Virtual DNA-Based Cryptography for Enhanced Security

[1]Toshitha Kannan, [2]M. Sindhu Madhuri

Department of Electronics and Communication Engineering, SSN College of Engineering, Kalavakkam

Abstract- DNA Cryptosystem involves using the principles of random yet structured DNA mechanisms to realise the encryption process. One such cryptographic scheme is proposed in our paper which incorporates the fundamental concept of DNA Hybridization and recombinant DNA technology. We imbibe the principles of restriction enzymes to implement a secure virtual cryptosystem with a two stage encryption scheme without the actual handling of DNA.

Keywords: DNA, Crytptosystem, Restriction enzymes

I. INTRODUCTION

Withholding the confidentiality and preventing unauthorised disclosure of information is one perspective of viewing communication theory. Cryptography is the study of mathematical techniques related to aspects of information security such as confidentiality, data integrity, entity authentication, and data origin authentication [1]. This method of protecting the message in its digital form requires two primary components - an algorithm and a key. The key required to crack the encrypted message is chosen such that it strengthens the algorithm against any attacks. Any cryptosystem can be represented by Fig.1.[2]

The message to be transmitted is encrypted into a set of symbols with the help of a key. After transmission, the receiver decrypts the plaintext to obtain the message using a key. If the same key is used for both encryption and decryption, then the cryptosystem is said to be symmetric. A breakthrough in the field of cryptography was the advent of the Public Key Encryption which uses two keys, a public key and private key. This path-breaking concept proposed by Whitfield Diffie and Martin Hellman in 1976, uses the public key for encryption and the private key for decryption where both the keys are different yet mathematically related to each other.

The success of a cryptosystem remains in its ability to minimize any errors or loss of information during transmission and its resistance to external attacks. Thus cryptanalysis is an integral component of testing a cryptographic algorithm's viability. A cryptosystem is made to deal with attacks and its response is noted. To name a few, brute force attack, cipher text only attack, known-plaintext attack, chosen-plaintext attack, chosen-cipher text attack etc are models used by cryptanalysts to deduce the part of the cipher or in a worst case, the key. Essentially the aim of every cryptosystem is to be considered a strong cryptosystem , almost impossible to break by ensuring the variability of the key.

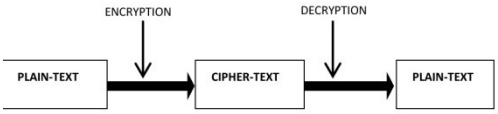

Fig. 1: Block diagram of a cryptosystem.

Deoxyribonucleic acid (DNA) [3] is a molecule that encodes the genetic instructions used in the development and functioning of all known living organisms and many viruses. DNA is a polymer comprising of four different monomers (linked together linearly in no definitive order) called nucleotides. Each nucleotide is composed of a nitrogen-containing base—Guanine (G), Adenine (A), Thymine (T), or Cytosine (C)—as well as a monosaccharide sugar called deoxyribose and a phosphate group. The nucleotides are joined to one another in a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone. According to base pairing rules (A with T and C with G), hydrogen bonds bind the nitrogenous bases of the two separate polynucleotide strands to make double-stranded DNA.

Restriction enzymes are enzymes isolated from bacteria that recognize specific sequences in DNA and then cut the DNA to produce fragments, called restriction fragments [4].

These enzymes are found naturally in bacteria to provide a defence mechanism against viruses. They recognise certain palindrome sequences known as 'restriction sites' in the foreign DNA and selectively cleave the same. At the same time, the host DNA is methylated and thus unharmed by these restriction enzymes. The whole mechanism is dubbed as 'restriction modification system.'

Restriction enzymes are of five types: type 1, type 2, type 3, type 4 and type 5. Among these, type 2 restriction enzyme alone cleaves within the restriction site. The rest cleave outside the restriction sites. Apart

from these, artificial restriction enzymes have also been synthesised. Artificial restriction enzymes can be generated by fusing a natural or engineered DNA binding domain to a nuclease domain (often the cleavage domain of the type IIS restriction enzyme FokI.) Such artificial restriction enzymes can target large DNA sites (up to 36 base pairs) and can be engineered to bind to desired DNA sequences[5]. Over 3800 restriction enzymes have been studied in detail and over 600 of them are available commercially [6].

The use of restriction enzymes as a tool for recombining, or joining, different DNA fragments led to recombinant DNA technology or rDNA technology. This gives molecular biologists powerful tools to create nearly limitless combinations of recombinant DNA. [7]

The polymerase chain reaction (PCR), invented in 1985 is used to amplify specific DNA sequences. PCR has been used in a variety of applications in the past 10 years because of the sensitivity, specificity, speed, and simplicity of the reaction,; such applications include characterising the structure and expression of genes; identification of disease-causing genes and pathogens; diagnosing inherited disease prenatally; and DNA fingerprinting in forensics, agriculture, and archaeology [8]. The PCR process utilises DNA primers for strand extension.

These primers are usually short, chemically synthesized oligonucleotides, with a length of twenty bases. A primer is a strand of nucleic acid that serves as a starting point for DNA synthesis. It is required for DNA replication because the enzymes that catalyse this process, DNA polymerases, can only add new nucleotides to an existing strand of DNA. The polymerase starts replication at the 3'-end of the primer, and copies the opposite strand. The primers are hybridized to a target DNA, which is then copied by the polymerase. The underlying principle behind PCR technique to amplify a DNA strand using primers is DNA hybridization.

DNA ligase [9] is a specific enzyme found in living cells that facilitates the joining of DNA strands by catalysing the process of formation of a phospho-di-ester bond. It is a powerful class of enzymes used extensively in molecular biology, once it is extracted and purified, and is used in recombinant DNA technology to adhere the recombinant DNA molecule on to a vector or host DNA, thereby creating the rDNA molecule.

This idea of recombinant DNA technology based on use of restriction enzymes is the main principle behind the suggested crypto system in this paper. While the encryption employs the principle of restriction, the decryption involves use of primers and the concept of DNA hybridization.

II. ALGORITHM

A. Encryption

In the first stage of encryption we use the principle of rDNA technology and restriction enzymes. The message in DNA form is the 'gene of interest' and a DNA sequence from the database is the 'vector' which is used for 'cloning.'

In the second stage of encryption, a DNA sequence is virtually generated as the key and the BLAST [10] is used to make sure this doesn't match the sequence we have taken from the database, completely.

- Step 1: Select a DNA sequence from a database. (Say, a complete sequence of a particular chromosome of some organism.)
- Step 2: Analyse and identify the restriction sites present in the selected sequence.
- Step 3: Choose an appropriate restriction enzyme after considering the restriction sites.
- Step 4: The sequence is divided into 'n' fragments using the restriction enzyme.
- Step 5: The binary message signal is converted into DNA using the following conversion, considering two bits at a time:

00=A

01=C

10=G

11=T

- Step 6: This message DNA i.e. plain text is divided into (n-1) fragments.

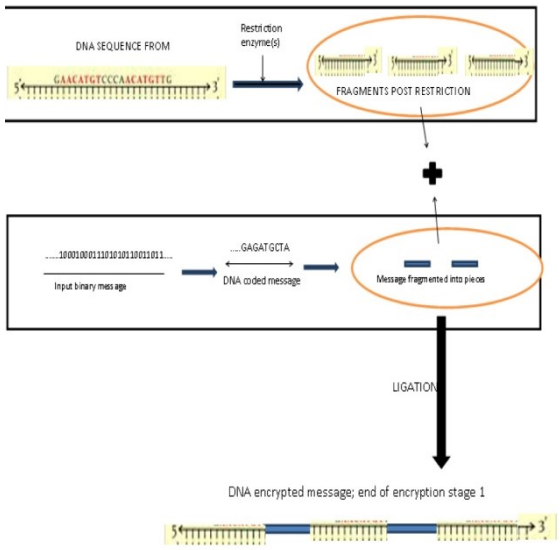

Fig. 2: First stage of encryption

- Step 7: These fragments are concatenated with the DNA fragments obtained post the restriction enzyme activity.
- Step 8: This new DNA strand is re-converted to binary as shown in Fig. 2.

- Step 9: Another virtually generated DNA sequence is taken as the key.
- Step 10: This sequence is first compared with the DNA encrypted message using BLAST tool [11].
- Step 11: The OTP is converted into binary.
- Step 12: This is then XOR'ed with the binary strand obtained in the previous step.(Fig.3)

The output of the XOR system is the double encrypted message, i.e. cipher text to be transmitted.

B. Decryption

The virtually generated DNA sequence that was used in the sender's end for the XOR operation is also present in the receiver's end as the key. XNOR is the first stage decryption.

Now, for second stage decryption the portions of the DNA sequence which contain the plaintext in DNA form should be identified. For this, primers are used; two primers for each fragment of message flanked by the originally cleaved DNA sequence on either side. The first primer is an oligonucleotide with 6-8 nucleotide bases which are complementary to the DNA sequence preceding the DNA-coded plaintext. The second primer is a similar oligonucleotide with bases complementary to the DNA sequence succeeding the DNA-coded plaintext. The last nucleotide of this primer is a di-deoxy ribonucleic acid (dDNA) to ensure chain termination.

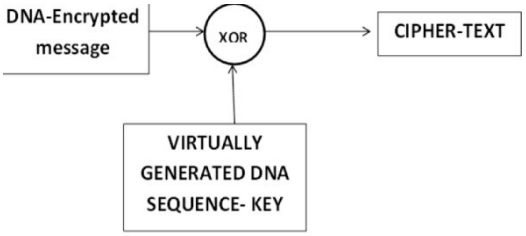

Fig. 3: Flow chart for Second stage encryption.

- Step 1: XNOR operation is performed on the binary sequence to obtain the DNA-encrypted message.
- Step 2: Use suitable primers to identify the portion of DNA which is the DNA coded plaintext.
- Step 3: The complementary strand of the sequence which is the plaintext is obtained and isolated.
- Step 4: The DNA-coded plain text is retrieved by complementing this complementary strand.
- Step 5: The DNA decryption is performed by converting the sequence back into binary form. This is the original message.

III. SIMULATION RESULTS

Constructing a DNA cryptosystem based on restriction fragments was implemented using MATLAB for a portion of a DNA sequence from a database and the output obtained after running the program is given below. Only portion of a sequence and a short message were used; that is scaling down was performed due to length constraints while implementing the program.

>>Encryption

Enter the message string'00101101'

Sequence from database:

ATGTTAAGTCCTAGAAGAACAAAATTCCGCAAACAG CAACGGGGCAGGATGCGGGGTCTAGCGGAGCGG G

Restriction enzyme used: XapI

DNA-coded message

AGTC

MERGED DNA STRAND 'ATGTTAAGTCCTAGAAGAACAAAGTCAATTCCGCA AACAGCAACGGGGCAGGATGCGGGGTCTAGCGGA GCGGG'

Result of restriction-enzyme encryption

0011101111000010110101110010000010000001000000101110100001111010110010000000100100100000110101010010010100011100110101010110111001001101 0001001101010

Before executing the actual encryption code, the sequence is analysed for restriction sites and the restriction enzymes that can be used on the sequence are determined. This can be done using a MATLAB function rebasecuts. This command when used yielded the following list of restriction enzymes:

'AciI'	'MaeI'	'BtsCI'	'TasI'
'AcsI'	'MaeI'	'CdpI'	'Tru1I'
'ApoI'	'MboII'	'FokI'	'Tru9I'
'BfaI'	'MseI'	'FspBI'	'Tsp509I'
'BfaI'	'SimI'	'FspBI'	'TspEI'
'BseGI'	'Sse9I'	'HaeIV'	'Tth111II'
'BspACI'	'SsiI'	'HaeIV'	'XapI'
'BstF5I'	'StsI'	'Hin4I'	'XspI'

Fig.5: List of restriction enzymes that can be used on this sequence

It is observed that even a short sequence from the database has even restriction sites to allow the usage of over 30 restriction enzymes. The choice of restriction enzymes that is used is made known to the sender only. Thus it is almost impossible to figure out which restriction enzyme is used and henceforth decrypt it, even if the encrypted message falls in the hands of intruders. Also, the database of DNA sequences itself is huge and unless the intruder knows which DNA sequence from which database is being operated on, decryption isn't plausible.

This cryptosystem is further encrypted by XOR'ing with a randomly generated sequence and this is the key known to only the sender and receiver; can be dubbed as a one-time pad of its own kind. This adds to the security of the message.

In the decryption process, after the XNOR operation using an OTP the DNA decryption using primers was

carried out using MATLAB for the same example. The output is as follows:

>>Decryption

ciphertext =

ATGTTAAGTCCTAGAAGAACAAAGTCAATTCCGCAA ACAGCAACGGGGCAGGATGCGGGGTCTAGCGGAG CGGG

Primer1 =

TTCTTGTT

Primer2 =

TTAAGGCG

DNA coded plaintext:

AGTC

Retrieved plaintext:

00101101

IV. ANALYSIS USING PROBABILITY

This method suggests the use of huge sequences such as chromosome sequences. Consider the sequence from the database that is initially considered for restriction action to be the complete sequence of chromosome #1 of Arabidopsis Thaliana (a small, flowering plant native to Eurasia).

Now this sequence contains 30,427,671 base pairs, that is, its single strand contains 30,427,671 nucleotides [12]. Minor 'attacks' on the DNA cryptosystem may be viewed as mutations (change in nucleotides; single or multiple at a particular position) or deletions. In this example the length of the DNA sequence is originally 70 nucleotides and the message to be inserted is of length 4. Assuming this as the upper limit for the size of the message direct variation is used to calculate the maximum length of message to be hidden in the chromosome.

Plaintext length	Length of the original DNA sequence
4	70
X ↓	30427671

Table 1: Direct variation table

X= × 4 = 1738724

A single nucleotide mutation or deletion is considered. The maximum probability that this nucleotide is deleted from the DNA-coded plaintext is:

(approx.)

This is considered as an 'upper-limit' scenario. The number of bits in the input message is assumed to be in the order of 32768, then the DNA coded plaintext would be of the length 16384 nucleotides. As a result, the probability that mutation or deletion occurs to the plain text is P (mutation or deletion of plaintext)

= = 0.0005845,

This implies that the chances of mutation or deletion of the plaintext is very less, especially if the plaintext is not very long when compared to the DNA sequence taken.

V. RESISTANCE AGAINST BRUTE FORCE ATTACK

This attack model involves the prediction and subsequent cracking of the cipher text by determining the key by exhausting all possible permutations. [13]

In this particular DNA cryptosystem, there are three keys. The primary key is the virtually generated DNA OTP present in both, the sender's and the receiver's end which is used in the XOR-XNOR system. The second 'key' is the basic DNA sequence obtained from the database which is operated on initially and the corresponding restriction enzyme(s) used for cleaving the DNA. This is present only at the sender's end. The third key is the set of primers used to extract each fragment of plain-text after XNOR'ing and this is present with the receiver only.

Brute force attack is used to deduce the primary key and either one of the remaining keys which is enough to crack the system. The primary key is of length N which is of the length of the encrypted message. The number of ways to deduce sequence of the primary key is $2^N$and in this suggested system which utilises the entire chromosome sequences for this purpose N is usually in the order of 30,000 (or greater).

⇨ No. of possible permutations required to figure out the sequences= Y

⇨ $Y=2^N$which is a very large number.

As it is observed, the odds against cracking the primary key using brute force attack are extremely less. In addition to this, there is the secondary key(s) to be cracked.

DNA restriction mapping is a biological problem whose brute force solutions are not practical. [14]

If the attacker is interested in determining the restriction enzyme(s) he has a choice of over 3800 enzymes to choose from. The sender might have also used these enzymes in combinations.

Therefore,

Let 'Y' be the number ways to determine the restriction enzyme used,

If one enzyme is used

$Y = {}^{3800}C_1 = 3800$

If two enzymes are used:

$Y= {}^{3800}C_2 =$ = 7218100

If three enzymes are used:

$Y = {}^{3800}C_3 =$ = 9138115000

The complexity of determining the restriction enzymes being used increases as the number of enzymes used in combination increases. This further makes it difficult to determine this secondary key.

Suppose the attacker wants to determine the restriction sites instead, then he has to know the DNA sequence taken from the database or ascertain each nucleotide which consumes effort as discussed above.

Now, the attacker also has the choice of determining the primers used for decryption which can also be the secondary key. Assume each primer is of the length of 8 nucleotides.

No. of ways of finding the nucleotides in order in one primer=$4^8= 2^{16}$

Each fragment of the plaintext is flanked by two primers. Therefore, if,

No. of restriction fragments= z

Then, No. of Message fragments= z-1

Therefore, No. of primers= 2×(z-1)

No. of ways to determine all the primers= $(2^{16})\times2\times(z-1)$, which is large when 'x' is a large number.

Techniques such as PCR typically use primers of length 20 nucleotides. Since the principle of PCR and hybridization is used here as a virtual implementation, if primers of length 20 are applied here, then it will increase the security further and inhibit easy exposure of the secondary key.

The number of ways to determine all the primers= $(2^{40})\times2\times(z-1)$, which is still a large number.

VI. PERFORMANCE ANALYSIS OF THE DNA CRYPTOSYSTEM

S. No.	Parameter	DNA Cryptosystem	Conventional Cryptosystem
1	Key size	Key 1: Very large and random; order of 10^6	Small. 109-3072, usually, except for OTPs.
		Key 2: Moderate in size; enhances security.	
2	Cryptographic strength	High; based on wide choice of sequences, enzymes, etc., from database and the size and randomness of the key.	Relatively less. Prone to brute force attack and the like.
3	Error during transmission	Very less probability for the message to be erroneous due to enormous size of cipher text when compared to the plain text.	Probability of error during transmission is greater. Part of plaintext maybe lost.
4	Security level	Two-fold	Single-fold
5	Time complexity	Greater than conventional cryptosystems.	>Few seconds

Table II: Performance Analysis

VII. FUTURE SCOPE

The system proposed in this paper is a static system. But there is a possibility to make it dynamic.

The idea now is to implement this principle of 'jumping genes' or transposons [15] and their function in this cryptosystem. Consider the plain text fragments in the DNA encrypted message to be the exons (functional part of genome) and the remaining DNA to be introns. If jumping genes were present in the intron portion of the system, then that would mean the system is dynamic; that is constantly changing. The rate at which it changes [16] could be programmed to be less than the time that taken by attackers to determine the secondary key using brute force attack so that the sequence changes before the code is cracked! The system could also be modified such that it identifies that there is an attack and 'triggers the transposition activity.' This kind of dynamicity would make this DNA cryptosystem very ideal and unbreakable.

VIII. CONCLUSION

Despite being a relatively new field, DNA Cryptography has a wide scope in the future of Cryptography. The sheer numbers of the base pairs of the sequence and the variability within the sequences ensures that the task of attack is tedious. This comparison of this system with the conventional ones justifies this field's promise in future. Having divulged into the role of restriction enzymes in Encryption and the use of primers in Decryption, this paper tries to unify the aspects of DNA with the basics of Cryptography.

REFERENCES/CITATIONS:

[1] Handbook of Applied Cryptography, by A. Menezes, P. van Oorschot, and S. Vanstone, CRC Press, 1996. 1-33

[2] Applied cryptography: protocols, algorithms, and source code in C, by Bruce Schneier,Wiley-India,2008,151-156

[3] Watson J.D. and Crick F.H.C. (1953). "A Structure for Deoxyribose Nucleic Acid,"Nature, vol. 25(1953), pp. 737-738

[4] Roberts RJ (November 1976). "Restriction endonucleases". CRC Crit. Rev. Biochem. 4(2): 123–64.

[5] Y G Kim, J Cha, and S Chandrasegaran: Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC40048/

[6] REBASE—enzymes and genes for DNA restriction and modificationRichard J. Roberts,* Tamas Vincze, Janos Posfai, and Dana Macelishttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC1899104/

[7] Restriction Enzymes By: Leslie A. Pray, Ph.D. © 2008 Nature Education, Citation: Pray, L. (2008) Restriction enzymes. Nature Education 1(1):38 http://www.nature.com/scitable/topicpage/restriction-enzymes-545

[8] PCR, second edition, By: Newton CR, Graham A

[9] "Ligases". Enzyme Resources Guide. Promega Corporation. pp. 8–14.

[10] S. Altschul, W.Gish, W.Miller, E.Myers, and J.Lipman. Basic local alignment search tool. Journal of Molecular Biology, 215:403–410, 1990.

[11] BLAST TOOL IS AVAILABLE AT: http://blast.st-va.ncbi.nlm.nih.gov/Blast.cgi

[12] Arabidopsis thaliana chromosome 1, complete sequence: http://www.ncbi.nlm.nih.gov/nuccore/NC_003070.9

[13] Adleman, Leonard M.; Rothemund, Paul W.K.; Roweis, Sam; Winfree, Erik (June 10–12). "On Applying Molecular Computation To The Data Encryption Standard". Proceedings of the Second Annual Meeting on DNA Based Computers (Princeton University)

[14] An Introduction to Bioinformatics Algorithms by Neil C. Jones, Pavel A. Pevzner(2004),chapter 4, p.83

[15] McClintock, Barbara (June 1950). "The origin and behavior of mutable loci in maize".Proc Natl Acad Sci U S A. 36 (6): 344–55

[16] Paquin CE, Williamson VM (5 October 1984). "Temperature effects on the Rate of Ty Transposition". Science 226 (4670): 53–55.