Pattern discovery methods for protein topology diagrams

David Gilbert (1) and Juris Viksna (2)

1 Department of Computing, City University, Northampton Square, London EC1V 0HB, UK. email: drg@cs.city.ac.uk
2 Institute of Mathematics and Computer Science, University of Latvia, Riga LV-1459, Latvia. email: jviksna@cclu.lv

We are carrying out research into developing several approaches to pattern discovery in protein topology diagrams, and comparing them. The underlying motivation is to efficiently automatically generate patterns classifying sets of proteins and to apply this to characterising databases of protein structure. We are using TOPS protein topology diagrams, which we have formalised as a restricted kind of graph in [3]. In general pattern discovery in TOPS diagrams can be reduced to subgraph isomorphism/maximal common subgraph problems in oriented labelled graphs. We are using two different approaches (illustrated in an experiment software suite at http://www.tops.ebi.ac.uk/learn.html):

Learning (pattern extension) is based on searching of common patterns constructed by pattern extension and on repeated pattern matching; in this sense it is similar to "pattern driven" approaches for strings [1]. In general this method is very simple: by starting with empty pattern we try to extend it in all possible ways and discard extensions that do not match a given set of examples, until we find a largest pattern that cannot be extended further. Such an approach seems in general to be far too inefficient for graphs of size that is typical to TOPS diagrams; however a somewhat limited version of this method (that does not always guarantee the finding of best possible pattern) works surprisingly fast for most of the real examples from TOPS database. The good property of this approach is that algorithm complexity grows proportionally to the size of the set of examples in learning set. We are now working on improved versions on pattern matching and pattern learning algorithms and expect that we will be able to obtain practically efficient pattern discovery method that will guarantee the finding of maximal pattern(s).

We have been experimenting with learning patterns for β and $\alpha\beta$ domains at the H level in the CATH hierarchy. We are evaluating our results by determining how many false positives are returned by a search over the database using the discovered pattern; initial results are encouraging in many cases.

Clique detection works by finding of common patterns via maximal clique detection in edge product graphs, using for example the Bron-Kerbosch algorithm [2]. We have implemented a modified version for this approach that takes advantage of the ordering of graphs and reduces the number of edges in the edge product graph by the factor of 6. Unfortunately, this modification also does not allow us to use Koch and Lengauer variant (algorithm 3 in [4]) that produces cliques only for connected patterns and thus additionally reduces complexity.

Clique detection is at present the only approach that guarantees finding all maximal patterns. However it is only practical for pairwise comparisons and additional methods are required for larger sets of examples, still leading to exponential increase of complexity with respect to the size of the set of examples. In practice this implementation tends to be slower (up to 10 times) in comparison with the repeated matching algorithm described above, although it sometimes produces better results.

Complexity results. The worst-time complexity for the learning algorithm (based on repeated matching) is $k*n^n$, where k is the number of sequences, n the number of secondary structures (helices and strands) in in sequence. The maximal clique method will give $(n^2)!$ for the same n and $k = 2$. 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. However, these are upper bounds and it is not clear whether they can not be improved.

References

[1] 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.

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

[3] 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.

[4] 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.