DNA Based Cryptography and Steganography

Ginu Alexander
Department of Computer Science and Engineering
SAINTGITS College of Engineering, Kerala

Jeenu Clara Joseph
Department of Computer Science and Engineering
SAINTGITS College of Engineering, Kerala

Jerin George David
Department of Computer Science and Engineering
SAINTGITS College of Engineering, Kerala

Vishnu Prasad
Department of Computer Science and Engineering
SAINTGITS College of Engineering, Kerala

Abstract

The most common and widely used technique in the communication security and computer security fields is cryptography. Cryptography includes converting some data to incomprehensible format so that a non-intended recipient cannot determine its intended meaning. In this project, a binary form of data, such as plaintext message transformed into sequences of DNA nucleotides. The proposed method is a two-step procedure. First it converts the text messages to a DNA format, subsequently the nucleotides are passing through a Playfair Encryption process based on amino acids structure and in the second phase hide these message in a reference DNA sequence using 3:1 hiding rule.

Keywords- Cryptography, DNA, Nucleotides, Amino Acid, Reference DNA Sequence

I. INTRODUCTION

Information security is of increasing importance with the fast developing era as well as its confidentiality. Consequently, high level of security is required as it is a critical feature for thriving networks. So, the research concerning data hiding techniques has been increased continuously, due to the necessary need for powerful data protection in different applications. Applications such as annotation, ownership protection, copyrighting, authentication and military. Data hiding requires a carrier to hide the data in it such as image, video and audio. In the proposed method data hiding carrier is a DNA reference sequence.

For achieving maximum protection and powerful security with high capacity and low modification rate, Deoxyribonucleic acid (DNA) is explored as a new carrier for data hiding. Various biological properties of DNA sequences can be exploited for obtaining successful secured data embedding process [7][11]. This leads to a new born research field based on DNA computing. The most common and widely used techniques in the communication security and computer security fields are cryptography and steganography.

The character form of a message can be easily transformed to the form of bits. This binary form can be transformed to DNA form and then to amino acid form based on standard conversion tables. Playfair is based on the English alphabetical letters. DNA contains four bases that can be given an abbreviation of only four letters (adenine (A), cytosine (C), guanine (G) and thymine (T)). On the other side, we have 20 amino acids with additional 3 codons to represent the Stop of coding region. Each amino acid is abbreviated by a single English character. So we are able to stretch these 20 characters to 26 characters, we will be able to represent the English alphabet. Then, we have to convert the DNA form of data to amino acid form so that it can go through a classical Playfair cipher. Through this conversion process, we have to keep in mind the problem of ambiguity; that most amino acids are given more than possible codon.

II. DNA CONCEPTS

A. Deoxyribo Nucliec Acid – DNA and Genetic Code

DNA is a nucleic acid that contains the genetic instructions used in the development and functioning of all known living organisms. The main role of DNA molecules is the long-term storage of information. DNA is often compared to a set of blueprints or a recipe, or a code, since it contains the instructions needed to construct other components. The DNA double helix is stabilized by hydrogen bonds between the bases attached to the two strands. The four bases found in DNA are adenine (A), cytosine (C), guanine (G) and thymine (T). The genetic code consists of 64 triplets of nucleotides. These triplets are called codons. With three exceptions, each codon encodes for one of the 20 amino acids used in the synthesis of proteins. That produces some redundancy in the code that is most of the amino acids being encoded by more than one codon. The genetic code can be expressed as RNA codons. The DNA Codons is read the same as the RNA codons Except that the nucleotide thymidine (T) is found in place of uridine (U).

III. DNA BASED ENCRYPTION AND DECRYPTION

A. Encryption Algorithm

The encryption involves two phases. First phase consists of encoding to the DNA form and converting into amino acids using the standard conversion table. Then, the DNA and amino acids based Playfair cipher encrypts the encrypted message. Then, the ciphered message is hidden in a selected DNA sequence from NCBI database [5] using the LSBase method. So, the system is providing double layer security.

1) Phase 1: Data Encryption

The encryption step is preferred before hiding the original format of the secret message in the DNA for achieving double layer of security .The proposed technique uses DNA and amino acids Playfair cipher to encrypt the secret message. Conventional Playfair cipher is a symmetric encryption technique that encrypts a text message using a 5*5 table. It is constructed using a secret key word and the remaining letters of the alphabet that are not included in the key word. Playfair cipher encrypts pairs of letters instead of single letters.

At first the input message M is mapped into its corresponding ASCII and then to binary Mbin, using 4-bits coding rule. Mbin is then mapped to DNA nucleotides using a binary coding rule (BCR) to be encrypted using DNA and amino acids Playfair cipher. The naive binary coding rule maps each 4 bits to two DNA nucleotide as shown in Table 1.

Table 1: 4-Bit Binary Coding Rule

DNA Nucleotides	*Binary Representation*
AA	*0000*
AC	*0001*
AG	*0010*
AT	*0011*
CC	*0100*
CA	*0101*
CG	*0110*
CT	*0111*
GG	*1000*
GA	*1001*
GC	*1010*
GT	*1011*
TT	*1100*
TA	*1101*
TC	*1110*
TG	*1111*

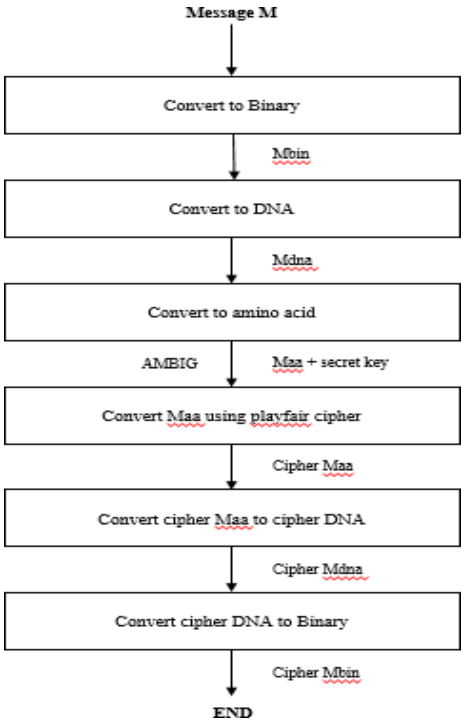

Fig. 1: Data Encryption flowchart

The output DNA of the secret message is converted to amino acids according to the new distribution of the alphabet with their corresponding new codons in [13]. The new distribution is derived from the standard universal table of amino acids and their DNA codons representation. Since each amino acid is associated with multiple codons as in [13] and the message is converted from DNA to amino acids. There should be something that refers to the index of each DNA codon corresponding to each amino acid to be able to retrieve the correct codon by the receiver in the decryption phase when amino acids convert to DNA. These indices are called AMBIG which refer to ambiguity. So, for successful retrieval, keep in track the ambiguity of each amino acid in AMBIG.

Playfair cipher is applied using the secret key to encrypt the amino acids form of the secret message formed from the last step into cipher amino acids form. The formed ciphered amino acids is converted back to DNA by selecting the first codon corresponding to each amino acid to form the cipher DNA format of the message. The overall encryption process is illustrated in Fig. 1.

2) Phase 2: Data Hiding

Least significant base is data hiding methodology proposed in [14]. In a DNA sequence each three adjacent nucleotides constitute a unit called codon. LSBase method depends on hiding the secret message bits in the least signification bit of each codon of the reference sequence. Any sequence is a combination of some purine bases (A & G) and pyrimidine bases (C & T). In order to hide the cipher message bits, the following steps are applied. The formed DNA (cipher Mdna) from the encryption phase is converted into binary again to form cipher binary message (cipher Mbin) by using 4-bits representation. binary coding rule as following: The AMBIG as well is converted to binary AMBIGbin and since the maximum number of codons corresponding to an amino acid is 4, indexed from 0 to 3 so it can be represented in maximum 2 bits as shown in table. 2. Select a DNA reference sequence from one of the public databases such as EBI or NCBI and convert it to RNA by substituting each T with U Hide cipher Mbin and AMBIGbin using LSBase method. Since, hiding methodology depends on the message bits and the LSBase of each codon in the DNA The LSB of S is checked and if it is a purine base (A & G), it is substituted by (G) to encode 1 of the secret message or (A) to encode 0.

Table 2: Ambiguity Conversion

Ambig	*Binary Representation*
0	*00*
1	*01*
2	*10*
3	*11*

If the LSB of S is a pyrimidine base (C & T), it is substituted by (C) to encode 1 or (U) to encode 0. LSBase algorithm neglects the following codons: UGA, UGG, AUA and AUG during the hiding process since according to the standard distribution of DNA codons to amino acids, Trp and Met amino acids have a single codon which are AUG and UGG respectively. Also, stop has only one codon which is UGA which will be neglected too. Finally it is coded by three codons: AUU, AUC and AUA .So, AUA is neglected and AUU and AUC will be used in data hiding.

Output from phase I, not only the cipher binary message but also the ambiguity results from converting DNA format of the message to amino acids. The objective of the proposed method is to hide the secret message and the ambiguity required by the receiver to retrieve the secret message from the DNA without additional information. This because, the ratio of the length of the binary cipher message to the length of the binary ambiguity is 3:1 as will be discussed in subsection C. So, hiding the message with the ambiguity in the DNA sequence using 3:1 ratio avoids adding additional data to mark the starting position of the message in the DNA reference sequence and the starting position of the ambiguity as well. Consequently, the data required to be sent is minimized and it is the faked sequence.

B. Decryption Algorithm

The sender sends the faked DNA sequence to the receiver. Then, the receiver applies the Playfair cipher to decrypt the message using the secret key. Both sender and receiver will share the secret key from the beginning. But sharing a secret key may possess a problem since it needs to be interchanged before applying the encryption process. To avoid this problem, the proposed method can be modified to hide the secret key within the faked sequence. At the receiver side, when the faked DNA sequence is sent to the receiver without any additional data which enhances the algorithm"s security, the receiver should apply two phases: data extraction and message decryption in order to retrieve the secret data which is contained in the faked DNA sequence.

1) Phase 1: Data extraction

The extraction process is simply the inverse of the embedding algorithm where LSBase method is used and sequence is divided into codons. Check the least significant base of each codon to retrieve the hidden bits of the secret message. If the LSBase is either „T" or „A" then the embedding bit was "0". If it is „C" or „G" then the embedding bit is „1". Each three extracted consecutive bits by LSBase method are added to the secret message and the next bit is added to the ambiguity of the secret message till Mbin and AMBIGbin are completely extracted from sequence.

2) Phase 2: Decryption

Decryption is the inverse of the encryption phase where Mbin is converted to DNA using proposed 4-bits binary coding rule. Then, the ciphered DNA format is converted to amino acids to apply the Playfair cipher on it using the secret key. Decrypted amino acids form is generated from Playfair cipher then AMBIGbin is converted to decimal digits by mapping each two bit to number. Use each ambiguity number with each amino acid character to retrieve the corresponding codon to this char associated with this ambiguity number. Finally, a sequence of DNA is retrieved, by converting it into binary using 4-bits binary coding rule then to ASCII to get the corresponding plain text which is the original form of the secret data as shown in fig 2.

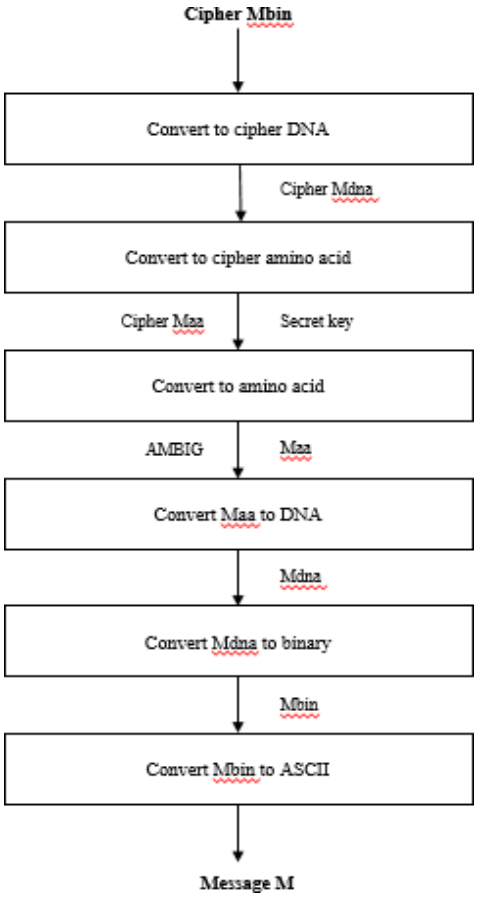

Fig. 2: Data Decryption flowchart

IV. CONCLUSIONS

This paper proposes a data hiding method by combining the means of cryptography and steganography as well. This achieves double layer security of the system. A new binary coding rule is proposed that assigns 4 bits to each combination of 2 nucleotides instead of assigning two bits to only one nucleotide which strengthen the algorithm"s security. Due to using LSB method in hiding the cipher bits of the message and the ambiguity, the proposed algorithm is still blind as the embedded data can be extracted without the need to the original DNA reference sequence. The sequence it preserves the original DNA sequence length as it depends on substitution only so the algorithm"s payload is zero which avoids attracting the attention to the faked sequence.

REFERENCES

[1] G. Hamed et al., "DNA Based Steganography: Survey and Analysis for Parameters Optimization,"in Applications of Intelligent Optimization in Biology and Medicine, Springer, 2015, ISSN: 1868–4394, pp. 47–89.
[2] Ghada Hamed, Mohammed Marey, Safaa Amin El-Sayed Mohamed Fahmy Tolba, "Hybrid Technique for Steganography-based on DNA with N-Bits Binary Coding Rule" 2015 Seventh International Conference of Soft Computing and Pattern Recognition (SoCPaR 2015)
[3] M. Skariya and M. Varghese, "Enhanced Double Layer Security usingRSA over DNA based Data Encryption System," International Journal of Computer Science & Engineering Technology (IJCSET), ISSN: 2229–3345, vol. 4, Issue No. 06, pp. 746–750, Jun 2013.
[4] Y. A. Yunus, S. Ab Rahman and J. Ibrahim,"Steganography: A Reviewof Information Security Research and Development in Muslim World," American Journal of Engineering Research (AJER), ISSN: 2320–0936,vol. 02, Issue No. 11, pp. 122–128, 2013.
[5] C. Guo, C. Change and Z. Wang, "A New Data Hiding Scheme based on DNA Sequence," International Journal of Innovative Computing, Information and Control, ISSN: 1349–4198, vol. 8, Issue No. 1, pp. 139–149, Jan 2014.

[6] I. K. Maitra, "Digital Steganalysis: Review on Recent Approaches," Journal of Global Research in Computer Science, ISSN: 2229–371X,vol. 2, Issue No. 1, pp. 1–5, Jan 2011.
[7] J. Taur, H. Lin, H. Lee and C. Tao, "Data Hiding in DNA Sequences based on Table Lookup Substitution," International Journal of Innovative Computing, Information and Control, ISSN: 1349–4198, vol. 8, IssueNo. 10, pp. 6585–6598, Oct. 2012.
[8] A. Atito, A. Khalifa and S. Z. Rida, "DNA-based Data Encryption and Hiding using Playfair and Insertion Techniques," Journal of Communications and Computer Engineering, ISSN: 2090–6234, vol. 2, Issue No. 3, pp. 44–49, 2012.
[9] A. K. Kaundal and A. K. Verma, "DNA based Cryptography: A Review,"International Journal of Information and Computation Technology, ISSN: 0974–2239, vol. 04, Issue No. 7, pp. 693–698, 2014.
[10] H.J. Shiu, K.L. Ng, J.F. Fang, R.C.T. lee and C.H. Huang, "Data Hiding Methods based upon DNA Sequences," Journal of Information Sciences: an International Journal, vol. 180, Issue No. 11, pp. 2196–208, June 2010.
[11] A. Khalifa and A. Atito, "High-Capacity DNA-based Steganography," in Informatics and Systems (INFOS), 8th International Conference on 2012, May 2012, pp. BIO–76.
[12] M. Sabry, M. Hashem, T.Nazmy and M. E. Khalifa, "A DNA and Amino Acids-based Implementation of Playfair Cipher," International Journal of Computer Science and Information Security, ISSN: 1947–5500, vol. 8, Issue No. 3, pp. 129–136, 2010.
[13] A. Khalifa, "LSBase: A key Encapsulation Scheme to Improve Hybrid Crypto-systems using DNA Steganography," in 8th International Conference on Computer Engineering & Systems (ICCES), Cairo, Egypt,Nov. 2013, pp. 105–110.
[14] NCBI Database, Bank for real DNA reference sequences, http://www.ncbi.nlm.nih.gov/

