Information Sciences Letters

Volume 9
Issue 3 *Sep. 2020*

Article 7

2020

# DNA-based steganography using genetic algorithm

Marghny H. Mohammed
*Department of Computer Science, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt*, alaa.abdelrazek@compit.aun.edu.eg

Alaa Abdel-Razeq
*Department of Computer Science, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt*, alaa.abdelrazek@compit.aun.edu.eg

Inf. Sci. Lett. **9**, No. 3, 205-210 (2020)



Information Sciences Letters
*An International Journal*

http://dx.doi.org/10.18576/isl/090307

# DNA-based steganography using genetic algorithm

*Marghny H. Mohammed and Alaa Abdel-Razeq**

Department of Computer Science, Faculty of Computers and Information, Assiut University, Assiut 71526, Egypt

Received: 2 Dec. 2019, Revised: 20 June 2020, Accepted: 27 June 2020
Published online: 1 Sep. 2020

**Abstract:** Development in steganography cares about increasing the amount of secret data embedded with a carrier. DNA file as a cover media is the best choice for that manner to increase the data embededd. This paper presents steganography with DNA files and proposes a new method to embed secret data in DNA files using genetic algorithm to enhance the performance and care with the steganography performance measures. The proposed method solves the problems of substitution method, using genetic algorithm to choose the best positions in the DNA file, to embed secret data, make the modification rate equal 0 in most cases and the lowest in other cases, and reduce the list generated. The cracking probability of the algorithm is very low. For more security, the secret data are encrypted with RSA algorithm before embedding. Final experiment results show an improvement in modification rate of the carrier.

**Keywords:** Steganography, security, data hiding, DNA files, genetic algorithm.

## 1 Introduction

Recently, the Internet has become important and extensively used in various fields, so development and innovation in security are required to keep sensitive data more secure from unauthorized access [1,2]. In security, the most used techniques are cryptography and steganography. Cryptography changes the form of secret information, while steganography hides the secret information from unauthorized access, so steganography is preferred and more secure in public and insecure channel [3,4,5]. Steganography model, which is shown in Fig. 1, consists of two algorithms: embedding and extracting. The embedding algorithm is used to hide the secret message in a cover media to generate a stego media that will be sent to the receiver. The extracting algorithm is used to extract the secret message from the stego-media. It is optional to use a key in the embedding and extracting algorithms as a level of security [6].

Steganography hides the secret message inside a cover media which is text, image, audio, video, DNA files,... etc. The most widely used is image, but DNA files have been recently preferred by researchers because a huge capacity of data can be stored. DNA (Deoxyribonucleic acid) holds genetic information and features of organism [7,8,9]. DNA is composed of two long strands known as a double helix, and each nucleotide comprises a purine or a pyrimidine base. The purine bases are Adenine (A) and Guanine (G), while the pyrimidine Thymine (T) and Cytosine (C). The critical feature behind using DNA files as a media for steganography is the very low visibility level of DNA sequences which makes finding secret messages in DNA sequences extremely difficult. In addition, the high randomness property of DNA efficiently hide any message [10].

The requirements of any steganography model that measure strength are robustness, capacity and imperceptibility. Robustness is the ability to recover the secret data from stego-carrier. Capacity is the quantity of secret data that can be embedded. Imperceptibility is the level where the stego-carrier more is close to the original. The most important requirement of steganography is the high imperceptibility followed by high capacity [11].

Clelland et al. [12] succeeded in achieving DNA steganography by encrypting the secret message in a DNA strand, then the strand is flanked by polymerase chain reaction (PCR) primer sequences within the enormous complexity of human genomic DNA. After that, it is hidden in a microdot. Shiu et al. [13] proposed three DNA based steganography methods: insertion, substitution, and complementary pair. The insertion method is based on inserting the secret message into the DNA sequence. It could be inserted as a block or the secret message is divided into segments and each segment is inserted into the divided parts of the DNA sequence. The substitution method is based on replacing the DNA

* Corresponding author e-mail: alaa.abdelrazek@compit.aun.edu.eg

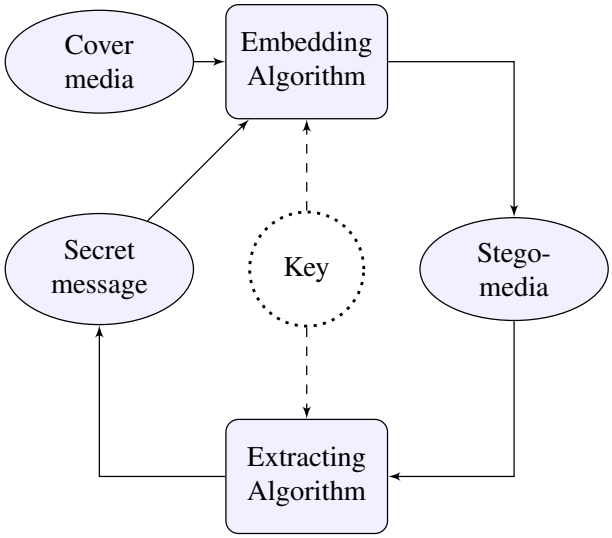


**Fig. 1:** Steganography Model.

nucleotide in the DNA sequence with the message bits according to a specific rule. The complementary pair method is based on using complementary pairs for the secret message segments and inserting them into the DNA sequence. Mousa et al. [14] applied reversible information hiding scheme on deoxyribonucleic acid sequence using the reversible contrast mapping technique. Abbasy et al. [15] divided the secret message into two complementary DNA nucleotides segments and search for the indices of each segment in the DNA reference sequence. Khalifa and Atito [16] encrypted the secret message by a DNA-based Playfair cipher algorithm. Then, the algorithm uses a substitution method to hide the encrypted secret message depending on a generic two-by-two complementary rule. Taur et al. [17] proposed an improved algorithm for substitution method called Table look up substitution method that uses look up table for the DNA characters to embed secret data in the DNA file using substitution method. Wang et al. [18] used an encryption algorithm to encrypt a message and decompose the encrypted message into two parts. One part is sent using DNA steganography. If the set part contaminated, the message will be encrypted and decomposed again and again until the microdot is not contaminated. Then, the other part will be publicly sent. Hamed et al. [20] applied the algorithm in two phases: First phase, the secret message is converted into DNA using generic N-bits binary coding rule. Then, DNA and amino acids playfair is applied to encrypt the DNA of the encrypted message in ambiguity. The second phase, the encrypted message is hidden in a DNA sequence at random positions. Malathi et al. [21] developed the DNA steganography using an improved DNA insertion algorithm that XOR the message repeatedly with the DNA sequence. The problems of algorithms using substitution method are the high modification rate with the original DNA file and the list generated increases with increasing message length, and with the insertion method the size of the DNA fileis expanded. Moreover, most algorithms are not blind and its a weakness point. In this paper, the proposed algorithm improves substitution method to hide the secret message in the DNA sequence using genetic algorithmto select the best positions to embed secret data, and RSA algorithm to encrypt secret data (RSA is a very strong encryption algorithm).

### *1.1 Genetic Algorithm*

Genetic Algorithm is adaptive heuristic exploration algorithm based on mechanics of the theory of natural selection and natural genetics [23]. It works , as follows:
**1.** Generate initial population with p chromosomes.
**2.** Fitness function is used to evaluate the populations of each individual chromosome.
**3.** Select top n of the population to survive based on the fitness function evaluation.
**4.** Use cross over and mutation to generate new children from n chromosome selected as a parent.
**5.** Go to step2 and repeat till finding the best solution.

Fitness function is a measure process to select a chromosome which will survive. Cross over is a genetic operator used to induce variation from one generation to another. Mutation is a genetic operator applied to a set of population to maintain its genetic diversity. Genetic algorithm is used in the proposed algorithm to generate a list of positions in DNA sequence which could hide the message segment. Selection of positions is evaluated and measured to represent the lowest modification as possible. In this paper, Section 2 presents the proposed method with the embedding and extracting algorithms in details. Section 3 addresses performance measures i.e., (cracking probability, capacity, payload and Bit per nucleotide), Section 4 presents the experimental results. Section 5 is dedicated to conclusion.

## 2 Proposed method

The model of DNA based steganography using genetic algorithm solution is presented in **Fig. 2**. It has a pair of algorithms embedding and extracting. The embedding algorithm uses a DNA file from EBI or NCIB as a carrier. The secret message is in a form of ASCII code. First, we encrypt it using RSA with key1public key and the result will be converted to a binary sequence, and using the binary rule $(A = 00, C = 01, G = 10, T = 11)$ to convert the binary sequence of the secret message into DNA sequence. Then, we divide the DNA sequence with key2 (1, 2, 4, 8 etc.) into segments. Using of genetic algorithm, we could obtain the best solution of choosing positions in

Inf. Sci. Lett. **9**, No. 3, 205-210 (2020) / www.naturalspublishing.com/Journals.asp 

the DNA file to embed these segments of the secret message. Finally, the positions list, key2 and the fake DNA are sent to the receiver. The extracting algorithm uses the positions list to seek in the fake DNA file, get the segments in these positions with length key2, concatenate those to get the DNA sequence of the encrypted secret message transform the sequence into ASCII code. Then RSA is used to decrypt it with key1 private. The result is the secret message in ASCII form.

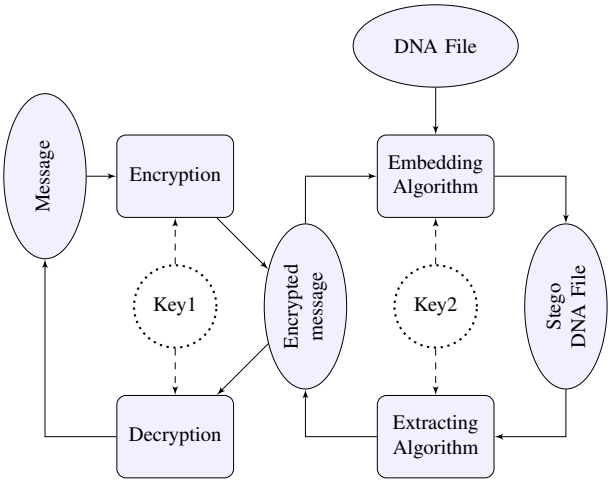


**Fig. 2:** Proposed method.

### *2.1 Embedding Algorithm*

**Input:** secret message M, DNA sequence S, binary coding rule, key1public, key2 (segments).
**Output:** fake DNA file, list.
Algorithm
**Step1:** encrypt the message using RSA Encryption algorithm with key1public.
**Step2:** convert the encrypted message into binary sequence Mbin.
**Step3:** convert the binary sequence Mbin using binary rule $(A = 00, C = 01, G = 10, T = 11)$ to DNA sequence.
**Step4:** Divide the message into segments using key2 (1, 2, 4 or 8).
**Step5:** Use genetic algorithm to find the best positions that make the smallest mean square error MSE modifications in file as the following steps:
**1.** Initialize the positions list P with random integer values from 0 to DNA file length segments multiply by 2 and generate N list of positions. Number of positions in a list P is (secret message length/ key2 = segments).
**2.** Apply the fitness function: using the positions list seek in the DNA file and get the segments (of length key2) Concatenate the segments with the same order in positions list and compare it with the secret message to calculate the MSE for all solutions N and select the lowest 20. MSE: Mean Square Error
**3.** Using cross over and mutation on the top 20 solutions to generate new children having the right positions from parent and generate randomly the rest of them.
**4.** New lists will generate then go to step 2 and repeat these steps till the MSE equal 0 or reach a specific number of iterations.
**Step6:** If the MSE not equal to 0 modify in the DNA file the not matched positions.
**Step7:** Send the list, fake DNA file, key2.

### *2.2 Extracting Algorithm*

**Input:** fake DNA file, positions list, key1private, key2, binary rule.
**Output:** original message.
Algorithm
**Step1:** Using the positions list seek in the fake DNA file and get the segments of length key2.
**Step2:** Concatenate the segments with the same order in positions list.
**Step3:** Convert the DNA sequence to binary sequence using binary rule.
**Step4:** convert the binary sequence to ASCII form.
**Step5:** Using RSA algorithm and key1private to decrypt the message to get the original form.

## 3 Performance measures

### *3.1 Cracking probability*

It is the total probability to predict the confidential information hidden inside the reference DNA sequence. The attacker needs the following information to crack the secret message hidden in the reference DNA [21]. The probability to predict these factors:
**Factor1:** Reference DNA sequence

$$\frac{1}{163 \times 10^6}. \tag{1}$$

The reference DNA available is about 163*106 (EBI or NCBI)
**Factor2:** Find the message in the DNA reference

$$\frac{1}{(y-1)}. \tag{2}$$

**Factor3:** Segments

$$\frac{1}{2^(m-1)}. \tag{3}$$

**Factor4:** Binary coding(A, C, G, T)

$$\frac{1}{24}. \quad (4)$$

**Factor5:** List of positions

$$\frac{1}{l^{(}y-1)}. \quad (5)$$

Thus, the total probability to find the message hidden in the DNA sequence using the proposed method is

$$\frac{1}{163\times 10^6}\times\frac{1}{(y-1)}\times\frac{1}{2^{(}m-1)}\times\frac{1}{24}\times\frac{1}{l^{(}y-1)}. \quad (6)$$

From the formula, we predict that it seems impossible for the attacker to predict the message hidden in the DNA sequence using the proposed method as the cracking probability is very low in addition to the cracking probability of RSA algorithm which is extremely hard to crack.

**Table 1:** Comparison of cracking probability methods.

| Method | Cracking Probability |
|---|---|
| Insertion based method [13] | $\frac{1}{1.63\times10^8}\times\frac{1}{24}\times\frac{1}{(n-1)}\times\frac{1}{(2^m-1)}\times\frac{1}{2^{s-1}}$ |
| Complementary based method [13] | $\frac{1}{1.63\times10^8}\times\frac{1}{24^2}$ *or* $\frac{1}{3^n}$ |
| Insertion based method [21] | $\frac{1}{1.63\times10^8}\times\frac{1}{24}\times\frac{1}{(n-1)}\times\frac{1}{(2^m-1)}\times\frac{1}{2^{s-1}}\times\frac{1}{2^{8m}}$ |
| Substitution based method [13] | $\frac{1}{1.63\times10^8}\times\frac{1}{6}$ |
| The Proposed algorithm | $\frac{1}{163\times10^6}\times\frac{1}{(y-1)}\times\frac{1}{2^{(}m-1)}\times\frac{1}{24}\times\frac{1}{L^{(}y-1)}$ |

**x** is the number of alphabets in the reference DNA sequence.
**y** is the number of alphabets in the fake DNA sequence.
**m** is the number of alphabets in the secret message.
**l** is the length of the positions list.
**s** is the number of segments of the secret message.

### *3.2 Capacity*

It is the length of the fake DNA sequence. The proposed method uses enhanced substitution based method to embed the secret message, so the original DNA equals the fake DNA reference in size.

### *3.3 Payload*

It is the length of the embedded data. The proposed method uses the enhanced substitution based method to embed the secret message, so the payload equals zero.

### *3.4 Bit per nucleotide (BPN)*

It is the average number of secret bits hidden per character.

## 4 Experimental Results

Tables 2, 3, 4, 5 and 6 show the experimental results performance of using 8 DNA sequences of cow and mice implemented on the proposed algorithm with message length 20000 ASCII character.

**Table 2:** Performance measures using proposed algorithm.

| Sequence | Number of DNA characters | Capacity | Payload | BPN |
|---|---|---|---|---|
| AC153526 | 200117 | 200117 | 0 | 2 |
| AC166252 | 149884 | 149884 | 0 | 2 |
| AC167221 | 204841 | 204841 | 0 | 2 |
| AC168874 | 206488 | 206488 | 0 | 2 |
| AC168897 | 200203 | 200203 | 0 | 2 |
| AC168901 | 191456 | 191456 | 0 | 2 |
| AC168907 | 194226 | 194226 | 0 | 2 |
| AC168908 | 218028 | 218028 | 0 | 2 |

Table 2 shows that for all samples, payload is zero and capacity equals to the original DNA file which means that the length of the fake DNA file is the same as original, so no expansion occurs in the file, while Malathi [21] method payload equals the secret message length and with the increase of secret data amount, the payload of Malathi [21] increases and the proposed method payload remains 0. The BPN is 2 and 2 bits are embedded in a nucleotide.

**Table 3:** Positions list for 1 character.

| Sequence | Number of DNA characters | Length of list | Modification in file |
|---|---|---|---|
| AC153526 | 200117 | 109440 | 0 |
| AC166252 | 149884 | 109440 | 0 |
| AC167221 | 204841 | 109440 | 0 |
| AC168874 | 206488 | 109440 | 0 |
| AC168897 | 200203 | 109440 | 0 |
| AC168901 | 191456 | 109440 | 0 |
| AC168907 | 194226 | 109440 | 0 |
| AC168908 | 218028 | 109440 | 0 |

The secret message is embedded as segments (1, 2, 4 or 8) and character (character= 2 bits) represents position for a segment. Tables 3, 4, 5 and 6 show the length of positions list and the modification rate. Table 3 shows the results of 1 character in a position; Table 4 shows the results of 2 characters in a position; Table 5 shows the

Inf. Sci. Lett. **9**, No. 3, 205-210 (2020) / www.naturalspublishing.com/Journals.asp 

**Table 4:** Positions list for 2 character.

| Sequence | Number of DNA characters | Length of list | Modification in file |
|---|---|---|---|
| AC153526 | 200117 | 54720 | 0 |
| AC166252 | 149884 | 54720 | 0 |
| AC167221 | 204841 | 54720 | 0 |
| AC168874 | 206488 | 54720 | 0 |
| AC168897 | 200203 | 54720 | 0 |
| AC168901 | 191456 | 54720 | 0 |
| AC168907 | 194226 | 54720 | 0 |
| AC168908 | 218028 | 54720 | 0 |

**Table 5:** Positions list for 4 character.

| Sequence | Number of DNA characters | Length of list | Modification in file |
|---|---|---|---|
| AC153526 | 200117 | 27360 | 0 |
| AC166252 | 149884 | 27360 | 0 |
| AC167221 | 204841 | 27360 | 0 |
| AC168874 | 206488 | 27360 | 0 |
| AC168897 | 200203 | 27360 | 0 |
| AC168901 | 191456 | 27360 | 0 |
| AC168907 | 194226 | 27360 | 0 |
| AC168908 | 218028 | 27360 | 0 |

**Table 6:** Positions list for 8 character.

| Sequence | Number of DNA characters | Length of list | Modification in file |
|---|---|---|---|
| AC153526 | 200117 | 13680 | 0.158 |
| AC166252 | 149884 | 13680 | 0.231 |
| AC167221 | 204841 | 13680 | 0.147 |
| AC168874 | 206488 | 13680 | 0.148 |
| AC168897 | 200203 | 13680 | 0.159 |
| AC168901 | 191456 | 13680 | 0.194 |
| AC168907 | 194226 | 13680 | 0.122 |
| AC168908 | 218028 | 13680 | 0.161 |

results of 4 characters in a position; and Table 6 shows the results of 8 characters in a position, when increasing the character number in the segment, the list of positions decreases. Thus, the sent data decrease. The most important advantage of the proposed algorithm is the improvement of modification rate. The results show that with segments (1, 2 and 4), the modification rate is zero which means the fake DNA file, is the original DNA file and with segment equal 8 characters, the modification rate is very low. Taur J-S [17] has a very high modification rate. However, in the proposed method almost has no or low modification. Furthermore, the positions list with Taur J-S [17] is 80000 position and increases with increasing secret data amount but the proposed algorithm, the list could decrease with increasing segments of secret data. The position list is dynamic as for each time the algorithm is applied on the same DNA file with the same secret message, the positions list values change. Using RSA algorithm to encrypt message before embedding is a strong point. The extracting algorithm is blind because it does not need the original DNA file to extract secret message. The algorithm is is easily implemented. Robustness, capacity, and imperceptibility using the proposed algorithm are achieved efficiently.

## 5 Conclusion

This paper introduces a new method to hide critical information inside a carrier. The proposed method uses genetic algorithm to choose the best positions in the DNA reference to hide the information where no or low modification occurs most or some cases of the DNA sequence compared to others methods and pay load equals zero. Furthermore, the list of positions of sent data. Using RSA to encrypt message before embedding makes the security level of the algorithm strong. The cracking probability of the proposed algorithm is very low. The algorithm is blind and achieves the requirements of robustness, capacity, and imperceptibility.

## Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

## References

[1] M. Mueller, A. Schmidt, B. Kuerbis, Internet Security and Networked Governance in International Relations, *International Studies Review*, **15**, 86–104 (2013).

[2] S. Notra, M. Siddiqi, H. H. Gharakheili, V. Sivaraman, R. Boreli, *An experimental study of security and privacy risks with emerging household appliances*, Communications and Network Security (CNS) 2014 IEEE Conference on, pp. 79-84, 2014.

[3] A.Joseph Raphael,Dr.V Sundaram, Cryptography and Steganography – A Survey, *Int. J. Comp. Tech. Appl.*, **2** 626-630 (2014).

[4] W. F. William, A. O. Osofisan and M. O. Asanbe, A Lookup XOR Cryptography for High Capacity Least Significant Bit Steganography, *International Journal of Applied Information Systems*, **10** 16-22 (2016).

[5] H. A. Prajapati, N. G. Chitaliya, Secured and Robust Dual Image Steganography: A Survey, *International Journal of Innovative Research in Computer and Communication Engineering* **3** 30-37 (2015).

[6] M. S. Subhedara, V. H. Mankar, Current status and key issues in image steganography: A survey, *Computer Science Review*, **13–14** 95-113 (2014).

[7] P. Vijayakumar, V. Vijayalakshmi and G. Zayaraz, An Improved Level of Security for DNA Steganography Using Hyperelliptic Curve Cryptography, *Wireless Pers Commun* **89** 1221–1242 (2016).

[8] E. Md. S. Hossain, K. Md. R. Alam, Md. R. Biswas, Y. Morimoto, *A DNA cryptographic technique based on dynamic DNA sequence table*, Computer and Information Technology (ICCIT) 2016 19th International Conference on, pp. 270-275, 2016.
[9] M. Shyamasree, S. Anees, *Highly secure DNA-based audio steganography*, Recent Trends in Information Technology (ICRTIT) 2013 International Conference on, pp. 519-524, 2013.
[10] G. Hamed, M. Marey, S. El-Sayed and F. Tolba, DNA based steganography: survey and analysis for parameters optimization, In: Hassanien AE., Grosan C., Fahmy Tolba M. (eds) Applications of Intelligent Optimization in Biology and Medicine. Intelligent Systems Reference Library, vol 96, 47-89 (2016) Springer, Cham.
[11] M. Mohamed , F. Al-Afari, MA. Bamatraf, Data Hiding by LSB Substitution Using Genetic Optimal Key-Permutation, *Int Arab J e-Technol* **1**, 11-7 (2011).
[12] C. T. Clelland, V. Risca, C. Bancroft, Hiding messages in DNA microdots, *Nature* **399**, 533 (1999).
[13] H. J. Shiu, K. L. Ng, J. F. Fang, R. C. Lee and C. H. Huang, Data hiding methods based upon DNA sequences, *Information Sciences*, **180**, 2196-2208 (2010).
[14] H. Mousa, K. Moustafa, W. Abdel-Wahed, MM. Hadhoud, Data hiding based on contrast mapping using DNA medium, *Int Arab J Inf Technol.*, **2**, 147-54 (2011).
[15] M. R. Abbasy, P. Nikfard, A. Ordi, and M. R. N. Torkaman, DNA base data hiding algorithm, *International Journal of New Computer Architectures and their Applications (IJNCAA)* **2**, 183-192 (2012).
[16] A. Khalifa and A. Atito, *High-capacity DNA-based steganography*, in Informatics and Systems (INFOS), 8th International Conference on. 2012. IEEE **7**6-80 (2012).
[17] J. S. Taur, H. Y. Lin, H. L. Lee and C. W. Tao, Data hiding in DNA sequences based on table lookup substitution, *International Journal of Innovative Computing, Information and Control* **8**, 6585-6598 (2012).
[18] Z. Wang, X. Zhao, H. Wang, G. Cui, editors, *Information hiding based on DNA steganography*, Software Engineering and Service Science (ICSESS), 4th IEEE International Conference 2013. IEEE (2013).
[19] N. Muhammad, N. Bibi, Z. Mahmood, DG. Kim, Blind data hiding technique using the Fresnelet transform, *SpringerPlus* **4**, 832 (2015).
[20] G. Hamed, M. Marey, SE-S. Amin, MF. Tolba, editors, *Hybrid Randomized and Biological Preserved DNA-Based Crypt-Steganography Using Generic N-Bits Binary Coding Rule* International Conference on Advanced Intelligent Systems and Informatics 2016. Springer(2015).
[21] P. Malathi, M. Manoaj, R. Manoj, V. Raghavan, and R. E. Vinodhini, Highly Improved DNA Based Steganography, *Procedia Computer Science* **115**, 651-659 (2017).
[22] G. Hamed, M. Marey, SE-S. Amin, F. Tolba, Hybrid, randomized and high capacity conservative mutations DNA-based steganography for large sized data, *Bio Systems* **167**, 47-61 (2018).
[23] S. Kalsi, H. Kaur, V. Chang, DNA Cryptography and Deep Learning using Genetic Algorithm with NW algorithm for Key Generation, *Journal of medical systems* **42**, 17 (2018).42(1):17.