International Journal of Multidisciplinary Innovative Research
ISSN: 2583-0228 Volume 4, Number 4 (Oct' 2024) pp. 15-20
© CIIR, Noida, INDIA (www.ciir.in)
https://www.ijmir.org
Article Received: 29 July' 2024; Accepted: 14 Aug' 2024.

Review Article

A Graph Theoretical Approach in Similarity/Dissimilarity Study of Proteins

D. Vijayalakshmi

Assistant Professor, Department of Mathematics, Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Kanchipuram, Tamilnadu, India.

Email: guruviji97@gmail.com

ABSTRACT: Similarity/Dissimilarity between protein sequences is determined graph theoretically in this paper. The proteins are converted to protein graphs. The Similarity/Dissimilarity between protein graphs is measured using the maximal common sub-graphs (MCS) and union of the graphs (UG). The edges - size of the protein graph measures the percentage of similarity between them. Results obtained by these two measures are compared with the blast sequence site results and this proves the efficiency of these methods.

AMS subject classifications: 05C76, 05C60 2010: 46F12, 44A10

KEYWORDS: Protein Graphs, Similarity Dissimilarity, Maximal Sub Graph, Union of Graphs.

INTRODUCTION

Protein, the biopolymer is an essential and a highly complex substance present in every living organism. Protein study is a wide area of research. In this, similarity/dissimilarity study occupies an important place as it allows to observe the nature of a new protein whose primary and secondary structure is known. Mathematical methods are effective in summarizing and predicting biological characteristics with lower cost. Among various kinds of mathematical methods, graph theory is an essential one, which owns advantages in various protein structure identification problems including predicting protein structure. In this paper, the similarity study of protein gets converted into a graph problem. Some similarity study of proteins we see in the following part of the paper.

In [1] Amine *et al*. summaries the similarity measure based on the union of graphs and common maximal sub graph in detail. Bunke et al [2] proposes a novel measure based on distance on maximal common sub graph with algorithm. Graph edit distance for similarity and pseudo sub graph isomorphism elaborated with algorithm by Huahai in [3]. In [4], graph indexing using frequent sub tree for undirected labeled graph is narrated in detail. Bunke *et al.* [5] briefs the error correcting and error occurring graph matching and proves maximum common sub graph computation in equivalent to graph edit distance. In [6], the distance measure on semantics set for semantic similarity is described in detail.

Algorithm for graph comparison based on maximal common sub graph to verify chemical structures is described in [7] by Edmund et al. Algorithm based on maximal common edge sub graph, maximum common induced sub graph and maximum cliques were also discussed in detail. In [8], John et al briefs maximum common edge sub graph detection algorithm to

determine degree and composition of similarity which can be directly applied to any graph. This can be applied to search and predict biological activity.

In [9], Minot *et al*. similarity is measured using maximum common induced subgraph along with triangulation and 2- triangulation techniques. Distance metric based on maximum common sub graph is explained in [10] by Dwallis *et al*. In [11] Meng *et al*. narrates measures based on path, information, features. The semantic similarity measure is also explained along with advantages and disadvantage of the measure.

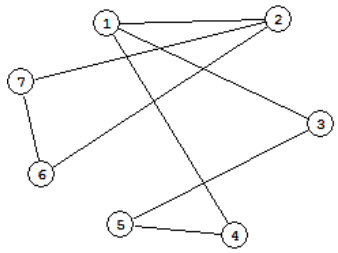
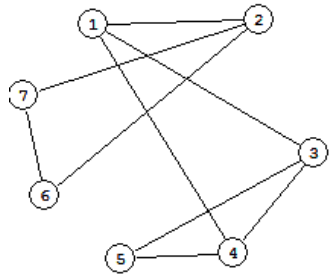
The text similarity measurement comparison is discussed in [12]. The degree of similarity is measured using string, corpus, knowledge and hybrid. Knowledge based measure is a path-based measure which is also known as edge counting measure. In [13], the adaption of six existing domain independent measure to biomedical domain is performed. The measures include path-based measure, information, content measure, context vector measure.

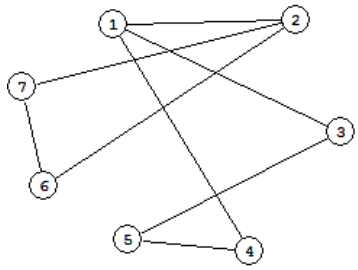
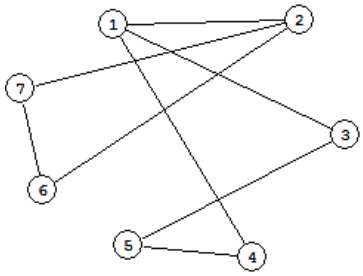
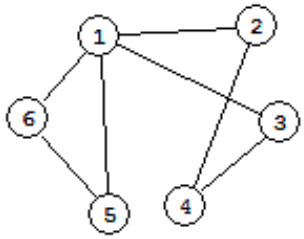
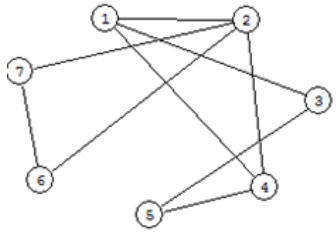
In [14] Luh yen et al counts similarity/ dissimilarity based on weight of graph. The dissimilarity between every pair of nodes also determined. In [15] Slimani briefs semantic similarity methods based on feature, hybrid, structure, information content and helps to choose the best method that fits the requirement. The paper is framed as below.

The protein graphs constructed and studied in [16] are used in this paper. The Similarity/Dissimilarity is calculated using maximal common sub-graphs and union of graphs. Following this the results and the conclusion are discussed. The results obtained by the above methods are compared with blast sequence site results.

PROTEIN DATA

Similarity/dissimilarity of proteins is studied using maximum common sub graph and union of the graphs. The formulas applied are discussed below.

PDBID	Protein data	Graph of proteins
1JXT	Crambin mixed sequence form at 160 K. Protein/water substates.	
1JXX	Crambin mixed sequence form at 200 K. Protein/water substates.	

1JXY	Crambin mixed sequence form at 220 K. Protein/water substates.	
1JXW	Crambin mixed sequence form at 220 K. Protein/water substates.	
1JXU	Crambin mixed sequence form at 240 K. Protein/water substates.	
1CCN	Direct NOE refinement of crambin from 2D NMR data using a slow-cooling annealing protocol.	

Method using Maximum common sub graph

$$SIM_{CMS}(G_1, G_2) = \frac{|CMS(G1,G2)|}{MAX\{|G1|,|G2|\}}$$

$|CMS(G1, G2)|$ = The number of edges in CMS (G_1, G_2)

$MAX\{|G1|, |G2|\}$ =The number of edges which is maximum among G_1 and G_2.

For the proteins 1JXT & 1JXX similarity / dissimilarity calculation based on MCS

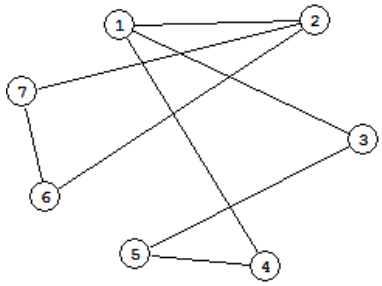
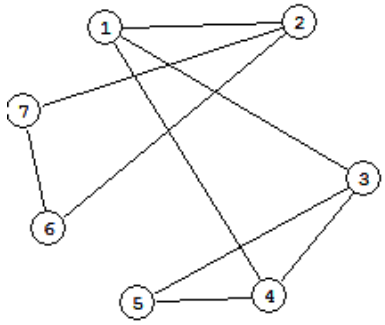
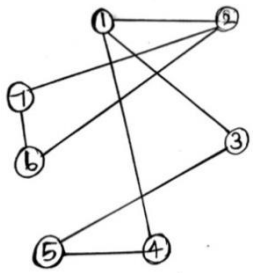
$|CMS(G1, G2)|$ = The number of edges in CMS (G_1, G_2) = 8

$MAX\{|G1|, |G2|\}$ =The number of edges which is maximum among G_1 and G_2 = 9

$|CMS(G1, G2)|$ = 8

$MAX\{|G1|, |G2|\}$= 9

$$SIM_{CMS}(G_1, G_2) = \frac{|CMS(G1,G2)|}{MAX\{|G1|,|G2|\}} = \frac{8}{MAX(8,9)} = \frac{8}{9} = \mathbf{0.88}$$

1JXT: G_1	1JXX: G_2	MCS (1JXT,1JXX):
		

1JXT & 1JXX-CMS

Method using Union of Graphs

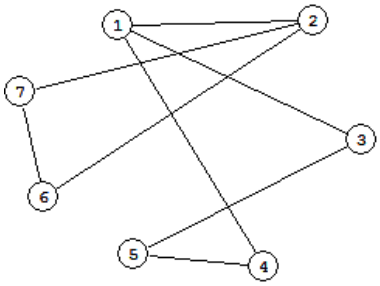
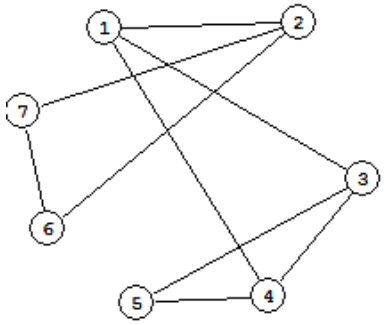
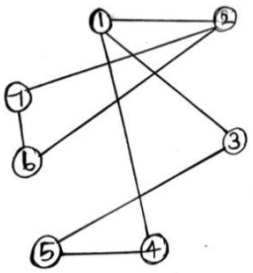
$$\mathbf{Sim_{UG}(G_1, G_2)} = \frac{|CMS(G1,G2)|}{|G1|+|G2|-|CMS(G1,G2)|}$$

$|CMS(G1, G2)|$ =The number of edges in CMS (G_1, G_2)

$|G_1|$= The number of edges in G_1

$|G_2|$= The number of edges in G_2

For the proteins 1JXT &1JXX similarity / dissimilarity calculation based on union of graphs

1JXT: G_1	1JXX: G_2	MCS(G_1,G_2)
		

1JXT & 1JXX-UOG

$|CMS(G1, G2)|$ =The number of edges in CMS (G_1, G_2) = 8

$|G_1|$= The number of edges in G_1=8

$|G_2|$= The number of edges in G_2=9

$$SIM_{UG}(G_1, G_2) = \frac{|CMS(G1,G2)|}{|G1|+|G2|-|CMS(G1,G2)|} = \frac{8}{8+9-8} = \frac{8}{9} = \mathbf{0.88}$$

Result

Based on the sim_{CMS} and sim_{UG} the similarity percentage between each pair of protein is calculated and is compared with blast sequence results. The details are shown below.

PDBID	**1JXT**	**1JXX**	**1JXY**	**1JXW**	**1CCN**	**1JXU**
1JXT	1	**100** 0.88 0.88	**100** 1 1	**100** 1 1	**95-98** 0.88 0.88	**95-98** 0.88 0.88
1JXX		1	**100** 0.88 0.88	**100** 0.88 0.88	**95-98** 0.88 0.8	**95-98** 0.88 0.8
1JXY			1	**100** 1 1	**95-98** 0.88 0.88	**95-98** 0.88 0.88
1JXW				1	**95-98** 0.88 0.88	**95-98** 0.88 0.88
1CCN					1	**100** 1 1

Similarity/Dissimilarity percentage of proteins

Values in **Bold letters** represent the result obtained by from blast sequence site and the values below are result by our methods. i.e, Underlined result is obtained by Method-1(CMS) and the values below are by Method-2(UG).

CONCLUSION

In this part the similarity is measured based on the common maximal sub-graphs and the union of graphs. The number of edges in protein graphs is taken as parameter. This is a novel and simple method to measure similarity. The results obtained by these two methods are exactly equal to the results of blast sequence site. These two methods prove their efficiency in measuring similarity by consuming very less time for calculation.

REFERENCES

[1] Amine Labriji, Salma Charkaoui, Issam Abdelbaki, Abdelouhaed Namir, ElHoussine Labriji. "Similarity Measure of Graphs", IJES – Vol. 5, No. 2, 4 April 2017.

[2] Horst Bunke, Kim Shearer, "A graph distance metric based on the maximal common subgraph"1998 Elsevier Science B. V. Pattern Recognition Letters 19, (1998) 255–259.

[3] Huahai He, Ambuj K. Singh, "Closure-Tree: An Index Structure for Graph Queries", Proceedings of the 22nd International Conference on Data Engineering (ICDE'06) 2006 IEEE.

[4] Shijie Zhang, Meng Hu, Jiong Yang, "Tree Pi: A Novel Graph Indexing Method"2007 IEEE.

[5] H. Bunke, "On a relation between graph edit distance and maximum common Subgraph", Pattern Recognition Letters.1997 Elsevier Science B.V.

[6] RoyRada, HafedhMili, Ellen Bicknell, And Maria Blettner, "Development and Application of a Metric On Semantic Nets" IEEE Transactions On Systems. Vol19. No 1. January/February 1989.

[7] Edmund Duesbury1, John D. Holliday1, Peter Willett. "Maximum Common Subgraph Isomorphism Algorithms" MATCH Communications in Mathematicaland in Computer Chemistry, ISSN 0340 – 6253.

[8] John W. Raymond, Eleanor J. Gardiner, Peter Willett. "RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Sub-graphs", The Computer Journal, Volume 45, April 2002.

[9] M. Minot and S. N. Ndiaye, "Searching for a maximum common induced sub-graph by decomposing the compatibility graph" by LIRIS, France.

[10] W. Dwallis, P. Shoubridge, M. Kraetz, D. Ray, "Graph distances using graph union" Pattern recognition letters 22 (2001) Elsevier Science B.V.

[11] Lingling Meng, Runqing Huang, Junzhong Gu. (2013) "A Review of Semantic Similarity Measures in Wordnet." International Journal of Hybrid Information Technology. (Grant No. 11530700300).

[12] Didik Dwi Prasetya, Aji Prasetya Wibawa, T. Sukasa Hirashima, (2018) "The performance of text similarity algorithms" International Journal of Advances in Intelligent Informatics. ISSN: 2442-6571.

[13] Ted Pedersen, Serguei V.S. Pakhomov, Siddharth Patwardhan, Christopher G. Chute (2007) "Measures of semantic similarity and relatedness in the biomedical domain", Journal of Biomedical Informatics.

[14] LuhYen, Amin Mantrach, Masashi Shimbo, (2008) "A Family of Dissimilarity Measures between Nodes Generalizing both the Shortest Path and the Commute time Distances", Conference Paper.

[15] Thabet Slimani, "Description and Evaluation of Semantic similarity Measures Approaches", Computer Science Department Taif University & LARODEC Lab.

[16] Vijayalakshmi D., SrinivasaRao and K. Sivakumar, (2013) "Methods of construction of a graph for a protein using secondary structural elements" presented at XVII Ramanujan, Symposium organized by Ramanujam Institute for advanced study in Mathematics, pp: 25-27.