Clustering of proteins based on their three-dimensional shape, using Maximum Common Subgraph

July Diana Banda Tapia
Cristian José López del Alamo
School of Computer Science - Arequipa, Perú

Abstract.- One of the main problems to clusterize proteins is given by the comparison method between them, in this sense, the present work propose the clustering of proteins using the maximum common subgraph in order to extract the maximum common subgraph between them and apply a metric that allow to determine the value of similarity. In this case, for the previous experiments was used an algorithm of hierarchical clustering and the McGregor algorithm to find the maximum common subgraph.

1. Motivation

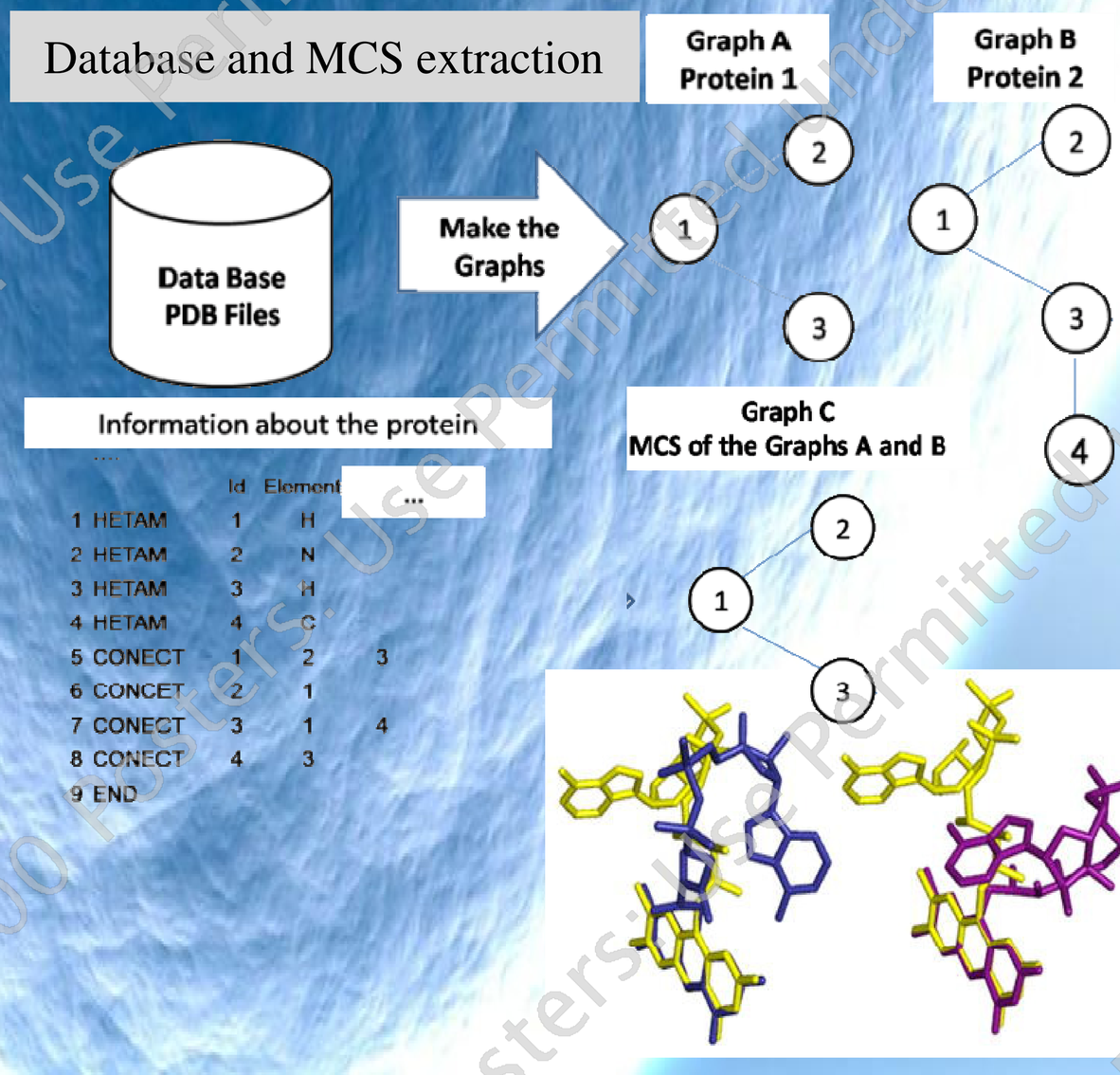
The clustering of proteins based on their 3D-shape is an important tool in the comparison of proteins with low similarity between their primary structure. Likewise, allows to infer the functionality and discover relations between them. All of these because the function of a protein is determined by its 3D-shape.

2. Proposal

A. Database, the database was obtained of the Protein Data Bank (PDB) and organized in 3 groups: of 97, 76 and 67 proteins.
B. Extraction of the Maximum Common Subgraph, using McGregor algorithm.
C. Comparison Process and Clustering, the score of similarity given by the maximum Common Subgrap between the graphs A and B is computed using the next equation:

$$\frac{V(MSC_{A,B}) + E(MSC_{A,B})}{V(A) + V(B) + E(A) + E(B)}$$

Database and MCS extraction



3. Conclusions and Results

Number of Clusters	Test Group 1	Test Group 2	Test Group 3
2	0.13	0.69	0.87
3	0.13	0.22	0.47
4	0.24	0.18	0.47

In order to improve the results the process should be done considering only the atoms of the backbone of the protein and non all of them, because a maximum common subgraph found outside the backbone has a different impact related with a maximum common subgraph of the backbone chain.

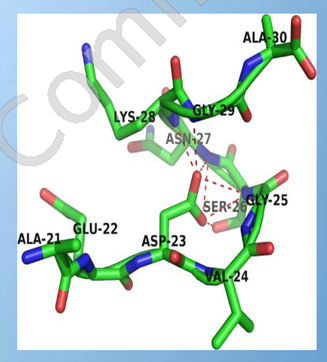
References

H.-C. Ehrlich and M. Rarey, "Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review," vol. 1, pp. 68–79, January/February 2011.

Yin-Hung Lin and Y.-L. Lin, "A study on tools and algorithms for 3d protein structure alignment and comparison," Int. Computer Symposium, 2004, Taipei, Taiwan.

Matriz of similarity

	Protein 1 Graph A	Protein 2 Graph B
Protein 1 Graph A	___	Sim(A,B)
Protein 2 Graph B	Sim(A,B)	___



Clustering