoff between speed (computing time) and accuracy. The Dalilite, FATCAT and SSAP methods provide good accuracy, however these methods are very slow and computationally more expensive compared with all the other comparison methods. Our TOPS+ strings method has a lower degree of complexity than all the other structure-based methods, with time complexity O(nm), where n and m represents the number of SSEs in the source and the target proteins, respectively.

The performance of our method on the large PDB40 dataset gives comparable results with the existing TOPS comparison method. However, we have demonstrated the better biological significance and performance of our method through the example of the ‘ligand-bound’ protein domains and our case study on functional classification of the NAD(P)+-binding proteins. This research opens new doors to exciting improvements to our TOPS+ strings models and associated comparison method with additional features. Currently, we are working on improving the performance of the system by introducing an optimal statistical scoring value in the TOPS+ strings element-matching process within our comparison method. A recent development is the TOPS++FATCAT system that exploits the TOPS+ comparison method to speedup the FATCAT protein structure alignment program in order to provide fast flexible structural alignment, while preserving the accuracy of the original FATCAT method (Veeramalai et al., 2008).

ACKNOWLEDGEMENTS

We would like to thank the TOPS Project at the Universities of Glasgow and Leeds for TOPS data resources, Juris Viksna advice on our method.

Funding: PhD studentship from the University of Glasgow (to MV).

Conflict of Interest: none declared.

REFERENCES


