# A mathematical aspect to establish phylogenetic network using hamming distance approach

**Mannu, Rinku Mathur** *

Department of Mathematics, School of Chemical Engineering and Physical Sciences,

Lovely Professional University, Phagwara - 144411, Jalandhar, India

## Abstract

In this work, DNA (de-oxy ribonucleic acid) sequences are transformed to binary sequences to explore the dissimilarities between different species. A mathematical approach has been reported and implemented to establish phylogenetic networks by using the concept of hamming distances on binary sequences to explore the evolutionary history of the species. A program is also coded in python language for the conversion of DNA sequences as well as to find the hamming distances among them.

**Keywords**: Hamming distance, phylogenetic network, DNA sequences, binary sequences

## 1. Introduction

Evolutionary studies have been of prime importance to life scientists since ancient times. The growth of genomic and evolutionary data in databases poses challenges to biologists to understand and analyze it [2, 5]. Computer scientists and mathematicians needs to come forward and to develop some new theories and models to analyze the large amount of genomic data. A phylogenetic network is any graphical representation used to envision formative associations between characteristics, chromosomes, genomes, or species under consideration. They are used when reticulation events, for instance, hybridization, recombination, or quality duplication and adversity are acknowledged to be incorporated. They differentiate from phylogenetic trees by incorporating short paths (showing hybridizing or recombination events) among different nodes of the evolutionary tree. Many attempts have been reported in literature to generate such type of models and algorithms representing the evolutionary histories among different species [2-4, 6-8].

Information of a particular gene of every living cell of the organisms has been carried out by DNA sequences which are comprised of four nucleotides called Adenine, Thymine, Guanine and Cytosine [9]. With the complexity of biological data, it is the need of hour to develop the fast and efficient methods to explore the evolutionary relation among sequences generally represented by phylogenetic networks [5, 10].

In this work, a program has been coded in python language to convert biological sequences into binary sequences and then same set of binary sequences has been analyzed to find the hamming distances among them. Hamming distances among the sequences are then used to cluster the different species depending upon their distances. Finally, a phylogenetic network among biological species can be inferred by following the approach proposed which is fast and alignment free and applied over the DNA sequences of six organisms.

## 2. Preliminaries

Some of the basic concepts used in this study are discussed in subsections:

### 2.1 De-oxy ribonucleic acid (DNA)

De-oxy ribonucleic acid (DNA) is an atom which carries the directions, a life needs to grow, live and duplicate. These directions can be visulized inside each cell, and are passed down from ancestors to their offsprings. DNA is composed of particles called nucleotides and each nucleotide contains a phosphate gathering, a sugar gathering and a nitrogen base. The four kinds of nitrogen bases are adenine (A), guanine

(G), thymine (T), and cytosine (C) [9]. In line with DNA, ribonucleic acid (RNA) is also composed of four base pairs in which only thymine (T) is replaced by Uracil (U) [5, 9-10].

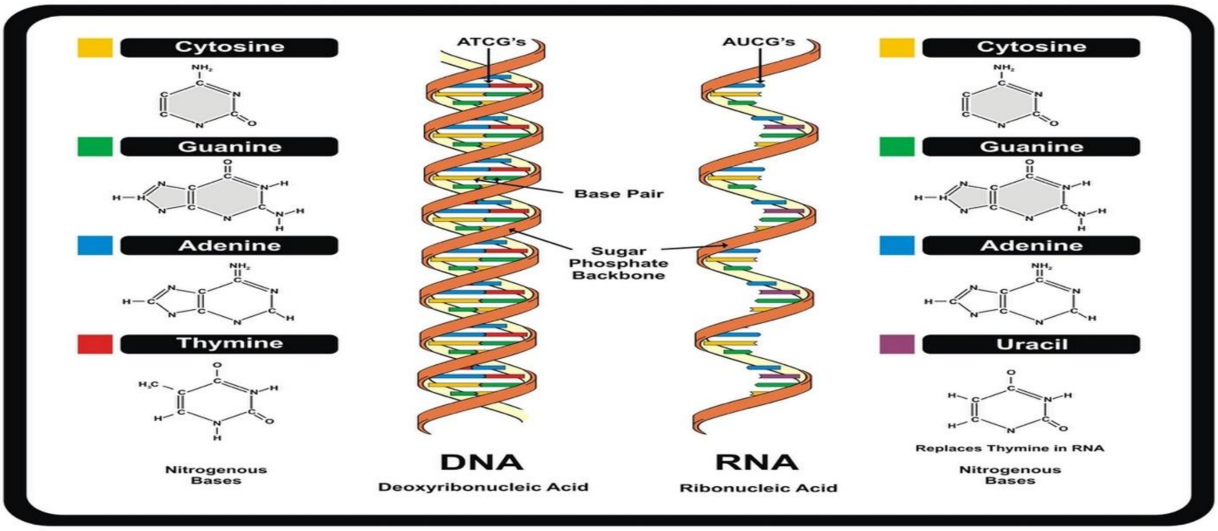


Figure 1: Chemical structure of DNA and RNA

## 2.2 Graph

Let Z be a set of species (or taxas), then a graph is a well-ordered pair G = (V, E), where V, is a set of objects called nodes or vertices. Each node is either a species belonging to a set Z or an intermediate node belonging to V −Z and E is a set of edges shows the path among pair of vertices (species) [9].

## 2.3 Path

A succession of vertices $(v_0, v_1, v_2, \ldots., v_k)$ such that for all $i$, $(v_i, v_{i+1}) \in E$ is called a path.

## 2.4 Cycle

A cycle is a path of vertices with $k > 2$, $v_0 = v_k$ and $v_i = v_j$ for $i \leq 0 < k$ [3, 5].

## 2.5 Connected Graph

If every pair of vertices $(x, y) \in V$ in G is connected by at least one path, then the graph is said to be connected, otherwise G is disconnected.

## 2.6 Hamming Distance

The total number of different positions in the alignment of two binary strings (sequences) of equal length is called hamming distance among them. It can also be found by using XOR gate on corresponding bits or equivalently [4, 9].

For instance, in the two binary sequences or strings:

P → 0 1 0 0 1 0 1 1 0

Q → 1 1 0 1 0 1 0 1 0

P XOR Q → 1 0 0 1 1 1 1 0 0

Hamming distance (HD) between these 9-bit strings is 5.

### 2.7 Phylogenetic Tree

A phylogenetic or developmental tree is a diagrammatic portrayal of the transformative relationship among different taxa. The terms developmental tree, phylogenetic tree, and cladogram are regularly utilized reciprocally to mean something very similar - that is, the transformative connections among taxa.

Given a set of taxa $Z$ , a phylogenetic tree $T$ on $Z$ is a connected graph with a single path between pair of nodes (species) with no loops. Any vertex in phylogenetic tree is called a leaf its degree $d(v) = 1$ and all internal nodes have degree $d(v) = 3$ .

### 2.8 Phylogenetic network

Phylogenetic systems or networks can portray the connections among qualities, chromosomes, genomes, people, species or other taxa. An extension of phylogenetic tree with edge lengths and cycles is called a network. It is represented as a triplet $(V, E, l)$ , where $l$ is a function of edge lengths assigning real non-negative numbers to the edges [8, 9].

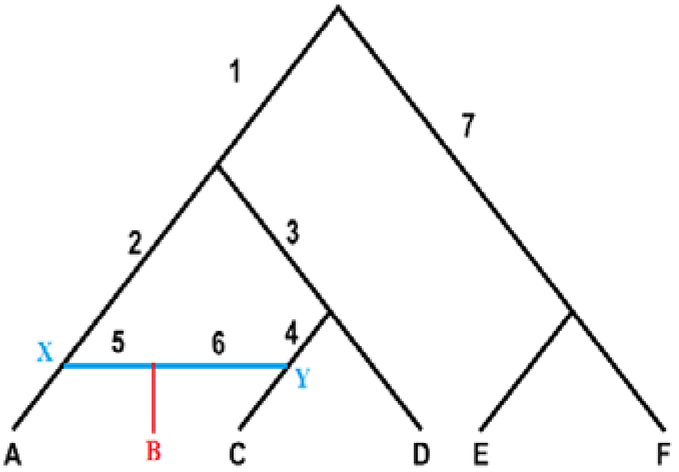


Figure 2: A Phylogenetic network with a single hybrid species B [1, 8]

## 3. Methodology of Proposed Approach

### 3.1 Conversion of DNA sequences to binary sequences

For conversion, the $'n'$ number of DNA sequences has been considered to find the evolutionary relationships between them. As these sequences are made up of four bases i.e. (A, C, G, T). The minimum number of bits required to convert these sequences to binary form is two [4, 9]. To convert the DNA sequences into binary form, a program has been coded in Python language.

### Algorithm

Now, an algorithm to convert the DNA sequences into binary sequences has been proposed.

- First store the DNA code sequences in any char-array variable say “a”.
- Now using “replace” command replace all the A’s in the DNA code sequence by 00.
- Similarly using “replace” command replace other DNA code sequence alphabets i.e. C, G and T by 01, 10, 11 respectively.
- Now the elements of variable “a” has been converted to digits.
- Lastly print the variable “a”.

Table 1: Nucleotides bases with their corresponding binary values

| S.No | Base Pair | Binary Value |
|---|---|---|
| 1 | C | 01 |
| 2 | T | 11 |
| 3 | A | 00 |
| 4 | G | 10 |

### 3.2 Complementary Sequence

The binary form of the given sequences can be changed to its complementary form by swapping 0 by 1 and 1 by 0 [9].

For Example:

| | Sequence | Binary Form |
|---|---|---|
| Original | A C G T | 00 01 10 11 |
| Complementary | T G C A | 11 10 01 00 |

After getting the binary strings of the species considered in the study, hamming distances presenting similarities/dissimilarities among them has been calculated which are then used to construct evolutionary tree and hence phylogenetic network will be established by clustering the species based on their similarities. The species having more similarity will be clustered first and vice versa.

## 4. Results and discussions

The result of the purposed approach has been discussed with the help of its implementation to a real set of biological sequences.

### Implementation of Proposed Approach to DNA sequences of six species

To test the utility of hamming distance approach purposed above, the DNA sequences of equal lengths of exon-1 of the β-globin gene for 6 real life species has been considered which are already published by many researchers in their work. Table 2 shows the accession numbers and other details of sequences available in NCBI database.

Table 2: ID Information for Exon-1 of β - globin gene of 6 species

| Species | ID/ Accession | Database | Length |
|---|---|---|---|
| Human | U01317 | NCBI | 92 |
| Lemur | M15734 | NCBI | 92 |

| | | | |
|---|---|---|---|
| Rat | X06701 | NCBI | 92 |
| Rabbit | V00882 | NCBI | 92 |
| Opossum | J03643 | NCBI | 92 |
| Gallus | V00409 | NCBI | 92 |

A program has been also coded in python language to get binary strings corresponding to their DNA sequences which are shown in Table 3. Following these binary sequences, hamming distances among them are obtained by the program and is presented in Table 4.

After finding the hamming distances, we begin the construction of the connected graph that represents the phylogenetic network.

Table 3: Conversion of DNA sequences to Binary Sequences

| Label | DNA Sequence | Binary Sequence |
|---|---|---|
| Human | ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG | 0011101011100100010111100001110101111000101000100000101101111001011011110001111001010111101110101010010000101011100000011011101000111000001011111010111010111000101001010111101010010010 |
| Lemur | ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG | 0011100001111111100111100010111001111000101000100000111001110100111011010001011101110111101110101010010000101011101000111011001000100000000101111101011101001100010100101111110101001 0010 |
| Rat | ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGCAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG | 0011101011100100010111000001111000111001111000100000101001110001111011110010111010010111101110101001000000101011100000010101111000110000111011111010011001111000101001010111101010010010 |
| Rabbit | ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTCAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGCAG | 0011101011100100110111101101010010111000101000100000101101111001101011010001111001010111101110101010010000101011010000111011101000001000000101111101011101011100010100101011110101001 0010 |
| Opossum | ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG | 0011101011100100011111100001111101111000101000100000100000011110010011010001110001010011011110101101110000101011100100101011111000010100100001111010111010111000101001010111111010010010 |
| Gallus | ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG | 0011101011100100011110100001111001111000101000100000100100100111010011010001011010010111011110101010010000101011010000111011101001011000001110111010101001011000001001010111101001010010 |

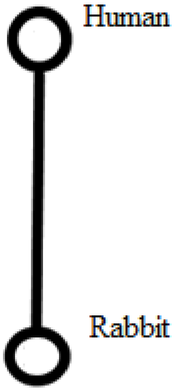
Table 4: Hamming distance (HD) among species

| HD | Human | Lemur | Rat | Rabbit | Opossum | Gallus |
|---|---|---|---|---|---|---|
| Human | 0 | 29 | 26 | 15 | 30 | 31 |
| Lemur | | 0 | 43 | 32 | 45 | 44 |
| Rat | | | 0 | 35 | 42 | 43 |
| Rabbit | | | | 0 | 43 | 34 |
| Opossum | | | | | 0 | 37 |
| Gallus | | | | | | 0 |

The species having smallest distance among them can be clustered first and so on. When two species are joined together, one species having small decimal value of the binary string will be considered as parent and other with high decimal value as child. Based on hamming distances obtained in Table 4, different cases have been explored below:

(1) When HD = 15

As the hamming distance 15 lies between the species human and rabbit and hence these species will be clustered (joined) first and human will be considered as parent of rabbit because, the decimal value of binary string of human is less than rabbit.



( a )

Figure 3: One component of the graph (network) when HD = 15

(2) When HD = 26

| Parent | Child |
|---|---|
| Rat | Human |

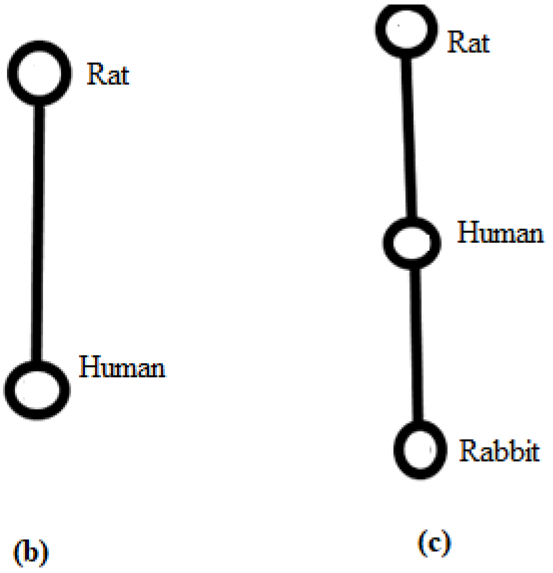


Figure 4: One component of graph for HDs=26 represented by (b) and updated component of graph showing its ancestral relationship represented by (c).

(3) When HD = 29

| **Parent** | **Child** |
|---|---|
| Lemur | Human |

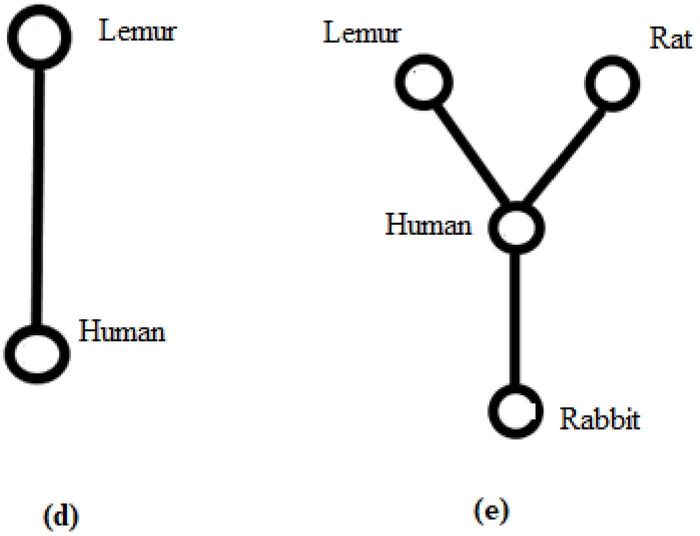


Figure 5: One component of graph for HD = 29 represented by (d) and updated component of graph showing its ancestral relationship represented by (e).

(4) When HD = 30

| **Parent** | **Child** |
|---|---|
| Human | Opossum |

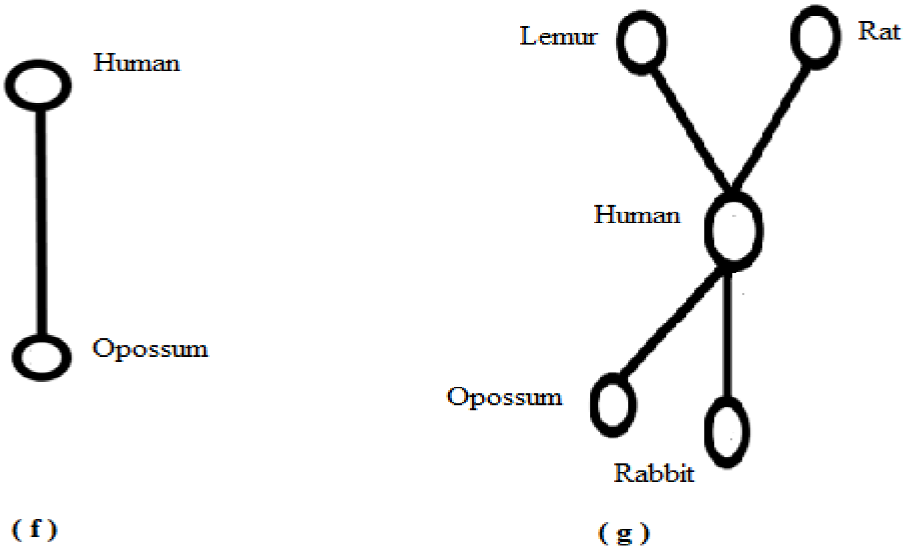


Figure 6: One component of graph for HDs=30 represented by (f) and updated component of graph showing its ancestral relationship represented by (g).

(5) When HD = 31

| **Parent** | **Child** |
|---|---|
| Human | Gallus |

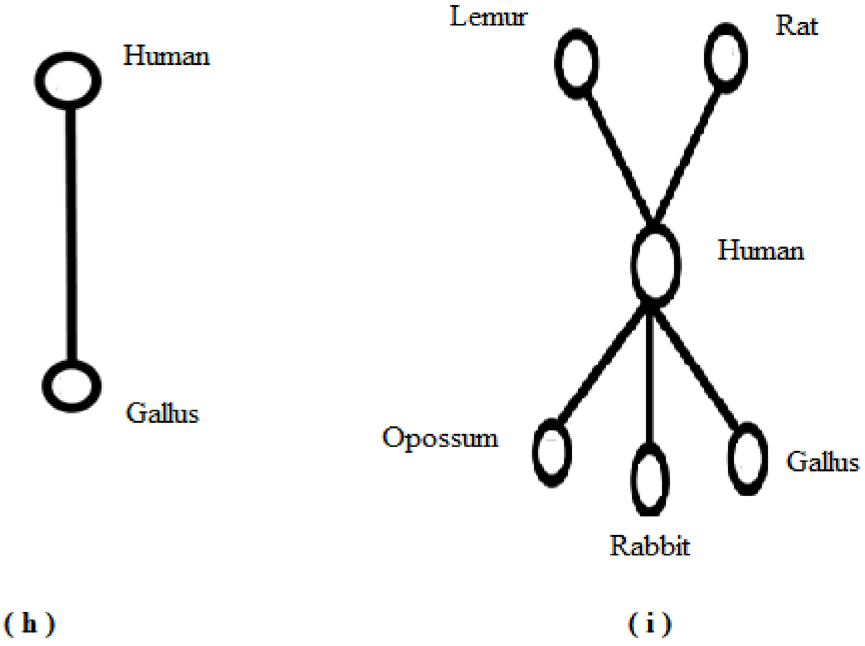
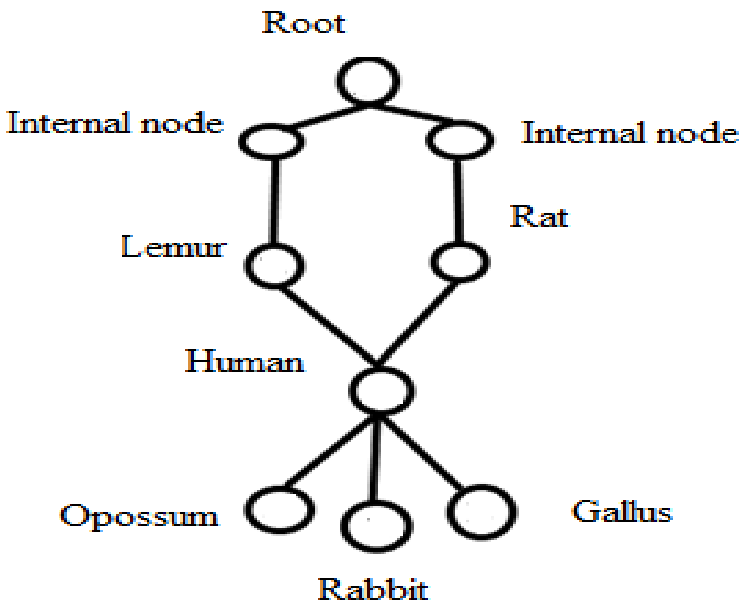


Figure 7: One component of graph for HDs=31 represented by (h) and updated component of graph showing its ancestral relationship represented by (i).

When all the species taken considered in the study are covered, then we can stop updating the graph and hence we got the final graph representing the concerned phylogenetic network among the species. So, the final network with hidden (called internal nodes) nodes and rooted nodes obtained in this work has been presented in Figure 8.



( j )

Figure 8: Phylogenetic network of all the input nodes (species) with rooted node and its ancestral relationships shown in (j).

## 5. Conclusion

In the proposed method, alignment of DNA sequences is not required to construct phylogenetic network among them. The limitation of this method is that it works only if the length of different DNA sequences is same. It works only on the extant (present) species instead of considering the ancestral ones but finds the evolutionary relationship among them and can be applied and extended to the larger dataset of sequences for wider applicability. It gives us the desired phylogenetic network based on the mathematical concept of hamming distances and can be very suitable and applicable to the biological community in future.

## References

[1]. C. R. Linder and L. H. Rieseberg, Reconstructing patterns of reticulate evolution in plants, *American Journal of Botany*, 91(10): 1700-1708, 2004.

[2]. D. Bryant and V. Moulton, Neighbor-net: an agglomerative method for the construction of phylogenetic networks, *Molecular Biology and Evolution*, 21(2): 255–265, 2004.

[3]. D. H. Huson, T. Dezulian, T. Klopper, and M. A. Steel, Phylogenetic super-networks from partial trees, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(4): 151–158, 2004.

[4]. M. A. H. Zahid, A. Mittal and R. C. Joshi, A pattern recognition-based approach for phylogenetic network construction with constrained recombination, *Pattern Recognition*, 39: 2312–2322, 2006.

[5]. O. Gascuel, *Mathematics of evolution and phylogeny*. Oxford University Press, UK, 2005.

[6]. Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, Similarity computation strategies in the micro RNA-disease network: a survey, *Briefings in Functional Genomics*, 15(1):55–64, 2016.

[7]. Q. Zou, X. B. Li, W. R. Jiang, Z.Y. Lin, G. L. Li, and K. Chen, Survey of MapReduce frame operation in bioinformatics, *Briefings in Bioinformatics*, 15(4): 637–647, 2014.

[8]. R. Mathur and N. Adlakha, A fuzzy weighted least squares approach to construct phylogenetic network among subfamilies of grass species, *Journal of Applied Mathematics & Bioinformatics*, 3(2):137-158, 2013

[9]. R. Mathur and N. Adlakha, Binary sequences-based approach for construction of evolutionary network, *International Journal of Biomathematics*, 7(2):14, 2014, DOI: 10.1142/S1793524514500120.

[10]. X. Guo, M. Randic and S. C. Basak, On the characterization of DNA primary sequences by triplet of nucleic acid bases, *Journal of Chemical Information and Computer Sciences*, 41(3): 619-626, 2001.