Cryptographic Algorithm based on DNA and RNA Properties

Sally Safaa Nafea [1], Prof. Dr. Mohmood Khalel Ibrahem [2]

***Abstract*— The cryptography algorithms play a key role in the security of information which apply complex logics and mathematical procedures to develop efficient encryption algorithms. It can also be defined as an art that allows people to hide their personal information in the world of electronics. DNA cryptography is very current state of the art and a new upcoming domain that is emerging based on DNA computing, to encrypt big message in compact volume. In this paper, a cipher algorithm is proposed using biological processes and arithmetic operations. The biological processes are used to create symmetric key generation system based upon converting of DNA to RNA. The arithmetic operation used is XOR operation as an encryption technique. The proposed algorithm is a simulation of biological operation on digital computers which may be not exploit biological composition in term of storage and parallel processing.**

***Index Terms*— Biological Operations, Central Dogma of Molecular Biology, DNA Cryptography, DNA One-Time Pad Key Generation, Protein Form, RNA, XOR Operation.**

I. INTRODUCTION

There has been a tremendous growth in the number and type of attacks that should be dealt with by data security specialists to protect sensitive data vulnerable to unauthorized disclosure or undetected modification during transmission or while in storage [1]. Over time many technologies have been researched for securing the data using cryptography these technologies are either involves core knowledge of physics or mathematical [2].

Cryptography is a method of coding/decoding data so that it becomes unreadable or accessible by unauthorized users, which is often used to protect data during their transmission or while in storage [3]. Cryptography is specialize in building and analyzing protocols that resist the influence of enemies and which are connected to various aspects in information security such as authentication ,data confidentiality, non-repudiation and data integrity [4].

Information security experts found that binary computers (digital computers) have various physical constraints, especially in data storage and computation processes so they focused on DNA computers (bimolecular computers) and quantum computers [5].

DNA computing is a new field which is growing in the modern days and it is providing a brand new data structure and evaluating techniques for the parallel processing capabilities of molecules. The path of DNA cryptography started with the development of DNA computing [4]. DNA computing was discovered by Dr. Leonard M. Adleman [6] for the purpose of solving complex computational problem in 1994.

In this paper, DNA and its components and sequence are explained, also DNA cryptography is explained and compared with traditional cryptography. The biological operations on DNA has been also explained and used in the proposed algorithm. The Key generation, encryption and decryption in the proposed algorithm has been described, represented and implemented.

II. DNA

DNA is a biological molecule term, referred to Deoxyribo Nucleic Acid, which is the basic building block of the human body which represents the genetic blueprint of living creatures. DNA is unique for each individual and is a collection of the complex organic molecules. There is DNA in every cell of the organism which is important for the identify the identity of any living being[7]. DNA is a sequence of nucleotides, this exact sequences of the nucleotides determine the code of each gene. DNA sequences represent biological information such as skin color, weight, nose shape, eye, and hair as well as other features[8]. In 1953, James Watson formed the first 3D structure of DNA depend on an X-Ray print. Most DNA molecules consist of two biopolymer strands coiled around each other to form a double helix / stranded like a spiral ladder, as depicted in Fig. 1,[9].

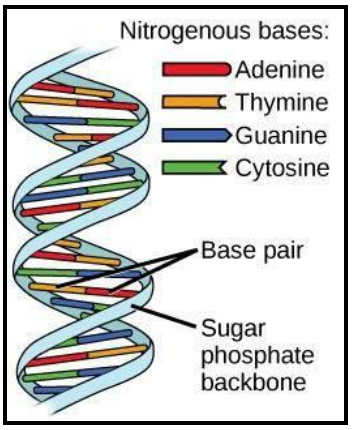


Figure 1. DNA Structure [5].

***Manuscript received Nov, 2018*.**

***Sally Safaa Nafea**, College of Information Engineering, Al-Nahrain University., Ramadi, Iraq.*

***Mohmood Khalel Ibrahem**, College of Information Engineering, Al-Nahrain University., Baghdad, Iraq.*

The two DNA strands are known as polynucleotides, since they are composed of simpler units called nucleotides. Each nucleotide is composed of nucleobase which is either Guanine (G),Adenine (A), Thymine (T) or Cytosine (C) as well as a phosphate group and a monosaccharide sugar called deoxyribose[4]. In DNA structure, there is a base pairing rules which determine that Guanine pairs with Cytosine and forms three hydrogen bonds, Adenine pairs with Thymine and forms two hydrogen bonds[4]. This base pairing rule is also called complementary theory of Watson-Crick as shown in Fig. 2 below.

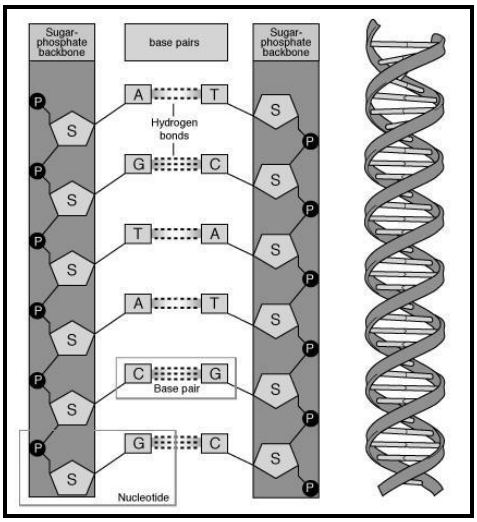


Figure 2. Hydrogen bonds in DNA[19].

DNA sequence has two strands, an individual strand as single stranded DNA (ssDNA) and double stranded DNA (dsDNA) [11]. There is a process called Hybridization that combine two strands ssDNA, which are anti-parallel to each other, and form dsDNA.The two ssDNA in dsDNA must be complementarily. This complementary makes DNA a unique data structure for computation and can be used in many ways[7].

In DNA strands there is directionality where one end of a DNA polymer contains an exposed hydroxyl group on the deoxyribose; this is known as the 3' end of the molecule. The other end contains an exposed phosphate group; this is the 5' end. Directionality of DNA is vitally important to many cellular processes, since double helices are necessarily directional [9] as illustrated in Fig. 3.

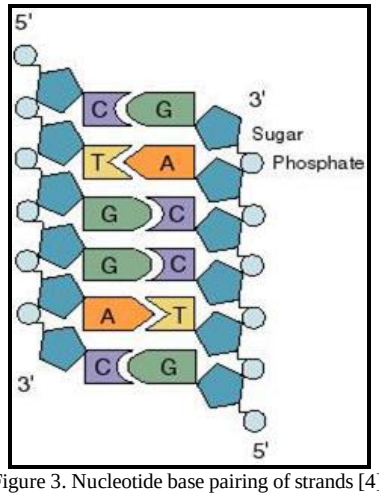


Figure 3. Nucleotide base pairing of strands [4].

The sequence of the DNA bases specifies the information for building or configuring an organism, similar to the way the letters of the alphabet appear in a specific order to form words and sentences [12] as an example shown in Fig. 4 .

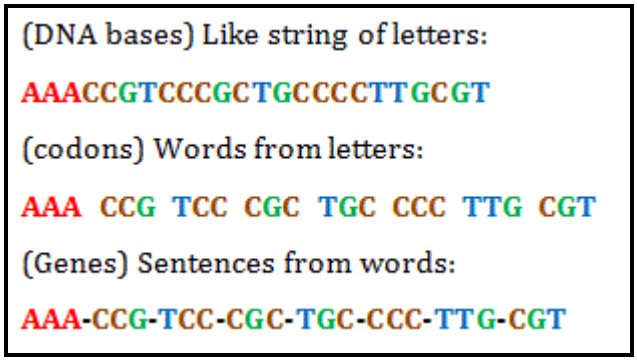


Figure 4. DNA sequence terminology.

Genes are the segments from DNA that contain the genetic information. The other segments from DNA have structural purposes and they are contributed in modifying the use of the genetic information. As Known, a string of binary data is encoded with ones and zeros, also DNA strands is encoded with four bases and are represented by letters AGTC[10].

DNA computing has many advantages over the silicon machines. These advantages mainly include the size, high parallelism and speed of computation. One gram of DNA contains 1021 DNA bases which can store about 108 Tera bytes of data. It is obviously exceeds the capacity of traditional storage media such as magnetic media, optical, electronic etc.. Every molecule of DNA can act as a single processor thus providing parallelism. Operations can thus be done in parallel, which increases the speed of computations. At the same time, DNA is highly energy efficient i.e., 1019 operations per Joule[5].

The DNA sequence ACGT has 4!=24 possible pattern each of them has different numeric encoding format (e.g.,0123 for ACGT, 0132 for ACTG, 0213 for AGCT, etc.), and consequently each encoding format will have different binary representation [4].

III. DNA CRYPTOGRAPHY

DNA cryptography is a process of securely hiding data in DNA sequences, it is explain the use of DNA as an information carrier also explain the use of modern biotechnology as a measure to convert plaintext to cipher text. Thus, biotechnology plays an essential role in the field of DNA cryptography[4].

The area of DNA cryptography is rather an untouched one. Over the years, many initiatives have been proposed to explore the process of DNA cryptography, but few have been implemented[5]. The cryptography and molecular biology have originally irrelevant relation, but with deep study of modern biotechnology and DNA computing, the relation became more closely. DNA cryptography and information science was discovered after research of DNA computing by Adleman. Many scholars around the world have done a big number of studies on DNA cryptography[4].

The major development in the area of cryptography is the establishment of DNA computing to the traditional

cryptographic technique[5]. Currently, DNA cryptography is not more influential than traditional cryptography but it can give a hybrid security by combining DNA cryptography with traditional cryptography[9]. DNA cryptography based on conventional cryptographic consists of key generation, encryption and decryption process[14]. During the last two decades, many DNA based algorithms have been developed and used for data cryptography and cryptographic key generation [10]. DNA can be used efficiently in computing and cryptography by replacing silicon chip with DNA chips or bio-chips in future, the comparison between DNA cryptography and traditional cryptography has tabulated in Table I [15].

TABLE I: comparison between dna cryptography and traditional cryptography [15]

Properties	The Traditional Cryptography	The DNA cryptography
System based	Silicon chips-based	DNA biochips-based
Performance Dependency	Implementation and system configuration	Environmental conditions
Processing Time	Less	High
Information Storage	Silicon computer chips	DNA strands
Storage Capacity	1 gm silicon chip carries 16 Mega Bytes	1 gm of DNA contains 10^8 Tera Bytes

IV. BIOLOGICAL OPERATIONS

DNA cryptography make use of biological operations, arithmetic operations or both. Arithmetic operations are used to manipulate the data entered the cryptography system. Addition, subtraction, complement, XOR, substitution, insertion, random number etc. are the most commonly used. The biological operations on DNA are used to solve computational and mathematical difficulties[5]. These operations are as follows:

A. Hybridization (Anneal)

In this process single stranded DNA chain or sequence are combined with other single stranded DNA to form double stranded DNA, where they are complementary strands[16] as shown in Fig. 2.

B. Transcription

Transcription is when two DNA strands in double stranded DNA are separated by an enzyme, the separation forms a single strand messenger RNA by mapping from DNA sequences. RNA which consist of (U, A, G, C) is complementary to DNA strand which consist of (T, A, G, C). The transcription processes noncoding segments called introns which are removed by splicing and the remaining segments called exons, that encode information for protein, synthesis and assembled in mRNA [17].

RNA is a biological molecule term, referred to Ribonucleic Acid, which contain the nucleotides C, A, G, and U. DNA is different from RNA in the nucleotides where Thymine(T) nucleotide in DNA is changed to Uracil(U) nucleotide in RNA. RNA has two types, the mRNA (Messenger Ribonucleic Acid) and tRNA (Transfer Ribonucleic Acid). In this paper we used mRNA type[18].

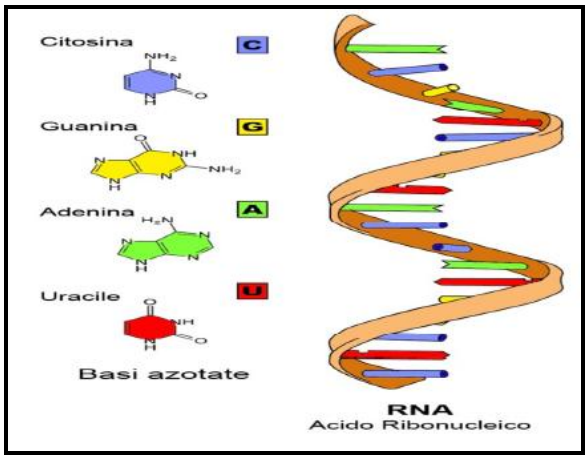


Figure 5. Structure of RNA [19].

Fig. 6 illustrates the simple concept of transcription where Thymine in DNA get replaced by Uracil to form mRNA.

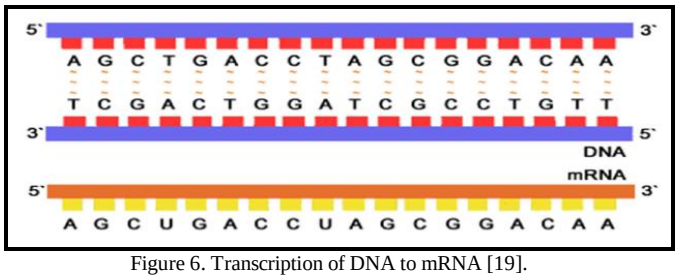


Figure 6. Transcription of DNA to mRNA [19].

C. Translation

Central Dogma of Molecular Biology is an operation of converting DNA molecules into protein sequence. Genetic code is made up of three letter codes and called codons, where DNA and RNA have these codons. To convert DNA to protein we have two stages Transcription and Translation. First stage transcription is as mentioned in previous section, which convert DNA molecules to mRNA. Second stage is translation which convert mRNA sequence to protein form[18]. Fig. 7 shows the process of DNA Central Dogma of Molecular Biology and an example of a protein construction through transcription and translation, which converts to amino acids to construct a protein cell. Only one strand of DNA is copied. A single gene may be transcribed thousands of times. After transcription, the DNA strands re-join to form amino acids and subsequently protein.

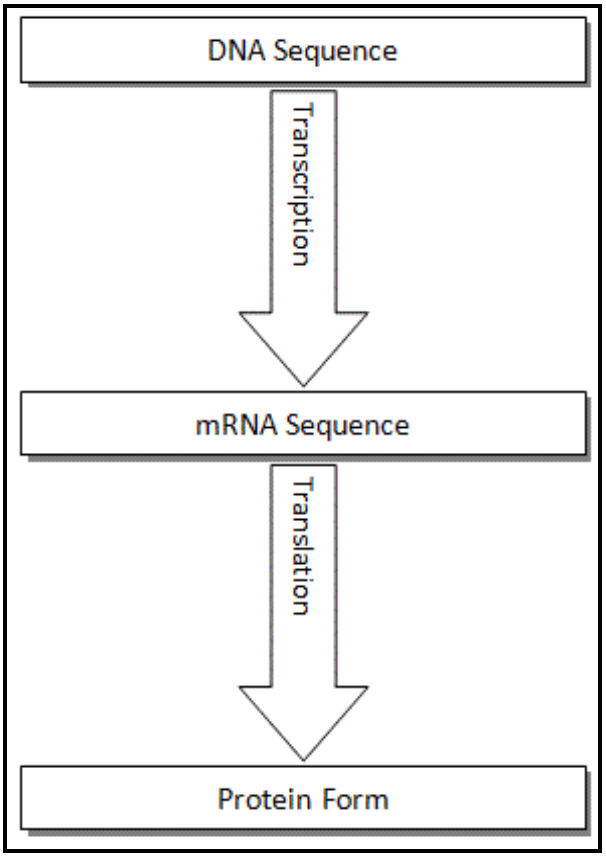


Figure 7. The process of DNA Central Dogma.

All genetic information are held in DNA, which are needed for an organism to have a form and perform functions. Genetic information encoded by the order of DNA nucleotides. Three nucleotide bases is called Codon, the combinations of different Codon make amino acid, and combination of different amino acids make protein[13]. So the order of codons is important to create a protein as shown in Fig. 8 below.

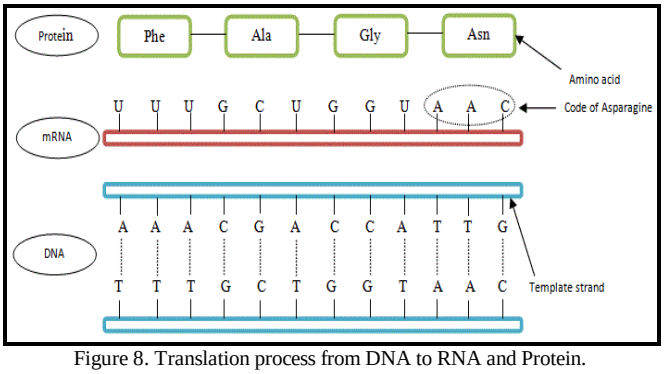


Figure 8. Translation process from DNA to RNA and Protein.

There are some instance in RNA codons represents same amino acid as shown in table II, which is contain 64 RNA codons, there are 3 stop codons and 61 RNA codons for a specific amino acid. Clearly many different RNA codons which generate the same amino acid, therefore only 20 amino acids are made by 61 RNA codons[16].

TABLE II: AMINO ACID REPRESENTATION OF RNA CODONS

First Nucleotide	Second Nucleotide								Third Nucleotide
	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys	C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop	A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg	C
	CUA	Leu	CCA	Pro	CAA	Gin	CGA	Arg	A
	CUG	Leu	CCG	Pro	CAG	Gin	CGG	Arg	G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser	C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg	A
	AUG	Start	ACG	Thr	AAG	Lys	AGG	Arg	G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly	C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly	A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly	G

GUU, GUC, GUA and GUG within the table above all represent the amino acid 'Valine' with representing three letters 'Val'. Substitution of the last letter will not make any changes to genetic information or in the protein chain. However changing the first letter of GUU to A will would create amino acid 'Isoleucine'. This would result in malfunctioning of the represented protein, this chemical reaction of DNA sequences sometimes has no effects but this can be valuable and help some problem theories.

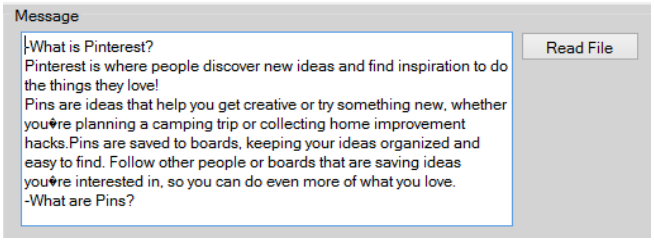
V. PROPOSED ALGORITHM

The proposed algorithm have been implemented using Visual C# on laptop Windows 8.1 Pro, Intel® Core™ 2 Duo CPU, with speed 2.2 GHz and RAM 2 GB. In our proposed Algorithm, we use OTP (One Time Pad) key generation. OTP is randomly generated key which is used only one time for encryption and decryption, that is changed in another encryption and decryption. Also XOR operation is used as an encryption and decryption technique.

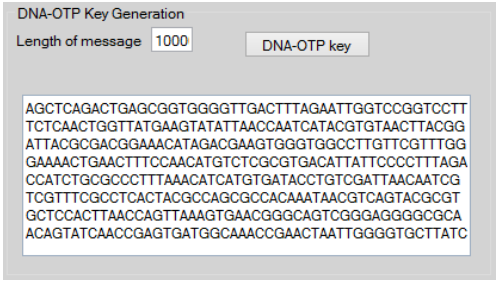
A. Key Generation

The generated OTP key is represented by DNA sequence of nucleotides named as DNA-OTP key. The steps of DNA-OTP key generation process is shown in Fig. 9, also Fig. 10 illustrate the flowchart of key generation as follows :

1. Read the text input message, as shown in Fig. 9-a.
2. Compute the length of message (L) and randomly generate DNA nucleotide sequence equal to the length of message, as shown in Fig. 9-b.
3. Apply the process of annealing to DNA-OTP key to create double helix DNA-OTP key by using base pairing rule to pair A to T and C to G, as shown in, Fig. 9-c.
4. Apply the process of transcription to convert double helix DNA-OTP key to mRNA sequence by converting every T in DNA-OTP key to U in mRNA, as shown in Fig. 9-d.
5. Final step, apply the process of translation to mRNA codons using Table II to create protein key, as shown in Fig. 9 e.

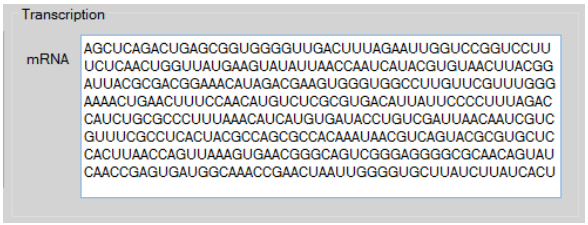


a. Reading the message.



b. DNA-OTP key generation

c. Annealing process.



d. Transcription process.

e. Translation process.

Figure 9-(a, b, c, d, e) The DNA-OTP key generation.

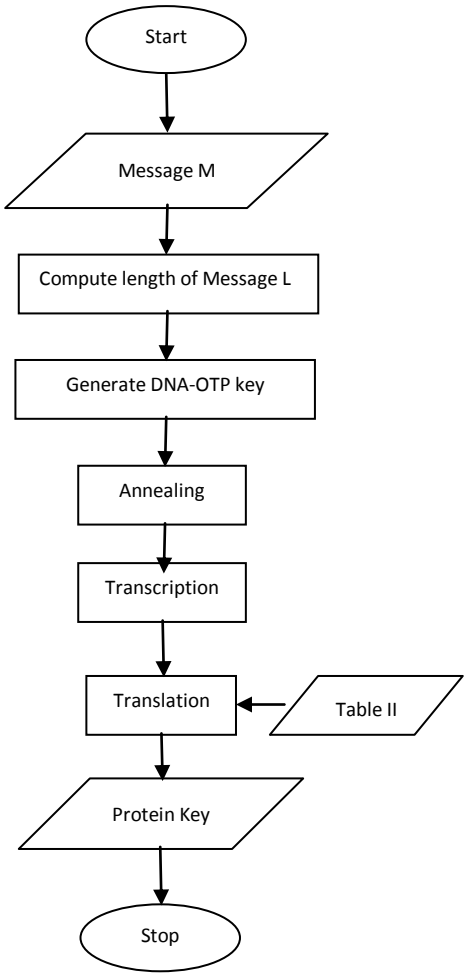


Figure 10. Flow chart of DNA-OTP key generation.

B. Encryption

The proposed algorithm use the arithmetic operation XOR as an encryption technique. The steps of encryption process is performed as shown in Fig. 11, also Fig .12 illustrate the flowchart of encryption system as follows :

1. The message is converted to ASCII code then converted to Binary, as shown in Fig .11-a.

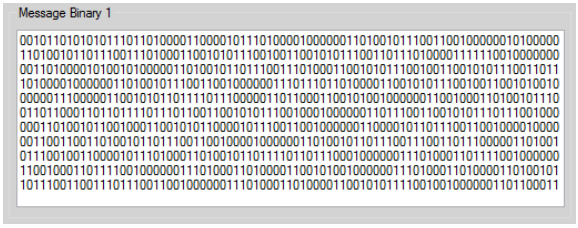
2. The generated protein key is converted to ASCII code then converted to Binary as shown in Fig .11-b.
3. Apply XOR operation between the binary of the message and the binary of the protein key, as shown in Fig. 11-c and described by equation below, where the protein key is K, the message is M and the resulted cipher text is C.

$$M \oplus K = C \quad (1)$$

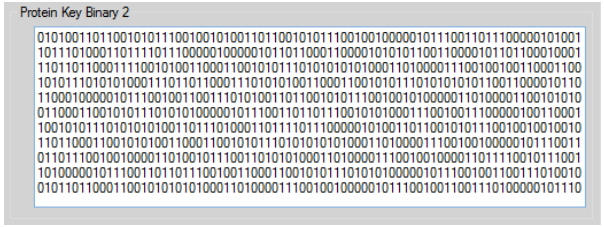
4. Finally, cipher text in Binary format is converted to DNA format by using Table III, as shown in Fig. 11-d.

TABLE III: BIT ENCODING INTO DNA NUCLEOTIDES

Bits	Nucleotide
00	A
01	C
10	G
11	T



a. Binary format of Message.



b. Binary format of protein keys.

c. XOR operation.

d. Cipher text.

Figure 11-(a, b, c, d) Encryption Process.

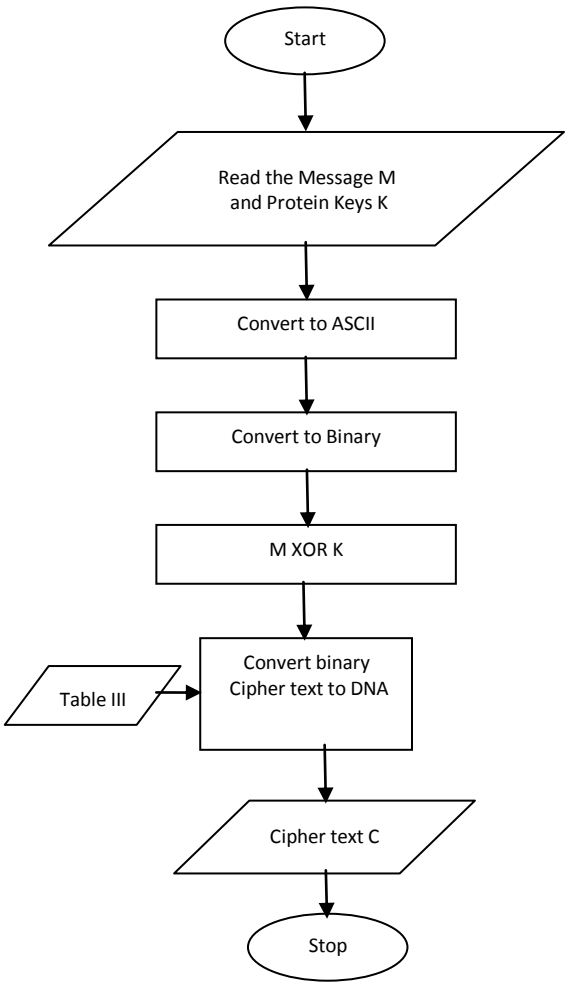


Figure 12. Flow char of Encryption System

C. Decryption

The message can be recovered by using the arithmetic operation XOR as a decryption technique. XOR operation is used because if XOR operation applied twice, the result is the message again e.g., XOR[K, XOR(K,M)]=M, where K is the protein key and M is the message. In this example, the inner XOR operation represent encryption, where the decryption is the outer XOR operation. Thus, XOR operation can be applied for both encryption and decryption . Fig. 13 illustrate the flowchart of decryption system. The steps of decryption process is as the following:

1. The cipher text in DNA format is converted to Binary format by using Table III.
2. The generated protein keys is converted to ASCII code then converted to Binary.
3. Apply XOR operation between the binary of the cipher text and the binary of the protein key, as described by equation below, where the protein key is K, the cipher text is C and the resulted message is M.

$$C \oplus K = M \quad (2)$$

4. Finally, the message in binary format is converted into ASCII code then to text message.

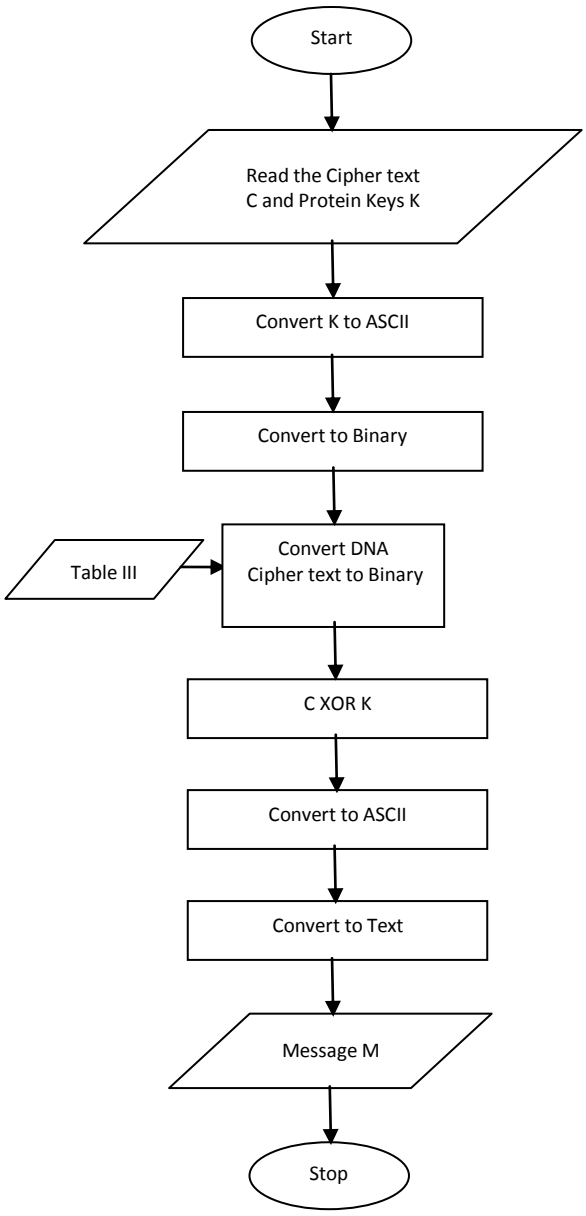


Figure 13. Flow chart of Decryption System.

.

VI. ALGORITHM ANALYSIS

A. Randomness

The NIST(National Institute of Standard and Technology) test suite is used to test the randomness of protein keys generated in the proposed algorithm and the randomness of the cipher text. The NIST Test Suite is a statistical package consisting of 15 tests that were developed to test the randomness of (arbitrarily long) binary sequences produced by either hardware or software based cryptographic random or pseudorandom number generators. These tests focus on a variety of different types of non-randomness that could exist in a sequence. Some tests are decomposable into a variety of subtests.

These tests depend on the probability value(P-value) to determine the randomness of the sequence. If the computed P-value is less than 0.01 then conclude that the sequence is non-random, otherwise conclude the sequence is random. It should be noted that some tests (Runs, Random Excursions and Random Excursions Variant) are not always applicable, because these tests are applied only if the sequence meets certain criteria. The randomness test of the protein key is tabulated as shown in Table IV.

TABLE IV: RANDOMNESS TEST OF PROTEIN KEY

No.	Test Type	P-value	State
1	Frequency	*0.739918*	Pass
2	Block Frequency	*0.035174*	Pass
3	Linear Complexity	*0.350485*	Pass
4	Longest Run	*0.534146*	Pass
5	Rank	*0.911413*	Pass
6	FFT	*0.122325*	Pass
7	Non Overlapping	*0.213309*	Pass
8	Overlapping Template	*0.350485*	Pass
9	Maurer's Test	*0.350485*	Pass
10	Runs	*0.739918*	Pass
11	Serial Test	*0.739918*	Pass
12	Approximate Entropy	*0.739918*	Pass
13	Cumulative Sums	*0.122325*	Pass
14	Random Excursions	--------	Not Applicable
15	Random Excursions Variant	--------	Not Applicable

B. Execution Time

The performance of the cryptographic algorithm can be measured by analyzing the encryption time and decryption time. The performance of the proposed algorithm has been compared with Triple Data Encryption algorithm(Triple DES). Encryption and decryption time is computed with two file size and the result is tabulated as shown in Table VI .The first text file with size of 157 Byte. The second text file with size of 2,549 Byte.

TABLE VI: COMPARISON BETWEEN PROPOSED ALGORITHM AND TRIPLE DES ALGORITHM

Algorithm	Text file size(Byte)	Cipher size(Byte)	Encryption time(ms)	Decryption time(ms)
Proposed Algorithm	4,327	36,036	12958.858	14010.0687
	6,457	53,512	35287.2749	36798.0848
Triple DES	4,327	13,319	289.5198	5.1738
	6,457	19,559	579.684	9.7612

VII. CONCLUSION

The proposed algorithm is implemented using the

biological operations on DNA-OTP sequence. It should be mentioned that the biological operations occur in laboratories and the proposed algorithm is a simulation process using a digital environment that cannot achieve the required properties in the biological environment such as storage and speed, that occur in laboratories. In the proposed algorithm the key size depends upon user's message by creating a protein key using translation operation, which makes cryptanalysis even harder. Encryption is implemented using XOR operation between the protein key and the message. Future work might consist of analyzing and comparing the performance of all the DNA cryptographic techniques based on secure data transmission processes. Also more biological operations can be used and computation techniques can be contained for improving the proposed algorithm.

REFERENCES

[1] M. Zhang, M. X. Cheng and T. J. Tarn, "A mathematical formulation of DNA computation," *IEEE Transactions on Nano Bio science*, vol.5, no. 1, pp. 32-40, 2006.
[2] Michael N. Leuenberger and Daniel Loss, "Quantum computing in molecular magnets," *Nature*, vol. 410, pages 789–793, 12 April 2001.
[3] P. Saxena, A. Singh and S. Lalwani, "Use of DNA for computation, storage and cryptography of information," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 3, issue 2, 2013.
[4] P.Surendra Varma and K.Govinda Raju, "Cryptography Based on DNA Using Random key Generation Scheme," *International Journal of Science Engineering and Advance Technology (IJSEAT)*, vol.2, issue 7, July – 2014.
[5] Nileena Ouseph, "A Survey on Diverse DNA Cryptographic Techniques," *International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE)*, vol. 4, issue 9, Sep. 2016.
[6] L. Adleman, "Molecular computation of solutions to combinatorial problems," *Science, JSTOR*, vol. 266, 1994.
[7] Sanjeev Dhawan and Alisha Saini, "Secure Data Transmission Techniques Based on DNA Cryptography," *International Journal of Emerging Technologies in Computational and Applied Sciences*, pp. 95-100, Aug. 2012.
[8] S. V Kartalopoulos, "DNA-Inspired cryptographic method in optical communications, authentication and data mimicking," *Milcom 2005 - 2005 IEEE Military Communications Conference, Atlantic City, NJ*, vol. 2, pp. 774-779, 2005.
[9] S. Jeevidha, Dr. M. S. Saleem Basha and Dr.P. Dhavachelvan, "Analysis on DNA based Cryptography to Secure Data Transmission," *IJCA*, vol. 29, no.8, Sep.2011.
[10] Bonny B Raj and Panchami, "DNA Based Cryptography Using Permutation and Random Key Generation Method," *International Conference On Innovations & Advances In Science, Engineering And Technology [IC - IASET 2014]*, vol.3, Special Issue 5, July 2014.
[11] Tausif Anwar, Sanchita Paul and Shailendra Kumar Singh, "Message Transmission Based on DNA Cryptography: Review," *International Journal of Bio-Science and Bio-Technology*, vol. 6, no.5, Issue 30, Oct.2014.
[12] Genetic home reference, a service of the U.S. National Library of Medicine, cited on: *https://ghr.nlm.nih.gov/primer/basics/dna*
[13] Needleman, S. B., "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *J. Mol. Biol.* , 1970.
[14] Tausif Anwar, Abhishek Kumar and Sanchita Paul, "DNA Cryptography Based on Symmetric Key Exchange," *International Journal of Engineering and Technology (IJET)*, vol.7, no.3, Jun-Jul 2015.
[15] B. Anam, K. Sakib, Md. A. Hossain and K. Dahal, "Review on the Advancements of DNA cryptography," *ARXIV*, 1 Oct. 2010.
[16] Mazhar Karimi, Waleej Haider, "Cryptography using DNA Nucleotides," *International Journal of Computer Applications*, vol. 168, no.7, June.2017.
[17] S. Lloyd and Q. O. Snell, "Sequence Alignment with Trace back on Reconfigurable Hardware," *2008 Int. Conf. Reconfigurable Computer. FPGAs*, pp. 259–264, Dec. 2008.
[18] B.Murali Krishna, Habibulla Khan, G.L.Madhumati, K.Praveen Kumar, G.Tejaswini, M.Srikanth and P.Ravali, "FPGA Implementation of DES algorithm using DNA cryptography," *Journal of Theoretical and Applied Information Technology*, vol.95, no. 10, 31 May 2017.
[19] Tushar Mandge, Vijay Choudhary, "A Review on Emerging Cryptography Technique: DNA Cryptography," *International Conference in Recent Trends in Information Technology and Computer Science (ICRTITCS)*, 2012.

Sally Safaa Nafea holds B.Sc degree in "Computer Engineering" from "University of Technology, Iraq". Currently she's is doing M.Sc in "Information and Communication Engineering" in "Al-Nahrain University , Iraq" .

Prof. Dr. Mahmood Khalel Ibrahem Prof. in "Collage of Information Engineering, Al-Nahrain University , Iraq" .