International Journal of Bioinformatics and Biomedical Engineering
Vol. 1, No. 3, 2015, pp. 292-299
http://www.aiscience.org/journal/ijbbe

Amino Acid Graph Representation for Efficient Safe Transfer of Multiple DNA Sequence as Pre – Order Trees

M. Yamuna*, A. Elakkiya

School of Advanced Sciences, Vellore Institute of Technology, Vellore, India

Abstract

DNA sequencing is important to apply to the human genome. It allows scientists to sequence genes and genomes. Once genes are identified and analyzed from sequence information, scientists can look for mutations that cause disease, thereby providing valuable medical information. In this paper we proposed a method of encrypting DNA sequence using pre – order tree traversal.

Keywords

Decryption, Encryption, Tree Traversal, Pre – Order

Received: September 11, 2015 / Accepted: October 10, 2015 / Published online: November 11, 2015

1. Introduction

The security to a system is essential nowadays with the growth of the Information Technology and with the emergence of new techniques the number of threats a user is supposed to deal with grew exponentially. It doesn't matter if we talk about bank accounts, social security numbers or a simple telephone call. It is important that the information is known only by the intended persons, usually the sender and the receiver. This is where the cryptography comes into picture. Cryptography is the basis of security of all the information. Cryptography is the art and science of achieving security by encoding the simple message to make it unreadable.

The security of a system is essential nowadays. With the growth of the information technology power, and with the emergence of new technologies, the number of threats a user is supposed to deal with grew exponentially. A DNA sequence is a sequence composed of four distinct letters, A, C, G and T. Each nucleotide contains a phosphate attached to a sugar molecule (deoxyribose) and one of four bases, adenine (A), cytosine (C), guanine (G), or thymine (T). It is the arrangement of the bases in a sequence, for instance like ATTGCCAT, that determines the encoded gene. Ms. Amruta D. Umalkar et. al., proposed a message cryptography formula supported deoxyribonucleic acid (Deoxy ribo Nucleic Acid) sequence for presenting during this paper. The most purpose of this formula is to write the message with the premise of complementary rules deoxyribonucleic acid sequence [1]. Behnam Bazli et. al., proposed a DNA encryption schemes and use of biological alphabets to manipulate information by employing the DNA sequence reaction, to autonomously make a copy of its threads as an extended encryption key [2]. Snehal Javheri et. al., provided the concept of DNA is being used in encryption and decryption process. The theoretical analysis shows this method to be efficient in computation, storage and transmission; and it is very powerful in certain attacks [3]. H. Z. Hsu et. al., presented three methods, the insertion method, the complementary pair method and the substitution method using DNA sequence [4]. Ashish Gehani et. al., presented some procedure for DNA – based cryptography based on one – time – pads that are in principle unbreakable [5]. Grasha Jacob, A. Murugan, presented an

* Corresponding author
E-mail address: myamuna@vit.ac.in (M. Yamuna), elakkiyaappu@gmail.com (A. Elakkiya)

encryption scheme with DNA technology and JPEG Zigzag Coding for secure transmission of images [6].

2. Preliminary Note

2.1. Graph

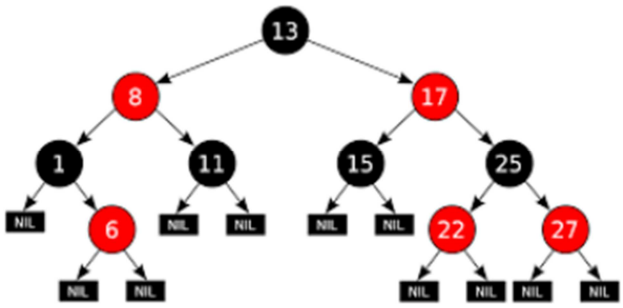
In the most common sense of the term, a graph is an ordered pair G = (V, E) compromising a set V of vertices or nodes together with a set E of edges or links, which are 2 – elements subset of V (that is an edge is related with two vertices, and the relation is represented as an unordered pair of the vertices with respect to the particular edge) [7].

2.2. Tree

A tree is an undirected graph in which any two vertices are connected by exactly one path [8].

2.3. Rooted Tree

A rooted tree is a tree with a countable number of nodes, in which a particular node is distinguished from the others and called the root node [9]. In a rooted tree, the parent of a vertex is the vertex connected to it on the path to the root; every vertex except the root has a unique parent. A child of a vertex v is a vertex of which v is the parent. A descendent of any vertex v is any vertex which is either the child of v or is (recursively) the descendent of any of the children of v. A sibling to a vertex v is any other vertex on the tree which has the same parent as v [10]. Snapshot – 1 [11] provides an example for rooted tree.



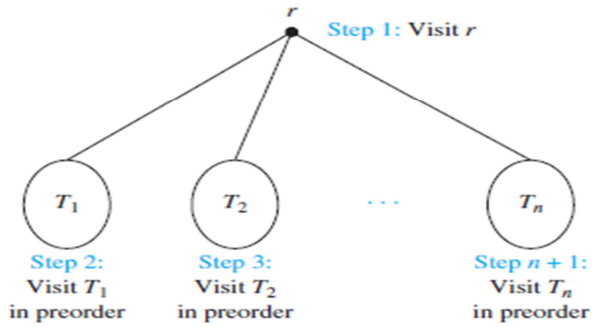
Snapshot 1. Rooted tree.

2.4. Tree Traversal

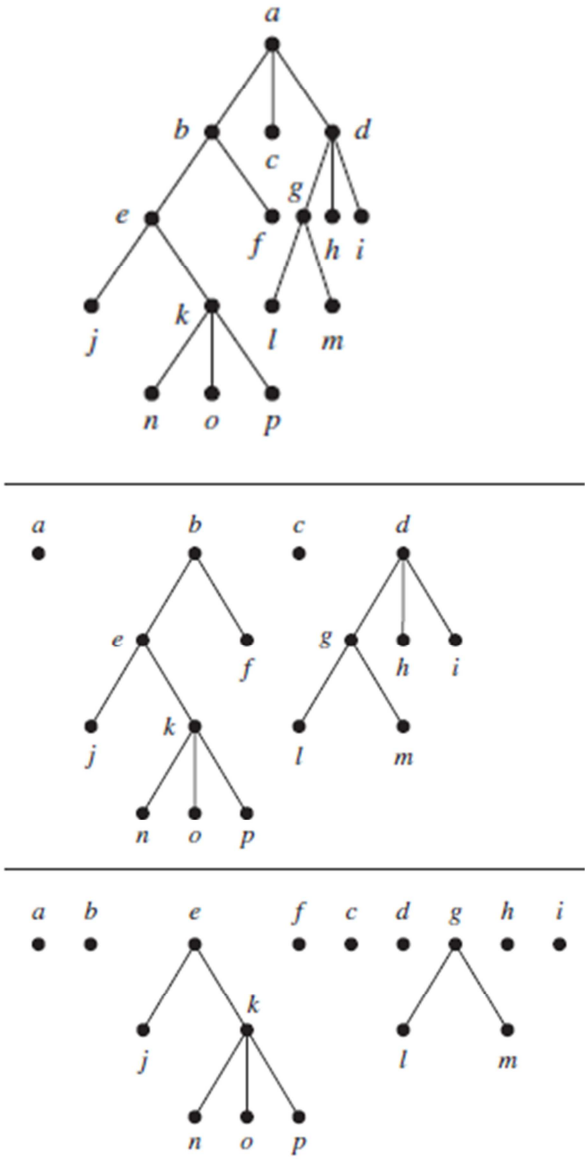
Tree traversal also known as Arkiletian tree search is a form of graph traversal and refers to the process of visiting each node in a tree data structure, exactly once, in a systematic way. Such traversals are classified by the order in which the nodes are visited [12].

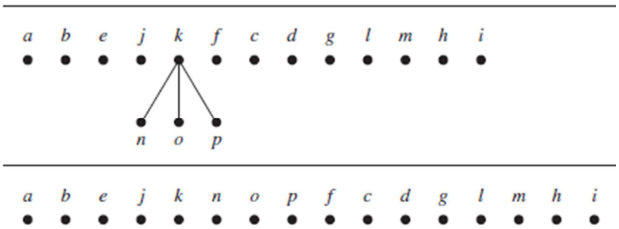
2.5. Pre – Order

Let T be an ordered rooted tree with root r. If T consists only of r, then r is the preorder traversal of T. Otherwise, suppose that T_1, T_2, . . . , T_n are the subtrees at r from left to right in T . The preorder traversal begins by visiting r. It continues by traversing T_1 in preorder, then T_2 in preorder, and so on, until T_n is traversed in preorder [13]. Snapshot - 2 [13] provides an example of pre – order tree traversal.



Preorder traversal: Visit root, visit subtrees left to right

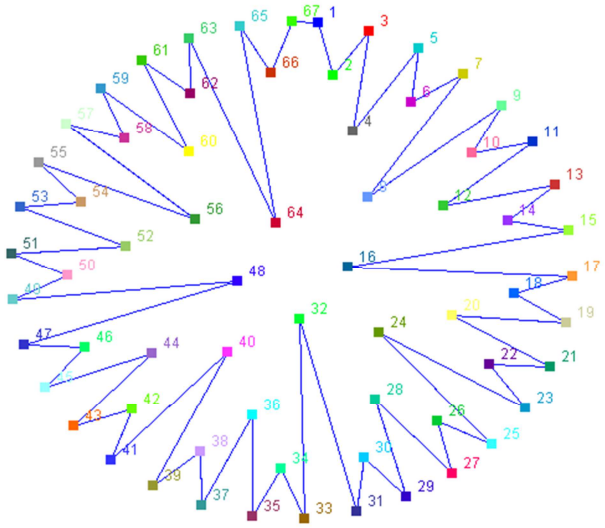




Snapshot 2. Pre – Order tree traversal technique.

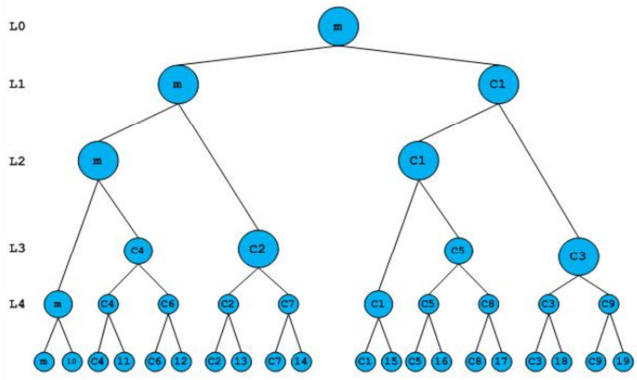
2.6. Degree

The degree (or valency) of a vertex of a graph is the number of edges incident to the vertex, with loops counted twice. The degree of a vertex v is denoted deg (v) or deg v. The maximum degree of a graph G, denoted by Δ(G), and the minimum degree of a graph, denoted by δ(G), are the maximum and minimum degree of its vertices [14]. In Snapshot [15] degree of each vertices is 2.



Snapshot 3. Graph with same degree vertices.

2.7. Tree Levels



Snapshot 4. Tree level illustration.

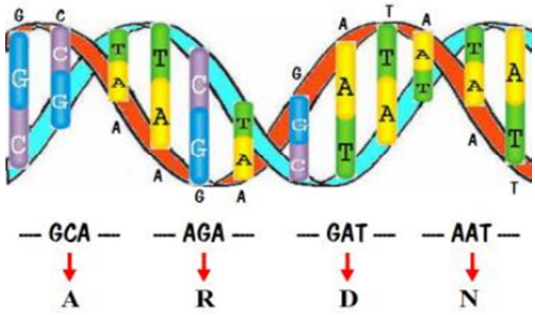
The level of a tree as the number of parent nodes a tree node has. The root of the tree, therefore, is at level 0. Root's children are at level 1, etc. In general, each level of a binary tree can have, at most, 2^N nodes, where N is the level of the tree [16]. Snapshot [17] provides an example of tree levels.

2.8. DNA Sequencing

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases — adenine, guanine, cytosine, and thymine - in a strand of DNA [18].

2.9. Protein Sequencing

Protein sequencing is a technique to determine the amino acid sequence of a protein, as well as which conformation the protein adopts and the extent to which it is complexed with any non-peptide molecules [19]. Snapshot - 3 [20] provides an example for DNA sequence to protein sequence.



Snapshot 5. DNA to protein.

3. Graph Construction for Chemical Structure

Consider the chemical structure. Replace each bond (either single or double) by an edge. Each bond is represented as an edge between two vertices.

For example consider the chemical structure of Phenylalanine. This is converted into the graph as seen in Fig. 1.

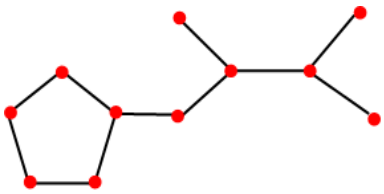
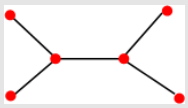
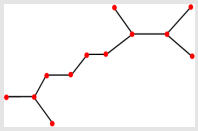
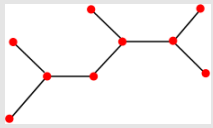
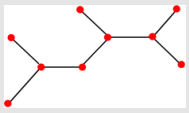
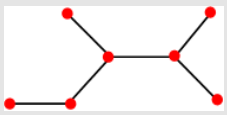
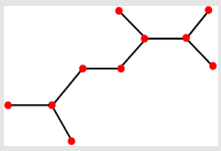
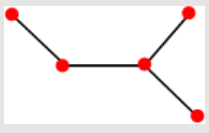
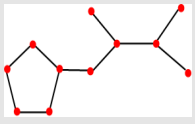
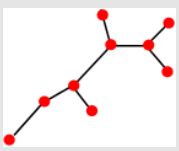
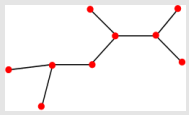
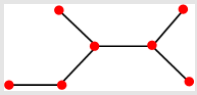
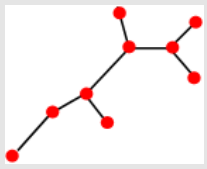
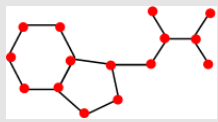
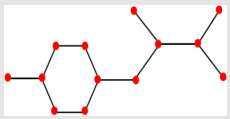
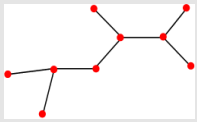


Fig. 1. Graph of phenylalanine

Similarly by using the same procedure the periodic table generated is as seen in Table - 1. Snapshot – 4 [21] represents the amino acids and its corresponding chemical structure.

4. Graph - Periodic Chart of Amino Acids

Table 1. Graph representation of amino acids.

S.No	Amino Acids	Conversion	Graph Representation
1	Alanine	07	
2	Arginine	03	
3	Asparagines	16	
4	Aspartic acid	02	
5	Cysteine	09	
6	Glutamine	06	
7	Glutamic acid	04	
8	Glycine	08	
9	Histidine	01	
10	Isoleucine	17	
11	Leucine	12	
12	Lysine	10	
13	Methionine	14	
14	Phenylalanine	05	
15	Proline	20	
16	Serine	15	
17	Threonine	11	
18	Tryptophan	19	
19	Tyrosine	13	
20	Valine	18	

5. Sequence Determination Using Periodic Chart

1. Randomly assign numbers from 1 to 20 to the amino acids in the table.
2. Since two decimal numbers are included, we label the

amino acids numbers as 01, 02,…,09, 10,…20.

3. Construct the string of numbers XY where X represents the amino acid number, Y a random sequence of numbers, length of Y = number of vertices in the graph representing the amino acids.
4. Denote these sequence as S_1, S_2, …, S_{20}. For the amino acid Glycine, S_8: X_8Y_8: 0815389

6. Encryption Algorithm

Let M be the sequence to be encrypted.

Let M = GAA UCU ACU GGT ACU GGC be the sequence to be encrypted.

Step 1 Convert the DNA sequence into protein sequence M1.

In our example for the sequence M

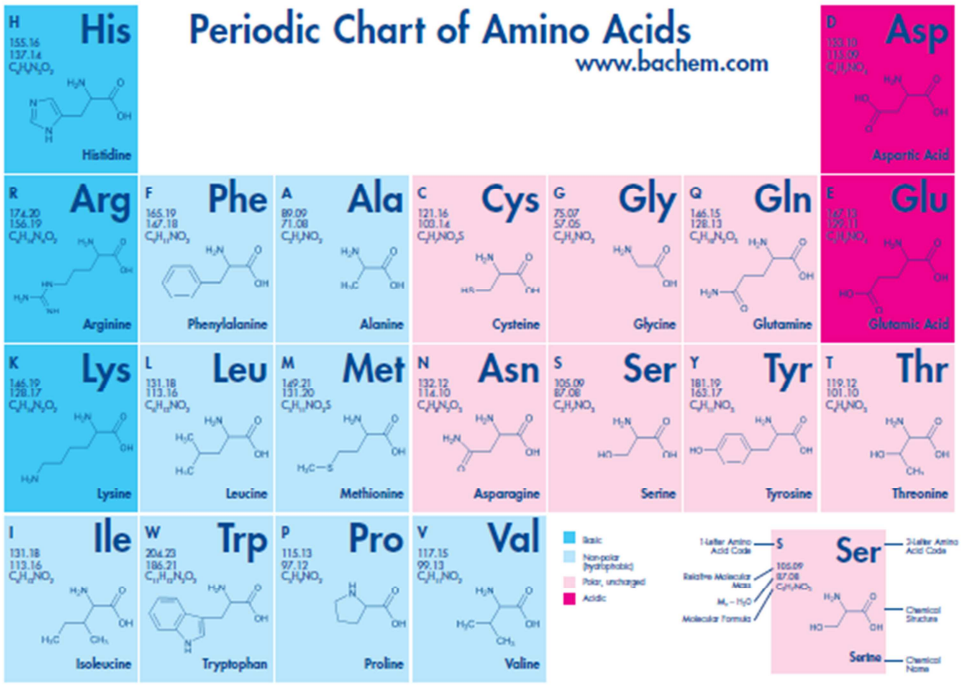
M1: glu ser thr gly thr gly.

Step 2 Replace each protein sequence by its sequence value Si as explained in above section

(Sequence determination from the periodic chart).

For our example M1= 043457839865 153894564 11237658468 0823748 11237658468 0823748. (Numbers 1 to 9 are represented as 01 to 09).

Step 3 Obtain a sequence M2 by concatenating M1.



Snapshot 6. Periodic chart of amino acids.

For our example M2: 043457839865153894564 112376584680823748112376584680823748

Step 4 Choose a tree T with k – vertices where k = Length of M2.

In our example length of M2 = length of (043457839865153894564112376584680823748112376584 680823748) = 51. So k = 51. (This count is based on the fact that the numbers Xi from 01 to 20 are considered as a single count highlighted in blue colour)

Step 5 Given the sequence 1 to 51, the pre – order tree generated for this sequence as explained in preliminary note is as seen in Fig. 2.

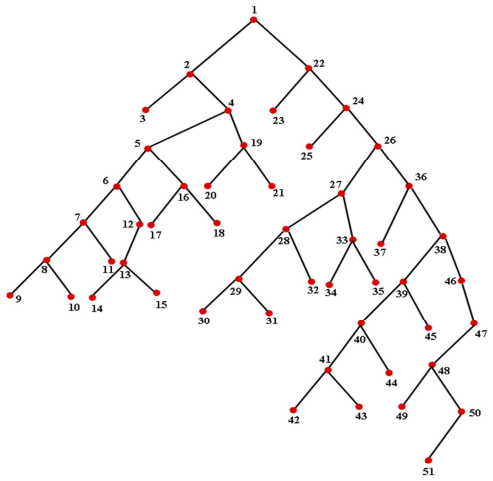


Fig. 2. Pre – Order tree with 51 vertices.

Converting these numbers into our original sequence

04 3 4 5 7 8 3 9 8 6 5 15 3 8 9 4 5 6 4 11 2 3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

7 6 5 8 4 6 8 08 2 3 7 4 8 11 2 3 7 6
23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40

5 8 4 6 8 08 2 3 7 4 8
41 42 43 44 45 46 47 48 49 50 51

We generate the encrypted tree as seen in Fig. 3.

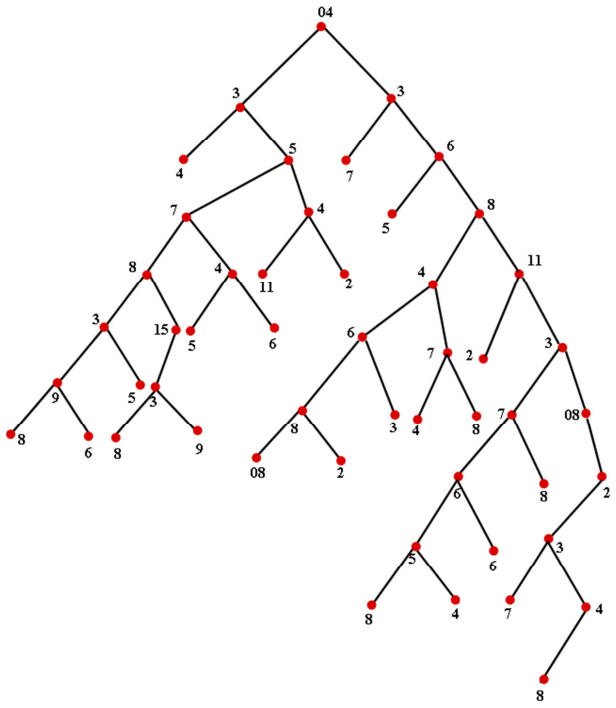


Fig. 3. Encrypted tree.

Step 6 Send the tree seen in Fig. 3 to the receiver.

6.1 Multiple sequences Encryption

Let S1, S2, …, Sk be the sequences to be encrypted. For example let S1 =GCTTGCGGA, S2 = CCCTCGGGG be the sequence to be encrypted.

Let M = GCTTGCGGACCCTCGGGG (Each sequence is assigned different colours for understanding purpose).

Step 1 Convert the DNA sequence into protein sequence M1 as in step 1 of section 6.

In our example for the sequence M

M1: Ala Cys Gly Pro Ser Gly.

Step 2 As in step 2 of above section replace each protein by its corresponding sequence value Si. For our example M1 = 07345671 091235789 0813695 2056798123 151456897 0813695.

Since we need to encrypt k sequence, we choose a tree T as follows.

Step 3 Obtain a sequence M2 by concatenating M1.

For our example M2: 07345671091235789 0813695205679812315145689708136 95.

Step 4 Since the number of sequence to be encrypted is k. We choose deg (r) = k.

In our example, since we are encrypting two sequence degree of root = 2.

Let level of r = 0. By the way we have picked our tree there are k vertices in level one. Label them as n1, n2, …, nk.

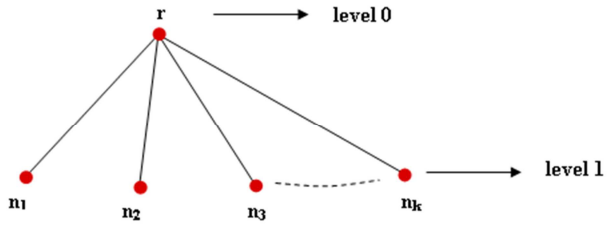


Fig. 4. Level 0, level 1 for encryption.

Step 5 Pick the first sequence S_1, starting from the root r, assign sequence S_1 as vertex label to n_1 and its descendants using pre – order labeling, (as explained in 2.6) starting from n_2, assign S_2 as vertex labels to n_2 and its descendants using pre order labeling, …, starting from n_k, assign S_k as vertex labels to n_k and its descendants using pre order labeling. Note that the root r is included only in sequence S_1. In general n_i and its descendants represent sequence S_i.

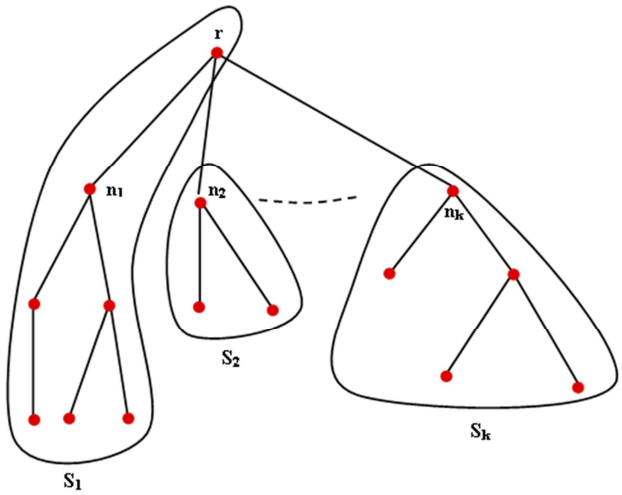


Fig. 5. Representation for multiple sequence encryption.

For our example, since two sequence are to be encrypted and length of S_1 + length of S_2 = 44, we choose a tree with 44 vertices, 2 vertices at level 1. Using pre – order labeling, we assigned numbers 1 to 44 as seen in Fig. 6.

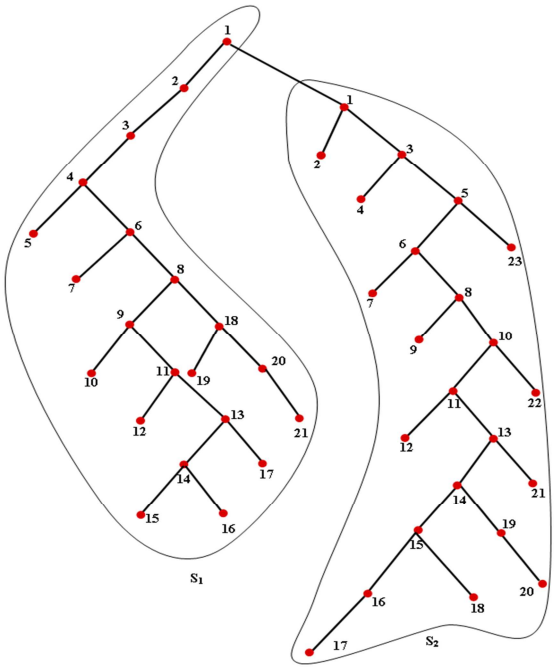


Fig. 6. Pre Order tree with 44 vertices.

Converting these numbers into our original sequence

07	3	4	5	6	7	1	09	1	2	3	5	7	8	9	08	1	3	6	9	5
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21

20	5	6	7	9	8	1	2	3	15	1	4	5	6	8	9	7	08	1	3	6	9	5
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23

We generate the encrypted tree as seen in Fig. 7.

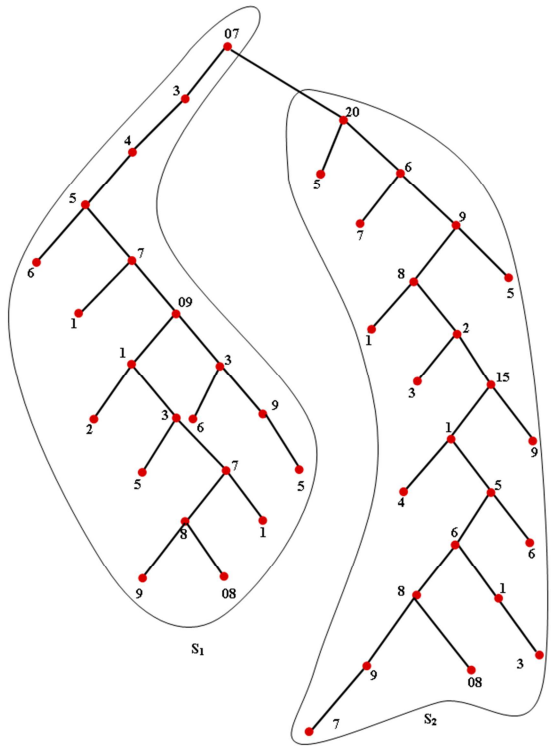


Fig. 7. Encrypted tree.

Step 6 Send the tree seen in Fig.7 to the receiver.

7. Decryption Algorithm

For decrypting the sequence we reverse the procedure.

Suppose the received sequence is as seen in Fig. 7 and the receiver is intimated that the tree represents the single sequence,

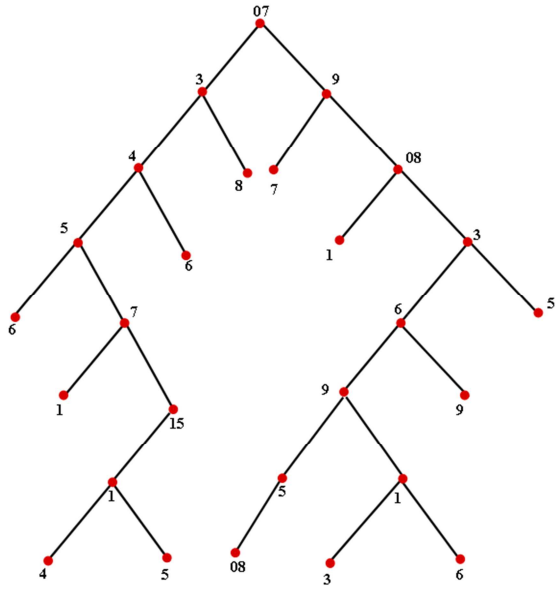


Fig. 8. Encrypted tree.

Step 1 For the tree in Fig. 8, the pre – order tree is

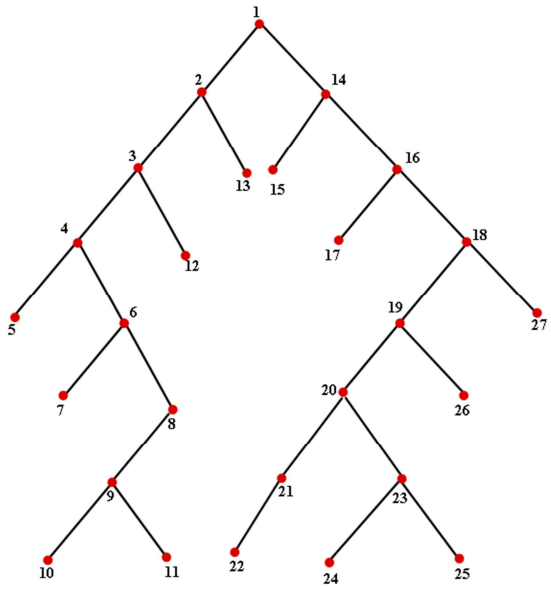


Fig. 9. Pre – Order tree with 27 vertices.

So the numerical sequence generated is 07345671 151456897 0813695 0813695.

Step 2 Convert these numbers into sequence as explained in the encryption algorithm.

In our example the first number is 07, this represents Ala. Looking into Table – 1 the graph representation of Alanine is

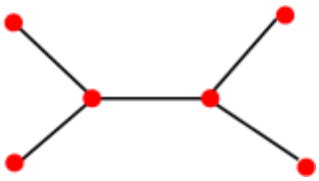


Fig. 10. Graph of alanine.

The number of vertices in the graph representing Alanine is 6. So in our sequence we separate sequence of length 6 after 07. The resulting sequence is 07345671. The next number is 15, this represents Ser. Looking into Table – 1 the graph representation of Serine is

Fig. 11. Graph of serine.

The number of vertices in the graph representing Serine is 7. So in our sequence we separate sequence of length 7 after 15. The resulting sequence is 151456897. The next number is 08, this represents Gly. Looking into Table – 1 the graph representation of Glycine is

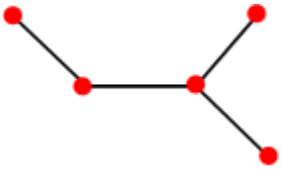


Fig. 12. Graph of glycine

The number of vertices in the graph representing Glycine is 5. So in our sequence we separate sequence of length 5 after 08. The resulting sequence is 0813695. The next number is 08, this represents Gly. Looking into Table – 1 the graph representation of Glycine is

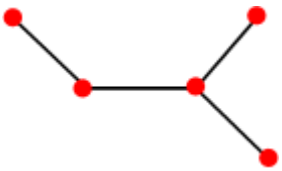


Fig. 13. Graph of glycine

The number of vertices in the graph representing Glycine is 5. So in our sequence we separate sequence of length 5 after 08. The resulting sequence is 0813695.

The sequence is decrypted as Ala Ser Gly Gly and its corresponding DNA sequence is GCTTCTGGTGGC.

8. Conclusion

DNA sequence plays an important role inventing medicines and so safe transfer of DNA sequences is necessary. The method is seemed to be so secure that it would be very difficult for any intrude to break the encrypted message and retrieve the actual message.

References

[1] Amruta D. Umalkar, Pritish A. Tijare, Data Encryption Using DNA Sequences Based On Complementary Rules – A Review, International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014.

[2] Behnam Bazli, Mustafa Anil Tuncel and David Llewellyn-Jones, Data Encryption Using Bio Molecular Information, International Journal on Cryptography and Information Security (IJCIS), Vol. 4, No. 3, September 2014.

[3] Snehal Javheri, Rahul Kulkarni, Secure Data communication and Cryptography based on DNA based Message Encoding, International Journal of Computer Application, Volume 98– No.16, July 2014.

[4] par.cse.nsysu.edu.tw/~algo/paper/paper06/A21.pdf.

[5] https://www.cs.duke.edu/~reif/paper/DNAcrypt/DNA5.DNAcrypt.pdf.

[6] Grasha Jacob, A. Murugan, An Encryption Scheme with DNA Technology and JPEG Zigzag Coding for Secure Transmission of Images.

[7] http://en.wikipedia.org/wiki/Graph_theory.

[8] https://www.google.co.in/?gws_rd=ssl#q=tree+in+graph+theory.

[9] https://proofwiki.org/wiki/Definition:Rooted_Tree.

[10] https://en.wikipedia.org/wiki/Tree_% 28graph_theory%29.

[11] https://encryptedtbn3.gstatic.com/images.

[12] https://en.wikipedia.org/wiki/Tree_traversal.

[13] Kenneth H. Rosen, Discrete Mathematics and its Application, 7th edition McGraw-Hill, New York, 2012.

[14] https://en.wikipedia.org/wiki/Degree_%28graph_theory%29.

[15] http://www.mathcove.net/petersen/lessons/images/c_67.gif.

[16] https://www.eecis.udel.edu/~breech/contest.inet.fall.01/problems/bin-tree-level.html.

[17] http://geeksforgeeks.org/wp-content/uploads/forkPuzzle4.jpg

[18] https://en.wikipedia.org/wiki/DNA_sequencing.

[19] https://en.wikipedia.org/wiki/Protein_ sequencing.

[20] http://stevemorse.org/genetealogy/dna _files/image014.jpg.

[21] https://sites.google.com/site/bioboy7/ AA. jpg.