



Journal of Information and Optimization Sciences

ISSN: 0252-2667 (Print) 2169-0103 (Online) Journal homepage: https://www.tandfonline.com/loi/tios20

Information hiding using artificial DNA sequences based on Gaussian kernel function

Eman I. Abd El-Latif & M. I. Moussa

To cite this article: Eman I. Abd El-Latif & M. I. Moussa (2019): Information hiding using artificial DNA sequences based on Gaussian kernel function, Journal of Information and Optimization Sciences, DOI: 10.1080/02522667.2017.1413041

To link to this article: <u>https://doi.org/10.1080/02522667.2017.1413041</u>



Published online: 10 Jul 2019.



📝 Submit your article to this journal 🗹



View Crossmark data 🗹

Journal of Information & Optimization Sciences ISSN 0252-2667 (Print), ISSN 2169-0103 (Online) DOI : 10.1080/02522667.2017.1413041



Information hiding using artificial DNA sequences based on Gaussian kernel function

Eman I. Abd El-Latif * Department of Mathematics Faculty of Science Benha University Benha Egypt

M. I. Moussa⁺ Department of Computer Sciences Faculty of Computers and Informatics Benha University Benha Egypt

Abstract

Cryptography is one of the major concerned areas of computer networks and data security. An efficient direction of providing data security can be termed as DNA based on cryptography. In this paper, a new reliable and more secure approach for data hiding based on DNA sequences is introduced. The proposed approach has two rounds of encryption. This encryption is similar to an existing encryption method called the Data Encryption Standard (DES) algorithm. Through cryptography algorithm two secret keys are used for encryption the message. The first key is induced from the elliptic curve cryptography (ECC) and gaussian kernel function (GKF). The second key is constructed based on an arbitrary injective mapping on the second characters repeated in the first key. Finally, the encryption message randomly hides in the second DNA sequence based on the numbers from GKF.

Subject Classification: 14H52, 35J75 *Keywords:* DNA Cryptography, DNA Sequences, DES, ECC, GKF.

^{*}E-mail: eman.mohamed@fsc.bu.edu.eg (Corresponding Author) *E-mail: mahmoud.mossa@fci.bu.edu.eg



1. Introduction

DNA cryptography is hiding a data in terms of DNA sequences, which can be done using several DNA technologies with the biological methods. In recent years, the use of data hiding approach has become popular in transmitting secret messages. Data hiding approaches based on the DNA sequence attracted much attention to avoid malicious intruding and fulfill a safe transmission. The DNA sequence is made up of four different types of bases, Guanine-G, Adenine-A, Thymine-T and Cytosine-C. Each base is attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide. For example, segment of LN611623.1 sequence retrieved from the European Bioinformatics Institute (EBI) [1] as shown figure 1.

The order of these bases determines the information available for building and maintaining an organism. DNA has much more storage capacity which is equal to (1gm = 10^8 TERA bytes). It means small amount of DNA can stores world's information. DNA cryptography method is one of the new techniques in cryptographic field that can provide higher security of the information. Cryptographic schemes use one key to encrypt and decrypt the message is referred as symmetric-key. A few well-known symmetric key cryptography examples are; the Data Encryption Standard (DES), Triple-DES (3DES) and the Advanced Encryption Standard (AES). The Cryptographic schemes use two keys; one is used for encryption "called public key" and the other one for decryption "called private key" are referred as asymmetric key encryption. In the majority of cryptographic

Figure 1

A segment of length 840 bases from the LN611623.1 of length 2,634 bases

applications in practical systems, symmetric and asymmetric algorithms are used together to construct hybrid schemes.

To deal with data hiding problem, many cryptographic schemes were proposed with cryptographic based on DNA such as a symmetric DNA-based cipher approach [2][3]. Secret message was encrypted using RSA algorithm and then hidden in DNA sequence using complementary character [2]. A new scheme to hide two secret bits from a message by replacing one character in DNA sequence is described recently. The researchers used a kind of mapping between one complementary rule and the two secret bits to hide the message and then send the fake DNA sequence [3]. An algorithm is proposed by using software point of view for implementing data hiding based on DNA sequences. Both of DNA's features and binary coding technology beside complementary pairing rules are needed. Data hiding is started by applying three different and separate steps to prepare cipher message [4]. A session keys are shared between the sender and the receiver instead of sharing the actual keys between them. These keys contain the information about the actual key that is used for encrypting the message. This DNA sequence is one of the key to encrypt the message in next step and then get the fake DNA sequence. Add some extra bits at the beginning and at the ending of the faked DNA sequence and send the total form of DNA sequence to the receiver[5][6]. New method provides a secure and reliable data transmission has 3 subphases. They are the key generation, data encryption and the use of DNA encoding. The key generation is done by selected two 128-bit DNA sequence randomly from publicly available DNA sequences. These two selected DNA sequences will produce two encryption keys after performing a large number of computations on it. The data encryption technique is proposed where two rounds encryption has been carried out among the plain text and the generated two secret keys. A DNA encoding is converted every hex digit into a corresponding DNA representation of 2 DNA bases [7].]. Eman and M. I. Moussa have proposed an algorithm based on two dimension chaotic system and DNA sequence. This algorithm uses the two dimensional chaotic map to generate two artificial DNA sequences S1and S2. The sender uses the first sequence S1 for the encryption and uses the second sequence S2 to hide the cipher message randomly in a real third sequence S3, which is selected from DNA database [8]. Neil Koblitz and Victor Miller [9, 10] discovered Elliptic curve cryptography (ECC) in 1985. ECC is an approach to public-key cryptography based on the algebraic structure of elliptic curves over finite fields. One of the main benefits in comparison with non-ECC cryptography is the same

level of security provided by keys of smaller size. Fatma and M. I. Moussa introduced a new data hiding algorithm based on deoxyribonucleic acid (DNA) sequence, where a DNA coding is used to encode plaintext instead of the classical 8-bit ASCII coding. The algorithm based on two DNA sequences [11]. Based on the properties hiding data in DNA sequence has been attracting much attention and research work has been carried out to propose several new methods [12, 13, 14, 15, 16].

The paper is organized as follow. Section 2 briefly introduces elliptic curve cryptography and the secret keys. Section 3 presents the proposed encryption and decryption algorithms. Section 4 introduces the security analysis; the experimental and comparisons results are given in Section 5. Finally, section 6 is the conclusion.

2. The Proposed Approach

The idea of the proposed scheme is to use ECC to generate points (x, y) that applied on GKF to generate the position of hiding the message and the numbers to select the character from DNA sequence to generate the first key. The second key is generated by indicated the second characters repeated in the first key and then establish a kind of injective mapping between one character from first key and one complementary rule. The secret message is encrypted through two levels using DES and the two keys and then hides the encryption message into another DNA sequence.



2.1 Proposed System and Complementary Rules

Let *P*, *Q* be prime numbers and choosing the variables *a*, *x*, *y* and b within the field of *Fp*. The elliptic curve generates all points (x, y) which satisfy the elliptic curve equation modulo p as follows in equation (1):

$$y^2 \mod P = x^3 + ax + b \mod P \tag{1}$$

All points (x, y) from equation (1) are applied in equation (2) and then choose prime number Q to generate the position to hide the message as follows equation (3):

$$F_{i}(x_{i}, y_{i}) = \exp\left(-\frac{|x_{i} - y_{i}|^{2}}{2\mu^{2}}\right)$$
(2)

$$Zi = mod(Fi^*10^14, Q)$$
 (3)

The adjustable parameter μ plays a major role in the performance of the Gaussian function, and should be carefully tuned to the problem at hand. When the value of μ is equal to 0, the values of Z_i appear sequential and there is no difference in places such. When the value of μ . is lies between 0 and 1, the values of Z_i is differed slight. When the value of μ . is greater than 1, the values of Z_i are completely different, and this choice is more suitable for hiding the secret in the reference DNA sequence.

For each letter s of a DNA sequence, all the following s, $\Gamma(s)$, $\Gamma(\Gamma(s))$ and $\Gamma(\Gamma(\Gamma(s)))$ are different, where $\Gamma(s)$ is the complementary of s. This case induced a one to one and onto map Γ . For example; if we have a complementary rule defined as (A-T) (C-G) (G-A) (T-C), where $\Gamma(A) = T$, $\Gamma(C) = G$, $\Gamma(G) = A$, $\Gamma(T) = C$.

2.2 Generating of Keys

In this algorithm, we generated two keys that are used in encryption algorithm. Elliptic curve equation is used for generate a sequence of pair

Algorithm 1: Keys Generation

Input : A reference DNA sequence $S = \{s_1, s_2, ..., s_k\}$ a secret message $M = \{m_1, m_2, ..., m_n\}$ and the proposed system and complementary rules

Output : Key K_1 and key K_2 .

Step 1: Choose a prime number *P*.

Step 2 : Use equation (2) from the proposed system and complementary rules to generate a sequence of pair (x_i, y_i)

Step 3 : Compute GKF

$$F_i(x_i, y_i) = \exp\left(-\frac{|x_i - y_i| 2}{2\mu^2}\right)$$

Step 4 : Compute $Z_i = mod (F_i * 10^{14}, Q)$

Step 5 : We sort the randomly generated sequence Z_i in ascending order and store them in an array Z, then add the value Z[i] to the index i.

Step 6 : Calculate $K_1 = s_{Z_1} \oplus s_{Z_2} \oplus s_{Z_3} \oplus ... \oplus s_{Z_p}$

Step 7 : For integer j = 1 to *P*

Apply the complementary rule $\Gamma(s)$ on the second repeated letter in K_1

$$K_1 = s'_{Z_1} \oplus s'_{Z_2} \oplus s'_{Z_3} \oplus \dots \oplus s'_{Z_p}$$

 (x_i, y_i) and then put this pairs in GKF to generate sequence of integer numbers. Select nuclide from DNA sequence based on integer numbers to generate first key. The second key is generated by applying the complementary rule on the second repeated letter in first key.

2.3 Generating the Expanded Number

In algorithm 2, we generate expanded numbers to expand the message by using GKF. GKF is generating sequence of random numbers and then we take this numbers and apply mod on it. We read sequence of numbers from left to right where every number appears exactly twice.

Algorithm 2 : Expanded Numbers Generation

Step 1: Use equation (2) and equation (3) to generate integer numbers

Step 2: $N = mod(Z_i, R)$, where R is the length of IC_2 in encryption algorithm

Step 3 : Take numbers from step 2 and read them from left to right where each number appears twice.

2.4 Message Cryptography Algorithm

Obviously, there is an original message M (M has two parts L and R, where L is left side and R is right side) which the sender decides to send via a network to another person who is called receiver. So, there are two levels to get the final from of the message. Two levels used two keys that are generated from algorithm 1.

LEVEL 1 : Encrypt the secret message using K_1

Step 1 : Convert Message into ASCII Binary Code

Step 2 : Divide M into Two Halves, Left and Right Hand Side

$$M = M(L) \oplus M(R)$$

Step 3 : Apply Bitwise X-OR Operation between K_1 and M(R) to produce IC_1

$$IC_1 = M(R) \oplus K_1$$

Step 4 : IC_2 = Apply S_0 -BOX into IC_1

Step 5: Expand IC_2 by random number generation (IC_3)

Step 6: Apply Bitwise XOR Operation between IC_3 and M(L) to get IC_4

Step 7 : Put IC_4 in the left half and right half is M(R)

 $IC_5 = merge(IC_4, M(R))$

Step 8 : swap the halves to generate IC_6

 $IC_6 = merge(M(R), IC_4)$



Message encryption diagram of level 1

LEVEL 2 : Encrypt the secret message using K_2 Step 1 : Divide IC_6 into Two Halves, Left and Right Hand Side $IC_6 = IC_6 (L) \oplus IC_6 (R)$ Step 2 : Apply X-OR between right side and K_2 $IC'_2 = IC_6 (R) \oplus K_2$ Step 3 : Apply S_0 -BOX into IC'_2 to get IC'_3 Step 4 : IC'_4 = Expand IC'_3 Step 5 : Apply X-OR between IC'_4 and IC_6 (L) $IC'_5 = IC'_4 \oplus IC_6$ (L)

Step 6 : Concatenate output of IC'_4 in left hand and $IC_6(R)$ in right hand to Produce IC'_5



Message encryption diagram of level 2

2.5 Message Hiding Algorithm

In Algorithm 3, we select another DNA sequence to hide the encryption message inside it. When using GKF a sequence of ascending numbers are generated to replace each nuclide from DNA sequence by nuclide from artificial DNA sequence(IC'_6). Finally send fake DNA sequence to the receiver.

Algorithm 3 : Generation position for hiding (*a*, *b*, *p*, *q*) **Input :** A reference DNA sequence S_2 and IC'_6 **Output :** The hidden secret message *M* **Step 1 :** For i = 1 to length (*M*)

Compute $z_i(x_i, y_i) = \exp\left(-\frac{|x_i - y_i|^2}{2\mu^2}\right) \mod Q$

end for

```
Step 2 : For i = 1: length (IC'_6)
```

Change $S(z_i)$ to be $IC'_6(i)$

end for

Step 3 : send fake DNA sequence to the receiver.

3. Example

We explain the message cryptography algorithm in this example through four steps. The first step is used to generate the first key and the second key for using in encryption the message. The second step is used to encrypt the secret message using first key through level 1 and third step is encrypt the secret message using second key through level 2. The finally step is used to hide the secret message into DNA sequence.

Step 1: Generate the first key and the second key

- 1. Choose S₁ = ATCGAATTCGGGCTGAGTCACAATTCGCG CTGAGTGAACC
- 2. With a = 5, b = 5 and p = 37, the elliptic curve the equation over F_{37} is; $y^2 = x^3 + 5x + 5$. The set of points which satisfy this equation are: (2,7) (2,19) (5,5) (5,21) (6,11) (6,15) (9,5) (9,21) (11,13) (12,5) (12,21) (15,7) (15,19) (18,5) (18,21) (19,11) (19,15) (22,5) (22,21) (24,13) (25,5) (25,21)

- 3. Compute $z_i = \{1, 4, 5, 11, 16, 18, 19, 20, 22, 23, 24, 28, 29, 32, 36, 37\}$ where $\mu = 2$
- 4. $K_1 = S_1(z_i) = AGAGATCAAATCGGGA$
- 5. Apply the complementary rule on K_1 according to the injective map Γ and get $K_2 = AGAGATCAAATCGTTA$

Step 2 : Encrypt the secret message using *K*₁

- 1. Let the message (M) = "hi" = AGAGCGGACGGCAGAG
- 2. Divide *M* into left M(L) and right M(R)

M(L) = AGAGCGGA

M(R) = CGGCAGAG

- 3. $IC_1 = K_1 \oplus M(R) = CAGTACCG$
- 4. Apply S_0 –BOX into IC_1 to generate IC_2 = TTAC
- 5. Expand IC_2 to get $IC_3 = TATCATCT$
- 6. $IC_4 = IC_3 \oplus M(L) = TGTTCCTT$
- 7. IC_6 = Concatenate *R* and IC_4 = CGGCAGAGTGTTCCTT

Step 3 : Encrypt the secret message using *K*₂

- 1. $IC'_2 = IC_6(R) \oplus K_2 = TATCCGGT$
- 2. Apply S_0 -Box into IC'_2 to get $IC'_3 = TCCT$
- 3. Expand IC'_{3} to get $IC'_{4} = CCCTCTTT$
- 4. $IC'_5 = IC'_4 \oplus IC_6(L) = ATTGCCTC$
- 5. Concatenate IC_5' and $IC_6(R)$ to get IC_6'

 $IC_{5}' = ATTGCCTCTGTTCCTT$

Step 4 : Hide the secret message (IC'_6) using

 $z_i = \{1, 4, 5, 11, 16, 18, 19, 20, 22, 23, 24, 28, 29, 32, 36, 37\}$

Select another DNA sequence $S_2 = TACCACGTCGTGTC CCA$ GGACCATACGGTGAACGTAAACGCTTAAAATTTAGGGC TCCCAGTCG

After hiding message we get the fake DNA sequence:

Fake DNA sequence = AACCTTGTCGGGTCCCACTCCTGT ACGTCGACCGTTTACGCTTAAAATTTAGGGCTCCCAGT CG

4. Decryption and Data Recovery Algorithm

The receiver uses the following algorithm to get the plaintext. First, we get the artificial DNA from fake DNA sequence by using the numbers that generated from GKF. By applying decryption steps in artificial DNA to get original message.

Algorithm 4-1: Data Recovery Algorithm

Step 1: Extract the artificial DNA sequence (IC'_{6}) according to the sequence

 $z_i(x_i, y_i)$ **Step 2 :** Divide (*IC*'₅) into two halves, *IC*'₆ (*R*), *IC*'₆ (*L*)

Step 3 : Apply XOR between $IC'_6(R)$ and K_2

$$IC'_{2} = K_{2} \oplus IC'_{6}(R)$$

Step 4 : Apply S_0 –BOX into IC'_2 to produce IC'_3

Step 5 : IC'_4 = Expand IC'_3

Step 6 : Apply XOR between IC'_4 and $IC'_6(L)$

 $M(R) = IC'_4 \oplus IC'_6(L)$

Step 7 : Concatenate M(R) and $IC'_{6}(R)$ and then swap them

Step 8 : Apply XOR between right (step 7) and K_1 to get IC_1

Step 9 : Apply S_0 –BOX in IC_1 to get IC_2

Step 10 : Expand *IC*₂ to get *IC*₃

Step 11 : Apply XOR between left (step 7) and IC_3 to get M(L)

Step 12 : Put output of step 11 in left side and put the output of step 6 in rights to get original message

5. Security Analysis

This section discusses the strength, the robustness and the security issues of the proposed scheme. There are roughly 163 million DNA sequences available publicly. Thus, the probability of an attacker making a successful guess is $\frac{1}{1.63 \times 10^8}$. Another DNA sequence which is used to hide the secret message, so the probability of the attacker making a successful guess for the second selection is $\frac{1}{1.63 \times 10^8}$. The probability of an attacker to make a successful guess for the complementary rule is $\frac{1}{6}$. The number of points in the elliptic curves over the field F_p is $E(F_p) | \le 1 + p + 2\sqrt{p}$. The final probability of guessing the secret message is $\left(\frac{1}{1.63 \times 10^8}\right) \mathbf{x} \left(\frac{1}{6}\right) \mathbf{x} \left(\frac{1}{1+p+2}\right)$

Table 1

The experimental results of the proposed scheme

٢

Locus	Specifies definition	No. of nucleotides	Capacity C	Payload	bpn= M /C [11]	bpn= M /C the propose
153526	Mus musculus 10 BAC RP23-383C2	200, 117	200,117	0	0.579	1.600
166252	Mus musculus 6 BAC RP23-100G10	149, 884	149,884	0	0.730	1.610
167221	Mus musculus 10 BAC RP23-3P24	204, 841	204,841	0	0.566	1.600
168874	Bostaurus clone CH240-209N9	206, 488	206,488	0	0.561	1.601
168897	Bostaurus clone CH240-190B15	200, 203	200,203	0	0.579	1.599
168901	Bostaurus clone CH240-18511	191, 456	191,456	0	0.605	1.604
168907	Bostaurus clone CH240-19517	194, 226	194,226	0	0.597	1.602
168908	Bostaurus clone CH240-95K23	218, 028	218,028	0	0.529	1.597

INFORMATION HIDING USING DNA SEQUENCES

6. Experimental Results

A series of experiments carried out to evaluate the performance of the proposed scheme. Table 1 displays the experimental results in terms of the parameters used to evaluate the performance (capacity, payload and bpn). As shown, eight DNA sequences are used as the test sample in first column. These DNA sequences are publicly available by accessing the National Center for Biotechnology Information database (NCBI). The third column shows the number of nucleotides before hiding the secret message, and the fourth column shows the total length of the faked sequence after hiding the secret message, the fifth column shows the remaining length of new sequence after extracting out the reference DNA sequence. The bpn columns show the number of bits hidden per characters by applying the previous approaches in [8] but last column shows results of applying the proposed approach. Capacity and payload show that the length of the fake reference DNA sequence is not expanded. Furthermore, the proposed scheme has an acceptable embedding capacity, which is stable with different reference DNA sequences. Capacity and payload show that the length of the fake reference DNA sequence is not expanded. Furthermore, as bpn is within [1.59, 1.60], the proposed scheme has an acceptable embedding capacity, and the embedding capacity is stable with different reference DNA sequences.

7. Conclusion

Basically, the purpose of cryptography and steganography is to provide a very high degree of security for the data. Before encryption the message we must generate two keys, the first key is generated by GKF and the second key is generated by indicated the second characters repeated in first key and then establish a kind of injective mapping between one character and one complementary rule. The secret message is encryption in two levels using DES. Finally hide message in another DNA sequence.

8. Reference

- European Bioinformatics Institute, http://www.ebi.ac.uk/, 13.50pm, 18 September 2015.
- [2] B. A. Mitras and A. K. Aboo, "Proposed Steganography Approach Using Dna Properties", vol.14, (2012).

- [3] C. Guo, C. -C. Chang, and Z. -H. Wang, "A new data hiding scheme based on DNA sequence," Int J Innov Comput Inf Control, vol. 8, pp. 1-11, (2012).
- [4] M. R. Abbasy, P. Nikfard, A. Ordi, and M. R. N. Torkaman, "DNA Base Data Hiding Algorithm," *International Journal of New Comput*er Architectures and their Applications (IJNCAA), vol. 2, pp. 183-192, (2012).
- [5] N. Kar, A. Majumder, A. Saha, A. Jamatia, K. Chakma, and M. C. Pal, "An improved data security using DNA sequencing," in Proceedings of the 3rd ACM MobiHoc workshop on Pervasive wireless healthcare, vol.10, pp. 13-18, (2013).
- [6] B. Roy and A. Majumder, "An Improved Concept of Cryptography Based on DNA Sequencing," IJECCE, vol. 3, pp. 1264-1267, (2012).
- [7] A. Majumdar and M. Sharma, "Enhanced Information Security using DNA Cryptographic Approach," *International Journal of Innovative Technology and Exploring Engineering* (IJITEE), vol. 4, pp. 72-76, (2014).
- [8] Eman I. Abd El- Latif, and M. I. Moussa, "Chaotic Information- hidingAlgorithm based on DNA", *International Journal of Computer Applications*, vol. 122, no. 10, pp. 41-45, (July 2015).
- [9] D. Hankerson, S. Vanstone, and A. J. Menezes, "Guide to elliptic curve Cryptography", Springer Science & Business Media, (2004).
- [10] V. Miller, "Use of elliptic curves in cryptography," in Advances in Cryptology—CRYPTO'85 Proceedings, pp. 417-426, (1986).
- [11] Fatma E. Ibrahim, M. I. Moussa, and H. M. Abdalkader, "Enhancing the Security of Data Hiding Using Double DNA Sequences", presented at Industry Academia Collaboration Conference (IAC), 6-8 April, Cairo, Egypt, (2015).
- [12] E. I. Fatma, I. M. Mahmoud and S. A. Hatem, "A Symmetric Encryption Algorithm based on DNA Computing," *International Journal of Computer Applications*, vol. 97, no. 16, pp. 41-45, (2014).
- [13] D. Bhattacharyya and S. K. Bandyopadhyay, "Hiding Secret Data in DNA Sequence "," *International Journal of Scientific & Engineering Research*, vol. 4, (2013).
- [14] Jin-Shiuh Taur, Heng-Yi Lin, Hsin-Lun Lee, Chin-Wang Tao," Data hiding in DNA sequences based on table lookup substitution", *International Journal of Innovative Computing, Information and Control*, vol. 8, no.10(A), pp. 6585–6598, (2012).

- [15] E. Bashier, G. Ahmed, H.-A. Othman, and R. Shappo, "Hiding Secret Messages using Artificial DNA Sequences Generated by Integer Chaotic Maps," *International Journal of Computer Applications*, vol. 70, pp. 1-5, (2013).
- [16] C.-C. Chang, T.-C. Lu, Y.-F. Chang, and R. Lee, "Reversible data hiding schemes for deoxyribonucleic acid (DNA) medium," *International Journal of Innovative Computing, Information and Control*, vol. 3, pp. 1145-1160, (2007).

Received May, 2016