Volume 2, No. 4, April 2011
Journal of Global Research in Computer Science
RESEARCH PAPER
Available Online at www.jgrcs.info
ISSN-2229-371X

IMAGE HIDING IN DNA SEQUENCE USING ARITHMETIC ENCODING

Prof. Samir Kumar Bandyopadhyay[1*] and Mr. Suman Chakraborty

[1]Dept. of Computer Sc. & Engg, University of Calcutta
92 A.P.C. Road, Kolkata – 700009, India
skb1@vsnl.com
[2]B.P. Poddar Institute of Management and Technology
137, V.I.P. Road, Kolkata – 700052, India

Abstract: Recently, biological techniques become more and more popular, as they are applied to many kinds of applications, authentication protocols, biochemistry, and cryptography. One of the most interesting biology techniques is deoxyribo nucleic acid and using it in such domains. Hiding secret data in deoxyribo nucleic acid becomes an important and interesting research topic. Some researchers hide the secret data in transcribed deoxyribo nucleic acid, translated ribo nucleic acid regions, or active coding segments where it doesn't mention to modify the original sequence, but others hide data in non-transcribed deoxyribo nucleic acid, non-translated ribo nucleic acid regions, or active coding segments. Unfortunately, these schemes either alter the functionalities or modify the original deoxyribo nucleic acid sequences. DNA has the ability to store large amount of digital data. This paper presents a method to hide an image in DNA sequence using arithmetic encoding.

Keywords **:** DNA, mRNA, Arithmetic Encoding and Decoding

INTRODUCTION

Today, network technologies have improved a lot so that more and more people access the remote facilities and send or receive various kinds of digital data over the Internet. However, the Internet is a public but insecure channel to transmit data. Thus, important information must be manipulated to be concealed while delivered via the Internet such that only the authorized receiver can get it. There are two main methods for concealing secret message traditional encryption and steganography.

The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences.

DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder's rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder [1-4].

An important property of DNA is that it can replicate, or make copies of itself. Each strand of DNA in the double helix can serve as a pattern for duplicating the sequence of bases. This is critical when cells divide because each new cell needs to have an exact copy of the DNA present in the old cell. This is shown in figure 1.

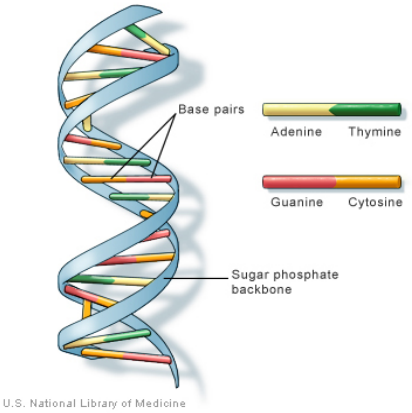


Figure 1 DNA is a double helix formed by base pairs attached to a sugar-phosphate backbone.

mRNA is transcribed from DNA, carrying information for protein synthesis. Three consecutive nucleotides in mRNA encode an amino acid or a stop signal for protein synthesis. The trinucleotide is known as a **codon**. This is shown in figure 2. 2.

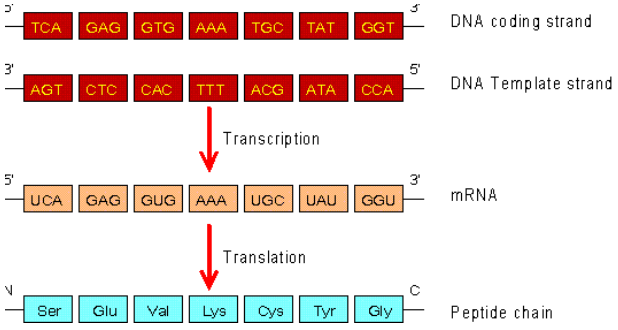


Figure -2 The sequence relationship of DNA, mRNA and the encoded peptide

The sequence of mRNA is complementary to DNA's template strand, and thus the same as DNA's coding strand, except that T is replaced by U.

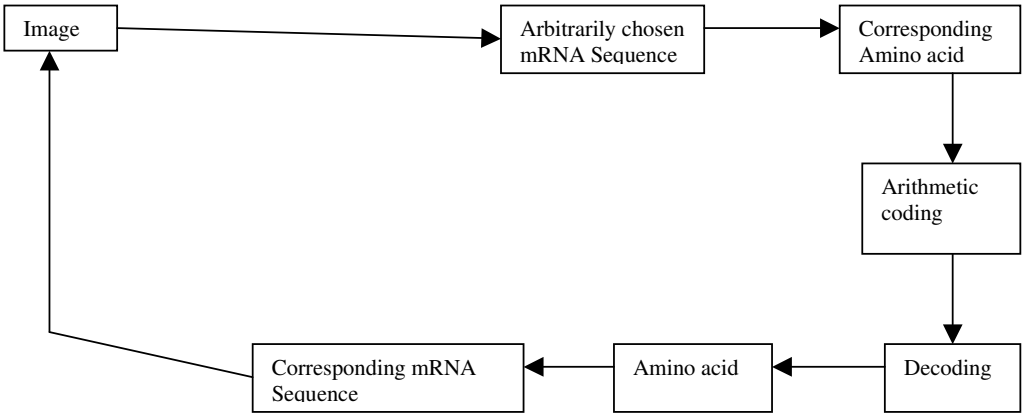


Figure 3 Image Hiding Scheme

Algorithm 1: Image embedding technique in mRNA

Step1:Choose a arbitrary mRNA sequence for binary sequence of image.

Step2: Convert mRNA sequence into amino acid sequence.

ARITHMETIC CODING

In arithmetic coding, a message is represented by an interval of real numbers between 0 and 1. As the message becomes longer, the interval needed' to represent it becomes smaller, and the number of bits needed to specify that interval grows. Successive symbols of the message reduce the size of the interval in accordance with the symbol probabilities generated by the model. The more likely symbols reduce the range by less than the unlikely symbols and hence add fewer bits to the message [5-6].

Before anything is transmitted, the range for the message is the entire interval [0, l), denoting the half-open interval 0 5 x < 1. As each symbol is processed, the range is narrowed to that portion of it allocated to the symbol.

For example, if we are going to encode the random message "BILL GATES", we would have a probability distribution that looks like this:

Character	Probability
SPACE	1/10
A	1/10
B	1/10
E	1/10
G	1/10
I	1/10
L	2/10
S	1/10
T	1/10

Once the character probabilities are known, the individual symbols need to be assigned a range along a *probability line*, which is nominally 0 to 1. It doesn't matter which characters are assigned which segment of the range, as long as it is done in the same manner by both the encoder and the decoder. The nine character symbol set use here would look like this:

Character	Probability	Range
SPACE	1/10	0.00 - 0.10
A	1/10	0.10 - 0.20
B	1/10	0.20 - 0.30
E	1/10	0.30 - 0.40
G	1/10	0.40 - 0.50
I	1/10	0.50 - 0.60
L	2/10	0.60 - 0.80
S	1/10	0.80 - 0.90
T	1/10	0.90 - 1.00

Each character is assigned the portion of the 0-1 range that corresponds to its probability of appearance. Note also that the character "owns" everything up to, but not including the higher number. So the letter 'T' in fact has the range 0.90 – 0.9999….

The most significant portion of an arithmetic coded message belongs to the first symbol to be encoded. When encoding the message "BILL GATES", the first symbol is "B". In order for the first character to be decoded properly, the final coded message has to be a number greater than or equal to 0.20 and less than 0.30. What we do to encode this number is keep track of the range that this number could fall in. So after the first character is encoded, the low end for this range is 0.20 and the high end of the range is 0.30.

After the first character is encoded, we know that our range for our output number is now bounded by the low number and the high number. What happens during the rest of the encoding process is that each new symbol to be encoded

will further restrict the possible range of the output number. The next character to be encoded, 'I', owns the range 0.50 through 0.60. If it was the first number in our message, we would set our low and high range values directly to those values. But 'I' is the second character. So what we do instead is say that 'I' owns the range that corresponds to 0.50-0.60 in the new subrange of 0.2 – 0.3. This means that the new encoded number will have to fall somewhere in the 50th to 60th percentile of the currently established range. Applying this logic will further restrict our number to the range 0.25 to 0.26.

Algorithm:

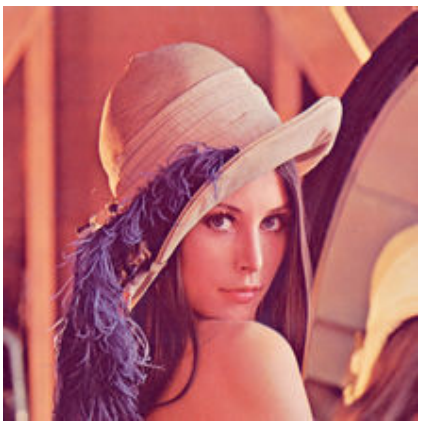
The algorithm to accomplish this for a message of any length is shown below:

```
Set low to 0.0
Set high to 1.0
While there are still input symbols do
    get an input symbol
    code_range = high - low.
    high = low +
range*high_range(symbol)
    low = low + range*low_range(symbol)
End of While
output low
```

The algorithm for decoding the incoming number looks like this:

```
get encoded number
Do
    find symbol whose range straddles
the encoded number
    output the symbol
    range = symbol low value - symbol
high value
    subtract symbol low value from
encoded number
    divide encoded number by range
until no more symbols
```

Example



For the above image corresponding mRNA sequence is "CUU CCG UGC GAU GUA GCC GGU AUC UUU GGA CAU UGG UAU AUU UCA UGC" which is chosen arbitrarily .

Converting this mRNA sequence into amino acid sequence with the help of Table-1.The amino acid sequence is " Leu Pro Cys Asp Val Ala Gly Ile Phe Gly His Trp Tyr Ile Ser Cys' .

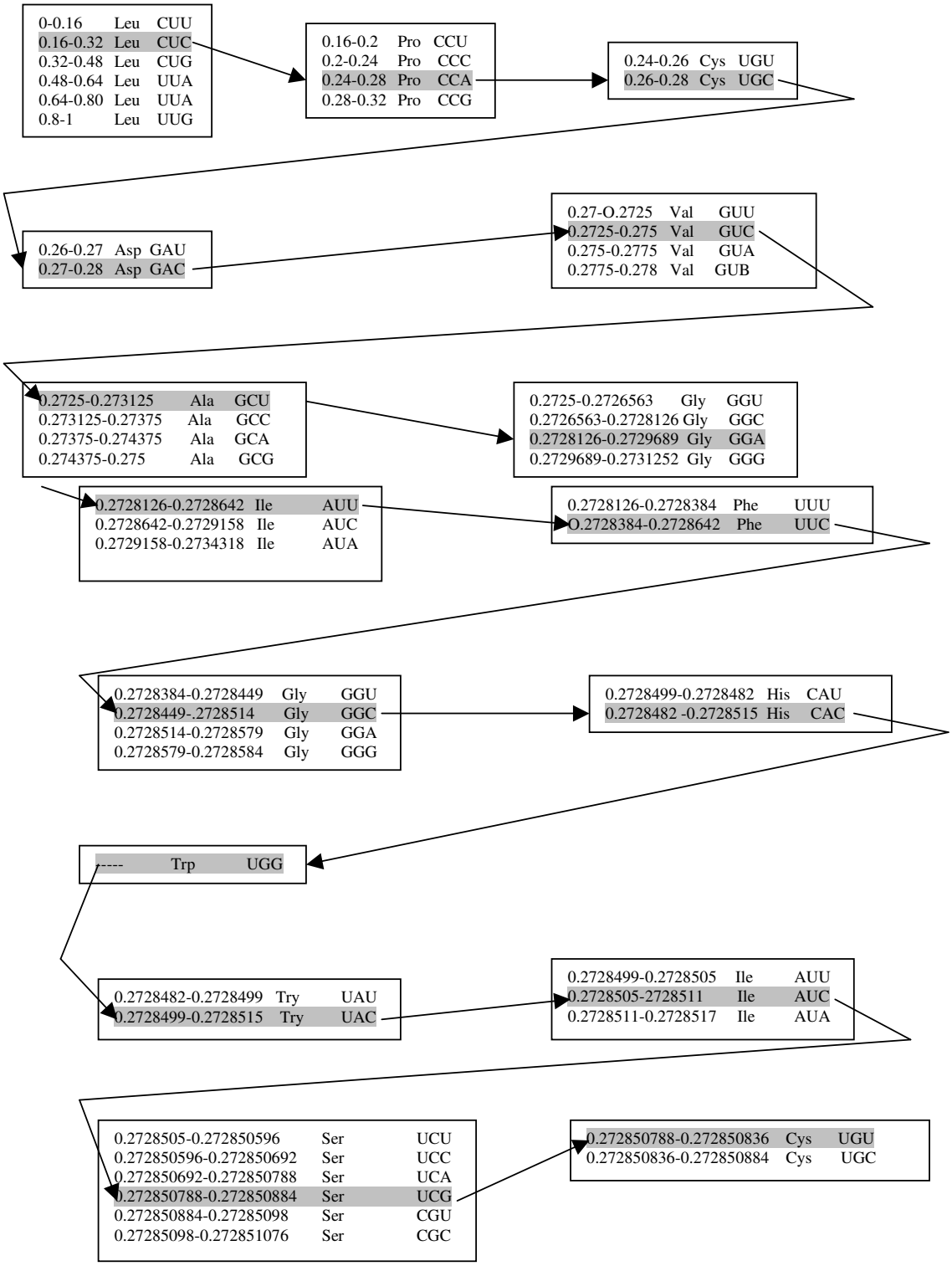
UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
UUA Leu	UCA Ser	UAA Stop	UGA Stop	A
UUG Leu	UCG Ser	UAG Stop	UGG Trp	G
CUU Leu	CCU Pro	CAU His	CGU Arg	U
CUC Leu	CCC Pro	CAC His	CGC Arg	C
CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
AUG Met	ACG Thr	AAG Lys	AGG Arg	G
GUU Val	GCU Ala	GAU Asp	GGU Gly	U
GUC Val	GCC Ala	GAC Asp	GGC Gly	C
GUA Val	GCA Ala	GAA Glu	GGA Gly	A
GUG Val	GCG Ala	GAG Glu	GGG Gly	G

Third letter

Table-1.mRNA to amino acid mapping

There are only twenty distinct amino acids, encoded from the mRNA codon. This clearly shows that some codons might be mapped to the same amino acids. For example, the codons 'GCU', 'GCC', 'GCA' and 'GCA' are mapped to the same amino acid Leu. Due to this repetition we can get different mRNA sequence for same amino acid sequence. Arithmetic coding is used for generating different mRNA and converting the image into a decimal number.

ARITHMETIC ENCODING



From this method got the different mRNA sequence of same amino acid sequence. The sequence is "CUC CCA UGC GAC GUC GCU GGA AUU UUC GGC CAG UGG UAC AUC UCG UGU". It will produce the same image.

CONCLUSION

Today, network technologies have improved a lot so that more and more people access the remote facilities and send or receive various kinds of digital data over the Internet. However, the Internet is a public but insecure channel to transmit data. Thus, important information must be manipulated to be concealed while delivered via the Internet such that only the authorized receiver can get it. There are two main methods for concealing secret message traditional encryption and steganography.

DNA has many characteristics which make it a perfect steganographic media. These techniques depend on the high randomness of the DNA to hide any message without being noticed. This paper presents a method to hide an image in DNA sequence using arithmetic encoding.

REFERENCE

[1] Peterson P., "Hiding in DNA," in Proceedings of Muse, pp. 22, 2001.
[2] Rijmen P., "Advanced Encryption Standard," in Proceedings of Federal Information Processing Standards Publications, National Institute of Standards and Technology, pp. 19-22, 2001.
[3] Rivest L., Shamir A., and Adleman L., "A Method for Obtaining Digital Signature and Public Key Cryptosystem," Computer Journal of Communications of the ACM, vol. 21, no. 2, pp. 120-126, 1978.
[4] Saeb M., El-abd E., and El-Zanaty M., "On Covert Data Communication Channels
Employing DNA Recombinant and Mutagenes is based Steganographic Techniques," Computer Journal of Bio Systems, vol. 57, no. 2, pp. 13-22, 2000
[5] Shimanovsky B., Feng J., and Potkonjak M., Hiding Data in DNA, Springer, UK, 2003.
[6] Smid E. and Branstad M., "Data Encryption Standard," in Proceedings of Federal
Information Processing Standards Publications, National Institute of Standards and Technology, pp. 550-559, 1988

