Physica A 452 (2016) 69–78

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

# Analysis of amino acids network based on distance matrix

Tazid Ali, Adil Akhtar*, Nisha Gohain

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

## HIGHLIGHTS

- Amino acids E, D, N, K, Y play crucial role in the evolutionary process.
- Betweenness centrality is independent of the other centrality measures.
- The evolutionary information flow is relatively slow in the neighbourhood of L and R.
- This network is assortative and consequently evolutionary information flow be easy.
- The degree of distribution of the amino acids network follows uniform distribution.

## ARTICLE INFO

*Article history:*
Received 24 January 2015
Received in revised form 29 September 2015
Available online 13 February 2016

*Keywords:*
Amino acids
Codons
Centrality measure
Clustering coefficient
Distance matrix

## ABSTRACT

In this paper we have constructed a distance matrix of the amino acids. The distance is defined based on the relative evolutionary importance of the base position of the corresponding codons. From this distance matrix a network of the amino acids is obtained. We have argued that this network depicts the evolutionary pattern of the amino acids. To examine the relative importance of the amino acids with respect to this network we have discussed different measures of centrality. We have also investigated the correlation coefficients between different measures of centrality. Further we have explored clustering coefficient as well as degree of distribution.



## 1. Introduction

All existing organisms consist of cells. In each cell there is the same set of chromosomes. Chromosomes are strings of DNA and serves as a model for the whole organism. A chromosome consists of genes, block of DNA. Each gene encodes a particular protein. Proteins are the basic constructional blocks and functional elements of living organisms. Amino acids are the building blocks of proteins. Each protein is formed by a linear chain of amino acids. 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 importance of base position is suggested by the error (accepted mutation) frequency found in the codons [1]. The frequency of errors decreases from the third base to first base and then next to the second base. That is the second base is biologically most relevant and third base is least relevant base in the codon. Also the second position (most significant base position) of codons is connected with the hydrophobicity of amino acids. The genetic code is a series of codons that specify which amino acids are required to make up a specific protein. As there are four bases, (Adenine (A), Cytosine (C), Guanine (G) or Thymine (T/U)) this gives us 64 codons. This means that there is some overlap i.e., more than one codon codes for the same amino acid. The codons that code for the same amino acids are known as synonymous codons. We can consider this as a function of many to one carrying codons to amino acids.

* Corresponding author.
*E-mail addresses:* tazidali@dibru.ac.in (T. Ali), adil.akhtar19@gmail.com (A. Akhtar), gohainnisha@gmail.com (N. Gohain).

http://dx.doi.org/10.1016/j.physa.2016.01.074
0378-4371/

Also out of these 64, the three triplets UAA, UAG and UGA are known as stop codons or nonsense codons and their role is to stop the biosynthesis.

Researchers have made significant contributions in the field of biological network. Kundu [2], discussed hydrophobic and hydrophilic networks separately. In these networks amino acids are considered as vertices and two amino acids are linked by an edge if any two atoms from two different amino acids lie within 5A distance. He showed that hydrophobic and hydrophilic networks satisfy "small world property" within protein. He also observed that hydrophobic network has large average degrees of nodes than the hydrophilic network. Further he has shown that the degree of distributions for long-range interactions of both the networks follow scale-free behaviour. Aftabuddin and Kundu [3] discussed three types of networks (hydrophobic, hydrophilic and charged) within protein. They have showed that the average degree of the hydrophobic network has a significantly larger value than the other two networks. The average degree of the hydrophilic network is slightly higher than that of the charged network. The average strength of the nodes of hydrophobic network 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. Each of the three types of networks follows the "small-world" property. Also they have shown that all the networks are of assortative type. Jiao et al. [4] discussed about the weighted amino acid network based on the contact energy. They have shown that weighted amino acid network satisfies "small-world" property. Akhtar and Ali [5] considered amino acids network based on mutation of the codons. They have discussed different centrality measures and observed that the high hydrophilicity amino acid R (Arginine) and least hydrophilicity amino acid S (Serine) have the highest centrality values irrespective of the centrality measures. Further they observed that the degree of distribution follows Weibull distribution pattern. Wuchty and Stadler [6] 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. Newman [7] discussed about assortative mixing property in the protein interaction networks, neural network and food webs. He also established that the information can be easily transferred through an assortative network as compared to a disassortative network. Bagler and Sinha [8] observed that average clustering coefficients of long range scales show a good negative correlation with the rate of folding of protein. Schreiber and Koschutzki [9] 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 this paper we have attempted to explore some graph theoretic notions in amino acids network.

The paper is organized as follows: in Section 2 we give some preliminary concepts of graph theory on which we operate and briefly review the various centrality measures. In Section 3 we define graph in amino acids based on distance matrix and discuss about various centrality measures. Also we discuss the bivariate correlation in between different centrality measures. Further, a comparison of amino acid distances based on codon similarity and properties is done in this section. In Section 4 we discuss some network parameters. In Section 5 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 $E \subseteq V \times V$ 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 neighbourhood $N(u)$ 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 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.

### 2.1. Centrality in graph

In graph theory, centrality measure of a vertex represents its relative importance within the graph. 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.

#### 2.1.1. 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. The nodes directly connected to a given node $u$ are also called first neighbours of the given node.