Data Encryption and Hiding using Playfair and Complementary Techniques based on DNA Sequences

Ahmed Elhadad[1]

Computer Science and Information Department, College of Science and Arts, Jouf University, Saudi Arabia

Summary
In this paper, we propose a novel algorithm to communicate data securely. The proposed technique is a composition of both encryption and data hiding using some properties of **Deoxyribonucleic Acid** (DNA) sequences. Hence, the proposed scheme consists mainly of two phases. In the first phase, the secret data is encrypted using a DNA and Amino Acids-Based Playfair cipher. While in the second phase the encrypted data is steganographically hidden into some reference DNA sequence using a complementary technique. The proposed algorithm can successfully work on any binary data since it is actually transformed into a sequence of DNA nucleotides using some binary conversion rule. Subsequently, these nucleotides are represented as an amino acids structure in order to pass through the specially designed Playfair Cipher and encrypt it into another DNA sequence. Then, this encrypted DNA data is substituted using some reference DNA sequence to produce a faked DNA sequence with the encrypted data hidden. In order to recover the embedded secret data, the receiver can carry out the inverse process with the help of the both the secret key and the reference DNA sequence.

Key words:
DNA, Playfair, complementary technique, reference DNA, Playfair cipher.

1. Introduction

The growth of computers and communications systems brought with it a demand from the private sector for means to protect information in digital form and to provide security services. **Information security** means protecting information and information systems from unauthorized access, use, disclosure, disruption, modification, or destruction [1, 2]. Some of data may be secret information which is candidate to unauthorized access. In order to keep the unauthorized user away, variety of techniques have been used such as cryptography and data hiding.

Until modern times cryptography referred almost exclusively to encryption, which is the process of converting ordinary information (plaintext) into unintelligible gibberish (i.e., ciphertext). Decryption is the reverse, in other words, moving from the unintelligible ciphertext back to plaintext. A cipher is a pair of algorithms that create the encryption and the reversing decryption [2, 3].

Steganography is the art and science of writing hidden messages within another seemingly innocuous message, or carrier. The carrier could be any medium used to convey information, including wood or slate tablets, tiny photographs or word arrangements [4]. Modern Steganography techniques using digital information offering wonderful opportunities not only to hide information, but also to develop a general theoretical framework for hiding different kinds of data such as sound tracks, images, videos, and even 3D objects [5].

In biology, a **Deoxyribonucleic acid** (DNA) is the master molecule whose structure encodes all the information needed to create and direct the chemical machinery of life [6]. In 1953, the structure of DNA was correctly predicted by Watson and Francis Crick that DNA molecule consists of two long polynucleotide chains each of these chains is known as a DNA chain, or a DNA strand which is made from simple subunits, called nucleotides. Each nucleotide consists of a sugar-phosphate molecule with a nitrogen-containing side group, or base. The bases are of four types (adenine, guanine, cytosine, and thymine), corresponding to four distinct nucleotides, labeled A, G, C, and T [6].

The DNA-based cryptography is a new and very promising direction in cryptography research. DNA can be used in cryptography for storing and transmitting the information, as well as for computation [7]. Recently, a number of cryptographic techniques have been proposed to utilize the DNA digital format in the ciphering process itself. Some of these techniques such as One-time pads [8], RSA Algorithm [9, 10], Playfair cipher [11], DNA-based Encryption using pointers [12] and DNA Encryption using PCR [13].

Although different methods of Data hiding techniques were introduced including: invisible inks, microdots, digital signatures, and spread spectrum communications [14], DNA-based Data hiding techniques have been recently added to that list. These techniques depend on the high randomness of the DNA to hide any message without being noticed. In fact, DNA has many characteristics which make it a perfect Data hiding media. These characteristics have two significant facts; the DNA has tremendous information storage capacity. In addition, any DNA sequence can be synthesized in any desirable length [8].

Manuscript received January 5, 2022
Manuscript revised January 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.1.91

Inspired by the microdots used during the 2nd world war, Clelland et al. developed an extension of this principle [15] using DNA. Leier et al. in [16] encoded binary information into DNA sequences. The resulting DNA sequence is mixed with dummy strands and can only be detected and isolated if the primer sequence is known. In [17] Saeb et al. proposed two Biotechnological methods for hiding message into DNA using DNA Recombinant and DNA Mutagenesis. Hayam Mousa et al. introduced a reversible information hiding scheme for DNA sequence based on reversible contrast mapping in [18]. There are three more data hiding methods proposed in [19] based upon properties of DNA sequences.

In this paper, we will illustrate a DNA- based cryptography method as combined with data hiding techniques for an increased level of security. These methods are implemented mainly on two stages: the first encrypts the plaintext using Amino acid and DNA based Playfair cipher. The second applies a secure complementary method to hide the encrypted DNA ciphertext using some reference DNA sequence.

2. Encrypting data using a DNA and Amino Acids-Based Playfair cipher

Playfair is multiple letter encryption cipher, in which diagrams in the plaintext are treated as single units and these units are translated into cipher text diagrams. The traditional Playfair algorithm is based on the use of a 5 x 5 matrix of letters constructed using a keyword. The Playfair cipher is a great advance over simple monoalphabetic ciphers[3]. Cryptanalysis of the Playfair cipher is much more difficult than normal simple substitution ciphers, because digraphs (pairs of letters) are being substituted instead of monographs (single letters)[3]. However, Playfair cipher has some disadvantages when concerning secure data communications:

- Still leaves much of language structure such as numbers and punctuation.
- The frequency analysis of digraphs is still possible.
- The Playfair cipher can be easily cracked if there is enough text for cryptanalysis.
- Easy to break nowadays using computer processing.

So, modern researches focused their efforts in using tools to make Playfair cipher more secure and avoid some its disadvantages. One of these suggestions is Playfair cipher based on DNA. It introduces some modifications to the Playfair cipher processing by using some Biological concepts such as DNA and amino acids structures to the improve Playfair ciphering process[11].

Therefore, we apply Playfair cipher as the first phase of the proposed technique. In this phase, the secret text "plaintext" is encrypted using DNA and Amino Acids concepts [11] as follows: Convert plaintext to binary form such as 8-bits coding. After that, a binary coding scheme used to transform binary form into DNA alphabets A, C, G and T. For instance, the following may be a binary coding used: ((A 00) (C 01) (G 10) (T 11)). It should be noted that more digits may be used. Next, The DNA form is transferred to the Amino acids form according to a standard universal table of Amino acids and their codons representation in the form of DNA [20] [21] and the new distribution of the alphabet with the corresponding new codons in [11]. Now, English alphabets form of Amino Acids can go through traditional Playfair cipher process using the secret key [3]. Amino Acids form of encrypted data transferred to DNA Sequence form. Ultimately, at this phase we get encrypted DNA sequence.

3. complementary the encrypted data using reference DNA sequence

A *palindrome* is a word, phrase, number, or other sequence of units that can be read the same way in either direction[22]. For example, "Eva, can I stab bats in a cave?". The meaning of palindrome in the context of genetics is slightly different from the definition used for words and sentences. Since a double helix is formed by two paired strands of nucleotides that run in opposite directions in the 5'-to-3' sense, and the nucleotides always pair in the same way (Adenine (A) with Thymine (T) for DNA, with Uracil (U) for RNA; Cytosine (C) with Guanine (G)), a (single-stranded) nucleotide sequence is said to be a palindrome if it is equal to its reverse complement. For example, the DNA sequence ACCTAGGT is palindromic because its nucleotide-by-nucleotide complement is TGGATCCA, and reversing the order of the nucleotides in the complement gives the original sequence [23].

Based on the above property, the proposed complementary method hides the encrypted DNA ciphertext [DC] using palindrome complementary definition of DNA sequence in the reference DNA sequence [Ref]. For instance, assume in [Ref] the reference DNA sequence there is longest complementary palindrome substrings sequence with length of six such as "GTTAAC". This method must therefore insert complementary palindromes substring pairs with a length of seven padded with character 'T' before and after each one into [Ref] to ensure that the longest palindromes complementary substrings are the newly inserted ones in [Ref]. Next, for each longest complementary palindrome substrings in [Ref] insert [DC] nucleotides immediately before 'T' (which padded before) for example "**G**T**T**ACAATGT**T**". The sender sends the final faked DNA sequence together with many other DNA, or DNA-like, sequences to the receiver. The receiver then finds all the longest complementary substrings, extracts the encrypted message and hence recovers the

original sequence as it was hidden. At this phase we will use the encrypted DNA sequence and reference DNA sequence as input of the complementary technique. The complementary technique was originally introduced in [24]; however, we modified the technique to deal with DNA alphabets instead of the binary data form.

4. The Playfair – complementary Algorithm

The following is Playfair – complementary method algorithm based upon previous defined legal palindrome rule and reference DNA sequence [Ref] to hide the encrypted ciphertext [DC] using DNA Playfair cipher and figure 1 shows this diagram:

1- Convert plaintext [P] to binary form [BP] such as 8-bits coding.
2- Code binary data form into a DNA sequence [DP] by using the binary coding rule.
3- Transfer DNA sequence to Amino acids [AP] and save Ambiguity number AMBIG [AP].
4- Construct the Playfair matrix using the secret key [SK] and then apply traditional Playfair encryption process to get Amino acid of cipher text [AC].
5- Transfer Amino acid to DNA sequence and add Ambiguity number into nucleotide form [DC].
6- Find the longest complementary palindromes in [Ref] and let its length is [K]then Let [L] is the length of [DC].
7- Generate a set [A] = $\{a_1, a_2, a_3, \ldots, a_L\}$ of random complementary palindromes with length [K] +1 and pad each one with character 'T'.
8- Insert each nucleotide of [DC] before the first 'T' in set [A] complementary palindromes without overlapping.
9- Insert each substring in set [A] into [Ref] to form final faked DNA [C].

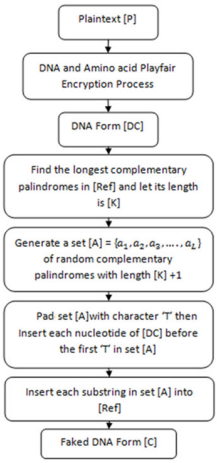

Fig. 1 Playfair – complementary method diagram.

Now, it is easy to recover the plaintext [P] by using Playfair secret key [SK] and the reference sequence [Ref] as the following algorithm:

1- Extract all longest complementary palindromes from [C]
2- Concatenate all nucleotides before 'T' to form [DC].
3- Extract Ambiguity number from [DC] after that Transfer remaining DNA sequence to the Amino acids [AC].
4- Construct the Playfair matrix using the secret key [SK] and then apply Playfair decryption process to get Amino acid of plain text [AP].
5- Transfer Amino acids [AP] to DNA sequence [DP] by using Ambiguity number.
6- Code DNA sequence form [DP] into binary form [BP] by using the binary coding rule.
7- Convert binary form [BP] to plain text [P].

So that there are two important things in the process of transferring secret data using this algorithm, first is the Ambiguity problem in decrypted data to construct the DNA form [DP] of plain text from the amino acid form [AP]. In[11], this problem has been solved by using two additional bits for each amino acid character to demonstrate which codon to choose. Our algorithm supports this solution to avoid the problem of ambiguity. While the second is hiding the faked DNA sequence [C] with many other DNA sequences. As in [24] the receiver processes every sequence received, extracts the encrypted DNA sequence [DC`] and recovers the reference DNA sequence [R`]. If the recovered reference DNA sequence [R`] is not a prefix of the original reference DNA sequence [R], it means that the receiver should test some other received sequences until the recovered sequence is exactly a prefix of [R].

5. Experiments and results

Here, the Playfair – complementary method experiment includes 1, 2, 3, 4, 5 Kilo bytes of plaintexts as various input and tables 2 and 3 listed these results.

Table 1: Playfair – complementary method times performance

LOCUS	Plaintext	Total processing time (Sec)
AEEX01000100	**1K**	27.05
	2K	113.71
	3K	258.90
	4K	463.83
	5K	761.99
AL157382	**1K**	24.05
	2K	97.77
	3K	224.25
	4K	403.25
	5K	655.03

Table 2: capacity, payload and bpn using Playfair – complementary method

LOCUS	Plaintext	Encrypted DNA Sequence	Capacity (c)	Payload (p)	Total bpn	Actual bpn
AC153526	1KB	5464	582597	382480	0.005	0.03
longest palindromes complementary 66	10KB	54616	4023237	3823120	0.051	0.27
	20KB	109228	7846077	7645960	0.102	0.55
	50KB	273068	19314877	19114760	0.256	1.36
	100KB	546136	38429637	38229520	0.512	2.73
AC167221	1KB	5464	576393	371552	0.005	0.03
longest palindromes complementary 64	10KB	54616	3918729	3713888	0.050	0.27
	20KB	109228	7632345	7427504	0.100	0.53
	50KB	273068	18773465	18568624	0.250	1.33
	100KB	546136	37342089	37137248	0.500	2.67
AC168901	1KB	5464	759712	568256	0.005	0.03
longest palindromes complementary 100	10KB	54616	5871520	5680064	0.053	0.29
	20KB	109228	11551168	11359712	0.107	0.57
	50KB	273068	28590528	28399072	0.267	1.43
	100KB	546136	56989600	56798144	0.535	2.85
AC168908	1KB	5464	2256100	2038072	0.005	0.03
longest palindromes complementary 369	10KB	54616	20589796	20371768	0.047	0.25
	20KB	109228	40960072	40742044	0.094	0.50
	50KB	273068	102072392	101854364	0.235	1.25
	100KB	546136	203926756	203708728	0.470	2.50

As shown in table (1), the results of total performance times for the Playfair- complementary method were not regular for the same plaintext because of various complementary processes in each reference. On the other hand, the total performances time for various plaintexts was approximately equal in each reference DNA sequence. The results in table (2) shows that, the length of the resultant faked DNA sequences denoted by the capacity (C) depended on the length of the longest complementary palindromes in each reference DNA sequence. This is in turn has its effect on the payload (P).

6. Conclusions

In this paper, we proposed a novel algorithm to communicate data securely. The Playfair- complementary technique is a composition of both encryption and data hiding using some properties of Deoxyribonucleic Acid (DNA) sequences. Hence, the proposed scheme consists mainly of two phases. In the first phase, the secret data is encrypted using a DNA and Amino Acids-Based Playfair cipher. While in the second phase the encrypted data is steganographically hidden into some reference DNA sequence using a complementary technique.

The scheme was implemented and tested on different DNA sequences for plaintext of sizes1, 2, 3, 4, and 5 Kilo Bytes. Although the proposed method includes two phases "Playfair and complementary", the proposed method gave a better time performance than the original encryption algorithm proposed in[11]. In addition, the values of Actual bpn were found to be approximately the same as the bpn values published in [19] despite of using an encrypted plaintext.

In conclusion, the proposed scheme can not only encrypt secret information into DNA sequences but also hide the encrypted data into another reference DNA sequence. Therefore, it is difficult for an attacker to detect whether or not there are secret messages hidden in a DNA sequence without knowing the embedding parameters. Even if an attacker knows that secret messages are present in the fake DNA sequence, it is still virtually impossible for the sequence to be correctly decrypted without Playfair cipher secret key.would like to itemize some parts of your manuscript, please make use of the specified style "itemize" from the drop-down menu of style categories

References

[1] "United States Code: Title 44,3542. Definitions | LII / Legal Information Institute," *Cornell university-Law school*.

[2] A. Menezes, P. v. Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*, 1997.

[3] W. Stallings, *Cryptography and Network Security Principles and Practices, Fourth Edition*, 2005.

[4] P. Wayner, *Disappearing Cryptography - Information Hiding Steganography and Watermarking Second Edition*, 2002.

[5] K. RAMA, K. THILAGAM, M. P. .S, A.JEEVARATHINAM, and K. .LAKSHMI, "SURVEY AND ANALYSIS OF 3D STEGANOGRAPHY," vol. 3, no. 1, 2011.

[6] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts *et al.*, *Molecular Biology of The Cell Fifth Edition*, 2008.

[7] X. Guozhen, L. Mingxin, Q. Lei, and L. Xuejia, "New field of cryptography: DNA cryptography," vol. 51 No. 12, 2006.

[8] A. Gehani, T. LaBean, and J. Reif, "DNA-Based Cryptography," 2000.

[9] X. Wang, and Q. Zhang, "DNA computing based Cryptography," 2009.

[10] "DNA encryption techniques," *www.scribd.com*, http://www.scribd.com/doc/55154238/11/DNA-encryption-techniques.

[11] M. Sabry, M. Hashem, T. Nazmy, and M. E. Khalifa, "A DNA and Amino Acids-Based Implementation of Playfair Cipher," vol. 8 No. 3, 2010.

[12] S. T. Amin, M. Saeb, and S. El-Gindi, "A DNA-based Implementation of YAEA Encryption Algorithm," *My Home Page: Magdy Saeb*, http://www.magdysaeb.net/images/DNAYAEAaminsaeb.pdf.

[13] D. Prabhu, and M. Adimoolam, "[1101.2577v1] Bi-serial DNA Encryption Algorithm(BDEA)," *arXiv.org*,

http://arxiv.org/ftp/arxiv/papers/1101/1101.2577.pdf, [1 13, 2011].
[14] H. T. Sencar, M. Ramkumar, and A. N. Akansu, *Data Hiding Fundamentals and Applications: Content Security in Digital Media*: Elsevier Academic Press, 2004.
[15] C. T. Clelland, V. Risca, and C. Bancroft, "Hiding messages in DNA microdots," vol. 399, 1999.
[16] A. Leier, C. Richter, W. Banzhaf, and H. Rauhe, "Cryptography with DNA binary strands," vol. 57, 2000.
[17] M. SAEB, E. EL-ABD, and M. E. EL-ZANATY, "On Covert Data Communication Channels Employing DNA Recombinant and Mutagenesis-based Steganographic Techniques," 2007.
[18] H. Mousa, K. Moustafa, W. Abdel-Wahed, and M. Hadhoud, "Data Hiding Based on Contrast Mapping Using DNA Medium," vol. 8 No. 2, 2011.
[19] H. J. Shiu, K. L. Ng, J. F. Fang, R. C. T. Lee, and C. H. Huang, "Data hiding methods based upon DNA sequences," vol. 180, 2010.
[20] W. JD, B. TA, B. SP, G. A, L. M *et al.*, *Molecular Biology of the Gene*, 2004.
[21] "DNA codon table - Wikipedia, the free encyclopedia," *Wikipedia*.
[22] "Palindrome - Wikipedia, the free encyclopedia," *Wikipedia, the free encyclopedia*.
[23] "Palindromic sequence - Wikipedia, the free encyclopedia," *Wikipedia, the free encyclopedia*.
[24] H. J. Shiu, K. L. Ng, J. F. Fang, R. C. T. Lee, and C. H. Huang, "Data hiding methods based upon DNA sequences," *Information Sciences*, 2010.

Ahmed Elhadad received the BSc, MSc, and PhD degrees from South Valley University, Qena, Egypt, in 2007, 2010, and 2015 respectively. He was working as an associate professor in the Computer Science Department, Faculty of Computer and Information sciences – South Valley University, Egypt. In 2017, He worked as a Post-Doc Researcher at CRACS & INESC - Porto LA, Faculdade de Ciências, Universidade do Porto, Porto - Portugal. In 2012, he was granted mobility scholarship to (IST), Lisbon - Portugal. Currently, he is working in the College of Science and Arts, Jouf University, Saudi Arabia. His research is focused on the area of information security and privacy.