# Investigating Evolutionary Relationships between Species through the Light of Graph Theory based on the Multiplet Structure of the Genetic Code

Antara Sengupta
Dept. of MCA
MCKV Institute of Engineering
Howrah, India
antara.sngpt@gmail.com

Jayanta Kumar Das
Applied Statistics Unit
Indian Statistical Institute
Kolkata, India
dasjayantakumar89@gmail.com

Pabitra Pal Choudhury
Applied Statistics Unit
Indian Statistical Institute
Kolkata, India
pabitrapalchoudhury@gmail.com

***Abstract*—Investigating evolutionary relationship between various species through similarity/dissimilarity analysis is a fundamental method. In this present work firstly 20 canonical amino acids and 3 stop codons (terminations) are classified into five different classes depending upon their frequency mapping with 64 codons of genetic table. Secondly, each DNA sequence is represented by a weighted directed multi graph based on that classification. Thirdly, the procedure has been implemented to find out the evolutionary relationship between various species of alpha globin and beta globin genes. Here a new mathematical tool has been constructed to derive similarity/dissimilarity matrix, to get suitable phylogenetic trees for each data set. It is completely alignment free approach and hence the time complexity is directly proportional to the sequence length, that is O(N). Moreover the classification rule will decrease the complexity of graph constructions.**

***Keywords—Weighted Directed Multi Graph; Alpha Globin; Beta Globin; Similarity Matrix; Phylogenetic Tree.***

## I. INTRODUCTION

Genetics is becoming more complex day by day for the evolutionists due to the exponential growth of DNA database and its diversity. Scientists from almost every field (diverse field) are trying to make quantitative analysis of DNA, RNA sequences and even gene regulatory networks [19]. DNA Sequence analysis is the process to understand its characteristics, whereas similarity and dissimilarity analysis of DNA sequences throughout the various species is a straight forward way to find out evolutionary relationships among them. Till date numerous approaches are reported in computational biology to characterize DNA sequences and to find out evolutionary relationships. When the multiple alignment of any DNA sequence is time consuming, application of graph theory can give easy visual inspection of it. Large numbers of computer algorithms are available both for DNA and protein sequence analysis and it has its very long history too[1-20].Three-dimensional representation of DNA sequences using Z curve is an old way, where as some authors [17]have tried to make DNA sequence analysis in terms of random walk. Some authors also have tried to make 2D as well as 3D representation of DNA sequences[8-12]and tried to get similarity matrix to construct phylogenetic trees. Further, in [20] a mathematical approach is applied to understand the degeneracy of genetic code and identified a new way to classify protein family. The basic procedure which are common in almost all papers are firstly mathematical denotations of DNA/RNA sequences, then based on those denotation represent the sequence in terms of graphs and finally derive similarity matrix to make evolutionary relationships among them.

Here in this paper DNA sequences are read as codons and represented as weighted directed multi graph. G=(V, A) is a Weighted Directed Multi Graph which allows multiple arcs between the same source and target nodes, where V is the set of all nodes or vertices and A is a multi set of ordered pairs of vertices called arcs.

Codon consists of three adjacent nucleotides on the mRNA that specifies the genetic code which translate the encoded information stored in DNA sequences into proteins. Thus 61 codons and 3 stop codons are there and each codon codes for a specific Amino acid and each amino acid is responsible for a specific protein synthesis. As a result it is clear that they do not have one-to-one mapping between them, instead some amino acids map with more than one codon and there is a talk about degeneracy which initiates 'multiplet structure'[18]. According to genetic table there are 2 singlets (tryptophane, methionine), 2 triplets (isoleucine, stop codon), 5 quartets (threonine, valine, glycine, proline, alanine), 3 sextets (serine, arginine, leucine,), and finally 9 doublets (phenylalanine, cysteine, asparagine, tyrosine, aspartic acid, lysine, glutamic acid, glutamine, histidine).

In this proposed work DNA sequences are read as codons (triplet) and the codons are grouped according to their mapping with amino acids and stop codons. This new approach decreases the complexity of graph construction because 64 codons have been classified into 5 classes based on the multiplet structure of amino acids along with 3 stop codons which are specified in Table I. The novelty lies in the efficiency and simplicity of the method which can lead biologists to find evolutionary relationships throughout the species without any biological interventions.

Facts and findings of the paper have been organized as follows: In section II the methodology has been described thoroughly. The codon classification is stated and weighted directed multi

978-1-5090-1560-3/17 $31.00 © 2017 IEEE
DOI 10.1109/IACC.2017.166

graph has been constructed based on that classification rule. Then the graphical information is represented in a matrix of size 5 by 5. A formula is constructed here in this section to draw similarity/dissimilarity matrix for DNA sequences of various species. In section III, the procedure stated in section II is applied on two data sets and experimental results are discussed clearly. The whole work has been concluded in section IV.

## II. METHODOLOGY

Reading DNA sequence to construction phylogenetic tree is a step by step procedure which seeks graph theoretical effort along with mathematical formulation. Hence the methodology is described using flow chart as described in Fig. 1.

As we know that the DNA sequences are nothing but the combinations of four nucleotides A, C, T and G, so let $S_i \in \{s_1, s_2, ..., s_k,\}$ be an arbitrary DNA sequence, where for any $s_i \in \{A, C, T, G\}$ for any $i \in \{1,2,...,k\}$. Then, it is possible to read a DNA sequence as codons, such that codon sequence $C_i \in \{c_1, c_2, ..., c_j\}$, where $j = \left[\frac{k}{3}\right]$ and $c_i$ is a codon for any $i \in \{1,2,..., j\}$.

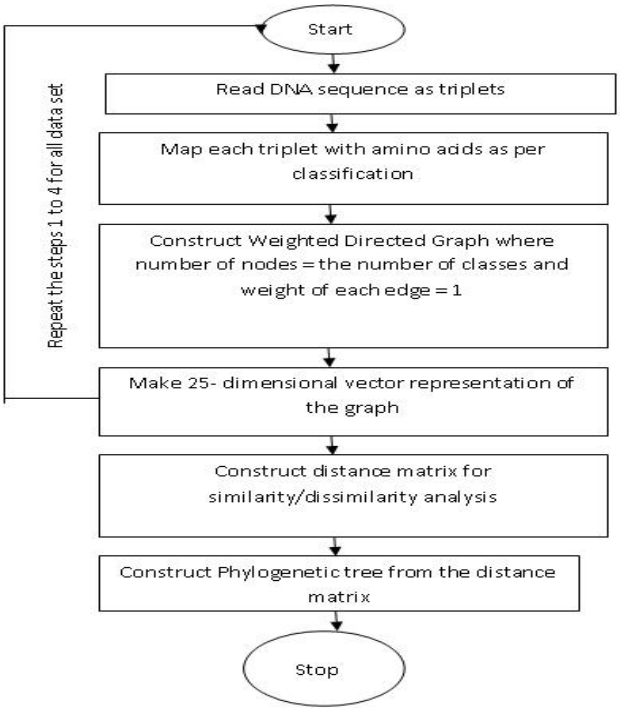


Figure 1. Flow chart to show the methodology

According to genetic code table five possible multiplet structures are there(singlet, doublets ,triplet ,quartets and sextets). The following table shows how are the 64 codons classified as per multiplet structural properties they belong to.

TABLE I. THE CLASSIFICATION OF 64 CODONS AS PER THEIR MULTIPLET STRUCTURAL PROPERTIES. HERE AMINO ACID IS ABBREVIATED AS AA.

| CLASS | 1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| AA | Met/M | Trp/W | | | | | | | |
| CODON | ATG | TGG | | | | | | | |
| CLASS | 2 | | | | | | | | |
| AA | Asn/N | Asp/D | Cys/C | Gln/Q | Glu/E | His/H | Lys/K | Phe/F | Tyr/Y |
| CODON | AAT | GAT | TGT | CAA | GAA | CAT | AAA | TTT | TAT |
| | AAC | GAC | TGC | CAG | GAG | CAC | AAG | TTC | TAC |
| CLASS | 3 | | | | | | | | |
| AA | Ile/I | STOP | | | | | | | |
| CODON | ATT | TAA | | | | | | | |
| | ATC | TAG | | | | | | | |
| | ATA | TGA | | | | | | | |
| CLASS | 4 | | | | | | | | |
| AA | Ala/A | Gly/G | Pro/P | Thr/T | Val/V | | | | |
| CODON | GCT | GGT | CCT | ACT | GTT | | | | |
| | GCA | GGA | CCA | ACA | GTA | | | | |
| | GCC | GGC | CCC | ACC | GTC | | | | |
| | GCG | GGG | CCG | ACG | GTG | | | | |
| CLASS | 6 | | | | | | | | |
| AA | Arg/R | Leu/L | Ser/S | | | | | | |
| CODON | AGA | TTA | AGT | | | | | | |
| | AGG | TTG | AGC | | | | | | |
| | CGT | CTT | TCT | | | | | | |
| | CGA | CTA | TCA | | | | | | |
| | CGC | CTC | TCC | | | | | | |
| | CGG | CTG | TCG | | | | | | |

### A. *Construction of Weighted Directed Multi Graphs to represent DNA sequences*

Let Gm (V, A) be the directed multi graph such that each class of codons represented as vertices, such that the vertex set $V(Gm) = \{1,2,3,4,6\}$. For any set of codons of a DNA sequence say $C_i \in \{c_1, c_2, ..., c_j\}$, $v_1$ and $v_2$ are any pair of codon classes for which directed multi graphs may have parallel arcs from $v_1$ to $v_2$, where $A^m_{v1,v2} \neq 0$ is the set of arcs from $v_1$ to $v_2$. Now define the weight of each arc as 1 from $v_1$ to $v_2$ .Now if we assign weight for every arc from $v_1$ to $v_2$, then total weight of the arcs from $v_1$ to $v_2$ of graph Gm is,

$$w_m(v_1, v_2) = \sum w(v_1, v_2) \qquad (1)$$

Where, $A^m_{v1,v2} \neq 0$.

### B. *Representation of Weighted Directed Multi Graphs in terms of 25-dimentional vector*

For any weighted directed multi graph say $G_m$, we can get its corresponding (5X5) adjacent matrix M for each pair of vertices. The representation of such weighted directed multi graph in terms of adjacency matrix is shown below.

TABLE II. REPRESENTATION OF WEIGHTED DIRECTED MULTI GRAPHS IN TERMS OF ADJACENCY MATRIX.

| | 1 | 2 | 3 | 4 | 6 |
|---|---|---|---|---|---|
| 1 | w(1,1) | w(1,2) | w(1,3) | w(1,4) | w(1,6) |
| 2 | w(2,1) | w(2,2) | w(2,3) | w(2,4) | w(2,6) |
| 3 | w(3,1) | w(3,2) | w(3,3) | w(3,4) | w(3,6) |
| 4 | w(4,1) | w(4,2) | w(4,3) | w(4,4) | w(4,6) |
| 6 | w(6,1) | w(6,2) | w(6,3) | w(6,4) | w(6,6) |

The adjacency matrix M of each DNA sequences mentioned in section above can be represented as 25-dimentional vectors[1] in row order say $\xrightarrow[R]{}$ such that $\mathrm{R} = (\mathrm{w}(1,1),...\mathrm{w}(1,6),\ldots,\mathrm{w}(6,1),\ldots\mathrm{w}(6,6))$.

*C. Similarity/Dissimilarity Analysis of a given set of DNA sequences based on their weight matrix*

Now for any two DNA sequences S1 and S2 the corresponding set of vectors are $\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_{25}\}$ and $\mathrm{Y} = \{\mathrm{y}_1, \mathrm{y}_2, \ldots \mathrm{y}_{25}\}$ respectively, then weight deviation (WD) between the two sequences S1 and S2 is as follows:-

$$WD(S_1, S_2) = \frac{\sum_{i=1}^{25} |X_i - Y_i|}{25} \qquad (2)$$

Thus we will get Similarity/Dissimilarity matrix D for a set of DNA sequences $S_1, S_2, \ldots S_n$ as in Table 3. According to the equation (2), the weight deviation between any two DNA sequences say S1 and S2, is inversely proportional to the degree of similarity of those sequences.

TABLE III. THE SIMILARITY/DISSIMILARITY MATRIX FOR THE SEQUENCES S1, S2…..Sn.

| **Sequences** | S1 | S2 | …… | Sn |
|---|---|---|---|---|
| S1 | | | | |
| S2 | | | | |
| : | | | | |
| Sn | | | | |

## III. APPLICATIONS AND RESULTS

The graph theoretical method introduced in above section has been applied on two different data sets to verify the method itself and experimental results are discussed accordingly.

*A. Data Set Specification*

Alpha and beta globins are the globin proteins which have significant role to make hemoglobin in mammals. In this present work primary DNA sequences of two gene members from globin protein family alpha globin type 1and beta globin are taken to carry out the experiments. DNA sequence of Alpha globin type 1 of 5 species from NCBI database (Table IV) and DNA sequences of beta globin of 11 different species have been taken from paper reported previously [10].

TABLE IV. DNA SEQUENCES OF ALPHA GLOBIN TYPE 1 OF 5 SPECIES (DATA SET 1)

| Species Name | Accession No. |
|---|---|
| Mus musculus | NM_008218 |
| Pongo Abeli | NM_001132429 |
| Homo Sapiens | JQ423459 |
| Papio Anubis | NM_001168816 |
| Macaca Mulatta | NM_001266776 |

*B. Construction of Weighted Directed Multi Graphs and Representation of Weighted Directed Multi Graphs in terms of 25-dimensional vector for each data set.*

Weighted Directed Multi Graphs for 5 species of data set 1 and 11 species of data set 2 are constructed based on the classification rule. Thus all the 16 graphs are represented in terms of 25-dimensional vectors. As an example, the Weighted Directed Graphs of DNA sequences of Human alpha globin and beta globin have been shown in Fig. 2 and Fig. 3 respectively. Their corresponding adjacency matrices are also shown in Table V and Table VI respectively. Both of them reflect the interrelationships of multiplets which take place in those DNA sequences.

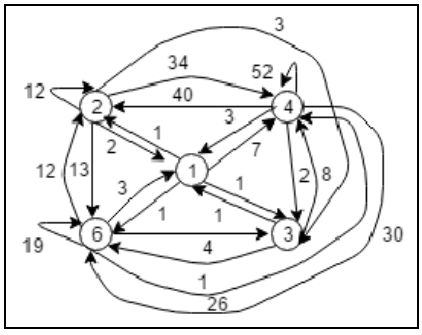


Figure 2. Weighted Directed Multi Graph for the data set 1 (human alpha globin DNA sequence).

The graphical representation stated above gives a pictorial view of the DNA sequence of human alpha globin, where the arcs between any two vertices are shown according to the equation 1 stated at section II. As an example from vertex 4

to vertex 2 the graph has 40 parallel arcs and thus as weight of each arcs are considered as 1, so here the summation of weight of the arcs between those vertices is 40.

TABLE V. CORRESPONDING ADJACENT WEIGHT MATRIX OF THE DIRECTED MULTI GRAPH OF DATA SET 1(HUMAN ALPHA GLOBIN).

| Weight → / ↓ | 1 | 2 | 3 | 4 | 6 |
|---|---|---|---|---|---|
| **1** | (0) | (1) | (1) | (7) | (1) |
| **2** | (2) | (12) | (3) | (34) | (13) |
| **3** | (1) | (0) | (0) | (8) | (1) |
| **4** | (3) | (40) | (2) | (52) | (30) |
| **6** | 35) | (12) | (4) | (26) | (19) |

The weighted directed multi graph stated below in Fig. 3 describes how multiplets are arranged themselves to form the DNA sequence of human beta globin. As an example from vertex 4 to vertex 2 the graph has 13 parallel arcs and thus using Equation 1, the the summation of weight of the arcs between those vertices is 13.

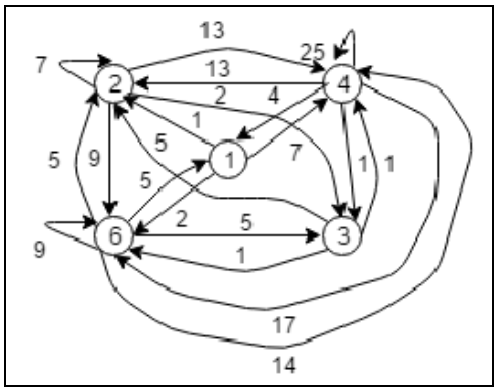


Figure3. Weighted Directed Multi Graph for the data set 2 (human beta globin DNA sequence).

The corresponding adjacent matrix of the graph is shown at Table VI.

TABLE VI. CORRESPONDING ADJACENT WEIGHT MATRIX OF THE DIRECTED MULTI GRAPH OF DATA SET 2(HUMAN BETA GLOBIN).

| Weight → / ↓ | 1 | 2 | 3 | 4 | 6 |
|---|---|---|---|---|---|
| **1** | (0) | (1) | (0) | (7) | (2) |
| **2** | (0) | (7) | (2) | (13) | (9) |
| **3** | (0) | (5) | (0) | (1) | (1) |
| **4** | (4) | (13) | (1) | (25) | (17) |
| **6** | (5) | (5) | (5) | (14) | (9) |

Thus graphical representations of data set 1 and dataset 2 have been made along with their corresponding adjacent matrices. As a result in this subsection row vectors are ready for each graph for analysis of evolutionary relationships between various species of two data sets.

*C. Similarity/ Dissimilarity Matrices based on the experiment*

In this subsection it is tried to derive distance matrix for each data set to make similarity/dissimilarity analysis between Species It is worthy enough to say that the pair of species having smaller deviation will have more similarities between them in their respective DNA sequences.

TABLE VII. SIMILARITY/DISSIMILARITY MATRIX OF DATA SET 1(DNA SEQUENCES OF ALPHA GLOBIN TYPE 1 OF 5 SPECIES).

| 5 Species → / ↓ | Macaca Mulatta | Papio Anubis | Homo Sapiens | Pongo Abeli | Mus Musculus |
|---|---|---|---|---|---|
| Macaca Mulatta | 0 | 23.33 | 37 | 25 | 14.66 |
| Papio Anubis | 23.33 | 0 | 32.33 | 9.66 | 26.66 |
| Homo Sapiens | 37 | 32.33 | 0 | 30.66 | 37.66 |
| Pongo Abeli | 25 | 9.66 | 30.66 | 0 | 27 |
| Mus Musculus | 14.66 | 26.66 | 37.66 | 27 | 0 |

It can be observed from the distance matrix of Table VII that the DNA sequence of Macaca mulatta has closest evolutionary relationship with Mus musculus, where as Papino Anubis and Homo sapiens has similarity with Pongo abeli. The phylogenetic tree construction (Fig. 4) reflects that fact accordingly.

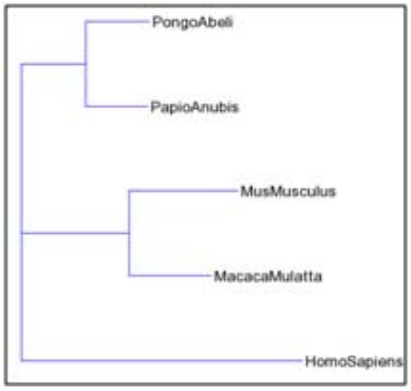


Figure 4. Phylogenetic tree constructed for Data Set1

The distance matrix constructed in Table VIII reflects the strong evolutionary relationships among Human- Gorilla-Chimpanzee, House mouse-Norway Rat and Bovin with Rabbit, whereas, as an example, there have huge dissimilarities between Human -House mouse and Human-Norway rat. The phylogenetic tree derived in Fig. 5 verifies the result found. It gives a clear pictorial view of the evolutionary relationships between all 11 species.

TABLE VIII. SIMILARITY/DISSIMILARITY MATRIX OF DATA SET2(HUMAN BETA GLOBIN DNA SEQUENCES OF 11 SPECIES)

| 11 Species → ↓ | Human | Black Lemur | House Mouse | Goat | Rabbit | Norway Rat | Bovin | Opossum | Gallus | Chimpanzee | Gorilla |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Human | 0 | 1.36 | 1.52 | 1.68 | 1.24 | 1.36 | 1.6 | 2.32 | 1.36 | 1.08 | 1.08 |
| Black Lemur | | 0 | 1.52 | 1.8 | 1.52 | 1.6 | 1.44 | 2.24 | 2.08 | 1.72 | 1.8 |
| House Mouse | | | 0 | 2 | 1.36 | 1.2 | `.36 | 1.92 | 1.44 | 1.96 | 1.8 |
| Goat | | | | 0 | 1.56 | 1.68 | 1.04 | 1.84 | 1.6 | 1.8 | 1.56 |
| Rabbit | | | | | 0 | 1.84 | 0.96 | 2.32 | 1.76 | 1.8 | 1.72 |
| Norway Rat | | | | | | 0 | 1.76 | 2.16 | 1.6 | 2.04 | 1.96 |
| Bovin | | | | | | | 0 | 2.24 | 2.08 | 2.12 | 2.12 |
| Opossum | | | | | | | | 0 | 2.24 | 2.36 | 2.12 |
| Gallus | | | | | | | | | 0 | 2 | 1.88 |
| Chimpanzee | | | | | | | | | | 0 | 0.64 |
| Gorilla | | | | | | | | | | | 0 |

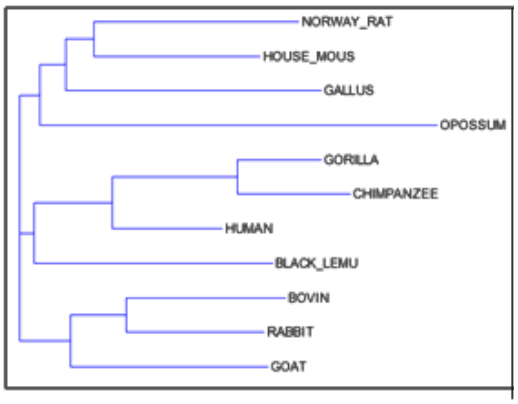


Figure 5. Phylogenetic tree constructed for Data Set 2

## IV. CONCLUSION AND DISCUSSION

In this present work firstly 20 canonical amino acids and terminations (or stop codons) are classified into five classes depending upon their mapping with 64 codons of genetic table. An emphasis has been given to the numbers of codons capable to represent a single amino acid. Secondly it is tried to represent the total set of given DNA sequences through the light of graph theoretical methods. The novelty of this approach is the simplification of graphical representation of DNA sequence, which provides simpler way to view and to compare various gene structures. Moreover it is completely alignment free approach and hence the time complexity is directly proportional to the sequence length, that is O(N). Not only that but also it is possible to find out similarities between DNA sequences over various species using this proposed mathematical model without any biological interventions, which can help biologists as a whole to investigate evolutionary relationships between those species.

## REFERENCES


[1] R. Wu, Q. Hu, R. Li, G. Yue, "*A novel composition coding method of DNA sequence and its application"*, MATCH Commun. Math. Comput. Chem. 67 (2012) 269–276.

[2] S. A. M. A. Junid et al, "Potential of Graph Theory Algorithm Approach for DNA Sequence Alignment and Comparison", IEEE Computer Society Washington, DC, USA 2012 ISBN: 978-0-7695-4668-1 doi 10.1109/ISMS.2012.123

[3] X. Q. Liu, Q. Dai, Z. Xiu, T. Wang, "PNN-curve: A new 2D graphical representation of DNA sequences and its application", J. Theor. Biol. 243 (2006) 55–56.

[4] P. He, D. Li, Y. Zhang, X. Wang, Y. Yao," A 3D graphical representation of proteinsequences based on the Gray code", J. Theor. Biol. 304 (2012) 81-87.

[5] N. Jafarzadeh, A. Iranmanesh, "C-curve: A novel 3D graphical representation of DNA sequence based on codons", Math. Biosci. 241 (2013) 217–224.

[6] E. Blaisdell, "A measure of the similarity of sets of sequences not requiringsequence alignment", Proc. Nat. Acad. Sci. USA 83 (1986) 5155–5159.

[7] S. Y. Wang, J. Yuan, S. Lin, "DNA labelled graphs with DNA computing", Sci. Chin.Ser. A: Math. 51 (2008) 437–452.

[8] W. Imrich, S. Klažvar, "Product Graphs: Structure and Recognition", Wiley, NewYork, 2000.

[9] X. Guo, A. Nandy, "Numerical characterization of DNA sequences in a 2-D graphical representation scheme of low degeneracy" Chem. Phys. Lett., 369 (2003), pp. 361–366.

[10] B. Liao, T. Wang, "Analysis of similarity of DNA sequences based on triplets",J. Chem. Inf. Comput. Sci., 44 (2004), pp. 1666–1670

[11] A Nandy, "A new graphical representation and analysis of DNA sequence structure: I.Methodology and application to globin genes",Curr. Sci., 66 (1994), pp. 309–314

[12] H.-J.Yu," Similarity Analysis of DNA Sequences Based on Three 2-D Cumulative Ratio Curves", DOI: 10.1007/978-3-642-24553-4_61 Conference.

[13] N. Helal, R.A. Moneim, M. Fathi," Mathematical Modeling Of P53 Gene Based On Matlab Code", IJRRAS_11_2_10, IJRRAS 11 (2) May 2012.

[14] B. Liao , C. Zeng, F. Li , Yong Tang, "Analysis of Similarity/ Dissimilarity of DNA Sequences Based on Dual Nucleotides", MATCH Commun. Math. Comput. Chem. 59 (2008) 647-652 ISSN 0340 – 6253.

[15] A. Sengupta , S.Hassan, P.P.Choudhury. Article: "Quantitative Understanding of Breast Cancer Genes". IJCA Proceedings on National

Conference cum Workshop on Bioinformatics an Computational Biology NCWBCB(1):15-17, May 2014.

[16] X. Qi, Q. Wu, Y. Zhang, E. Fuller, C-Q Zhang, "A Novel Model for DNA Sequence Similarity Analysis Based on Graph Theory", Evolutionary Bioinformatics 2011:7

[17] P. M. Leong, S. Morgenthaler," Random walk and gap plots of DNA sequences", Comput. Appl. Biosci. 11 (1995) 503–507.

[18] T. Négadi, "The genetic code degeneracy and the amino acids chemical composition are connected", arxiv.org/pdf0903.4131, 2009

[19] A.Sengupta, C. Ray, P.P.Choudhury," Study of miRNA Let-7 Involvement in Breast Cancer through the Light of Graph Theory and Canalizing Function", IJCA Proceedings on National Conference cum Workshop on Bioinformatics an Computational Biology NCWBCB (3):26-29, May 2014.

[20] J. K. Das, A. Majumder, P.P.Choudhury, B. Mukhopadhyay, "Understanding of Genetic Code Degeneracy and New Way of Classifying Protein Family: A mathematical Approach", 2016 IEEE 6th International Conference on Advanced Computing, DOI 10.1109/IACC.2016.57.