Article

Analysis of amino acids network based on transition and transversion mutation of codons

Tazid Ali, **Chandra Borah**
Department of Mathematics, Dibrugarh University, Assam 786004, India
E-mail: tazid@dibru.ac.in, chandra92borah@gmail.com

Received 12 March 2021; Accepted 29 March 2021; Published 1 September 2021

Abstract
In this paper, we have developed a network of 20 amino acids based on a distance matrix of amino acids. This distance matrix is obtained by considering the transition and transversion mutation of codons. We have proposed that the evolutionary pattern of amino acids is reflected throughout this network. We have discussed different measures of centrality: degree centrality, closeness centrality, betweenness centrality and eigenvector centrality, concerning this network and investigated the comparative impact of the amino acids. We have also explored the correlation coefficients between the different centrality measures checking the assortativity of the network. Further, we have explored three network parameters: namely clustering coefficient, degree of distribution and skewness.

Keywords amino acids; genetic code; centrality measure; correlation coefficient; network parameter; distance matrix.

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

All life forms in existence are composed of cells. In each cell, there is a set of chromosomes that serves as a blueprint for the whole life entity. A chromosome consists of genes (sequence of DNA), where a gene encodes a specific protein. A protein consists of a linear sequence of amino acids, the essential building blocks and functional components of living organisms. Twenty different amino acids have been found to date that exists in proteins. The three sequencing bases is a unit called a codon that specifies an amino acid. Since there are four bases, that give us a total of 64 codons. So, there must be some similarity, i.e., more than one codon code for the same amino acid. Codons that code for the same amino acids are classified as synonymous codons. This can be observed as a mapping of many to one taking codons to amino acids. Also, out of these 64 codons, UAA, UAG and UGA triplets are known as stop codons, and their task is to terminate the translation process.

The flow of information from DNA to protein is carried out via transcription and translation (Shu, 2017). As a consequence of mutation, the sequencing bases are not duplicated exactly in replicating the DNA strand.

This influences the formation of proteins. Deletion, insertion, inversion, point and frame shift are different types of mutation in genetics. The point mutation is a replacement of one base of its genetic sequence. Transition, in genetics and molecular biology, corresponds to a point mutation that shifts the purine {A and G} to purine or pyrimidine {C and U} to a pyrimidine. The point mutation that switches purine to pyrimidine or vice-versa is referred to as transversion.

A wide range of contributions have been made by various researchers to the field of the biological networks (Bagler and Sinha, 2007; Khansari et al., 2016; Zhang, 2016). Kundu (2005) explored that the hydrophobic and hydrophilic networks fulfill the "small-world properties" of proteins. In these networks, amino acids are taken as vertices and any two amino acids have a link within a distance of 5A. He also noticed that the hydrophobic network has a greater average degree of nodes than the hydrophilic one. Aftabuddin and Kundu (2007) explained three kinds of protein networks (hydrophobic, hydrophilic and charged). They have shown that the average degree of a hydrophobic network is significantly higher than that of the other two networks. The hydrophilic network's average degree is slightly higher than that of the charged network. All of the three types of networks reflect "small-world" property. Based on contact energy, Jiao et al. (2007) explored the weighted amino acids network and shown that the weighted amino acids network satisfies the small-world property. Akhtar and Ali (2014) observed a network of amino acids dependent on codon mutations. Their analysis reveal that amino acids Arginine (high hydrophilic) and Serine (low hydrophilic) have the largest centrality values regardless of centrality measurements. Wuchty and Stadler (2003) explored multiple centrality measures for the biological network. They concluded that only the degree of vertex centrality is not enough to distinguish between lethal protein and viable protein. Newman (2002) explored assortative mixing characteristics in network of protein associations, neural networks and food networks. He has also observed that the information can be transmitted efficiently across an assortative network in contrast to a disassortative network. Koschutzki and Schreiber (2004) analyzed the centralities for the biological networks, namely the transcriptional network and the PPI network. Their research suggested that different centrality measures should be considered in the study of biological networks. Ali and Akhtar (2016) constructed a network of amino acids, depicting the evolutionary pattern of the amino acids. They have discussed different centrality measures for that network and noted that the hydrophobic amino acid Tyrosine (Y) has the highest centrality values considering the centrality measures such as degree centrality, closeness centrality, betweenness centrality and eigenvector centrality. They have also studied the correlation coefficients among various centrality measures. Zhang (2016) screened node attributes that significantly influence node centrality in the network. Zhang and Zhang (2019) constructed the PPI network and made centrality analysis on insecticide resistance molecular mechanism in *Drosophila melanogaster*. Xin and Zhang (2020) constructed the PPI network for the olfactory system of the silkworm *Bombyx mori* and made centrality analysis.

In this paper, we are interested in the analysis of amino acids network based on codon mutation. For that, we have attempted to explore certain graph theoretic notions in the network of the amino acids.

The paper is structured as follows. In section 2, we include some introductory principles of the graph theory in which we work and briefly examine the various centrality measures. In section 3, we describe the graph of amino acids based on transition and transversion mutation of bases of codons. We obtain a network of 20 amino acids, where we compare various centrality measures. In section 4, we are interested in the study of some of the important network parameters. We have the conclusion of the paper in section 5.

2 Preliminary Graph Concepts

An undirected graph $G = (V, E)$ is a finite set V of nodes or vertices and a set E of edges or sides where $E \subseteq V \times V$ (Bertman and Jungck, 1979; Zhang, 2018). For any edge $e = (u_1, u_2)$, the vertices u_1 and u_2 are

said to be incident on the edge e and adjacent to one another. The neighbourhood of a vertex u, denoted by $N(u)$, is the set of all vertices adjacent to u. In the case of a directed graph $G = (V, E)$, any edge $e \in E$ has a direction.

The adjacency matrix $\boldsymbol{A}$ of a graph $G = (V, E)$, with vertex set $V = \{u_1, u_2, \cdots, u_n\}$ is an $(n \times n)$ matrix, where $a_{ij} = 1$ if and only if there is an edge from vertex u_i to vertex u_j and $a_{ij} = 0$ otherwise. An undirected simple graph's adjacency matrix is symmetric. For any graph $G = (V, E)$, the degree of a vertex v, denoted by $d(v)$ or $deg(v)$ is the number of edges incident to v. A graph G is connected if there is a path in G between any given pair of vertices, otherwise it is disconnected.

For any graph $G = (V, E)$, a walk is a finite alternating sequence of vertices and edges, starting and ending with vertices. A walk of length n is a non-empty alternating sequence $u_0 e_0 u_1 e_1 \cdots e_{n-1} u_n$ of vertices and edges in G such that $e_i = \{u_i, u_{i+1}\}$ for all $i < n$. If $u_0 = u_n$, then the walk will be closed. A path is a walk where there are no repeated vertices. A path with the minimum length between two vertices u and v is the shortest or geodesic path between the vertices. A connected graph has a walk between every pair of vertices.

2.1 Centrality in graph

In graph theory, the centrality measure indicates the relative significance of a vertex (Zhang, 2018). A centrality is defined as a real-valued function on the vertices of a graph. More formally, the centrality is a function f which assigns a real value $f(v)$ to each vertex v of the given graph G. The four most widely used centrality measures, namely degree centrality, closeness centrality, betweenness centrality and eigenvector centrality are discussed in the following sections.

2.1.1 Degree centrality

The degree centrality is the simplest measure of centrality. For any vertex u, it is defined as the number of vertices to which the vertex u is directly linked (Freeman, 1978; Zhang, 2018) and denoted by $C_d(u)$. Degree centrality indicates that a large number of interactions are involved in an important vertex. It is mathematically defined as

$$C_d(u) = \deg(u)$$

In real-world applications, the degree centrality is not a realistic measure to assess the value or importance of a node. In real scenario, a significant node may be linked implicitly to many other nodes.

2.1.2 Closeness centrality

Closeness centrality measures how connected a node is to the rest of the nodes in the network on a global scale (Freeman, 1978). If a node is close to other nodes, it can communicate easily with all other nodes. The centrality of closeness is calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. It is mathematically defined as

$$C_c(u) = \frac{(n-1)}{\sum_{v \epsilon V} d(u, v)}$$

where n is the number of vertices or nodes of the network and $d(u, v)$ gives the shortest path distance between the pair of nodes u and v. It is evident from the above definition that if the node has the lowest cumulative shortest path distance, the node has the highest centrality of closeness. The maximum closeness centrality node is quite well associated with all other nodes.

2.1.3 Betweenness centrality

In network theory, betweenness centrality measures the magnitude to which a vertex lies on the paths between the other vertices (Freeman, 1978). Vertices with a high betweenness can have significant impact within the

network as a result of their control over information passing between others. The betweenness centrality of a vertex v is the number of shortest paths that pass through v (Watts and Strogatz, 1998). It is mathematically defined as

$$C_b(v) = \sum_{s \neq v \epsilon V} \sum_{t \neq v \epsilon V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} count the number of shortest paths containing s and t as their end vertices, while $\sigma_{st}(v)$ is the number of those shortest paths from s to t that pass through v.

Betweenness centrality reflects the recognition of vertices that make the most of the network's information flow. A significant vertex will lie on a high percentage of paths between most of the other vertices on a network. We can monitor the information on the network from this vertex. A high degree vertex has a high betweenness centrality since plenty of the shortest paths will run across them. However, a high betweenness centrality vertex may not always be a high degree vertex.

2.1.4 Eigenvector centrality

The eigenvector centrality is another great important measure of centrality (Bonacich, 1972). In graph theory, it is way of measuring the dominance of a node in a network. For any square matrix $\boldsymbol{A}$, λ is an eigenvalue if $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$, and $\boldsymbol{I}$ is the identity matrix of the same order as $\boldsymbol{A}$. Eigenvector centrality is identified as the principal eigenvector of the corresponding graph's adjacency matrix. We can write the eigenvector equation as

$$\boldsymbol{A}X = \lambda X$$

where $\boldsymbol{A}$ is the adjacency matrix for the graph, λ is the eigenvalue (constant term) and X is the corresponding eigenvector. The eigenvector of the greatest eigenvalue is the eigenvector centrality (Bonacich, 1972).

2.2 Network parameters

Various network parameters are used in biological networks. We have discussed three basic network parameters: clustering coefficient, degree of distribution and skewness.

2.2.1 Clustering coefficient

Suppose there is a node v of degree k in the undirected graph G, and there is e number of edges between the k neighbours of v in G. Then the clustering coefficient of v in G is defined as

$$C_v = \frac{2e}{k(k-1)}$$

So, C_v calculates the ratio between the edge numbers among the neighbours of v and the total potential edge numbers: $k(k-1)/2$, where $0 \leq C_v \leq 1$.

2.2.2 Degree distribution

The degree of a vertex for an undirected graph is the number of links or edges the vertex has to the other vertices (Zhang, 2018). Then the degree distribution, $P(k), k = 0,1,\cdots$, which calculates the ratio of vertices in the network having degree k. Mathematically,

$$P_k = \frac{n_k}{n}$$

where n is the size of the network and n_k is the total number of vertices of degree$k$ in the network.

2.2.3 Skewness

In 1895, Karl Pearson first suggested the measuring of skewness. The situation of skewness, which implies the absence of symmetry, exists in a curve when the mean, median, and curve mode are not the same. Depending on the vertices and relative location of the mode, mean and median, two forms of skewness emerge in the distribution, respectively positive skewness and negative skewness. In our analysis, we consider Karl Pearson's coefficient of skewness, which is denoted by S_k and defined by the following formula

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

The value of the skewness is inside the range of -3 to $+3$.

3 Graph of Amino Acids

The four bases Adenine (A), Guanine (G), Cytosine (C) and Uracil (U) can be classified into two groups: purine {A, G} and pyrimidine {C, U}. A purine base has a double carbon-nitrogen ring, while pyrimidine has one such carbon-nitrogen ring. We know that the transversion mutation of codons causes extreme physicochemical properties change to the amino acids in comparison to the transition mutation of codons. Firstly, we obtain a distance matrix of amino acids based on the transition and transversion mutation of codons at various base positions.

3.1 Distance between amino acids based on mutation

A codon consists of three bases, and we define the distance between codons as follows. For any two codons, transition mutation in any of the base position is assigned a score of 1, whereas for transversion mutation, we give a score of 2.

For example, to find the difference between the codons ACG and GCC, we have the score 1 for the first base position (A and G), score 0 for the second base position (C and C), and a score of 2 for the third base position (G and C). So, the distance between the codons ACG and GCC is $1 + 0 + 2 = 3$.

Now we can find the distance between amino acids by calculating the mean distance between their respective codons.

Table 1 Distance between the amino acids Proline (P) and Tyrosine (Y).

	CCA	**CCC**	**CCG**	**CCU**
UAC	5	3	5	4
UAU	5	4	5	3

For example, we compute the distance between the amino acids Proline (P) and Tyrosine (Y) is 4.25. The codons that code Proline are CCA, CCC, CCG, CCU and the codons that code Tyrosine are UAC, UAU. In Table 1, we measure the distances between the codons. The distance between the amino acids P and Y is obtained by considering the mean distance between the above codons.

In Table 2, we obtain the distance between each pair of amino acids.

The distance matrix we obtain above is symmetric, and it has 210 data points. We have constructed a network of amino acids from the distance matrix above. In consideration of the 210 data points, we have found that the mean is 3.496. Since "mean" implies that the data points tend to cluster around it, this "mean" value

(i.e., 3.496) is assumed to be the threshold value for determining the relationship between pairs of amino acids. So, we define two amino acids as comparable, i.e., connected by an edge if their distance is less or equal to 3.496. The subsequent graph G is represented in Fig. 1.

Table 2 Distance matrix of amino acids based on transition and transversion mutation.

	R	**K**	**E**	**Q**	**D**	**N**	**H**	**P**	**Y**	**S**	**T**	**G**	**W**	**A**	**M**	**C**	**F**	**L**	**V**	**I**
R	0.00	3.33	3.67	2.67	4.17	3.83	3.17	3.92	4.00	4.00	4.50	2.92	2.33	4.92	4.33	2.83	4.83	4.05	5.08	4.72
K	3.33	0.00	1.50	2.50	2.00	2.00	4.00	5.25	4.00	3.83	3.25	4.25	3.50	4.25	2.50	5.00	6.00	5.00	4.25	3.50
E	3.67	1.50	0.00	2.50	2.00	3.00	4.00	5.25	4.00	4.83	4.25	3.25	3.50	3.25	3.50	5.00	6.00	5.00	2.25	4.50
Q	2.67	2.50	2.50	0.00	4.00	4.00	2.00	3.25	4.00	4.50	5.25	5.25	2.50	5.25	4.50	4.00	5.00	3.33	5.25	5.50
D	4.17	2.00	2.00	4.00	0.00	1.50	2.50	5.25	2.50	4.33	4.25	2.50	5.00	3.25	5.00	3.50	4.50	5.50	3.25	4.00
N	3.83	2.00	3.00	4.00	1.50	0.00	2.50	5.25	2.50	4.00	3.25	3.25	5.00	4.25	4.00	3.50	4.50	5.50	4.25	3.00
H	3.17	4.00	4.00	2.00	2.50	2.50	0.00	3.25	1.50	4.00	5.25	4.25	4.00	5.25	6.00	2.50	3.50	3.83	5.25	5.00
P	3.92	5.25	5.25	3.25	5.25	5.25	3.25	0.00	4.25	3.50	3.25	5.25	4.25	3.25	4.25	4.25	3.25	2.58	4.25	4.25
Y	4.00	4.00	4.00	4.00	2.50	2.50	1.50	4.25	0.00	3.33	5.25	4.25	3.00	5.25	6.00	1.50	2.50	3.83	5.25	5.00
S	4.00	3.83	4.83	4.50	4.33	4.00	4.00	3.50	3.33	0.00	3.33	4.25	3.50	3.58	4.17	3.00	3.00	3.78	4.25	3.83
T	4.50	3.25	4.25	5.25	4.25	3.25	5.25	3.25	5.25	3.33	0.00	4.25	5.25	2.25	2.25	5.25	4.25	4.25	3.25	2.25
G	2.92	4.25	3.25	5.25	2.50	3.25	4.25	5.25	4.25	4,25	4.25	0.00	3.25	3.25	4.25	3.25	5.25	5.25	3.25	4.25
W	2.33	3.50	3.50	2.50	5.00	5.00	4.00	4.25	3.00	3.50	5.25	3.25	0.00	5.25	4.00	2.00	4.00	3.67	5.25	5.67
A	4.92	4.25	3.25	5.25	3.25	4.25	5.25	3.25	5.25	3.58	2.25	3.25	5.25	0.00	3.25	5.25	4.25	4.25	2.25	3.25
M	4.33	2.50	3.50	4.50	5.00	4.00	6.00	4.25	6.00	4.17	2.25	4.25	4.00	3.25	0.00	6.00	2.50	3.00	2.25	1.67
C	2.83	5.00	5.00	4.00	3.50	3.50	2.50	4.25	1.50	3.00	5.25	3.25	2.00	5.25	6.00	0.00	2.50	4.17	5.25	5.00
F	4.83	6.00	6.00	5.00	4.50	4.50	3.50	3.25	2.50	3.00	4.25	5.25	4.00	4.25	2.50	2.50	0.00	2.17	3.25	3.00
L	4.05	5.00	5.00	3.33	5.50	5.50	3.83	2.58	3.83	3.78	4.25	5.25	3.67	4.25	3.00	4.17	2.17	0.00	3.25	3.33
V	5.08	4.25	2.25	5.25	3.25	4.25	5.25	4.25	5.25	4.25	3.25	3.25	5.25	2.25	2.25	5.25	3.25	3.25	0.00	2.25
I	4.72	3.50	4.50	5.50	4.00	3.00	5.00	4.25	5.00	3.83	2.25	4.25	5.67	3.25	1.67	5.00	3,00	3.33	2.25	0.00

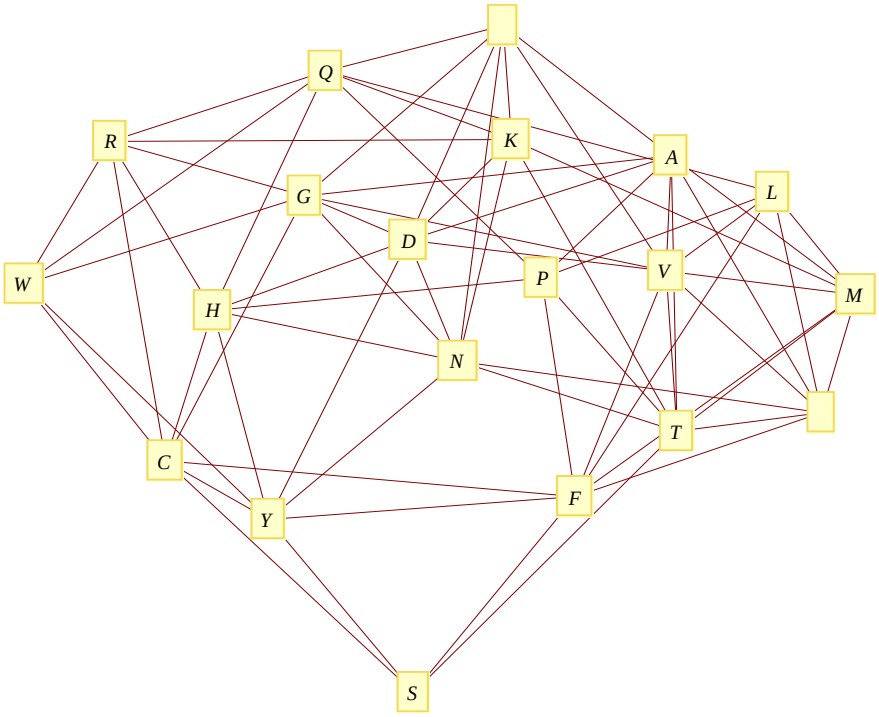

Fig. 1 Graph of amino acids (G) (based on transition and transversion mutation).

Since the distance in the matrix is dependent on the distinctions between the respective codons of the two amino acids, it can be assumed that two amino acids are compatible if they are bound by an edge. If two amino acids are bound by an edge, there is a high probability that one amino acid will evolve from the other. The evolution of the amino acid from the other is regulated by the mutation of the related codons. So, we assume that the graph shows the evolutionary trend of amino acids. We obtain the corresponding adjacency matrix of amino acids for the graph in the following

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

Here, $\boldsymbol{M} = \boldsymbol{M}^{\boldsymbol{T}}$. Observing $\boldsymbol{M}$, we can say that graph G is connected as no row or column of the lower (or upper) triangular matrix is zero.

3.2 Centralities in amino acids graph

Here, we have computed different centrality measures to analyze the amino acids graph G (Fig. 1) and displayed all values in Table 3. Table 3 shows that the degree centrality, the eigenvector centrality, and the closeness centrality give the amino acid Valine (V) the highest rank, while the betweenness centrality assigns the highest rank to the amino acid Phenylalanine (F).

Table 3 Different centrality measures for the 20 amino acids.

Vertex	Degree Centrality	Closeness Centrality	Betweenness Centrality	Eigenvector Centrality
R	6	0.576	3.564	0.164
K	7	0.613	6.206	0.225
E	7	0.594	3.283	0.246
Q	7	0.594	9.563	0.187
D	8	0.633	6.724	0.271
N	8	0.633	7.861	0.261
H	7	0.594	6.292	0.201
P	6	0.594	5.294	0.185
Y	7	0.613	7.111	0.194
S	4	0.528	1.900	0.119
T	8	0.613	9.039	0.256
G	8	0.633	11.649	0.255
W	5	0.528	1.667	0.136
A	8	0.633	5.546	0.276
M	7	0.576	3.492	0.238
C	7	0.613	7.267	0.180
F	8	0.633	12.553	0.229
L	6	0.594	3.928	0.191
V	9	0.655	9.318	0.304
I	7	0.576	3.305	0.242

For any amino acid X (say), the degree centrality is determined by the number of first neighbours of X. For example, the amino acid Valine (V) has the degree centrality 9. Accordingly, V is like to be the immediate antecedent or follower of nine acids in the evolutionary process, i.e., D, G, E, L, F, T, A, I, M.

If the closeness centrality of an amino acid is high, it can interact easily with all other amino acids. Thus, higher value of the closeness centrality for an amino acid indicates that the evolutionary mechanism is readily shared with other amino acids. The amino acids V and D have the closeness centrality value of 0.633 and 0.593, respectively. So, we can assume that the evolutionary mechanism is better mediated by V than through D, i.e., more amino acids precede or succeed V than D in the process of evolution.

The betweenness centrality of the amino acid is an estimate of the contribution it has made in the course of expressing the evolutionary process. Higher value of the betweenness centrality for an amino acid represents the identification of amino acids that render much of the network's information flow. For example, the

betweenness centralities for the amino acids F and G are 12.552 and 11.649, while for the amino acids S and W are 1.900 and 1.667. Thus, more pairs of amino acids are related through the evolutionary mechanism by F and G than through S and W, i.e., amino acids F and G appear as an intermediate between more pairs of amino acids relative to S and W.

Eigenvector centrality is more striking and well-preserved than the degree of centrality in a network. It's large for a node if it either has many neighbours and/or important neighbours. The top 4 amino acids with the highest eigenvector centrality are D, N, A and V, as the sum of the direct and indirect bonds of amino acids D, N, A and V is maximum. Consequently, in the evolutionary process, the contribution of neighbours and neighbours, and so on, to these amino acids is higher (Chakrabarty and Parekh, 2014). So, we can say that the top 4 amino acids with highest eigenvector centrality D, N, A and V play a crucial role in the evolutionary process. The amino acids D and N are hydrophilic and one can be obtained from another by a 1st base transition mutation, while A and V are hydrophobic and one can be obtained from another by 2nd base transition.

3.3 Bivariate correlation between various centralities

Here, we have discussed the bivariate correlation of different measures of centralities for the amino acids network. Correlation is the most significant element in the study of assortative or disassortative networks. A network is called assortative if the nodes with higher degree appear to communicate with other nodes that also have a high degree of connectivity. In a disassortative network, the nodes with higher degree appear to communicate with other nodes with low degree connectivity (Newman, 2002). In Table 4, we obtain the correlation coefficients for all the centrality measures.

Table 4 Correlation coefficients for the different centrality measures.

	C_d	C_c	C_b	C_λ
C_d	1	0.916	0.733	0.796
C_c	0.916	1	0.775	0.702
C_b	0.733	0.775	1	0.418
C_λ	0.796	0.702	0.418	1

Here, all the correlation coefficients (r) are computed using Pearson's method. The value of r ranges from $+1$ to -1. In the case of an assortative network, we have $r > 0$, and for a disassortative network, we have $r < 0$. From Table 4, we note that all the centrality measures are strongly correlated with each other except betweenness centrality with eigenvector centrality. We observe that correlation coefficient is positive for each pair of centrality measures, and so our network G (in Fig. 1) is assortative. Consequently, the evolutionary information transmits efficiently through this network.

4 Network Parameters

We use different network parameters to analyze biological networks. In the following sections, we tackle a few of them to interpret network's communication pattern.

4.1 Clustering coefficients of amino acids

The clustering coefficient is a metric that indicates the tendency of a graph to be split into clusters. A cluster is a group of nodes that involved several links connecting these nodes. The high clustering coefficient of a node represents a close association between adjacent nodes. The clustering coefficient of a node seems to have an

impact on the neighbouring node of that node and hence stabilizes the flow of information (Sengupta and Kundu, 2012).

Table 5 displays the clustering coefficients of all amino acids for the network G.

Table 5 Clustering coefficients of the amino acids

R	K	E	Q	D	N	H	P	Y	S	T	G	W	A	M	C	F	L	V	I
0.467	0.333	0.524	0.286	0.464	0.357	0.333	0.267	0.381	0.500	0.357	0.393	0.500	0.464	0.571	0.428	0.250	0.533	0.472	0.571

The clustering coefficient of the amino acid depends on the degree of amino acid as well as the number of direct interactions between the neighbouring amino acids. For the network G, we observe that large hydrophobic amino acids I and M have a high clustering coefficient value of 0.571. The whole network has a clustering coefficient value of 0.422, approximately the same as the hydrophobic amino acid C (Cysteine). The clustering coefficient is getting higher with the higher number of links between neighbours. So, the higher clustering coefficient values of the network slow down the flow of evolutionary messages. From the clustering coefficient of the whole network and the clustering coefficients of the amino acids, we can say that the evolutionary mechanism is comparatively slow in the vicinity of M and I in comparison to the whole network.

4.2 Degree of distribution

In this section, we compute the degree of distribution of the nodes (amino acids) for the network G. The degree of a node in a network is the number of links that the node has to the other nodes. If there are n number of nodes in a network and n_k of them have degree k, we have the degree distribution $P(k) = n_k/k$. The degree distribution value of a node describes the probability that the selected node will have exactly k connections. In Table 6, we have shown the degree of distribution values of different amino acids.

Table 6 Degree distribution of amino acids.

R	K	E	Q	D	N	H	P	Y	S	T	G	W	A	M	C	F	L	V	I
0.15	0.40	0.40	0.40	0.30	0.30	0.40	0.15	0.40	0.05	0.30	0.30	0.05	0.30	0.40	0.40	0.30	0.15	0.05	0.40

4.3 Skewness

As described above, skewness is a measure of the asymmetry of a variable distribution (Zhang, 2018). The Karl Pearson's coefficient of skewness (S_k) is given by the following formula

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

If $S_k = 0$, then the distribution is symmetrical. If $S_k > 0$, then the distribution is positively skewed, and if $S_k < 0$, then it is negatively skewed.

Here, we assume the degree of distribution as a variable (X) and the number of amino acids carrying the same distribution as frequency (f). We have Table 7 (using Table 6), where we compute Karl Pearson's coefficient of skewness.

Table 7 Calculation of Karl Pearson's coefficient of skewness.

X	f	Xf	$d_x = X - \bar{x}$	fd_x	fd_x^2
0.05	3	0.15	-0.23	-0.0115	0.1587
0.15	3	0.45	-0.13	-0.0195	0.0507
0.30	6	1.80	0.02	0.0060	0.0024
0.40	8	3.20	0.12	0.0480	0.1152

From Table 7, we have mean is $\bar{x} = 0.28$ and that median is 0.30. Also, the standard deviation is 0.1279. So, Pearson's coefficient of skewness is $-0.469 < 0$. We see that, the coefficient of skewness is negative. Accordingly, the degrees of distribution of the amino acids are negatively skewed.

5 Conclusion

In this paper, we have explored the evolutionary mechanism of amino acids based on codon mutation. For this reason, we have considered the fact that the transversion mutation of codons induces extreme physicochemical property variations in amino acids as compared to the transition mutation of codons. We have constructed a graph structure of 20 amino acids that specifies the compatibility relationship based on the amino acids distance matrix. Different centrality measures have been discussed, and we observe that the hydrophobic amino acid Valine (V) has the highest degree of centrality, closeness centrality and eigenvector centrality value. Phenylalanine (F) has the highest betweenness centrality value, which suggests it has the highest contribution in communicating the evolutionary process. So, more pairs of amino acids are connected through F than through the rest of the amino acids.

Next, we have obtained correlation coefficients for the various centrality measures of amino acids and noticed that all centrality measures are closely correlated except betweenness centrality with eigenvector centrality. Also, the correlation coefficient is positive for each pair of centrality measures which indicates that our network is an assortative one. Consequently, the evolutionary data flow will be smooth.

We have also observed that large hydrophobic amino acids M and I have high clustering coefficient values. So, the rate of the evolutionary process is comparatively slow in the vicinity of M and I. Lastly, we have observed that the degree of distribution of the 20 amino acids is negatively skewed.

Acknowledgement

Chandra Borah, the second author of this research work, a recipient of JRF, would like to express my sincere gratitude towards the University Grants Commission (India) for the financial assistance in doing my research.

References

Aftabuddin M, Kundu S. 2007. Hydrophobic, hydrophilic and charged amino acids networks within protein. Biophysical Journal, 93(1): 225-231

Akhtar A, Ali T. 2014. Analysis of unweighted amino acids network. International Scholar Research Notices, Article ID: 350276

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

Bagler G, Sinha S. 2007. Assortative mixing in protein contact networks and protein folding kinetics.

Bioinformatics, 23(14): 1760-1767

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

Bonacich P. 1972. Factoring and weighted approaches to status and clique identification. The Journal of Mathematical Sociology, 2(1): 113-120

Chakrabarty B, Parekh N. 2014. Graph centrality analysis of structural ankyrin repeats. International Journal of Computer Information Systems and Industrial Management Applications, 6: 305-314

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

Jiao X, Chang S, Li C, Chen W, Wang C. 2007. Construction and application of the weighted amino acid network based on energy. Physical Review E, Article ID: 051903

Khansari M, Kaveh A, Heshmati Z. 2016. Centrality measures for immunization of weighted networks. Network Biology, 6(1): 12-27

Koschutzki D, Schreiber F. 2004. Comparison of centralities for biological networks. In: Proceeding of the German Conference of Bioinformatics (GCB'04). Lecture Notes in Informatics LNI P-53. 199-206, Springer

Kundu S. 2005. Amino acid network within protein. Physica A, 346: 104-109

Newman MEJ. 2002. Assortative mixing in networks. Physical Review Letters, 89(20): 2087011-2087014

Sengupta D, Kundu S. 2012. Role of long and short range hydrophobic, hydrophilic and charged residues contact network in protein's structural organization. BMC Bioinformatics, 13: 142

Shu JJ. 2017. A new integrated symmetrical table for genetic codes. Biosystems, 151: 21-26

Watts DJ, Strogatz SH. 1998. Collective dynamics of small-world networks. Nature, 393: 440-442

Wuchty S, Stadler PF. 2003. Centers of complex networks. Journal of Theoretical Biology, 223(1): 45-53

Xin SH, Zhang WJ. 2020. Construction and analysis of the protein protein interaction network for the olfactory system of the silkworm *Bombyx mori*. Archives of Insect Biochemistry and Physiology, 100(1): e21737

Zhang GL, Zhang WJ. 2019. Protein-protein interaction network analysis of insecticide resistance molecular mechanism in *Drosophila melanogaster*. Archives of Insect Biochemistry and Physiology, 100(1): e21523

Zhang WJ. 2016. Screening node attributes that significantly influence node centrality in the network. Selforganizology, 3(3): 75-86