Int. J. Adv. Appl. Math. and Mech. 2(3) (2015) 31 - 37 (ISSN: 2347-2529)
Journal homepage: www.ijaamm.com

IJAAMM

International Journal of Advances in Applied Mathematics and Mechanics

Graph theoretic approach to analyze amino acid network

Research Article

Adil Akhtar *, Nisha Gohain

Department of Mathematics, Dibrugarh University, Dibrugarh-786004, India

Received 11 February 2015; accepted (in revised version) 06 March 2015

Abstract: To understand the evolution process of twenty natural amino acids is one of the most important area of the research in biological networks. In this manuscript we have considered the amino acid networks as a biological network based on amino acids properties. To analyze the faster or slower evolutionary process of the amino acids we have discussedclustering coefficient as a graph theoretic tool. We have also investigated the correlation coefficients between degrees of the amino acids. Finally we discuss degree of distribution of the amino acids.

MSC: 92B05 • 05C62 • 05C50

Keywords: Amino acid • Clustering coefficient • Correlation coefficient, • Graph

© *2015 IJAAMM all rights reserved.*

1. Introduction

Now a day in natural and artificial systems networks appear almost everywhere. Amino acids play a central role in cellular metabolism, and organisms need to synthesize most of them. There are 20 different amino acids being found till now that occurs in proteins. Each amino acid is a triplet code of four possible bases. A sequence of three bases forms a unit called codon. A codon specifies one amino acid. The genetic code is a series of codons that specify which amino acids are required to make up specific protein. As there are four bases, this gives us 64 codons. In this manuscript we have analyze amino acid networks through graph theoretic approach. Graph theory is one of the most important parts of mathematics, it is a non-numerical branch of modern mathematics considered part of topology, but also closely related to algebra and matrix theory. Application of the various tools of graph theory in biological networkis more important, which help to easily describe and analyze the relevant structures and which lead to findings of many important aspect of the structures. A biological network is any network that apply in biological system such as protein-protein interaction (PPI), gene regulatory network, neuronal network etc.Several groups have studied in this field. Kundu [1] discussed that hydrophobic and hydrophilic network satisfy âĂIJsmall world property" within protein. Also he has discussed that hydrophobic network have large average degrees of nodes than the hydrophilic network. Aftabuddin and Kundu discussed about three types of networks within protein and gives some idea about all three types of networks [2]. Newman discussed correlation of degree of centrality and betweenness centrality [3]. Also in Newman, discussed about assortative mixing property in the protein interaction networks, neural networks and food webs [4]. He also discussed that the information can be easily transferred through an assortative networks as compared to a disassortative network. Sinha and Bagler observed that average clustering coefficients of long range scales shows good negative correlation with the rate of folding, indicating that clustering of amino acids that participate into long range interaction with slow down folding process [5]. Fell and Wagner considered a graph with metabolites as vertices and edges connecting any two metabolites that appear in the same reaction [6] . They have examined whether metabolites with highest degree may belong to the oldest part of the metabolism. Jeong et al. discussed about the lethality and centrality of the PPI network [7] . Where proteins consider as a node and edges defined by direct physical interaction between nodes. They have shown that the deletion

* Corresponding author.
E-mail address: adil.akhtar19@gmail.com

Table 1. Amino acids property

A.A.	Hpho	Hphi	P	N P	Al	Ar	N (A)	(+)ve	(-)ve	N (B)
G	1	0	0	1	0	0	1	0	0	1
A	1	0	0	1	0	0	1	0	0	1
V	1	0	0	1	1	0	0	0	0	1
M	1	0	0	1	0	0	1	0	0	1
W	1	0	1	0	0	1	0	0	0	1
L	1	0	0	1	1	0	0	0	0	1
I	1	0	0	1	1	0	0	0	0	1
F	1	0	0	1	0	1	0	0	0	1
P	1	0	0	1	0	0	1	0	0	1
Y	1	0	1	0	0	1	0	0	0	1
S	0	1	1	0	0	0	1	0	0	1
T	0	1	1	0	0	0	1	0	0	1
E	0	1	1	0	0	0	1	0	1	0
C	0	1	1	0	0	0	1	0	0	1
N	0	1	1	0	0	0	1	0	0	1
Q	0	1	1	0	0	0	1	0	0	1
D	0	1	1	0	0	0	1	0	1	0
K	0	1	1	0	0	0	1	1	0	0
H	0	1	1	0	0	1	0	1	0	0
R	0	1	1	0	0	0	1	1	0	0

of the central vertices of the PPI networks, which are usually important functionally, is related to lethality. Wuchty and Stadler discussed various centrality measures in biological network. They concluded that the degree of vertex centrality alone is not sufficient to distinguish lethal protein from viable ones [8]. Also Schreiber and Koschutzki compared centralities for biological networks namely PPI network and transcriptional network (TR) of Escherichia coli [9]. In PPI networks considered proteins as a vertices and interactions as an edges and in TR networks of Escherichia coli models operons as vertices and regulation between transcription factors and operons as an edges. As a result of their study, it was observed that in the analysis of biological networks various centrality measures should be considered.

2. Basic concepts of graph

An undirected graph $G=(V,E)$ consists of a finite set V of vertices and a finite set $E \subseteq V \times V$ of edges. 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 consists of a set V of vertices and a set E of edges such that $e \in E$, if each edge of the graph G 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 an end point. 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. Also a clique in an undirected graph G is a subgraph G_1 which is complete.

3. Network in amino acids

Every amino acid exhibits different physico-chemical properties. In this study we have considered four properties of the amino acids. In Table 1 we have shown different physico-chemical properties of the amino acids. Here, under the column of any property, we have assigned 1 to an amino acid if the amino acid posses that property and 0 otherwise. When an amino acid which is neither aliphatic nor aromatic then we defined it as neutral (A). Similarly, we define neutral (B) to represent an amino acid which is neither positively charged nor negatively charged. We will consider two approaches. In the first approach we define an edge between two amino acids if there are at least two properties common to both. In the second approach two amino acids will be joined by an edge if there is at least one common property.

The difference in their properties basically depends upon their structures. If two amino acids have more common properties they will be structurally more similar than when they have less number of common properties. An amino acid will have more tendency or likelihood to evolve into another amino acid having structural similarity. In terms of properties we can say that an amino acid will have tendency to evolve into another if they have more common properties and less otherwise. So the graph based on properties will give a rough picture of a smooth transition or

evolution of the amino acids. The corresponding graphs are depicted below. Following the construction of the two graphs of Fig. 1 and Fig. 2, obtained by using the Table 1, the two networks are obviously connected graphs.

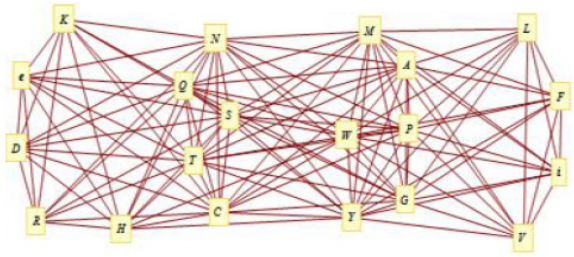

Fig. 1. First Approach

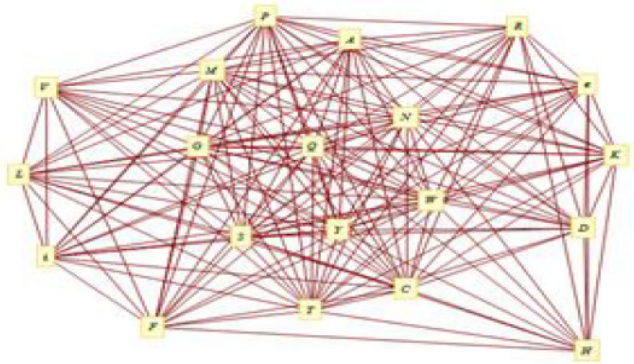

Fig. 2. Second Approach

Adjacency matrix for first approach:

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

Adjacency matrix for second approach:

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

4. Network parameters

There are various network parameters which are used in the biological network. In this paper we have discussed basically three network parameters; namely clustering coefficient, degree of distribution and Pearson's skewness. Clustering coefficient is the measurement that shows the tendency of a graph to be divided into cluster. A cluster is a subset of vertices that contains lots of edges connecting these vertices to each other. The clustering coefficient C_i of a node i is the ratio between the total number (e_i) of links actually connecting its nearest neighbours and the total number (the number of such links is $K_i(K_i-1)/2$, where K_i is the degree of node i) of all possible links between these nearest neighbours. It is given by $C_i = 2e_i/K_i(K_i-1)$.Also nodes with less than two neighbors are assumed to have a

clustering coefficient of 0. It takes values as $[0,1]$. The clustering coefficient of the whole network is the average of all individual C_i. The higher clustering coefficient of a node represents strong relationship in between neighbouring nodes. That is the higher value of the clustering coefficients of a node represents more number of connection among its neighbours. Since the maximum clustering coefficient value is 1, so we consider if the clustering coefficient value of the whole network is less than half of the maximum clustering value than it has small clustering value otherwise higher clustering value. In the following Table 2 and Table 3, we have shown clustering coefficients of all the amino acids.

Table 2. Clustering coefficient of the amino acids for first approach

G	A	V	L	I	M	P	F	W	Y	N	Q	S	T	C	D	E	K	R	H
0.78	0.78	1	1	1	0.78	0.73	1	0.73	0.73	0.73	0.73	0.73	0.73	0.73	1	1	1	1	0.85

Table 3. Clustering coefficient of the amino acids for second approach

G	A	V	L	I	M	P	F	W	Y	N	Q	S	T	C	D	E	K	R	H
0.88	0.88	1	1	1	0.88	0.88	0.93	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.93	0.93	0.93	0.93	0.94

From the above Table 3, it is clear that clustering coefficient of an amino acid depend upon degree of the amino acids as well as number of direct connection in between two neighbouring amino acids. From the above clustering values we observed that all the aliphatic amino acids form clique in both the approaches. Again the clustering coefficient of whole amino acids network for first and second approach is 0.85 and 0.91 respectively. Since the higher values of clustering coefficients of a network gives large effect on the nodes of the network and slowing down the information spread. From the above value it is clear that the information can be sent slow in these amino acids network. Again we can say that the first approach is faster than the second approach.

Next, it is of interest to investigate the nature of the node of the distribution of degrees of nodes for both patterns. The spread in the number of links a node has is characterized by a distribution function $P(k)$. The degree distribution $P(k)$ of a network is defined to be the fraction of nodes in the network with degree k. If there are n nodes in total in a network and n_k of them have degree k, we have $P(k) = n_k/n$. Generally the degree distributions value of a node represents the probability that a selected node will have exactly k links. In the following Table 4 and Table 5 we have shown degree of distribution values of different amino acids.

Table 4. Degree of distribution of amino acids for first approach

G	A	V	L	I	M	P	F	W	Y	N	Q	S	T	C	D	E	K	R	H
0.2	0.2	0.4	0.4	0.4	0.2	0.2	0.4	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.4	0.4	0.4	0.4	0.05

Also another well known parameter is skewness. Skewness is a measure of the symmetry or asymmetry of the distribution of a variable. The measuring skewness was first suggested by Karl Pearson in 1895. There are various measures of skewness. In this paper we have used only Pearson's coefficient of skewness.In normal curve the mean, the median and the mode all coincide and there is perfect balance between the right and the left side of the curve. The situation of skewness which means lack of symmetry occurs in a curve when the mean, median and mode of the curve are not coincident. Skewness describes the shape of the distribution. Symmetry means that the variables are equi-distance from the central value on the either side. Again asymmetrical means either positively skewed or negatively skewed. Skewness is denoted in mathematical notation by S_k. Based on the values and relative position of the mode, mean and median there are two types of skewness that appear in the distribution namely positive skewness and negative skewness.. If mean is maximum and mode is least and the median lies in between the two then it is called positive skewed distribution. Again if mode is maximum and the mean is least and the median lies in between the two then it is called negative skewed distribution. There are various relative measure of skewness. In this study we have discussed about Karl Pearson's coefficient of skewness, which is given by the following formula.

$$s_k = \frac{3(\text{Mean - Median})}{\text{Standard daviation}}$$

The value of the measure of the skewness lies within the range of - 3 to +3. If $S_k = 0$, then the distribution is symmetrical i.e., normal. $S_k > 0$, then the distribution is positively skewed. $S_k < 0$, then the distribution is negatively skewed. Here we assume the degree of distribution as variable (X) and number of the amino acids which contains same value of the distribution as frequency (f). Then we have following Table 6 and Table 7, for first and second approaches respectively, where 0.2 considered as assumed mean in both the approaches.

Table 5. Degree of distribution of amino acids for second approach

G	A	V	L	I	M	P	F	W	Y	N	Q	S	T	C	D	E	K	R	H
0.2	0.2	0.15	0.15	0.15	0.2	0.2	0.25	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.25	0.25	0.25	0.25	0.05

Table 6. Calculation of PearsonâĂŹs coefficient of skewness for first approach

X	f	$d_x = X - 0.2$	fd_x	fd_x^2
0.4	8	0.2	1.6	0.32
0.35	7	0.15	1.05	0.1575
0.2	4	0	0	0
0.05	1	-0.15	-0.15	0.0225

Table 7. Calculation of PearsonâĂŹs coefficient of skewness for second approach

X	f	$d_x = X - 0.2$	fd_x	fd_x^2
0.35	7	0.15	1.05	0.1575
0.25	5	0.25	0.25	0.0125
0.2	4	0	0	0
0.15	3	-0.15	-0.15	0.0075
0.05	1	-0.15	-0.15	0.0225

From the first approach we have mode is 0.4(because the highest frequency. i.e. 8) and median is 0.35. Also the standard deviation is 0.097. Therefore the Pearsons coefficient of skewness is -0.78 < 0. Again in second approach we have mode is 0.35(because the highest frequency. i.e. 7) and median is 0.283. Also the standard deviation is 0.086. Therefore the Pearson's coefficient of skewness is -1.15< 0. From here we concluded that degrees of distribution of the amino acids networks in both the approaches are negatively skewed distribution.

5. Correlation coefficient between various degrees of the amino acids

Here we discuss about correlation between the first and second approach for degree of the amino acids. The correlation coefficients between first and second approaches are shown in Table 8. All correlation coefficients are based on Pearson's method. It is observed that the correlation coefficients are highly correlated. Again we know that 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 higher degree have the tendency to connect with other vertices with low degree then the network is called disassortative. The range of r-value is between +1 and -1. If $r > 0$ then the network is assortative whereas if $r < 0$ then the network is disassortative.

Table 8. Correlation coefficients for the degrees of the amino acids

	First approach	Second approach
First approach	1	0.996
Second approach	0.996	1

Also from the above correlation coefficients we observed that these two networks form anassortative type ($r > 0$) network. Therefore the information can be easily transferred through that network.

6. Conclusion

In this paper we have equipped the amino acids with graph structure by defining compatibility relation based on properties. We have observed that the graph is connected. We have discussed clustering coefficient of the amino acids in both the approaches. we also observed that the aliphatic amino acids form a clique. Also we have discussed clustering coefficient of the whole networks, then we observed that the information can be sent slow in these amino acids network. Again from the correlation coefficient we observed that these two networks form an assortative type network. Finally we have observed that in both the approaches the degree of distribution is negatively skewed. Then by KS- test we observed that the degree of distribution for first and second approach follows beta and uniform distribution pattern respectively. Since our networks are based on the properties of the amino acids, the network shows a general picture of the evolution of the amino acid.

Acknowledgment

The authors would like to thank Prof. Tazid Ali, Department of Mathematics, Dibrugarh University for their valuable suggestion and encouragement to develop this paper.

References

[1] S. Kundu, Amino acid network with in protein, Physica A, 346 (2005) 104- 109.
[2] M. Aftabuddin, S. Kundu, Hydrophobic hydrophilic and charged amino acid networks within protein, Biophysical journal 93 (2007) 225-231.
[3] M. E. J. Newman, A measure of betweenness centrality based on random walks, Social networks, 27(1) (2005) 39-54.
[4] M.E.J. Newman, Assortative mixing in networks, Physical review letter, 89 (2002) (208701).
[5] G. Bagler, S. Sinha, Assortative mixing in Protein Contact Networks and protein folding kinetics, Bioinformatics 23 (2007)1760-1767.
[6] D. A. Fell, A. Wagner, The small world of metabolism, Nature Biotechnology, 18 (2000) 1121-1122.
[7] H. Jeong, S. P.Mason, A. L. Barabasi, Z. N. Oltvai, Lethality and centrality in protein networks, Nature 44 (2001) 411.
[8] S. Wuchty, P.F. Stadler, Centers of complex networks, Journal of Theoretical Biology, 223 (2003) 45- 53.
[9] F. Schreiber, D. Koschutzki, Comparison of centralites for biological networks, Proc German conf Bioinformatics (GCB) P-53 of LNI, 2004.
[10] B. Chakrabarty, N. Parekh, Graph centrality analysis of structural ankyrin repeats,International Journal of Computer Information Systems and Industrial Management Applications 6 (2014) 305-314.
[11] M. W. Hahn, G. Conant, A. Wagner, Molecular evolution in large genetic networks: connectivity does not equal importance, Technical report, Santa Fe Institute (2002) 02-08-039.