A Highly Secure DNA Based Image Steganography

Prasenjit Das
Dept. of Computer Science & Engineering
National Institute of Technology Agartala, India
pj.cstech@gmail.com

Nirmalya Kar
Dept. of Computer Science & Engineering
National Institute of Technology Agartala, India
nirmalya@nita.ac.in

***Abstract*—DNA steganography is one of the emerging technologies in the field of covert data transmission. In this paper we are proposing a novel DNA based steganography which uses images as primary cover media. A theoretical single stranded DNA (ssDNA) or oligonucleotide is extracted from the image which is used as the secondary cover providing a huge amount of background noise for the secret data. This is why we call it a dual cover steganography. The pixel sequence contributing the ssDNA codon series construction is determined by a two dimensional chaotic map. Performance of the algorithm is tested against several visual and statistical attacks and parameterized in terms of both security and capacity.**

***Index Terms*—DNA, image, steganography, logistic map, primer, DNA algebra, LSB, histogram, PSNR, neighbourhood.**

I. Introduction

When it comes to secured and convenient data transmission over internet, steganography is one of leading technologies being used around the globe for long time. With inclusion of newer cover media and stronger algorithms we can deal with the latest attacks. DNA steganography is one such name. Adleman [1], Clelland [2], Leier [3] are some of the leading researchers in this field working with organic DNA. Some more relevant work has been proposed in [4] - [7]. All the works done on biological DNA till date are very worthwhile, yet they have some drawbacks like the biological errors(e.g. mutation) and difficulty of implementation [8]. One solution to this problem is to use theoretical model of DNAs and utilize its natural properties to strengthen the existing hiding techniques. Some of the worth mentioning works are discussed next.

A combination of cryptography and DNA steganography is utilized in [8], where steganography is used for hidden symmetric encryption key distribution on every new communication. Hayam Mousa *et al.* devised a reversible information hiding scheme for DNA sequence based on reversible contrast mapping [9]. DNA encoding and watermarking is used to embed an offline handwritten signature in the form of a watermark in [10]. The watermarked image is embedded using the principles of spread spectrum watermarking and further encrypted as DNA sequences. In [11] an image (Secret image) is hidden with another image (Cover image) by creating 256 combinations of DNA bases using 4 nucleotides and replacing them with the original pixel values. Suman *et al.* in [12] proposed a double cover DNA based steganography using magic numbers as the forward tracking algorithm. [13] propose a lossless staganography method in which DNA sequence is used to represent secret image.

Amal *et al.* in [14] illustrate a DNA-based steganography method combined with a DNA cryptography technique for secure exchange of data in DNA carriers. In [15] secret message is hidden inside a reference DNA strand collected from a publicly available DNA database and the indices (locations) of message is sent to the receiver.

II. DNA Preliminaries

1) DNA: Chemically, DNA is formed by two backbone strands helicoidally twisted around each other, and mutually attached by means of two nitrogenous base sequences. The four bases are adenine (A), cytosine (C), thymine (T), and guanine(G). Artificially synthesized single-stranded DNA chains are named oligonucleotides.

2) DNA encoding: A binary coding scheme by which we can represent the 4 nucleotides by 2bit/3bit equivalent codes. The umber of possible coding patterns is 4! = 24. One such is- A = 0(00), T = 1(01), C = 2(10), G = 3(11)[9].

3) Codon: Triplets of consecutive bases in a base sequence are called codons. There are $4^3 = 64$ possible codons. Each codon encode for one of the 20 amino acids, used in the proteins synthesis, except TAA, TAG and TGA, indicate codon STOP [17].

4) Degenerative codons: Degenerative codons are those which differ in their third base yet code for the same amino acid. By replacing third base, we can hide data without effecting the resulting amino acid [17], which is known as silent mutation.

5) Addition & subtraction algebraic operation on DNA: Addition and subtraction operation can be performed on DNA sequences according to traditional addition and subtraction in the Z_2 (mod 2) [16]. For example, 3 + 2 = 1, so if (C,A,T,G) = (0,1,2,3), we have G + T = A. Table I & Table II shows addition and subtraction matrices.

TABLE I: Addition Operation for DNA Sequence

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

III. Proposed Algorithm

The proposed dual cover steganography process consists of two algorithms - embedding algorithm and extraction algorithm.

TABLE II: Subtraction Operation for DNA Sequence

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

A. Embedding algorithm

The entire process flows as follows- after verifying the cover and secret media, we perform a capacity check of the cover. Using the user defined logistic map parameters as key, we get a pixel sequence to construct a DNA sequence. By comparing the identified degenerative codon count with the half message length we determine the capacity of the cover. At the end we embed the message bits by replacing third bases of degenerative codons, perform DNA algebric addition with a key primer and hide the DNA back to the cover. The secret data bits are encrypted prior to embedding into the DNA, thus providing another layer of security.

a) Keys used: The entire algorithm uses 3 keys - Key 1(K_{E1}) consists of 6 parameters of 2D logistic map which are $x_0, y_0, \mu_1, \mu_2, \gamma_1$ and γ_2. Key 2 (K_{E2}) is a primer (a short DNA sequence), which is added to the mutated ssDNA after message embedding. Key 3 (K_{E3}) is a variable length key for RC4 encryption.

This embedding procedure has the following sub procedures.

1) Constraints for cover image & secret message: Some constrains are imposed on both the cover and secret media for getting robust output.

a) Constrains on Cover image: Due to the spatial domain operations involved the cover image should be inconspicuous, thus not drawing attention to the fact that it could be concealing information. Neither it should not contain large blocks of one color or smooth homogeneous areas, nor be a 'lossy' file format.

b) Constrains on secret message: The cover contains both data and meta-data. So we can say, Effective message length (L_E)

$$L_E = 4 \times L + metadata\_length \tag{1}$$

Here, L = number of characters in text message. Considering capacity of 2 bpp & 1 character converted to an equivalent ASCII value of 8 bits, we have,

$$L_E < Wid_C \times Hgt_C \tag{2}$$

Here, Wid_C = Cover image width. Hgt_C = Cover image height.

Let N be the number of iterations for 2D map, required to hide all the message bits, then we have.

$$N \geq 4 \times L + 3 \tag{3}$$

2) Perform capacity check with respect to effective message length (L_E): In this step the ssDNA is constructed from the cover image pixels, degenerative codons are identified and their count is compared with the effective message length (L_E).

The pixel used to construct DNA codon is chosen based on the sequence generated by the 2D logistic map. The 2D Logistic map [18] is given by the Eq. (4):

$$\begin{cases} x_{i+1} = \mu_1 x_i(1 - x_i) + \gamma_1 {y_i}^2 \\ y_{i+1} = \mu_2 y_i(1 - y_i) + \gamma_2({x_i}^2 + x_i y_i) \end{cases} \tag{4}$$

When $2.75 < \mu_1 \leq 3.4$, $2.75 < \mu_2 \leq 3.45$, $0.15 < \gamma_1 \leq 0.21$, $0.13 < \gamma_2 \leq 0.15$; the system generates two chaotic sequences in the region (0, 1]. The scaled values of the values gives the pixel locations in the cover image. To make the generated chaotic sequence more random, the map is preprocessed using the following [18],

$$X_i = 10^k X_i - floor(10^k X_i),\ where\ X_i\ =\ x_i\ or\ y_i \tag{5}$$

The chaotic sequence S' =$\{(x_i, y_i) \mid i = 1, 2, ...\}$ generated by (4) & (5) sometimes cross the positive bounds of 1. Hence it is further transformed using the following.

$$X_i = X_i\ mod\ 1, \text{if}\ X_i > 1 \tag{6}$$

The DNA extraction from cover image pixels and thereafter capacity check operations consists of the following steps.

i. Input the 6 variable parameters of 2D logistic map namely $x_0, y_0, \mu_1, \mu_2, \gamma_1$ and γ_2.
ii. Calculate x_{i+1}, and y_{i+1} from the 2D logistic map and scale them up to indicate corresponding pixel coordinate in the image.
iii. Take the pixel value from the cover image at scaled position (x_{i+1}, y_{i+1}). Let after the i^{th} iteration $x_{i+1} = 0.6439, y_{i+1} = 0.3758$ and cover image resolution is 600×800, then the chosen pixel is at $(\lfloor(0.6439 \times 600)\rfloor, \lfloor(0.3758 \times 800)\rfloor) = (\lfloor(386.34)\rfloor, \lfloor(300.64)\rfloor) = (386, 300)$.
iv. Extract the 2 LSBs from each color component i.e. red, green and blue.
v. The 2 least significant bits of each color component are converted to a corresponding nucleotide base, using DNA encoding technology; T = 0, A = 1, C = 2, G = 3.
vi. The 2 LSBs from red color component is used to construct the 1st base of a codon. The 2 LSBs from green color channel is used to construct the 2nd base and the 3rd base of the codon is constructed from the 2 LSBs from blue color component. The reason for choosing blue color as the 3rd base of a codon is that, in case of degeneracy the 3rd base of the codon is replaced. So if the blue color value is changed in the 0-255 range then it affects the luminance of the pixel with minimum distortion. Let, the pixel at (386, 300) gives us 10, 00 and 11 as LSBs from red, green and blue color. By DNA encoding we extract the codon as CTG.

vii. If the generated codon has degeneracy property then, we increment the counter for capacity by 1. Otherwise we move on to creating the next codon without affecting the current one. The codons which show degeneracy, their corresponding indices in the DNA strand is stored in memory for faster replacement process in the message embedding process.
viii. We perform the steps from (i) to (vii) unless capacity counter reaches $4 \times L + 3$ or all possible pixel positions generated of chaos map are overlapped. For the later, if the difference between calculated capacity and message length is marginal then we can change the 2D logistic map parameters or the cover image and repeat the entire process.
ix. Let D_l be length of the final synthesized DNA.

3) Encrypt the secret message: The secret message and the meta-data are converted to byte values and passed through RC4 encryption algorithm. The symmetric encryption key (K_{E3}) is also known to the receiver of the message.

4) Embed Stego Data: In this step subsequent 2 bits of a secret message are embedded in the 3rd base of a degeneracy codon using the following steps.

i. The encrypted byte values are converted to 8 bit binary values.
ii. Select 2 bits (m_0, m_1) from encrypted data bit stream.
iii. Choose the next degeneracy codon's index from the stored set generated during capacity check process.
iv. Replace the 3rd base of the codon with the codon represented by the 2 message bits.
v. Repeat steps (ii) to (iv) until all the message bits are embedded in the generated oligonucleotide or ssDNA.
vi. Choose a primer with length l bases ($l \geq 20$, for better security) to be used as secondary key to the algorithm.
vii. Perform algebraic addition operation between the modified DNA and the primer using Table I. The rows and columns of the table are indexed from 0 - 3, giving us a equivalent map for the DNA bases. The decimal equivalent of the mutated DNA base indicates the row number, while the decimal equivalent of the primer base denotes column number. The value contained in the cell identified by this row and column number replaces the current DNA base from the mutated DNA.
The addition is performed in sequence. As generally $l \ll D_l$, the primer is repeated after performing a single addition operation with the modified DNA strand.
viii. Next the 2D logistic sequence is generated again and all the codons are embeded to their corresponding locations, i.e. the LSBs of all the color components of all the pixels are replaced by the corresponding nucleotide decoding bits. The algebraic addition operation is a confusion operation, which eliminated any trace of codon degeneracy, which increases the key space for steganalysis purpose.

B. Extraction Algorithm

The extraction process has two parts- header extraction and data extraction. During the header extraction process all the meta data related to the secret message and embedding process are extracted. The steps involved in secret message extraction are explained below.

i. Extract the logistic map parameters from K_{E1} and run the map until N to extract the ssDNA.
ii. From the pixels extract the DNA codons which can be generated by DNA encoding process using the 2 LSBs of each color component of every extracted pixel. With the help of these codons construct the ssDNA.
iii. Perform the algebraic primer subtraction operation on the ssDNA using K_{E2}. Same Table I is used because it is self-invertible.
iv. From the final DNA strand locate the degeneracy codons and extract their third nucleotide base.
v. The third nucleotide base contains the secret message bits. We can arrange them in order and from subsequent 8 bits we construct the encrypted message byte stream.
vi. The encrypted message byte stream is then passed through RC4 cypher to give the original message byte.
vii. The bytes are converted to their corresponding ASCII characters and the original message is restored.

IV. Experimental Results

We conducted the following tests on our algorithm to check its effectiveness of hiding against some of the very common to most sophisticated attacks available today. Our secret message length varies from 32 bases to 264 bases. The algorithm hides 2 bits (1 base) per pixel. The key primer length is 20-40 bases. As a result distinct number of possible primers ranges from 4^{20}- 4^{40}. The primer, along with the 6 parameters for 2D map provides a huge key-space for steganalysis purpose.

A. Metrics of Distortion

The following metrics are calculated in terms of embedding capacity and distortion made to the image to assess the quality of our stego images. Table III shows that with increase in payload amount change in image distortion is quite less.

TABLE III: Metrics of distortion values at different payloads

Payload	MSE	PSNR	AD	LMSE	NAE
10%	0.101041	58.08582	0.001017	0.001663	0.0003
20%	0.200184	55.116513	0.002415	0.003292	0.00059
30%	0.298948	53.374852	0.003599	0.004917	0.00089
40%	0.396935	52.143615	0.004509	0.006509	0.00118
50%	0.496807	51.168926	0.005939	0.008124	0.00147
60%	0.594551	50.388912	0.006953	0.009705	0.00177
70%	0.682023	49.792811	0.008261	0.011125	0.00202
80%	0.789327	49.158234	0.010679	0.012865	0.00234
90%	0.886463	48.654197	0.012638	0.014406	0.00263

B. Visual Attack

We applied the LSB enhancement method [19] to check any visually perceivable change between the carrier and the stegogramme. Fig. 1(a) and 1(b) shows the result. As we can see that even with about 78% payload, difference between the two LSB planes is barely distinguishable.

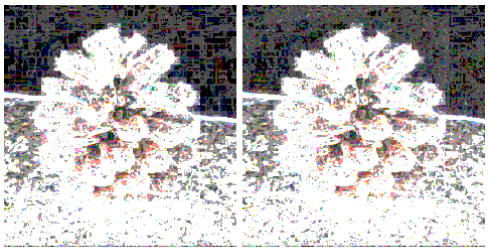

Fig. 1: LSB planes (a) original pine cover (b) with secret data

C. Statistical Attack

1) Histograms Analysis: We performed histogram analysis on luminance channel to detect significant changes in frequency of appearance of the colors (see Fig. 2). Results show that with 59% of payload our algorithm preserves the general shapes of the histograms. The changes in several other parameters (such as mean, standard deviation, energy and relative entropy) are also negligible.

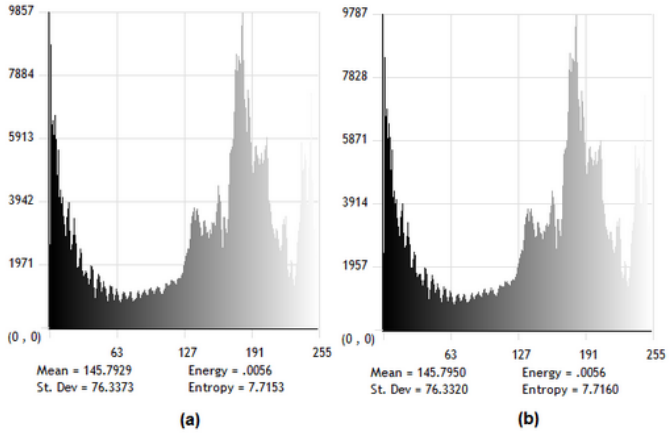

Fig. 2: Histogram analysis (a) original cover (b) stegogramme

2) Neighbourhood Histogram: We performed the neighbourhood histogram [20] analysis to test the effect of LSB embedding on neighbourhood of each pixel. As we can see, by our algorithm with 76% payload, change in neighbours count is negligible, although the frequency hike is remarkable (see Fig. 3).

3) Difference Image Histogram: We analyzed the difference image histograms [21] for all our images with and without 62% payload. Results show that no visible Pairs of Values (PoVs) are identified in the histogram and it retains the natural Gaussian shape by maintaining the slope of bars in each range (see Fig. 4).

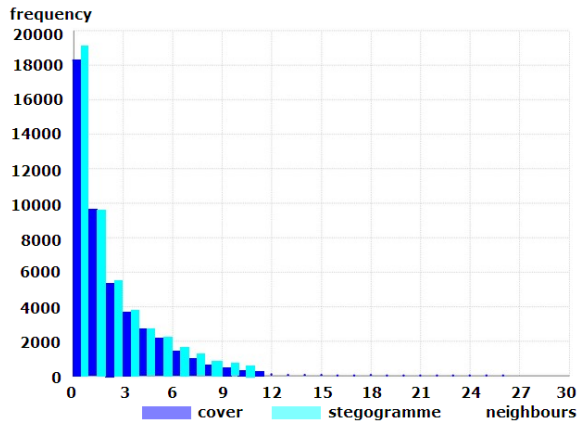

Fig. 3: Neighbourhood Histogram

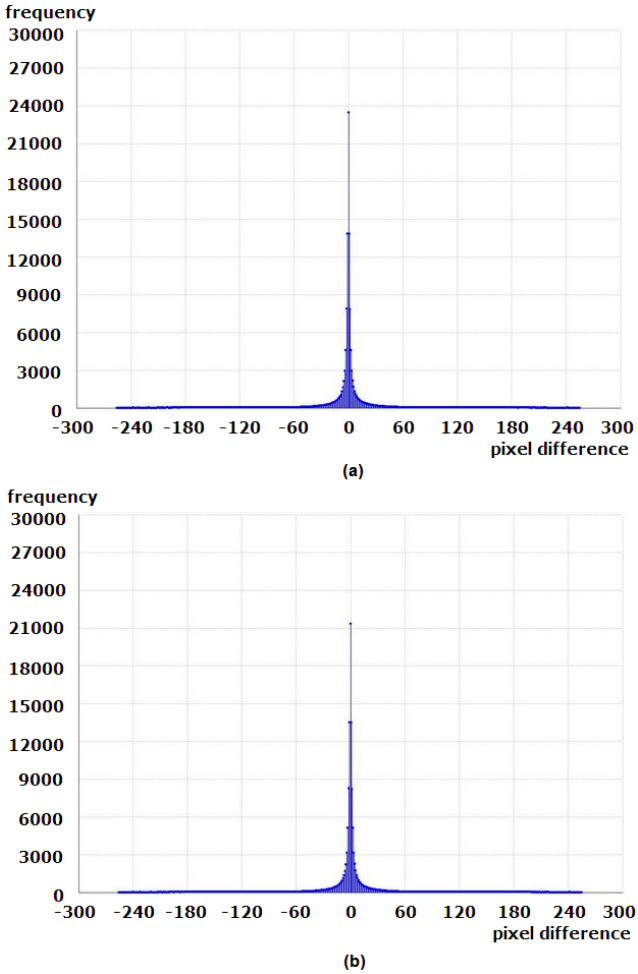

Fig. 4: Difference Histogram (a) original cover (b) stegogramme

V. Conclusion & Future Work

The dual cover steganography algorithm proposed in the paper provides better security than many of the existing algorithms. The 2D logistic map secures the generated random sequence with the 6 parameters with vast steganalysis domain. Also the primer addition provides an effect equivalent to cryptographic confusion. Application of DNA properties changes the operation domain. Test results show that- with moderate amount of payload it is highly impossible to detect the existence of secret message. Despite the fact of being a secure hiding method, security of the key elements is also

equally important. That is why the focus of our future work will be to develop a secure key exchange protocol for our proposed steganography algorithm.

REFERENCES

[1] L. M. Adleman, "Molecular computation of solutions to combinatorial problem", *Science*, vol. 266, no. 5187, pp. 1021-1024, 1994.

[2] C. Clelland, V. Risca and C. Bancroft, "Hiding messages in DNA microdots", *Nature*, 399(6736), pp. 533-534, 1999.

[3] A. Leier, C. Richter, W. Banzhaf and H. Rauhe, "Cryptography with DNA binary strands". *BioSystems*, vol. 57, pp. 13-22, 2000.

[4] H. Shiu, K. Ng, J.F. Fnag, R. Lee and C. Huang, "Data hiding methods based upon DNA sequences," *Information of Sciences*, vol. 180, no. 11, pp. 2196-2208, 2010.

[5] B. Shimanovsky, J. Feng and M. Potkonjak, "Hiding Data in DNA," *In Procs. of the 5th International Workshop in Information Hiding*, LNCS, vol. 2578, pp. 373-386, 2002.

[6] M. Arita and Y. Ohashi, "Secret signatures inside genomic DNA," *Biotechnology Progress*, vol. 20, No.5, pp. 1605-1607, 2004.

[7] C. Chang, T. Lu., Y. Chang and C. Lee, "Reversible Data Hiding Schemes for DEOXYRIBONUCLEIC ACID Medium," *International Journal of Innovative Computing, Information and Control*, vol. 3, no. 5, pp. 1-16, 2007.

[8] Mohammad Reza Najaf Torkaman, Pourya Nikfard, Nazanin Sadat Kazazi, Mohammad Reza Abbasy and S. Farzaneh Tabatabaiee, "Improving Hybrid Cryptosystems with DNA Steganography", *DEIS 2011*, pp. 42-52, 2011.

[9] Hayam Mousa, Kamel Moustafa, Waiel Abdel-Wahed and Mohiy Hadhoud, "Data Hiding Based on Contrast Mapping Using DNA Medium", *The International Arab Journal of Information Technology*, Vol. 8, No. 2, pp. 147-154, 2011.

[10] Meenakshi S Arya, Nikita Jain, Jai Sisodia and Nukul Sehgal, "DNA Encoding Based Feature Extraction for Biometric Watermarking", *International Conference on Image Information Processing (ICIIP 2011)*, 2011.

[11] Samir Kumar Bandyopadhyay and Suman Chakraborty, "IMAGE STEGANOGRAPHY USING DNA SEQUENCE", *Asian Journal Of Computer Science And Information Technology*, vol. 1, No.2, pp. 50 52, 2011.

[12] Suman Chakraborty and Prof. Samir Kumar Bandyopadhyay, "Two Stages Data-Image Steganography Using DNA Sequence". *International Journal of Engineering Research and Development*, Volume 2, Issue 7, PP. 69-72, August 2012.

[13] Suman Chakraborty, Sudipta Roy and Prof. Samir K. Bandyopadhyay, "Image Steganography Using DNA Sequence and Sudoku Solution Matrix", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 2, Issue 2, February 2012.

[14] Amal Khalifa and Ahmed Atito, "High-Capacity DNA-based Steganography", *The 8th Internationa.l Conference on INFOrmatics and Systems (INFOS2012)*, Bio-inspired Optimization Algonthms and Their Applications Track, May, 2012.

[15] Mohammad Reza Abbasy, Pourya Nikfard, Ali Ordi and Mohammad Reza Najaf Torkaman, "DNA Base Data Hiding Algorithm", *International Journal on New Computer Architectures and Their Applications (IJNCAA)*, vol. 2, No. 1, pp. 183-192, 2012.

[16] Piotr Wasiewicz, lan J. Mulawka, Witold R. Rudnicki and Bogdan Lesyng, "Adding Numbers with DNA", *International Conference on Systems, Man and Cybernetics*, pp. 265-270, 2000.

[17] S. Jiao and R. Goutte, "Code For Encryption Hiding Data into Genomic DNA Of Living Organisms," *9th International Conference on Signal Processing (ICSP2008)*, pp. 2166-2169, 2008.

[18] Hongjuan Liu, Zhiliang Zhu, Huiyan Jiang and Beilei Wang, "A Novel Image Encryption Algorithm Based on Improved 3D Chaotic Cat Map", *The 9th International Conference for Young Computer Scientists*, pp. 3016-302l, 2009.

[19] A. Westfeld and A. Pfitzmann, "Attacks on steganographic systems," *Third International Workshop on Information Hiding*, vol. 1768, pp. 61-76, 1999.

[20] A. Westfeld, "Detecting low embedding rates," F.A.P. Petitcolas (Ed.), Information Hiding. 5th International Workshop, Springer-Verlag Berlin Heidelberg, pp. 324-339, 2003.

[21] T. Zhang and X. Ping, "Reliable detection of lsb steganography based on the difference image histogram," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 03)*, Vol. 3, 2003.