The TOPStrings Protein Structure Comparison Method - A Novel Approach

Mallika Veeramalai\textsuperscript{1} and David Gilbert\textsuperscript{1}

\textsuperscript{1}Bioinformatics Research Centre, Department of Computing Science, University of Glasgow, Glasgow G12 8QQ, Scotland, UK.

Here, we introduce TOPStrings (TOPS+), a highly abstract string-based model of protein topology which 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 and outgoing arcs, 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 (Gilbert et al., 2000; Gilbert et al., 2001; Viksna and Gilbert, 2001). Our TOPStrings 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 NP-completeness associated with graph problems.

We have developed a string version of our general graph model, called TOPStrings, where the relationships between SSEs and SSEs and ligands are reduced into incoming arcs (into an SSE from another SSE earlier in the chain, or from a ligand) and outgoing arcs (from an SSE to an SSE further on in the chain) SSEs, retaining their arc type properties. Ligands are represented indirectly via their connections to SSEs rather than as first class objects. For example, Fig 1 (a1) and (a2) illustrate the visual representation of our enhanced TOPS model and reduced TOPStrings model for the protein domain 1fnb01. Here the triangles represent the beta strands; red curves represent the alpha helix; circles indicate loop regions and green arcs indicate hydrogen bonds between two beta strands, called the anti parallel beta sheet. The length of a TOPStrings is given by the number of SSEs; thus the length of 1fnb01 is 18.

We have chosen this representation because this linear representation of protein topology can be described by a regular grammar, permitting the use of efficient and tractable string-based matching algorithms for matching and comparison rather than graph-based approaches which are NP-complete. We note that the original TOPS graphs of Gilbert et al. had a strict linear ordering on the nodes, which was exploited in the subgraph isomorphism matching algorithm developed by Viksna et al (Viksna and Gilbert, 2001). Moreover, their approach was tractable for these TOPS graphs with an average of 50 nodes or less. 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 TOPStrings representation is sufficient to perform biologically meaningful protein structure comparison and has the advantage of fast comparisons.

Fig 1 (a1) general enhanced TOPS model, (a2) TOPStrings model.
The TOPStrings model permits pairwise structure comparison by computing the edit distance between two proteins using dynamic programming. The current version of our comparison method considers global alignment (Needleman and Wunsch Algorithm). Our TOPStrings model records topological features, so that a global match is appropriate for proteins that are known to share similarity at the fold or superfamily level over their whole domain. However, local alignment (Smith and Waterman Algorithm) can be applicable to find the local structural similarity or patterns such as similar SSE-ligand interactions at a local level across difference folds. 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 either based on local or global alignment according to the user’s requirement. In our current version which we describe here we have considered ligand arc information without recording the actual ligand name. In work not reported here we have improved our comparison method by parameter-tuning, and have also developed a method for ligand-pattern discovery (Veeramalai, 2005).

Our comparison method comprises five major steps:

1. Recursive definition of optimal match (alignment) score (based on SSEArc+ Match 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 (Longest Common Substructure)
5. Computation of the comparison score.

We have considered a large PDB40 dataset, which contains both ligand-bound and ligand-free protein domains. We obtained this dataset corresponding to SCOP version 1.61 from the ASTRAL database (Brenner, et al., 2000; Chandonia, et al., 2002; Chandonia, et al., 2004). The original PDB40 dataset contains 4774 protein domains from all classes; we considered the SCOP classes 1 to 4, which comprise 4220 protein domains. This dataset was subsequently reduced into 2620 domains w.r.t 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). Because of the very large size of this dataset and the fact that in our evaluation process we wished to compare our method with the performance of several other comparison methods we performed a class-filter. Here the class-filter indicates the grouping of protein domain pairs such that both belong to same SCOP structural class (1 to 4), i.e. allalpha(α), allbeta (β), alpha/beta (α/β), alpha+beta (α+β). 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 dataset was reduced to 940383 protein domain pairs (see Table1).

Table 1 Homologous & non-homologous statistics

<table>
<thead>
<tr>
<th>SCOP Classes</th>
<th>Homologous Domain pairs</th>
<th>Non_Homologous Domain pairs</th>
<th>Total</th>
<th>Hom(%)</th>
<th>nonHom(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All alpha (α)</td>
<td>1550</td>
<td>85750</td>
<td>87300</td>
<td>2</td>
<td>98</td>
</tr>
<tr>
<td>All beta (β)</td>
<td>4528</td>
<td>90180</td>
<td>94708</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>Alpha/beta (α/β)</td>
<td>9720</td>
<td>379490</td>
<td>3892102</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>Alpha+beta (α+β)</td>
<td>2532</td>
<td>366633</td>
<td>3691651</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>18330</td>
<td>922053</td>
<td>940383</td>
<td>98</td>
<td></td>
</tr>
</tbody>
</table>
For the PDB40 dataset the performance computed using Receiver Operating Characteristics (ROC) curve analysis shows that our TOPStrings method gives similar results compared to TOPS in the case of SCOP class 4 and class 1 proteins and slightly worse results for classes 2 and 3 (see Fig 2). Overall results shows that 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.

The TOPStrings (TOPS+) comparison server is available from the website http://balabio.dcs.gla.ac.uk/mallika/WebTOPS/topsplus.html and for further information contact mallika@dcs.gla.ac.uk or mallikav@burnham.org.


