International Journal of ChemTech Research
CODEN(USA): IJCRGG ISSN : 0974-4290
Vol.6, No.1, pp 53-63, Jan-March 2014

Genetic Code for Amino Acids using Huffman Trees

M. Yamuna*, B. Joseph Sasikanth Reddy,

Nithin Kumar Reddy, Paladugula Raghuram

VIT University, Vellore, Tamilnadu, India

***Corres.author: myamuna@vit.ac.in**
Mobile: 9894205471

Abstract: The genetic code consists of 64 triplets of nucleotides called codons. The genetic code can be expressed as either RNA codons or DNA codons. Today communication system demands transfer of various details in public domain. So need to encrypt any kind of detail becomes unavoidable. Encryption of any DNA sequence is also necessary in many cases because it carries all the genetic information. In this paper we provide genetic code for amino acids using Huffman Codes and use it for encrypting any DNA sequence.
Keywords: DNA, RNA, Amino Acid, Genetic Code, Huffmann Code.

1. INTRODUCTION

The genetic code is the set of rules by which information encoded within genetic material(DNAor mRNA sequences) is translated into proteins by living cells. The genetic code is highly similar among all organisms and can be expressed in a simple table with 64 entries[1]. There are many circumstances, like DNA testing etc, where sending information about a DNA strand becomes a need. Many times it need to be send confidentially. The main aim of this paper is to provide a genetic code for the twenty amino acids then provide a new table with all the 64 entries and hence use it to encrypt any DNA sequence.

1.1 PRELIMINARY NOTE

In this section we provide a brief discussion about amino acids, binary trees and Huffmann code that is used in the construction of the proposed genetic code.

AMINO ACID

Amino acids play central roles both as building blocks of proteins and as intermediates in metabolism. The 20 amino acids that are found with proteins convey a vast array of chemical versatility. The chemical properties of the amino acids of proteins determine the biological activity of the protein[2]. All amino acids can be converted into tree structures. The amino acids and their tree structures is provided in[2].

BINARY TREE

A node is a structure which may contain a value or condition, or represent a separate data structure. An internal node (also known as an inner node, in node for short, or branch node) is any node of a tree that has child nodes. Similarly, an external node (also known as an outer node, leaf node, or terminal node) is any node that does not have child nodes. The topmost node in a tree is called the root node. The height of a node is the length of the longest downward path to a leaf from that node. The height of the root is the height of the tree[3].

HUFFMAN CODE

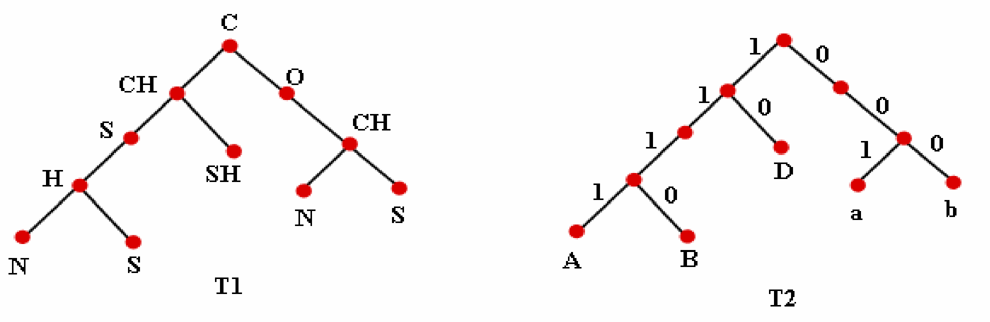
In computer science and information theory, Huffman coding is an entropy encoding algorithm used for lossless data compression. The term refers to the use of a variable-length code table for encoding a source symbol (such as a character in a file) where the variable-length code table has been derived in a particular way based on the estimated probability of occurrence for each possible value of the source symbol[4].

3. RESULTS AND DISCUSSIONS

We propose to find a new genetic code for amino acids, so that any detail regarding amino acids can be encrypted. We use Huffman codes for this.

3. 1 HUFFMAN CODE FOR CHEMICAL STRUCTURES

Consider any chemical structure which can be represented as a tree. Now fix the root node. From this node determine the binary tree. To each left child assign a value 1 and to each right child assign a value 0. Label the leaf nodes on the left of the root node using A, B, C… and those on the right by a, b, c… from the last to first. For example consider the chemical tree structure T1(random chemical tree does not represent any chemical structure) and its Huffman representation T2.



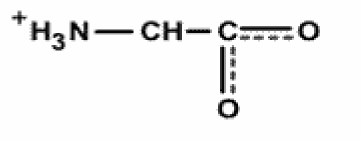
Now we represent A – 1111; B – 1110; D – 10; a – 001; b – 000

Note that a missing alphabet means that there is no leaf node at that level. Here C is missing represents that there is no leaf node in that level. This tree can be represented as **4ABD3ab**.

3.2 CONSTRUCTION OF HUFFMAN TREES FOR AMINO ACIDS

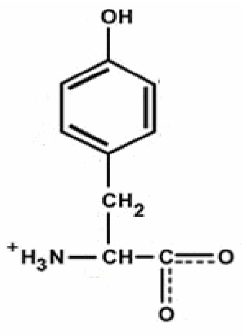
We use the above Huffman tree code for constructing Huffman trees for amino acids.

Observing the amino acid trees we notice that all the amino acids contain as a common base.

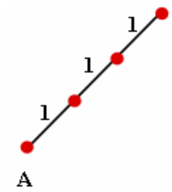


So we fix this as the root of the Huffman tree. Vertices represent the different chemical combination. The vertices represent one of C, CH, CH_2, CH_3, N, NH, NH_2, H, SH, O, OH.

Some amino acids contain cycles also as a part of the structure. This part is also included as a vertex. Since the amino acids have only one main branch, we fix it to the left of the root node. The tree is constructed as explained in section 3.1. For example consider the tree for Thyrosine



Fix the basic carbon group as the root. The other two vertices represent CH_2, OH and the cycle. So the tree can be converted as

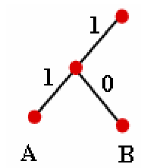


3.3 GENETIC CODE FOR AMINO ACIDS

From the Huffman tree constructed we generate the genetic code. The vertices represent one of C, CH, CH_2, CH_3, N, NH, NH_2, H, SH, O, OH and cycles of length 5, 6 and double cycles. We use the table 1 to represent them.

In the genetic code the first number represents the number of ones in the tree. As discussed in section 3.1 the number is followed by A. Then string following A represents the vertices from down from table. If there are any right leaves, they are represented by B, C … as discussed in section 3. 1. Following each alphabet the vertex labeling from down is provided.

For example the Huffman tree for Valine is



Now there are 2 ones in the tree, so the first number in the code is 2 followed by A. From the vertex labeling we see that the vertex representation following A is α3α1. The only left leaf is B. So the next alphabet in the code is B followed by α3. So the genetic code of Valine is **2Aα3α1Bα3** (coloured characters represent vertices following each alphabet)**.**

Table 2 represents the complete Huffman tree and genetic code for the basic amino acids.

3.4 ENCRYPTION ALGORITHM

Step 1 Decide the amino acid and hence the amino acid tree.

Step 2 Construct the corresponding Huffman tree.

Step 3 Write the genetic code for the amino acid using table .

Any received message can be decrypted by reversing the encryption.

4. APPLICATION

The new genetic code generated can be used to encrypt details regarding any DNA sequence. For this we construct a DNA codon table using the genetic code we have generated.

The usual DNA codon table is given in given in snapshot 1. We observe that some amino acids represent more than one codon. We suffix each occurrence of the amino acid by integers for identification purpose. Let us denote the stop codon by **1Aα** (Note that this is not used to represent any amino acid.

Table 3 provides the DNA codon table constructed using snapshot 1.

nonpolar | polar | basic | acidic | (stop codon) — **Standard genetic code**

1st base	2nd base T		2nd base C		2nd base A		2nd base G		3rd base
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)	TGA	Stop (*Opal*)	A
	TTG		TCG		TAG	Stop (*Amber*)	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

Snapshot 1

For example if the following sequence represents a part of the DNA of a human,

ATCGAATTCGCGCTGAGTCACAATTCGCGC

Dividing this into segments of length k = 3 we get

ATC GAA TTC GCG CTG AGT CAC AAT TCG CGC

Using table 1 this can be converted as

3Aα3α2α1Cα32 4Aγαα2α2Bγ1 2A6α22 1Aα34 3Aα3α1α2Bα36 2Aγ1α21 2A5α22 3Aγαα2Bβ21 2Aγ1α24 1A54

Let us use λ to differentiate between the amino acids, that is a sequence between two λ denotes an amino acid. In the above example inserting λ we obtain the sequence

3Aα3α2α1Cα32λ4Aγαα2α2Bγ1λ2A6α22λ1Aα34λ3Aα3α1α2Bα36λ2Aγ1α21λ2A5α22λ3Aγαα2Bβ21λ2Aγ1α24λ1A54 which would be send to the receiver (red colour is used to understand the blankspace which will not be used while encrypting).

Table 1: Conversion Table

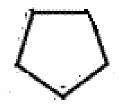
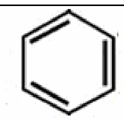
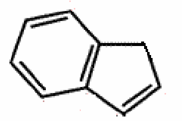
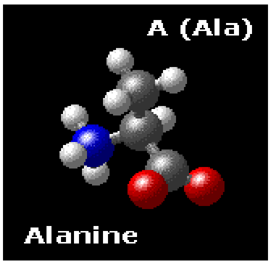
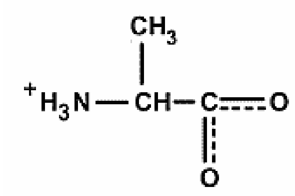
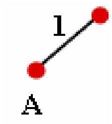
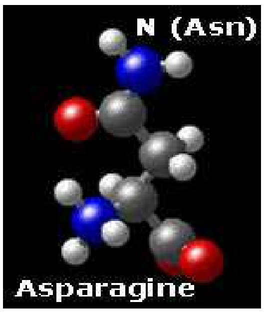
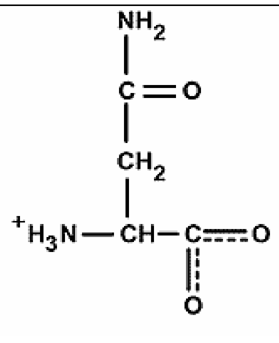
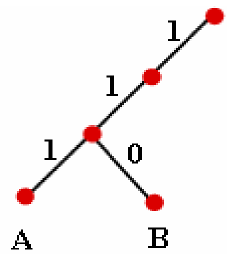
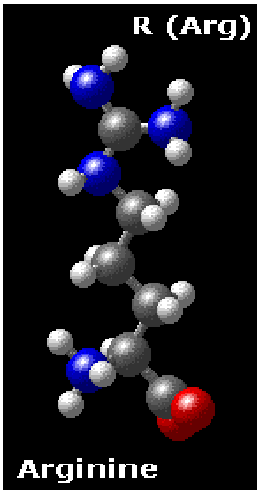
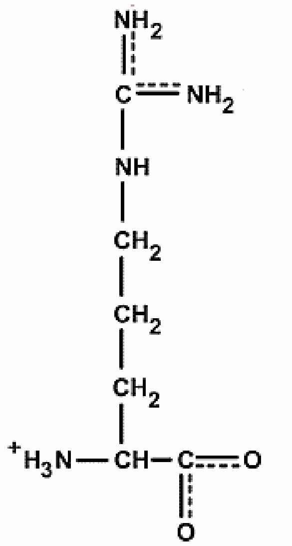
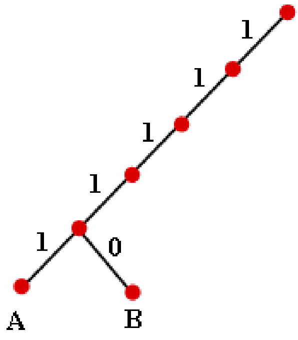
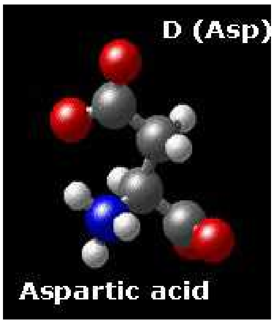
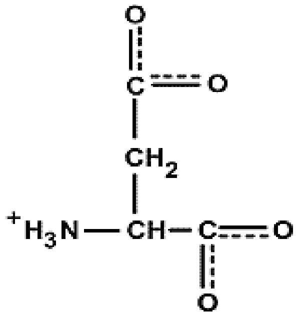
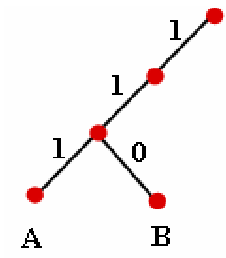
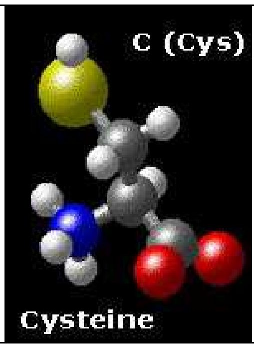
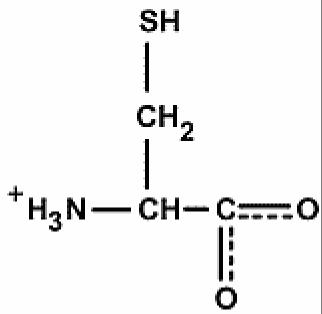
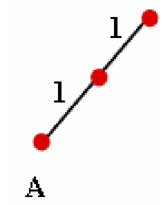
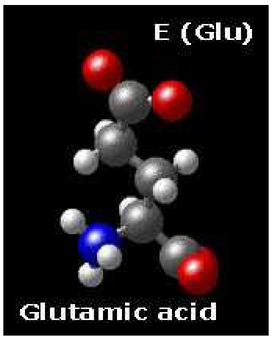
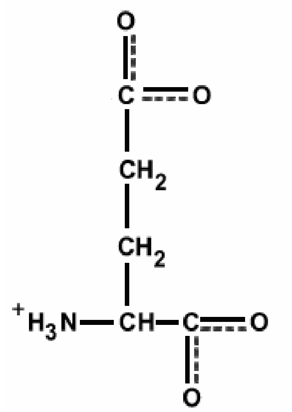
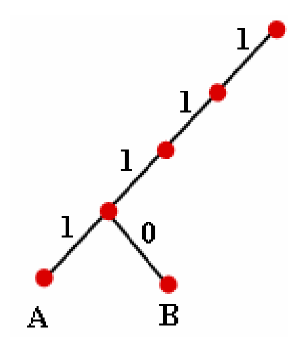
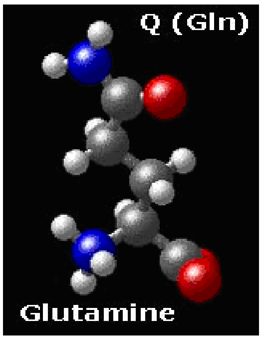
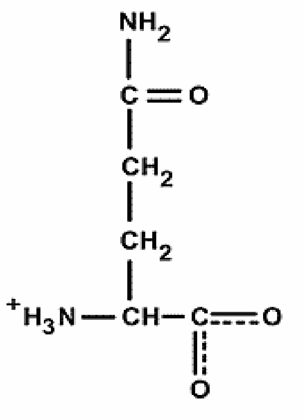
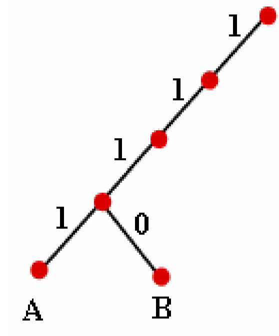
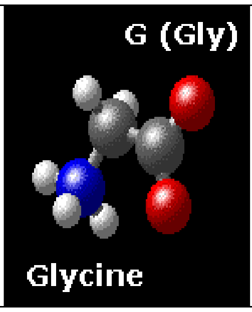
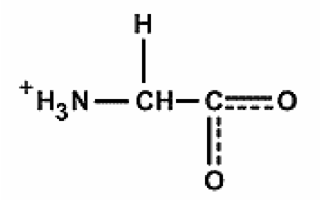
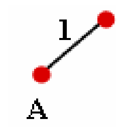
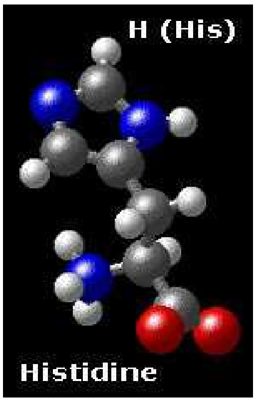
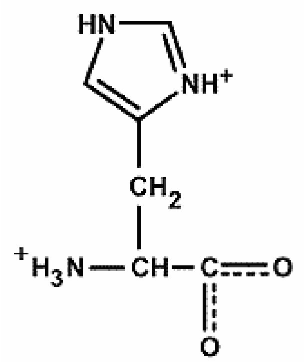
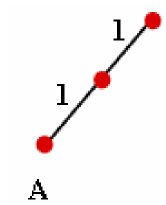
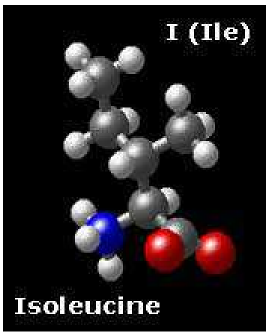
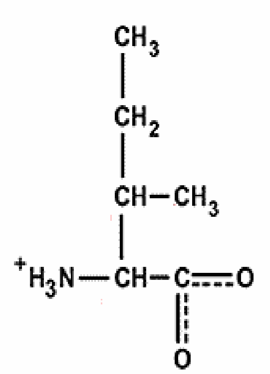
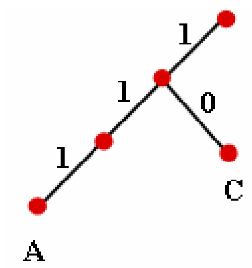
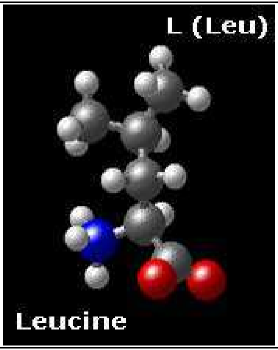
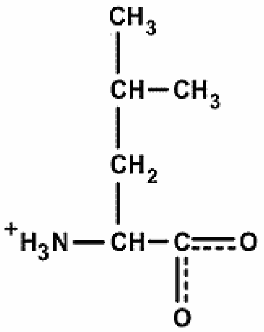
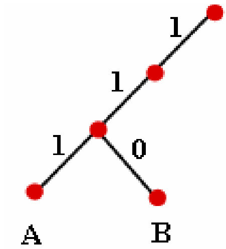
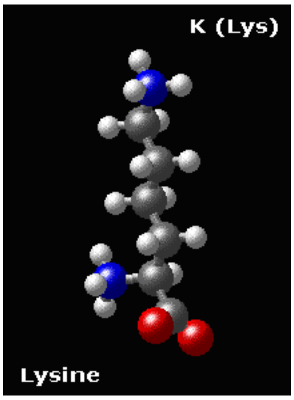
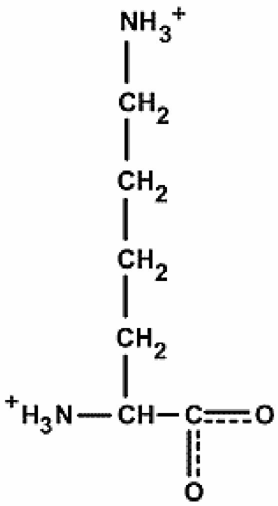
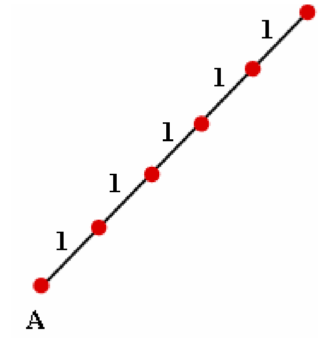
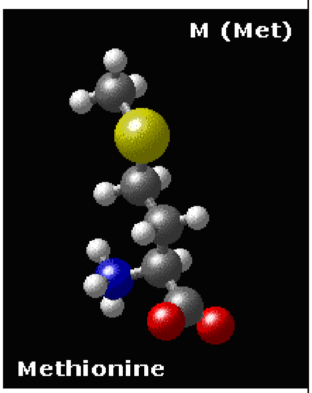
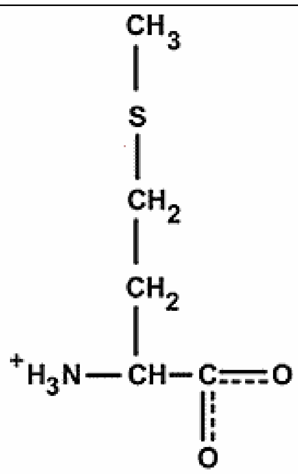
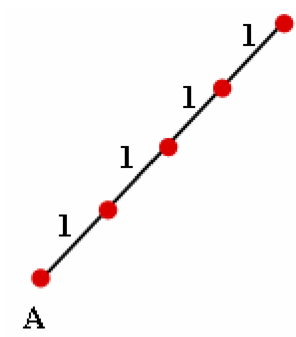
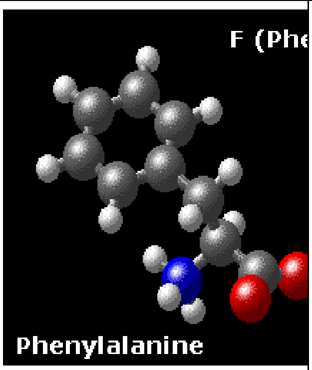
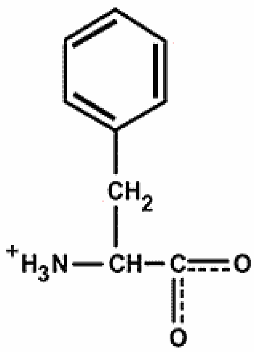
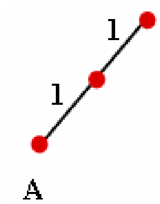
Molecule	Conversion
C	α
CH	α1
CH_2	α2
CH_3	α3
N	β
NH	β1
NH_2	β2
NH_3	β3
H	ε
S	δ
SH	δ1
O	γ
OH	γ1
	5
	6
	65

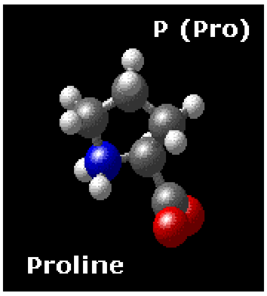
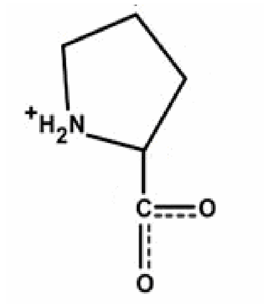
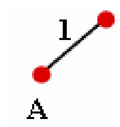
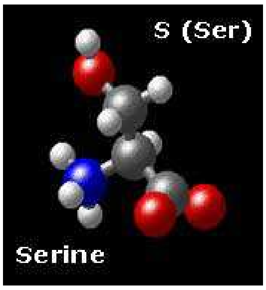
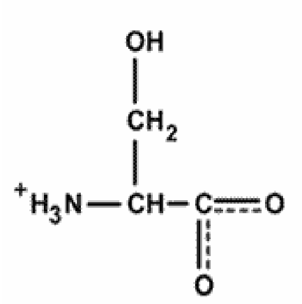
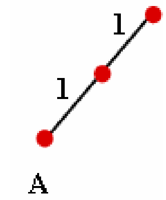
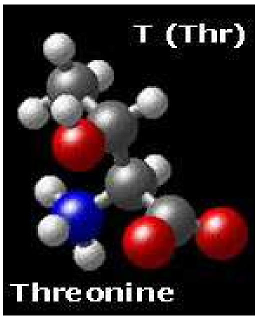
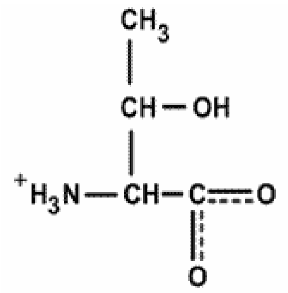
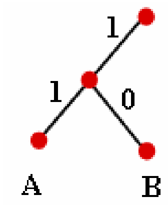
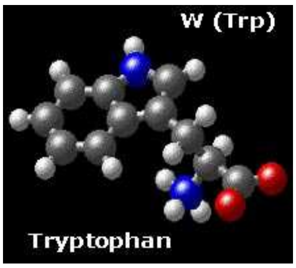
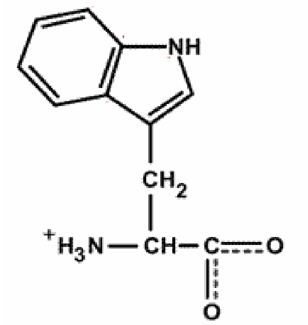
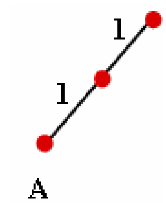
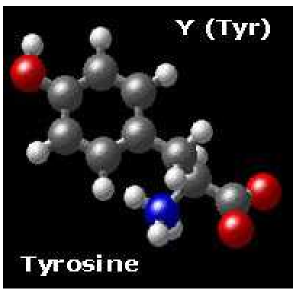
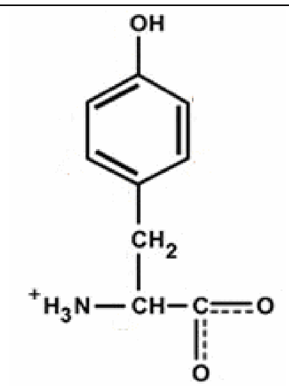
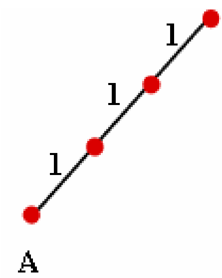
Table 2: Genetic Code Table

Amino Acid	Amino Acid Tree	Huffman Tree	Message
			1Aα3
			3Aβ2αα 2Bγ

			5Aβ2αβ 1α2α2B β2
			3Aγαα2 Bγ
			2Aδ1α2
			4Aγαα2 α2Bγ

<table>
<tr><td>
</td><td>
</td><td>
</td><td>4Aβ2αα
2α2Bγ</td></tr>
<tr><td>
</td><td>
</td><td>
</td><td>1Aε</td></tr>
<tr><td>
</td><td>
</td><td>
</td><td>2A5α2</td></tr>
<tr><td>
</td><td>
</td><td>
</td><td>3Aα3α2
α1Cα3</td></tr>
</table>

<table>
<tr>
<td></td>
<td></td>
<td></td>
<td>3Aα3α1
α2Bα3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5Aβ3α2
α2α2α2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4Aα3δα
2α2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2A6α2</td>
</tr>
</table>

<table>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
<td>1A5</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
<td>2Aγ1α2</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
<td>2Aα3α1
Bγ1</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
<td>2A65α2</td>
</tr>
<tr>
<td>
</td>
<td>
</td>
<td>
</td>
<td>3Aγ16α2</td>
</tr>
</table>

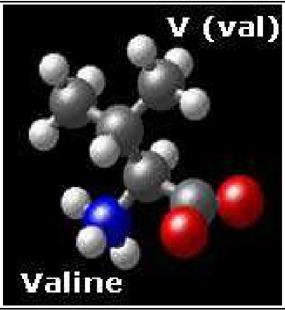
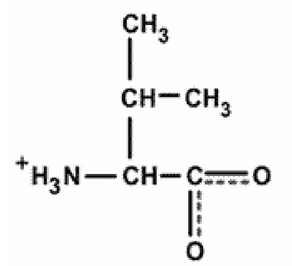
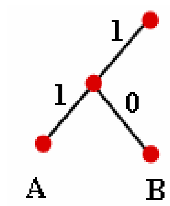
			2Aα3α1 Bα3

Table 3 : DNA Codon Table

	T		C		A		G		
T	TTT	2A6α21	TCT	2Aγ1α21	TAT	3Aγ16α21	TGT	2Aδ1α21	T
	TTC	2A6α22	TCC	2Aγ1α22	TAC	3Aγ16α22	TGC	2Aδ1α22	C
	TTA	3Aα3α1α2 Bα31	TCA	2Aγ1α23	TAA	1Aα1	TGA	1Aα3	A
	TTG	3Aα3α1α2 Bα32	TCG	2Aγ1α24	TAG	1Aα2	TGG	2A65α21	G
C	CTT	3Aα3α1α2 Bα33	CCT	1A51	CAT	2A5α21	CGT	5Aβ2αβ1α2 α2Bβ21	T
	CTC	3Aα3α1α2 Bα34	CCC	1A52	CAC	2A5α22	CGC	5Aβ2αβ1α2 α2Bβ22	C
	CTA	3Aα3α1α2 Bα35	CCA	1A53	CAA	4Aβ2αα2α 2Bγ1	CGA	5Aβ2αβ1α2 α2Bβ23	A
	CTG	3Aα3α1α2 Bα36	CCG	1A54	CAG	4Aβ2αα2α 2Bγ2	CGG	5Aβ2αβ1α2 α2Bβ24	G
A	ATT	3Aα3α2α1 Cα31	ACT	2Aα3α1Bγ 11	AAT	3Aγαα2Bβ 21	AGT	2Aγ1α21	T
	ATC	3Aα3α2α1 Cα32	ACC	2Aα3α1Bγ 12	AAC	3Aγαα2Bβ 22	AGC	2Aγ1α22	C
	ATA	3Aα3α2α1 Cα33	ACA	2Aα3α1Bγ 13	AAA	5Aβ3α2α2 α2α21	AGA	5Aβ2αβ1α2 α2Bβ25	A
	ATG	4Aα3δα2α 2	ACG	2Aα3α1Bγ 14	AAG	5Aβ3α2α2 α2α22	AGG	5Aβ2αβ1α2 α2Bβ26	G
G	GTT	2Aα3α1Bα 31	GCT	1Aα31	GAT	3Aγαα2Bγ 1	GGT	1Aε1	T
	GTC	2Aα3α1Bα 32	GCC	1Aα32	GAC	3Aγαα2Bγ 2	GGC	1Aε2	C
	GTA	2Aα3α1Bα 33	GCA	1Aα33	GAA	4Aγαα2α2 Bγ1	GGA	1Aε3	A
	GTG	2Aα3α1Bα 34	GCG	1Aα34	GAG	4Aγαα2α2 Bγ2	GGG	1Aε4	G

5. CONCLUSION

Amino acids are first converted into chemical trees, then into Huffmann trees, and then genetic code is provided based on the chemical representing the vertices. This enables decoding the amino acid. Moreover DNA is encrypted as a sequence containing numbers and roman numbers. One first need to identify that λ is used to separate to codons. Then one need to understand that it represents a Huffmann tree. Moreover, the roman numbers represent molecules. So decoding the DNA sequence becomes not possible. So the proposed method is safe for representing any DNA sequence.

REFERENCES

1. http://en.wikipedia.org/wiki/Genetic_code
2. http://www.biology.arizona.edu/biochemistry/problem_sets/aa/aa.html
3. http://en.wikipedia.org/wiki/Leaf_node#Terminology
4. http://en.wikipedia.org/wiki/Huffman_coding
5. http:// ecr-bio-lin-wikispaces.com.

*****