A Data Concealing Method Characterized by DNA coding

Tina Babu
Federal Institute of Science And Technology

Nimisha Abraham
Federal Institute of Science And Technology

Mehbooba P Shareef
Federal Institute of Science And Technology

Reshma KV
Federal Institute of Science And Technology

ABSTRACT

Now-a-days many erogenous personal data may be sent through internet that may be tapped. Different techniques have been used to ensure that secret messages would reach their intended recipient without being tampered with along the way. **A Data Concealing Method Characterized by DNA coding** proposes a data hiding method build on DNA coding using word document or PDF as transporter. Firstly, we encode the plaintext by DNA coding to a DNA sequence. Secondly, we encrypt the DNA sequence by another equal-length DNA sequence generated by Chebyshev maps, and attach the result to a primer DNA sequence. After circularly shifting the whole sequence for finite times, we embed them into the Word document by modifying the forecolour of the characters. Each character in the Word document can be embedded one character, i.e. 6-bit DNA coding. The initial values and parameter of the Chebyshev maps and shifting times are all served as keys, and the plaintext can be extracted successfully from the host document. This technique ensures the plaintext to be embedded into the characters, and can also skip the non-character contents in the Word document, such as image and object, the same to the extracting process. Thus the method is well suited to some data hiding applications, such as fragile watermarking, secret communication, and online content distribution systems. This paper is a modification of a paper titled 'A novel data hiding method based on deoxyribonucleic acid coding' [1] aimed at increasing the security of the data hiding method proposed in it.

Keywords

DNA sequences, DNA coding, Chebyshev maps, one time cryptography, security, encryption, DNA cryptography.

1. INTRODUCTION

Various communications over the Internet such as electronic mail, or the use of world wide web browsers are not secure for sending and receiving sensitive information as they may be tampered. The users would like to have a secure, private communication with the other party. Online users may need private and secure communications for other reasons as well. They may simply not want third parties to browse and read their e-mails or alter their content. Information Technology is the most essential aspect in today's world. Based on this fact computer application is still developing to handle securely the financial as well as the personal data more effectively. These data are extremely important from every aspect and we need to secure this from unauthorized access.

The technique of preventing and detecting unauthorized use of data or computer or network is termed as security. To stop unauthorized users from accessing any part of computer system, prevention measures help us. Detection helps to determine whether or not someone attempted to break into the system, if they were successful, and what they may have done. We may use various cryptography and steganography techniques to achieve that security. However, today data encryption is not everything or cannot achieve strong security through this and also need to secure the presence of data. Thus comes the demand of Steganography.

Steganography is the art and science of communicating in a way which hides the existence of the communication. The messages are hidden inside other harmless messages such that it does not allow any enemy to even detect that there is a second message present. In a digital world, Steganography and cryptography are both intended to protect information from unwanted parties. Both Steganography and Cryptography are excellent means by which to accomplish this but neither technology alone is perfect and both can be broken. It is for this reason that most experts would suggest using both to add multiple layers of security.

In this paper, a new data concealing method by embedding secret message into Word document is proposed. Firstly, DNA sequence is generated by encoding the plaintext by DNA coding. Then, we encrypt the DNA sequence by another equal-length DNA sequence generated by Chebyshev maps, and attach the result to a primer DNA sequence. We embed them into the Word document by modifying the fore color of the characters, after circularly shifting the whole sequence for finite times,. Each character in the Word document can be embedded one character, i.e. 6-bit DNA coding. The keys for the encryption are taken by the initial values and parameter of the Chebyshev maps and shifting times. The plaintext can be extracted successfully from the host document by the reverse process.

2. RELATED WORK

The security of data hiding techniques has attracted increasing research attention as people are becoming aware of the threat to their own privacy that they feel the need to use (more commonly) encryption, and (more recently) steganography to overcome this threat.

PDF file can be used as carrier of cipher text so PDF file is widely used to transmit information as the content of PDF file cannot be modified directly. Zhong et al.2007 presented Data hiding in a kind of PDF texts for secret communication [2]. It is a novel steganographic technique for hiding data in a kind of PDF texts. First step is to point out the secret channels in a kind of PDF English texts, which are generated from documents that make the texts justified to occupy the full line width and position each character individually. In succession, the steganographic system PDFStego is described in which several strategies are applied to improve security, such as making use of redundancy to complement security; constituting two chaotic maps to meet the Kerckhos principle and to prevent statistical attacks, and applying the secure hash algorithm to enable integrity service and blindly extracting service. PDFStego can be used to exchange sensitive data securely or to add copyright information to the PDF files.

Lee et al., 2010 designed a new approach to covert communication via PDF files [3] by embedding secret messages in PDF files is proposed. A message is regarded as a string of bits or characters, and encoded with a special ASCIIcode by binary or unitary coding. The encoding results are then embedded at the between-word or between-character locations in the text part of a cover PDF file. The embedding results in the resulting stego-PDF document are found to be invisible in this study in the windows of common PDF readers, creating a steganographic effect and achieving the purpose of secret communication.

In recent years chaos has attracted much attention. Thus Franois M, Grosges T. proposed Image Encryption Algorithm Based on a Chaotic Iterative Process based on a coupling of chaotic function and xor operator was presented [4]. The main advantages of such a method are the abilities to produce a large key space to resist brute-force attacks, and to encrypt securely images with any entropy structure assuring indistinguishability, confusion and diffusion properties in the corresponding cipher images.

Kanso A, Ghebleh M. designed A novel image encryption algorithm based on a 3D chaotic map [5]. The design of the proposed algorithm is simple and efficient, and based on three phases which provide the necessary properties for a secure image encryption algorithm including the confusion and diffusion properties. In phase I, the image pixels are shuffled according to a search rule based on the 3D chaotic map. In phases II and III, 3D chaotic maps are used to scramble shuffled pixels through mixing and masking rules, respectively. Simulation results show that the suggested algorithm satisfies the required performance tests such as high level security, large key space and acceptable encryption speed. These characteristics make it a suitable candidate for use in cryptographic applications.

Seyedzadeh SM, Mirzakuchaki S. presented A fast color image encryption algorithm based on coupled two-dimensional piecewise chaotic map [6]. This paper proposes a novel chaos-based image encryption algorithm to encrypt color images by using a Coupled Two-dimensional Piecewise Nonlinear Chaotic Map, called CTPNCM, and a masking process. Distinct characteristics of the algorithm are high security, high sensitivity, and high speed that can be applied in encryption of color images. In order to generate the initial conditions and parameters of the CTPNCM, 256-bit long external secret key is used. Computer simulations confirm that the new algorithm has high security and is very fast for practical image encryption.

Massive parallelism, huge storage and ultra-low power consumption properties of DNA computing had lead to many researches. Gehani A, LaBean TH, Reif JH proposed DNA-based cryptography where some procedures for DNA cryptography based on one time pads that are in principle unbreakable [7]. DNA provides a much more compact storage media and an extremely small amount of DNA suffices for huge one-time pads.

Zhang Q, Guo L, Wei XP. designed an Image encryption using DNA addition combining with chaotic maps [8]. First, a DNA sequence matrix is obtained by encoding the original image, then, divide the DNA sequence matrix into some equal blocks and use the DNA sequence addition operation to add these blocks. Next, perform the DNA sequence complement operation to the result of the added matrix by using two Logistic maps. Finally, decode the DNA sequence matrix from the third step, and we can get the encrypted image. The simulation experimental results and security analysis show that this scheme not only can achieve good encryption, but can also resist exhaustive attack, statistical attack and differential attack.

In most of the previous works, ASCII coding of the data is done which is 8 bit coding and has got less hiding capacity as they are embedding data in between the words and characters in the PDF file.

3. DNA CODING

Instead of the traditional binary encoding methods that are taking on for DNA coding here a coding technique that will grant a coding length up to 6 bit which is much shorter than the classical coding techniques is made use of. Complementary relationship between the sequences of bases on the two intertwined chains gives DNA its self-encoding character.

The bases in DNA fall into two classes, mainly Purines and Pyrimidines. The Purines are adenine(A) and guanine(G), and the Pyrimidines are cytosine(C) and thymine(T)[9].The paring is done mainly between adenine and thymine and between guanine and cytosine. Complimentary pairs of binary system 0 and 1 and 0 and 3,1 and 2 form the complimentary pairs. In broad 24(4!) kinds of coding is possible and eight of them can meet the complimentary rule.0123 can be articulated as CTAG, CATG, GTAC, GATC, TCGA, TGCA, ACGT and AGCT. Decide on one of them at random to encode the plaintext.

Unlike the traditional binary encoding for DNA, here it make draw on of the base-4 numeral system ie, quaternary code, in which each character can be denoted by three nucleotides say eg: S=ACG ,O=TTA as shown in the table 1 ; translation table from alphabets to DNA bases proposed by Clell and Risca[10] and DNA coding can be expressed only for capital letters, digits and for punctuation marks . if 0 , 1 , 2 ,3 are used to represent the nucleotides C,A,T and G respectively, we can easily encode the plaintext message M to MDNA ,for eg: “DNA” the plain text can be expressed as TTGTCTCCA(101011100010000001) and by decoding we will acquire the resultant plaintext.

The translation table, the mapping and the base combination used to denote each number can be redefined as per the interest of the parties involved in data transfer and this too contributes to the complexity of the method.

Table 1. Translation table for the alphabets ,digits, and punctuations marks to DNA bases

A=CGA	H=CGC	O=GGC	V=CCT	2=TAG	9=GCG
B=CCA	I=ATG	P=GGA	W=CCG	3=GCA	=ATA
C=GTT	J=AGT	Q=AAC	X=CTA	4=GAG	,=TCG
D=TTG	K=AAG	R=TCA	Y=AAA	5=AGA	.=GAT
E=GGT	L=TGC	S=ACG	Z=AAT	6=GGG	:GCT
F=ACT	M=TCC	T=TTC	0=TTA	7=ACA	;=ATT
G=TTT	N=TCT	U=CTG	1=ACC	8=AGG	_=ATC

4. CHEBYSHEV MAPS FOR DNA ENCODING

4.1. Chebyshev maps

The most important property of a cryptographic system is chaotic nature. It ensures that cryptanalysis will be nearly impossible or really difficult. Chebyshev maps are efficient at providing the needed chaotic nature. The expression of chebyshev map is as follows,

$$z_{i+1} = \cos\left(w \cos^{-1} z_i\right), \qquad -1 \le z_i \le 1 \qquad (1)$$

Where w is the degree of the Chebyshev map. If $w \in [2,6]$, the Lyapunov exponent of Chebyshev map will be positive and it indicates that Chebyshev maps are chaotic The sequences generated by Chebyshev maps are real numbers and are orthogonal polynomials. Their correlation functions are δ functions too.

4.2 Generating pseudo-random sequences using Chebyshev maps

For enhancing security, the cipher text obtained has to be encrypted and concealed. For this real, orthogonal polynomial sequences generated by Chebyshev maps can be used. One time keys are the back bone of most secure encryption algorithm. For this random numbers are generated using mouse movement and chaotic cryptographic methods as

$$S_{keys} = \{(w_1, z_1), (w_2, z_2), \ldots, (w_n, z_n)\} \qquad (2)$$

To ensure chaotic nature each time two different pairs (w_i, z_i) and (w_j, z_j) are taken randomly from S_{keys} and w_i from one group act as the initial degree and z_j from the other group act as initial parameter for Chebyshev map. Now two sequences x_{XOR} and y_{PRIMER} are generated using equation 1. x_{XOR} is defined as x_{XOR} = { $x_1, x_2, \ldots, x_{LM}$ } and is used for encryption of DNA coded plain text. x_{XOR} is obtained after iterating equation 1 'm' times and is defined as

$$x_i = \begin{cases} 00, & (z_i > -1) \ and \ (z_i \le -.05), \\ 01, & (z_i > -.05) \ and \ (z_i \le 0), \\ 10, & (z_i > 0) \ and \ (z_i \le 0.5), \\ 11, & (z_i > .5) \ and \ (z_i \le 1) \end{cases}$$

Whereas y_{PRIMER} is defined as y_{PRIMER} = { $y_1, y_2, \ldots, y_{LW-LM}$ } and is used as the primer of the DNA sequence . It is obtained after iterating equation 1 'n' times and is defined as

$$y_i = \begin{cases} 00, & (z_i > 0.5) \ and \ (z_i \le 1), \\ 01, & (z_i > 0) \ and \ (z_i \le 0.5), \\ 10, & (z_i > -0.5) \ and \ (z_i \le 0), \\ 11, & (z_i > -1) \ and \ (z_i \le -0.5) \end{cases}$$

5. MESSAGE HIDING AND RECOVERY ALGORITHM

5.1 Encrypt the DNA coding sequence

With the growth in DNA computing, Researchers propose some biological as well as algebraic operations based on DNA sequence such as XOR, addition, subtraction operation etc. Here 00, 01, 10, 11 is used to denote C, A, T, G respectively.

Rules of addition and subtraction operation:-

Table 2 : Addition operation for DNA sequences

+	T	A	C	G
T	C	G	T	A
A	G	C	A	T
C	T	A	C	G
G	A	T	G	C

To perform the encryption and decryption of DNA sequence using addition and subtraction operation respectively.

Let M_{DNA} is the DNA sequence. Then

$$M_{DNA} = M'_{DNA} + X_{XOR}$$
$$M'_{DNA} = M_{DNA} - X_{XOR}$$

Table 3: Subtraction operation for DNA sequences

-	T	A	C	G
T	C	G	T	A
A	A	C	G	T
C	T	A	C	G
G	G	T	A	C

5.2 Data hiding algorithm

If we hide the cipher text sequence M'_{DNA} in the Word document by selecting the start location randomly, the start location can be easily detected by the color analysis, so we take two measures to modify the fore color of all the characters.

Firstly, attach M'_{DNA} to a primer sequence Y_{PRIMER} to get sequence M'Y, for that set the total length of M'Y equal to the number of characters in the Word document, so the fore color of all the characters will be substituted. Secondly, circularly shift M'Y to the right or left for sn ϵ [1,2000] times, then embed the whole sequence into the Word document through substituting each characters forecolor. Depending on the position in the pixel, a bit can contain different amount of information. For example, “1"at the 8th bit of a pixel represents 128 (2^7), but it only represents 1 (2^0) at the first bit. According to Equation

$$p(i) = \frac{2^{i-1}}{255}, \quad i = \{1,2,3...,8\}$$

the higher four bits (8th, 7th, 6th and 5th) carry 94.125% of the total information, and on the other hand, the lower two bits (2nd and 1st) carry only 1.1765% of the total information, so it is usually used to hide information. The percentage of the pixel information is shown in Table.

Table 4 : Percentage of pixel information contributed by different bits

Bit position *i* in the	Percentage *p(i)* of the pixel
1	0.3922
2	0.7843
3	1.5686
4	3.1373
5	6.275
6	12.55
7	25.10
8	50.20

The algorithm is designed to ensure the plaintext to be embedded into the characters, and can also skip the non-character contents in the Word document, such as image and object, the same to the extracting process. The embedding process for one character in the plaintext is shown in Figure 5.1. The changes in the fore color can still be remained and cannot be directly modified, even if the Word document is converted to its corresponding PDF, so the PDF file can also be served as the host file. For the sender, the data encryption and hiding algorithm can be divided into the following steps,

Step 1: The plaintext M has got the length L_M, and the number of characters in the Word document W is L_W.

Step 2: Shift the translation code ‘sn’ times to get new translation code.

Step 3: Encode M to the DNA sequence M_{DNA} by DNA coding using new translation code.

Step 4: By the Chebyshev maps with one-time keys from sequence S_{KEYS} two pseudo-random DNA sequences of X_{XOR} and Y_{PRIMER} are generated

Step 5: Set $M'_{DNA} = M_{DNA} + X_{XOR}$.

Step 6: The sequence M’Y is generated by attaching M'_{DNA} to the right of Y_{PRIMER} , then circularly shift it to the left or right for sn times to get the DNA sequence E, the shift times sn ϵ [1,2000].

Step 7: By substituting each character’s forecolor, embed E into word document.

Step 8: Send the Word document to the receiver directly, or convert it to a PDF file firstly.

5.3. Data extracting and recovery algorithm

The Word document or PDF file is sent to the receiver from sender. If the receiver receives the Word document, he can extract the plain message from the Word document directly, if he receives the PDF file, he can also convert the PDF file to Word document firstly, then extract the plain message. The keys can be transmitted from the sender to the receiver through a secure channel. According to the same keys, the receiver can extract the plaintext. The steps are shown as follows.

Step 1: For the Word document, go to Step 2. For the PDF file, copy the contents to a new Word document firstly.

Step 2: Extract the DNA sequence E from the least 2-bit of the three color components of each character.

Step 3: Circularly shift E to the reverse direction for ‘sn’ times to get the sequence E’. According to the length L_M of M extract the sequence M'_{DNA}.

Step 4: Generate the pseudo-random chaotic sequence X_{XOR} by the Chebyshev maps with the same keys as the hiding process.

Step 5: Set $M_{DNA} = M'_{DNA} + X_{XOR}$.

Step 6: Shift the triplets in M_{DNA} in reverse direction ‘sn’ times.

Step 7: Decode M_{DNA} by DNA coding to get the plaintext M.

One of the most important advantages of using DNA coding is that frequency attacks will be almost impossible as every letter is coded in terms of the four DNA bases. It helps to increase confusion and diffusion too. A single change to the DNA code of a letter results in lot of change all over the document.

The translation code used for each letter can be shifted sn times before the characters are encoded in DNA sequence for increasing the complexity of the algorithm even more. In this way, even if the translation table is tapped by the attacker, it is ensured that he cannot successfully decode it as the translation code is shifted again 'sn' times as number 'sn' is unknown to attackers. All these factors contribute in making this cryptographic method unbreakable and resistant to many known attacks.

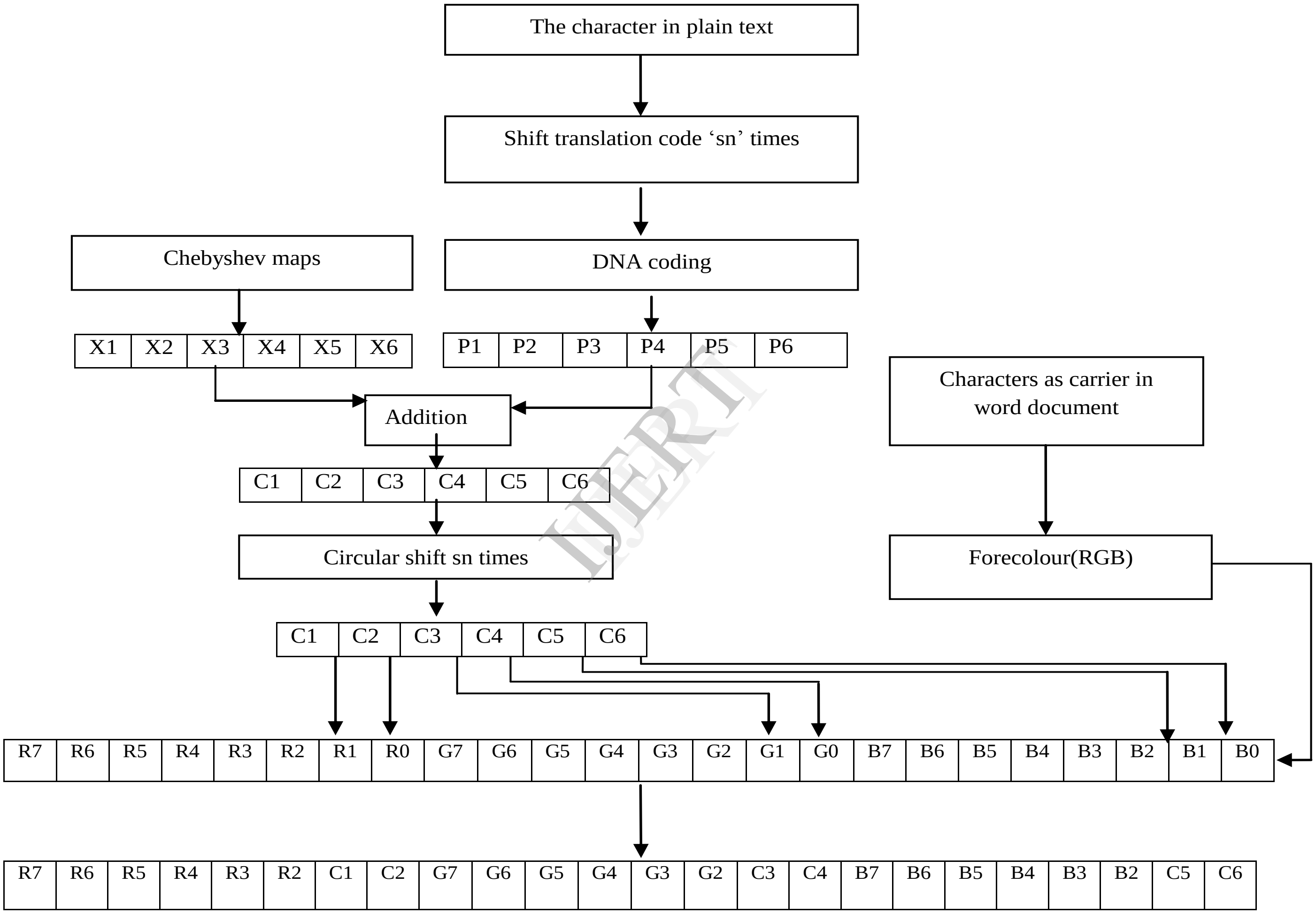
The character in plain text
Shift translation code 'sn' times
Chebyshev maps
DNA coding
X1 X2 X3 X4 X5 X6
P1 P2 P3 P4 P5 P6
Characters as carrier in word document
Addition
C1 C2 C3 C4 C5 C6
Circular shift sn times
Forecolour(RGB)
C1 C2 C3 C4 C5 C6
R7 R6 R5 R4 R3 R2 R1 R0 G7 G6 G5 G4 G3 G2 G1 G0 B7 B6 B5 B4 B3 B2 B1 B0
R7 R6 R5 R4 R3 R2 C1 C2 G7 G6 G5 G4 G3 G2 C3 C4 B7 B6 B5 B4 B3 B2 C5 C6

Figure 5.1: The embedding process for one character in plain text

6. PERFORMANCE ANALYSIS

6.1 Variation of host file size

Both the Word document and the PDF file can all be served as the host file. Their growth ratios slowly increase with the number of hidden characters, the variation tendency is nonlinear.

6.2. Hiding Capacity

Each character in the Word document can be hidden in a 6-bit binary DNA coding. So the embedding capacity of this method is 100%. If more than least significant 2-bit are substituted, the visual quality will be decreased with the increase of hiding capacity.

6.3 The key space

We assume that the algorithm is known to the public, so the solution to security is the keys. This algorithm actually does have some of the following keys:

1. The total combinations Sc of translation from alphabets to DNA nucleotides. 2
2. Initial condition (wi, zi) for Chebyshev maps.
3. Iteration times m for Chebyshev maps.
4. The shift direction, and the shift times sn,
5. Length of plaintext LM.

In the receiver, the group (wj, zj) and iteration times n are useless, so they are not served as keys. There are total $S_c = P_{64}^{42} \approx 5.2 \times 10^{70}$ kinds of combination of translation from alphabets to DNA nucleotides. Only the group (wi, zi) serves as the key, the variation of the parameter w in the chaotic region is between 2 and 6 with a step of 10^{-7}, so $S_{wi} = 4 \times 10^{-7}$.For any chaotic system is sensitive depending on the initial conditions, even the initial value z_i is changed with a tiny step of 10^{-16}, their orbits of the Chebyshev maps will be completely different, so we set the key space for initial value zi to $s_{zi} = 10^{-16}$. We set the iteration times m ∈ [100,1000], so $S_m = 9 \times 10^{52}$, the shift direction is left or right, so d = 2. We set the shift times sn ∈ [1,2000], so $S_{sn} = 2 \times 10^3$. The length of plaintext LM ∈ [1,1000], then $S_{LM} \approx 10^3$. The key space is $S = dS_c S_{w0} S_{z0} S_m S_{sn} S_{LM} S_{sn} \quad \approx 1 \cdot 497 \times 10^{111}$, which is much larger than 2^{100}. In order for the brute force attacks to be ineffective Alvarez and Li suggested that the key space should be at least $2^{100}$and hence this data hiding method based on Chebyshev maps and DNA encoding is quite efficient.

6.4. Possible attacks

If the Word document is served as the carrier of secret message, the active attacker can only get the DNA coding sequence E, but cannot decode it. For the active attacker, we can compute the message digest 5 (MD5) hash value of the Word document or the PDF file and send it to the receiver by secret channel, if these files are modified by attackers, their MD5 hash values will be different completely. For the passive attacker, we have the large key space to resist brute force attacks. Even though he gets the whole ciphered DNA sequence, it is unable to decide where the starting location of the secret message is, and how to decrypt it.

7. CONCLUSION

A novel data hiding method based on DNA coding and using Word document as host file where the cipher sequence is hidden into a Word document by substituting the least significant 2-bit of the three colour components. The secret message can be extracted successfully if it is a Word Document or PDF file. To enhance the robustness three measures were taken -
i. To shorten the cipher text, the plaintext was encoded using DNA coding instead of using 8-bit ASCII coding.
ii. To encrypt and conceal the cipher sequence, two random aided DNA sequences were generated.
iii. By substituting the least significant 2-bit of the three colour components the cipher sequence are hidden into a Word document.
Main issues are
i.Skips the non-character contents in the Word document, such as image and object.
ii. Attacks are possible
The proposed method is well suited to some data hiding applications, such as fragile watermarking, secret communication, and online content distribution systems.

REFERENCES

[1] Hongjun Liu, Da Lin, Abdurahman Kadir, “A novel data hiding method based on deoxyribonucleic acid coding”

[2] Shangping Zhong, , Xueqi Cheng, Tierui Chen “Data Hiding in a Kind of PDF Texts for Secret Communication” International Journal of Network Security, Vol.4, No.1, PP.17–26, Jan. 2007

[3] Lee IS, Tsai WH. “A new approach to covert communication via PDF files” Journal
Signal Processing Volume 90 Issue 2, February, 2010 Pages 557-565

[4] M. François, T. Grosges, D. Barchiesi and R. Erra, "Image Encryption Algorithm Based on a Chaotic Iterative Process," Applied Mathematics, Vol. 3 No. 12, 2012, pp. 1910-1920

[5] A. Kanso, M. Ghebleh. “A novel image encryption algorithm based on a 3D chaotic map”. Communications in Nonlinear Science and Numerical Simulation 2012 17:7, 2943-2959.

[6] Seyed Mohammad Seyedzadeh, and Sattar Mirzakuchaki. “A fast color image encryption algorithm based on coupled two-dimensional piecewise chaotic map”.Signal Processing 92(5):1202-1215 (2012)

[7] Gehani A, LaBean TH, Reif JH “DNA-based cryptography” Aspects of Molecular Computing Lecture Notes in Computer Science Volume 2950, 2004, pp 167-188

[8] L. Zhang, X. Liao, and X. Wang, "An image encryption approach based on chaotic maps," Chaos, Solitons & Fractals, vol. 24, pp. 759-765, 2005.

[9]. N.Lewis,P Weinberger DNA Computing october 1995 1st

[10] Clelland CT, Risca V, Bancroft C. Hiding messages in DNA microdots. Nature 1999;399(6736):533–4