Chaotic Information-hiding Algorithm based on DNA

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

Mahmoud I. Moussa
Department of Computer Sciences,
Faculty of Computers and Informatics,
Benha University, Benha, Egypt

ABSTRACT
This paper proposes a cryptography technique based on two dimension 2D chaotic system and DNA sequence. The 2D chaotic map generates two artificial DNA sequences $S_1$and S_2 to get a cipher message by encrypting the plain message using the first sequence S_1. The sender uses the second sequence S_2 to hide the cipher message randomly in a real third sequence S_3 that is selected from DNA database. The hamming code approach is applied on the original message M to ensure reliable secret communication. The proposed approach provided a more security method compared to previous approaches.

General Terms
Chaotic maps, Hamming Code.

Keywords
DNA, Cryptography, Chaotic maps, and Hamming code.

1. INTRODUCTION
Deoxyribonucleic acid (DNA) is a molecule that encodes the genetic instructions used in the development and functioning of all known living organisms and many viruses. Each nucleotide is composed of a nucleobase (guanine, adenine, thymine, and cytosine), recorded using the letters G, A, T, and C. DNA does not usually exist as a single molecule, but as a pair of molecules that are held tightly together. DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide.

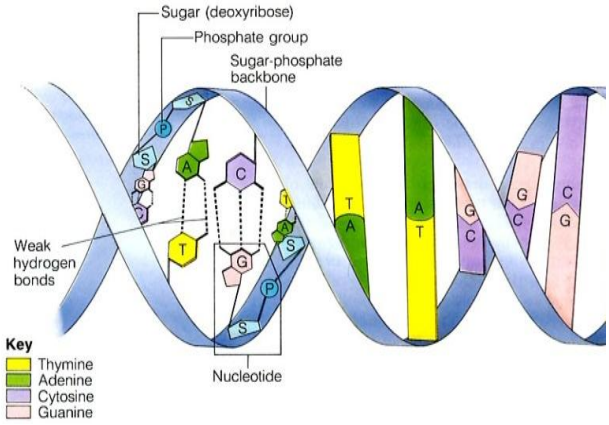

Figure 1.Base pairing in DNA Structure, Nucleobase (C paired with G, A paired with T)

DNA sequences have some inherent properties and a computational ability, which can be utilized to hide data because it is difficult to distinguish between a real DNA sequence and a fake one. Other knows advantage of the DNA is that one gram of DNA is capable to store 10^8terabytes of data. If we use the four DNA bases A, T, C and G, which can be encoded based on 0 or 1 by four digits: 0=A(00), 1=C(01), 2=G(10), 3=T(11), then each alphabet symbol is encoded by a triplet of DNA bases elements in such a way that the natural construction of the amino acids is simulated. There are 64 known 3-letter combinations of the DNA coding units T, C, A and G, which are used either to encode one of these amino acids or as one of the three stop codons that signals the end of the DNA sequence.

In 2000, Leier et al. proposed a robust scheme using a special key sequence, called a primer. A primer is a short complementary substring of a DNA sequence. It used to decode an encrypted DNA sequence. The receiver has the same technological capabilities as the sender, without the primers and the designated sequences it is not possible to correctly decode the binary data[1]. In 2012, new method was presented to hide the message in DNA sequence. Secret message was encrypted using RSA algorithm and then hidden in DNA sequence using complementary character[7]. In 2010, Shiu et al. proposed three data-hiding methods based on the DNA sequence: the insertion method, the complementary pair method and the substitution method. In each method; the secret message M is embedded into a reference DNA sequence S resulting a new reference sequence with data hidden S'. The insertion method increases the redundancy and expands the length of the DNA reference sequence because bits from secret message M are inserted one at a time into the reference DNA sequence. DNA sequence resulting from the substitution method has a highly modification rate. The fake sequence S′ maintains the original sequence length by replacing the nominated characters. However, the reference sequence S sent to the receiver to identify and extract the message hidden in S′. In addition, the replacement process is related to one secret message bit embedded in the DNA sequence, for the long secret message, the sequence will suffer a high medication rate. The complementary pair method maintains the DNA sequence for embedding the secret message based on a particular complementary rule[2, 3]. In 2012, a novel method uses both encryption scheme and data hiding together to communicate data securely. The secret message is encrypted using DNA and Amino Acids-Based Play fair cipher, and then encrypted data is hidden into some DNA sequence using an insertion technique. In order to recover the embedded secret data, the receiver carries out the inverse process with the help of the both the secret key and the reference DNA sequence[4]. In 2013, a new algorithm is proposed by using the byte values of the secret message and multi-keys. The encryption algorithm runs in two levels of encryption and contains a new method for generating the key[5, 6]. In 2014, a new research work has been carried out to propose new method using One Time Pad (OTP) encryption algorithm in [9].

The rest of the paper is organized as follow. Section 2 briefly introduces related work. Section 2 presents the proposed method. Section3 introduces the security analysis; the experimental results are given in Section 4. Finally, section 5 is the conclusion.

2. THE PLAINTEXT SECURITY APPROACH BASED ON CHAOTIC MAPS

Chaotic maps have been used extensively in cryptography during the past two decades, particularly for multimedia encryption. The proposed chaotic system is associated with the two-dimensional can be written as follows:

$$x_{i+1} = \mu_1 x_i(1 - x_i) + \gamma_1 y_i^2 \quad \text{(I)}$$

$$y_{i+1} = \mu_2 y_i(1 - y_i) + \gamma_2(x_i^2 + x_i y_i)$$

Chaos (I) is the state of disorder, which is very dependent on the initial condition. It increased the quadratic coupling of the items $y_i^2, x_i^2, x_i y_i$ and provided more security to the system. When $2.75 < \mu_1 < 3.4$, $2.7 < \mu_2 < 3.45$, $0.15 < \gamma_1 < 0.21$, and $0.13 < \gamma_2 < 0.15$, the system comes into chaotic state and can generate a chaotic sequence in the region (0, 1]. So all the parameters (μ_1, μ_2, γ_1 , and γ_2) are used as secret keys. Figure 2; figure 3 show the chaotic behavior of Chaos (I) system

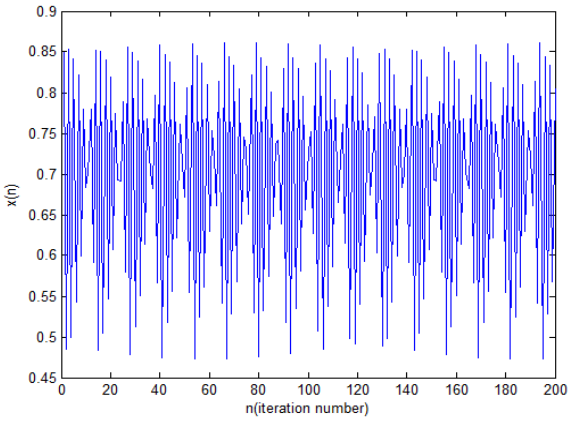

Figure 2: Plot of X component of 2D logistic map.

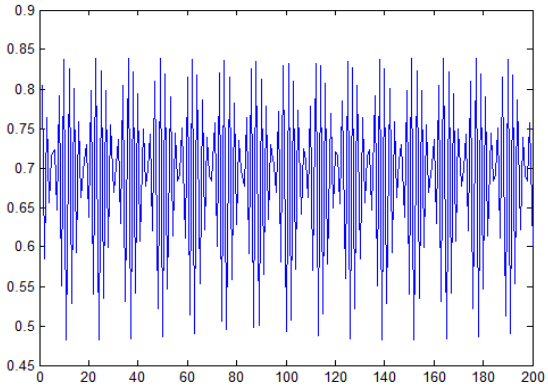

Figure 3: Plot of Y component of 2D logistic map.

2.1. Overview of the proposed approach

The idea of the proposed scheme is to have three DNA Sequences (S_1, S_2, and S_3). The first two sequences S_1 and S_2 are generated from two-dimensional chaotic map 2D and the third sequence S_3 is selected from DNA database. The proposed algorithm converts the original message M to binary form and carries out the hamming code resulting M´ from M. Then the first sequence S_1 is used for encryption M´ using XOR operation to produce the encrypted message M´´. The random sequence numbers S_2 is used to determine the positions in S_3 to hide M´´.

Figure 1 shows the diagram of the proposed method.

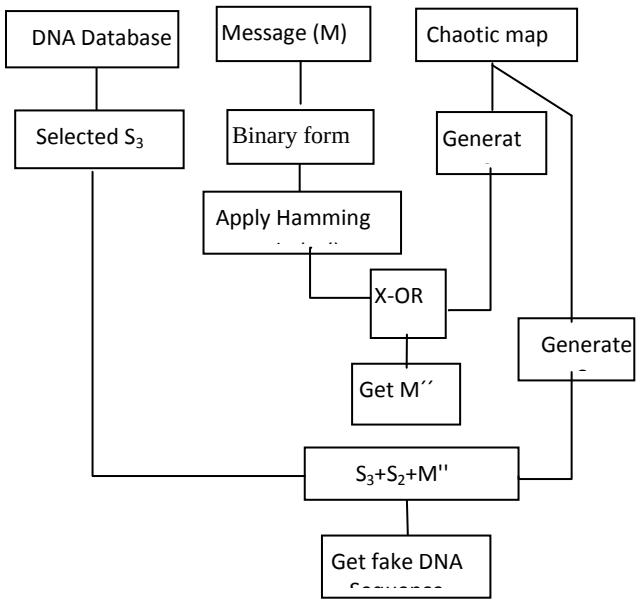

Figure 4: The diagram of the overall data hiding and encryption process

2.2. Generate two truly random sequences based on chaos.

Based on the chaotic system in formula (I), generate a set of completely random sequences x_i and y_i, then choose two prime numbers p and q to get the sequences.

$w_i = x_i \bmod p$
and
$z_i = y_i \bmod q$;
The integers w_i are converted into the binary form $Bi(w_i)$. The DNA (Num_format) transfers each two binary bits of $Bi(w_i)$ to the DNA nucleobase using the replacements in Table 1. Without loss of generality, we consider all the values of z_i are distinct then there will be only one position in the selected sequence S_3 to embed one or more letter from the secret message.

Table 1:DNA (Num_format) representation of bits with the nucleobase letters.

DNA nucleobase letters	Bit 0	Bit 1
A	0	0
T	1	1
C	0	1
G	1	0

2.3. Hamming Code

Hamming (M, R) is a linear error-correcting code that encodes R bits of data into M bits by adding R parity bits. For any positive whole $R \geq 3$, the hamming adds M-R additional check bits to every R data bits of the message.

- The length of the code N equals 2^R -1.
- The length of the original message M equals 2^R – R - 1.
- The number of the parity bits R equals N - M.

Algorithm 2-3-1 Hamming Step(M, |M|)

1. Generate the binary 01, 10, 11, 100...corresponding the counting numbers 1,2,3,4...
2. Compute the redundant bits using the inequality $2^R \geq |M|+R+1$
3. All the numbers1,2,4,8,..,| M | have only one bit 1 in their binary form 1,10,100,1000
4. All other numbers with two or more 1bits are message bits.
5. Check the even (odd) parity bits
6. Calculate the different collection of bits and nominate the bit to switch.

2.4. Encryption Algorithm

The sender uses Algorithm 2-4-1 to encrypt and hide the secret message.

Algorithm 2-4-1 Data Hiding Algorithm

Input: the 2D chaotic system, S_3 and the secret message M.

Output: A faked DNA sequence S_3 with secret message M hidden.

Step 1: Apply Hamming Step (M, |M|) on M and get M´.

Step 2: Use Table 1 to encode the secret message M to DNA sequence DM.

Step3: Generate the first artificial DNA sequences S_1.

Step4: Generate the second sequences S_2 of distinct integer numbers.

Step4: Randomly select the DNA Sequence S_3 from DNA database.

Step 5: Perform the bitwise XOR of the secret message M´ and the first sequence S1 to generate the encrypted message M´´

$$M'' = M' \text{ XOR } S_1.$$

Step 8: Use the integer numbers of S_2 to determine the positions in S_3 and replace M´´ characters to get fake DNA Sequence S'_3 as follow:

For i = 1 to the length length(M´´)
$S_3(h(i)) = M''(i)$;
end for

Step 9: Send Fake DNA Sequence S'_3 to the receiver.

Example

The following example illustrates the encryption and hiding processes in details:

1. The parameters values are:
μ1=2.77985671895587; μ2=2.71589456247256; x(1)=0.97489215436257; ɣ1=0.15796875214124; ɣ2=0.14897521465128; y(1)=0.85492315687125; q=29; L=12and p=1949

2. Generate the artificial DNA sequences
S_1:
TCTCTCAGACGATGCACGGTCTATGGACACGGAATC GATGCGCCAGATAACTATTAGAAT

3. Randomly select the realistic DNA sequence S_3:
S_3=ATCGAATTCGGGCTGAGTCACAATTCGCGCTGAG TGAACC

4. Let the original message (M) =1100111100110000, where |M| = 16

5. The redundant bits (R) where $2^R \geq |M| + R + 1 = 17+R$ then R=5

6. Apply Hamming Step (M, 16) to get the message M´=00111001111001110000, where |M '|=R+|M|=21, add 0 to the most left of M´=001110011111001110000, and the length of |M´| becomes 22.

7. Apply the binary XOR operation between M´ and S_1 and delete the extra unused nucleotide to get the following encrypted nucleotides
M´´=M´ XOR S_1 =1110010000100001100110 = TGCAAGACGCG.

8. Generate the distinct integer sequences S_2: 3, 4, 12, 23, 24, 27, 29, 33, 34, 35, 38

9. For i = 1 to | M" |
Swipe (M'' (i), S_3[S_2[i]])

S'_3=ATTGAATTCGGCCTGAGTCACAAATCGCACTGCG CGAGCC

10. Send S'_3 to receiver.

2.5.Decryption Algorithm

The data recovery procedure is similar to that of the encryption algorithm but in the reversed order. The sender sends the faked DNA sequence without any other DNA sequences, DNA-like sequences, or any sequences of numbers to the receiver. The receiver processes the Data Recovery Algorithm 2-5-1 to recover the original message.

Algorithm 2-5-1 Data Recovery Algorithm

Input: A faked DNA sequence S'_3 and the 2D chaotic system.
Output: The secret message M.

Step 1: Use 2D chaotic map to generate S_1 and S_2.
Step 2: Using S_2 to get M´´ from fake DNA sequence S'_3
For i = 1: to the length(M´´)
$M''(i) = S'_3(S_2(i))$
End for
Step 3: Perform the bitwise XOR of the secret message M" and the first sequence S_1 to generate the encrypted message M´
M´= M´´ X-OR S_1
Step 4: Calculate number of redundant bits (R) in M´ with the following inequality

$$2^R >= M+R+1$$

Step 6: Delete R bits from M´ to get the original message M.

3. SECURITY ANALYSIS

The key space should be large enough that make brute-force attacks infeasible. In our experiment, the initial conditions parameters$\mu 1 = 2.77985671895$, $\mu 2 = 2.71589456247, x1 = 0.97489215436, ɣ10.15796875$ $ɣ2 = 0.14897521465$ and $y(1) = 0.85492315687$ are the secret key. If the precision of each key is $10^{(-14)}$, the key size is bigger than 2^{128}. The total number of the different keys, which can be used in the encryption, is defined as the size of the key space of the

logistic map is 10^{6x14}.There are roughly 163 million DNA sequences available publicly.

Table 2. The experimental results of the proposed scheme

Locus	Specifies definition	No. of nucleotides	Capacity C	Payload	bpn= |M|/C [8]	bpn= |M|/C [10]	bpn= |M|/C the propose
AC153526	Mus musculus10 BAC RP23-383C2	200,117	200,117	0	0.434	0.577	0.579
AC166252	Mus musculus6 BAC RP23-100G10	149,884	149,884	0	0.442	0.580	0.730
AC167221	Mus musculus10 BAC RP23-3P24	204,841	204,841	0	0.424	0.563	0.566
AC168874	Bostaurus clone CH240-209N9	206,488	206,488	0	0.446	0.560	0.561
AC168897	Bostaurus clone CH240-190B15	200,203	200,203	0	0.451	0.565	0.579
AC168901	Bostaurus clone CH240-18511	191,456	191,456	0	0.439	0.583	0.605
AC168907	Bostaurus clone CH240-19517	194,226	194,226	0	0.444	0.580	0.597
AC168908	Bostaurus clone CH240-95K23	218,028	218,028	0	0.443	0.583	0.529

Therefore, the probability of the attacker making a successful guess for the third DNA sequence is$\frac{1}{1.63\times10^{8}}$. The binary coding rules are 4×3×2×1=24. The likelihood of making correct guess by the attacker is 1/24.The hamming code is used to transform a character of 8 bits into 12-bit code word, the number of bits increase by a factor of 1.5, and the number of combinations may be 2^{12}.The probability of guessing the secret message is :$(\frac{1}{1.63\times10^{8}})$ $(\frac{1}{24})$ x $(\frac{1}{2^{12}})$ x $(\frac{1}{10^{6x14}})$

4. EXPERIMENTAL RESULTS.

A series of experiments carried out to evaluate the performance of the proposed scheme. Table 2 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, 10]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 [0.56, 0.73], the proposed scheme has an acceptable embedding capacity, and the embedding capacity is stable with different reference DNA sequences

5. CONCLUSION

The current approach is more secure than the precious's cryptographic approaches. We used three sequences in the proposed scheme.2D chaotic map generated two DNA sequences, the first sequence encrypted the plain message and the second sequence determined the position of hiding in third realistic DNA sequence which is publicly available by accessing the NCBI. The security analysis shown that, for all practical purposes, it is impossible for an intruder to recover the secret message.

6. ACKNOWLEDGEMENTS

The authors express their deep sense of gratitude to Prof. Dr. Maher Sh. Zayed for providing valuable guidance, encouragement, support and advising technical points from time to time during the preparation of this research paper.

7. . REFERENCES

[1] A Leier, C Richter, W Banzhaf, H Rauhe, Cryptography with DNA binary strands.in BioSystems (2000*)*, pp. 13-22.

[2] H Shiu, K Ng, J Fang, R Lee, C Huang, Data hiding methods based upon DNA sequences. in *Information Sciences(2010),* pp. 2196-2208.

[3] B Shimanovsky, J Feng, M Potkonjak, Hiding data in DNA. in Information Hiding(2003), pp. 373-386.

[4] A Atito, A Khalifa, S Rida, Dna-based data encryption and hiding using playfair and insertion techniques. in Journal of Communications and Computer Engineering(2011), p. 44- 49.

[5] N Kar, A Majumder, A. Saha, A Jamatia, K Chakma, M Pal, An improved data security using DNA sequencing. in Proceedings of the 3rd ACM MobiHoc workshop on Pervasive wireless healthcare(2013), pp. 13-18.

[6] B Roy , A Majumder, An Improved Concept of Cryptography Based on DNA Sequencing.(IJECCE,*2012)* , pp. 1264-1267.

[7] BA Mitras , AK Aboo, Proposed Steganography

Approach Using DNA Properties. in InternationalJournal of Information Technology BusinessManagement (2012), pp. 96-102.
[8]C Guo, C Chang , Z Wang, A new data hiding scheme based on NA sequence. Int J Innov Comput Inf Control(2012),p. 1-11.

[9] FE Ibrahim, MI Moussa , HM Abdalkader ,A Symmetric Encryption Algorithm based on DNA Computing, in InternationalJournalofComputerApplications(2014), pp. 41-45.

[10] Fatma E. Ibrahim, M. I. Moussa, 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.