Article

Evolutionary aspect of protein sequence network based on the 2D representation of amino acids

Sanjay Sharma, **Birinchi Kumar Boruah**, **Tazid Ali**
Department of Mathematics, Dibrugarh University, Assam, India 786004
E-mail: snjyshrma90@gmail.com

Received 7 June 2022; *Accepted 15 July 2022*; *Published online 16 August 2022*; *Published 1 December 2022*

Abstract
For the comparative analysis of proteins, their proper clustering, and evolutionary relationships require analysis of their sequences. We used a mathematical parameter termed a *similar factor* to create a similar degree matrix of ND6 protein sequences taken from eight different species in this paper. We built a network out of the matrix to analyze their evolutionary and similarity trends with each other. By observing the various centrality measures, the correlation between multiple centrality measures and different network parameter shows that our network is consistent with the known evolution fact of ND6 protein sequences.

Keywords amino acids; similar factor; similar degree matrix; characteristic vector; centrality measures; clustering coefficient.

Network Biology
ISSN 2220-8879
URL: http://www.iaees.org/publications/journals/nb/online-version.asp
RSS: http://www.iaees.org/publications/journals/nb/rss.xml
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

The biological systems in nature are formed up of cells, with each cell's chromosomes serving as a core for living beings. Chromosomes are the aggregation of genes (DNA sequences), each creating a specific protein. Proteins are linear sequences of amino acids, essential building blocks and operating components of living organisms. The three-base genetic code consists of four nitrogenous bases: thymine T or uracil U, adenine A, cytosine C, and guanine G. A codon is a unit that consists of three nitrogenous bases. A codon is a three-letter code that represents each amino acid. The genetic code is a set of codons that specifies which amino acids are needed to generate a particular protein. Overall, 61 codon triplets correspond to the 20 amino acids Phenylalanine (P), Leucine (L), Isoleucine (I), Methionine (M), Valine (V), Serine (S), Proline (P), Threonine (T), Alanine (A), Tyrosine (Y), Histidine (H), Glutamine (Q), Asparagine (N), Lysine (K), Aspartic acid (D), Glutamic acid (E), Cysteine (C), Tryptophan (W), Arginine (R), Glycine (G), with the other three triplets, UAA, UAG, and UGA, being known as stop codons or nonsense codons. These nonsense and stop codons do not affect forming an amino acid. Codon degeneracy is a mechanism in which different codons code for the same amino acid. Amino acid sequence analysis is crucial for understanding protein structure and function in

the cell. So, the study of protein sequences for similarity and dissimilarity is essential because proteins with similar sequences usually have identical structures.

There are two methods of sequence comparisons: one is alignment-based computer oriented and the other is alignment-free. The alignment-based approaches, however, have a considerable computational cost. Alignment-free graphical representation contributes equally to outcomes and has a low processing cost. Since 1983, several researchers have represented DNA and protein sequences in various dimensional spaces (Hamori and Ruskin, 1983; Hamori, 1985; Gates, 1986; Leong and Morgenthaler, 1995; El-Lakkani and El-Sherif, 2013; Yao et al., 2009). Many distance matrices and mathematical descriptors were provided for numerical and graphical comparisons by treating each nucleotide and amino acid in a given DNA and protein sequence as a point in different dimension spaces (Bai and Wang, 2006; Abo-Elkhier et al., 2019; Pearson, 2013; Bajusz et al., 2021). Xie et al. (2012) proposed a novel approach for assessing the similarity/dissimilarity of protein sequences based on the protein sequence's conditional probability. The unique method was demonstrated using the protein sequences of eight species' ND6 (NADH dehydrogenase subunit 6) proteins. Gupta et al. (2014) introduced a two-dimensional graphical representation of protein sequences. The recommended approach is evaluated on ND6 protein sequences from eight different species. Their graphical representation is used to build a probabilistic distribution of protein sequences and measure sequence similarity using relative entropy (Kullaback-Leibler divergence). Ali et al. (2016) created an amino acid distance matrix and demonstrated that the network derived from the distance matrix depicted the amino acid evolutionary trend.

Similarly, the evolutionary homology of the ND6 protein sequences taken from eight different species is investigated in this paper by various graph theories concept based on aspects of their network. Here, we restrict our study to only a few protein sequences homology. The following is how the paper is structured: In section 2, we go over some basic graph theory concepts and several centrality measures. Section 3 designs a network in protein sequences based on a similar degree matrix and examines several centrality measurements in the network. Further, we discuss several network parameters of the protein sequence network in section 4, and finally, we gave the paper's conclusion in section 5.

2 Material and Methods

2.1 Some basic concepts of graphs

An undirected graph $G = (V, E)$ (Bertman and Jungck, 1979) has a finite number of vertices V and a finite number of edges $E \subseteq V \times V$. The vertices u and v are said to be incident with the edge e and next to each other if an edge $e = (u, v)$ links them. The neighbourhood $N(u)$ of u is the collection of all vertices near u. A directed graph, or digraph G, is made up of a set V of vertices and a set E of edges, where $e \epsilon E$ is the direction of each edge in the graph G. If no edge connects a vertex to itself, the graph is said to be loop-free. A graph's adjacency matrix A is a $(n \times n)$ matrix, with a_{ij}= 1 if and only if $(i, j) \in E$, and $a_{ij} = 0$ otherwise (Zhang, 2016, 2018). Any undirected graph's adjacency matrix is symmetric. The degree of a vertex v is defined as the number of edges with v as one of their end nodes. A walk is a finite alternating series of vertices and edges that starts and ends with vertices and has each edge coincide with the vertices before and after it. In a walk, no edges emerge more than once. However, a vertex can occur several times. Starting and ending vertices are initial and terminal vertices in a walk. If the initial and terminal vertices coincide, the walk is closed; otherwise, it is open. A walk is said to be a path, if it does not have any repeating vertices. A walk is said to be the trail, if it has no repeated edges. A path with the shortest or geodesic length between two vertices u, v is called a shortest or geodesic path. A graph is linked if every pair of its vertices can be walked between.

2.2 Centrality in protein sequence network graph

2.2.1 Degree centrality $(deg(k))$ (Freeman, 1978) is the optimal topological index, referring to the range of

nodes directly adjacent to a given node v, where neighbouring means are attached. The nodes that are directly linked to a given node v are also known as the node's first neighbours. As a result, the degree refers to the number of adjoining corners of the incidence (Shams and Khansari, 2014; Zhang, 2016, 2018; Haliki, 2021). Biologically, the degree allows a quick estimation of a node's regulatory importance. For example, in a protein sequence network, a very high degree protein sequence is linked with other sequences, indicating a crucial role in the origin and their evolution.

2.2.2 Closeness Centrality $(C_{clo}(v))$ (Freeman, 1978) of a node v is derived by calculating the optimal route between the node and the rest of the network's nodes, and then combining them. When this value is obtained, the reciprocal value is calculated to ensure that the higher values in terms of node closeness have a positive significance (Zhang, 2016, 2018; Haliki, 2021). Biologically, the closeness of nodes in the protein sequence network can be viewed as the vital node that can easily commute with the other nodes that share a close evolutionary relationship.

2.2.3 Betweenness Centrality $(Cbtw(v))$ (Freeman, 1978) measures how central an edge is. Given an edge e, the couples of nodes $(v1, v2)$ and the number of shortest paths between $v1$ and $v2$ and going through the edges e are considered while calculating. The values are then connected to the total number of shortest paths that connect $v1$ and $v2$. As a result, an edge can only be traversed by one path connecting $v1$ and $v2$, but if this path is the only one connecting $v1$ and $v2$, the edge will have a greater betweenness value (Shams and Khansari, 2014; Zhang, 2016, 2018; Haliki, 2021). Biologically, betweenness centrality is the process of determining which nodes in a protein sequence network control the flow of evolutionary information. More similar pairs of protein sequences (nodes) are connected to it.

2.2.4 Eigenvector Centrality (Freeman, 1978) is an index of node centrality. It assigns comparative results to all network nodes based on the notion that connections to high-scoring nodes contribute more to the node score than equal links to low-scoring nodes. The definition is recursive: a high Eigenvector value means that a node has a several neighbourhoods with high Eigenvector value. A high Eigenvector centrality indicates that the nodes are being visited while traversing and are well connected. Biologically, the eigenvector centrality of nodes in a protein sequence network contributes to the regulatory flow of evolutionary information to it's neighbouring sequence and neighbours of neighbour's and so on. In other words, the flow of information occurs between the closely evolved neighbour's that have similar or identical protein sequences.

2.3 Network of ND6 protein sequences

Characterization of protein sequences is done to preserve the genetic information of different species that share a considerable amount of information in their protein-coding sequence, resulting in significant homology. So there arises the question of sequence similarity/dissimilarity and the distances between the sequences. The protein sequences are the linear chain of amino acids, and the degeneracy of the amino acids plays a vital role in evolution. We have created a similar degree matrix and a network to analyze their evolutionary trends by introducing a mathematical parameter termed as *similar factor* by considering amino acids degeneracy, their locations in the quadrants and their angular values with respect to $x-$axis (Fig. 1).

In this section, we define the mathematical parameter *similar factor* for a protein sequence to create a similar degree matrix of eight ND6 protein sequence. Following several steps below, the degree of similarity between arbitrary protein sequences, say $P1$ and $P2$ is calculated.

I. Amino acids are arranged alphabetically in descending order according to their degeneracy number $6,4,3,2,1$ obtain from the standard genetic code table

$$L < R < S < A < G < P < V < T < I < C < D < E < F < H < K < N < Q < Y < M < W$$

II. Treating each amino acids above arrangement confined to first quadrant only, a two dimensional $(2D)$ vector representation all having equal positive $x-$ co-ordinate and assigned angles $3°, 6°, 9°, \ldots, 60°$ with the positive direction of $x-$axis (Fig. 1).

III. The vector representations of the amino acids based on step I and II are depicted in the Fig. 1 below.

IV. *Characteristic Vectors:* For an arbitrary protein sequence: *WTFESRNDPAKDPVILWLNGGPGCSSLTGL*, the corresponding set of characteristic vectors is $\{\vec{V}_W\ \vec{V}_T\ \vec{V}_F \ldots \vec{V}_L\}$

V. We define a matrix M for N different protein sequences denoted as $P1, P2, \ldots, PN$ whose elements m_{ij} are calculated as

$$m_{ij} = \sum_{i=1}^{n}\left(1 - \frac{|\ f(\vec{V}_k^{\,i}) - f(\vec{V}_k^{\,j})\ |}{90^\circ}\right) \quad (1)$$

$$d_{ij} = \log_{10}\left(m_{ij}\right) \quad (2)$$

where $\vec{V}_k^i$ and $\vec{V}_k^j$ represent the k^{th} characteristics vector of the i^{th} and k^{th} proteins sequences, m_{ij} is the *similar factor* between the i^{th} and k^{th} protein sequences, and the function $f(\vec{V}_k^i)$ represents the angles between the x-axis and the kth characteristics vector. Similar degree matrix associated with the similar factor whose elements are given by the equation 2.

VI. The eight ND6 protein sequences along with their accession number of Human (AP_000650), Gorilla (NP_008223), Common Chimpanzee (NP_008197), Harbor Seal (NP_006939), Gray Seal (NP_007080), Rat (AP_004903), Mouse (NP_904339) and Wallaroo (NP_007405) taken from https://www.ncbi.nlm.nih.gov/ were used for analyzing their evolutionary trends.

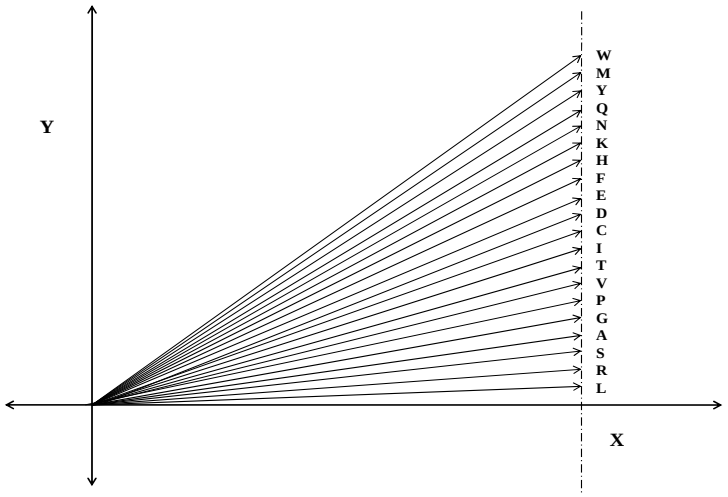


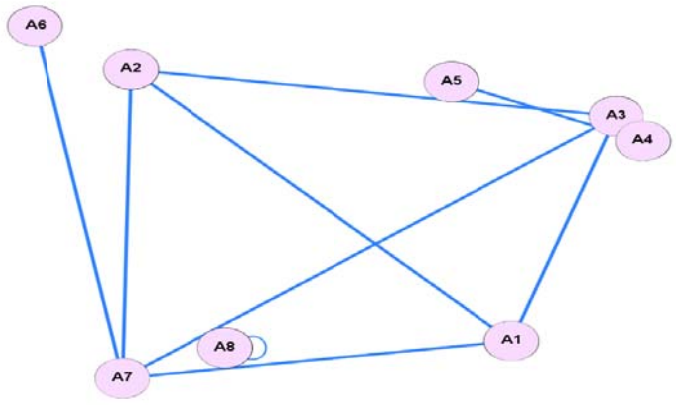
Fig. 1 Vector representation of twenty amino acids.

3 Results

Many methods (Maaty et al., 2010; Wen and Zhang, 2009; Hea et al., 2012) based on distances were proposed to study the evolutionary relationship of the protein sequences graphically and theoretically. Here we use a matrix termed as similar degree matrix because the *similar factor* between two protein sequences is calculated based on the angular value of the characteristics vector. By following the several steps as mentioned above, we have constructed a similar degree matrix in Table 1 below. The matrix consists of 36 data points and their average degree value is 2.17. By considering 2.17 as threshold value, we constructed a protein sequence network. The two protein sequences (nodes) are linked to each other if their degree values are greater than or equal to 2.17, and subsequently, we obtained the network P in the Fig. 2. The corresponding adjacency matrix of the network is described by matrix M below.

Table 1 Protein sequence degree matrix (A1= Human, A2= Gorilla, A3=Chimpanzee, A4=H seal, A5=G seal, A6=Rat, A7=Mouse, A8=Wallaroo).

Species	A1	A2	A3	A4	A5	A6	A7	A8
A1	2.24054	2.23720	2.23695	2.14134	2.14061	2.16613	2.17240	2.12428
A2	2.23720	2.24054	2.23594	2.14176	2.14124	2.16554	2.17026	2.12472
A3	2.23695	2.23594	2.24054	2.13998	2.13924	2.16682	2.17016	2.12199
A4	2.14134	2.14176	2.13998	2.24303	2.24146	2.12990	2.12990	2.16385
A5	2.14061	2.14124	2.13924	2.24146	2.24303	2.12893	2.12893	2.16375
A6	2.16613	2.16554	2.16682	2.12990	2.12893	2.23552	2.21642	2.12276
A7	2.17240	2.17026	2.17016	2.12990	2.12893	2.21642	2.23552	2.11859
A8	2.12428	2.12472	2.12199	2.16385	2.16375	2.12276	2.11859	2.22271



$$M = \begin{bmatrix} 0&1&1&0&0&0&1&0 \\ 1&0&1&0&0&0&1&0 \\ 1&1&0&0&0&0&1&0 \\ 0&0&0&0&1&0&0&0 \\ 0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&1&0 \\ 1&1&1&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0 \end{bmatrix}$$

Fig. 2 ND6 Protein sequence network.

From the network P, we can see that the species primates $(A1, A2\ A3)$ and rodents $(A6, A7)$ each are connected and are close to each other. Also, carnivorus $A4$ and $A5$ are connected separately to each other. Marsupial $(A8)$ has no connections with others being the remote species as they are not placental mammalian species. By the known fact of the evolution we can depict that closely evolves species are connected to each other.

4 Discussion

4.1 Centralities of the protein sequence network

In this section, we have calculated the different centrality measures in Table 2 for the protein sequence network and analyze the results in terms of the evolutionary importance of the node. The number of nodes in a network that are likely direct descendants or antecedent of a protein sequence, say $P1$, is determined by its degree of centrality. From the network in Fig. 2 we see that the node $A7$ has the highest degree centrality. So in the process of evolution the nodes $A1, A2, A3$ and $A6$ which are first neighbour's of $A7$ are probable immediate predecessor or successor i.e. the protein sequences of $A1, A2, A3\ and\ A6$ have close evolutionary relationship with $A7$.

Table 2 Centrality measure of protein sequence network

Vertex	Degree Centrality (C_d)	Closeness Centrality (C_{cl})	Betweenness Centrality (C_{bwt})	Eigenvector Centrality (C_λ)
A1	3	0.8	0	0.876543
A2	3	0.8	0	0.876543
A3	3	0.8	0	0.876543
A4	1	1	0	0.154321
A5	1	1	0	0.154321
A6	1	0.571429	0	0.320988
A7	4	1	3	1
A8	0	0	0	0

Table 3 Correlation coefficient of centrality measures

	C_d	C_{cl}	C_{bwt}	C_λ
C_d	1	0.1226	0.606977	0.919545
C_{cl}	0.1226	1	0.305126	0.331888
C_{bwt}	0.606977	0.305126	1	0.468935
C_λ	0.919545	0.331888	0.468935	1

Closeness centrality of the node in a network refers to the close distances to all other nodes through which flow of evolutionary information occurs between the nodes (Yang and Zhang, 2022). Higher the closeness centrality value of the node the more actively it can communicate with other nodes in the network. In our network, the closeness centrality of the node A7, A4, and A5 are higher than another node. So in the process of evolution more genetic information is easily communicable to the rest of the nodes through them or vice versa.

The centrality of a node's betweenness in a network gives us the contribution it has in transmission of the evolutionary information in the network (Xin and Zhang, 2021). Higher the betweenness centralities of the node more the similar pairs of nodes are connected to it through evolutionary process. In our network, $A7$ has the highest betweenness centrality value and highest degree value, so many shortest paths are linked through it connecting the similar protein sequences node in the network.

Eigenvector centrality in a network estimates the distributive impact of nodes. A high eigenvector centrality measure of a nodes indicates that the nodes itself is connected to its neighbours which themselves have high eigenvector centrality value. In our network, $A7$ has the highest eigenvector centrality followed by $A1, A2, A3$ and $A6$, indicating that information flow occurs between the closely evolved neighbour's that have the similar protein sequences.

4.2 Correlation analysis of the four centrality measures

In this section, we analyze the four correlation centrality measure that we have used in our protein sequence network. Assortative or dissasortative analysis of a network can be known through the use of the correlation coefficient(r). If the vertices with a higher degree of connectedness have a tendency to connect with other

vertices with a high degree of connectivity, the network is called assortative. When high-degree vertices have a tendency to connect with low-degree vertices, the network is said to be disassortative. The correlation coefficient of the four centrality measures is shown in Table 3 and all the values are in the range of -1 to 1. Further, we observe from Table 3 all correlation values between the different centrality measures are strongly correlated except degree with closeness centrality. All values of correlation coefficients are positive, so our network in Fig. 2 is of assortative type where the evolutionary information is transmissible more easily than the dis-assortative type network.

4.3 Network parameters

In biological networks, a variety of network parameters are used. We only looked at three network parameter in this paper: clustering coefficient, degree of distribution, and skewness (Zhang, 2018). The clustering coefficient is a measurement that indicates how likely a network is to be separated into groups. A cluster is a collection of vertices that contains a large number of edges linking them. The ratio between the total numbers of edges e_i of a node i actually linking its nearest neighbours to the total number of all conceivable links between these nearest neighbours is the clustering coefficient C_i. The total number of conceivable link is given by $k_i(k_i-1)/2$. Mathematically, $C_i = \frac{2e_i}{k_i(k_i-1)}$. A clustering coefficient has a value in the range of $0 \leq C_i \leq 1$, where $C_i = 0$ for the nodes i which have fewer than two neighbours and $C_i = 1$ for nodes i and its neighbours which are part of a group, or a completely connected group of nodes. A node with a greater clustering coefficient has a strong association with its neighbours, i.e. the greater the clustering coefficients of a node, the more connections there are among its neighbours. The clustering coefficients of all amino acids are listed in Table 4. It is obvious from this that protein sequences clustering coefficient is determined by their degree as well as by the number of direct links between two neighbouring protein sequences.

Table 4 Clustering coefficients of Protein sequence network

Nodes(i)	C_i
A1	1
A2	1
A3	1
A4	0
A5	0
A6	0
A7	0.5
A8	0

Table 5 Degree distribution of the protein sequence network

Nodes(i)	C_i
A1	0.375
A2	0.375
A3	0.375
A4	0.375
A5	0.375
A6	0.375
A7	0.125
A8	0.125

In our network, all the primates $(A1, A2, A3)$ has the highest equal value 1, $A7$ has 0.5 and the rest have zero values. Again we found that the clustering coefficient of the whole network is 0.4375 which is almost nearer to node $A7$. The clustering coefficient increases with the number of connections between neighbours. Higher clustering coefficients in our network have a big effect on the network's nodes and delay the spread of evolutionary information near the neighbourhood of similar protein sequences, signifying that their betweenness centrality value are zero.

Next, we investigate the degree of distribution of each node in our network. The spread in a number of links or edges a node has to other nodes determines its degree in a network. The ratio of nodes in a network of

degree k is defined as the network's degree distribution $P(k)$. If a network has n nodes in total and n_k of them have degree k, we have $P(k) = n_k/n$. In general; a node's degree distribution value indicates the probability that a specific node would have exactly k relationships. In Table 5 above we have shown the degree of distribution of protein sequences as nodes in our network.

Finally, we investigate the third network parameter for our network known as *skewness*. Karl Pearson proposed the concept of quantifying skewness in 1895. Skewness is a method to estimate whether a variable's distribution is symmetric or asymmetric. If on either side of a curve, the variables are equidistant from the centre value, which is referred to as symmetry. Asymmetric refers to skewed data which is either positively or negatively skewed. Skewness is represented by the letter S_k. There are two forms of skewness in the distribution: positive skewness and negative skewness, which are determined by the values and relative positions of the mode, mean, and median. A positive skewed distribution is one in which the mean is the highest, the mode is the lowest, and the median is in the middle. Negative skewed distribution occurs when the mode is the highest, the mean is the lowest, and the median is in the middle. Throughout this study, we used Karl Pearson's skewness coefficient, which is calculated using the formula

$$S_k = \frac{3(Mean-Median)}{Standard\ Deviation}$$

The skewness measure has a value in the range of -3 to $+3$. $S_k = 0$ the distribution is symmetrical, or normal. $S_k > 0$, the distribution is positively skewed. $S_k < 0$, the distribution is skewed negatively. From Table 5, their Pearson's coefficient of skewness is found to be -1.62. The negative value led us to conclude that the degree of distribution of the protein sequences are skewed negatively.

5 Conclusion

We attempted to study in this paper the evolutionary aspect of few protein sequences by constructing a network and analyzing their similarity/dissimilarity with the known sequences. Similarity analysis helps us in saving time and effort in re-determining the function, structure and relationship of the new sequence. The numerical characterization of the protein sequences helps in the construction of the ND6 protein's network, providing us a simple and intuitive way for analyzing and sorting the sequences. By observing the various centrality measures of the network, the protein sequence of the mouse has the highest centrality measure viz. degree, closeness, betweenness and eigenvector centrality. So we can conclude that the protein sequences of the mouse and their neighbouring linked sequences in the network revealed that they have some close evolutionary relationship.

We have also studied the correlation coefficient of the various centrality measures in our network and found that degree and closeness centrality is not strongly correlated. As a result, we may conclude that degree centrality is independent of the closeness centrality measures in the analysis of ND6 protein networks based on similar degree matrix and must be explored individually.

By examining the clustering value of protein sequences, we can see $A1$, $A2$ and $A3$ have a high clustering coefficient. As a result, in comparison to rest of the node of the network, the flow of evolutionary information to other node through the neighbourhood of $A1$, $A2$ and $A3$ is comparatively slow. This network is assortative type so evolutionary information transfer is easy. Finally, we found that the degree of distribution is negatively skewed. In further work, we would like to investigate the evolutionary aspects of more protein sequences, corona virus sequences based on different physico- chemical, classification of amino acids.

References

Abo-Elkhier MM, Abd Elwahaab MA, Abo El Maaty MI. 2019. Measuring similarity among protein sequences using a new descriptor. BioMed Research International , 2019: 2796971

Ali T, Akhtar A, Gohain N. 2016. Analysis of amino acids network based on distance matrix. Physica A, 452: 69-78

Bai F, Wang T. 2006. On graphical and numerical representation of protein sequences. Journal of Biomolecular Structure and Dynamics, 23(5): 537-546

Bajusz D, Miranda-Quintana RA, Rácz A, Héberger K. 2021. Extended many-item similarity indices for sets of nucleotide and protein sequences. Computational and Structural Biotechnology Journal, 19: 3628-3639

Bertman MO, Jungck JR. 1979. Group graph of the genetic code. Journal of Heredity, 70(6): 379-384

El-Lakkani A, El-Sherif S. 2013. Similarity analysis of protein sequences based on 2D and 3D amino acid adjacency matrices. Chemical Physics Letters, 590: 192-195

Freeman LC. 1978. Centrality in social networks conceptual clarification. Social Networks, 1(3): 215-239

Gates M. 1986. A simple way to look at DNA. Journal of Theoretical Biology, 119: 319-328

Gupta MK, Niyogi R, Misra M. 2014. A 2D Graphical representation of protein sequence and their similarity analysis with probabilistic method. MATCH Communications in Mathematical and in Computer Chemistry, 72: 519-532

Haliki E. 2021. Centralities of galaxies in the weighted network model of the local group. Selforganizology, 8(3-4): 7-20

Hamori E. 1985. Novel DNA sequence representations. Nature, 314: 585-586

Hamori E, Ruskin J. 1983. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. Journal of Biological Chemistry, 258(2): 318-327

Hea P, Weia J, Yaob Y, Tie Z. 2012. A novel graphical representation of proteins and its application. Physica A, 391: 93-99

Leong PM, Morgenthaler S. 1995. Random walk and gap plots of DNA sequences. Bioinformatics, 11(5): 503-507

Maaty MI, Abo-Elkhier MM, Abd Elwahaab MA. 2010. 3D graphical representation of protein sequences and their statistical characterization. Physica A: Statistical Mechanics and Its Applications, 389(21): 4668-4676

Pearson WR. 2013. An introduction to sequence similarity ("homology") searching. Current Protocols in Bioinformatics, 0471250953.bi0301s42

Shams B, Khansari M. 2014. Using network properties to evaluate targeted immunization algorithms. Network Biology, 4(3): 74-94

Wen J, Zhang Y. 2009. A 2D graphical representation of protein sequence and its numerical characterization. Chemical Physics Letters, 476: 281-286

Xie LZ, Yu Y, Liang L, Guo M, Song J, Yuan Z. 2012. Protein sequence analysis based on hydropathy profile of amino acids. Journal of Zhejiang University Science B, 13(2): 152-158

Xin SH, Zhang WJ. 2021. Construction and analysis of the protein-protein interaction network for the detoxification enzymes of the silkworm, *Bombyx mori*. Archives of Insect Biochemistry and Physiology, 108(4): e21850

Yang S, Zhang WJ. 2022. Systematic analysis of olfactory protein-protein interactions network of fruitfly, *Drosophila melanogaster*. Archives of Insect Biochemistry and Physiology, 110(2): e21882

Yao YH, Dai Q, Li L, Nan XY, He PA, Zhang YZ. 2009. Similarity/Dissimilarity studies of protein sequences based on a new 2d graphical representation. Journal of Computational Chemistry, 31(5): 1045-1052

Zhang WJ. 2016. Selforganizology: The Science of Self-Organization. World Scientific, Singapore

Zhang WJ. 2018. Fundamentals of Network Biology. World Scientific Europe, London, UK