ANALYSIS ON SIMILARITY/DISSIMILARITY OF PROTEIN USING TRIVIALTHAT THE DA MATRIXDA(L)

[1]K. Pushpa Latha, [2]G.Sampath Kumar, [3]Ravi Padma
[3]Assistant Professor, [1,2,3]Department of H & S, Brilliant Institute of Engineering and Technology, Hyderabad, India

ABSTRACT
This study of the similarity and dissimilarity of proteins makes use of protein graphs with secondary structure elements as a parameter. The key to this study is a novel Distance- Adjacency matrix and associated Eigen values. The longest path between any two vertices is used as a constraint while building the distance matrix. For protein graphs, the distance-adjacency matrix's first Eigen value is thought to represent the degree of similarity or dissimilarity between proteins.

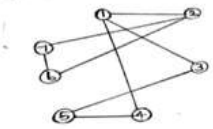
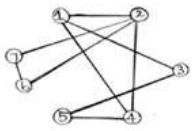
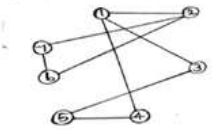
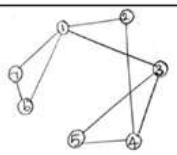
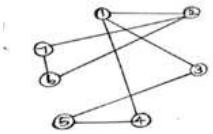
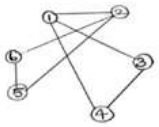
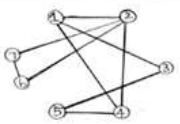
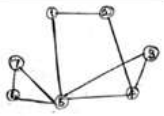
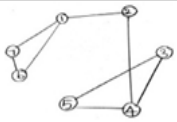
INTRODUCTION

By using graph theory, Yan Yan [1] offers a number of approaches to the problem of identifying protein structures. Spectral techniques are used to identify side chain clusters in protein structures. The second-lowest Eigen value and associated vector of the Laplacian matrix are used to create clusters. From the top Eigen value and associated vector, side chains that increase the number of interactions in a cluster are obtained. A 2D graphical representation of a protein sequence based on the Huffman tree approach was described by Qi et al. [2]. Based on the singular values decomposition of a matrix mapped from the original amino sequences, Yu et almethod .'s [3] reduced main sequences into 20 tuple mathematical descriptors. Based on the physical characteristics of twenty amino acids, He et al. [4] suggested a graphical depiction of protein sequences. Dragos [5] discusses various spectra of graphs. The similarity between the proteins using spectral distance is explained in detail (ie) if the spectral distance is small, the graphs are similar, if zero there are cospectral and if the distance is high the graphs are dissimilar. In normalized laplacian of biological network graph is studied. D.vijayalakshmi, K. srinivasaraoand K.sivakumar [7] have constructed graph for protein in three different ways using secondary structural elements as vertices. The edges are drawn based on average distance between 3D coordinates of nitrogen, alpha carbon atom, beta carbon atom of the amino acids in the SSE.Charalambschrysostomou[8] developedComplex information Spectrum Analysis for Protein Sequences (CISAPS) and its web based server is also developed and presented. CISAPS is constructed to consider and provide results in three forms including absolute, real, imaginary spectrum.Sheng-lung peng and yu-weiTsay [9]study the protein structural similarity using the spectra of adjacency matrix, Laplacian matrix, signlesslaplacian matrix, seidaladjacency matrix. The similarity measured based on the Euclidean distance between the spectra of seidal adjacency matrix yields a better result. Sheng-lung peng and yu-weiTsay measures the stability of the graph constructed for proteins.In peng et.al [10] the proteins are represented as graph, the laplacian matrix for the graph are considered.The similarity between proteins are measured using Euclidean distance of laplacian spectra. The important fact discussed in this journal is the stability of the graph constructed for the protein and the stability verified using entropy.In [11] ,proteins are converted to graph and Degree Distance matrix of the protein graph is obtained. The least positive Eigen value of the DD matrix is considered as parameter to measure similarity between proteins. In this paper, the graphs for protein are constructed in the way as specified in[7]. The similarity / dissimilarity between proteins are measured using Distance – Adjacency matrix DA (l) of protein graphs. We narrate a novel method to measure

similarity/dissimilarity between proteins in which the distance matrix is constructed taking the longest path between the vertices as parameter.

DA matrix and protein similarity/dissimilarity
Protein graphs considered in this part are constructed using secondary structure elements as vertices and edges are drawn using centroids of the vertices. The centroids are the average of 3-D co-ordinates of the central carbon atom of the amino acids in the particular structure[7].We consider 10 proteins and they are 1jxt, 1jxy, 1jxw, 1jxx, 1ccn, 1jxu, 3u7t, 2eya, 1ab1, 1wuw

PROTEINS AND THEIR DETAILS

Protein Structure	Graph for Protein	Protein Structure	Graph for Protein
1.1jxt		6.1jxu	
2.1jxy		7.1ab1	
3.1jxw		8.1wuw	
4.1ccn		9.2eya	
5.3u7t			

Proteins and Protein Graphs

METHOD
Longest Path & Modulus of First Eigen Value (LP&MFEV)
Distance Matrices also known as path length matrices as it gives details about vertices of path length>1

$$Dist(i,j) = \begin{cases} \text{longest length of path between } v_i \text{ and } v_j & \text{otherwise} \\ o & i=j \ (v_i \text{ is adjacent to } v_j) \end{cases}$$

$$a_{ij} = \begin{cases} 1 & if\ (i,j) \in E \\ 0 & otherwise \end{cases}$$

We construct the Distance-Adjacency Matrix definedby Distance matrix – adjacency matrix and it is denoted by DA(l). Let □1 □ □2 □ □3 □ .....□□*n* be the Eigen value of DA(l) matrix.The Eigen Value of the matrices satisfy the following condition.

□ $\sum_x = 0$ − 1

Sum of Values lies between 0 and -1

Distance-Adjacency Matrix DA(l) is

a. DA matrix is a non-singular matrix
b. This is a real symmetric matrix
c. The diagonal entries are zero
d. Indefinite matrix

The Modulus of First Eigen Value (MFEV) of the DA (l) matrix is taken into account to measure similarity/dissimilarity of protein.First we find the DA(l) matrix for all protein graphs. Then we calculate the Eigen Values of all DA (l)matrix .Next we consider set of MFEV of all matrices. For a protein the difference between MFEV of each other proteins are calculated based on that the similarity/dissimilarity is measured.
The Modulus of First Eigen Values of the DA (l) matrices of proteins are given below.

S.No	Protein structure	MFEV □ 1
1	1JXT	12.4877
2	1JXY	12.4877
3	1JXW	12.4877
4	2EYA	11.2511
5	1AB1	11.2187
6	3U7T	11
7	1JXU	11.3808
8	1CCN	11.3808
9	1WUW	8

Modulus of First Eigen Value (MFEV) of DA(l).

The Similarity and Dissimilarity of proteins are measured using the differences between MFEV. For a Protein, the difference value (DV) between its MFEV and remaining Protein's MFEV are calculated. Based on this DV the percentage of similarity is measured. The details of range of DV and its corresponding similarity percentage are given below in the table.

S.No	DV	% Similarity
1	0-0.5	100%
2	0.5-1	96%-98%
3	1-1.5	90% - 95%
4	1.5-2	80% - 90%
5	2-2.5	70% - 80%
6	2.5-3	65% - 70%
7	3 -3.5	60%-65%
8	3.5-4	55% - 60%

Difference Value and Similarity percentage

We start the study of similarity/ dissimilarity with 2eya Protein

Protein Name	Difference Value
1ccn,1jxu	0.1297
1ab1	0.0324
1jxt,1jxy,1jxw	1.2366
3u7t	0.2511
1wuw	3.2511

DV of 2eya protein with remaining proteins

The percentage of similarity of 2eya with the remaining proteins based on the DV values is given below.

Proteins Name	**% Similarity**
1ccn, 1jxu	100
1ab1	100
1jxt,1jxy,1jxw	90–95
3u7t	100
1wuw	58-60

Percentage of similarity of 2eya with other proteins

RESULT

Based on the DV the similarity percentage between each pair of protein is calculated and is compared with **blast sequence** results. The details are shown below

	1jxt	1jxy	1jxw	1ccn	1jxu	1ab1	3u7t	2eya	1wuw
1jxt	0	**100** 100	**100** 100	**90-95** 95-98	**90-95** 95-98	**90-95** 100	**95** 100	**90-95** 96	**50-54** 56-58
1jxy		0	**100** 100	**90-95** 95-98	**90-95** 95-98	**90-95** 100	**95** 100	**90-95** 96	**50-54** 56-58
1jxw			0	**90-95** 95-98	**90-95** 95-98	**90-95** 100	**95** 100	**90-95** 96	**50-54** 56-58
1ccn				0	**100** 100	**100** 95-97	**100** 95-97	**100** 100	**58-60** 53-56
1jxu					0	**100** 95-97	**100** 95-97	**100** 100	**58-60** 53-56
1ab1						0	**100** 100	**100** 100	**58-60** 56-58
3u7t							0	**100** 96	**60** 55
2eya								0	**58-60** 56
1wuw									0

Similarity/Dissimilarity percentage of Proteins
Values in **Bold letters** represent the result obtained by **our method** and the values below are result from blast sequence site.

Conclusion

It is trivial that the Eigen values of the DA matrix DA (l) serve as a novel and exclusive parameter in determining similarity. The findings acquired using this method is extremely similar to those from the website, demonstrating their accuracy. The approach, despite being easier to use, is effective at detecting similarities and differences across proteins. The DA matrix is built using the path length between the vertices, whereas protein graphs are built using the distance between the vertices. Because of this, our matrix is a unique parameter, and our method is a straightforward and accurate way to determine how similar or dissimilar two proteins are analyzed.

REFERENCES

1. Y. Yan &shengguizhang ,et al.: proteomo science 2011 9(suppl 1):S17 https://doi.org/10.1186/1477-5956-9-S1-S17, Application of graph theory in protein structure identification
2. Qi ZH, Feng J, Qi XQ, Li L. Application of 2D graphical representation of protein sequences and its application. J computChemk 2011;32:2539-44.
3. Yu HJ, Huang DS. Novel 20-D descriptors of protein sequences and its application in similarity analysis. ChemphysLett 2012;531:261-6.
4. He PA, Wei JZ, Yao YH, Tie ZX. A novel graphical representation of proteins and its application. Phys A: Stat Mechappl 2012;391:93-9.
5. DragosCvetkovic, Spectral Recognition of graphs, Yugoslav Journal of operation research 22(2012), November2,\15-16\DOI:10.2298\yJOR120925025c.
6. The spectrum of the graph Laplacian as a tool for analyzing structure and Evolution of networks dissertation by Banerjee 2012, www.iiserkd.ac.in>Banerjee-ph.D-Thesis.
7. D. Vijayalakshmi, K.SrinivasaRao and K.Sivakumar, Methods of construction of a graph for a protein using secondary structural Elements, proceeding of XVII Ramanujan symposium of Recent Trends in Dynamical System and Mathematical Modeling, Organized by Ramanujan institute for Advanced study in Mathematics, University of Madras, Chennai, during,25-27(2013).
8. Charalamboschrysostomou,huseyinseker, et al:advance in bioinformatics volume 2015,Article ID 909765, http://dx.doi.org/10.1155/2015/909765, CISAPS: Complex Information Spectrum for the Analysis of Protein Sequences
9. sheng-lungpeng, Yu Wei TsayAdjusting Protein graphs based on graphs entropy BMC Bio Informatics. 2014 (suppl15):S6 http\\www.biomedcentral.com 1471-2105\15\S15\S6.
10. Sheng-Lung Peng and Yu-Wei Tsay on the usage of graph spectra in protein structural similarity, www.csroc.org,tw\journal\JOC23-2-abs\JOC 23-2-7-abs.pdf.
11. D.Vijayalakshmi, K.SrinivasaRao, DD Matrix with Least Positive Eigen Value and Protein Similarity, International journal of pure and Applied Mathematics, volume 115 No.9 2017,331-342 ISSN:1311-8080(printed version);ISSN:1314-3395(online version)