Multi-Layer Data Encryption Using Residue Number System in DNA Sequence

M. I. Youssef, A. Emam and M. Abd Elghany

Faculty of Engineering, Department of Electrical engineering
Al-Azhar University, Egypt

a_emamm@yahoo.com, mohamedgheth@yahoo.com

Abstract

In this paper, we will merge between the usages of DNA sequences and Residue number system in encryption systems. The message which is coded will be secretly impeded inside the DNA sequence. This merge will be leaded to perform multilayer encryption with different keys - that can be used as a hash function - versatile alternatively to increase the security and more flexibility, with less complexity. As the security is one of the most important issues in communication systems, the evolvement of cryptography and cryptographic analysis are considered as the fields of ongoing research. This field is becoming very promising. Thus, a straight forward algorithm that achieves efficiency as multi-layer encryption techniques are implemented.

Keywords: *DNA, Encryption, and Residue number system*

1. Introduction

In recent years, much research work has been done on DNA based encryption schemes [1 - 3]. A DNA sequence is a sequence consisting of four alphabets: A, C, G and T. Each alphabet is related to a nucleotide. It is usually quite long. For instance, the DNA sequence of "Litmus", its real length is with 2856 nucleotides long:

ATCGAATTCGCGCTGAGTCACAATTCGCGCTGAGTCACAATTCGCGCTGAGTC
ACAATTGTGACTCAGCCGCGAATTCCTGCAGCCCCGAATTCCGCATTGCAGAGAT
AATTGTATTTAAGTGCCTATCGATACAATAAACGCCATTTGACCATTCACCACATT
GGTGTGCACCTCCAAGCTCGCGCACCGTACCGTCTCGAGGAATTCCTGCAGGATA
TCTGGATCCACGAAGCTTCCCATGGTGACGTCAC [4].

From this sequence a several useful properties could be shown:

a) There is almost no difference between a real DNA sequence and a faked one.

b) There are a large number of DNA sequences publicly available in various web-sites [4]. A rough estimation would put the number of DNA sequences publicly available to be around 55 million [4].

By using the above facts, we designed a DNA based encryption methods. This method would secretly select a reference sequence S from publicly available DNA sequences. Only the sender and the receiver are aware of this reference sequence. The sender would transform this selected DNA sequence S into a new sequence S' by incorporating the DNA sequence S with the secret message M. This transformed sequence S' is sent by a sender to the receiver together with many other DNA sequences. The receiver would then examine all of the received sequences, identify S' and recover the secret message M.

We shall introduce a method in the section 4. It is assumed that the secret message M is a binary sequence and the binary coding scheme which transforms alphabets A, C, G and T into binary codes and vice versa is described in Table 1.

Table 1

Alphabet	Binary representation
A	00
C	01
G	10
T	11

We shall propose to convert the plain text message to residue number system with different moduli, which are used to convert from binary to RNS. In addition the moduli order (arrangement) is kept secret to increase the security issues.

2. Bimolecular Technology Backgrounds

DNA, the major support of genetic information (genetic blueprint) of any organism in the biosphere, is composed of two long strands of nucleotides, each containing one of four bases (A – adenine, C – cytosine, G – guanine, T – thymine), a deoxyribose sugar and a phosphate group. The DNA strands have chemical polarity, meaning that on each end of a molecule there are different groups (5' – top end and 3' – bottom end) [5].

A DNA molecule has double-stranded structure obtained by two single-stranded DNA chains, bonded together by hydrogen bonds: A = T double bond and C ≡ G triple bond. The double helix structure is configured by two single antiparallel strands (Figure 1). The DNA strands that bond each other through A-T and C-G bonds are known as complementary strands.

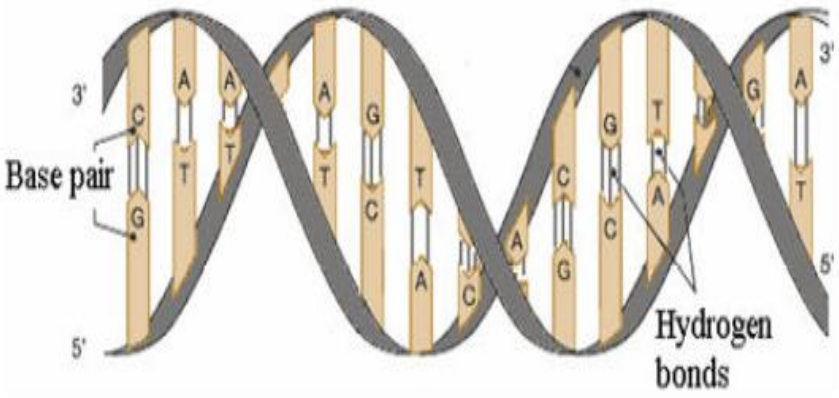

Figure 1. DNA Structure

The DNA strands can be chemically synthesized using a machine, known as DNA synthesizer. The single-stranded chains obtained artificially with the DNA synthesizer are named oligonucleotides having usually 50-100 nucleotides in length.

3. Residue Number System (RNS)

A residue number system (RNS) [6 - 8] represents a large integer using a set of smaller integers, so that computation may be performed more efficiently. It relies on the Chinese remainder theorem (CRT) [8] of modular arithmetic for its operation, a mathematical idea from Sun Tsu Suan-Ching (Master Sun's Arithmetic Manual) in the 4^{th} century AD.

The residue number system is defined by the choice of *v* positive integers m_i (i = 1, 2, 3 … v) referred to as moduli. If all the moduli are pair-wise relative primes, any integer N, describing a non-binary message in this letter, can be uniquely and unambiguously represented by the so-called residue sequence (r_1, r_2 ..r_v) in the range 0<N<MI ,where $r_i = N \pmod{m_i}$ represents the residue digit of N upon division by m_i, and $M_I = \Pi\ m_i$ is the information symbols' dynamic range. Conversely, according to the Chinese Reminder Theorem, for any given *v*-tuple (r_1, r_2 ..r_v) where $0 \leq r_i < m_i$; there exists one and only one integer N such that $0 \leq N < M_i$ and $r_i = N \pmod{mi}$ which allows us to recover the message N from the received residue digits.

Residue number system has two inherent features that render the RNS attractive in comparison to conventional weighted number systems, such as for example the binary representation. These two features are [7]: The carry-free arithmetic and Lack of ordered significance amongst the residue digits.

The first property implies that the operations related to the individual residue digits of different moduli are mutually independent because of the absence of carry information. The second property of the RNS arithmetic implies that some of the residue digits can be discarded without affecting the result, provided that a sufficiently "high dynamic range" is retained in the "reduced" system in order to unambiguously contain the result.

4. Proposed Encryption Scheme

In this section varies DNA encryption schemes are utilized together with residue number system in order to provide a more secure and flexible encryption system. Two main methods are used; insertion and complementary pair approach methods. In the following the two methods are explained.

4.1 Insertion Method

Starting with the simplest approach called the insertion approach. Suppose the secret 16-bit message M is 1010010101101100. Let S be TTCATAGCACGGATTATCGGAGTTTCGTAT.

The coding steps are as follows:

1) Using a secret RNS system, convert the message to another binary form M'. So, for a RNS moduli [15 13 11 8 7], the modified message becomes: 0011 0111 1001 0100 0101.
2) Choose the segmentation scheme for the DNA code. Suppose k is 3.
3) Depending on the length of the modified message M' and the selected segmentation scheme k, the sequence S length are selected. In this example, the length of S is 60 bits: 111101001100100100011010001111001101101000101111110110110011.
4) Divide S into segments whereby each segment contains k bits. Then we have the following segments: 111, 101, 001, 100, 100, 100, 011, 010, 001, 111, 001, 101, 101, 000, 101, 111. 110, 110, 110, 011, 110, 110, 110, 011
5) Insert bits from M', once at a time, into the beginning of segments of S. The result is as follows: 01 11, 0101, 1001, 1100, 0100, 1100, 1011, 1010, 1001, 1111, 0001, 0101, 1101, 0000, 1101, 0111, 0110, 0011, 0110, 1110, 0110, 1011. We should ignore those segments without any secret message inserted. Thus, concatenating the above segments, we have the following binary sequence: 01 11 01 01 10 01 11 00 01 00 11 00 10 11 10 10 10 01 11 11 00 01 01 01 11 01 00 00 11 01 01 11 01 10 00 11 01 10 11 10 01 10 10 11

   We use the binary code scheme to produce the following faked DNA sequence:

   S'=GTCCGCTACATAGTGGGCTTACCCTCAATCCTCGATCGTGCGGT

   It is shown that this sequence is quite different from S.
6) We send the above sequence S' to the receiver.

This proposed scheme has two layers of encryption as seen in figure 2, first conversion of the message data to RNS system, and the second is the insertion of the RNS message secretly in the DNA sequence.

Each one of these two layers has its own security levels, where in the first step the conversion to RNS, the security is implemented in the number of modules used, values of the selected moduli and in the order of these moduli's. While in the second step were the faked DNA sequence is generated, the security is achieved in the DNA code selected, the located of the inserted bits inside the DNA code, and finally in the segmentation used.

Figure 2. Implemented Security Layers Using Insertion Method

4.2 Complementary Pair Approach

Suppose the secret message M is 0110. Let S be ACGGTCGTTCCCTAGTTG.
The coding steps are as follows:

1) Using a secret RNS system (15, 13, 11, 8, 7), convert the message to another binary form M'.
2) Generate a DNA sequence consisting of A, C, G and T only. Assume that the sequence is L=ACGGTCTCATCAATGCTTCAGT.
3) Divide M into segments such that each segment contains even number of bits. Thus, in our case, we have 01 and 10.
4) Generate set of complementary strings with length k and insert them into L. The number of complementary strings depends on the message size. Assume k=5 and we have the following two complementary strings: (AAGCT TTCAG) and (ACCTG TAAGC). The sequence L now becomes L'= ACG AAGCT GTCT TTCAG CAT ACCTG CAAT TAAGC GCTTCAGT.
5) Insert the first (second) alphabet of the secret message one alphabet before the first (second) complementary string. The string becomes: L'= ACCG AAGCT GTCT TTCAG CAGT ACCTG CAAT TAAGC GCTTCAGT.\
6) Use a random number generator to select two positive integers j and i. Assume j=2 and i=4. Insert substrings. S[j,j+i]=CGGTC and S[2j,2j+i]=GTCTC one alphabet

after the first and second complementary substrings. L' becomes: L"=ACCGA AGCTGTCTTTCAGCCGGTCAGTACCTGCAATTAAGCGGTCTCCTTCAGT.

7) We send the above sequence L' to the receiver.

In this scheme we have four layers of encryption as seen in Figure 3, first conversion of the message data to RNS system, the second is the generation of the complementary pairs in the DNA code, the third is the insertion of the RNS message secretly in the DNA sequence and the fourth is the usage of a random number generator that is also added to the DNA code to generate the transmitted faked DNA code.

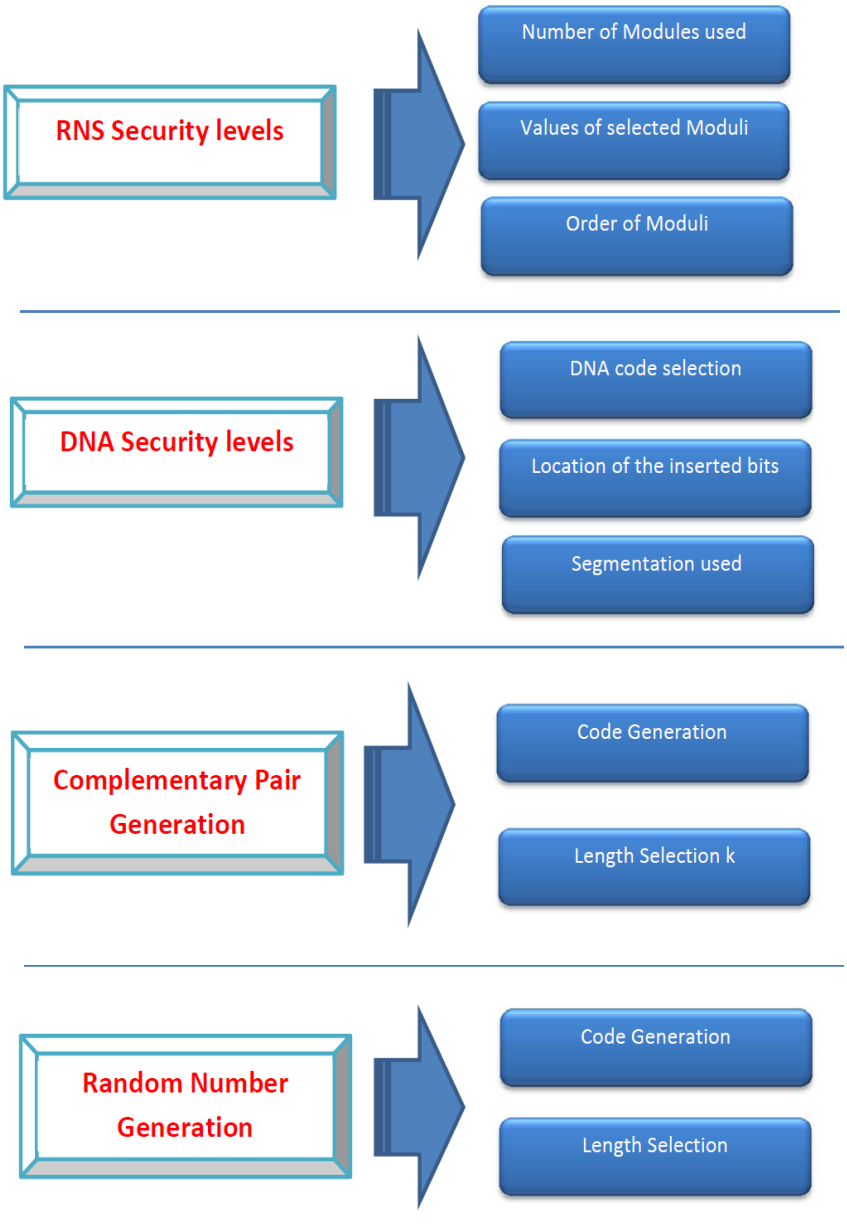

Figure 3. Implemented Security Layers using the Complementary Approach

5. System Model

In this section, a basic transmission system as shown in Figure 4, is proposed and analyzed when the system is designed with and without RNS.

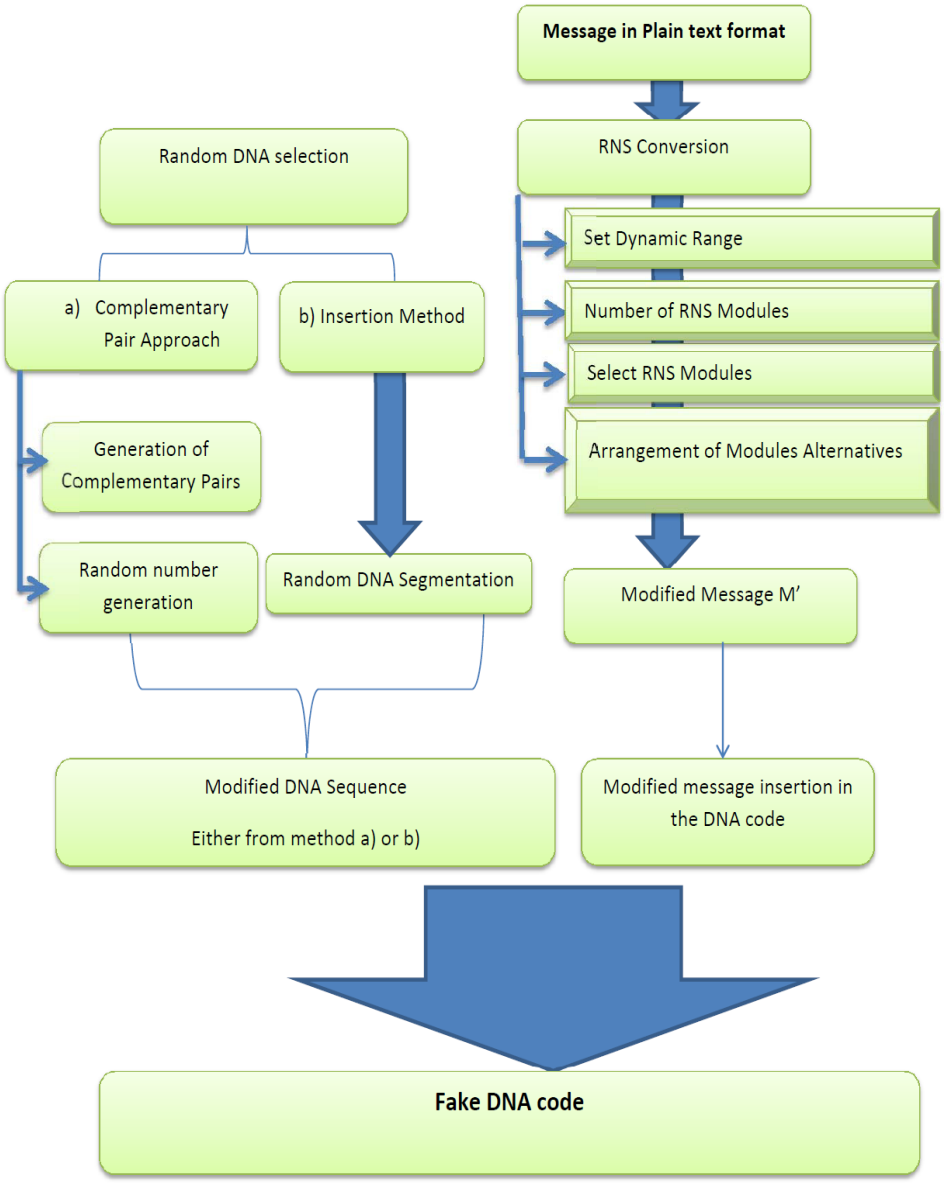

Figure 4. DNA Encryption System

6. Simulation Results

Using a 16-bit message various simulations were performed on the transmission system for both insertion and complementary approach methods, with and without RNS. Both the cross correlation and autocorrelation properties are measured and compared as shown in the following.

6.1 RNS Performance

RNS modules [15 13 11 8 7] are selected. The cross correlation properties for different Tx systems are measured. The results are shown in Figures 5, 6, 7, 8 and summarized in Table 2.

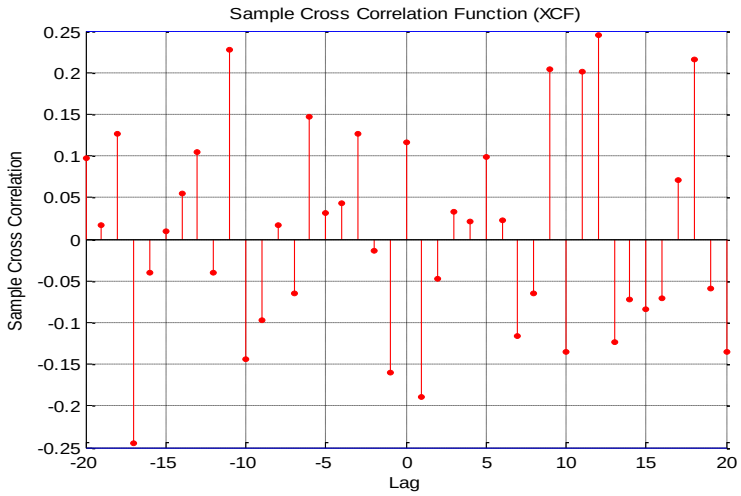

Figure 5. Cross Correlation for Encrypted System using Insertion Method without RNS

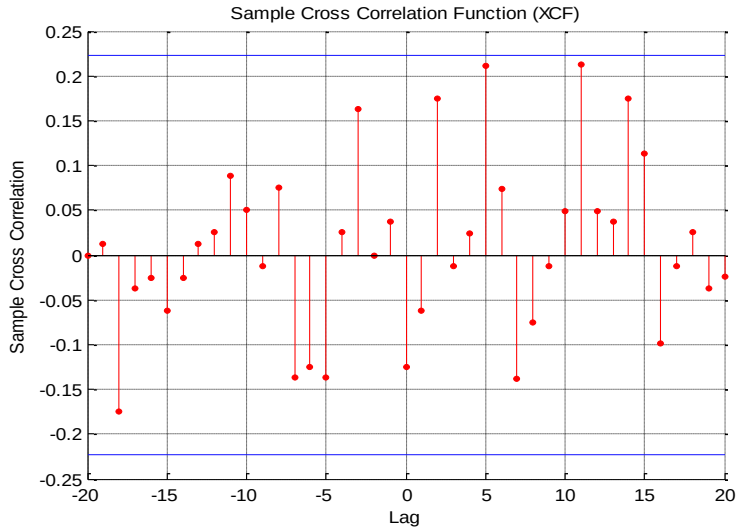

Figure 6. Cross Correlation for Encrypted System using Insertion Method with RNS

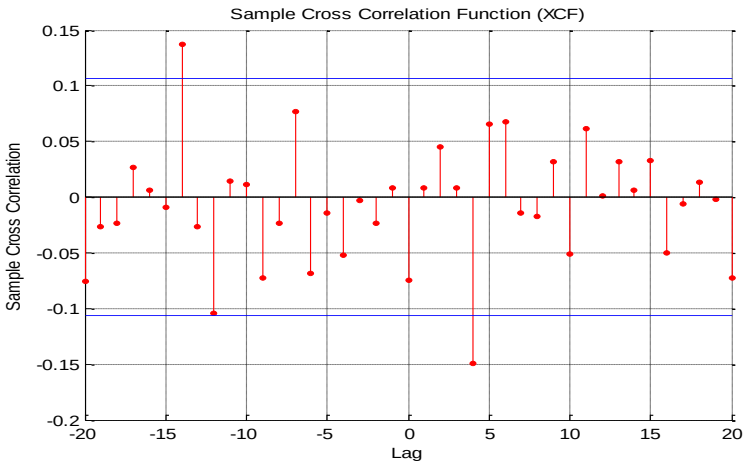

Figure 7. Cross Correlation for Encrypted System using Complementary Pair Method without RNS

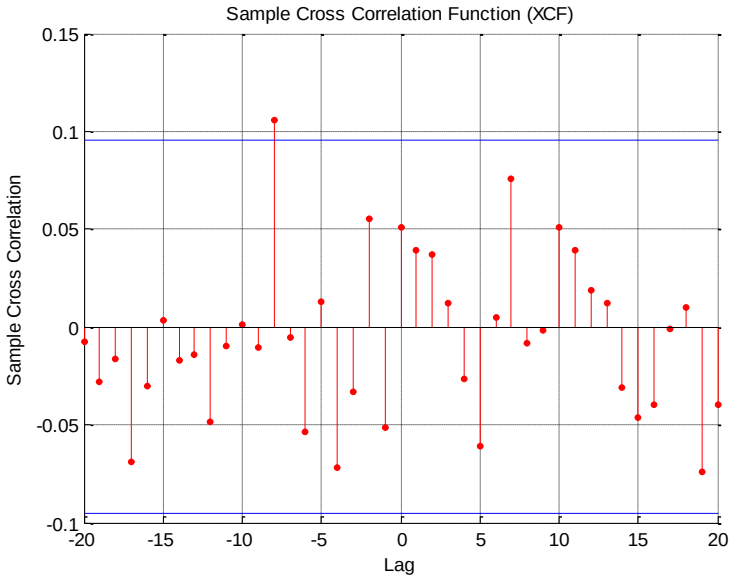

Figure 8. Cross Correlation for Encrypted System using Complementary Pair Method with RNS

The cross correlation property could be summarized in the next table;

Table 2

Max Cross-correlation property		
DNA Method	Without RNS	With RNS
Insertion method	0.2448	0.212
Complementary Pair approach	0.1369	0.106

As seen from Table 2, the RNS implementation improves the cross correlation property of the system and thus enhances the security aspect. Also, the utilization of the complementary pair approach improves the cross correlation property compared to the insertion method.

Then auto-correlation function for different Tx systems are measured as seen in Figure 9 and 10 and also summarized in Table 3.

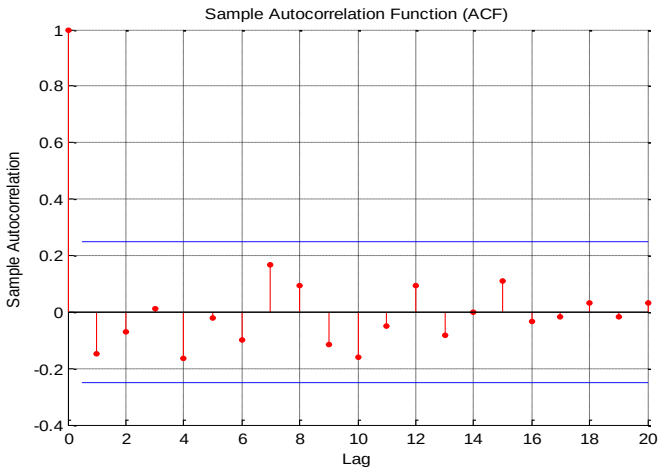

Figure 9. Auto Correlation for Encrypted System using Insertion Method without RNS

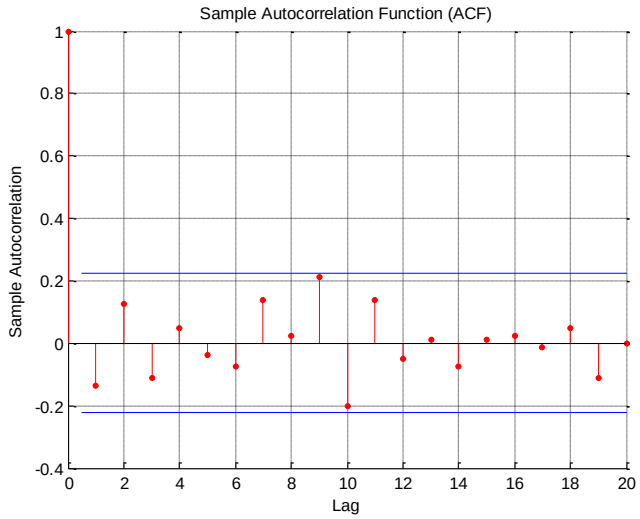

Figure 10. Auto Correlation for Encrypted System using Insertion Method, with RNS

Table 3

Mean Auto-correlation property		
DNA Method	Without RNS	With RNS
Insertion method	0.02637	0.04642
Complementary Pair approach	0.027	0.04

As seen from Table 3, the RNS implementation improves the Auto correlation property of the system and thus simplifies the reception system.

6.2 Different RNS modules Performance

It is seen from Figure 11 & 12 that changing the RNS modules changes the cross and auto correlation properties of the generated codes.

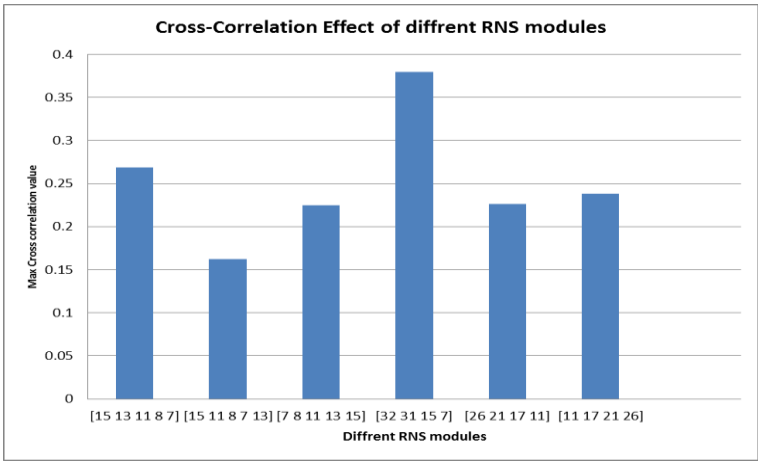

Figure 11. Cross Correlation of Different RNS Modules

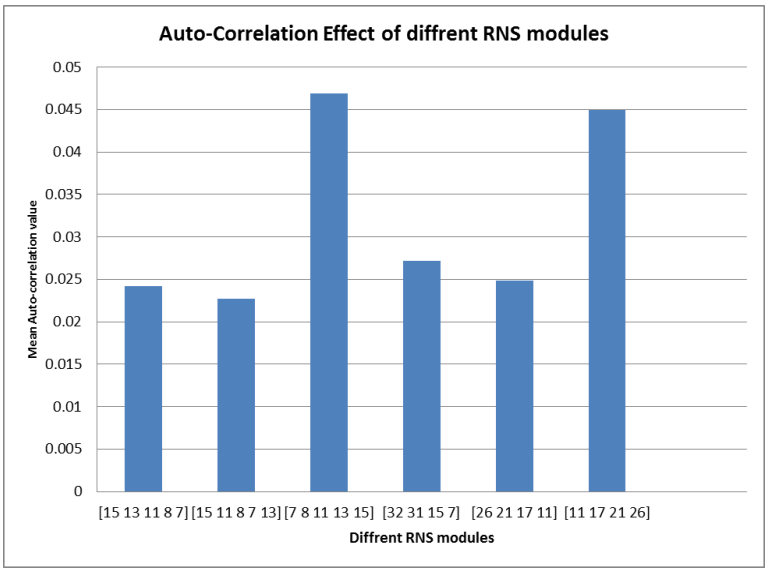

Figure 12. Auto Correlation of Different RNS Modules

6.3 Segmentation Performance

It is seen from Figure 13 that increasing the segmentation bits would improve significantly the cross-correlation property.

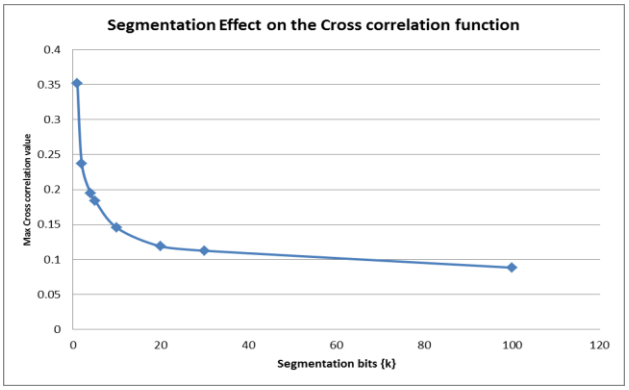

Figure 13. Segmentation Effect on Cross Correlation

7. Conclusion

In this paper, we have pointed out that the DNA sequences have the special properties which we can utilize for encryption purposes.

Two methods (insertion and complementary approaches methods) had been proposed and demonstrated, which is based upon a reference sequence known only to the sender and the receiver. This reference sequence can be selected from any web-site associated with DNA sequences. Since there are many web-sites and roughly 55 million publicly available DNA sequences, it is virtually impossible to guess this sequence.

Residue number system is merged with the DNA sequence, which added more permutations and combinations that provides more security, flexibility with less complexity.

The complementary Pair approach provides additional improvements in the cross correlation property compared to that using Insertion method.

8. Future Work

For further research, we shall investigate some mathematical properties of our approach, and also, try to insert an image as a secret data to be hidden inside the DNA sequence and see the effect on the security aspects.

References

[1] C. T. Clelland, V. Risca and C. Bancroft, "Hiding Messages in DNA Microdots", Nature, vol. 399, **(1999)**, pp. 533-534.
[2] A. Leier, C. Richter, W. Banzhaf and H. Rauhe, "Cryptography with DNA Binary Strands", BioSystems, vol. 57, **(2000)**, pp. 13-22.
[3] B. Shimanovsky, J. Feng and M. Potkonjak, "Hiding Data in DNA", Revised Paper from the 5th International Workshop on Information Hiding, Lecture Notes in Computer Science, vol. 2578, **(2002)**, pp. 373-386.
[4] European Bioinformatics Institute, URL: http://www.ebi.ac.uk.
[5] M. Schena, "Microarray Analysis", Wiley-Liss, **(2003)** July.
[6] K. W. Watson, "Self-checking computations using residue arithmetic", Proc. IEEE, vol. 54, **(1966)** December, pp. 1920–1931.
[7] E. D. D. Claudio, G. Orlandi and F. Piazza, "A systolic redundant residue arithmetic error correction circuit", IEEE Trans. Computers, vol. 42, **(1993)** April, pp. 427–432.
[8] H. Krishna and J. D. Sun, "On theory and fast algorithms for error correction in residue number system product codes", IEEE Trans. Computers, vol. 42, **(1993)** July, pp. 840–852.
[9] R. Kohno, R. Meidan and L. B. Milstein, "Spread Spectrum Access Methods for Wireless Communications", IEEE Communication magazine, **(1995)** January.
[10] M. I. Youssef, M. Zahara, A. E. Emam and M. Abd ElGhany, "Chaotic Sequences Implementations on Residue Number Spread Spectrum System", International journal of communication, vol. 2, Issue 2, **(2008)**.