Article

Analysis of amino acid network based on bond energy using graph mining techniques

Nasrin Irshad Hussain[1], **Kuntala Boruah**[1], **Adil Akhtar**[2]

[1]Department of Computer Application, Sibsagar University, Assam-785665, India

[2]Department of Mathematics, Golaghat Engineering College, Golaghat, Assam-785621, India

E-mail: hussainnasrin531@gmail.com, kuntala17@gmail.com, adil.akhtar19@gmail.com

Received 6 August 2024; *Accepted 15 September 2024*; *Published online 20 October 2024*; *Published 1 March 2025*

Abstract

Amino acids are building blocks of proteins which are essential for all biological functions. Each amino acid has different physico-chemical characteristics. The interconnection between one amino acid with other amino acids in surrounding environment creates a network which is very complex in nature. Graph mining is most efficient and effective methods for analyzing various complex datasets. However, the knowledge discovery from those heterogeneous datasets is a nontrivial task. The computational method automate the prediction of patterns and less expensive than experimental methods in terms of resources and time. This paper presents amino acid network based on bond energy where different graph mining techniques are used to extract information from the network. Further, different centrality measures are calculated to analyze and determine the amino acids relative importance in this network. Also we have examined the clustering coefficient, correlation coefficient among different centrality measures, eccentricity and center of the network. Our results reveal significant insights into the evolutionary roles of amino acids, highlighting the centrality of hydrophilic amino acids and the unique positions of specific residues like Tryptophan and Glycine. These findings provide a deeper understanding of amino acid interactions and their evolutionary implications, demonstrating the effectiveness of computational methods in biological research.

Keywords bond energy; centrality measure; clustering coefficient; eccentricity.

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

1 Introduction

The living organisms made up of cells and each cell from chromosomes which are identical in each cell. The chromosomes are made up by genes which are blocks of DNA. These Genes are codes for proteins. Protein is made up through twenty amino acids. Each amino acid is expressed by a triplet code. A codon is a three-base sequence that makes up a unit. In the genetic code, the second base position in a codon is biologically most

important, while the important of third base position is least. The chain of amino acids takes on different shapes for different proteins. The DNA/RNA is the chemical in the cells of animals and plants that carries genetic information. The nucleotides are the basic building blocks of DNA/RNA. Each nucleotide is made of sugar group (deoxyribose), phosphate group, and nitrogenous base such as Adenine (A), Cytosine (C), Guanine (G), or Thymine (T); one from each group. The helical backbone is composed of sugar and phosphate groups. The purines (Adenine and Guanine) and pyrimidines (Cytosine and Thymine) make the double helix in DNA/RNA.The nucleotides are connected through hydrogen bonds.

The main goal of this work is to consider a network of different 20 essential amino acids on the basis of bond energy. In the area of interdisciplinary research, researchers made important contribution in the field of computational biology where complex biological networks are analyzed. Jiao et al. (2007) discussed the contact energy-based weighted amino acid network. Hussain and Boruah (2024) constructed a graph of amino acids based on property similarity and discussed different centrality measures to extract important amino acids which play an vital role in evolution of amino acids. They proved that a network of weighted amino acids can satisfy the “small-world property”. Aftabuddin and Kundu (2007) studied the hydrophobic, hydrophilic, and charged networks in protein. The hydrophobic network has the vertices which are hydrophobic amino acids, while the hydrophilic and charged network have hydrophilic and charged amino acids, respectively. The amino acids within a cutoff distance of $5A^{o}$ are considered. They observed that in the hydrophobic network the average degree is much higher than the average degree of hydrophilic and charged networks. The hydrophilic networks have a slightly higher average degree than the charged networks. The average strength of nodes in hydrophobic networks is almost comparable to that of charged networks, but hydrophilic networks have a lower average strength. Mukhopadhyay and Maulik (2014) discussed the viral-human protein interaction network and Chakrabarty and Parekh (2014) studied the protein network and discussed different centrality analysis of structural ankyrin repeats (ANK). They have considered amino acid network where alpha carbon atom as vertices of each amino acid and two vertices have an edge between them if their Euclidean distance between alpha carbon atoms within $7A^{0}$. According to Gohain et al. (2015), genetic code has a partial ordering and a lattice structure has been built from it. The structure of the genetic code and partial ordering in the base set are influenced by the codon-anticodon interaction, hydrogen bond number, and chemical classes of bases. There were some demonstrated connection in between the lattice structure of the genetic code and the physico-chemical properties of amino acids. Ali et al. (2016) developed a distance matrix of amino acids. The range is determined by the relative evolutionary relevance of the respective codons' nucleotide positions. A network of amino acids is constructed by using distance matrix. They propose that this network shows the amino acids evolutionary pattern. They also focused about the relative relevance of amino acids in this network. Bora et al. (2020) discussed structural similarity of amino acids (Ali and Borah, 2021) constructed a network on mutation of purine and pyrimidine. Jin et al. (2021) discussed deep learning on biological networks and give some future development prospects in this field. Calderer and Kuijjer (2021) discussed a bipartite graph in biological network. They have discussed the communities detect in bipartite biological networks. Ali Akhtar (2022) discussed balanced and unbalanced network in amino acids. In Kornev et al. (2022, some network parameters are studied to explain amino acid networks.

This paper is arranged as follows: in section 2 discuss the concepts and techniques of graph and graph mining used in this paper. In section 3 construction of amino acid network and in section 4 analysis of amino acid network are discussed. In section 5 the conclusion of this paper is given.

1 Concepts of Graph and Graph Mining Techniques

An undirected graph *G* has no orientation which consists of a finite sets of vertices (also called nodes) and

edges (also called links). Two nodes u and v are said to be adjacent to each other if they have an edge connecting the two vertices. The neighbourhood of u is the collection of all nodes that are adjacent to u. The number of edges at a vertex v is its degree with v as one of their endpoints. A directed graph is a graph where each path has a direction. If no path links a node to itself, then the graph is called loop-free. An adjacency matrix of a graph G, is a matrix A, when $a_{ij} = 1$ if $(i, j) \in E$ and $a_{ij} = 0$ or else. For any undirected graph, the adjacency matrix of is symmetric. A finite alternating pattern of nodes and edges is called walk that starts and ends with nodes. In a walk, no edges present more than once. Deo (2002) stated that if a walk exists between each pair of its nodes then the graph is connected.

1.1 Centrality measures

The centrality of a vertex in a network reflects the relative significance of a vertex inside the network (Zhang, 2016, 2018, 2023). A centrality measure is a function $f(v)$ where $v \in V$ is a vertex of a given network, and $f(v)$ is a real value. The four most frequently used centrality measurements are discussed in the following.

1.1.1 Degree of centrality

The degree of centrality $c_d(u)$, is the simplest basic centrality of the vertex (u) and defined as the number of vertices that are directly connected to vertex u (Bonacich, 1972; Zhang, 2016, 2018, 2023). It is mathematically defined as:

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

Degree of centrality indicates the immediate importance or relative risk of the vertex in the corresponding network. However in real world problem the degree of centrality is not a reliable indicator for finding importance or risk of a vertex. An important node may have indirect connection with other vertex.

1.1.2 Eigenvector centrality

Eigenvector is another important centrality (Bonacich, 1972). The eigenvector is $det(A\text{-}I) = 0$ for square matrix A, where I represents the identity matrix which is same order as A. In the associated network, the highest value of the eigenvector in the adjacency matrix is the eigenvector centrality. The matrix-vector notation is shown as below

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

The Eigenvector centrality may give the contribution of the vertices in respect of the neighbours and neighbours of neighbours, and so on.

1.1.3 Closeness centrality

Closeness centrality refers to a vertex's near to all other vertices, not just its immediate neighbours, but also on a global scale (Freeman, 1978; Zhang, 2016, 2018, 2023). When a vertex is central, it is in close proximity to all remaining vertices. If a vertex is adjacent to other vertices, it will be able to interact with them fast. It is mathematically defined as:

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

Here, n represent number of nodes inside the network and $d(u,v)$ represent between the nodes u and v, the shortest path distance.

1.1.4 Betweenness centrality

Betweenness centrality refers to the process of determining which vertices in a network pass the most information flow (Zhang, 2016, 2018, 2023). The betweenness centrality between two non-adjacent vertices is

controlled by the other vertex. How many shortest paths are passing through a vertex *u* is its betweenness centrality (Freeman, 1978). It is mathematically 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} denotes the number of shortest pathways between vertex *s* and vertex *t*, and $\sigma_{st}(u)$ denotes the shortest paths between *s* and *t* that pass through *u*. An important vertex or node will be connected to other nodes in the network through a large number of paths. We can manage the network's information from this node. Without these nodes, two neighbors would not be able to communicate with one another.

1.2 Eccentricity

The eccentricity of a vertex *v* in a network is calculated by computing the shortest path distance between *v* to all other vertices in a network. Then the value of the longest shortest path is called eccentricity of a vertex in a network. If the length of a vertex *v* to any node *w* is *dist (v, w)*, then the eccentricity $C_e(v)$ of a node $v \in V$ is defined as:

$$C_e(v) = \max\{dist\ (v,w) |\ w \in V\}$$

1.3 Clustering coefficient

In graph theory, a clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. The Clustering coefficient of a vertex is high that represents strong relationship in between neighbouring vertices. The clustering coefficient C_i of a vertex '*i*' is the ratio between the total number (e_i) of links that are connected to nearest neighbours and the total number of all possible links to neighbouring vertices.

Here, the total number of links (e_i) which connecting its nearest neighbours $e_i = \frac{K_i(K_i-1)}{2}$, where K_i is the degree of node '*i*'. It is given by $C_i = \frac{2e_i}{K_i(K_i-1)}$ where $0 \leq C_i \leq 1$. Nodes with less than two neighbors are assumed to have a clustering coefficient of 0.

2 Bond Energy of Amino Acids

Bond energy can be defined as gross amount of energy needed to break one mole of a specific bond. The bond energy of amino acids depends on the type of bond and the amino acid. The following tables of bond energy values are calculated using *Gaussian* software. Amino acids are distinguished by the attached functional group R. The difference of the bond energies between any two amino acids are calculated and shown in the Table 1.

Table 1 Distance between pair of amino acids based on bond energy.

	G	A	V	L	M	I	S	T	C	P	N	Q	F	Y	W	K	H	D	R	E
G	0	19.799	59.824	79.268	104.727	79.273	42.534	62.299	65.125	56.569	80.767	100.539	121.517	150.008	180.208	100.536	110.533	82.313	140.262	100.191
A	19.799	0	40.026	59.47	84.929	59.476	22.828	42.534	45.327	36.577	61.019	80.772	100.752	130.214	160.914	80.751	93.464	62.365	120.405	82.135
V	59.824	40.026	0	19.445	44.90	19.466	17.866	3.879	5.319	3.303	21.439	40.790	67.453	90.220	120.926	40.799	53.448	22.595	80.757	42.121

	4	6			8	1					2	6		7	88	4	3	3	7	5
L	79.268	59.47	19.45	0	25.461	0.507	36.964	17.307	14.145	22.705	5.199	21.737	48.083	70.775	103.242	21.463	34.09	4.986	61.23	22.941
M	104.727	84.929	44.908	25.461	0	25.457	62.373	42.609	39.604	48.165	24.633	6.615	22.633	45.357	77.79	5.32	9.073	23.078	36.065	5.053
I	79.273	59.476	19.461	0.507	25.457	0	36.971	17.321	14.162	22.723	5.231	21.733	48.086	70.777	103.244	21.46	34.095	5.002	61.229	22.937
S	42.534	22.828	17.66	36.964	62.373	36.971	0	19.799	22.896	14.46	38.237	58.022	84.977	107.551	140.093	58.067	70.789	39.624	97.743	59.418
T	62.299	42.534	3.879	17.307	42.609	17.321	19.799	0	4.173	6.41	18.495	38.243	65.195	87.752	120.296	38.276	50.993	19.852	77.967	39.63
C	65.125	45.327	5.319	14.145	39.604	14.162	22.896	4.173	0	8.562	16.277	35.645	62.226	84.906	117.384	35.506	48.188	17.399	75.293	36.937
P	56.569	36.77	3.303	22.705	48.165	22.723	14.46	6.41	8.562	0	24.562	44.137	70.784	93.458	125.941	44.038	56.73	25.794	83.817	45.454
N	80.767	61.019	21.392	5.199	24.633	5.231	38.237	18.495	16.277	24.562	0	19.805	47.009	69.397	101.973	20.002	32.674	2	59.514	21.266
Q	100.539	80.772	40.906	21.737	6.615	21.733	58.022	38.243	35.645	44.137	19.805	0	27.41	49.612	82.195	2.245	13.003	18.446	39.734	2
F	127.351	107.552	67.53	48.083	22.633	48.086	84.977	65.195	62.226	70.784	47.009	27.41	0	22.827	55.163	27.069	14.497	45.516	14.66	25.867
Y	150.008	130.214	90.207	70.775	45.357	70.777	107.551	87.752	84.906	93.458	69.397	49.612	22.827	0	32.589	49.512	36.784	67.968	10.759	48.178
W	182.508	162.709	122.688	103.242	77.79	103.244	140.093	120.296	117.384	125.941	101.973	82.195	55.163	32.589	0	82.059	69.333	100.531	42.872	80.745
K	100.536	80.751	40.794	21.463	5.32	21.46	58.067	38.276	35.506	44.038	20.002	2.245	27.069	49.512	82.059	0	12.741	18.598	39.8	2.617
H	113.253	93.464	53.483	34.09	9.073	34.095	70.789	50.993	48.188	56.73	32.674	13.003	14.497	36.784	69.333	12.741	0	31.264	27.176	11.613
D	82.	62.	22.	4.9	23.	5.0	39.	19.	17.	25.	2	18.	45.	67.	10	18.	31.	0	58.	19.

	13	365	593	86	078	2	624	852	399	794		446	516	968	0.531	598	264		169	805
R	140.262	120.493	80.577	61.23	36.065	61.229	97.743	77.967	75.293	83.817	59.514	39.734	14.6	10.759	42.872	39.8	27.176	58.169	0	38.427
E	101.912	82.135	42.215	22.941	5.053	22.937	59.418	39.63	36.937	45.454	21.266	2	25.867	48.178	80.745	2.617	11.613	19.805	38.427	0

There are 210 data in above distance matrix because it is symmetric. We generate an amino acids network residue using this data. When we have looked the mean value of the 210 dataset and get 45.31 as threshold value. We consider this mean value (i.e., 45.31) as a cutoff level to indicate the link between different amino acids because mean denotes that data points tend to cluster around it.

3 Method

In this part, we create an amino acids network based on bond energy. We proposed an algorithm to create a network from the above table 1. As threshold value is set to 45.31 then if the distance between two amino acids is less than or equal to ($\leq$) 45.31, then they are connected by an edge otherwise no edge between them.

Symbols and notations used here are as follows:

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

Algorithm for construction of amino acid network is

Algorithm-1 Construction of Amino acids network based on bond energy:

Input: Give the values of A_i A_j from the Table 1
Output: Returning Amino acid network, A_G

Step1. Start

Step2. Read the input values A_i and A_j where i,j as variable with initial value 1

Step3. Declare ***Thr*** as threshold value with initial value 45.31

Step4. For all *(i,j)* less then equal to 20

And *A[ij] <=thr then*

A[ij] =E(A_{ij}) an edge is drawn

Step5. Increase *A[ij]* by 1

Step6. Go to Step4 until *(i,j)* equal to 20

Step7. Return A_G

Step8. Stop

Pseudocode 1: Construction of Amino acid network

```
1: Input as A_i A_j
Read thr=45.31
2: for each A_i, i =1, i ->1-20
3: for each A_j, j=1, j ->1-20
4:      if (A[ij]<=thr) then
5:           A[ij]=E(A_ij)
6:           A[ij]+
7:      else
8:           A[ij]=0
9:      end if
10: for each A_i and A_j
11:     if A[ij]=E(A_ij)
12:          E(A_ij)==1
13:     else
14:          E(A_ij)=0
15:     end if
16: return
```

Output of the algorithm: the network of amino acid as shown below:

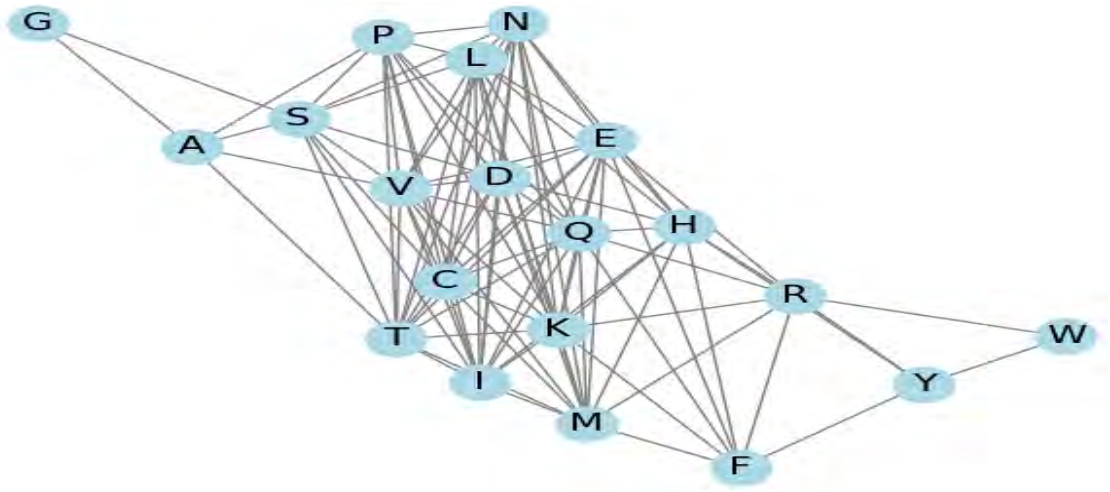

Fig. 1 Amino acid network based on bond energy.

The corresponding adjacency matrix of the network is given below:

	G	A	V	L	M	I	S	T	C	P	N	Q	F	Y	W	K	H	D	R	E
G	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
A	1	0	1	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0
V	0	1	0	1	1	1	1	1	1	1	1	1	0	0	0	1	0	1	0	1
L	0	0	1	0	1	1	1	1	1	1	1	1	0	0	0	1	1	1	0	1
M	0	0	1	1	0	1	0	1	1	0	1	1	0	0	0	1	1	1	1	1
I	0	0	1	1	1	0	1	1	1	1	1	1	0	0	0	1	1	1	0	1
S	1	1	1	1	0	1	0	1	1	1	1	0	0	0	0	0	0	1	0	0
T	0	1	1	1	1	1	1	0	1	1	1	1	0	0	0	1	0	1	0	1
C	0	0	1	1	1	1	1	1	0	1	1	1	0	0	0	1	0	1	0	1
P	0	1	1	1	0	1	1	1	1	0	1	1	0	0	0	1	0	1	0	0
N	0	0	1	1	1	1	1	1	1	1	0	0	0	0	0	1	1	1	0	1
Q	0	0	1	1	1	1	0	1	1	1	1	0	1	0	0	1	1	1	1	1
F	0	0	0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	0	1	1
Y	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	1	0
W	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
K	0	0	1	1	1	1	0	1	1	1	1	1	1	0	0	0	1	1	1	1
H	0	0	0	1	1	1	0	0	0	0	1	1	1	1	0	1	0	1	1	1
D	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1
R	0	0	0	0	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	1
E	0	0	1	1	1	1	0	1	1	0	1	1	1	0	0	1	1	1	1	0

4 Analysis of Network of Amino Acids

This section gives the analysis of amino acid network by using the following network parameters. Based on bond energy of 20 essencial amino acids, we built a network. Now we are going to analyze this network using some graph mining techniques as follows:

4.1 Centrality measures

Different centrality (Fig. 1) such as degree, closeness, betweenness and eigenvector centralities are calculated to analyzed the amino acid network using python programming language are shown in the following Table 2.

Table 2 Various amino acids' centrality scores.

Amino Acid	Degree Centrality (C_d)	Closeness centrality (C_{cl})	Betweenness centrality (C_{bwt})	Eigenvector centrality (C_e)
G	2	0.380	0.000	0.022
A	5	0.475	0.018	0.083
S	10	0.575	0.089	0.189
V	13	0.703	0.042	0.272
T	13	0.703	0.042	0.272
P	11	0.655	0.023	0.232
L	13	0.730	0.023	0.282
M	13	0.730	0.043	0.268
I	13	0.730	0.023	0.282
C	12	0.678	0.010	0.266
N	12	0.730	0.023	0.282
Q	14	0.760	0.060	0.286
K	14	0.760	0.060	0.286
D	13	0.730	0.023	0.263
E	12	0.730	0.043	0.248
H	11	0.655	0.080	0.208
F	7	0.542	0.014	0.122
R	8	0.558	0.103	0.123
Y	4	0.441	0.010	0.039
W	2	0.372	0.000	0.013

From the above Table 2, the centrality measures of amino acids are mentioned as first column shows degree centrality, here we observed that Q and K, has highest score. Like degree centrality, closeness centrality, and eigenvector centrality have same results, whereas R has the highest score betweenness centrality measures. The plots for degree, closeness, betweenness and eigenvector centrality measures in the following Fig. 2.

The number of residues of amino acids that are probably near to an amino acids predecessors or successors, say X, is determined by its degree of centrality. In particular, the amino acid residue G has 2 degree of centrality. As a result, in the evolutionary process, G is most likely the immediate predecessor or successor of two different amino acids viz., A and S.

The eigenvector centrality of a vertex is high if it has several neighbours and/or significant neighbours. The connectivity of a vertex depends on, neighbour's and their neighbours's of neighbour's so on (Chakrabarty and Parekh, 2014) which give the eigenvector value. The amino acids Q and K have the highest eigenvector centrality so the contribution of these amino acids in neighbours and their neighbours' of neighbours in the

progress of evolution is high.

In the network of amino acid, the greater value of the closeness centrality implies that it is smoothly interacts with several other residues of amino acids throughout the evolution; which may or may not be immediate predecessors or successors. In Table1, amino acids residue K and G have the closeness centrality value 0.760 and 0.380 respectively where K has the higher score than G. As a result, we can conclude that K is easily communicated than G when it comes to the evolutionary process.

The greater betweenness centrality value of amino acids indicates that through the evolutionary process, more pairings of amino acid residues are connected. From Table 2, we observed that the betweenness centrality value of amino acid residues R and S have 0.103 and 0.089 respectively. These values indicate that amino acids R and S have more pairings of amino acid in the evolutionary of amino acids than other residues in this particular network.

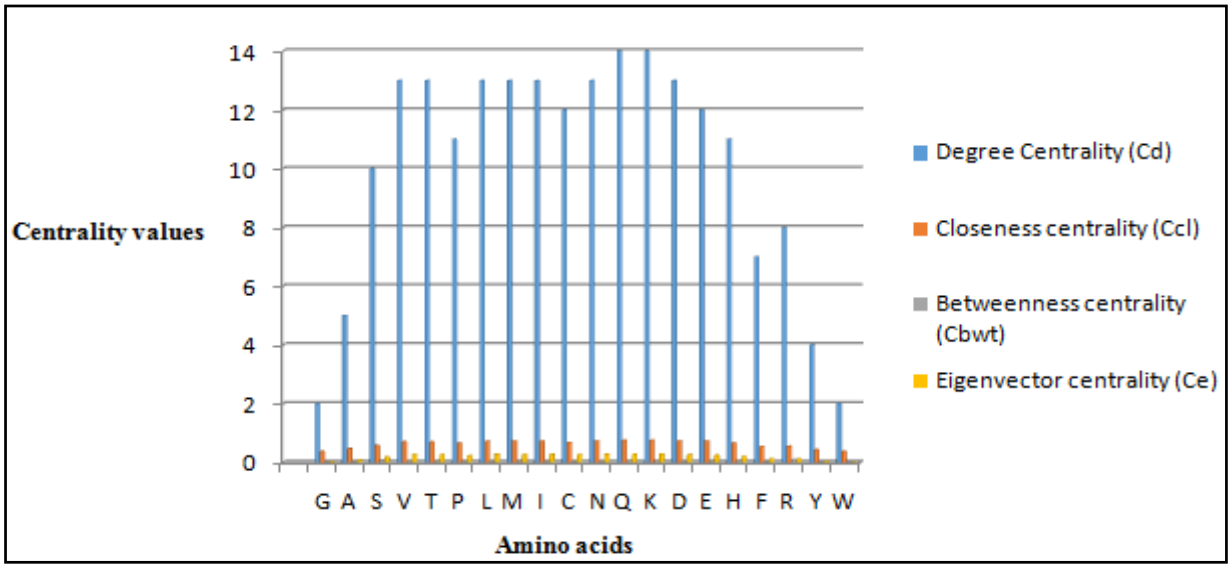

Fig. 2 Bar diagram for centrality measures.

4.2 Correlation coefficient between different centralities

Here, we compare various centralities which have been studied in earlier part. For that, we studied bivariate correlation between several measures of centrality. The most crucial characteristic when studying assortative network or disassortative network is correlation. “If the nodes with a higher degree of connectivity have a tendency to link with other nodes with a high degree of connectivity, then the network is called assortative network. If high-degree nodes have a tendency to link with low-degree nodes, the network is said to be disassortative network” (Newman, 2002). In Table 3, shows correlation coefficients (r) for different centrality measures. The value of r in between +1 and -1. If the value of r is positive ($r > 0$), then that network is called assortative network and if the value of r is negative ($r < 0$), then that network is called disassortative network.

Table 3 Correlation coefficients for different centrality measures.

	C_d	C_{cl}	C_{bwt}	C_λ
C_d	1	0.991	0.377	0.992
C_{cl}	0.991	1	0.324	0.985
C_{bwt}	0.377	0.324	1	0.286
C_λ	0.992	0.985	0.286	1

The degree of centrality (C_d), closeness centrality (C_{cl}) and eigenvector centrality (C_λ) are all highly correlated; on the other hand betweenness centrality (C_{bwt}) is not highly correlated with any of the other measures as seen in Table 3. The information transferred can be easier in an assortative network than through a disassortative network (Newman, 2002). The correlation coefficient is positive (r>0) hence the network is assortative in nature, as a result flow of evolutionary information will be easy.

4.3 Eccentricity and clustering coefficient

In this part we have discussed two important network parameters; namely eccentricity, clustering coefficient which also play a vital role in information flow. The measure of eccentricity in the network gives how accessible a vertex from other vertices. For biological networks, eccentricity with lower value gives more significant than higher one and it presume a positive meaning in term of node proximity. If the eccentricity of the node v is low, this means that all other nodes in the network are nearby (Scardoni and Laudanna, 2012). In contrast, if the eccentricity is high, it means that there is at least one node (and all its neighbors) that is far from node v. Thus, eccentricity is a more meaningful parameter when the behaviour of the graph is analyzed.

An amino acid is functionally accessed by all other amino acids in the network can be used to determine a node's eccentricity in that particular network, such as an amino acid network. Therefore, the activity of neighbouring amino acids will have direct effect on an amino acid with a low eccentricity in this network. On the other hand, a high eccentricity may affect a minimal functional importance. The eccentricities of all the amino acids have been calculated by using python language and shown in Table 4.

Table 4 Eccentricity of the amino acids.

E	D	M	K	W	N	I	L	Q	V	A	T	C	R	P	F	Y	S	G	H
3	3	3	3	5	3	3	3	3	3	4	3	3	4	3	4	4	4	5	3

From Table 4 the average eccentricity value is 3.45, and then the hydrophilic amino acids R and S and the hydrophobic amino acids W, A, F, Y and G have higher eccentricities than the average eccentricity value, whereas W and G have the highest eccentricity value among all the amino acids in the table. Therefore amino acids W and G spread evolutionary information slower than the remaining eighteen amino acids.

Further, the center of a network is the collection of nodes whose eccentricity equals the radius, and the radius of a network is the least eccentricity over all of its nodes. Thus we can see from the aforementioned eccentricity value in Table 4 that amino acids E, D, M, K, N, I, L, Q, V, T, C, P and H represent the network's center as they have approximately equal to radius. So we conclude these amino acids in the center of a network have more capable to carry evolutionary information.

For a vertex the clustering coefficient quantifies how densely connected a network is around a particular node. High clustering coefficient of a vertex represents strong relationship in between neighbouring nodes. That means a node represents more number of connection among its neighbours. The clustering coefficient of a node gives a larger effect on the network (Dhriti and Sudip, 2012) as its high value indicates that the information flow is decreases. Again the clustering coefficient of a network is defined as the average value of the clustering coefficient of vertices in the network. In Table 5 we have shown clustering coefficients of all the amino acids by using python language.

Table 5 Clustering coefficient of 20 amino acids.

E	D	M	K	W	N	I	L	Q	V	A	T	C	R	P	F	Y	S	G	H
0.7	0.8	0.7	0.7	1	0.8	0.8	0.8	0.7	0.7	0.	0.7	0.8	0.6	0.8	0.8	0.6	0.7	1	0.6
7	6	6	4		4	4	4	4	9	7	9	9	4	3	0	6	1		9

We observed from the above mentioned Table 5 that the amino acid W (Tryptophan) and G (Glycine) have highest coefficient viz., 1. The clustering coefficient of the whole network is 0.79. Hence we can conclude that W and G flow of evolutionary process is relatively slower in comparison to the whole network. Also from betweenness centrality we got the value of W and G is zero which means there is no shortest path passing through it.

5 Conclusion

Based on the bond energy of the amino acids, we have explored the evolutionary process of amino acids in this work. The graphical structure of the amino acids, we have equipped the amino acids by defining a compatibility relation based on bond energy matrix of the amino acids. As we noticed via our analysis of several centrality measures hydrophilic amino acids Q and K have the highest centrality values among all the other centrality measures, with the exception of betweenness centrality. If we see the betweenness centrality hydrophilic amino acid S has highest centrality value. Thus, we can draw conclusion that the hydrophilic amino acids Q, K, and hydrophilic amino acid S plays a key role in the evolutionary process of amino acids.

Next we analyze the correlation coefficients of the different centrality measures of amino acids and as a result, it is observed that all centrality measures are highly correlated except betweenness centrality (C_{bwt}). There betweenness centrality is independent of the other centrality measures so needs to be explored separately.

Further from the center and eccentricity of the amino acids network ,we have observe that amino acids W & G is not at the center of the network so evolutionary information spread is relatively slower than all other amino acids. When we extract that the amino acids W and G have high clustering coefficient value, then we may conclude that in the neighbourhood of W and G flow of evolutionary process is relatively slow in comparison to the whole network.

Finally, from eccentricity and clustering coefficient values of the amino acid network, we can conclude that at center (E, D, M, K, N, I, L, Q, V, T, C, P, H) evolutionary process is faster in comparison to the whole network.

From the above discussion we have done extensive analysis of amino acid network at different centrality measures as well as network parameters base on bond energy. We have also suggested an algorithm and give an analysis framework which will give a theoretical significance only. This network can be construct by using random connection of different amino acids, by weighted network and also can be construct hydrophobic as

well as hydrophilic amino acids network etc. Further, it can be study how the crucial amino acids based on different centrality measures and network parameters are change as the protein structure shift from folded to unfolded state.

References

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

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

Ali T, Borah C. 2021. Analysis of amino acids network based on mutation and base positions. Gene Reports, 24: 101291.

Application of Algebraic Graph Theory in Genetic Code [WWW Document]. https://www.ijaresm.com/application-of-algebraic-graph-theory-in-genetic-code (accessed 10.16.24)

Bonacich P. 1972. Factoring and weighting approaches to status scores and clique identification. The 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

Borah C. 2023. Algebraic structures and distance based analysis of genetic code. Network Biology, 13(1): 17-36

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

Calderer G, Kuijjer ML. 2021. Community detection in large-scale bipartite biological networks. Frontiers in Genetics, 12: 649440

Chakrabarty B. Parekh N. 2014. Graph centrality analysis of structural ankyrin repeats. International Journal of Computer Information Systems and Industrial Management Applications, 6: 1-10

Dhriti S, Sudip K. 2012. Role of long-and short-range hydrophobic, hydrophilic and charged residues contact network in protein's structural organization. BMC Bioinformatics, 13: 142

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

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

Hussain NI, Boruah K. 2024. Analysis of amino acids network based on graph mining. Network Biology, 14(3): 242-253

Jiao X., Chang S, Li C, Chen W, Wang C. 2007. Construction and application of the weighted amino acid network based on energy. Physics Review E, 75: 051903. https://doi.org/10.1103/PhysRevE.75.051903

Jin S, Zeng X, Xia F, Huang W, Liu X. 2021. Application of deep learning methods in biological networks. Briefings in Bioinformatics, 22: 1902-1917

Kornev, A.P., Aoto, P.C., Taylor, S.S., 2022. Calculation of centralities in protein kinase A. Proceedings of the. National Academy of Sciences USA, 119: e2215420119. https://doi.org/10.1073/pnas.2215420119

Koschützki D, Schreiber F. 2004. Comparison of centralities for biological networks. German Conference on Bioinformatics 2004, GCB 2004. Gesellschaft für Informatik, Bonn, Germany

Kundu S. 2005. Amino acid network within protein. Physica A: Statistical Mechanics and its Applications, Statphys - Kolkata V: Proceedings of the International Conference on Statistical Physics: Complex Networks: Structure, Function and Processes, 346: 104-109

https://doi.org/10.1016/j.physa.2004.08.055

Mukhopadhyay A, Maulik U. 2014. Network-based study reveals potential infection pathways of hepatitis-C leading to various diseases. PLoS one, 9: e94029

Newman MEJ. 2002. Assortative mixing in networks. Physics Review Letter, 89: 208701. https://doi.org/10.1103/PhysRevLett.89.208701

Quan R, Wang W, Ma F, Fan H, Yang Y. 2024. Clustering for protein representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 319-329, CVF Open Access

Rachman A. Brennan MA, Morton J, Torrico D, Brennan CS. 2023. In-vitro digestibility, protein digestibility corrected amino acid, and sensory properties of banana-cassava gluten-free pasta with soy protein isolate and egg white protein addition. Food Science and Human Wellness, 12: 520-527 https://doi.org/10.1016/j.fshw.2022.07.054

Rahimian K, Arefian E, Mahdavi B, Mahmanzar M, Kuehu DL, Deng Y. 2023. SARS2Mutant: SARS-CoV-2 amino-acid mutation atlas database. NAR Genomics and Bioinformatics, 5: lqad037

Scardoni G, Laudanna C. 2012. Centralities based analysis of complex networks. In: New Frontiers in Graph Theory. IntechOpen, UK

Scheiner S, Kar T, Gu Y. 2001. Strength of the Calpha H..O hydrogen bond of amino acid residues. Journal of Biological Chemistry, 276: 9832-9837. https://doi.org/10.1074/jbc.M010770200

Sun H, Ma X, Fei L, Cao Z, Zhong H, Wang S. 2022. Amide Group Enhanced Self-Assembly and Adsorption of Dicarboxylic Amino Acid-Based Surfactants On a Rhodochrosite Surface Through Intermolecular Weak Interaction. https://doi.org/10.2139/ssrn.4221763

Wang K, Zhang S, Zhou X, Yang X, Li X, Wang Y, Fan P, Xiao Y, Sun W, Zhang P. 2024. Unambiguous discrimination of all 20 proteinogenic amino acids and their modifications by nanopore. Nature Methods, 21: 92-101

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

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