International Journal of Multidisciplinary Innovative Research
ISSN: 2583-0228 Volume 4, Number 1 (Jan' 2024) pp. 1-7

https://www.ijmir.org
Article Received: 4 Oct' 2023; Accepted: 6 Nov' 2023.

Review Article

Protein Study using Spectral Radius and Canberra Distance

D. Vijayalakshmi

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

Email: guruviji97@gmail.com

ABSTRACT: ***Protein study utilizing spectral radius and Canberra distance is an innovative approach in the field of bioinformatics and computational biology. The primary structure–amino acid sequence of protein plays the vital role in this study. The spectral radius of amino acid is considered as parameter. The value of spectral radius is assigned to the amino acid of proteins under consideration as a vector. The Canberra distances between each pair of proteins are calculated. For a protein the other proteins are placed in an assumed neighborhood based on the distance obtained to study the similarity /dissimilarity of protein. Integrating these two techniques, scientists can gain a deeper understanding of protein structures and their functional implications, making significant strides in drug design, disease research, and the broader realm of molecular biology.***

KEYWORDS: ***Spectral radius, Canberra distance, Amino acid sequence, similarity/dissimilarity study, Protein Study.***

INTRODUCTION

Similarity/dissimilarity study of protein based on sequence can be categorized into two methods namely alignment-based methods and alignment free methods. As alignment-based method requires more time and more memory the alignment free method developed as an alternate to overcome the issues. In some alignment free method, the amino acids are given a numerical value based on their physicochemical properties. The features of proteins were extracted to achieve the goal. In some methods the amino acids sequence is represented as 2 D graphs and similarity / dissimilarity analysis were done. In this part Some alignment free methods are in [1] the power spectra of protein sequences are obtained from Discrete Fourier transform and Dynamic time wrapping where the protein sequences are converted to numerical sequence based on hydropathy and isoelectric point properties of amino acids. Phylogenetic tree was constructed to classify different species. In [2] pKa (NH3+) value of amino acids plays a key role in 2D graph representation of protein and deep analysis is carried out to learn similarity/ dissimilarity between proteins.

In [3], a novel 2D graphical representation of protein sequence based on pKa value is constructed and similarity/dissimilarity between proteins are obtained. In [4] amino acids are arranged in a cyclic order based on their physicochemical according to chaos game representation. Similar fragments of amino acids are identified using Euclidean distance. Based on that the similarity /dissimilarity of protein is measured. In [5] the amino acids in the protein

sequences are characterized by six physicochemical properties. In addition to similarity study this representation encodes the structure of proteins.

In [6] the amino acids in protein sequence are represented by right cone with unit base radius and unit height. The spatial median (mx.my.mz) represents the protein residues and characterize protein numerically. By using correlation angle or Euclidean distance between protein descriptor the similarity / dissimilarity is studied. In [7], based on five letter model of the 20 amino acids a new 2D representation of protein is constructed. Phylogenetic tree of 56 corona virus spike protein is also constructed.

In this paper the amino acids in the protein sequence are represented by their spectral radius. These spectral radius of amino acids forms a vector representation of proteins. The Canberra distance between each pair of vectors of proteins are obtained [8]. For each protein the remaining proteins are placed in a neighbourhood based the distance value [9]. The neighbourhood where the proteins are placed measures the similarity percentage between proteins.

METHODOLOGY

The spectral radius of amino acids stated in [8] are used to represent amino acids in the protein sequence. The values of spectral radius are shown in Table 1 and the protein data used is shown in Table 2. The graphical representation of proteins based on spectral radius are shown in Figure 1, Figure 2, Figure 3 and Figure 4.

Table 1: Spectral radius value of Amino acids

Amino acids	Radius value	Amino acids	Radius value
A	6.04	M	2.24
C	2.5	N	3.26
D	3.02	P	8.13
E	2.46	Q	2.46
F	3.26	R	8.32
G	8.13	S	5.19
H	2.93	T	3.15
I	5.15	V	3.61
K	5.1	W	2.24
L	7.42	Y	3.26

Table 2: Illustrates protein data

PDB ID	Protein name
1jxx	Crambin mixed sequence form at 200 k. Protein/water substates
1jxw	Crambin mixed sequence form at 180 k. Protein/water substates

1cbn	Atomic resolution (0.83 angstroms) crystal structure of the hydrophobic protein crambin at 130 k
1ccn	Direct noe refinement of crambin from 2d nmr data using a slow-cooling annealing protocol
1jxu	Crambin mixed sequence form at 240 k. Protein/water substates
2fd7	X-ray crystal structure of chemically synthesized crambin

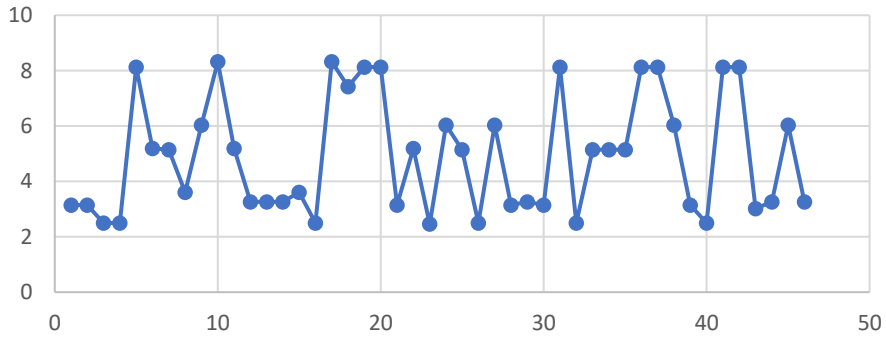

Figure 1: 2D graphical representation of protein sequence 1JXX

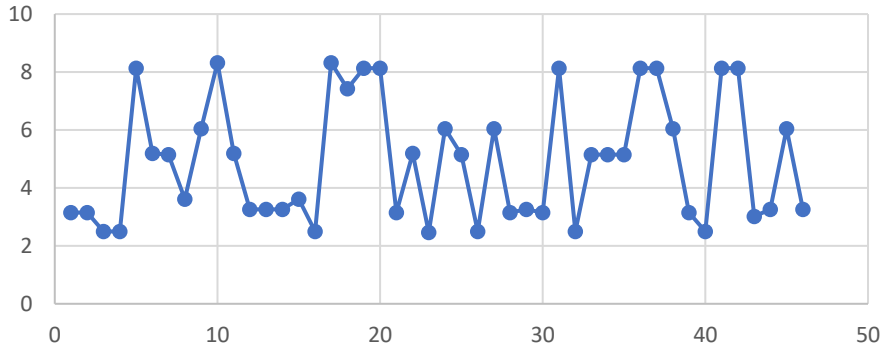

Figure 2: 2D graphical representation of 1JXW

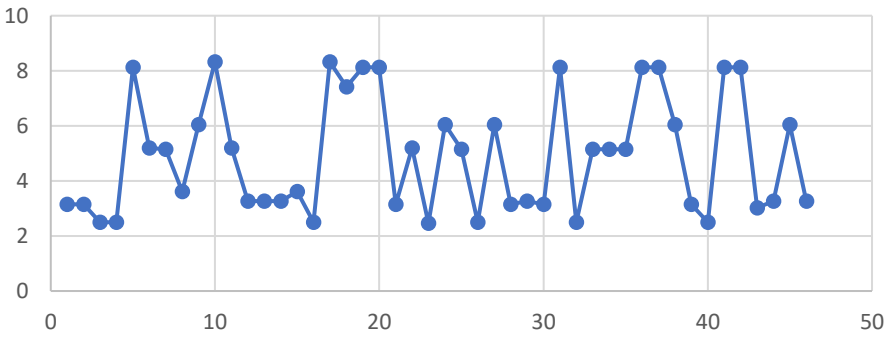

Figure 3: 2D graphical representation of 1CBN

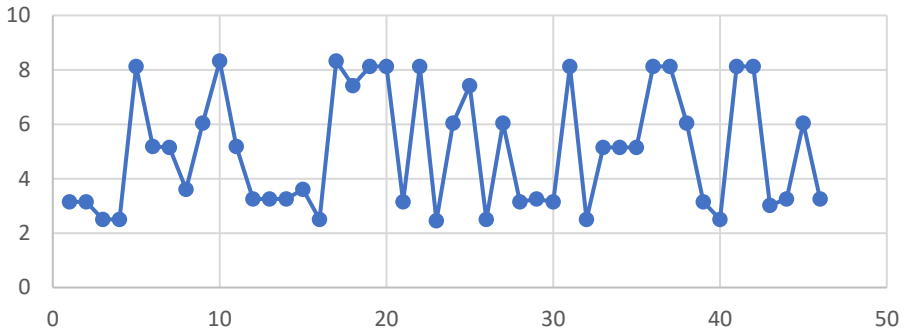

Figure 4: 2D graphical representation of 1CCN

Canberra distance

The Canberra distance, a metric for measuring dissimilarity between two data sets, is applied to compare protein sequences or structures, enabling researchers to assess the evolutionary relationships or structural variations between proteins. Canberra distance measures the distance between pair of points represented as vectors.

$$C(p,q)=\sum_{i=1}^{n}\frac{|p_i-q_i|}{|p_i+q_i|}$$

Where, p_i&q_i are the corresponding elements two vectors. Here p_i&q_i are the corresponding amino acid values in corresponding vectors (protein sequence). The Canberra distance between each pair of proteins is calculated and the values obtained are given below.

Table 3: Canberra Distance

	1JXW	1CBN	1CCN`	1JXU`	2FD7
1JXX	0	0	0.4013	0	0.4013
1JXW		0	0.4013	0	0.4013
1CBN			0.4013	0	0.4013
1CCN				0.4013	0
1JXU					0.4013

Based on the above values the proteins are placed in neighborhood. The classification of neighborhood is shown in table 4. It can be observed that as the distance value increases the percentage of similarity decreases.

Table 4: Neighborhood and percentage of similarity

Distance value	**Neighborhood**	**Percentage of similarity**
0	0-Neighborhood	100
0.1-0.2	0.2 Neighborhood	98-99
0.3-0.4	0.4 Neighborhood	96-97
0.5-0.6	0.6 Neighborhood	94-95

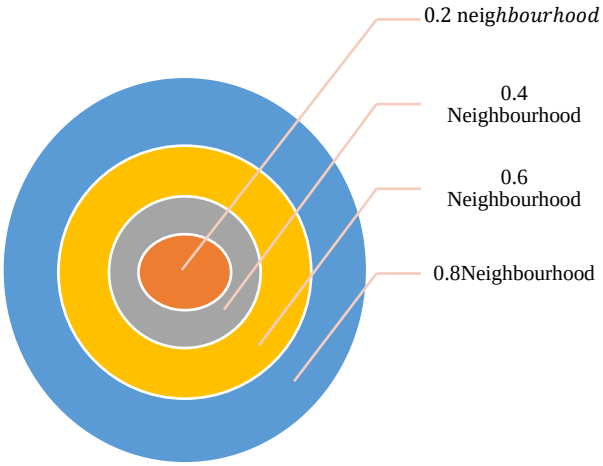

Figure 5: Illustrates the Neighborhoods

The calculation of similarity on comparing 1JXX with other proteins is shown in Table 5. Canberra distance between 1jxx and remaining proteins are given in the Table 5.

Table 5: Distance of 1JXX

Protein	Distance
1JXW	0
1CBN	0
1CCN	O.4013
1JXU	0
2FD7	0.4013

Based on the values of the Canberra distance for 1JXX the proteins are placed in neighborhoods and it is shown in the below diagram.

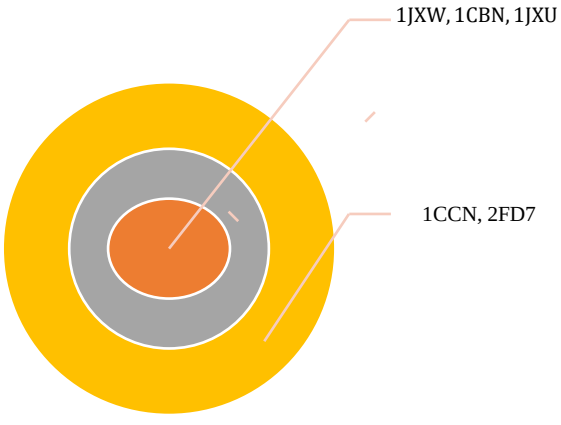

Figure 6: Neighborhoods of 1JXX

1JXW,1CBN,1JXU protein lies in the 0- Neighborhood. Therefore, these proteins are 100% similar to 1JXX. 1CCN and 2FD7 lies in 0.4 Neighborhood of 1JXX, therefore these 2 proteins 96% similar to 1JXX. In the Similar way the percentage of similarity is measured and

percentage obtained is given in the table 6. In Table 6, the bold values are the data obtained from blast protein sequence site.

Table 6: Results obtained.

	1JXW	1CBN	1CCN	1JXU	2FD7
1JXX	100 **100**	100 **100**	96 **95.65**	100 **100**	96 **95.65**
1JXW		100 **100**	96 **95.65**	100 **100**	96 **95.65**
1CBN			96 **95.65**	100 **100**	96 **95.65**
1CCN				96 **95.65**	100 **100**
1JXU					96 **95.65**

CONCLUSION

Protein study, a critical field in bioinformatics and molecular biology, often relies on advanced mathematical and computational techniques to analyze and compare protein structures. Spectral radius, a mathematical concept derived from the eigenvalues of a matrix, can be employed to analyze the structural properties of proteins, providing insights into their stability and conformation. By representing proteins as graphs, with amino acids as nodes and their interactions as edges, spectral radius can help identify key structural features critical for protein function. This is a novel method in measure of similarity/ dissimilarity between proteins. Determining the distance based on corresponding amino acids in the corresponding protein vector adds value to this method. As the similarity is measured based on primary amino acid sequence also adds value to this method. The result obtained by this method is compared with existing result and this proves the efficiency of method. Therefore, this method is simple and efficient. Additional, together these two methods provide a robust framework for studying proteins at a structural level, shedding light on their roles in biological processes and aiding in the development of novel therapeutics and treatments.

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] Wenbing Hou, Qiuhui Pan, Qianying Peng, Mingfeng He, “A new method to analyze protein sequence similarity using Dynamic Time wraping “, Genomics 109(2) 123 2017

[3] Guohua Huang, Jerry Hu, “Similarity/Dissimilarity analysis of protein sequences by new graphical representation “, curr.bioinf 8(5) 539-544 2013

[4] Yangfan Li, Guahua Huang, BeLiao, Zanbo Liu “H-L: A novel 2D graphical representation of protein sequences”, match Commun. Math.Comput.chem 61 , 519-532, 2009

[5] Ping An He, Yan Ping Zhang, Yu- Hua Yao, Yi-Fa Tang, Xu- Ying Nan," The graphical representation of protein sequences based on the physicochemical properties and its applications J comput chem (11), 2136-42, 2010

[6] Yu. Hua Yao, Qi Dai, Ling Li, Xu Ying Nan, Ping- AnHe, Yao Zhou Zhang," Similarity/ dissimilarity studies of protein sequences based on a new 2D graphical representation ", J. comput. chem. ,31(5),1045-52,2010

[7] Mervat M Abo Elkhier," Similarity/dissimilarity analysis of protein sequences using spatial median as a descriptor", Journal of Biophysical chemistry vol 3, 142-143,2012

[8] Chun Li, Lili XING, Xin Wang,"2D graphical representation of protein sequences and its application to coronavirus – phylogeny" Korean society for biochemistry and molecular biology vol 41 issue 3, 217-222, 2008

[9] Chuanyan Wu, Rui Gao, Yang De Mariris,Yusen Zhang,"A novel method for protein sequence similarity analysis based on spectral radius", Journal of theoretical biology, 446,61-70,2011