International Journal of Latest Trends in Engineering and Technology
Vol.(7)Issue(4), pp.102-107
DOI: http://dx.doi.org/10.21172/1.74.014
e-ISSN:2278-621X

ENHANCED LS BASE: A DNA BASED SECURITY SCHEME TO AUGMENT AUDIO STEGANOGRAPHY

Rashmi M. Tank[1] and Prof. Vikram Agrawal[2]

Abstract – Security is important criteria for information transmission. This paper proposes highly secure method to hide the existence of secret message to prevent unauthorized access. The proposed method has three levels. Single level of encryption and two levels of steganography are used. First level makes use of DNA based RSA encryption. Second level hides the encrypted message inside the DNA sequence using randomized LSBase substitution technique. Third level hides the embedded DNA sequence inside the audio file using LSB encoding. The main objective of this method is that no one could be able to find the existence of secret message.
Keywords- Steganography, Data hiding, Information Security, DNA Sequence, Audio Steganography

I. INTRODUCTION

The explosive growth of computer system and their interconnection via computer networks such as internet has led to a heightened need for information and data security. The growing use of internet has led to a continuous increase in the amount of data that is being exchanged and stored in various digital media. In network it is very important to protect the data. In order to achieve this there are several technologies being used. One of the secured ways to protect the data is encryption and decryption method [10]. Cryptography is the science and art of protecting information by scrambling its content into unreadable format called *ciphertext*. Cryptography system can be broadly classified into *symmetric system* that use a single key for both encryption and decryption, and *asymmetric systems* that use one key for encryption and another for decryption. The Data Encryption Standard (DES) is the most common symmetric cryptography method that is in use today. Examples of asymmetric cryptography algorithm include RSA, Diffie and Hellman and Digital Signature Algorithm (DSA) which is a variant of ElGamal. The objective of secure communication is that actual data should not be revealed to any third party. Steganography techniques are most successful technique in supporting hiding of critical information in ways that prevent the detection of hidden messages. While cryptography scrambles the message so that it cannot be understood [10].

The conventional methods of encrypting are not strong enough today for providing the data security and reliable data transmission. Unauthorized user or intruders may attack and can interrupt or intercept the message for doing some malicious tasks. In order to enhance the data security effective encryption algorithms are required. Recent research has shown DNA as a medium for large scale computation system [2]. DNA computing is a new method of simulating bimolecular structure of DNA and computing by means of molecular biological technology which is a novel and potential growth. Adleman demonstrated the first DNA computing. It marked the beginning of a new stage in the era of information. DNA (Deoxyribonucleic Acid) is the germ plasma of all life styles [9]. Two different kinds of genetic material exist, Deoxyribonucleic acid (DNA) and Ribonucleic Acid (RNA) present in a cell. DNA contains four types of nucleotides like, Adenine (A), Thymine (T), Cytosine (C) and Guanine (G). James D. Watson and Francis Crick were the two co-discoverers of the structure of DNA in 1953 [14]. In a double helix DNA string, two strands are complementary in terms of sequence, that is A to T and C to G according to Watson-Crick rules [9]. There are a large number of DNA sequences publicly available in various domains of biological DNA. A rough estimation would put the number of DNA sequences publicly available in various websites are around to be 55 million [4].

Shyamasree C M and Sheena Anees in [13] proposed DNA based playfair cipher. A binary form of data are transformed into sequences of DNA nucleotides subsequently these nucleotides pass through a playfair encryption process based on amino acids structure. The relationship between the nucleotide sequences of genes and the amino acid sequences of proteins is determined by the rules of translation, known collectively as genetic

[1] *B.V.M. Engineering college, Gujarat Technological University, Vallabh Vidhyanagar, Gujarat, India*
[2] *B.V.M. Engineering college, Gujarat Technological University, Vallabh Vidhyanagar, Gujarat, India*

code. The genetic code consists of three-letter words called codons formed from a sequence of three nucleotides. Since there are 4 bases in 3-letter combinations, there are 64 possible codons. These encode the twenty standard amino acids, giving most amino acids more than one possible codon.

One of the latest techniques called LSBase is proposed by Amal Khalifa in [1] to improve hybrid cryptosystem using DNA steganography. The session key is hidden inside a chosen DNA sequence and hence can be securely exchanged between parties through public channels such as the internet. This technique uses codon degeneracy to hide the information within DNA sequences without actually affecting the type or structure of the protein they code for. One more technique called indexing technique is proposed by K Menaka in [5] uses 3 complementary rules.

Steganography is the art and science of hiding information such that its presence cannot be detected. The secret information is hidden in some carrier file and then transmitted. The carrier file can be an image, audio file, text file, video file etc. Among different steganography schemes, the image steganography is widely used. But increase use of voice over Internet Protocol (VoIP) and various Peer-to-Peer (P2P) audio services encourage researcher to choose audio steganography. Audio steganography requires a text or audio secret message to be embedded within a cover audio message. Audio files are considered to be excellent carriers for the purpose of the steganography due to presence of redundancy. Due to availability of redundancy, the cover audio message before steganography and stego message after steganography remains same. However, audio steganography is considered more difficult than video steganography because the Human Auditory System (HAS) is more sensitive than Human Visible System (HVS). A steganography system is more expected to meet three key requirements, namely transparency, capacity and robustness. Least Significant Bit (LSB) is one of the earliest and simpler methods used for information hiding of digital audio. Traditional, LSB is based on inserting each bit from the message in the LSB of binary sequence of each sample of cover digitized audio file [10].

The objective of this paper is to come up with an efficient method to preserve security of secret messages in a text file against unauthorized access by hiding the presence of the text file. The method proposed in this paper works in three levels. The three levels are encryption, DNA steganography and audio steganography. For encrypting the text file the proposed method uses DNA based RSA encryption algorithm. In the second level, the encrypted secret file is hidden in a DNA sequence taken from publicly available database on NCBI's website. In the third level the DNA sequence which is embedded with the encrypted message is hidden inside an audio file using Least Significant Modification technique.

The motivation of preparing this paper is given in next section. The section III gives detailed description of encryption phase. The section IV illustrates a detailed overview of the DNA steganography phase of the proposed method. Details of audio steganography are given in section V. The details of extraction of original text file from the stego audio file are given in VI. Section VII gives the experimental results of the proposed method.

II. MOTIVATION

Information security is one of the important fields in which researches are taking place. Main goal of it is to prevent unauthorized access to the secret message as well as to hide the existence of the secret message. Steganography is the main method to hide the message.

DNA due to its immense storage capacity and high randomness is used now in the field of steganography. This can be considered as recent technique in steganography. A large number of researchers take an initiative for implementing DNA encoding concept in the applications like cryptography, scheduling, clustering, GPU applications, multi-core architectures, forecasting and even trying to apply this in signal and image processing algorithm [13], [4].

Since Human Auditory System (HAS) is more sensitive than Human Visible System (HVS) [6], and since audio files are redundant and highly available, audio steganography is of high importance in the field of steganography. Hence more techniques are to be found out in this field [13].

III. DNA BASED ENCRYPTION

In this section encryption algorithm is illustrated which is used to convert the secret file into encrypted DNA sequence.

In the first level of proposed method, secret message is encrypted using DNA based RSA encryption algorithm. This secret file is input to the algorithm.

The secret message in a file is converted to DNA sequence using DNA digital coding pattern. From a computational point of view, any DNA sequence can be encoded using a binary coding scheme, in which anything can be encoded by a combination of the two states 0 and 1. Therefore simplest coding pattern to encode the 4 nucleotide bases (A, U, C, G) is: 0(00), 1(01), 2(10), 3(11) respectively. Obviously, there are 4!. So, among these 24 patterns, only 8 kinds of patterns (0123/CUAG, 0123/CAUG, 0123/GUAC, 0123/GAUC, 0123/UCGA, 0123/UGCA, 0123/ACGU and 0123/AGCU) which are topologically identical fit the complementary rule of the nucleotide bases. It is suggested that the coding pattern in accordance with the

sequence of molecular weight, 0123/CUAG, is the best coding pattern for the nucleotide bases as illustrated in table 1 [10].

TABLE 1: DNA DIGITAL CODING [10]

DNA Nucleotide	Decimal	Binary
A	0	00
C	1	01
G	2	10
U	3	11

The process of DNA based encryption has the following steps:

1. Convert the contents of the secret file in to binary form by taking the ASCII value.
2. Group the binary data into groups of two bits and code binary data into DNA sequence using table 1.
3. Group the alphabets in the DNA sequence into groups of three letters (i.e. codons) and map the codons into amino acids using table 2.
4. Store the ambiguity numbers in a file.
5. Encrypt the amino acid sequence using RSA cipher method.
6. Convert the amino acids in the encrypted sequence into binary form by taking the ASCII value.
7. Group the binary form of encrypted amino acid sequence into groups of two bits and code the groups into DNA sequence using table 1.
8. Store the resulting sequence in a text file.

TABLE 2: A MAPPING OF THE DNA CODONS INTO AMINO ACIDS

Codons	**Char**	**Codons**	**Char**
GCU, GCC, GCA, GCG	A	AAU, AAC	N
UAA.UAG,UGA	B	UUA,UUG	O
UGU,UGC	C	CCU,CCC, CCA,CCG	P
GAU,GAC	D	CAA,CAG	Q
GAA,GAG	E	CGU,CGC,CGA CGG,AGA,AGG	R
UUU,UUC	F	UCU,UCC,UCA, UCG,AGU,AGC	S
GGU,GGC,GGA, GGG	G	ACU,ACC,ACA, ACG	T
CAU,CAC	H	AGA,AGG	U
AUU,AUC,AUA	I	GUU,GUC,GUA, GUG	V
-	J	UGG	W
AAA,AGG	K	AGU,AGC	X
UUA,UUG,CUU, CUC,CUA,CAG	L	UAU, UAC	Y
AUG	M	-	Z

Table 2 is used to map the DNA sequence into amino acid sequence. There are 1 to 6 possible codons to represent each amino acid in table 2. This fact will create confusion on decryption that which codon should be selected corresponding to a particular amino acid. To counter this problem ambiguity numbers are used which are 0, 1, 2, 3, 4 and 5 corresponding to the first, second, third, fourth, fifth and sixth codon respectively. These ambiguity numbers are stored in a separated file to be used in the decryption process [10].
Hence the output of the DNA based encryption is DNA based encrypted sequence and sequence of ambiguity numbers.

IV. DNA STEGANOGRAPHY

This section gives the idea about the second level of the proposed method which is the DNA steganography. The encrypted DNA sequence obtained from the first level is hidden inside a reference DNA sequence taken from publicly available database. Randomized LSBase substitution technique is used for hiding a DNA sequence in other. A condition for applying the proposed algorithm for DNA steganography is that the reference DNA

sequence should have a length at least three times greater than the length of encrypted binary sequence. The codons in the reference DNA sequence are first randomized using Linear Congruential Generator (LCG).

TABLE 3: A MAPPING OF MESSAGE BITS TO DNA NUCLEOTIDE BASES

	Least Significant Base of Codon	
Message Bits	Purine(A/G)	Pyrimidine(T/C)
000	A,0	T,0
001	A,1	T,1
010	A,2	T,2
011	A,3	T,3
100	G,0	C,0
101	G,1	C,1
110	G,2	C,2
111	G,3	C,3

Table 3 shows a mapping table for hiding 3 encrypted binary bits at a time inside the least significant base of each codon. The seed and modulus of the LCG need to be sent to the receiver securely. For that purpose it is encrypted using Elliptic Curve Cryptography (ECC).
The embedding and extraction module using an example sequence is discussed below.
Here the property of codon degeneracy (i.e. multiple codons may produce same amino acid) is utilized in order to change the codon's last base while keeping its type (either purine or pyrimidine).
Four exception should be considered. First the tryptophan (Trp), and methionine (Met) have a single codon, so they can't be used for embedding. The same is true for the stop codon UGA. The fourth case actually appears in the amino acid Isoleucine (Ile), since it is coded by three codons: AUU, AUC and AUA. Therefore, only AUU and AUC can be used while the AUA is neglected.
The embedding process starts by converting DNA sequence into RNA by replacing each T with a U. Then the sequence is divided into codon triplets. At each codon, the least significant base is inspected and if it is pyrimidine base, it is changed into to (T) to encode 3 message bits having MSB 0 with ambiguity number 0, 1, 2 and 3 respectively. Also, if message bits have MSB 1 then LSBase is changed to C with ambiguity number 0, 1, 2 and 3 respectively. If the least significant base is purine base, it is changed into to (A) to encode 3 message bits having MSB 0 with ambiguity number 0, 1, 2 and 3 respectively. Also, if sequence of 3 message bits has MSB 1 then it is changed to G with ambiguity number 0, 1, 2 and 3 respectively.

Here LSBase technique is the blind technique. So the actual reference DNA sequence needs not to be stored for extraction process. The output of DNA steganography is embedded DNA sequence in which actual encrypted DNA sequence is hidden.

V. AUDIO STEGANOGRAPHY

This section illustrates the third level of proposed method. The embedded DNA sequence obtained as the output of DNA steganography is hidden in an audio file to hide the existence of the secret data.
Least Significant Bit (LSB) modification is used for audio steganography. Audio files used are in WAV format. The audio file read in a binary format and the embedded DNA sequence is stored in the lower half of each byte of the audio file. This will create no distortion to sound.
The process of audio steganography has the following steps:

1. Read the embedded DNA sequence and the audio file in binary format.
2. Sample the audio file.
3. Encode the length of the cipher in lower half of the first 32 audio samples.
4. Encode the cipher in lower half of the remaining audio samples.

VI. EXTRACTION OF SECRET FILE FROM AUDIO DATA

This section illustrates how the secret file is extracted from the stego audio file. The extraction process consists of the reverse of all the process that had been done to hide the secret file containing the secret message.
After the encryption phase and two steganography phases, there is a sequence of ambiguity numbers and seed and modulus of LCG as output. Ambiguity numbers from the encryption phase is necessary in the decryption phase. Seed and modulus from DNA steganography is necessary for generating random numbers during extraction process.
As the first step in extraction process, the stego audio file is sampled. From the first 32 samples the length of the cipher embedded in it is decoded. Then from the remaining samples, in which the cipher is encoded, is decoded. This cipher is decrypted to obtain embedded DNA sequence.

The extraction process has following steps:
1. Decrypt the seed and modulus using Elliptic Curve Cryptography.
2. Generate random numbers using linear congruential generator.
3. Convert embedded DNA sequence into codon triplets.
4. Randomize codons based on random numbers.
5. Inspect LSBase of codons and ambiguity numbers:

If the LSBase is A then embedded message bits are 000,001,010 and 011 if ambiguity numbers are 0, 1, 2 and 3 respectively.
If the LSBase is G then embedded message bits are 100,101,110 and 111 if ambiguity numbers are 0, 1, 2 and 3 respectively.
If the LSBase is T then message bits are 000,001,010 and 011 if ambiguity number is 0, 1, 2 and 3 respectively.
If the LSBase is C then embedded message bits are 100,101,110 and 111 if ambiguity numbers are 0, 1, 2 and 3 respectively.

6. Now the actual DNA sequence is extracted.

The actual DNA sequence is then decrypted to obtain the secret file. The ambiguity numbers obtained from the encryption phase is needed for decryption. The decryption process has following steps:
1. Convert the DNA sequence into binary form using table 1.
2. Group the binary data into groups of 8 bits and convert each group into alphabets corresponding to the ASCII values to obtain the encrypted amino acid sequence.
3. Decrypt the amino acid sequence using RSA cipher method to obtain the actual amino acid sequence.
4. Using ambiguity numbers and table 2 convert each amino acid in the sequence into corresponding DNA codons to obtain the DNA sequence.
5. Using table 1 convert the DNA sequence into binary form.
6. Group the binary form into groups of 8 bits and convert them into alphabets using corresponding ASCII value.
7. Store the output of step 6 to obtain the actual secret file.

VII. EXPERIMENTAL EVALUATION

The software tool used for study the evaluation for this work is JDK1.8. The text file is used for the input to the algorithm. The algorithms are implemented successfully and results are obtained.
The size of file after implementation is given in table 4.

TABLE 4: RESULT OF HIDING TEXT FILE

Size of text file (KB)	Size of file after Encryption (KB)	Size of file after DNA Steganography (KB)	Number of audio samples required to embed the file
1 KB	1.39 KB	11.1 KB	11469
2 KB	2.74 KB	22 KB	22494
3 KB	4.08 KB	32.7 KB	33519
4 KB	5.48 KB	43.8 KB	44967
5 KB	6.83 KB	54.6 KB	55989
10 KB	13.6 KB	108 KB	111534

The running time of algorithm is given in table 5.

TABLE 5: RUNNING TIME OF ALGORITHM

Size of text file (KB)	Running time for Encryption (ms)	Running time for DNA Steganography (ms)	Running time for Audio Steganography (ms)
1 KB	38	68	152
2 KB	67	140	260
3 KB	101	279	360
4 KB	130	479	485
5 KB	161	722	624
10 KB	319	2672	1133

VIII. APPLICATIONS

DNA due to its immense storage capacity and high randomness is used now in the field of steganography.DNA based algorithms can be used in various fields such as job scheduling for clusters, GPU applications, multi-core architectures, etc. Audio files are considered to be excellent carriers for the purpose of steganography due to presence of redundancy [13].

In the business world steganography can be used to hide a secret chemical formula or plans for a new invention. Steganography can also be used for corporate espionage by sending out trade secrets without anyone at the company being any the wiser. Terrorists can also use steganography to keep their communications secret and to coordinate attacks. There are a number of peaceful applications. The simplest and oldest are used in map making, where cartographers sometimes add a tiny fictional street to their maps, allowing them to prosecute copycats. A similar trick is to add fictional names to mailing lists as a check against unauthorized resellers. Most of the newer applications use steganography like a watermark, to protect a copyright on information. Photo collections, sold on CD, often have hidden messages in the photos which allow detection of unauthorized use. The same technique applied to DVDs is even more effective, since the industry builds DVD recorders to detect and disallow copying of protected DVDs [12].

IX. CONCLUSION

Communicating secretly without giving away any kind of crucial information is very important now a days in many fields. In this paper, we proposed a method to hide the secret messages stored in text files from unauthorized access. The method can be applied to text file. The proposed method has three levels. First level is DNA based encryption. Second level is DNA steganography which hides the encrypted message inside the reference DNA sequence. In the third level the embedded DNA sequence is hidden inside the audio file. Applications of proposed method are listed. In conclusion, in the proposed scheme the message has gone through encryption as well as two stages of steganography.

ACKNOWLEDGMENT

I am very grateful and would like to thank my guide for their advice and continued support to complete this paper and help to think beyond the obvious.

REFERENCES

[1] Amal Khalifa“ LSBase: A key encapsulation scheme to improve hybrid crypto-systems using DNA steganography ” IEEE 2013
[2] Ankur, Divyanjali and Vikas Pareek“A New Approach to Pseudorandom Number Generation” Fourth International Conference on Advanced Computing & Communication Technologies,IEEE 2014.
[3] Anupam Kumar Bairagi, Saikat Mondal and Amit Kumar Mondal“A Dynamic Approach In Substitution Based Audio Steganography”IEEE/OSA/IAPR International Conference on Infonnatics, Electronics & Vision , 2012
[4] Bama R, Deivanai S, Priyadharshini K“Secure Data Transmission Using DNA Sequencing” IOSR Journal of Computer Engineering (IOSR-JCE) Volume 16, Issue 2, Ver. II (Mar-Apr. 2014)
[5] K. Menaka “Message Encryption Using DNA Sequences ”IEEE 2014
[6] Muhammad Asad, Junaid Gilani and Adnan Khalid“An Enhanced Least Significant Bit Modification Technique for Audio Steganography”IEEE 2011
[7] Padma Bh, D. Chandravathi, P. Prapoorna Roja “Encoding and Decoding of a Message in the Implementation of Elliptic Curve Cryptography using Koblitz’s Method” International Journal on Computer Science and Engineering(IJCSE) Vol. 02, No. 05, 2010, 1904-1907
[8] Pratik Pathak, Arup Kr. Chattopadhyay and Amitava Nag“A New Audio Steganography Scheme based on Location Selection with Enhanced Security”IEEE.
[9] Rashmi M. Tank, Hemant D. Vasava, Vikram Agrawal “Literature Review on DNA based Audio Steganographic Techniques” International Journal of Engineering Trends and Technology(IJETT) 2014.
[10] Rashmi M. Tank, Hemant D. Vasava, Vikram Agrawal “DNA based Audio Steganography” International Journal of Computer Science and Technology (IJCST) 2015.
[11] Rohit Tanwar, Bhasker Sharma and Sona Malhotra“A Robust Substitution Technique to implement Audio Steganography ” International Conference on Reliability, Optimization and Information Technology, IEEE 2014
[12] Ronak Doshi, Pratik Jain, Lalit Gupta “Steganography and its Applications in Security”International Journal of Modern Engineering Research (IJMER) Vol.2, Issue.6, Nov-Dec. 2012 pp-4634-4638
[13] Shyamasree C M, Sheena Anees “Highly Secure DNA-based Audio Steganography” International Conference on Recent Trends in Information Technology (ICRTIT) IEEE 2013.
[14] Siddaramappa V“Data Security in DNA Sequence Using Random Function and Binary Arithmetic Operations”International Journal of Scientific and Research Publications, Volume 2, Issue 7, July 2012
[15] Swarnendu Mukherjee, Debashis Ganguly, Swarnendu Bhattacharya, Partha Mukherjee“A Cognitive Study on DNA Based Computation” International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009
[16] Tushar Mandge , Vijay Choudhary “A DNA Encryption Technique Based on Matrix Manipulation and Secure key Generation Scheme”IEEE.