Communications in Mathematics and Applications
Vol. 14, No. 2, pp. 561–573, 2023
ISSN 0975-8607 (online); 0976-5905 (print)
Published by RGN Publications
DOI: 10.26713/cma.v14i2.2140

http://www.rgnpublications.com

Research Article

Analysis of Amino Acids Network Based on Nucleotide of DNA

Sanjay Sharma*, Birinchi Kumar Boruah and Tazid Ali

Department of Mathematics, Dibrugarh University, Dibrugarh 786004, Assam, India
***Corresponding author:** snjyshrma90@gmail.com

Received: December 19, 2022 **Accepted:** May 1, 2023

Abstract. The sequence of amino acids in protein synthesis is determined by the order of monomers of DNA. A study on the physicochemical aspect of nucleotide of DNA gives us valuable characteristics related to amino acids and protein functions. We considered six different parameters for four nucleotide of DNA and weighing each of them we constructed a distance matrix for twenty amino acids and subsequently, an amino acids network. In this network, we looked into evolutionary pattern of amino acids based on nucleotide of DNA and how it plays significant roles in the function of protein stability and membrane proteins. Lastly, we investigate several centrality metrics and explored correlation coefficients to assess the network's assortativity for a comparative analysis of amino acids. We have also examined the clustering coefficient, degree distribution, and skewness as network parameters.

Keywords. Codon, Amino acids, Distance matrix

Mathematics Subject Classification (2020). 92B05, 92D20, 92D10

Copyright © 2023 Sanjay Sharma, Birinchi Kumar Boruah and Tazid Ali. *This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.*

1. Introduction

Most biological processes that make life depend on proteins as their building components because proteins have many different activities, from participating in structural cellular components to activating cellular reactions. The essential biological processes transform the genetic information from *Deoxyribonucleic acid* (DNA), which is located within chromosomes in the nucleus, via mRNA to protein within the cell through transcription and translation. The DNA molecule comprises two chains coiled around one another to form a double helix. Each chain is made up of monomeric units called as nucleotides, which are repeating units connected

by chemical bonds. Nucleotides are composed of (I) a pentose sugar, (II) a Phosphate group, (III) Four nitrogenous bases, adenine (A), thymine (T), guanine (G), and cytosine (C), and they differ from each other by their nitrogenous base used. A triplet (codon) is made up of three nucleotides in a row and codes for a single amino acid. As a result, each three-letter sequence represents amino acids. The genetic information in DNA is used to create proteins, which are polymers of amino acids. The genetic code is a biological mechanism that determines how DNA nucleotide sequences are transcribed into mRNA codon sequences, which are then translated into amino acid protein sequences (Sanchez [15]). The standard genetic code is redundant, i.e., 64 codons code for 20 amino acids. For example, the amino acid glycine (G) has degeneracy four, so there exists a subset of four codons $\{GGT, GGC, GGA, GGG\}$ or in other words, there are four synonymous codons coding the amino acid glycine (Godfrey-Smith and Kim [10]).

Examining the genetic code aims to provide insight into the evolution of amino acids and protein synthesis. Numerous research projects have been conducted in this area over the years. Numerous topics still require clarification, such as why there are synonymous codons, which codon positions are more or less crucial, how codon anticodons relate to physicochemical qualities through their H-bonding, etc. Different researchers have made many attempts to study the genetic code based on the properties of the most significant codons, i.e., the positions of the nucleotide bases within the codon, the quantity of H-bonds, and the chemical forms, purine and pyrimidine, all play significant roles in the research of the codon-anticodon mechanism (Ali and Borah [3], Ali *et al*. [4], Bashford and Jarvis [5], Bora *et al*. [6], Lehman [13], and Sanchez *et al*. [16]). In this paper, we investigate the genetic code mechanism by incorporating some physicochemical parameters of nucleotide bases that build up the DNA and which, in turn, build up the proteins. The genomic information comprises three nucleotide triplets (codons) in DNA that code for a specific amino acid or signals the end of protein synthesis. Firstly, we assign binary weight on the nucleotide bases based on their various parameters and then for each pair of codons, Hamming distances are determined. Secondly, by considering the mean distances between the synonymous codons of amino acids, we calculate distance between each pair of amino acids. Lastly, by applying the concept of graph theory and taking in to consideration of the parameters in the manuscripts, we explored the evolutionary pattern of the amino acids in the network by analyzing the triplets of nucleotides.

To establish a comprehensive arrangement of the genetic code, numerous researchers have contributed a lot of publications to the domain of biological networks (Aftabuddin and Kundu [1], Chang *et al*. [8], and Yan *et al*. [19, 20]). Ali *et al*. [2], considered different physicochemical properties of amino acids to investigate amino acid networks. They discuss many centrality measures to examine the amino acids relative significance. Using graph theory and various centrality measures, they analyze the biological networks and give rank to the vertices. Using hydrogen bond numbers and the chemical kinds of bases purines and pyrimidines, Sanchez *et al*. [17] define two dual Boolean codon lattices in their study, which shows hydrophobicity affects how each amino acids codon assignments are distributed. Ali and Borah [3] obtained a distance matrix and then an amino acids network after weighing the impact of base positions and codon transition and transversion mutation and investigated the characteristics that have

evolved in the amino acids, which are proteins' building blocks. The chemical types of bases, purine, pyrimidine, and their hydrogen bond number, as well as the differences between the base's natural placements inside the codon, have all been essential factors in the analysis of the genetic code. Bora *et al*. [6] examine the evolutionary characteristics of amino acids in protein synthesis by defining a distance matrix among the amino acids in their publication while considering these properties. From the above literature review, we can indicate that there is a wide range of research areas that the genetic codon study still leaves open, allowing one to better understand the evolution of amino acids and proteins.

2. Some Basic Concepts of Graph Theory and Properties of Nucleotide Bases

In this section, we review some preliminary concepts related to graphs, various centrality measures and physicochemical parameters of the nucleotide bases.

2.1 Some Elementary Concepts of Graphs

A finite set V of vertices and a finite set $E \subset V \times V$ of edges make up an undirected graph $G = (V, E)$. The vertices u and v are described as the incident with the edge e and next to one another if an edge $e = (u, v)$ links the two. The neighbourhood $N(u)$ of u is the collection of all vertices close to vertex u. If each edge of graph G has a direction, then graph G is a directed graph or digraph, with a set V of vertices and a set E of edges such that $e \in E$. A graph is said to be loop-free if an edge does not link one vertex to another. A graph's adjacency matrix A is an $(n \times n)$ matrix with $a_{ij} = 1$ if and only if $(i, j) \in E$ and $a_{ij} = 0$ in all other cases. Any undirected graph has a symmetric adjacency matrix. The number of edges with v as one of their endpoints is considered the vertex degree. A walk is a finite alternating sequence of vertices and edges where each edge coincides with the vertices before and after, except for the beginning and finished vertices. No edges appear more than once throughout a walk. A vertex can, however, seem more than once. Initial and terminal vertices in a walk are referred to as starting and end vertices. The walk is closed if the initial and terminal vertices coincide, otherwise, it is open. If a walk has no recurring vertices, it is considered to be a path. A walk with no repetitive edges is referred to as a trail. A path between two vertices (u, v) with the shortest or geodesic length is referred to as the most concise or geodesic path. Every pair of vertices in a graph can be walked between if the graph is linked (Ali *et al*. [4]).

2.2 Various Centrality Measures in Graphs

The centrality measure indicates a vertex's relative importance inside the graph in graph theory. Centrality is a real-valued function on a graph's vertices. In more technical terms, centrality is a function f that assigns a value $f(v) \in R$ to each vertex $v \in V$ of a given graph G. Here, we have covered the four most popular centrality measures below.

2.2.1 Degree Centrality Measure

Degree of centrality, or $C_d(u)$, is the most basic measure (Freeman [9]). It is described as the number of nodes directly connected to node u. First, neighbours of the given node refer to nodes directly linked to a particular node u. The degree of centrality indicates how many interactions

a key node participates in. This interaction shows the node's immediate importance or risk in the relevant network. Mathematically, $C_d(u) = \deg(u)$. However, in some specific problems, the degree of centrality is not a reliable indicator of its significance or risk. In the real world, significant nodes may indirectly connect to other nodes.

2.2.2 Eigen Value Centrality Measure

Eigenvalue centrality is a method for arranging graphs vertices based on the notion that the values of adjacent vertices influence the value of a single vertex. In contrast to different measures, the centrality values of a vertex neighbours are taken into account in addition to their location within the graph (Koschützki and Schreiber [12]). Mathematically, it is defined as $C_\lambda(u) = \sum_{v \in N(u)} C_\lambda(v)$, where λ is the Eigenvalue of the adjacency matrix and $N(u)$ is the neighbourhood of the vertices u. Considering graphs adjacency matrix representation A_{ij}, $C_\lambda(v_i) = \sum_{j=1}^{n} a_{ij} C_\lambda(v_j)$, which directly leads to the well-known eigenvector computing problem, where $AS = \lambda S$ and the greatest eigenvector is the eigenvector-centrality ($C := S$).

2.2.3 Betweenness Centrality Measure

The betweenness centrality is another popular centrality measure (Koschützki and Schreiber [11], and Watts and Strogatz [18]). Interactions between non-adjacent nodes that are between each other depending on the other node, typically on the pathways that connect them. A number of shortest paths of a node that cross through it, u, determines its betweenness centrality. Mathematically, it is defined as, $C_{btw}(u) = \sum_{s=u \in V} \sum_{t \neq u \in V} \frac{\sigma_{st}(u)}{\sigma_{st}}$, where σ_{st} is the number of shortest paths connecting vertex s to vertex t, and $\sigma_{st}(u)$ is the number of shortest pathways connecting s to t that run through u. Betweenness centrality shows which nodes contribute the most to the network information flow. A vital node will be located on numerous paths in the network connecting other nodes. We may manage the network information from this node. Two neighbours would be unable to communicate with one another without these nodes. Due to the possibility of many of the shortest paths passing through them, high-degree nodes typically have high betweenness centrality. However, a high-degree node is not always required by a high betweenness centrality node.

2.2.4 Closeness Centrality Measure

The opposite of the longest pathways between each node added together and every other node in the network is generally used to define closeness centrality. A node closeness centrality (Freeman [9], and Koschützki and Schreiber [11]) indicates how important it is and how readily it can connect or communicate with other nodes in the network. Closeness centrality measure how long it will take for information to propagate from one vertex to other reachable vertices in the network. To put it another way, this measurement enables the identification of nodes that, on average, require fewer steps to communicate with other nodes than their immediate neighbours. Mathematically, it is defined as $C_{cl}(u) = \frac{n-1}{\sum_{v \in V} d(u,v)}$, where $d(u,\ v)$ denotes the shortest path distance between the pair of vertices u and v and n represents the network size.

2.3 Physicochemical Parameters of the Nucleotide Bases

Two helical segments that are coiled around the same axis make up DNA. Depending on how they are coiled, DNA helices are categorized as either right-handed or left-handed. DNA

right-handed helices, however, are the most reliable and standard structure. In contrast to one another, the two helical strands are opposite to each other, where one stretches from $5'$ to $3'$ and the other from $3'$ to $5'$. The four nitrogenous bases create the strands (A, T, C, and G). The amino acids that make up proteins are encoded by these nucleotide bases, which act as genetic data banks. Here in this paper, we attempt to investigate the evolutionary behavior of amino acids by considering physiochemical parameters of nucleotide bases as shown in Table 1 that preserve their identity when translated into protein synthesis.

Table 1. Few physicochemical properties of four nucleotide bases of DNA

Properties	Adenine (A)	Guanine(G)	Cytosine(C)	Thymine (T)
Molar mass (g/mol)	135.13	153.13	111.10	126.11
Density (g/cm^3)	1.6	2.2	1.5	1.2
Melting point (℃)	365	360	325	317
Solubility (g/L)	1.03	Insoluble	7.69	3.82
Acidity (pK_a)	9.80	12.3	12.2	9.7
Strength[1]	Very weak	Strong	Very strong	Weak

3. Amino Acids Distance matrix and Network

Nucleotide sequences in DNA are symbolic sequences in mathematics and can be thought of as digital sequences to facilitate mathematical analysis. DNAs can be shown as binary strings because they are one-dimensional sequences. The nucleotide base set $\Lambda = \{A, C, G, T\}$ are categorized into six groups as shown in Table 1, based on some physicochemical parameters. For our analysis, we assigned 2-digit binary weights 00, 01, 10, 11 to uniquely identify the four bases as follows:

$$\Omega_1 : \Lambda \to \mathbb{Z}_2^2, \text{ where } \Omega_1(A) = 10,\ \Omega_1(G) = 11,\ \Omega_1(C) = 00,\ \Omega_1(T) = 01,$$
$$\Omega_2 : \Lambda \to \mathbb{Z}_2^2, \text{ where } \Omega_2(A) = 10,\ \Omega_2(G) = 11,\ \Omega_2(C) = 01,\ \Omega_2(T) = 00,$$
$$\Omega_3 : \Lambda \to \mathbb{Z}_2^2, \text{ where } \Omega_3(A) = 11,\ \Omega_3(G) = 10,\ \Omega_3(C) = 01,\ \Omega_3(T) = 00,$$
$$\Omega_4 : \Lambda \to \mathbb{Z}_2^2, \text{ where } \Omega_4(A) = 01,\ \Omega_4(G) = 00,\ \Omega_4(C) = 11,\ \Omega_4(T) = 10,$$
$$\Omega_5 : \Lambda \to \mathbb{Z}_2^2, \text{ where } \Omega_5(A) = 01,\ \Omega_5(G) = 11,\ \Omega_5(C) = 10,\ \Omega_5(T) = 00,$$
$$\Omega_6 : \Lambda \to \mathbb{Z}_2^2, \text{ where } \Omega_6(A) = 00,\ \Omega_6(G) = 10,\ \Omega_6(C) = 11,\ \Omega_6(T) = 01.$$

The binary weights functions Ω_1, Ω_2, Ω_3, Ω_4, Ω_5 and Ω_6 are defined according to molar mass, density, melting point, solubility, acidity and strength, respectively. Each nucleotide bases in the weight function is assigned a 2-digit binary weight according to their properties value given in Table 1. The highest value of the respective property is assigned with the highest weight, 11 and the least value is assigned with the least weight 00.

[1]D. Duplij and S. Duplij, Symmetry analysis of genetic code and determinative degree (in Russian), ArXiv: *Chemical Physics* (2000), URL: https://arxiv.org/abs/physics/0006062.

3.1 Distances Between Pairs of Codons

In this section, we calculate the total Hamming distance between the pair of codons based on the weight functions defined above. Each codon from the standard genetic code table is encoded with the 6-bit binary strings containing the physicochemical properties of the nitrogenous bases, which are translated into codons and then into amino acid sequences for protein synthesis.

Mathematically, if codons $X_1X_2X_3$, $Y_1Y_2Y_3 \in \Lambda^3$ where $X_j, Y_j = \{\mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}\}$, $j = 1,2,3$ then function $p : \Lambda^3 \to \mathbb{Z}_2^6$ denote encoding the codon to 6-bit binary string with $p(X_1X_2X_3) = p(X_1)p(X_2)p(X_3)$. Now, $d_{\Omega_i} : \Lambda^3 \times \Lambda^3 \to \mathbb{Z}_2^6 \times \mathbb{Z}_2^6$, $i = 1,2,\ldots,6$.
$d_{\Omega_i}(X_1X_2X_3, Y_1Y_2Y_3) = (\Omega_i(X_1)\Omega_i(X_2)\Omega_i(X_3), \Omega_i(Y_1)\Omega_i(Y_2)\Omega_i(Y_3))$ denote the transformation between respective physicochemical properties of the pair of a codon. Here we define

$$d_{H_{\Omega_i}} : \mathbb{Z}_2^6 \times \mathbb{Z}_2^6 \to \mathbb{R},$$

where $d_{H_{\Omega_i}}(\Omega_i(X_1)\Omega_i(X_2)\Omega_i(X_3), \Omega_i(Y_1)\Omega_i(Y_2)\Omega_i(Y_3)) =$ number of a different digit between them, denote the distance between the two pair of codons based on their properties.

The Hamming distance d_H between pair of codons will be the sum of the distances between their respective properties. That is, for two codons $X_1X_2X_3$, $Y_1Y_2Y_3$ we have:

$$d_H(X_1X_2X_3, Y_1Y_2Y_3) = \sum_{i=1}^{6} d_{H_{\Omega_i}}.$$

3.2 Amino Acids Distance Matrix

In this section, we obtained the distances between the 20 amino acids by considering the Hamming distances between the pair of codons. The distance between two amino acids which are coded by the synonymous codons, is given by the formula:

$$d(A_m, A_n) = \frac{\sum_{m=1,n=1}^{N} d_{H_{mn}}}{6N},$$

where

A_m, A_n = Amino acids coded by m and n codons,

$6N$ = Number of properties × product of the number of synonymous codon that code the amino acids A_m, A_n,

$d_{H_{mn}}$ = denote the Hamming distance between the mth codon and nth codon.

As an example, in Table 3, we calculate the distance between amino acids Alanine A (coded by GCT, GCC, GCA, GCG) and Histidine H (coded by CAT, CAC)

$$d(A_4, H_2) = \frac{18+24+27+27+24+18+27+27}{6\times 8} = \frac{192}{48} = 4.00.$$

By similar patterns, we calculated the distances between the pairs of twenty amino acids shown in Table 2.

Table 2. Hamming distance between codons

	CAT	CAC
GCT	18	24
GCC	24	18
GCA	27	27
GCG	27	27

3.3 Amino Acids Network

In Table 3, we present a symmetric distance matrix of 20 amino acids and calculated the mean distance as 3.08 unit. We created an amino acids network using this mean distance of 3.08 as a threshold value. If the distance between two amino acids is 3.08 units or less, they are linked in the network and the resulting graph G is shown in Figure 1. In general, the possibility of a link between a pair of amino acids with a lesser distance has higher possibility to be replaced or evolved during the evolution process.

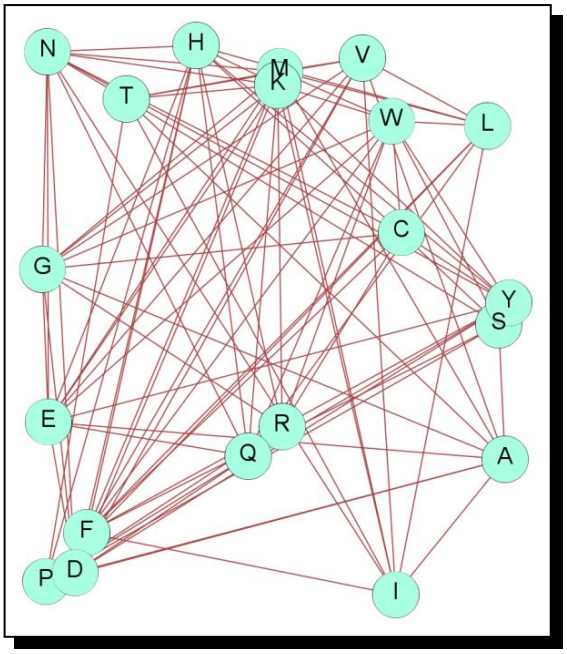

Figure 1. Amino acids graph (G)

3.4 Centrality Measures for the Analysis of Amino Acids Network

An essential graphical theoretical tool to study network is the centrality measures. As stated in Table 4, we used four centrality measures in this paper. According to Table 3, the amino acid Lysine (K) has the highest degree of centrality, closeness centrality betweenness centrality and eigenvector centrality. On the other hand, amino acids Arginine (R) have the lowest centralities values of all the amino acids but it has more significant impact in protein stability and membrane proteins than Lysine.

A high degree centrality value indicates the prominence of a node inside the network as it denotes the close direct relationship of the node. From Table 4, Lysine (K) has the highest and Arginine (R) has lowest degree centrality value, which justify the fact that due to their highly charged side chain they can provide both short-range and long-range interactions for protein folding and membrane protein binding (Li *et al*. [14]).

We calculate the eigenvector centrality of each node of graph G, which is based on how many connections the node has to other nodes. Due to shared neighbours eigenvector centrality of K is high in this instance. Therefore, for this node, the contribution of neighbours, neighbours of neighbours, and so on is same in the evolutionary process (Ali *et al*. [4], and Chakrabarty and

Parekh [7]). Additionally, the strong eigenvector centrality of K points to its central location, which may reflect its primary regulating function in the information flow process. Similar to this, the low eigenvector centrality of R and S denotes both their neighbour's low score and their extreme distance from the center of network.

Based on the significant number of shortest paths that travel through a node in the network, betweenness centrality estimates a node's evolutionary significance. K has the highest betweenness centrality value, while D, C, I, and R have low values. It is evident from Figure 1 that, in comparison to D, C, I, and R, K has the most significant number of shortest pathways and the greatest information flow due to its central location. The highest value of K denotes the topologically substantial position of the node and its greatest contribution to the evolutionary information flow in the network.

A highest closeness centrality value of a node indicates that it is closer to all other nodes and that information moves quickly through it. In our situation, the maximum number of predecessor and successor nodes in the evolutionary process is suggested by the larger proximity centrality value of K. We can acquire a general idea of K location since the proximity centrality of a protein interaction network is often high for the active site.

The Amino Acids lysine and arginine are influenced by the physicochemical properties of the nucleotides of DNA and are equally crucial in protein stability and membrane protein. We also notice specific significant characteristics of lysine and arginine from the current analysis, which may support our findings. A few of them are: Arginine and lysine, which are cations suitable for binding to the phosphate anion. They are frequently present in the regions of proteins called active centers, where they bind phosphorylated substrates and are crucial for preserving the overall charge balance of a protein. Lysine has a lysyl side chain and due to its geometric shape, proteins are less stable as electrostatic interactions occur less frequently and only in one direction. On the other hand, arginine has a side chain of an aliphatic straight chain with three carbons that end in a guanidino group. Due to its geometric structure, proteins are more stable as electrostatic interactions occur more frequently in three directions. The amino acid arginine is relevant to numerous membrane protein and membrane-active peptide events in which interactions between protein and lipids are significant (Li *et al*. [14]).

3.5 Correlation Among the Various Centrality Measures

This section examines the relationships between the four centrality measures discussed in the preceding section. We investigate a network's assortative and disassortative behavior using the correlation coefficient (r). The network is assortative if r is more significant than zero ($r > 0$), which indicates higher degree nodes tend to cluster with other high-degree nodes. Similarly, higher-degree nodes tend to associate with lower-degree nodes in a disassortative network when r is smaller than zero ($r < 0$). In Table 5, all centrality measures have high correlations with one another, except betweenness centrality, which has a slightly lower correlation than the other three. As in our scenario, all correlated values between each pair of amino acids are more significant than zero, suggesting that the network is assortative and that information transmission during evolution will be simple in this situation (Figure 1).

Table 3. 20 Amino acids distance matrix based on physicochemical properties of nitrogenous base

AA	M	W	I	S	L	R	G	A	N	P	V	T	C	D	K	Q	E	H	F	Y
M	0.00	3.00	1.34	3.34	2.25	3.34	2.87	3.00	3.00	3.50	2.00	1.87	4.50	4.00	2.00	3.50	3.00	4.50	3.00	4.50
W	3.00	0.00	4.34	2.67	3.00	2.00	2.50	4.00	4.00	3.50	4.00	4.00	1.50	4.00	2.00	2.50	3.00	3.50	3.00	2.50
I	1.34	4.34	0.00	3.12	3.05	3.38	3.45	3.00	2.34	3.50	2.00	2.00	3.50	3.78	2.67	4.16	3.67	3.91	2.34	3.84
S	3.34	2.67	3.12	0.00	3.16	3.16	3.34	2.81	3.16	2.67	3.50	2.50	2.34	3.50	3.50	3.67	3.84	3.42	2.50	2.67
L	2.25	3.00	3.05	3.16	0.00	3.17	4.00	3.50	4.17	2.34	2.50	3.50	3.34	4.17	3.84	2.75	3.84	3.00	1.84	3.34
R	3.34	2.00	3.38	3.16	3.17	0.00	2.34	3.84	3.17	3.00	3.84	3.50	2.04	3.50	2.83	2.33	3.17	2.67	3.83	3.33
G	2.87	2.50	3.45	3.34	4.00	2.34	0.00	2.50	3.00	4.00	2.50	3.50	2.50	2.00	3.00	3.50	2.00	3.50	4.00	3.50
A	3.00	4.00	3.00	2.80	3.50	3.84	2.50	0.00	3.50	2.50	2.00	2.00	3.50	2.50	3.50	4.00	2.50	4.00	3.50	3.50
N	3.00	4.00	2.34	3.16	4.17	3.17	3.00	3.50	0.00	4.00	3.50	2.50	3.00	1.50	2.00	3.00	2.50	2.00	3.50	2.00
P	3.50	3.50	3.50	2.67	2.34	3.00	4.00	2.50	4.00	0.00	3.50	2.50	3.50	4.00	4.00	2.50	4.00	2.50	3.00	3.50
V	2.00	4.00	2.00	3.50	2.50	3.84	2.50	2.00	3.50	3.50	0.00	3.00	4.00	2.50	3.50	4.00	2.50	4.00	2.50	4.00
T	1.87	4.00	2.00	2.50	3.50	3.50	3.50	2.00	2.50	2.50	3.00	0.00	4.00	3.50	2.50	4.00	3.50	4.00	3.50	3.50
C	4.50	1.50	3.50	2.34	3.34	2.04	2.50	3.50	3.00	3.50	4.00	4.00	0.00	3.00	4.00	3.50	4.00	2.50	2.00	1.50
D	4.00	4.00	3.78	3.50	4.17	3.50	2.00	2.50	1.50	4.00	2.50	3.50	3.00	0.00	2.50	3.00	1.50	2.00	3.50	2.00
K	4.00	2.00	2.67	3.50	3.84	2.83	3.00	3.50	2.00	4.00	3.50	2.50	4.00	2.50	0.00	2.00	1.50	3.00	4.50	3.00
Q	3.50	2.50	4.16	3.67	2.75	2.33	3.50	4.00	3.00	2.50	4.00	4.00	3.50	3.00	2.00	0.00	2.00	1.50	4.50	2.50
E	3.00	3.00	3.67	3.84	3.84	3.17	2.00	2.50	2.50	4.00	2.50	3.50	4.00	1.50	1.50	2.00	0.00	3.00	4.50	3.00
H	4.50	3.50	3.91	3.42	3.00	2.67	3.50	4.00	2.00	2.50	4.00	4.00	2.50	2.00	3.00	1.50	3.00	0.00	3.00	1.50
F	3.00	3.00	2.34	2.50	1.84	3.83	4.00	3.50	3.50	3.00	2.50	3.50	2.00	3.50	4.50	4.50	4.50	3.00	0.00	2.00
Y	4.50	2.50	3.84	2.67	3.34	3.33	3.50	3.50	2.00	3.50	4.00	3.50	1.50	2.00	3.00	2.50	3.00	1.50	2.00	0.00

Table 4. Different centrality values of twenty amino acids

Amino acids	Degree	Closeness centrality	Betweenness centrality	Eigen centrality
M	11	0.7037	0.0314	0.9125
W	11	0.7037	0.0421	0.8985
I	8	0.6333	0.0141	0.6637
L	8	0.6333	0.0192	0.6631
G	10	0.6785	0.0349	0.8350
A	9	0.6551	0.0319	0.7127
N	11	0.7037	0.0383	0.9206
V	9	0.6551	0.0234	0.7289
T	8	0.6333	0.0209	0.6475
K	12	0.7307	0.0480	1
E	11	0.7037	0.0278	0.9453
F	10	0.6785	0.0404	0.7956
S	7	0.6129	0.0195	0.5571
R	7	0.6129	0.0118	0.5403
C	8	0.6333	0.0172	0.6593
Q	10	0.6785	0.0263	0.8354
Y	10	0.6785	0.0263	0.8450
P	8	0.6333	0.0303	0.6211
H	11	0.7037	0.0386	0.8988
D	9	0.6551	0.0180	0.7766

Table 5. Correlation values of four centrality measures

	C_d	C_c	C_b	C_λ
C_d	1	0.9989	0.8299	0.9891
C_c	0.9989	1	0.8333	0.9885
C_b	0.8299	0.8333	1	0.7526
C_λ	0.9891	0.9885	0.7526	1

4. Network Parameters

To investigate biological networks, we employ different network parameters. We address a few of them in the following sections to clarify the network's behavioral pattern. Only three network parameters, the clustering coefficient, degree distribution and skewness, have been explored here.

4.1 Clustering Coefficients

The clustering coefficient is a measurement that reveals the propensity of a graph to cluster. The ratio of the number of links e_i that are actually used to connect nearest neighbours of node i to the total number of possible links $\frac{k_i(k_i-1)}{2}$, (where k_i is the degree of the node i) between these nearest neighbour is known as the clustering coefficient C_i of a node. It ranges between $[0,1]$. The average of each C_i represents the clustering coefficient for the entire network. Nodes with more significant clustering coefficients represent stronger relationships between neighbouring nodes. In other words, a more significant clustering coefficient of a node value indicates that there are more connections between its neighbours. We have provided the clustering coefficients for each amino acid in Table 6.

Table 6. Clustering coefficient of twenty amino acids

M	W	I	L	G	A	N	V	T	K	E	F	S	R	C	Q	Y	P	H	D
0.54	0.47	0.57	0.50	0.53	0.47	0.52	0.52	0.50	0.53	0.61	0.44	0.47	0.57	0.53	0.57	0.60	0.39	0.50	0.63

Table 6 shows that the amino acids Aspartic Acid (D) and Glutamic Acid (E) have the highest clustering coefficient values. Also, we observed that the degree centrality value of D and E are nine and eleven. The codons GAT, GAC and GAA, GAG codes the two amino acids with acidic group side chain respectively, which are occupied by the nitrogenous bases G and A and have a substantial variation in physicochemical parameters with bases T and C. The definition also shows that the clustering coefficient increases with the degree of amino acids and the number of direct connections between neighbours. Consequently, higher clustering coefficient values of a network significantly impact its nodes and delay the dissemination of information. Therefore, the information can quickly be transmitted in the amino acid network.

4.2 Degree of Distribution

Degree distributions ($P(i)$), $i = 0,1,\ldots$, is equivalent to m_i/m. As a result, m indicates the size of the network and m_i is the total number of degree i vertices. The degree distribution value of a

vertex reflects the likelihood that the selected vertices may have i links precisely. We determine the degree distributions for the network G in Table 7 using amino acids as vertices.

Table 7. Degree distribution of twenty amino acids

M	W	I	L	G	A	N	V	T	K	E	F	S	R	C	Q	Y	P	H	D
0.25	0.25	0.25	0.25	0.20	0.15	0.25	0.15	0.25	0.05	0.25	0.20	0.10	0.10	0.25	0.20	0.20	0.25	0.25	0.15

4.3 Skewness

In 1895, Karl Pearson made the initial suggestion for assessing skewness. When a curve's mode, median, and mean are not the same, skewness is also known as a lack of symmetry. In our analysis, we take into account Karl Pearson's coefficient of skewness, abbreviated S_k and calculated as follows:

$$S_k = \frac{3(\text{Mean-median})}{\text{Standard deviaion}}.$$

The measurement of skewness is between the range of -3 to $+3$. The distribution exhibits positive skewness when $S_k > 0$ and negative skewness when $S_k < 0$, which is symmetric with $S_k = 0$. We compute the distribution's mean, median and standard deviation, i.e., 0.2, 0.225, 0.0628. Using Karl Pearson's formula, we calculate the skewness value as -1.1937. So, based on the degree distribution, we conclude that amino acids have negative skewness.

5. Conclusion

We have interpreted a network of amino acids in this article by taking into account the physicochemical parameter of the nucleotide bases, which is an essential cofactor in numerous biological processes, including the synthesis of amino acids, proteins, cell divisions, etc. To investigate the influence of nucleotide triplet in amino acids evolution, we integrated the various parameters of nucleotides with binary weights and calculated the Hamming distances between two triplets. We then obtained a network of twenty amino acids after establishing a distance matrix and constructing a distance formula.

This network describes the amino acid evolution model that suggests the likelihood of how one amino acid evolves from another. Several centrality measurements are used as a graph-theoretical technique to examine the impact of each amino acid. These centrality measures allow us to conclude that Lysine contributes the most to communicating the evolutionary process. In addition, it was discovered that D and E had the highest clustering coefficient values, meaning that evolutionary information flowed here more slowly than it did for other amino acids. The degree distribution of the amino acids has finally been added. We intend to use a graph theoretic method in our future work to investigate the phylogeny of different species to study their similarity/dissimilarity of protein sequences based on various physico-chemical parameters of nitrogenous bases.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

All the authors contributed significantly in writing this article. The authors read and approved the final manuscript.

References

[1] M. Aftabuddin and S. Kundu, AMINONET - a tool to construct and visualize amino acid networks, and to calculate topological parameters, *Journal of Applied Crystallography* **43** (2010), 367 – 369, DOI: 10.1107/S002188981000110X.

[2] A. Akhtar and T. Ali, Analysis of unweighted amino acids network, *International Scholarly Research Notices* **2014** (2014), Article ID 350276, DOI: 10.1155/2014/350276.

[3] T. Ali and C. Borah, Analysis of amino acids network based on mutation and base positions, *Gene Reports* **24** (2021), 101291, DOI: 10.1016/j.genrep.2021.101291.

[4] T. Ali, A. Akhtar and N. Gohain, Analysis of amino acids network based on distance matrix, *Physica A: Statistical Mechanics and its Applications* **452** (2016), 69 – 78, DOI: 10.1016/j.physa.2016.01.074.

[5] J. D. Bashford and P. D. Jarvis, The genetic code as a periodic table: algebraic aspects, *Biosystems* **57**(3) (2000), 147 – 161, DOI: 10.1016/s0303-2647(00)00097-6.

[6] P. K. Bora, P. Hazarika and A. K. Baruah, Distance based amino acids network analysis, *Gene Reports* **21** (100933), DOI: 10.1016/j.genrep.2020.100933.

[7] B. Chakrabarty and N. Parekh, Graph centrality analysis of structural Ankyrin repeats, *International Journal of Computer Information Systems and Industrial Management Applications* **6** (2014), 305 – 314, URL: https://www.mirlabs.net/ijcisim/regular_papers_2014/IJCISIM_28.pdf.

[8] S. Chang, X. Jiao, X.-Q. Gong, C.-H. Li, W.-Z. Chen and C.-X. Wang, Evolving model of amino acid networks, *Physical Review E* **77** (2008), 061920, DOI: 10.1103/PhysRevE.77.061920.

[9] L. C. Freeman, Centrality in social networks conceptual clarification, *Social Networks* **1**(3) (1978-1979), 215 – 239, DOI: 10.1016/0378-8733(78)90021-7.

[10] P. Godfrey-Smith and S. Kim, Biological information, *The Stanford Encyclopedia of Philosophy*, Summer 2016 Edition, E. N. Zalta (editor), (2016), URL: https://plato.stanford.edu/archives/sum2016/entries/information-biological/.

[11] D. Koschützki and F. Schreiber, Centrality analysis methods for biological networks and their application to gene regulatory networks, *Gene Regulation and Systems Biology* **2** (2008), 193 – 201, DOI: 10.4137/grsb.s702.

[12] D. Koschützki and F. Schreiber, Comparison of centralities for biological networks, *German Conference on Bioinformatics 2004* (GCB 2004), Gesellschaft für Informatik, Bonn, 199 – 206 (2004), URL: https://dl.gi.de/handle/20.500.12116/28667.

[13] J. Lehmann, Physico-chemical constraints connected with the coding properties of the genetic system, *Journal of Theoretical Biology* **202**(2) (2000), 129 – 144, DOI: 10.1006/jtbi.1999.1045.

[14] L. Li, I. Vorobyov and T. W. Allen, The different interactions of lysine and arginine side chains with lipid membranes, *Journal of Physical Chemistry B* **117**(40) (2013), 11906 – 11920, DOI: 10.1021/jp405418y.

[15] R. Sanchez, Symmetric group of the genetic-code cubes. Effect of the genetic–code architecture on the evolutionary process, *MATCH Communications in Mathematical and in Computer Chemistry* **79**(3) (2018), 527 – 560, URL: https://match.pmf.kg.ac.rs/electronic_versions/Match79/n3/match79n3_527-560.pdf.

[16] R. Sanchez, E. Morgado and R. Grau, Gene algebra from a genetic code algebraic structure, *Journal of Mathematical Biology* **51** (2005), 431 – 457, DOI: 10.1007/s00285-005-0332-8.

[17] R. Sanchez, E. Morgado and R. Grau, The genetic code boolean lattice, *MATCH Communications in Mathematical and in Computer Chemistry* **52** (2004), 29 – 46, URL: https://match.pmf.kg.ac.rs/electronic_versions/Match52/match52_29-46.pdf.

[18] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393** (1998), 440 – 442, URL: https://www.nature.com/articles/30918.

[19] W. Yan, J. Zhou, M. Sun, J. Chen, G. Hu and B. Shen, The construction of an amino acid network for understanding protein structure and function, *Amino Acids* **46** (2014), 1419 – 1439, DOI: 10.1007/s00726-014-1710-6.

[20] W. Yan, M. Sun, G. Hu, J. Zhou, W. Zhang, J. Chen, B. Chen and B. Shen, Amino acid contact energy networks impact protein structure and evolution, *Journal of Theoretical Biology* **355** (2014), 95 – 104, DOI: 10.1016/j.jtbi.2014.03.032.