Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering pp. 21–24
DOI: 10.15439/2022R42
ISSN 2300-5963 ACSIS, Vol. 33

A DNA Steganography Algorithm Based on The DNA-XOR Technique

Vinh-Quy Nguyen
Hung Yen University of Technology and Education
Hungyen, Vietnam
vinhquynguyen@gmail.com

Dinh-Chien Nguyen*
Hung Yen University of Technology and Education
Hungyen, Vietnam
https://orcid.org/0000-0002-4939-6355

Viet-Hung Dang
Hung Yen University of Technology and Education
Hungyen, Vietnam
dangviethung1107@gmail.com

Thanh-Hue Nguyen Thi
Hung Yen University of Technology and Education
Hungyen, Vietnam
huentt1509@gmail.com

Thu-Hang Phan Thi
Tan Lap junior high school
Hungyen, Vietnam
phanthuhang.tk36@gmail.com

Dinh-Thinh Luu
Doan Thi Diem junior high school
Hungyen, Vietnam
luuthinh@yenmy.edu.vn

***Abstract*—Steganography is the technique of concealing secret data in a physical object, such as a database, video, image, audio, QR code, and DNA sequence…. The DNA sequence is also considered for data hiding problems with the strong development of Bioinformatics. Many proposed algorithms to conceal data in DNA and RNA sequences, but they could not increase the amount of hidden data. In this study, we propose a DNA steganography algorithm for improving the embedded capacity in the DNA sequence. The algorithm uses the DNA-XOR technique based on the XOR operation. The secret data are encoded to a DNA sequence and then matched with the original DNA sequence by DNA-XOR operation to make a new DNA sequence. By this algorithm, we can embed two bits for each nucleotide. Moreover, with many kinds of combinations of nucleotides, the algorithm shows that the proposed method also improves the security of hidden data.**

***Index Terms*—Steganography, DNA sequence, XOR operation, DNA-XOR.**

I. Introduction

Steganography is the technique of concealing a secret data within a message or physical object such as database, image, audio, video, etc. [1-6]. Nowadays, many researchers studied data hiding methods for digital images [1-3].

Biological computing, including Bioinformatics, has supported researchers exploit techniques, which hide secret data in DNA (Deoxyribose nucleic acid) sequences, RNA (Ribonucleic acid) sequences, and protein structures.

In 1953, two famous genetic scientists, Watson and Crick found the DNA sequence structure [7]. In the research, DNA is the genetic material in each organism, consisting of humans, animals, and plants. DNA data is stored in a computer system with four nucleotide bases consist of Adenine (A), Guanine (G), Cytosine (C) và Thymine (T). Nucleotides combined into a DNA sequence by rule A pair with T and C pairs with G.

In 2012, Church et al. [8] proposed a study by using DNA sequence to store data, based on DNA synthesis and sequencing in Next Generation DNA Sequencing (NGS) by using DNA microchips. With these discoveries, concealing data in a DNA microchip will be fairly rewarding for cryptographers. However, with current technologies, DNA microchips are still quite expensive, the scientists have implemented hiding data in DNA sequences, which are taken from the website of the Information Technology Center in National Center for Biotechnology Information (NCBI) [9] or the website of the European Bioinformatics Institute in Ensembl (EMBL-EBI) [10]. In 2002, Shimanovsky et al. [11] presented two methods for hiding data in DNA and RNA. The first technique allows embedding the secret data in a non-coding DNA sequence, which is a DNA sequence that has not been transcripted and transferred to the genetic sequence. The second technique is used by scientists to organize data in active coding without converting it to the Amino Acid sequence. The techniques in this study can be used in the protection of intellectual property rights in biotechnology. Based on the analysis of Shimanovsky, many researchers have studied hiding secret data in DNA sequences [12-18], especially in non-coding regions.

In 2019, AI-Harbi et al. [19] presented some techniques for DNA-based steganography in security analysis. The study showed the advantages and disadvantages of a few methods, then suggested ways to improve the techniques for the DNA hiding field. Singh and Sharam [20] provided a review of nineteen existing algorithms of DNA-based cryptography.

To improve the embedded capacity, this study proposes an algorithm for data hiding in DNA sequences by using DNA-XOR operation. The rest of our work deals with the related studies, the proposed algorithm, and the implementation results. Section II introduces the methods of hiding information in the study of Shiu et al. [12]. Improving the data hiding algorithm with a higher capacity of embedding data by substitution method will be presented in the third section. Section IV shows the implementation results of the study. Conclusions and future studies will be shown in Section V.

* — corresponding author

II. Related work

In 2010, Shiu et al. [12] proposed three methods for concealing data based on the properties of DNA sequences.

The first method, the Insertion method, was performed by coding a DNA sequence to a binary string by the rule A - 00, C - 01, G -10, T - 11, then dividing the binary string into segments with size k bits. One secret data bit is inserted at the beginning of each segment, then merged all segments into a new binary string, and decoded into a pseudo-DNA sequence. On the receiver's side, the hiding data will be extracted at the beginning segments by the same rule, and then merge into a binary string of hiding data.

The second method, Complementary Pair, uses dynamic programming to find the longest complimentary substring, and then uses complementary rule, ((AC)(CG)(GT)(TA)). Each complementary substring is placed between the same nucleotide, for example, TACGT. The secret data will be embedded before the complementary substring, by the rules 00-A, 01-C, 10-G, and 11-T. For instance, with the DNA sequence S=ACGTAGCTGTTCTGTTCTCCTTCATGGAT and the secret data m='10110100'='GTCA', we can find a complementary substring (bold characters), and then insert a nucleotide before each complementary substring. The embedded DNA sequence is S'=ACGGTAGCTGTTTCTGTTCCTCCTTCAATGGAT. The hidden data will be extracted by the same algorithm with the embedding process, finding a complementary substring and then extracting the nucleotide before the substring. Finally, the binary string will be decoded from extracted nucleotides.

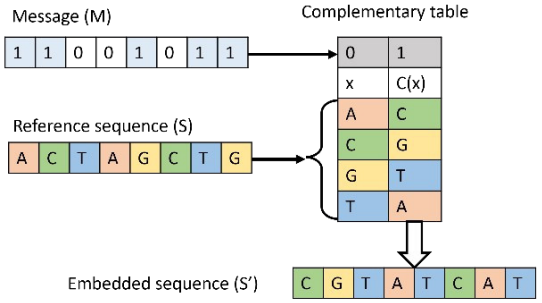

Fig. 1. Embedding process in the study [12]

The third method, the Substitution Method, employed the Complementary rule (x, C(x)). If the secret bit is 0, skip nucleotide, otherwise, alter the nucleotide x to nucleotide C(x), as in "Fig. 1".

On the receiver side, the DNA sequence, which brings hidden data (fake), is compared with the original sequence (Orig). If a nucleotide in fake is equal to a nucleotide in Orig, we get hidden bit 0, otherwise, the hidden bit is 1 "Fig. 2".

The methods did not improve the embedded capacity. So, in this study, we introduce a data hiding algorithm to improve the capacity, but still against massive attacks, in the following section.

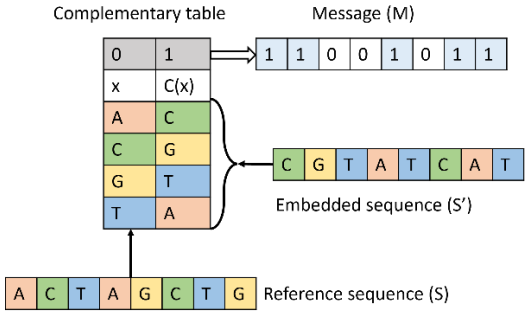

Fig. 2. Extraction process in the study [12]

III. Proposed Steganography algorithm

A. Embedding process

This study improves the embedding capacity by using DNA-XOR technique. We consider a nucleotide is presented by two bits, for example A - 00, C - 01, G – 10, and T -11. Based on XOR operation between two binary bits, we can build a table of DNA-XOR (Table 1). With this idea, we can embed two secret bits for each nucleotide, so that, the proposed algorithm can take more capacity than Shiu's [12] study.

Table 1. DNA-XOR Table

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

To embed message into a DNA Sequence, we can perform by following process,

Step 1: Encrypt Message to DNAseq by the rules, A - 00, C - 01, G – 10, and T -11

Step 2: Embed DNAseq into RefSeq using DNA-XOR table in Table 1

Fig. 3 shows the Embedding phase of the proposed algorithm.

Message may be a text or other format, so it must be encrypt to DNA sequence (Fake). To easy known, we assume that the message is in binary string.

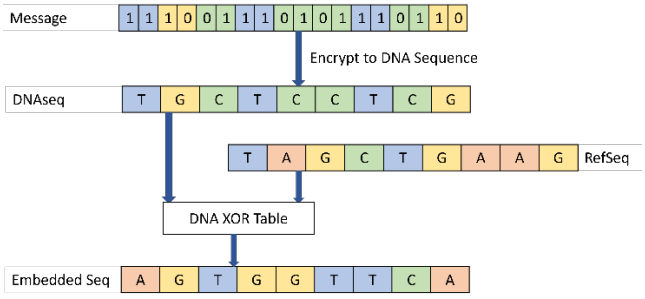

Fig. 3. The embedding process

B. *Extraction process*

To extract the hidden data,

Firstly, we match the Embedded Seq with the original DNA sequence (RefSeg) by using the DNA-XOR table (Table 1), and get the Extracted Seg.

Then, decrypt the Extracted Seg to bynary seqence (Message)

Fig. 4 presents Extraction process of this algorithm.

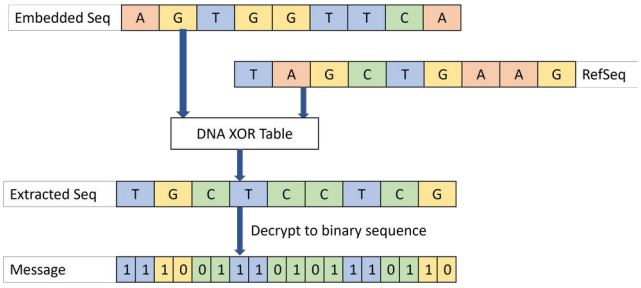

Fig. 4. Extraction process

IV. Experimental results

In this study, we implemented the data hiding algorithm by Matlab programming language on Core i5 Computer with 8GB RAM. DNA sequences database was downloaded from the NCBI website [9].

Compared with Shiu et al.'s algorithm, the proposed algorithm improved the embedded capacity of the secret data hidden in DNA sequences. (Table 2)

Table 2. Comparing the embedded capacity of the proposed algorithm with an algorithm in [12]

DNA Sequence	**Number of nucleotides**	**Shiu [12] algorithm**	**Proposed algorithm**
AC153526	200117	200117	400234
AC166252	149884	149884	299768
AC167221	204841	204841	409682
AC168874	206488	206488	412976
AC168901	191456	191456	382912
AC168907	194226	194226	388452
AC168908	218028	218028	436056
Average	***195005***	***195005***	***390010***

Fig. 5 shows the difference of nucleotide between the proposed method, presents by the yellow bars, and Shiu's method [12], shows in the green bars, with the Refence seqence. Although hiding data of the proposed method significantly changes the value of each nucleotide in the DNA sequence, the difference in nucleotides of both methods is negligible compared to the reference DNA sequence. For example, after embedding data in the DNA sequence with code AC166252, the number of Nucleotides A, C, G, and T are 39000, 34700, 35200, and 40100, respectively. The results show that the proposed algorithm can balance with the original DNA sequence, therefore, detecting data hiding in the reference DNA sequence will be more difficult.

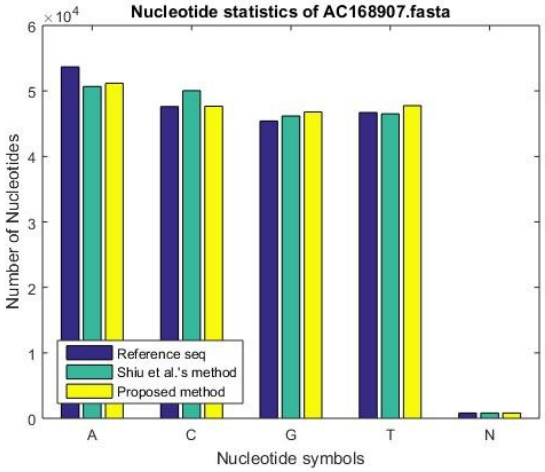

Fig. 5. The difference of nucleotide

Because we use the DNA-XOR table, the data extraction will be more difficult for the people who want to obtain the secret data. So, the proposed algorithm also improves the security of the hidden data in the DNA sequence. This characteristic is an important condition in data hiding problems

V. Conclusion

There are many data hiding algorithms in DNA sequences, but most of these algorithms either increase the size of the reference DNA sequence or can not improve the embedded capacity. The proposed algorithm can improve the capacity of embedding data, which hide in DNA sequences. For each nucleotide, the algorithm can embed two data bits. With a character (8 bits) in the text file, we use four nucleotides, so we can embed a 1MB text file into a DNA sequence of 4MB size. Moreover, the algorithm also improves the security for anyone who wants to get secrete data.

In 2020, Tabalabaei et al. proposed a new technique for recording information in DNA backbone [21]. This techniqe can change the storage devices technology, and helps researchers more improve DNA steganography methods.

References

[1] T. S. Nguyen, C. C. Chang, M. C. Lin, "Adaptive lossless data-hiding and compression scheme for SMVQ indices using SOC," Smart Comput. Review, vol. 4, nno. 3, pp. 230-245, 2014.

[2] J. Mielikainen, "LSB matching revisited," IEEE Signal Process. Letts., vol. 13, pp. 285–287, 2006.

[3] C. C. Chang, T. S. Nguyen, "A reversible data hiding scheme for SMVQ indices," Informatica, vol. 25, no. 4, pp. 523-540, 2014.

[4] C. V. Nguyen, D. Tay, and G. Deng, "A fast watermarking system for H.264/AVC video," in Proc. IEEE APCCAS, Dec. 2006, pp. 81–84.

[5] M. Fallahpour, M. David, "Reversible data hiding based on H. 264/AVC Intra prediction." Digital Watermarking. Springer Berlin Heidelberg, pp. 52-60, 2008.

[6] Chien. N. D, Son N. T, & Hsu F. R, “An algorithm for DNA sequence hiding in H. 264/AVC video.” In Proceedings of the Seventh Symposium on Information and Communication Technology ACM, pp. 229-234, December 2016.

[7] J.D. Watson, F.H.C. Crick, “Molecular structure of Nucleic acids: A structure for deoxyribose nucleic acid,” Nature 171 (1953), pp. 737, 738.

[8] Church, G. M., Gao, Y., & Kosuri, S., “Next-generation digital information storage in DNA,” Science, 337(6102), 1628-1628, 2012.

[9] National Center for Biotechnology Information, https://www.ncbi.nlm.nih.gov/

[10] Ensembl, http://www.ensembl.org/downloads.html

[11] Shimanovsky, B., Feng, J., & Potkonjak, M., “Hiding data in DNA” International Workshop on Information Hiding (pp. 373-386). Springer Berlin Heidelberg, 2002.

[12] Shiu, H. J., Ng, K. L., Fang, J. F., Lee, R. C., & Huang, C. H., “Data hiding methods based upon DNA sequences,” Information Sciences, 180(11), 2196-2208, 2010.

[13] Haughton, D., & Balado, F., “BioCode: Two biologically compatible Algorithms for embedding data in non-coding and coding regions of DNA,” BMC bioinformatics, 14(1), 121, 2013.

[14] Wang, Z., Zhao, X., Wang, H., & Cui, G., “Information hiding based on DNA steganography,” Software Engineering and Service Science (ICSESS), 2013 4th IEEE International Conference on (pp. 946-949). IEEE, 2013.

[15] Najaftorkaman, M., & Kazazi, N. S., “A method to encrypt information with DNA-based cryptography,” International Journal of Cyber-Security and Digital Forensics, 4(3), 417-427, 2015.

[16] UbaidurRahman, N. H., Balamurugan, C., & Mariappan, R., “A novel string matrix data structure for DNA encoding algorithm,” Procedia Computer Science, 46, 820-832, 2015.

[17] Huang, Y. H., Chang, C. C., & Wu, C. Y., “A DNA-based data hiding technique with low modification rates,” Multimedia tools and applications, 70(3), 1439-1451, 2014.

[18] Liu, H., Lin, D., & Kadir, A., “A novel data hiding method based on deoxyribonucleic acid coding,” Computers & Electrical Engineering, 39(4), 1164-1173, 2013.

[19] Al-Harbi, O. A., Alahmadi, W. E., & Aljahdali, A. O. “Security analysis of DNA based steganography techniques.” SN Applied Sciences, 2(2), 1-10, 2020.

[20] S. Singh and Y. Sharma, "A Review on DNA-based Cryptography for Data hiding," 2019 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, Tamilnadu, India, pp. 282-285, 2019.

[21] Tabatabaei, S. K., Wang, B., Athreya, N. B. M., Enghiad, B., Hernandez, A. G., Fields, C. J., ... & Milenkovic, O. (2020). DNA Punch Cards: Storing Data on Native DNA Sequences via Nicking. BioRxiv, 672394.