Scientific Computing Department
Faculty of Computer and Information Sciences
Ain Shams University

Deoxyribonucleic Acid Based Steganography

Thesis submitted as a partial fulfillment of the requirements for the degree of Master of Science in Computer and Information Sciences

By

Ghada Hamed Aly Kamel

Teaching Assistant at Scientific Computing Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Under Supervision of

Prof. Dr. Mohamed Fahmy Tolba
Professor in Scientific Computing Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Safaa Amin El-Sayed
AssociateProfessor in Scientific Computing Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Mohammed Abd El-Rahman Marey
Assistant Professor in Scientific Computing Department,
Faculty of Computer and Information Sciences,
Ain Shams University

January– 2017
Cairo

Acknowledgment

First of all, I would to like to thank GOD for his endless blessings, for giving me the power and strength to complete this work and for giving me supportive people.

Second, I would like to express my sincere gratitude to my supervisors; Prof. Dr. Mohamed Fahmy Tolba for his support, patience and guidance, Dr. Safaa Amin El-Sayed for her encouragement and Dr. Mohammed Marey for the special supervision experience he gave me. I am deeply thankful.

Third, I would like to thank my family for being available all the time and for the love they gave me through the years. Thank you for accepting me through the tough times and for always believing in me.

My dear friends who have helped me through the past time and kept on encouraging me to get this work done; Alaa Atef, Alaa Salah, Eman Hamdi, Hanan Yousry, Samar Ahmed and Yasmin Alaa without you it would have been much harder.

Last but not least, I would like to thank all my professors, colleagues and students who kept on encouraging me. Thank you for being in my life.

Ghada

Abstract

The research on developing data security techniques has been increased continuously, due to the indispensable need for powerful data protection in different applications, these include ownership protection, annotation, copyrighting, authentication and military applications. Cryptography and steganography are the most common and widely used techniques in computer security. Cryptography includes encrypting and decrypting some data into incomprehensible formats. While steganography aims to hide the existence of the message in a different media such as image, audio, video, etc. so that an intruder is mainly prevented from suspecting that the data is there. DNA is explored as a new carrier for data hiding due to its huge variations and its high capacity in which 10^6 TB of data can be stored in one gram of DNA. However, like every data storage device, hiding data in DNA requires a high protection level through a secure algorithm. This leads to a DNA steganography research field based on DNA computing, where the properties of biotechnology are exploited in steganography.

Mainly, there are two approaches in DNA based steganography topic. First approach: Techniques that provide high data hiding capacity but at the expense of the original biological properties of the used DNA reference sequence and it is the widest approach. This approach in most cases may result in dangerous side effects that may lead to the death of an organism related to the used sequence in the steganography process. The second approach cares about preserving the DNA biological and chemical properties

but at the expense of either the data hiding capacity allowed or the algorithms' cracking probability.

In this thesis, the second approach is achieved to allow merging the proposed algorithms with the industry in practical way and without any dangerous effects through two proposed algorithms. The first proposed one is a hybrid crypto-steganography algorithm is proposed that achieves double layers of security through two phases.In the first phase, the confidential data is converted to DNA using a proposed generic N-bits binary coding rule that leads to lower cracking probability compared by others. Then, DNA and amino acids Playfair is applied as a first phase to encrypt the DNA of the message resulting in ambiguity. In the second phase, the cipher text is placed with the ambiguity using 3:1 placement strategy. Then they are shuffled to be hidden in DNA at random positions generated by using a true real random number seed that is obtained from the atmospheric noise, thereby achieving very low cracking probability. The proposed technique is a blind preserved one as it achieves zero modification for the generated protein without extra data. The drawback of the first implemented algorithm in the steganography field is the low capacity which is overcome by merging the DNA based cryptography concepts with the proposed steganography technique that results in a double layer security level and low cracking probability. The achievement is that the selected cryptography algorithm results in the same size of the input data with no extra information.

The drawback of the first proposed algorithm gets a new objective which is providing higherdata capacity using DNA with preserving its main structure. For the sake of this, the second proposed algorithm is implemented to combinethe DNA conservative mutation concepts with the steganography.It results in utilizing the DNA in an effective and non-

negativity for neither the DNA nor the steganography process. The idea is that the conservative mutation is exploited in a way to support high capacity, thanks to the use of each DNA base to hide two bits that is considered the firstachievement. Since, most of the steganography algorithms support high capacity at the expense of the DNA modification rate which prevents it from being carried out practically. The second achievement in this study is selecting the conservative mutation specially to solve the tradeoff between the high capacity feature and preserving the original biochemical properties of the used sequence which is one of the main challenges considered in this research work. Besides minimizing the data that's sent to the receiver to strengthen the algorithm's security as compared with other techniques, results in sending only the carrier DNA sequence containing the secret data which is the final and third achievement in the steganography field.

After some deep biological view, the second proposed algorithm is modified to generalize the conservative mutation for each amino acid by getting all its possible substitutions with the other amino acids that have the same structure and functions.This proposed modification results in minimizing the algorithm's cracking probability by nearly fourteen billion trials. Even in the worst-case scenario, if the intruder success in taking away some parts of data, the extracted part is not meaningful, since the data is hidden in an arbitrary DNA bases which complexes the cracking process.

Besides that, for the first-time real large sized data to three megabytes of different formats are used to evaluate the performance of the proposed algorithm and investigate its scalability. The experimental results prove that the proposed technique is the one that overcomes the weakness points in the current steganography techniques as it merges the highest capacity feature with preserving the biochemical properties of the used DNA sequence

through a blind and a highly-secure algorithm without any generated extra information and with the lowest achieved cracking probability.

List of Publications

1- Ghada Hamed, Mohammed Marey, Safaa Amin and Mohamed Fahmy Tolba. "DNA based steganography: Survey and analysis for parameters optimization." Applications of intelligent optimization in biology and medicine, Springer International Publishing, pp 47-89, 2016.

2- Ghada Hamed, Mohammed Marey, Safaa Amin and Mohamed Fahmy Tolba. "Hybrid technique for steganography-based on DNA with n-bits binary coding rule." Soft Computing and Pattern Recognition (SoCPaR), pp 95-102, IEEE, November 2015.

3- Ghada Hamed, Mohammed Marey, Safaa Amin and Mohamed Fahmy Tolba. “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, pp 618-627, Springer International Publishing, 2016.

4- Ghada Hamed, Mohammed Marey, Safaa Amin and Mohamed Fahmy Tolba. “Hybrid, Randomized and High Capacity Conservative Mutations DNA-Based Steganography for Large Sized Data”. Information Sciences (Submitted).

5- Ghada Hamed, Mohammed Marey, Safaa Amin and Mohamed Fahmy Tolba. “Comparative Study for Various DNA Based Steganography Techniques with the Essential Conclusions about the Future

Research". The 11th IEEE International Conference on Computer Engineering and Systems, pp 220-225,IEEE, 2016.

6- Ghada Hamed, Mohammed Marey, Safaa Amin and Mohamed Fahmy Tolba. "Randomized DNA-based Crypto-Steganography Algorithm using Generic N-Bits Binary Coding Rule". Internet of Things and Big Data Technologies for Next Generation Healthcare, Springer International Publishing, 2016 (Accepted).

Table of Contents

List of Figures

List of Tables