Article

Centrality based analysis of amino acids network

Chandra Borah[1], **Tazid Ali**[2]
[1]Department of Mathematics, Maryam Ajmal Women's College of Science and Technology, Hojai, Assam 782435, India
[2]Department of Mathematics, Dibrugarh University, Assam 786004, India
E-mail: chandra92borah@gmail.com, tazid@dibru.ac.in

Received 29 September 2023; *Accepted 20 October 2023*; *Published online 1 November 2023*; *Published 1 June 2024*

Abstract
A network is a crucial asset in biology for capturing and exploring interaction data in biological systems of many types, such as protein-protein communications, amino acid associations, gene regulation, and cellular metabolism. In this article, we constructed an amino acid distance matrix by considering each base's positional relevance in a codon, chemical types: Purine and Pyrimidine, and H-bonding count. Based on the amino acid distance matrix, we eventually generated a twenty amino acid network having evolutionary significance. We reviewed multiple centrality metrics to assess the relative importance of amino acids in the proposed network: Degree Centrality, Closeness Centrality, Betweenness Centrality, Eigenvector Centrality, Eccentricity Centrality, and Radiality Centrality. We also looked at the correlation coefficients between the different centrality measures to figure out whether the network is assortative or disassortative. Furthermore, we examined the Clustering Coefficient and Degree Distribution as two effective network measures, and the results seem noteworthy.

Keywords amino acid; genetic code; codon redundancy; centrality measure; correlation; 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

An elementary function in molecular biology is called protein biosynthesis, in which genetic signals are transmitted from DNA to active amino acid sequences through the processes of transcription and translation. Each amino acid is a fundamental building block and operational element of all living organisms, and when they are incorporated, they form the protein structure united by peptide bonds. In general, twenty different amino acids are recognized to be involved in protein synthesis. All amino acids comprise a carboxyl group (COOH), an amino group (NH2), and an R-group in addition to a core carbon atom surrounded by hydrogen atoms. Amino acid networks (AAN s) are commonly generated using the cartesian coordinates of amino acid residues from protein molecules contained in a protein data bank (Yan et al., 2014).

DNA, or deoxyribonucleic acid, is the carrier of genetic information in humans and virtually all other organisms. DNA stores data as a code formed of four nitrogenous bases: Adenine (A), Guanine (G), Cytosine (C), and Thymine (T). The array of these bases specifies the data accessible for constructing and nourishing an organism in the same way that letters of the alphabet occur in specific ordering to construct words and phrases. The four DNA bases form specific pairs: A with T has two hydrogen bonds, and C with G has three hydrogen bonds. A genetic code is a triplet code from the four potential bases: A, G, C and T or U (in mRNA) that determines an amino acid out of 20 different amino acids. These four bases incorporate to generate 64 codons in total. In general, 61 codon triplets correlate to the 20 amino acids, whereas the extra triplets, UAA, UAG, and UGA, are referred to as termination codons. As a result, there must be some commonality, implying that more than one codon codes for the same amino acid. The phenomenon by which these synonymous codons code for the same amino acid is known as codon redundancy.

Genetic code research can provide vital insights into protein synthesis and amino acid evolution. The placement of nitrogenous bases in codons specifies the distinct physico-chemical characteristics of amino acids. The placements of Purine (A or G) and Pyrimidine (C or G) base in a codon and their H-bond count ($A = U$ or $C \equiv G$) are all essential factors to be considered in the mechanism of codon-anticodon interaction (Lehmann, 2000; Sanchez et al., 2005). Over the years, numerous researchers have strived to explore different genetic code enigmas: why there is codon redundancy, finding the most significant base location in a codon, the codon-anticodon interaction, the H-bonding count versus amino acid physicochemical aspects, and so on (Beland and Allen, 1994; Freeland and Hurst, 1998; Bashford and Jarvis, 2000). The natural distinction between base locations in a codon, the chemical types of bases, Purine and Pyrimidine, and their hydrogen bond count proven to be the most relevant codon features used in the genetic code research (Sanchez et al., 2005).We used all these codon features in this study to compute the mean distance between two codons and then explored the genetic codon structure. The average distance between the related codons is used to assess the distance between two amino acids. We are interested in three bases' positional relevance in our research since the second base is the biologically most important and the third base to be least important (Woese, 1965), base classifications: purine and pyrimidine, and their H-bonding count: two or three. Further, we developed an undirected amino acids network (AAN) in which the nodes are amino acids, and the links reflect the impact between amino acids.

In the recent years, networks have been used to unravel numerous complex systems in a variety of domains, including computer science, biology, technology, and social science. A network is a powerful tool in biology for representing and exploring interaction data of numerous forms in biological systems, such as protein-protein communications, biochemical and gene regulation (Barabási and Oltvai, 2004; Zhang, 2018, 2023). New insights into the molecular functions underlying these systems can be uncovered by analyzing interactions at the network level (Vidal et al., 2011). Numerous studies have been conducted over the years in the biological networks field to obtain a detailed description of the genetic code (Bertman and Jungck, 1979; Bashford et al., 1998; Jiao et al., 2007; Gohain et al., 2015). Using several centrality metrics, Schreiber and Koschutzki (2004) conducted a comparative analysis of biological networks. In their analysis, Sanchez et al. (2004) proposed a Boolean expression for the genetic codons in which physicochemical properties linked with the partial order of the codon set and Boolean deductions between codons. Recognizing the evolutionary relevance of base placements, Ali et al. (2016) looked at multiple centrality metrics in their amino acid network analysis while overlooking significant aspects like base type (purine and pyrimidine) and hydrogen bonding factor. Ali and Borah (2021) established an amino acids network and explored centrality metrics and network parameters by estimating the codon's transition and transversion mutation and the impact of base positions. Newman (2002) used multiple centrality measures and correlations to explore the protein-protein interaction networks, food web, and neural networks, as well as explained the biological aspects of data transmission in assortative and disassortative networks. In the light of the above discussion, it is evident that there are still numerous areas of genetic codon analysis that remain unexplored, and analyzing these will provide us with a comprehensive understanding of the proteins and amino acids' evolution.

Here, we provide a brief overview of the subsequent sections of the article. Section 2 covers network theory's fundamental concepts, followed by a review of six centrality measurements with their biological interpretations. In section 3, using the proposed distance measure, we construct a 20 amino acid network having evolutionary relevance concerning the base's positional significance, chemical types, and H-bonding count. We examined numerous centrality measures for the network and assessed the correlation coefficients between them. A statistical study of two network parameters is also featured. Section 5 serves as a summary and conclusion to the article.

2 Node Centralities: Definition, Description and Biological Significance

Centrality measure is a network theory approach to estimating the node scores and the significance of a particular node in a network (Haliki, 2021; Zhang, 2018, 2023). Each centrality metric offers unique information about a node. In biological networks, it is crucial, for example, to recognize core nodes or intermediate nodes that impact network topology, depending on the biological situation. Schreiber and Koschutzki (2004) investigated the centrality measures for biological networks, namely the transcriptional and PPI networks. Their findings showed that multiple metrics of centrality should be considered when studying biological networks.

In this section, we introduce some of the existing network centralities. For each centrality, we demonstrate the mathematical formula, a brief illustration, and the likely biological consequences in a protein network.

2.1 Some basic definitions

2.1.1 An undirected graph *G* is specified as a pair *(V, E),* where *V* is a set of vertices denoting the nodes and *E* is a collection of edges indicating the links between the nodes.

2.1.2 The nodes *u* and *v* in *G* are regarded as immediate neighbors if an edge *e* connects them, i.e., *e = (u, v).*

2.1.3 The collection of all the nodes adjacent to *u* defines its neighbor, i.e., *N(u).*

2.1.4 The degree of a node *u* in an undirected network is the number of links the node has to other nodes, and it is defined as *deg(u)* = |*N(u)*|, where *N(u)* count the nodes u's neighbors.

2.1.5 An adjacency matrix *M* of an undirected graph$G = (V, E)$is an$n \times n$ symmetric matrix, *n* being the no. of nodes, where each entry $a_{ij} = 1$ if and only if *(i, j) ∈E* and $a_{ij} = 0$ otherwise.

2.1.6 A walk is a finite alternating series of nodes and links for any graph $G = (V, E)$ that begins and ends with nodes. A walk of length n is a non-empty sequence $v_0 l_0 v_1 l_1 \cdots l_{n-1} v_n$ of nodes and links in G such that $l_i = (v_i, v_{i+1})$ for all $i < n \in \{0,1,2, \dots, n-1\}$.

2.1.7 A simple path is a walk with no repeated nodes.

2.1.8 If there is a path linking any two nodes in an undirected graph G, then the graph is said to be connected.

2.1.9 A path with the shortest distance between any two nodes, such as *u* and *v*, is said to be the shortest path between them.

2.2 Degree Centrality

The simplest centrality metric is the degree of centrality. It is the most basic topological index, relating to the number of nodes directly connected to a particular node *v* and denoted by $C_{deg}(v)$. It is mathematically defined as:

$$C_{deg}(v) = deg\ (v)$$

A significant node is said to have a lot of connections when it has a high degree of centrality. In general, nodes with more links are more crucial to the structure and have a significant impact on others.

Biological relevance: Nodes with a high degree of centrality are referred to as hubs, as they are linked to multiple neighbors (Zhang, 2018, 2023). Hubs are observed frequently in scale-free networks (Pavlopoulos, 2011). Hubs dominate scale-free networks, which are innately resistant to random attacks but vulnerable to specific changes (Albert et al., 2000). Degree centrality has been used extensively in biological network research. For example, Jeong et al. (2001) use it to describe the degree of a protein in the network with the fatality of its elimination. Aftabuddin and Kundu (2007) described three types of protein networks: hydrophilic, hydrophobic, and charged and revealed that the average degree of a hydrophobic network is substantially higher than that of the other two networks.

2.3 Closeness Centrality

The term "Closeness Centrality" refers to how closely a node is associated with the rest of the nodes in the network on a large scale (Freeman, 1978). A node may readily communicate with every other node if it is nearby. The reciprocal of the sum of the shortest paths between each node v and all other nodes in the network is defined as the node v's closeness centrality.

Mathematically,

$$C_{clos}(u) = \frac{(n-1)}{\sum_{v \in V} d(u,v)}$$

where $d(u,v)$gives the shortest path between the nodes u and v, and n counts the no. of nodes in the network. Following the above definition, a node has the highest closeness centrality if it has the lowest overall shortest path distance. The node with the highest closeness centrality is substantially associated with all other nodes.

Biological relevance: Closeness centrality has been utilized to detect the most core metabolites in genome-based largescale metabolic networks (Ma and Zeng, 2003; Zhang, 2018, 2023) and to obtain insight into the evolution of the metabolic organization in genome-based large-scale metabolic networks (Mazurie et al., 2010). It has been recognised as the best centrality metric for determining the metabolic core of a network (Silva et al., 2008). Wuchty and Stadler (2003) illustrate the correlation with the service facility location problem by applying closeness centrality to numerous biological networks.

In an amino acid network, the closeness centrality value of an amino acid reflects the possibility that it is functionally significant for several other amino acids while being irrelevant for a few others. An amino acid with the highest closeness in comparison to the network's average closeness centrality will be readily crucial to the control of other amino acid activity and may play an influential role in amino acid evolution.In their analysis, Ali et al. (2016) asserted that Tyrosine with the highest closeness centrality value plays a crucial role in amino acid evolution.

2.4 Betweenness Centrality

Betweenness centrality in network theory estimates the magnitude to which a node resides on the pathways linking the other nodes (Freeman, 1978). A node with a high betweenness centrality can have a significant effect on the network due to its control over information flow between other nodes.

Mathematically, betweenness centrality is defined as follows:

$$C_{bet}(u) = \sum_{v \neq u \epsilon V} \sum_{w \neq u \in V} \frac{s_{vw}(u)}{s_{vw}}$$

where s_{vw} refers to the no. of shortest routes with v and w their end nodes, and $s_{vw}(u)$ is the no. of shortest routes from v to w that pass via u. Betweenness centrality highlights the identification of nodes that enhance the network information flow. An important node will lie on a large percentage of the pathways connecting the most of the other nodes on a network. We can track data flow across the network by observing this node.

Biological relevance: A node's high betweenness in a biological network, such as a protein-signaling network, may indicate the importance of a protein in keeping communication proteins together (Zhang, 2018, 2023). Proteins with high betweenness centralities are referred to as "bottlenecks" because of their role as crucial connector proteins with critical functional and dynamic features (Barabási, 2011). Potapov et al. (2005) discussed betweenness centrality in mammalian transcriptional regulatory networks and noted that betweenness is an intriguing topological feature concerning the biological significance of various components. Bora et al. (2020) explored distance based amino acid network and noted that Tyrosine (Y) takes the highest betweenness centrality value suggesting its topologically important location in the network.

2.5 Eigenvector Centrality

Another significant metric of centrality is eigenvector centrality (Bonacich, 1972). Eigenvector centrality is defined as the largest eigenvector of the adjacency matrix of the associated network. If A is the network's adjacency matrix and X is an eigenvector corresponding to A's eigenvalue, we may write the equation:

$$AX = \lambda X$$

The eigenvector centrality is the eigenvector of the largest eigenvalue (Bonacich, 1972). The nodes with the highest eigenvector centrality are related to significant neighbors. It is a way of determining a node's dominance in a network.

Biological relevance: Eigenvector centrality metric has been utilized in biology to detect synthetic genetic linkages (Paladugu et al., 2008), gene-disease relationships (Özgür et al., 2008), and network hubs (Zotenko et al., 2008). Ali and Borah (2021) computed eigenvector centrality for the amino acid network and observed that polar amino acid Q, N, H and Y have higher values.

2.6 Eccentricity Centrality

Eccentricity centrality is a metric that indicates how quickly a node can be contacted from other nodes. Let $G = (V, E)$ be an undirected network. The eccentricity centrality is mathematically defined as follows:

$$C_{ecc}(v) = \frac{1}{max\{dist(u, v): u \in V\}}$$

Here, $dist(u, v)$ gives the shortest path distance between the nodes u and v. If the eccentricity of the node v is high, this indicates that all other nodes are nearby. On the contrary, a low eccentricity indicates that at least one node (and all its associates) is far from node v. So, if eccentricity is large, it becomes a more relevant metric.

Biological relevance: A node's eccentricity in a biological network, such as a protein-signaling network, may be defined as the mitigation with which all other proteins in the network can reach a protein. As a result, a protein with a high eccentricity in comparison to the network's average eccentricity is more frequently influenced by the function of other proteins or, conversely, might easily affect multiple other proteins. Lower eccentricity proteins, on the other hand, frequently serve a minor functional role in the system (Chavali et al., 2010).

2.7 Radiality Centrality

Radiality is a node centrality metric.Let$G = (V, E)$ be an undirected network.The radiality of a node u is determined by finding the shortest path between node u and every other node in the network. It is defined as follows:

$$C_{rad}(u) = \frac{\sum_{v \in V}(D_G + 1 - dist(u,v))}{n-1}$$

Here, D_G refers the diameter of the network G and $dist(u,v)$ is the shortest path length between the nodes u and v.

In general, a high radiality indicates that the node is relative to the other nodes concerning diameter, whereas a low radiality suggests that the node is peripheral. If a node has a high eccentricity, closeness as well as radiality value, it suggests the node's high central placement in the network (Scardoni and Laudanna, 2012).

Biological relevance: In a protein-signaling network, the radiality centrality of a node may be interpreted as a measure of the likelihood of a protein being biologically significant for multiple other proteins while being irrelevant for a few others. As a result, a protein with a high radiality compared to the network's average eccentricity will be somewhat more crucial in the regulation of other proteins, while some others will be unaffected by its action.

3 Analysis of Amino Acids Network

The analysis of genetic code features leads us to a biochemical interpretation of data transmission from DNA via mRNA to protein. We looked at the genetic code's unique features of positional distinction, the Purine and Pyrimidine classes of bases, as well as the Hydrogen bonding factor involved with base in our investigation. Ali et al. (2016) strived to explore only its positional features, which was insufficient for a comprehensive analysis of the genetic code. In this section, we proposed a new distance measure to assess the distance between two codons, which results in the distance between two amino acids. Using the proposed distance metric, we offered an overall idea of the evolutionary relevance of 20 Amino Acids in terms of positional importance, chemical kinds, and base's H-bonding number.

Sanchez et al. (2005) noted that the four RNA (or DNA) bases might be arranged or sorted based on the codon-anticodon associations. The physicochemical properties of bases: chemical classes (Purine and Pyrimidine) and the hydrogen bond numbers are the crucial factors of the codon-anticodon associations in order to obtain two sequences in the base set. Accordingly, two sequences of the base set are obtained: {A, C, G, U} and {U, G, C, A}. An addition operation is introduced on the first base set in such a way that it is isomorphic to the cyclic group $(Z_4, +)$(Sanchez et al., 2005).

Identifying each base with the corresponding integer in Z_4 as given by Table 1, we define the distance between any two bases X and Y as |X - Y|. For example, the distance between the bases A and G will be |A - G| = |0 - 2| = 2 (Table 2).

Table 1 Sum operation on the set B.

Sum	+	A	C	G	U
	A	A	C	G	U
	C	C	G	U	A
	G	G	U	A	C
	U	U	A	C	G

Table 2 Computing the distance between bases.

D = |X - Y|	A	C	G	U
A	0	1	2	3
C	1	0	1	2
G	2	1	0	1
U	3	2	1	0

As per evolutionary influence, the codon's second base is the most biologically important, and the third one to be the least. A codon has three base positions, and each one has a distinctive contribution to the corresponding amino acid.

Considering (1) the evolutionary value of the base locations in a codon, (2) the hydrogen bond number and the chemical form (Purine or Pyrimidine) of the base, the distance between the two codons $X_1X_2X_3$ and $Y_1Y_2Y_3$ is defined as follows:

1. If there is a difference between the first bases of the two codons, then give the value $2|X_1 - Y_1|$, otherwise give 0.
2. If there is a difference between the second bases of the two codons, then give the value $3|X_2 - Y_2|$, otherwise give 0,
3. If there is a difference between the third bases of the two codons, then give the value $1|X_3 - Y_3|$, otherwise give 0.

So, the distance between the codons $X_1X_2X_3$and $Y_1Y_2Y_3$ is given by $2|X_1 - Y_1| + 3|X_2 - Y_2| + 1|X_3 - Y_3|$ and we denote it by $D_C(X_1X_2X_3, Y_1Y_2Y_3)$.

i.e., $D_C(X_1X_2X_3, Y_1Y_2Y_3) = 2|X_1 - Y_1| + 3|X_2 - Y_2| + 1|X_3 - Y_3|$.

We find the distance between the codons ACC and UAC is

$$D_C(ACC, UAC) = 2|A - U| + 3|C - A| + 1|C - C| = 2|0 - 3| + 3|1 - 0| + 1|0 - 0| = 9$$

To measure the distance between any two amino acids, we compute the mean distance between the respective codons. We compute the distance between the amino acids Lysine (provided by AAA, AAG) and Tyrosine (provided by UAU, UAC) from Table 4.

Table 3 The distance between the codons.

D_C	UAU	UAG
AAA	9	8
AAG	7	6

Thus, the distance between Lysine (K) and Tyrosine (Y) is 7.50, obtained by taking the average of the above distance measures (D_C). The weighted Manhattan distance we described here is analogous to those presented in Sanchez et al. (2006).

The following Table 4 provides the distance between each pair of amino acids.The changes in physicochemical characteristics of amino acids increase as the distance values rise. There is a high difference in distance between a hydrophilic and hydrophobic amino acid. The distance between Lysine (strongly hydrophilic) and phenylalanine (highly hydrophobic) is the greatest, with a value of 16.50. A small distance value between the related amino acids indicates mutations with minor differences between the associated codons.

Table 4 Distance matrix for amino acids pairs.

	R	***K***	***E***	***Q***	***D***	***N***	***H***	***P***	***Y***	***S***	***T***	***G***	***W***	***A***	***M***	***C***	***F***	***L***	***V***	***I***
R	**0**	**8.50**	**9.83**	**7.83**	**10.00**	**8.50**	**8.00**	**4.91**	**11.83**	**6.89**	**5.41**	**3.92**	**5.67**	**6.91**	**5.33**	**6.00**	**9.60**	**6.33**	**6.83**	**5.67**
K	**8.50**	**0**	**5.00**	**3.00**	**5.50**	**1.50**	**3.50**	**6.25**	**7.50**	**9.33**	**4.25**	**11.25**	**13.00**	**8.33**	**10.00**	**13.50**	**16.50**	**13.50**	**14.25**	**10.33**
E	**9.83**	**5.00**	**0**	**3.00**	**1.50**	**5.50**	**3.50**	**6.25**	**3.50**	**8.00**	**8.25**	**7.25**	**9.00**	**4.25**	**14.00**	**9.50**	**12.50**	**12.00**	**10.25**	**14.33**
Q	**7.83**	**3.00**	**3.00**	**0**	**3.50**	**3.50**	**1.50**	**4.25**	**4.50**	**8.67**	**6.25**	**9.25**	**11.00**	**6.25**	**11.00**	**11.50**	**14.50**	**11.50**	**12.25**	**12.50**
D	**10.00**	**5.50**	**1.50**	**3.50**	**0**	**5.00**	**3.00**	**6.25**	**3.00**	**7.83**	**8.25**	**7.25**	**9.00**	**4.25**	**14.00**	**9.00**	**12.00**	**12.33**	**10.25**	**14.33**
N	**8.50**	**1.50**	**5.50**	**3.50**	**5.00**	**0**	**3.00**	**6.25**	**3.00**	**8.33**	**4.25**	**11.25**	**13.00**	**8.25**	**10.00**	**13.00**	**16.00**	**13.67**	**14.25**	**10.33**
H	**8.00**	**3.50**	**3.50**	**1.50**	**3.00**	**3.00**	**0**	**4.25**	**5.00**	**8.50**	**6.25**	**9.25**	**11.00**	**6.25**	**11.00**	**11.00**	**14.00**	**11.67**	**12.25**	**11.33**
P	**4.91**	**6.25**	**6.25**	**4.25**	**6.25**	**6.25**	**4.25**	**0**	**8.25**	**5.50**	**3.75**	**6.25**	**8.00**	**3.50**	**9.00**	**8.25**	**11.25**	**8.60**	**9.25**	**9.33**
Y	**11.83**	**7.50**	**3.50**	**4.50**	**3.00**	**3.00**	**5.00**	**8.25**	**0**	**7.17**	**10.25**	**9.25**	**7.00**	**6.25**	**16.00**	**7.00**	**10.00**	**11.67**	**12.25**	**16.33**
S	**6.89**	**9.33**	**8.00**	**8.67**	**7.83**	**8.33**	**8.50**	**5.50**	**7.17**	**0**	**6.17**	**5.83**	**5.33**	**4.83**	**10.33**	**5.17**	**8.17**	**9.11**	**8.83**	**9.20**
T	**5.41**	**4.25**	**8.25**	**6.25**	**8.25**	**4.25**	**6.25**	**3.75**	**10.25**	**6.17**	**0**	**8.25**	**13.00**	**5.25**	**7.00**	**10.25**	**13.25**	**10.08**	**11.25**	**7.33**
G	**3.92**	**11.25**	**7.25**	**9.25**	**7.25**	**11.25**	**9.25**	**6.25**	**9.25**	**5.83**	**8.25**	**0**	**3.00**	**4.25**	**8.00**	**3.25**	**6.25**	**6.25**	**4.25**	**8.25**
W	**5.67**	**13.00**	**9.00**	**11.00**	**9.00**	**13.00**	**11.00**	**8.00**	**7.00**	**5.33**	**13.00**	**3.00**	**0**	**6.50**	**9.00**	**1.00**	**4.00**	**6.67**	**6.00**	**10.00**
A	**6.91**	**8.33**	**4.25**	**6.25**	**4.25**	**8.25**	**6.25**	**3.50**	**6.25**	**4.83**	**5.25**	**4.25**	**6.50**	**0**	**11.00**	**6.25**	**9.25**	**9.25**	**7.25**	**11.25**
M	**5.33**	**10.00**	**14.00**	**11.00**	**14.00**	**14.00**	**11.00**	**9.00**	**16.00**	**10.33**	**7.00**	**8.00**	**9.00**	**11.00**	**0**	**10.00**	**7.00**	**4.33**	**5.50**	**1.33**
C	**6.00**	**13.50**	**9.50**	**11.50**	**9.00**	**13.00**	**11.00**	**8.25**	**7.00**	**5.17**	**10.25**	**3.25**	**1.00**	**6.25**	**10.00**	**0**	**4.00**	**5.67**	**6.25**	**10.33**
F	**9.00**	**16.50**	**12.50**	**14.50**	**12.00**	**16.00**	**14.00**	**11.25**	**10.00**	**8.17**	**13.25**	**6.25**	**4.00**	**9.25**	**7.00**	**4.00**	**0**	**4.00**	**3.25**	**7.33**
L	**6.33**	**13.50**	**12.00**	**11.50**	**12.33**	**13.67**	**11.67**	**8.60**	**11.67**	**9.11**	**10.08**	**6.25**	**6.67**	**9.25**	**4.33**	**5.67**	**4.00**	**0**	**3.25**	**4.00**
V	**6.83**	**14.25**	**10.25**	**12.25**	**10.25**	**14.25**	**12.25**	**9.25**	**12.25**	**8.83**	**11.25**	**4.25**	**6.00**	**7.25**	**5.50**	**6.25**	**3.25**	**3.25**	**0**	**3.33**
I	**5.67**	**10.33**	**12.50**	**12.50**	**14.33**	**10.33**	**11.33**	**9.33**	**16.33**	**9.20**	**7.33**	**8.25**	**10.00**	**11.25**	**1.33**	**10.33**	**7.33**	**4.00**	**3.33**	**0**

We can construct an amino acid network from the following distance matrix (Table 4), which is symmetric and contains 210 data points. We compute that the mean value of the distribution of 210 data points is 7.258. The mean value (i.e., 7.258) is deemed to be the target value for finding the links between any two amino acids, as the mean suggests that data points tend to cluster around it. We take all 20 amino acids as a set of nodes, with any two nodes u and v linked if their distance is less than or equal to 7.258. In Fig. 1, we displayed the resulting network G.

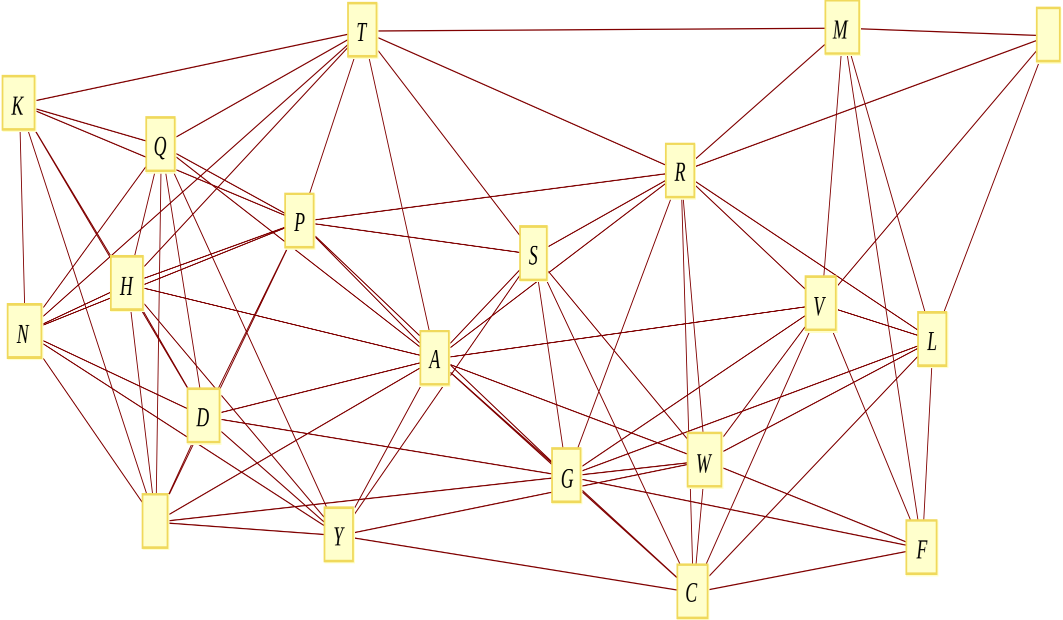

Fig. 1 Network of amino acids (*G*).

Here, we obtain a network where all the 20 amino acids are connected. We have observed that as the value of d (distance) increases, the likelihood of that a mutational event transforming a codon into another encoding for a different amino acid increase.

Because the distance metric in Table 4 is given by the differences in the corresponding codons of the two amino acids, it is plausible to conclude that two amino acids are compatible if they are connected by an edge. The likelihood of two edge-bound amino acids evolving into one another is high; related codon mutations govern this process. We have the adjacency matrix A for the network G presented below. We see, $A = A^T$.

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

3.1 Centrality based analysis of amino acidsnetwork

The centrality metric is an essential graph theoretical measure for analyzing the significance of nodes in a network. We used six centrality measures to examine the amino acid network G, as displayed in Table 6: degree, closeness, betweenness, eigenvector, eccentricity, and radiality. Table 5 reveals that amino acid A has the highest rank for degree centrality, betweenness centrality, eigenvector centrality, eccentricity centrality, and radiality centrality, while amino acid P has the highest value for closeness centrality.

The amino acids P, R, T, G, and A are the most influential of all amino acids, with considerably higher centrality measures concerning the average centrality measure. Each of these amino acids has a degeneracy value of 4 or above. Whereas I, K, N, M, F, and L are less influential since they have lower centrality ratings, and except for L, each one has a degeneracy value of less than or equal to 3.

As degree centrality counts a node's immediate link, a node with a high degree centrality score signifies its prominence within the network. We see that amino acid A has the highest degree centrality value, which is 13 (Table 6). The hydrophobic amino acid Alanine is significant because it is ambivalent and might exist inside or outside a protein molecule. Concerning the average degree of centrality, the hydrophilic amino acids R, E, Q, and D have a greater degree of centrality, while hydrophobic amino acids M, F, L, and I have a lower degree centrality value.Here, we follow the IMGT Physico-chemical class table of 20 amino acids (Pommié et al. 2004).

A high value of a node's closeness centrality implies a closer proximity distance to all other nodes along with the rapid data flow across the node. In our scenario, the higher closeness centrality values of P and R (0.576 and 0.513, respectively) indicate the highest number of predecessor and successor nodes in the data transformation route. Proline is an amino acid with a distinctive cyclic shape that helps many proteins fold but slows the rate of peptide bond production by the ribosome (Melnikov et al. 2016). The amino acid K has the lowest closeness centrality value of 0.218, which conveys that the data flow is better mediated through the rest of amino acids than through K.

Betweenness centrality estimates the comparative biochemical relevance of amino acids roughly based on the number of shortest routes in the network that pass through them. A higher score of an amino acid's betweenness centrality signifies the recognition of amino acids that contribute much of the network's data transmission. Due to their central placements in the network, the amino acids R, A, and G have greater betweenness centrality values (with A having the highest value of 36.883). It is worth mentioning that, Amino acid R is hydrophilic, A is hydrophobic, and G is neutral. The most hydrophobic amino acid I has the lowest betweenness centrality value of 0.000, indicating that it is the most away from the rest of the network. As a result, it appears as an intermediate between fewer pairs of amino acids than the others.

Table 5 Centrality measures for amino acids.

Node	*Degree*	*Closeness*	*Betweenness*	*Eigenvector*	*Eccentricity*	*Radiality*
R	11	0.513	31.896	0.256	0.500	2.579
K	7	0.218	0.974	0.181	0.333	2.053
E	9	0.372	7.362	0.242	0.333	2.368
Q	9	0.352	3.406	0.238	0.333	2.316
D	9	0.380	7.362	0.242	0.333	2.368
N	8	0.333	2.089	0.204	0.333	2.210
H	9	0.352	3.548	0.238	0.333	2.316
P	11	0.576	20.417	0.289	0.500	2.579
Y	9	0.279	17.879	0.235	0.333	2.368
S	8	0.352	5.227	0.224	0.500	2.421
T	9	0.432	28.289	0.226	0.500	2.474
G	11	0.452	32.928	0.275	0.500	2.579
W	9	0.365	9.599	0.222	0.333	2.421
A	13	0.475	36.883	0.337	0.500	2.684
M	6	0.271	10.224	0.115	0.333	2.158
C	9	0.365	9.599	0.222	0.333	2.421
F	6	0.327	2.058	0.130	0.333	2.105
L	8	0.292	6.169	0.163	0.333	2.210
V	9	0.339	11.565	0.196	0.333	2.368
I	4	0.306	0.000	0.079	0.333	1.842
Avg.	8.500	0.367	12.374	0.216	0.383	2.342

Eigenvector centrality is more apparent and noteworthy than the degree centrality. Eigenvector centrality of a node will provide us with a rating or score, dependent on the number of linkages a node has to other nodes. It tracks not just a node's connections but also the connection of its neighbors, the connection of its neighbors'

neighbors, and so forth (Ali et al., 2016; Chakrabarty and Parekh, 2014). The top three amino acids with higher eigenvector centrality are A, P, and G (with A having the highest value of 0.337), since the total of the immediate and indirect connections of A, P, and G is maximal. We assert that the amino acids: A, P, and G, having higher eigenvector centrality, are crucial in the amino acid biosynthesis process. It is evident from Table 6 that the strong hydrophobic amino acids M, F, L, V, and I have lower eigenvector centrality measures.

If a node's eccentricity is high, it suggests that all other nodes are nearby.An amino acid with a high eccentricity in comparison to the network's average eccentricity will be impacted more rapidly by other amino acids, or conversely, it may easily affect numerous amino acids. As per Table 6, the amino acids R, P, S, T, G, and A have the highest eccentricity centrality value of 0.500. i.e., they could easily interact with the several amino acids due to their more central positions in the network. It is also worth noting that all these amino acids have a degeneracy score of 4 or above.

A node with a high radiality centrality value is typically more proximate to the other nodes. The amino acids R, P, G, and A have greater radiality centrality values (with A having the highest value of 2.684) due to their high central placements in the network. As a result, they could control the flow of information throughout the network. These amino acids have high scores for closeness and eccentricity centrality (Scardoni and Laudanna, 2012). Likewise, the lower radiality values of I, K, and F indicate that these amino acids prefer to remain at the outer boundary location, where they have less chance of transmitting data.

From the perspective of the base's positional impact in a codon plus its chemical types and the factor of Hydrogen bonding count in DNA structures, we may infer that amino acids Alanine, Arginine, Glycine, Proline, and Threonine are relatively most important in the evolutionary process. Here, the most important one, i.e., **Alanine**, is thought to be one of the earlier amino acids to be contained in the genetic code standard **repository** (Higgs and Pudritz, 2009; Kubyshkin and Budisa, 2019a). From a biochemical standpoint, the "Alanine World" hypothesis describes the evolutionary preference of amino acids in the repository of the genetic code (Kubyshkin and Budisa, 2019b).In this concept, the amino acids used for ribosomal protein synthesis are restricted to Alanine derivatives that can be used to construct α-helix or β-sheet secondary structural components.

Arginine is the most hydrophilic amino acid. It is frequently found on the protein's surface, where the hydrophilic head group may interface with the polar environment, for example via hydrogen bonding and salt bridges (Barnes, 2007). As a result, it is frequently encountered at the interface of two proteins (Kleanthous, 2000). Glycine is the most basic stable amino acid, and is integral during the synthesis of alpha-helics in protein secondary structure. It may fit into either hydrophobic or hydrophilic situations due to its small side chain of one hydrogen atom. Glycine is thought to be generated by early genetic codes in the process of evolution (Trifonov, 2000; Higgs and Pudritz, 2009; Ntountoumi et al., 2019). Proline is the sole amino acid where the side chain is linked to the protein backbone twice, resulting in a nitrogen-containing ring with five components. Threonine is an amino acid that is used in protein synthesis. It is a polar, uncharged amino acid with a carboxyl group, an amine group, and a side chain containing a hydroxyl group.

3.2 Correlation between six centralities

Correlation is a statistical technique employed to estimate a potential linear relationship between two continuous variables. In this section, we analyze the correlation between six centrality measures for the amino acids network. The correlation coefficient (cc) allows us to examine a network's assortative and disassortative characteristics. If $cc > 0$, the network is assortative, which indicates that higher degree nodes tend to cluster with other high degree nodes; social networks are one such type (Newman, 2002). In a disassortative network, we have $cc < 0$, which implies that a higher degree node is more likely to cluster with a lower degree node. Newman (2002) addressed this phenomenon in the context of protein interaction networks, food webs and neural networks. The cc-value can range from -1 (absolute negative correlation) through 0 (no correlation) to +1 (absolute positive correlation) (Swinscow and Campbell, 2002).

The correlation coefficients between sixcentralitymetrics are shown in Table 6.

Table 6 Correlation coefficients.

	$\boldsymbol{C_d}$	$\boldsymbol{C_c}$	$\boldsymbol{C_b}$	$\boldsymbol{C_\lambda}$	$\boldsymbol{C_e}$	$\boldsymbol{C_r}$
$\boldsymbol{C_d}$	1	0.747	0.769	0.960	0.604	0.946
$\boldsymbol{C_c}$	0.747	1	0.719	0.705	0.773	0.787
$\boldsymbol{C_b}$	0.769	0.719	1	0.657	0.784	0.796
$\boldsymbol{C_\lambda}$	0.960	0.705	0.657	1	0.581	0.909
$\boldsymbol{C_e}$	0.604	0.773	0.784	0.581	1	0.700
$\boldsymbol{C_r}$	0.946	0.787	0.796	0.909	0.700	1

Here, we use Pearson's approach to evaluate all the correlation coefficients (r). Table 6 shows that all centrality metrics are highly associated except eigenvector centrality with eccentricity. Since each pair of the centrality metrics has a positive correlation coefficient, our network is an assortative type. As a result, information may be transmitted smoothly throughout this network.

3.3 Network parameters

We can examine a network's overall behavior by considering a variety of network parameters. In this paper, we solely look at two introductory network parameters: (1) the clustering coefficient and (2) the degree of distribution. The clustering coefficient is a statistical measure that reveals how likely a network is to be divided into clusters. A cluster is a subset of vertices with multiple edges linking them. Suppose that v is a node with $deg\ (v) = n$ in an undirected network *G* and that there are *l* links between *v*'s immediate neighbors in *G*, the clustering coefficient of v in *G* is calculated by: $C_c = 2l/n(n-1)$. As a result, clustering coefficient (C_c) calculates the ratio of the number of linkages between *v*'s immediate neighbors to the total number of potential links, i.e., $n(n-1)/2$. The value of C_c lies within the given range of: $0 \leq C_c \leq 1$. Similarly, while studying a network, degree distribution is a crucial parameter to examine. The number of connections or edges linking to a node in a network defines its degree. The probability distribution of these degrees throughout the entire network is known as the degree distribution. The degree distribution captures only a small amount of the network characteristics since it ignores how nodes are related. However, such information still gives significant insights into a network organization. Because of the degree distribution notable role in explaining network topology, much emphasis has been made on understanding the mechanisms that drive the shape of degree distribution in ecological networks (Jordano et al., 2003).

3.3.1 Clustering coefficients

A node with a high clustering coefficient has close associations with its neighbors, indicating that there are more number linkages between them. The clustering coefficient of a node appears to have an impact on its adjacent node, hence regulating the data flow throughout the network (Sengupta and Kundu, 2012). When compared to random networks, biological networks have a much greater average clustering coefficient.

Table 7 shows the clustering coefficient value for each of the 20 amino acids.

Table 7 Clustering coefficients values (for the network *G*).

R	*K*	*E*	*Q*	*D*	*N*	*H*	*P*	*Y*	*S*	*T*	*G*	*W*	*A*	*M*	*C*	*F*	*L*	*V*	*I*
0.545	0.905	0.750	0.805	0.750	0.821	0.805	0.618	0.555	0.643	0.528	0.545	0.694	0.526	0.600	0.694	0.800	0.714	0.750	1.000

The degree of the amino acid and the number of direct connections between the nearby amino acids affect the clustering coefficient of the amino acid. The more connections there are between neighboring amino acids, the greater the clustering coefficient value. Thus, a network with a higher clustering coefficient value slows down the information flow through the network.

For the network *G*, we notice that the most hydrophobic amino acid I has the highest clustering coefficient value of 1.000. In the network, I form a clique structure with the amino acids M, R, V, and L (Fig. 2). A clique is a subset of amino acids in the network *G* where each pair of distinctive amino acids are adjacent. Strikingly, I, M, L and V are all strong hydrophobic amino acids. Finding cis-regulatory patterns, matching three-dimensional molecular structures, and detecting groups of regularly co-expressed genes in microarray datasets are just a few biological uses for the identification and study of clique structures (Zhang et al., 2009; Voy et al., 2006).

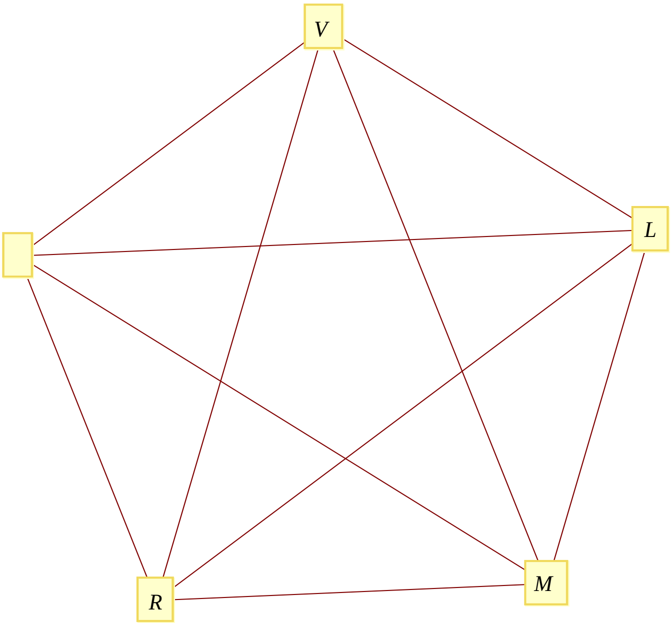

Fig. 2 A clique structure of amino acids from the network *G*.

The clustering coefficient of the entire network has a value of 0.702, which is the average of the 20 amino acids' clustering coefficient values. We note that most hydrophobic amino acids: F, L, V, and I and most hydrophilic amino acids: N, D, Q, E, and K have high clustering coefficient values than that of the whole network. As a result, compared to the overall network, the flow of information is rather sluggish in the proximity of the most hydrophilic and the most hydrophobic amino acids.

Also, we notice that all the neutral amino acids: G, T, S, Y, and P (except H)(IMGT Class Table, Pommié et al., 2004) have low clustering coefficient values than that of the whole network. Thus, compared to the entire network, the flow of information is relatively fast in the proximity of the neutral amino acids.

3.3.2 Degree of Distribution

Here, we assess the degree distribution values of the different amino acids for the network G. An amino acid's degree in the network is determined by how many links it has to the rest amino acids. The degree distribution is given by $P(k) = m_k/m$, if there are m nodes in a network and m_k of them, have degree k. The likelihood that the chosen amino acid will have precisely k connections is expressed by the degree distribution value of the amino acid. We have displayed the degree of distribution values for different amino acids in Table 6.

Table 8 Degree distribution values.

R	***K***	***E***	***Q***	***D***	***N***	***H***	***P***	***Y***	***S***	***T***	***G***	***W***	***A***	***M***	***C***	***F***	***L***	***V***	***I***
0.15	0.05	0.45	0.45	0.45	0.15	0.45	0.15	0.45	0.15	0.45	0.15	0.45	0.05	0.10	0.45	0.10	0.15	0.45	0.05

Here, the most hydrophobic amino acid I has the minimum distribution value of 0.05. The hydrophilic amino acids E, Q, and D, the neutral amino acids H, Y, and T, and the hydrophobic amino acids W, C, and V all have a distribution value of 0.45, which is greater than the average distribution value of 0.265.

4 Conclusion

In this article, we demonstrate how to employ ideas, models, and methods from the network theory world to discover hidden aspects and features of an amino acid network. Interactions at the network level can deliver novel insights into the genetic codon systems that support the biosynthesis of amino acids. Accordingly, we endeavored to assess the evolutionary significance of the 20 amino acids by analyzing six centrality metrics for the amino acid network.First, we established a distance metric between amino acids by estimating the positional importance, chemical kinds, and H-bonding count of each base. Following that, we developed a network model of 20 amino acids that describes the compatibility linkage depending on the amino acid distance matrix.

Different centrality metrics are used as a graph theoretical approach to investigate the impact of each amino acid. We used six centrality measures: degree, closeness, betweenness, eigenvector, eccentricity, and radiality centrality to analyze the amino acid network. We found that amino acid Alanine ranks top for the centralities: degree, betweenness, eigenvector, eccentricity, and radiality, whereas amino acid P ranks highest for closeness centrality. Our results provide credence to the “Alanine hypothesis,” as the amino acids employed in ribosomal protein synthesis are confined to Alanine derivatives that can be used to form α-helix or β-sheet secondary structural components (Kubyshkin and Budisa, 2019a). Alanine is also believed to be one of the primary amino acids discovered in genetic coding systems (Higgs and Pudritz, 2009). Overall, we observed that amino acids P, R, T, G, and A, all with higher codon redundancy values, are the most influential among all amino acids, with considerably greater centrality measures concerning the average centrality measure. We may infer that the amino acids Alanine, Arginine, Glycine, Proline, and Threonine played the most crucial roles in the evolutionary process. Amino acids I, K, N, M, F, and L, all with lower codon redundancy (except for L), are less influential since they have lower centrality ratings. The most hydrophobic amino acid I (Isoleucine) has the lowest betweenness centrality score of 0.000, suggesting that it is the most distant from the rest of the network. As a result, it occurs as an intermediate between fewer pairs of amino acids than the others.

Correlation coefficients explain the solidity and direction of an association between variables. We looked into the correlation coefficients between the six centrality measures and found that all centrality metrics are somewhat highly correlated. Our network is of the assortative type since each pair of centrality measures has a positive correlation coefficient. As a result, data may be transported smoothly throughout this network.

The clustering coefficient quantifies a network's tendency to be split into clusters. We noticed that the most hydrophobic amino acid I has the highest clustering coefficient value of 1.000, and it forms a clique structure with the amino acids M, R, V, and L (Fig. 2). Most hydrophobic amino acids, such as F, L, V, and I, and most hydrophilic amino acids, such as N, D, Q, E, and K, have greater clustering coefficient weights than the whole network. So, the rate of data transmission is somewhat sluggish in the proximity of the most hydrophilic and the most hydrophobic amino acids. Finally, we examined the degree distribution of the 20 amino acids.

References

Aftabuddin M, Kundu S. 2007. Hydrophobic, hydrophilic, and charged amino acid networks within protein. Biophysical Journal, 93(1): 225-231. https://doi.org/10.1529/biophysj.106.098004

Albert R, Jeong H, Barabasi AL 2000. Error and attack tolerance of complex networks. Nature, 406: 378-382

Ali T, Akhtar A, Gohain N. 2016. Analysis of amino acid network based on distance matrix. Physica A, 452: 69-78. https://doi.org/10.1016/j.physa.2016.01.074

Ali T, Borah C. 2021. Analysis of amino acid network based on mutation and base positions. Gene Reports, 24:101291. https://doi.org/10.1016/j.genrep.2021.101291

Barabási AL, Gulbahce N, Loscalzo J. 2011. Network medicine: a network-based approach to human disease. Nature Reviews Genetics, 12(1): 56-68. https://doi.org/10.1038/nrg2918

Barabási AL, Oltvai ZN. 2004. Network biology: Understanding the cell’s functional organization. Nature Reviews Genetics, 5: 101-113.https://doi.org/10.1038/nrg1272

Bashford JD, Jarvis PD. 2000. The genetic code as a periodic table: algebraic aspects. Biosystems, 57: 147-161. https://doi.org/10.1016/S0303-2647(00)00097-6

Bashford JD, Tsohantjis I, Jarvis PD. 1998. A supersymmetric model for the evolution of the genetic code. Proceedings of the National Academy of Sciences USA, 95(3): 987-992

Barnes MR. 2007. Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data. John Wiley & Sons, USA

Beland P, Allen TF. 1994. The origin and evolution of the genetic code. Journal of Theoretical Biology, 170: 359-365. doi: 10.1006/jtbi.1994.1198

Bertman MO, Jungck JR. 1979. Group graph of the genetic code. Journal of Heredity, 70: 379-384. https://doi.org/10.1093/oxfordjournals.jhered.a109281

Bonacich P. 1972. Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology, 2: 113-120. https://doi.org/10.1080/0022250X.1972.9989806

Bora PK, Hazarika P, Baruah AK. 2020. Distance based amino acids network analysis. Gene Reports, 21: 100933. https://doi.org/10.1016/j.genrep.2020.100933

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

Chavali S, Barrenas F, Kanduri K, Benson M. 2010. Network properties of human disease genes with pleiotropic effects. BMC Systems Biology, 4: 78

da Silva MR, Ma H, Zeng AP. 2008. Centrality, network capacity, and modularity as parameters to analyze the core-periphery structure in metabolic networks. Proceedings of IEEE, 96(8): 1411-1420. DOI:10.1109/JPROC.2008.925418

Freeman L. 1978. Centrality in social networks conceptual classification. Social Networks, 1(3): 215-239. https://doi.org/10.1016/0378-8733(78)90021-7

Freeland SJ, Hurst LD.1998. The genetic code is one in a million. Journal of Molecular Evolution, 47: 238-248. https://doi.org/10.1007/PL00006381

Gohain N, Ali T, Akhtar A. 2015.Lattice structure and distance matrix of genetic code. Journal of Biological Systems, 23(03): 485-504. https://doi.org/10.1142/S0218339015500254

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

Higgs PG, Pudritz RE. 2009. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. Astrobiology, 9(5): 483-490

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, 75(5Pt1): 051903.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature, 411(6833): 41-42

Jordano P, Bascompte J, Olesen JM. 2003. Invariant properties in coevolutionary networks of plant-animal interactions. Ecology Letters, 6: 69-81

Kleanthous C. 2000. Protein-Protein Recognition. Oxford University Press, UK

Kubyshkin V, Budisa N. 2019a.The Alanine World Model for the Development of the Amino Acid Repertoire in Protein Biosynthesis. International Journal of Molecular Sciences, 20(21): 5507

Kubyshkin V, Budisa N. 2019b. Anticipating alien cells with alternative genetic codes: away from the alanine world!. Current Opinion in Biotechnology, 60: 242-249

Lehmann J. 2000.Physico-chemical constraints connected with the coding properties of the genetic system. Journal of Theoretical Biology, 202: 129-144

Ma HW, Zeng AP. 2003. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics, 19(11): 1423-1430

Mazurie A, Bonchev D, Schwikowski B, et al. 2010. Evolution of metabolic network organization. BMC Systems Biology, 4: 59

Melnikov SV, Khabibullina NF, Mairhofer E, et al. 2019. Mechanistic insights into the slow peptide bond formation with D-amino acids in the ribosomal active site. Nucleic Acids Research, 47(4): 2089-2100

Newman MEJ. 2002. Assortative mixing in networks. Physical Review Letters, 89(20): 208701.

Ntountoumi C, Vlastaridis P, Mossialos D, et al. 2019. Low complexity regions in the proteins of prokaryotes perform important functional roles and are highly conserved. Nucleic Acids Research, 47(19): 9998-10009

ÖzgürA, Vu T, Erkan G, Radev DR. 2008. Identifying gene-disease associations using centrality on a literature mined gene interaction network. Bioinformatics, 24(13): i277-i285

Paladugu SR, Zhao S, Ray A, Raval A. 2008. Mining protein networks for synthetic genetic interactions. BMC Bioinformatics, 9: 426

Pavlopoulos GA, Secrier M, Moschopoulos CN, et al. 2011. Using graph theory to analyze biological networks. BioData Mining, 4(1): 10. https://doi.org/10.1186/1756-0381-4-10

Pommié C, et al. 2004. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. Journal of Molecular Recognition, 17: 17-32

Potapov AP, Voss N, Sasse N, Wingender E. 2005. Topology of mammalian transcription networks. Genome Inform, 16(2): 270-278.

Sanchez R, Morgado E, Grau R. 2004. The genetic code Boolean lattice. MATCH Communications in Mathematical and in Computer Chemistry, 52: 29-46

Sanchez R, Morgado E, Grau R. 2005. Gene algebra from a genetic code algebraic structure. Journal of Mathematical Biology, 51(4): 431-457. https://doi.org/10.1007/s00285-005-0332-8

Sanchez R, Grau R, Morgado E. 2006. A Novel DNA Sequence Vector Space over anextended Genetic Code Galois Field. MATCH Communications in Mathematical and in Computer Chemistry, 56(1): 5-20

Scardoni G, Laudanna C. 2012. Centralities Based Analysis of Complex Networks. New Frontiers in Graph Theory. Intech, London, UK. https://doi.org/10.5772/35846.

Schreiber F, Koschutzki D. 2004. Comparison of centralities for biological networks. In: Proceeding German Conference of Bioinformatics (GCB), LNI, P-53: 199-206

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

Swinscow TDV, Campbell MJ. 2002. Statistics at Square One. BMJ Books, London, UK. https://doi.org/10.1093/ije/dyg117

Trifonov EN. 2000. Consensus temporal order of amino acids and evolution of the triplet code. Gene, 261(1): 139-151. https://doi.org/10.1016/S0378-1119(00)00476-5

Vidal M, Cusick ME, Barabási AL, 2011. Interactome networks and human disease, Cell, 144: 986-998

Voy BH, Scharff JA, Perkins AD, et al. 2006. Extracting Gene Networks for Low-Dose Radiation Using Graph Theoretical Algorithms. PLoS Computational Biology, 2(7): e89

Woese CR. 1965. Order in the genetic code. Proceedings of the National Academy of Sciences USA, 54(1): 71-75. https://doi.org/10.1073/pnas.54.1.71

Wuchty S, Stadler PF. 2003. Centers of complex networks. Journal of Theoretical Biology, 223(1): 45-53. https://doi.org/10.1016/S0022-5193(03)00071-7

Yan W, Zhou J, Sun M, et al. 2014. The construction of an amino acid network for understanding for understanding protein structure and function. Amino Acids, 46: 1419-439

Zhang H, Song X, Wang H, Zhang X. 2009. MIClique: An algorithm to identify differentially coexpressed disease gene subset from microarray data. Journal of Biomedicine and Biotechnology, 2009: 642524.

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

Zhang WJ. 2023. netAna: A tool for network analysis. Network Biology, 13(4): 192-212

Zotenko E, Mestre J, O'Leary DP, Przytycka TM. 2008. Why do hubs in the yeast protein interaction network tend to be essential: re-examining the connection between the network topology and essentiality. PLoS Computational Biology, 4(8): 1000140