Utilizing DNA Strands for Secured Data-Hiding with High Capacity

https://doi.org/10.3991/ijim.v11i2.6565

Samiha Marwan
The British University in Egypt, Cairo, Egypt
samiha.abdelrahman@bue.edu.eg

Ahmed Shawish
The British University in Egypt, Cairo, Egypt
ahmed.gawish@bue.edu.eg

Khaled Nagaty
The British University in Egypt, Cairo, Egypt
khaled.nagaty@bue.edu.eg

Abstract—There are continuous threats to network technologies due to its rapidly-changing nature, which raises the demand for data-safe transmission. As a result, the need to come up with new techniques for securing data and accommodating the growing quantities of information is crucial. From nature to science, the idea that genes themselves are made of information stimulated the research in molecular deoxyribonucleic acid (DNA). DNA is capable of storing huge amounts of data, which leads to its promising effect in steganography. DNA steganography is the art of using DNA as an information carrier which achieves high data storage capacity as well as high security level. Currently, DNA steganography techniques utilize the properties of only one DNA strand, since the other strand is completely dependent on the first one. This paper presents a DNA-based steganography technique that hides data into both DNA strands with respect to the dependency between the two strands. In the proposed technique, a key of the same length of the reference DNA sequence is generated after using the second DNA strand. The sender sends both the encrypted DNA message and its reference DNA sequence together into a microdot. If the recipient receives this microdot uncontaminated, the sender can safely send the generated key afterwards. The proposed technique doubles the amount of data stored and guarantees a secure transmission process as well, for even if the attacker suspects the first-sent DNA sequence, they will never receive the key, and hence full data extraction is nearly impossible. The conducted experimental study confirms the effectiveness of the proposed.

Keywords—DNA Steganography, Hiding Capacity, Data Hiding, Data Storage.

1 Introduction

Nowadays, network technologies are rapidly improving in an extensive manner and internet applications are widely used in almost all fields. The data widespread and its rapid improvement transform it into a priceless resource resulting in continuous threats on the data itself and the way of its transmission. These continuous threats necessitate the employment and development of well-suited information security techniques to accommodate securing huge data amounts.

Recently, steganography becomes one of the promising techniques for data security. Steganography is the art of hiding data into a medium in a way that makes data unsuspicious. The hiding media used can be in the form of images, audios, videos and DNA. From nature to science, the idea that genes themselves are made of information stimulated the research in molecular deoxyribonucleic acid (DNA). DNA molecule has three main advantages that make it an efficient medium for data hiding and transmission. First of all, its high storage capacity; as proved in [1]. Secondly, the simplicity of converting data to DNA sequence. Third, the DNA sequence complexity and randomness provides a great uncertainty which makes DNA as a cover media is better than any other media for data hiding. By exploiting the advantages of a DNA as an efficient data carrier, researches ended up by many DNA steganography techniques for secure data communication and transmission [2, 3, 4, 5, 6].

In order to convert data into a DNA format, the most famous and simple method is done by converting each two binary bits into a DNA unit. Since DNA consists of four building units (A,G,C and T), and there exists four 2-bit binary combinations, each two binary bits can map to one DNA unit. Table 1 shows an example of DNA-binary representation. For example, a message: 0110010 is encoded using Table 1 to CGAG. DNA molecule has a double-stranded nature, where the second DNA strand is complementary to the first one. Due to the dependent relation between the two strands, currently, DNA steganography techniques utilize the properties of only one DNA strand for data hiding. The continuous need for securing huge amount of data as well as the limited. utilization of the doubly nature of DNA strands encourages us to investigate more about the possibility of hiding in the second DNA strand with respect to its dependency to the first strand and developing a hiding and extraction algorithms that enable data transmission securely without any data loss.

Table 1. DNA Letter Representation Of Binary Bits

DNA letter	Binary Representation
A	00
C	01
G	10
T	11

This paper presents a novel steganography technique, where data is hidden into both DNA strands. The proposed algorithm is processed as follows; firstly, any data type whether text, audio or images are transformed into their binary form. This result-

ed binary sequence is then encoded into a DNA format using the Binary-DNA representation method in Table 1. Afterwards, the encoded DNA message is then hidden into the non-coding regions of the two strands of a suitably chosen reference DNA sequence -real DNA sequence-. According to the previous step, a key sequence will be generated of length equal to that of the reference DNA sequence. The sender will send the resultant fake DNA sequence into a microdot and then the key sequence through a public channel. This key sequence will not be sent unless the previously sent microdot was uncontaminated. At the receiver's side, the reverse of the hiding process will take place. In the proposed algorithm, the amount of data stored is double that stored in the current DNA steganography techniques. Additionally, a secure data transmission is highly guaranteed, due to the avoidance of sending both the key and the fake DNA sequence together. For instance, if the attacker suspects the fake DNA sequence in the microdot, the microdot would be contaminated and therefore the sender will never send the key, hence full data extraction is nearly impossible. The performance of our proposed technique as well as its cracking probability have been carefully calculated and the conducted experimental study confirms our technique's effectiveness.

The rest of this paper is organized as follows. Section 2 overviews a biological background on DNA and related work on the current Steganography techniques. Section 3 presents the proposed technique in detail. Section 4 discusses its performance analysis. Finally, the paper is concluded in Section 5.

2 Background

In this section, we provide a brief biological background on the DNA and related work on DNA steganography.

2.1 DNA Overview

Deoxyribonucleic acid (DNA) is the hereditary material in all living organisms. It is considered as the building blocks of life since it contains the genetic instructions used in the development and functioning of all known living organisms [7]. The main role of DNA molecules is the long- term storage of the body information. DNA stores the body's information as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). DNA consists of two strands formed by pairing up each two bases together, namely 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, which is the building block of the DNA. Each three bases are called a codon, e.g. ACG, TAC, CGA, .. etc. There are 64 codons, since there exists only 4 DNA bases [8]. When a cell uses the information in a DNA, the DNA sequence is duplicated and then copied to a complementary RNA sequence (another important nucleic acid). Usually, this RNA copy is then used to make a matching protein sequence. This process is referred to as the central dogma.

A DNA sequence consists of two regions namely coding regions and noncoding regions. The coding regions are those DNA bases which are translated into proteins and therefore they are responsible for the DNA functionality, whereas the noncoding regions are those DNA bases which are not translated to any proteins, sometimes they are referred to as junk DNA bases. In this paper, we are focusing with the noncoding regions to avoid any mutation occurrence and consequently no loss in data. From a practical point of view in order to determine the nucleotides in a DNA strand, a process of DNA amplification must take place. Amplifying a DNA strand takes place by applying the PCR process, where this process uses primer sequences to select exactly the DNA strand to be amplified. The primer sequences are short DNA sequences from 10 to 30 bases maximum used for correct selection of DNA strands that needed to be sequenced or amplified.

2.2 Related Work

DNA steganography has started in 1999 by C.Catherine where data is encrypted in DNA and hid into microdots [4]. A. Leier in 2000 proposed a hiding technique where data is encoded into a DNA sequence, and can be only recovered if the primer sequence is known [2]. In 2007, C.Chang proposed two data hiding schemes where data are hidden into DNA sequences in a way that preserves the functionality of the original DNA [9]. In 2010, H.Shiu et~al proposed three data hiding schemes based on DNA sequence named as : the insertion method, the complementary base- pair method and the substitution method. The most significant one was the substitution method, yet its hiding capacity is not efficient enough [10]. In 2012, A.Khalifa proposed a steganography technique, where data is encrypted using DNA-based playfair cipher, then it's hidden into a real DNA sequence using a modified substitution technique to increase its hiding capacity. Although it achieves higher hiding capacity than the original substitution method, yet the hiding capacity achieved is not the optimal one [11]. In the same year, Taur et~al proposed another modified substitution technique achieving a high hiding capacity but without encrypting data which minimized the technique's security [12]. In 2013, Zicheng Wang et~al proposed an information hiding scheme, they designed some procedures to encrypt a message using the vigenere cipher and then decompose the cipher text into two parts. The sender uses DNA steganography to send one part. If the microdot is contaminated, the message will be encrypted and decomposed again and again until the microdot is not assumed or contaminated. If the microdot is not contaminated, the second part will be publicly sent [13]. In 2015, Rashmi introduced a novel technique named as the DNA-Based audio steganography technique where a combination between the DNA steganography and audio steganography is obtained [14]. In the same year, S.Marwan proposed a steganography technique using a modified version to the playfair cipher as an encryption method before hiding where this modification leads to higher hiding capacity and better performance as well [15]. In 2016, G.Hamed et~al presented a survey on various DNA-based steganography techniques. It compares different DNA-based steganography techniques based on the cracking probability of each technique, besides

their advantages and disadvantages and ended by proposing some recommendations for achieving optimization in the DNA steganography field [16].

3 Proposed Technique

The proposed DNA-based steganography technique consists of two phases, the first one is from the sender's side and the second one is from the receiver's side. In case of the first phase, a preprocessing step and a hiding step will take place. In order to hide the encoded message efficiently, a specific reference DNA sequence must be chosen as a hiding medium. If we assumed that the length of the message is ML, then the chosen reference DNA sequence must have non-coding regions of length NL such that:

$$1 \leq ML \leq 2NL \qquad (1)$$

Moreover, in order to guarantee higher security level, the message binary bits must be encrypted first by any of the cipher techniques such as the vigenere cipher or the playfair cipher using a randomly selected key.

3.1 Hiding Phase

Assume we have a message *Msg*, a reference DNA Ref of length *RL*, and an encryption key *EKey*, then the preprocessing step and the hiding process will work as follows:

Procedure PREPROCESSING(***Msg, EKey*****)**

1. Convert *Msg* to its binary form *B*.
2. Encrypt *B* using *EKey*, where *EB* sequence will result.
3. Encode *EB* into a DNA format using Table I to obtain the encoded DNA message *ED*.

End Procedure

The output of the preprocessing phase is the encoded DNA message *ED*, which will be used as the input to the following hiding process. Table 2 is used for substituting the encoded DNA letter -message letter- with a reference letter to generate the resulting fake letter, for instance if the encoded DNA letter is A and the reference DNA letter is G then substitution (A, G) is T. This substitution table construction is briefly explained in [12].

Afterwards, all the resultant fake DNA letters will be concatenated together to form the fake DNA sequence F and it will be considered as the first DNA strand. The second strand will be generated from substitutions of the complementary sequence of the reference DNA *CRef*. This second strand will be considered as the key sequence needed for correct extraction and it will be annotated as *HKey*. Consequently, the proposed hiding process will work as follows:

Table 2. Substitution Table [12]

Reference DNA letter	Encoded DNA letter	Substituted DNA letter
A	A	C
A	C	A
A	G	G
A	T	T
C	A	A
C	C	C
C	G	G
C	T	T
G	A	T
G	C	G
G	G	A
G	T	C
T	A	G
T	C	T
T	G	A
T	T	C

Procedure HIDING(***ED, Ref*****)**
j=1
  for i=1 to RL **do**
    if Ref_i is a noncoding letter **then**
      F_i= Substitution (Ref_i, ED_j)
      $HKey_i$= Substitution ($CRef_i$, ED_{j+1})
      $j=j+2$
    else
      concatenate Ref_i to F_i
      concatenate $CRef_i$ to $HKey_i$
    end if
  end for
End Procedure

Afterwards, F sequence is confined into a microdot in a paper and then this paper will be sent to the receiver. If the sent paper is not contaminated then the sender will send the *HKey*, *EKey* through a secure channel.

Otherwise, the sender will regenerate another encryption key and another reference DNA sequence and start the whole steganography process all over again. This guarantees that whenever the microdot sent becomes suspicious, completely new fake DNA and keys will be generated.

3.2 Extraction Phase

The extraction process is the second phase of our proposed algorithm. It is worth noting that the primer sequences and the reference DNA Ref are shared between the sender and the receiver. The receiver will use the primer sequences in order to correctly extract the fake DNA from the microdot. The following extraction process - reverse of the substitution process- will take place based on F, Ref, HKey, and EKey:

Procedure EXTRACTION(***F, Ref, HKey, S*****)**
j=1
  for i=1 to RL **do**
    if Ref_i is a noncoding letter **then**
      ED_j= Substitution (Ref_i, F_i)
      ED_{j+1}= Substitution ($CRef_i$, $HKey_i$)
      $j=j+2$
    end if
  end for
End Procedure

Finally, using the resultant encoded DNA message *ED* and the encryption key *EKey*, the reverse of the preprocessing process will take place as follows:

Procedure REVERSE-PREPROCESSING(***ED, EKey*****)**
1. Decode the *ED* sequence into its binary form using Table I, where *EB* will result.
2. Use *EKey* to decrypt *EB*, where the binary sequence *B* will result.
3. Convert *B* to the original message form.

End Procedure

3.3 Illustrative Example

From the sender's side, assume a message: 011000101100 and the encryption key *EKey*: 2411

1. Encrypt the message using the encryption key *EKey*, then the encrypted message *EB* is: 110100100110.
2. Encode it into a DNA sequence using Table 1, such that the resultant sequence *ED* is TCAGCG.
3. Assume the chosen reference DNA sequence *Ref* is: ACT***TA***CC***GAC***GA and its complementary sequence *CRef* is: TGA***AT***GG***CTG***CT. The bold DNA bases refer to the noncoding regions.
4. Using Table 2, the resultant fake DNA sequence *F* is: ACT***CC***CC***G***ACGA and the resultant Key sequence is *HKey* is: TGA***AA***GG***G***TGCT, where the message is hidden through the DNA bases in bold.

It can be noted that the length of the encoded message *ED* is 6 and the length of the noncoding regions in *Ref* is 5 and since we can hide through the *Ref* sequence and its complementary sequence *CRef*, therefore the number of bases that we can hide through is 10. It's clear that the encoded message ED is hidden through Ref_4, $CRef_4$, Ref_5, $CRef_5$, Ref_8, $CRef_8$.

From the receiver's side, given *F, Ref, HKey*, and *EKey:*

1. Using Table 2, the extracted hidden encoded DNA sequence *ED* is: ACT***TA***CC***GAC***GA.
2. Convert *ED* into its binary form using Table 1, such that *EB* is: 110100100110.
3. Using the encryption key *EKey*, decrypt *EB*, such that the original message is: 011000101100.

Consequently, by using the proposed DNA-based steganography technique we can notice the following:

1. An encryption technique is used before hiding, to decrease the risk of suspecting the hidden message.
2. Since we are utilizing the two DNA strands, double the data capacity can be stored as well.
3. Since sending the Key took place after the safe arrival of the fake DNA sequence to the receiver, therefore full data extraction by the attacker is nearly impossible.
4. The proposed algorithm preserves the DNA sequence functionality since no mutations can occur in a noncoding region, and therefore, there is no possibility for data loss.

4 Cracking Probability

In this section the cracking probability of our algorithm is briefly discussed. In order for an attacker to extract the hidden message correctly, the attacker needs 5 types of information to decrypt a message, binary representation, substitution table, reference DNA, hiding technique and the ciphering technique. Since there exist only 4 DNA bases, therefore the probability to get the binary scheme *b* is:

$$P(b) = \frac{1}{4!} \tag{2}$$

Since there exist 4 DNA bases and each base can be substituted by one of four DNA bases, therefore the probability to get the correct substitution table *st* is:

$$\mathrm{P}(st) = \frac{1}{4!^4} \tag{3}$$

Since there exist 1.6 x 10^8 DNA sequences on the DNA database [17], therefore the probability to get the Reference DNA *r*:

$$\mathrm{P}(r) = \frac{1}{1.6 \times 10^8} \tag{4}$$

Therefore, the overall cracking probability k is:

$$\mathrm{P}(k) = \mathrm{P}(b) \times \mathrm{P}(st) \times \mathrm{P}(r) = \frac{1}{4!} \times \frac{1}{4!^4} \times \frac{1}{1.6 \times 10^8} \quad (5)$$

It's worth noting that this cracking probability can occur in case the attacker knows the ciphering technique, hiding technique and their keys, which means that our technique's cracking probability is even much harder than the previous calculated one.

5 Experiments and Results

In this section, the results of our proposed technique after implementation are discussed. The proposed technique is tested on eight real DNA sequences representing a benchmark data adopted from the NCBI database [17]. Table III shows the performance of the proposed technique regarding the amount of data stored. To be more precise, we assumed that the introns proportion in each of the eight DNA sequences is 45%, which means that if we have a DNA sequence of length 100 nucleotide, we can hide data in only 45 nucleotides to avoid any interference with the nucleotides of the exons sequences and therefore up to a great extent there will be no change in the DNA sequence functionality.

As shown in Table 3, the first column shows the unique code number for each DNA sequence used. The second column shows the length of the corresponding DNA sequence in base pairs (bp). The third column shows number of introns base pairs (bp) and column four shows the amount of data that can be stored in each sequence using our proposed technique. For example, if we have a reference DNA sequence of length 200,117 bp and since 45% of it are introns, therefore we can use only 90,053 bp from it for hiding data in, eventually, the amount of bits that can be stored in it is 360,212 bits. This amount is resulted since each base can hide up to two bits and the proposed technique is hiding in both strands of the reference sequence.

6 Conclusions

This paper proposed a new DNA-based steganography algorithm utilizing the two DNA strands in order to both maximize the amount of data to be stored and achieve a high security level. The proposed algorithm encodes any data type into a DNA format, encrypts it, then hides it in the noncoding regions of a suitably chosen reference DNA sequence. After the hiding process, a key sequence is generated of the same length as the resultant fake DNA sequence. The proposed algorithm is capable of storing double the amount of data than the other steganography techniques. Moreover, it provides a high security level due to the use of encryption step before hiding as well as the prevention of sending the hiding key sequence together with the fake DNA sequence. This means that the data full extraction will be impossible in this case. Additionally, the proposed technique preserves the DNA sequence functionality and therefore it avoids any mutations that can lead to data loss. The conducted experimental studies confirmed the effectiveness of our proposed technique and opens a

space of new methods and algorithms in utilizing both DNA strands. As a future work, we should focus on developing algorithms that can hide data through all the two DNA strands -the coding and noncoding regions- and at the same time preserve the DNA functionality.

Table 3. Results Of The Proposed Technique Performance

DNA Sequence Number	DNA Sequence Length(bp)	Number of Nucleotides in the Introns(bp)	Amount of Data that can be Stored(bits)
ACI53526	200,117	90,053	360,212
AC166252	149,884	67,448	269,792
AC167221	204,841	92,178	368,712
AC168874	206,488	92,920	371,680
AC168897	200,203	90,092	360,368
AC168901	191,456	86,156	344,624
AC168907	194,226	87,402	349,608
AC168908	218,028	98,113	392,452

7 References

[1] L.Adleman, "Molecular Computation of Solutions to Combinatorial Problems", Science11, vol.266, pp1021-1024, 1994.

[2] A.Leier, C.Richter, W.Banzhaf and Hilmar Rauhe, "Cryptography with DNA binary strands", BioSystems57, 2000.

[3] B.Shimanovsky, J.Feng, and M.Potkonjak, "Hiding data in DNA", Revised Papers from the 5th International Workshop on Information Hiding, vol.2578, pp 373-386, 2002.

[4] C.Catherine, V.Risca and C.Bancroft, "Hiding Messages in DNA Microdots", Nature Magazine, vol.399, 1999.

[5] I.Peterson, "Hiding in DNA", Muse, 2001.

[6] M.SAEB, E.EL-ABD, and M.EL-ZANATY, "On Covert Data Communication Channels Employing DNA Recombinant and Mutagenesis- based Steganographic Techniques", CEA'07 Proceedings of the 2007 annual Conference on International Conference on Computer Engineering and Applications, pp 200-206, 2007.

[7] H.Yahya, "If Darwin Had Known About DNA???", 1st edition, 2007.

[8] B.Alberts and A.Johnson, "Molecular Biology of The Cell", The publishing company, 5th edition,US, 2008.

[9] C.Chang, T. Chuen, Y.Chang, and C. Lee, "Reversible data hiding schemes for deoxyribonucleic acid (DNA) medium", International Journal of Innovative Computing, Information and Control ICIC, vol.3, pp 1145-1160, 2007.

[10] H.I. Shiu, K.L. Ng, J.F. Fang, R.C.T. Lee and C.H. Huang, "Data hiding methods based upon DNA sequences" , ELSEVIER, vol.180, pp 2196-2208, 2010.

[11] A.Khalifa and A.Atito, "High Capacity DNA-based Steganography", The 8th International Conference on INFOrmatics and Systems, 2012.

[12] Jin-Shiuh Taur, Heng-Yi Lin, Hsin-Lun Lee and C.Tao, "Data Hiding in DNA Sequences Based on Table Look Up Substitution", International Journal of Innovative Computing, Information and Control, vol.8, pp 6585-6598, 2012.

[13] H.W.G.C. Zicheng Wang, Xiaohang Zhao, "Information hiding based on DNA steganography", 4th IEEE International Conference on Software Engineering and Service Science (ICSESS), 2013.
[14] Rashmi M. Tank, Hemant D. Vasava, and Vikram Agrawal, "DNA-Based Audio Steganography", Oriental journal of Computer Science and Technology, vol.8, pp 43-48, 2015.
[15] S.Marwan, A.Shawish and K.Nagaty, "An Enhanced DNA based Steganography Technique with a Higher Hiding Capacity", BIOSTEC, Bioinformatics, SciTePress, pp 150-157, 2015.
[16] G.Hamed, M.Marey, S.Elsayed, and F.Tolba, "DNA Based Steganography: Survey and Analysis for Parameters Optimization", Applications of Intelligent Optimization in Biology and Medicine, Springer International Publishing, vol.96, pp 47-89, 2016. https://doi.org/10.1007/978-3-319-21212-8_3
[17] NCBI database, {http://www.ncbi.nlm.nih.gov/}.

8 Authors

Samiha Marwan, Lecturer Assistant in the Faculty of Informatics and Computer Science, The British University in Egypt, (e-mail: samiha.abdelrahman@bue.edu.eg).

Ahmed Shawish, Associate Professor in the Faculty of Informatics and Computer Science in the British University in Egypt, (e-mail: ahmed.gawish@bue.edu.eg).

Khaled Nagaty, Professor of Computer Science and the head of the department of Computer Science in the Faculty of Informatics and Computer Science in the British University in Egypt, (e-mail: khaled.nagaty@bue.edu.eg

This article is a revised version of a paper presented at the BUE International Conference on Sustainable Vital Technologies in Engineering and Informatics, held Nov 07, 2016 - Nov 09, 2016 , in Cairo, Egypt. Article submitted 26 December 2016. Published as resubmitted by the authors 09 March 2017.