New Data hiding method based on DNA and Vigenere Autokey

Wafaa M. Abduallah[1], Subhi R. Mohmmed Zeebaree[2]

[1]Department of Computer Science, College of Computer & Information Technology, Nawroz University, Duhok, Kurdistan Region – Iraq
[2]Department of Communication and Computer Engineering, College of Engineering, Nawroz University, Duhok - Iraq

ABSTRACT

People are always looking for secure methods to protect valuable information against unauthorized access or use. That's why; disciplines like cryptography and steganography are gaining a great interest among researchers. Recent steganographic techniques hide data into digital media such as sound, images, and videos. However, steganography took a step further to utilize the DNA as a carrier of secret information. DNA-based steganography techniques could be for either authentication or data storage. There is a special property for real DNA sequences, that is, there is almost no difference between a real DNA sequence and a fake one. This property has been be exploited in this study. A distinct type of hiding approach is adopted in this work to be applied on the DNA sequences based on cryptographic method called Vigenere Autokey. The proposed mechanism hides the secret message via converting it along with the key to the DNA sequences and then applying the Autokey cipher using a special table created via making use of DNA-XOR operator to increase the security of the proposed mechanism. So, it can meet the requirements of good steganographic system (with high embedding capacity, good visual imperceptibility, and reasonable level of security).

KEYWORDS: Steganography, Cryptography, DNA, Vigenere Autokey, XOR.

1. INTRODUCTION

Nowadays, Data hiding is one of the techniques that have been developed in parallel to improvements in technology. Data hiding is a technique of hiding data that is known only by the sender and receiver. So, the hidden data is gained only by a person who has the stego key. Data can be hidden in apparently innocent multimedia files such as (image, audio signal, video, etc.) with the condition of not exceeding size of the cover media during transmission. At the receiver side, to obtain the hidden data, the receiver has to know the stego key and apply the operation of embedding procedure in reverse order (Leier, Richter, Banzhaf, & Rauhe, 2000), (Taur, Lin, Lee, & Tao, 2012).

Academic Journal of Nawroz University
(AJNU) Volume 6, No 3(2017), 5 pages
Received 1 May 2017; Accepted 20 August 2017
Regular research paper: Published 27 August 2017
Corresponding author's e-mail: heevy9@yahoo.com

for intellectual property applications. In recent years, most of the research that are related to secret data transmission are focused on data hiding techniques to avoid malicious attacks and accomplish a safe transmission. Where, the most common medium used for data hiding is the image and it is difficult to hide the secret message in the image without making noticeable effect to the original image (Shiu et al. 2010).

Image is the most common medium in data hiding without affecting the original image overmuch (Taur et al. 2012). Nonetheless, the original image structure is usually affect hiding data in the image (Shiu et al. 2010). Recently, hiding data in DNA sequences has been attracted by many researchers, and several methods have been proposed (Shiu et al. 2010), (Shimanovsky, Feng, & Potkonjak, 2002), (Abbasy & Shanmugam, 2011), (Guo ,Chang, & Wang, 2012), (Bhattacharyya, & Bandyopadhyay, 2013), (Mousa, Moustafa, Abdel-Wahed, & Hadhoud, (2011) in this field. Briefly speaking, DNA is two twisted strands composed of four bases, Adenine (A), Cytosine (C), Thymine (T) and Guanine (G). The four bases represent the genetic code. Where, (A) bonds with the complementary (T), (G) bonds with the Complementary (C), and vice versa. Thus one strand and the corresponding complementary strand forms DNA. For example, one strand is AACGTC, and the

other must be TTGCAG as shown in Fig. 1. The arrangement of amino acids which form a protein is determined by DNA sequence (Shimanovsky, Feng, & Potkonjak, 2002). In most DNA-based methods, the biological property of a DNA sequence has been manipulated. Nevertheless, using only biological properties may bound the performance of data hiding methods. In addition to the biological view, Chang et al. and Shiu et al. have established methods based on the fact that there are almost no differences between a real DNA sequence and a faked DNA sequence. As an emergent medium, DNA sequences have many advantages compared with images (Kencl, & Loebl, 2010). First, there is no need to worry about the distortion since, DNA sequences are composed of letters which are meaningless for most people. Second, the embedding capacity and computational efficiency is potentially for DNA sequence much better than that of the image. For instance, for a gray-level image, each pixel requires 8 bits memory while, for a DNA sequence, each nucleotide needs only 2 bits. Therefore, considering the memory requirement for hiding a 1-bit message, one bpn (bits per nucleotide) is equivalent to about four bpp (bits per pixel). In contrast, most image-based methods have bpp values less than three in order to have a reasonable PSNR and the values depend on the characteristics of the images. Consequently, regarding memory requirement, the DNA medium is more cost-effective than the image (Tuncer, & Avci, 2016).

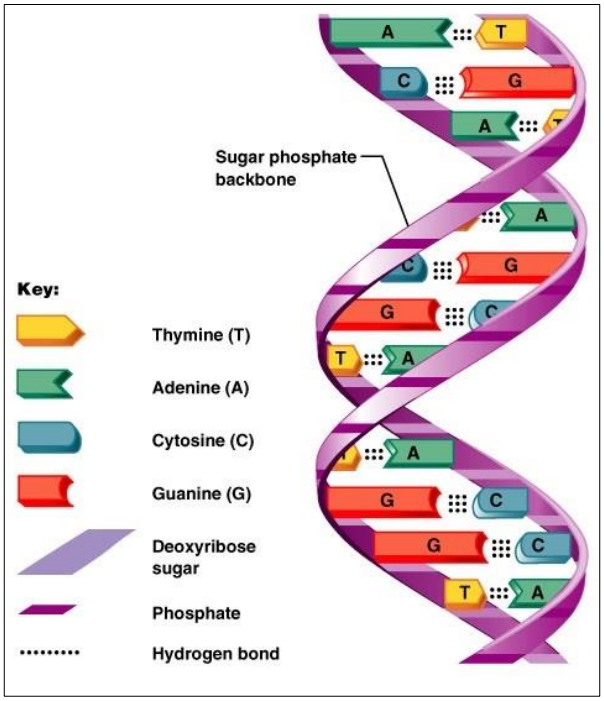

Fig. 1. The structure of part of a DNA.

In this work, a new type of hiding approach is applied on the DNA sequences based on cryptographic method called Vigenere Autokey, which can meet the requirements of good steganographic system (with high capacity, good visual imperceptibility, and reasonable level of security).

Following the introduction, in Section 2 the related works and motivation of this work are discussed. The DNA digital coding technology is presented in Section 3. Then, in Section 4, the proposed algorithm is presented. Section 5 presents the security analysis for the proposed scheme. Finally, Section 6 concludes the paper.

2. RELATED WORKS AND MOTIVATION

Recently, data hiding based on the DNA sequence has been appealing much attention, and many methods have been proposed (Taur et al. 2012), (Shiu et al. 2010), (Shimanovsky, Feng, & Potkonjak, 2002), (Abbasy & Shanmugam, 2011), (Guo ,Chang, & Wang, 2012), (Bhattacharyya, & Bandyopadhyay, 2013). More precisely, the original idea of hiding data in DNA and RNA has been proposed by (Shimanovsky, Feng, & Potkonjak, 2002). Where, two original techniques for hiding the data, along with the analysis and evaluation for different functions of hidden data have been presented. The first technique embeds data in non-coding DNA such as non-transcribed regions as well as non-genetic DNA for instance, DNA computing solutions. While, the second one, in theory, can actually be used to hide information directly into active genetic segments. Both techniques can be applied in order to protect the intellectual property in the realms of gene therapy, transgenic crops, tissue cloning, and DNA computing.

To increase the capacity, the authors (Guo ,Chang, & Wang, 2012) established an injective mapping between two secret bits and one complementary rule. Based on this mapping mechanism, the proposed scheme can hide two secret bits by replacing one character. Furthermore, the security analysis showed that it is impossible for an intruder to recover the secret message for all practical purposes. Finally, the experiments indicated the stable and efficient embedding capacity for the proposed scheme with a low modification rate and without expanding the length of the reference DNA sequence.

On the other hand, the authors (Taur et al. 2012), proposed an improved method based on the original substitution method that is presented by (Shiu et al. 2010) called table lookup substitution method (TLSM). The basic framework of the original method is adopted for the TLSM but, the complementary rule is replaced by the proposed rule table. Similar to the original method and according to the rule table with the secret bits to be hidden, the TLSM substitutes a specific letter for its corresponding letter in the reference DNA sequence. The proposed rule table is able to hide two secret bits for each letter conversion, whereas the original complementary rule can hide one bit only. Accordingly, the TLSM is basically improved to hide more information in each letter. Moreover, the TLSM has been extended to a general approach which can hide data in any sequences of letters or symbols. In addition, two approaches are proposed by the authors to further enhance the performance of the generalized TLSM; they are the Base-t TLSM and the Extended TLSM (ETLSM). The TLSM can utilize the substitution table more efficiently with the secret message expressed in base-t representation, since it can fully utilize all conversion entries with a proper radix.

While, the ETLSM can largely increase the safety level of the basic TLSM via taking additional letters into account. Finally, it is obvious that all the proposed methods by (Taur et al. 2012) provided improvement over the original substitution method in cracking probability and bits-per-letter.

Also, three data-hiding methods based on the DNA sequence are proposed by (Shiu et al): the insertion method, the complementary pair method and the substitution method. In these methods, a reference DNA sequence S is selected and incorporated with the secret message M to obtain S'. In the insertion method, bits from secret message M are inserted one at a time into the reference DNA sequence, but this scheme certainly leads to increase the redundancy and expand the length of the DNA sequence where, the notable expansion can easily attract the attention of intruders. While, in the substitution method, the fake reference DNA sequence S' preserves the original sequence length by replacing the designated characters. However, the reference DNA sequence S has to be sent to the receiver in order to be used to identify and extract the message hidden in S'. Additionally, replacement for each character is related to one secret message bit embedded in the reference DNA sequence, so if the reference DNA sequence needs to be embedded with a long secret message, the sequence will be exposed to high modification rate. As to the complementary pair method, it expands the reference DNA sequence considerably in the process of embedding the secret message. So, it can definitely attract the attention of intruders.

3. DNA DIGITAL CODING TECHNOLOGY

The most fundamental coding method in the information science is binary digital coding, which is anything can be encoded by two state 0 or 1 and a combination of 0 and 1. In DNA sequence, there are four kinds of bases, which are adenine (A) and thymine (T) or cytosine (C) and guanine (G). The simplest coding method to encode the 4 nucleotide bases (A, T, C, G) is by using of 4 digits: 0(00), 1(01), 2(10), 3(11). Where, using this encoding format, there will be 4!=24 possible coding patterns. Obviously, in a double helix DNA string and in terms of sequence, two DNA strands are held together complementary; that is A to T and C to G according to Watson-Crick complementarity rule. Taking in consideration the DNA digital coding, it should reflect the biological characteristics of 4 nucleotide bases so, the complementary rule that (~0)=1, and (~1=0) is proposed in this DNA digital coding (Cui, Qin, Wang, & Zhang, 2007), (Cui, Qin, Wang, & Zhang, 2008). According to this complementary rule, that is encode 0(00) to 3(11) and 1(01) to 2(10). Thus, among these 24 coding patterns, only 8 kinds of patterns (0123/CTAG, 0123/CATG, 0123/GTAC, 0123/GATC, 0123/TCGA, 0123/TGCA, 0123/ACGT, 0123/AGCT) which are topologically identical fit the complementary rule of the nucleotide bases. The authors (Cui et al. 2008) suggested that the coding pattern in accordance with the sequence of molecular weight, 0123/CTAG as a best coding pattern for the nucleotide bases. This pattern could perfectly reflect the biological characteristics of 4 nucleotide bases and have a certain biological significance.

There are many advantages for using binary digital coding of DNA sequences rather than character coding of DNA: (1) compared to the traditional character DNA coding, the binary coding of DNA decreases the redundancy of the information coding and improves the coding efficiency. (2) Using digital coding of DNA sequence makes the mathematical operation and logical operation very convenient and this may give a great impact on the DNA bio-computer. (3) After preprocessing the DNA sequence by DNA digital coding techniques will be able to do digital computing and adapt to the existing computer-processing mode, which consequently facilitates the direct conversion between biological information and encryption information to be used by the cryptography scheme. (4) Finally, by using the DNA digital coding technology, the traditional encryption method such as DES or RSA could be used to preprocess to the plaintext in the crsyptography scheme.

4. PROPOSED MECHANISM

From real DNA sequences, it is easy to discover one special property of a DNA sequence. That is, there is almost no difference between a real DNA sequence and a fake one []. This property has been exploited in this study in addition to complementary rule that has been discussed by (Shiu et al. 2010), to establish an injective mapping. That is, each character x is assigned a complement, denoted as C(x). For example, we can apply the following complementary rule:

(AT)(CA)(GC)(TG), where C(A) = T, C(C) = A, C(G) = C, C(T) = G.

There are six legal complementary rules as follows: (AT)(TC)(CG)(GA), (AT)(TG)(GC)(CA), (AC)(CT)(TG)(GA), (AC)(CG)(GT)(TA), (AG)(GT)(TC)(CA) and (AG)(GC)(CT)(TA).

The proposed scheme can be demonstrated in the following steps:

Step 1. Firstly, a table of DNA sequences is constructed via applying XOR operation among its' elements as further explained in Table 1, where the following rule can be used as a binary coding: ((A:00)(C:01)(G:10)(T:11)).

Table 1. Applying XOR operator for DNA sequences.

XOR	00	01	10	11
00	00	01	10	11
01	01	00	11	10
10	10	11	00	01
11	11	10	01	00

XOR	A	C	G	T
A	A	C	G	T
C	C	A	T	G
G	G	T	A	C
T	T	G	C	A

- **Step 2.** Next, the secret message has to be converted to binary format via using the equivalent ASCII code for each character then, the resulted integer value has also converted to binary format composed of (8 digits). Furthermore, the resulted binary sequence has to be converted to DNA sequence using the inverse function of the binary coding rule to produce a faked DNA sequence as explained in the following example:

Let the secret message be M="hi"

Converting it to ASCII code M=104, 105;

M_Binary=01101000 01101001

M_DNA Sequence= CGGACGGC

- **Step 3.** In this step, a secret key has to be chosen which can be any number or character or even a word then, the same procedure of step 2 has to be applied again on the key as well to convert it to DNA Sequence. For instance, if the selected key is (11) then converting it to binary will be (1011) and finally turning it to Small DNA sequence will produce (GT).
- **Step 4.** Then, applying Vigenere Autokey cipher on the M_DNA Sequence to produce a fake DNA Sequence that hides the secret message as further explained in the following:

M_DNA Sequence = CGGACGGC

| | |

Key = GTCGGACG

↓↓↓

Coded_DNA Sequence = TCTGTGTT = CGCACACC

The coded DNA sequence can be obtained via taking the first letter of M_DNA Sequence which is (C) to represent the column in table 1 with the equivalent first letter of the key which is (G) to represent the row in the same table then taking the intersection value between them as a result which is (T) in this case and so on for all other values.

Step 5. Finally, applying one of the six complementary rule on the Coded_DNA Sequence to add another layer of security as demonstrated in the following:

Coded_DNA Sequence = TCTGTGTT

After applying this complementary rule (AT)(TC)(CG)(GA): the final result will be:

Final Coded_DNA Sequence = CGCACACC

The Final Coded_DNA Sequence can be send to other party. On the other side, to recover the secret message, the receiver has to apply the same steps that have applied by the sender but in reverse order. Starting with applying the inverse of complementary rule then applying Vigenere Autokey cipher to get back the original sequence.

5. SECURITY ANALYSIS

The proposed algorithm has various steps to break and to get the original message. As per this study and the discussions of the proposed algorithm, it is very difficult to break and guess actually embedded data from the DNA sequence. Hence, any intruder who receives the intermediate message will never be able to retrieve the original message as intended by the sender. Taking in consideration a useful fact that there is a large number of DNA sequences publicly available on various web-sites such as the EBI database.according to European Bioinformatics Institute (EBI), (Mousa, Moustafa, Abdel-Wahed, & Hadhoud, (2011).

On the other hand, a good cryptographic security system has to be able to protect confidential data. The level of security that the proposed algorithm offers is its strength. Cryptographic attacks are a part of cryptanalysis and are designed to ruin the security of cryptographic algorithms, and they are used to attempt to decrypt data without knowing the key. As a good encryption technique should be robust against such attacks. One of these attacks is the Brute Force Attack that can be used by the intruder via checking all possible keys until finding the correct key. For the proposed algorithm the cryptographic used method is, Autokey cipher which overcomes the problem of periodically repeating a keyword and instead it simply add the plaintext to the keyword in the case of being the keyword is shorter than the plaintext. Moreover, the intruder doesn't have information about the utilized mechanism for DNA coding Technology as well as the used Complementary rules. Where all these information are exchanged secretly between two the sender and receiver.

6. CONCLUSION

In this paper, a secure, reversible and applicable data hiding algorithm is proposed based on DNA technology. Where, the proposed method made use of the special properties of DNA Sequences for data hiding which is, there is almost no difference between a real DNA sequence and a fake one therefore hiding data in DNA sequences and producing new sequence of DNA will not be discovered by the intruder since, there are approximately 163 million publicly available DNA sequences and it is virtually impossible to guess this sequence. The proposed hiding algorithm provides a new way of embedding mechanism which used Autokey cipher for hiding the secret message within a generated DNA sequence. The proposed algorithm utilized DNA-XOR operator for creating a table to be used by the mechanism of Autokey

cipher. To the best of our knowledge, this mechanism has not been adopted by other works.

REFERENCES

Taur, J. S., Lin, H. Y., Lee, H. L., & Tao, C. W. (2012). Data hiding in DNA sequences based on table lookup substitution. International Journal of Innovative Computing, Information and Control, 8(10), 6585-6598.

Shiu, H. J., Ng, K. L., Fang, J. F., Lee, R. C., & Huang, C. H. (2010). Data hiding methods based upon DNA sequences. *Information Sciences*, *180*(11), 2196-2208.

Shimanovsky, B., Feng, J., & Potkonjak, M. (2002, October). Hiding data in DNA. In *International Workshop on Information Hiding* (pp. 373-386). Springer, Berlin, Heidelberg.

Abbasy, M. R., & Shanmugam, B. (2011, July). Enabling data hiding for resource sharing in cloud computing environments based on DNA sequences. In *Services (SERVICES), 2011 IEEE World Congress on* (pp. 385-390). IEEE.

Guo, C., Chang, C. C., & Wang, Z. H. (2012). A new data hiding scheme based on DNA sequence. *Int. J. Innov. Comput. Inf. Control*, *8*(1), 139-149.

Bhattacharyya, D., & Bandyopadhyay, S. K. (2013). Hiding secret data in dna sequence. *International Journal of Scientific &Engineering Research*, *4*(2).

Mousa, H., Moustafa, K., Abdel-Wahed, W., & Hadhoud, M. M. (2011). Data hiding based on contrast mapping using DNA medium. *Int. Arab J. Inf. Technol.*, *8*(2), 147-154.

Tuncer, T., & Avci, E. (2016). A reversible data hiding algorithm based on probabilistic DNA-XOR secret sharing scheme for color images. *Displays*, *41*, 1-8.

Cui, G., Qin, L., Wang, Y., & Zhang, X. (2008, September). An encryption scheme using DNA technology. In Bio-Inspired Computing: Theories and Applications, 2008. BICTA 2008. 3rd International Conference on (pp. 37-42). IEEE.

Leier, A., Richter, C., Banzhaf, W., & Rauhe, H. (2000). Cryptography with DNA binary strands. *Biosystems*, *57*(1), 13-22.

Kencl, L., & Loebl, M. (2010). DNA-inspired information concealing: A survey. *Computer Science Review*, *4*(4), 251-262.

Cui, G., Qin, L., Wang, Y., & Zhang, X. (2007, April). Information security technology based on DNA computing. In Anti-counterfeiting, Security, Identification, 2007 IEEE International Workshop on (pp. 288-291). IEEE.

EMBL-EBI, The home for big data in biology. (n.d.). Retrieved from http://www.ebi.ac.uk/ [Last accessed on July 14, 2017]