Poster I-56

Fast Structural Similarity Search based on Topology String Matching

Authors:
Sung Hee Park (*Bioinformatics Research Centre, Department of Computing Science, University of Glasgow*)
Keun Ho Ryu (*Database Laboratory, Department of Computer Science, Chungbuk National University*)
David Gilbert (*Bioinformatics Research Centre, Department of Computing Science, University of Glasgow*)

Short Abstract: We describe a framework for fast similarity search based on a string representation of topological relationships of proteins and report on the comparison of protein structures based on these topology strings using bipartite graph matching. The system is implemented on top of the Oracle 9i spatial DBMS.

Long Abstract:
The diversity techniques in protein structure comparison since the late 1980s demonstrates that structure comparison continues to be a challenging topic in Structural Bioinformatics. It has been shown that the complexity of structure similarity search for 3D structures is NP hard. Therefore, current algorithms use a mixture of heuristic methods and exhaustive search for structure similarity search. Most of the existing structural alignment algorithms use different abstractions of protein structures and algorithms of alignment as well as their own scoring functions. Due to the exponential growth of the size of the 3D structure databases and the complexity of similarity search, the issue of speed will become ever more important, and tools for fast structure similarity search will be required. Some methods use fast pre-filtering before performing accurate alignment. Even in this case, due to the lack of pre-computed and stored data, each structure in the database has to be scanned at least once. Therefore, such an exhaustive search is expensive to perform. Furthermore, the existing systems for structural similarity search may require improvement in accuracy in order that they can be used practically for structure classification.
In our approach, we adopt a multi step query processing technique that comprises a preprocessing step, a filter step and a refinement step. The preprocessing step entails the representation of the geometry of protein structures as spatial objects, the identification of topological relationships between SSEs (Secondary Structure Elements) and the transformation of these strings. The filter step includes approximate database search and matching of SSE string pairs, which we call topology string matching. The refinement step computes structural alignments and structural cores using further accurate comparison methods over the small candidate sets that pass through the filter step.
Our method is based on the spatial theory of Geographic Information Systems: topological properties of spatial objects are invariant even if geometric features such as length and angle are easily changed. We can infer that topological properties are preserved in conserved structures and help to identify the fold family in similarity searching. The computational cost of a topological match is less expensive than that of atomic-coordinate based geometry matching, and the comparison of topological properties finds global similarities. Thus, we

consider that topological properties of proteins are suitable for the filtering step.

In this paper, we describe a method to identify the topological relationships between SSEs using 9IM (Intersection Matrix) in order to detect the connectivity and order of SSEs. In our representation, topological relationships are transformed into linear strings in order to reduce of the complexity of the comparison algorithm of structural similarity and database search. The notation of the topology string we have developed has the ability to describe non-linear as well as linear topological properties. In order to perform fast structural similarity search, pairwise comparisons are made to find the maximum matching pairs of topology strings using a bipartite graph matching algorithm.

The system is implemented on top of the Oracle 9i spatial DBMS. The performance evaluation was conducted on 36 proteins from Chew and Kedem data set and on a subset of the PDB40. Our method performs acceptably well in terms of the quality of matching whilst having the advantage of fast execution and being able to compute solutions in polynomial time. This work shows that the pre-computed string representation of topological properties between secondary structure elements in proteins using spatial relationships of spatial databases is practical for fast structural similarity search. The system can be used as a front-end to systems such as DALI and SSAP that are more accurate but slower than our comparison algorithm, in order to speed-up the large scale analysis of proteins for structure classification.