Article

Analysis of amino acids network based on graph mining

Nasrin Irshad Hussain, **Kuntala Baruah**
Department of Computer Application, Assam Rajiv Gandhi University of Cooperative management, Assam 785665, India
E-mail: hussainnasrin531@gmail.com; kuntala17@gmail.com

Received 23 February 2024; *Accepted 31 March 2024*; *Published online 10 May 2024*; *Published 1 September 2024*

Abstract
Applications of graph mining have proliferated across the research spectrum in recent years. Mining data to retrieve information is a big deal as data are unstructured and huge in size, volume and in different data-types as internet is available to everyone and anywhere. Therefore data is so rapidly increased and for that point this mining concept came. In graph mining, analysis of graph base data is considered. In different research fields use graph base mining as it give quick and efficient result of large datasets. Here we consider biological data which are very complex to describe and analyse to extract useful information, so now researchers use computational tools to mine the large datasets, graphs are the most efficiently used. We consider amino acid network to do graph mining and extract some useful patterns from the network. Amino Acid Networks (AANs) are undirected graphs where amino acids are act like vertices and their relationships connect two vertices in protein structures. Every amino acid exhibits different physico-chemical properties. The shift in R groups affects various characteristics of the amino acids. The shift in R groups affects the various characteristics of the amino acids. In this paper we have construct a graph of amino acids based on property similarity and discussed different measures of centrality. We have also investigated the correlation coefficients between different measures of centrality.

Keywords amino acids; centrality measure; correlation coefficients; data mining; graph mining.

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 introduction to Amino acid properties given in following description: Cells are the building blocks of all living things. Each cell contains a single set of chromosomes, which are long strands of DNA or RNA that serve as the model for the entire organism. Genes are DNA blocks which make up a chromosome. Every gene codes for a different protein. The fundamental building blocks and functional components of all living things are proteins. The building components of proteins are amino acids. Every amino acid is a three set code that can have four different bases. One amino acid is specified by a codon which is a unit made up of three base pairs. Based on the evolutionary significance of the genetic code, in a codon the second base position thought

to be biologically most relevant, whereas the third base is thought to be least significant. Amino acid chains create a linear chain that makes up each protein. There are 20 amino acids found till now that occurs in proteins. The chain of amino acids takes on different shapes for different proteins. Deoxyribonucleic acid (DNA) stores genetic information about how to construct or synthesize proteins. The DNA or RNA is the chemical in the cells of animals and plants that carries genetic information. It is made up of unit called nucleotides. Each nucleotide contains one sugar group (deoxyribose), one phosphate group and one nitrogenous base (Adenine (A), Cytosine(C), Guanine (G) or Thymine (T)). The sugar and phosphate group are responsible for the helical backbone of DNA. The purines (Adenine and Guanine) and the pyrimidines (Cytosine and Thymine) form the double helix. The bases are paired and joined together by hydrogen bonds.

In this article we have considered a network of amino acids. Based on the amino acid properties similarity distance, the networks compatibility relation is established. In the field of biology using computational techniques, the researchers give significant contributions to solve complex networks. Bora et al. (2020) constructed an amino acid networks based on properties that exhibits by amino acids. Correlating some physico-chemical and physical properties of amino acid, they analyzed the relative importance of different amino acids using centrality measures. Trying to capture uncertainty in relationships between users in a social network, the fuzzy graph model of Bloch et al. (2023) can do more precisely a social network than a deterministic graph. They generate centrality of each node as a fuzzy relation and find all possible centrality values and determine the corresponding truth degree of every node. In three such measures the centrality is expressed as fuzzy relation. In earlier studies Ali et al. (2016) developed an amino acid distance matrix. The relative evolutionary significance of the nucleotide positions of the corresponding codons is used to define the distance. A network of amino acids is derived from the distance matrix. According to their argument, this network illustrates the evolutionary pattern of amino acids. They have also discussed about how important amino acids are in relation to this network (Gohain et al. 2015). A partial ordering is equipped on the genetic code and a lattice structure has been developed from it. The codon-anticodon interaction, hydrogen bond number and the chemical types of bases play an important role in the partial ordering in the base set and on the structure of the genetic code. Some relations between the lattice structure of genetic code and physico-chemical properties of amino acids had established. Also Aftabuddin and Kundu (2007) discussed about hydrophobic, hydrophilic and charged network within the protein. Hydrophobic amino acids are considered as vertices in the hydrophobic network, whereas hydrophilic and charged amino acids are considered as vertices of hydrophilic and charged networks respectively. If any two atoms from different amino acids are within a cutoff distance $5A^{o}$, the amino acids are considered to be connected. Further they showed that the average degree of the hydrophobic networks has a significantly larger value than that of hydrophilic and charged networks. The average degree of the hydrophilic networks is slightly higher than that of the charged networks. The average strength of the nodes of hydrophobic networks is nearly equal to that of the charged network, whereas that of hydrophilic networks has a smaller value than that of hydrophobic and charged networks. The average strength for each of the three types of networks varies with its degree. The average strength of a node in a charged network increases more sharply than that of the hydrophobic and hydrophilic networks. Each of the three types of networks exhibits the ‘‘small-world’’ property. In earlier studies Jiao et al. (2007) discussed about the weighted amino acid network based on the contact energy. They have shown that weighted amino acid network satisfy “small-world” property. Considering hydrophobic and hydrophilic network, Kundu (2005) taked amino acids as vertices. Two amino acids (vertices) are connected if any two different amino acids are within $5A^{o}$ distance. The paper gave a explanation that small world property within protein satisfy within hydrophobic and hydrophilic network. It was also demonstrated that hydrophobic networks have higher average node counts than hydrophilic networks. Also Schreiber and Koschutzki (2004)

compared centralities for biological networks namely PPI network and transcriptional network. As a result of their study, it was observed that in the analysis of biological networks various centrality measures should be considered.

In earlier studies Baruah and Ali (2022) constructed identity sub-graph on the genetic code and they have studied different centrality measures, clustering coefficient and degree of distribution. Also, they construct a network based on distance matrix. The distance matrix is obtained by transition and transversion mutation of codons. Finally they proposed that this network reflects the evolutionary patterns of amino acids. Further, they have study different centrality measures, correlation coefficients. Also Yuan et al. (2022) suggested a deep graph-based network for predicting protein-protein interaction sites, in which the prediction problem is transformed into a graph classification challenge and resolved using deep learning. In deep learning, this method used initial residual and identity mapping techniques. As a result of the experiment they showed that deeper architecture allows significant performance improvement in comparison to other sequence based or structure based methods.

In present study, we try to explore some graph theoretic notions in amino acids network. The paper is organized as follows: in section 2 some preliminary concepts of graph theory on which we operate and briefly review the various centrality measures. In section 3 the methodology for construction of graph from amino acids is given and in section 4 different centralities measures of Amino Acids Graph is discussed. In section 5 correlation between different centralities are discussed and in section 6 we give the conclusion of this paper.

2 Basic Concepts of Graph

An undirected graph $G = (V, E)$ consists of a finite set V of vertices and a finite set E of edges (Zhang, 2016, 2018). If an edge $e = (u, v)$ connects two vertices u and v then vertices u and v are said to be incident with the edge e and adjacent to each other. The set of all vertices which are adjacent to u is called the neighborhood $N(u)$ of u. A directed graph or digraph G, if each edge of the graph has a direction. A graph is called loop-free if no edge connects a vertex to itself. An adjacency matrix A of a graph $G = (V, E)$ is a $(n \times n)$ matrix, where $a_{ij} = 1$ if and only if $(i, j) \in E$ and $a_{ij} = 0$ otherwise. The adjacency matrix of any undirected graph is symmetric. The degree, of a vertex v is defined to be the number of edges having v as one of its end point. A walk is defined as a finite alternating sequence of vertices and edges, beginning and ending with vertices, such that each edge is incident with the vertices preceding and following it. No edges appear more than once in a walk.

A vertex however may appear more than once. In a walk starting and end vertices are initial and terminal vertices. A walk is closed if the initial and terminal vertices coincide and open otherwise. A trail is a walk without repeated edges and path is a walk without repeated vertices.

A shortest or geodesic path between two vertices u, v is a path with minimal length. A graph is connected if there exists a walk between every pair of its vertices.

In graph theory, centrality measure of a vertex represents its relative importance within the graph (Zhang, 2018, 2021, 2023). A centrality is a real-valued function on the nodes of a graph. More formally a centrality is a function f which assigns every vertex $v \in V$ of a given graph G a value $f(v) \in R$. In the following we have discussed four most commonly used centrality measures.

2.1 Degree of centrality

The most simple centrality measure is degree of centrality, $c_d(u)$. It is defined as the number of nodes to which the node u is directly connected (Zhang, 2016, 2018, 2021, 2023). The nodes directly connected to a given node u are also called first neighbors of the given node. Degree Centrality shows that an important node is involved in a large number of interactions. This interaction gives the immediate importance or risk of the node in the corresponding network. It is defined as

$$c_d(u)=\deg(u).$$

However in real world problem the degree of centrality is not an actual measurement for finding importance or risk of a node. In real situation an important node may be connected indirectly with other nodes.

2.2 Eigenvector centrality

Another important measure of centrality is eigenvector centrality (Bonacich, 1972; Xin and Zhang, 2021; Yang and Zhang, 2022). An eigenvalue of a square matrix A is a value λ for which $det(A-\lambda I) = 0$, where I is the identity matrix of same order as A. Eigenvector centrality is defined as the principal eigenvector of the adjacency matrix of corresponding graph. In matrix-vector notation we can write

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

where $\boldsymbol{A}$ is the adjacency matrix of the graph, λ is a constant (the eigenvalue), and $\boldsymbol{X}$ is the eigenvector. In general, there will be different eigenvalues λ for which an eigenvector solution exists. However eigenvector of the greatest eigenvalue is the eigenvector centrality. Eigenvector centrality of vertices gives the contribution of the neighbour's and neighbour's of neighbour's and so on.

2.3 Closeness centrality

One of the three traditional centralities at the node or vertex level is closeness centrality, the other two are betweenness centrality and degree centrality (Rolito 2021; Zhang, 2018, 2021, 2023). Generally closeness centrality is defined as the inverse of the sum of the shortest path between each node to every other node in the network. The more central the vertex is, the more close to all other vertices of G. It is an idea how a vertex is close to all other vertices not only the first neighbour but also in global scale (Freeman, 1978). The closeness centrality of a node depicts an important node that can easily reach or communicate with other nodes of the network. It is defined as:

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

where, n is the number of vertices of the network and $d(u,v)$ is the shortest path distance between the pair of vertices u and v. From the above definition it is clear that if a node has minimum cumulative shortest path distance than that node has maximum closeness centrality. And maximum closeness centrality node is very well connected to all other nodes.

2.4 Betweenness centrality

The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex (Zhang, 2018, 2021, 2023). The betweenness centrality finds wide application in network theory: it represents the degree of which nodes stand between each other. The betweenness centrality interactions between two non adjacent nodes depend on the other node (Freeman, 1978), generally on those on the paths between the two. The betweenness centrality of a node u is the number of shortest path going through u. Mathematically it is defined as:

$$C_{btw}(u) = \sum_{s \neq u \in V} \sum_{t \neq u \in V} \frac{\sigma_{st}(u)}{\sigma_{st}}$$

where, σ_{st} is the number of shortest paths from vertex s to t, and $\sigma_{st}(u)$ is the number of shortest paths from s to t that pass through u. Betweenness centrality depicts identifying nodes that make the most information flow of the network. An important node will lie on a large number of paths between other nodes in the network. From this node we can control the information of the network. Without these nodes, there would be no way for two neighbors to communicate with each other. In general the high degree node has high betweenness centrality because many of the shortest paths may passes through them. However a high betweenness centrality node need not always be high degree node.

3 Methodology for Construction of Graph From Amino Acids

In this manuscript, we have proposed an algorithm for construction of amino acid graph based on physical properties. The organic compound which is composed of amine ($-NH_2$) and carboxylic acid (-COOH) are amino acids. Each amino acid expresses different properties which are changes with the change in R groups. The R group is often referred to as the amino acid side chain. Urbina et al. (2006) developed a measure of similarity between amino acids based on eight physical properties of amino acids. These properties are volume (Creighton, 1993), bulkiness, polarity and isoelectric point (Zimmerman et al., 1968), two different measures of hydrophobicity (Kyte and Doolittle, 1982; Engelman et al., 1986), surface area accessible to water (Miller et al., 1987), and fraction of accessible area lost when a protein folds (Rose et al., 1985). Based on these eight properties, we have done network analysis by using centrality measures to find out the relative importance between them. Further we calculate the correlation of centrality measures to give conclusion on basis of the results.

The similarity properties of amino acids in numerical values are shown in Table 1 (Urbina et al., 2006). The table shows a distance matrix which is symmetric in nature. From this distance matrix, there are 210 data points, we try to build a graph of the amino acids. Calculating the 210 data points we get the mean value is 3.461. Considering this mean as a threshold value which indicates that the data points have a tendency to cluster around that amino acid to define link between pair of amino acids. If the distance between two amino acid is less than or equal to 3.461 then linked by an edge.

Table 1 Distance between pair of amino acids based on physical properties.

	F	L	I	M	V	S	P	T	A	Y	H	Q	N	K	D	E	C	W	R	G
F	0	1.16	1.189	1.031	1.855	4.544	3.675	3.166	4.11	2.062	4.093	3.889	4.402	5.888	5.758	5.171	2.904	1.982	6.269	5.958
L	1.16	0	0.368	1.435	0.831	4.189	3.209	2.77	3.562	2.543	4.168	3.892	4.251	5.819	5.662	5.202	2.554	2.801	6.354	5.595
I	1.189	0.368	0	1.565	0.873	4.42	3.531	3.064	3.721	2.805	4.367	4.214	4.552	6.081	5.912	5.463	2.592	2.935	6.559	5.744
M	1.031	1.435	1.565	0	1.879	3.618	3.014	2.404	3.263	1.853	3.513	3.243	3.624	5.456	5.152	4.657	2.217	2.44	5.905	4.995
V	1.855	0.831	0.873	1.879	0	3.936	3.075	2.627	3.125	3.1	4.346	4.074	4.245	6.076	5.616	5.296	2.168	3.532	6.672	5.22
S	4.544	4.189	4.42	3.618	3.936	0	1.986	1.697	1.295	4.033	3.9	2.764	1.993	5.291	3.768	3.9	2.758	5.495	6.229	1.908
P	3.675	3.209	3.531	3.014	3.075	1.986	0	0.869	2.276	2.909	3.357	1.882	1.759	4.375	3.864	3.688	3.047	4.292	5.425	3.787
T	3.166	2.77	3.064	2.404	2.627	1.697	0.869	0	1.795	2.691	3.308	2.07	1.898	4.818	3.858	3.701	2.279	4.045	5.74	3.467
A	4.11	3.562	3.721	3.263	3.125	1.295	2.276	1.795	0	4.177	4.199	3.489	2.892	5.847	4.471	4.661	1.919	5.441	6.687	2.163
Y	2.062	2.543	2.805	1.853	3.1	4.033	2.909	2.691	4.177	0	3.105	2.24	3.077	4.398	4.633	3.882	3.591	1.643	4.837	5.704
H	4.093	4.168	4.367	3.513	4.346	3.9	3.357	3.308	4.199	3.105	0	2.912	3.164	3.027	3.692	3.134	4.094	4.033	3.352	5.115
Q	3.889	3.892	4.214	3.243	4.074	2.764	1.882	2.07	3.489	2.24	2.912	0	1.09	3.595	3.225	2.743	3.868	3.835	4.462	4.496
N	4.402	4.251	4.552	3.624	4.245	1.993	1.759	1.898	2.892	3.077	3.164	1.09	0	4.074	2.825	2.738	3.592	4.686	4.961	3.593
K	5.888	5.819	6.081	5.456	6.076	5.291	4.375	4.818	5.847	4.398	3.027	3.595	4.074	0	4.718	4.075	6.352	5.343	1.694	6.646
D	5.758	5.662	5.912	5.152	5.616	3.768	3.864	3.858	4.471	4.633	3.692	3.225	2.825	4.718	0	1.109	4.849	6.01	5.541	4.827
E	5.171	5.202	5.463	4.657	5.296	3.9	3.688	3.701	4.661	3.882	3.134	2.743	2.738	4.075	1.109	0	4.835	5.172	4.87	5.328
C	2.904	2.554	2.592	2.217	2.168	2.758	3.047	2.279	1.919	3.591	4.094	3.868	3.592	6.352	4.849	4.835	0	4.476	6.878	3.536
W	1.982	2.801	2.935	2.44	3.532	5.495	4.292	4.045	5.441	1.643	4.033	3.835	4.686	5.343	6.01	5.172	4.476	0	5.516	7.008
R	6.269	6.354	6.559	5.905	6.672	6.229	5.425	5.74	6.687	4.837	3.352	4.462	4.961	1.694	5.541	4.87	6.878	5.516	0	7.415
G	5.958	5.595	5.744	4.995	5.22	1.908	3.787	3.467	2.163	5.704	5.115	4.496	3.593	6.646	4.827	5.328	3.536	7.008	7.415	0

The symbols and notations used in the algorithm are described in the following Table 2.

Table 2 Symbols and notations.

Symbols	Description
A_i	Amino acids in row where *i*, goes to 1->20
A_j	Amino acids in column where *j* goes to 1->20
thr	Threshold value(mean value)
$E(A_{ij})$	An edge between two amino acids
A_G	Amino acid graph

Algorithm:

Input: A_i , A_j

Output: A_G

Step1. Start

Step2. Read the input values A_i and A_j

Step3. Read thr=3.416

Step4. For each A_i , i =1, i ->1-20

For each A_j , j=1, j ->1-20

If (A[ij]<=thr)

A[ij]=$E(A_{ij})$

A[ij]+

Else

A[ij]=0

End

Step5. For each A_i and A_j

If A[ij]=$E(A_{ij})$

$E(A_{ij})$==1

Else

$E(A_{ij})$=0

End

Step6. stop

The algorithm we have proposed is explained briefly in details as follows:

Step 1 to Step 4-

- We have put 20 amino acids in row as A_i and column as A_j.
- Taking the distance between similarity property for each pair of amino acids.
- The mean value of the above table i.e. 3.416 as threshold value

- We check each A_{ij} <=threshold then there will be a connection or edge between them.
- Now, we got the graph of amino acid based on distance between similarity properties.

The corresponding graph is depicted in Fig. 1.

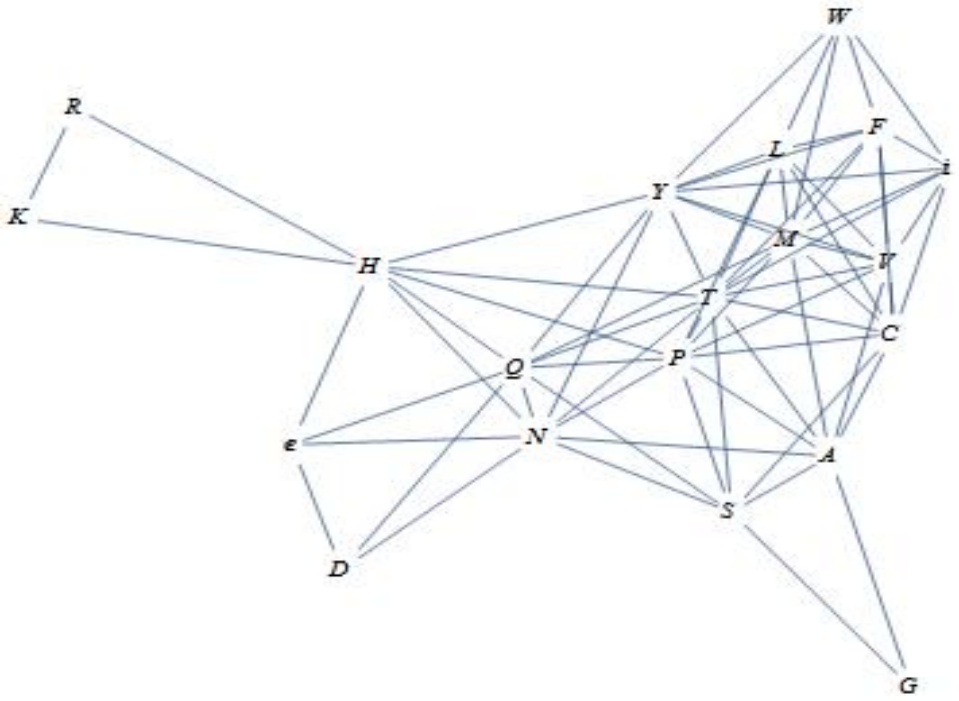

Fig. 1 The graph of amino acids.

Given that the matrix's distance between two amino acids is determined by how comparable their properties are, we are able to determine from the graph that two amino acids are compatible if they are connected. One amino acid is likely to evolve from the other if there is an edge connecting them. Therefore, we can claim that the network roughly illustrates the amino acid's evolutionary process.

Step5-To calculate the adjacency matrix of the above network

- We have check that each pair of the amino acids in the network have an edge or not
- If there is an edge then value set as 1 otherwise it set as 0.
- Now, we got the adjacency matrix of the network.

The corresponding adjacency matrix of the graph is given as the following matrix. As can be seen from the adjacency matrix bellow, it is clear that the graph (Fig. 1) is connected, then no row or column in the top (or lower) triangular matrix is zero.

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

4 Different Centralities in Amino Acids Network

To analyze the amino acids network (Fig. 1) we have calculated different measures of centrality by using python, taking the values in Table 3.

Table 3 Different centrality values for the amino acids.

Vertex	Degree Centrality (C_d)	Closeness centrality (C_{cl})	Betweenness centrality (C_{bwt})	Eigenvector centrality (C_λ)
F	8	0.542	1.092	0.248
L	8	0.542	1.859	0.256
I	7	0.527	0.950	0.222
M	11	0.655	10.006	0.325
V	9	0.575	2.453	0.286
T	13	0.760	22.327	0.369
Y	10	0.655	18.125	0.283
C	9	0.575	4.259	0.277
W	5	0.487	0.142	0.153
P	10	0.678	11.865	0.292
A	8	0.593	11.139	0.229
Q	9	0.655	16.206	0.227
S	7	0.558	8.667	0.190
N	9	0.655	16.166	0.217
G	2	0.395	0.000	0.048
H	8	0.612	36.737	0.173
K	2	0.395	0.000	0.022
E	4	0.463	1.000	0.077
R	2	0.395	0.000	0.022
D	3	0.431	0.000	0.059

Here, we've built a network of the 20 amino acids according to eight of their physical characteristics. From Table 3, we found that T receives the highest rank from degree, closeness, and eigenvector centralities, while H receives the highest rank from betweenness centrality.

The number of amino acid which are immediate predecessors or successor of another amino acid say x is the degree of centrality. The number of X's first neighbors determines it. The amino acid T, for instance, has a degree of centrality of 13. Consequently, T is most likely a direct predecessor or successor of thirteen amino acids (A, Y, H, Q, N, C, F, L, I, M, V, S, and P) in the evolutionary process.

The centrality of the eigenvector vector is better preserved and more recognizable than the degree of centralities in the same areas. A recursive form of degree of centrality is the eigenvector centrality. It is considered large for a node, if it has significant neighbors or a high number of neighbors. According to Chakrabarty and Parekh (2014), it preserves not only a node's connectivity but also that of its neighbors, neighbors' neighbors, and so forth. For the amino acid T, the eigenvector centrality is greatest. As such, the contribution of this amino is same with contribution of neighbour's and neighbour's to neighbour's, and so on. Otherwise, the evolutionary pattern of this amino acid is the same in its neighbour's and its neighbour's neighbour's, and so on.

Closeness centrality of an amino acid say, P gives the number of amino acid which may or may not be immediate of P and also predecessor or successor of P.

Increased closeness centrality values for amino acids indicate that evolutionary process is easily communicated between other amino acids. As an example, the closeness centrality values of the amino acids T and P are 0.760 and 0.678, respectively. Therefore, we might conclude that T facilitates easier communication of the evolutionary process than P does; that is, more amino acids precede or succeed T than P in the evolutionary process.

From betweenness centrality indication of an amino acid's contribution to the process of conveying the evolutionary process can be determined. More greater value of betwenness centrality of an amino acid, say, Q means more pairs of amino acids are connected by evolutionary process through it. According to the tabular results, the betweenness centrality of the amino acids H and E is 36.737 and 1, respectively. Therefore, amino acid H emerges as an intermediary between more pairs of amino acids than amino acid E, meaning that more pairs of amino acids interact by evolutionary process via H than through E.

5 Correlation Between Different Centralities

In this section, we have compared the different centrality measures that were discussed in the previous section. For this we have discussed the bivariate correlation of various measures of centrality of the amino acids network. Correlation is the most important character to study assortative or disassortative networks. A network is called assortative if the vertices with higher degree have the tendency to connect with other vertices that also have high degree of connectivity. If the vertices with high degree have the tendency to connect with other vertices with low degree then the network is called disassortative (Newman, 2002). The correlation coefficients for all the centrality measures are shown in Table 4. All correlation coefficients(r) are based on Pearsons's method. The range of r-value lies between +1 and -1. If $r > 0$ then the network is assortative whereas if $r < 0$ then the network is disassortative and $r=1$, then it is a perfect assortative mixing pattern.

Table 4 shows a strong correlation between the eigenvector centrality (C_λ), closeness centrality (C_{cl}), and degree of centrality (C_d), whereas there is no significant correlation seen betweenness centrality (C_{bwt}) and any of the other three variables. Assortative networks are recognized to facilitate information transfer more easily than disassortative networks (Newman, 2002). Based on the correlation coefficients table above, we can see that each pair of centrality measures has a positive correlation coefficient ($r>0$), indicating that the network is

of the assortative type and that the evolutionary flow of data will be easy.

Table 4 Correlation coefficients for the centrality measures.

	C_d	C_{cl}	C_{bwt}	C_λ
C_d	1	0.967	0.598	0.974
C_{cl}	0.967	1	0.724	0.895
C_{bwt}	0.598	0.724	1	0.430
C_λ	0.974	0.895	0.430	1

It has been observed that there is no overall relationship between property similarity and codon similarity as stated in this paper. However, for most amino acids (55% of total amino acids), property similarity correlates to codon similarity. We can therefore draw the conclusion that an evolutionary process based on codon similarity will present a different picture than one that is explained in relation to property similarity. Deciding which represents the true picture is not that simple. All that we have attempted to do is analyze the evolutionary process through property similarity.

6 Conclusion

In this manuscript we have attempt to describe the evolution of amino acids by considering their property similarity of the amino acids. It is possible to study the evolutionary process of amino acids using different concepts, such as structural similarity and codon similarity etc. We created a network structure of amino acids, through the definition of a compatibility relation based on the similarity distance of amino acids. After discussing various centrality metrics, we found that, with the hydrophilic amino acid T (threonine) has the highest centrality values except betweenness centrality, while hydrophilic amino acid H (histidine) has the highest centrality value in betweenness centrality. Thus, we can draw the conclusion that the hydrophilic amino acids H (histidine) and T (threonine) have been essential to the evolution of amino acids.

Next, we discussed correlation coefficients of different centrality measures of amino acids. We noticed that, all centrality measures show strong correlations except betweenness centrality (Cbwt). Thus, in the analysis of the amino acid network based on property similarity, we can conclude that betweenness centrality must be explored independently of the other centrality measures and studied separately. Also, our network is an assortative one since the correlation coefficient is positive for each pair of centrality measures. The evolutionary messages would therefore be following seamlessly.

References

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

Ali T, Akhtar A, Gohain N. 2016. Analysis of amino acids network based on distance matrix. Physica A, DOI: 10.1016/j.physa.2016.01.074

Bonacich P. 1972. Factoring and weighting approaches to status scores and clique identification, Journal of Mathematical Sociology, 2: 113-120

Boruah BK, Ali T. 2023. Algebric Structures and distance based analysis of genetic code. Network Biology, 13(1): 17-36

Boruah BK, Ali T. 2022. Special identity subgraph in genetic code. Network Biology, 12(2): 45-63

Boruah BK, Ali T. 2022. A study of the total graph in genetic code algebra, Network Biology, 12(1): 1-10

Boruah BK, Ali T. 2021. Analysis of Amino acids network based on transition and transversion mutation of

codons. Network Biology, 11(3): 125-136

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

Creighton TE. 1993. Proteins: Structures and Molecular Properties (2nd ed). W.H. Freeman, New York, USA

Engelman DA, Steitz TA, Goldman A. 1986. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annual Review of Biophysics and Biophysical Chemistry, 15: 321-353

Fell DA, Wagner A. 2000. The small world of metabolism. Nature Biotechnology, 18: 1121-1122

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

Gohain N, Ali T, Akhtar A. 2015. Lattice structure and distance matrix of genetic code. Journal of Biological Systems, 23: 485-504

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: 051903

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

Kyte J, Doolittle RF. 1982. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology, 157: 105-132

Miller S, Janin J, Lesk AM, Chothia C. 1987. Interior and surface of monomeric proteins. Journal of Molecular Biology, 196: 641-657.

Newman MEJ. 2002. Assortative mixing in networks. Physical Review Letter, 89: 2087011-2087014

Rolito G, Isagani S. 2021. Closeness centrality of some graph families. IJCMS, 16: 9160

Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. 1985. Hydrophobicity of amino acid residues in globular proteins. Science, 228: 834-838

Schreiber F, Koschutzki D. 2004: Comparison of centralites for biological networks. Proceedings of German Conference on Bioinformatics (GCB), P53 of LNI

Strub C, Alies C, Lougarre A, Ladurantie C, Czaplicki J, Fournier D. 2004. Mutation of exposed hydrophobic amino acids to arginine to increase protein stability. BMC Biochemistry, 5: 9

Taylor WR. 1986. The classification of amino acid conservation. Journal of Theoretical Biology, 119: 205-218

Urbina D, Bin T, Paul GH. 2006. The response of amino acid frequencies to directional mutation pressure in mitochondrial genome sequences is related to the physical properties of the amino acids and to the structure of the genetic code. Journal of Molecular Evolution, 62: 340-361

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

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

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

Yuan Q, Chen J, Zhao H, Zhou Y, Yang Y. 2022. Structure-aware protein–protein interaction site prediction using deep graph convolutional network. Bioinformatics, 38(1): 125-132

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

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

Zhang WJ. 2021. Construction and analysis of the word network based on the Random Reading Frame (RRF) method. Network Biology, 11(3): 154-193

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

Zimmerman JM, Eliezer N, Simha R. 1968. The characterization of amino acids sequences in proteins by statistical methods. Journal of Theoretical Biology, 21: 170-201