A Novel Encryption Scheme based on DNA Computing

M. Shyam, N. Kiran, V. Maheswaran

Abstract — *A novel method to encrypt data by applying DNA based computing technology is discussed. Binary data may be encoded in DNA strands by use of an alphabet of short oligonucleotide sequence. The encryption of natural DNA and the encryption of DNA encoding binary data are discussed.A pure XOR based DNA computing scheme is also discussed and the limitations of the proposed scheme are observed.*

Index Terms — DNA sequencing, cryptography, one-time pad, BMC Methods.

I. INTRODUCTION

1.1 Encryption

Encryption process deals with secure transmission of data. The necessity of using encryption in a communication system is to protect the data from being received by persons other than the intended recipient. Thus encryption is an indispensable block in the transmission section of any communication system.

Mathematically encryption involves mapping of a plain text to a cipherext using a well-defined function. A common algorithm is agreed upon by both the sender and the receiver and is used both to encrypt the plain text and to decrypt the cipher text. The efficacy of any encryption technique can be gauged by the length and type of its key. In symmetric key cryptography, the keys used for encryption and decryption are the same, whereas they are different in asymmetric key cryptography. The proposed scheme is based on symmetric key cryptography.

1.2 Deoxyribonucleic acid

Deoxyribonucleic acid (DNA) is a biochemical macromolecule that contains genetic information necessary for the functioning of living beings. DNA is organized as chromosomes in a cell's nucleus and the chromosomes make up the genome, the entire hereditary information about a cell.

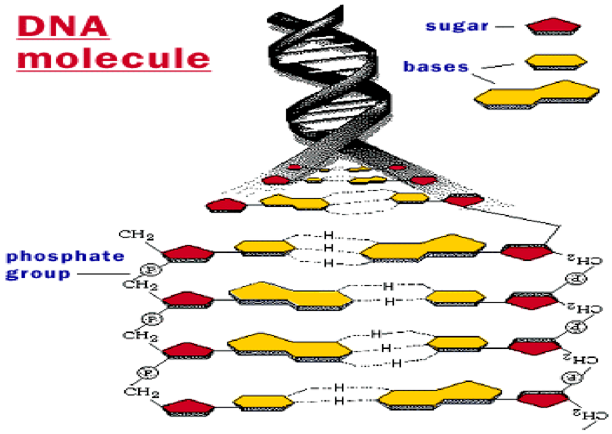

Fig. 1: A DNA molecule representation using double-helical structure

Chemically, DNA is made up of monomers called nucleotides, sugar and phosphorous. A DNA molecule consists of two strands of nucleotides twisted together to form a double helix. Four kinds of bases are found in the two strands, namely adenine, guanine, thymine and cytosine. A strand contains a sequence of bases in a specific pattern. The other strand contains the complementary nucleotides of the first strand. Adenine pairs with thymine using a double bond $(A = T)$, while thymine and cytosine pair with each other using a triple bond ($G \equiv C$). The genomic sequencing information is contained in the nucleotide bases.

1.3 DNA Computing

This field was initially developed by Leonard Adleman of the University of Southern California. In 1994, Adleman demonstrated a proof-of-concept use of DNA as a form of computation which solved the seven-point Hamiltonian path problem. Since the initial Adleman experiments, advances have been made and various Turing machines have been proven to be constructible.

This paper deals with bio-molecular encryption techniques for audio and image files. The idea is then implemented and the performance is verified.

M. Shyam is with the Department of Electronics and Communication and Engineering, College of Engineering, Guindy, Chennai – 600025, INDIA (e-mail: scintillatingstuffs@yahoo.co.in)
V. Maheswaran is with the Department of Electronics and Communication Engineering, College of Engineering, Guindy, Chennai – 600025, INDIA (e-mail: kvmakes@yahoo.com)
N. Kiran is with the Department of Computer Science and Engineering, College of Engineering, Guindy, Chennai – 600025, INDIA (e-mail: nkiran87@yahoo.com)

II. DNA CRYPTOGRAPHY

2.1 DNA Cryptography

Our DNA encoded messages are manipulated such that the plaintext strands are converted to cipher strands by adding message words to one-time pads. There are a number of possible methodologies for construction of cipher words used for the cryptosystems. One methodology is the random assembly of one-time pads in solution. Such methods are less favorable due to the difficulty of achieving both full coverage and yet still avoiding possible conflicts by repetition of plaintext and/or cipher words. A favorable method is to employ a DNA chip technology. Such DNA chips are currently commercially available; and chemical methods for construction of custom variants are well of immobilized developed. The DNA chip has an array of DNA strands, so that multiple copies of a single sequence of a microscopic array. These are grouped together in microscopic arrays of the DNA chip and are carbon nano-tube (CNT) probe addressable. There is known technology for growing distinct DNA sequences at each site of the array. DNA synthesis can be conducted in parallel. Therefore, the number of sequences prepared far exceeds the number of chemical reactions required. For preparation of oligonucleotides of length L, the 4L sequences are 4n chemical reactions. For example, the synthesized in 65,000 sequences of length 8 require 32 synthesis cycles. The plaintext and cipher pairs can he constructed so that there is a nearly unique word mapping between plaintext and cipher pairs. These resulting cipher word, plaintext word pairs can be assembled together in random order on a long DNA strand by a number of known methods, e.g., simply blunt end ligation. For higher efficiency, the use of a technique of hybridization assembly with complemented pairs has been done in Adleman's original DNA experiment. Cloning or polymerase chain reaction (PCR) may be used to amplify the resulting one-time so pad. Again, the cipher words should be constructed that there is nearly complete coverage for each pad and a nearly unique cipher word.

2.2. DNA-based Addition

A general algorithm for DNA-based modulo-2 addition of any two non-negative rational binary numbers is presented. We begin with the DNA representation of all possible pairs of input non-negative binary bits implementation standards.

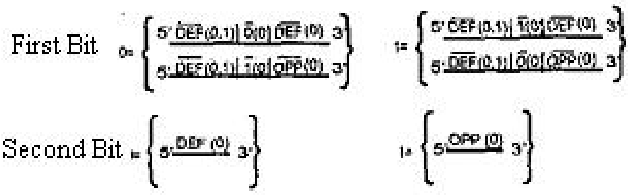
First Bit
0=
5' DEF(0,1) 0(0) DEF(0) 3'
5' DEF(0,1) 1(0) OPP(0) 3'
1=
5' DEF(0,1) 1(0) DEF(0) 3'
5' DEF(0,1) 0(0) OPP(0) 3'
Second Bit
5' DEF(0) 3'
1=
5' OPP(0) 3'

Fig. 2: Representation of 2 bits in DNA computing

The "first bit" and "second bit" at a given 0 or 1 of bit to be position refer to the value of either added in the original message and in the cipher word, respectively. DNA sequences are single-stranded, unique, and non-complementary, except that over-lining indicates a complementary DNA sequence, for example, m(0,l) is complementary to DEF(0,l). A number in parentheses refers to a position, whereas a number not in parentheses refers to the value of the digit at that position. Here the position formation provides CNT-probe accessibility for data write-in/read-out. The first bit is presented by two DNA strands, each containing (from the 5' end) a "position transfer operator".

The second digit is represented by a single DNA strand with the sequence DEF or PP if the digit is either 0 or 1, respectively. This strand represents an operator that serves as a primer in a primer extension reaction. As a schematic example, we illustrate how to add bits 1 + 1.

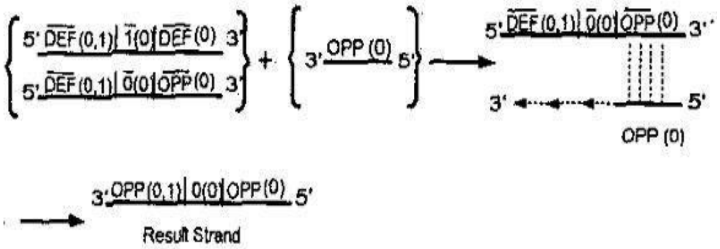
5' DEF(0,1) 1(0) DEF(0) 3'
5' DEF(0,1) 0(0) OPP(0) 3'
+
3' OPP(0) 5'
5' DEF(0,1) 0(0) OPP(0) 3'
3'
5'
OPP (0)
3' OPP(0,1) 0(0) OPP(0) 5'
Result Strand

Fig. 3: Addition of 1 and 1 through DNA computing

In the bio-chemical reaction, the operator (primer) vertical dotted lines represent hybridization between complementary DNA elements, and reiterated arrows represent primer extension. Adding the bit i in original message sequence (W,) hybridizes to the appropriate DNA strand representing the coded bit i (C;). Primer extension yields result DNA strand.

This successive reaction represents an example of a process of horizontal chain reaction, in which input DNA sequences serve as successive templates for extension of result strands.

III. DESIGN OF ENCRYPTION SYSTEM

3.1 Cryptography Mechanism

In classical cryptography [13], the Vernam cipher (now as the XOR one-time-pad cryptosystem) is deployed by generating a sequence, S, of R independently distributed random bits known as a one-time-pad. The one-time-pad is replicated, and stored one copy at the source and one at the destination. Let L be the number of bits of S that remain unused, where initially L = R. Recall that XOR is the operation that essentially is Boolean modulo-2 addition discussed previously.

When a plaintext binary message M which is n < L bits long needs to be sent, each bit Mi is XOR'ed with the message bit Ki to produce encrypted bits for i =n C; = M; + K; The n bits of S that have been consumed are then destroyed at the source. The encrypted sequence C = (Cl, Cz, , , . , C,) is then despatched to the destination. At the destination, the identical process is repeated. The sequence C is used in the place of M. We perform bitwise module2 addition with bits from S, and then the bit,s S are destroyed after they are consumed. The commutative property of the addition results in the initial C, + Ki = M;. In principle, we wish to convert one test tube of DNA strands (the plaintext messages) into another set of entirely different strands (the encrypted messages) in a random yet reversible way. Each of the plaintext messages are assumed to have appended unique prefix index tags of fixed length LO. Each of the one-time-pad DNA sequences are also assumed to have appended unique prefix index tags of the same length LO that form the complements of the plaintext message tags. By use of known recombinant DNA techniques (e.g., annealing and ligation), each corresponding pair of a plaintext message and pad sequence, with the same tag, can be combined into a single DNA strand. Our DNA encoded messages area bit-wise modulo-2 addition modified in this case by computation, so that fragments of the plaintext are converted to cipher strands using the one-time pad DNA sequences, and the plaintext strands are removed afterwards.

3.2 Parallel DNA Cryptosystem

In order to take advantage of the massive parallel processing capabilities of bio-molecular computation, the following method for basic operations such as arithmetic (addition and subtraction) permit chaining of the output of these operations into the inputs to further operations, and to allow operations to be executed in massive parallel fashion. Generalization of the addition to two non-negative vectors with n-digit binary numbers is straightforward. The two bits in each of the positions 2' through 2" are as shown in Figure 2 represented with the following modification.

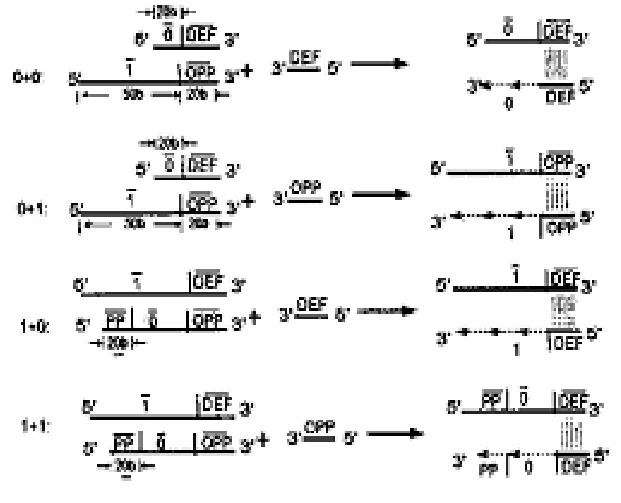

Fig. 4: Representation of Single Bit Addition

At a position i other than 1, unique DNA sequences represent the values O(i) and l(i), and operators are replaced appropriately; for example, DEF(1) and DEF(1,2) by DNA sequences representing EF(i) and DEF(i,i+l). The modulo-2 addition operation is in theory exactly as described above. This operation yields a final result strand longer but with the same basic structure. This more general algorithm can readily be extended to the addition of any two n-digit positive rational numbers. Addition is performed by combining a test tube primer extension reagents plus DNA one-time- two numbers pad strands that appropriately represent numbers to be added. The addition is then followed by a primer extension reaction. For 0+ 0, the 20-base input operator DEF hybridizes specifically to the 40-base input strand. Primer extension elongates the DEF operator to yield a 40-base result strand that encodes DEF and the result 0. For 0 + 1, the 20-base input operator OPP hybridizes to the 70-base "opposite" strand. Primer extension yields a 70-base result strand that encodes OPP and the result 1. Similarly, input of 1 + 0 yields a 70-base result strand encoding DEF plus the result 1. For 1 + 1, the first primer extension reaction yields the 60-base result FF strand and transforms the potential primer creator into the actual primer PP. This is the simplest example of the horizontal chain reaction. For module2 addition, since we don't care about the carrier, the addition takes O(1) to combine two sequences with length of n in a massive parallel fashion.

IV. SIMULATION

All performance analyses of this section have been carried out using Matlab 7.0 software package on a Pentium 4 processor with clock speed of 3 GHz and the system has 512 MB RAM.

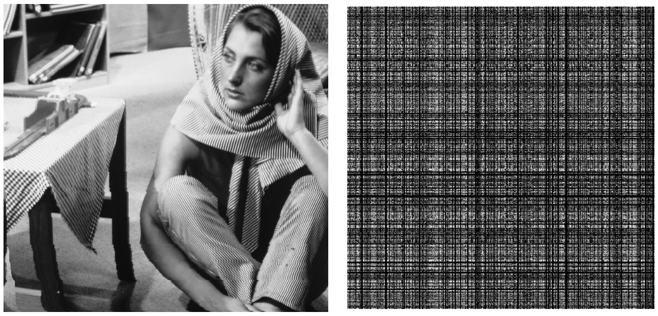

Fig. 5: Original and Encrypted Images through DNA Computing

Mean Squared Error (MSE) and Peak Signal to Noise Ratio (PSNR) are calculated for the encrypted audio file using these formulae.

$$MSE = \frac{1}{N}\sum_{i}\left|x(i)-e(i)\right|^{2}$$

$$PSNR = 20\log_{10}\left(\frac{65535}{\sqrt{MSE}}\right)$$

Here x and e are the input and encrypted signals respectively and N is the number of samples in the audio signal. An extremely high value of MSE of 7.157E8 and a correspondingly low PSNR of 7.77dB were obtained for images while MSE of 3.52E8 and PSNR of 10.86 dB for audio files were obtained.

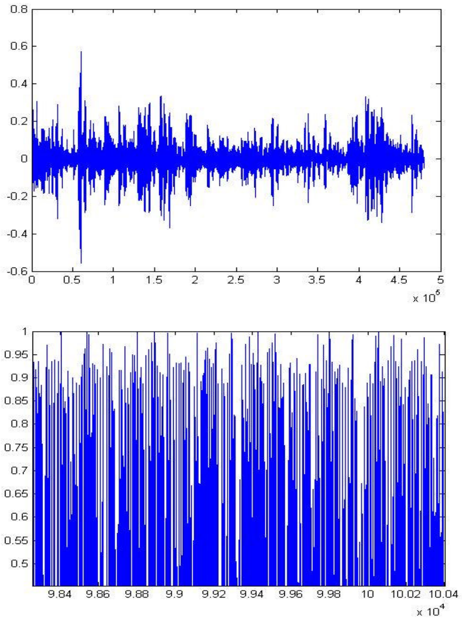

Fig. 6 : Audio encryption through the prescribed method: Original and encrypted signals as a function of time

Considerable scope for improvement is available from this work. The work of this paper could help biotechnologists and researchers from all over the world who are busy with building a DNA computer.

ACKNOWLEDGMENT

The authors would like to acknowledge the professors and lecturers of the Departments of Electronics and Communication Engineering and Computer Science and Engineering who have guided us through this work.

REFERENCES

[1] L.Adleman, “Molecular computation of solutions to combinatorial problems”, *Science*, vol. 266, no. 5187, pp.1021-1024, 1994.

[2] L.Adleman, “Computing with DNA”, *Sci. Amer.*, vol. 279, no. 2, pp.54-62, 1998.

[3] M.Arita, A.Nishikawa, M. Hagiya, K. Komiya, H.Gouzu, and K.Sakamoto, “Improving sequence design for DNA computing”, in *Proc.Genetic and Evolutionary Computation Conf.* 2000, 2000, pp.875-882.

[4] E.Ben-Jacob, Z.Hermon, and S. Caspi, DNAtransistor and quantum bit element: realization of nano-biomolecular logical devices, *Phys.Lett. A*, vol. 263, no. 3, p.199- 202, 1999.

[5] Y.Benenson, T. Paz-Elizur, R.Adar, E. Keinan,Z.Livneh, and E. Shapiro, “Programmable and autonomous computing machine made of biomolecules”, *Nature*, vol. 414, no. 6862, pp.430-434, 2001.

[6] S.Brennerand R.Lerner, “Encoded combinatorial chemistry”, *Proc. Natl. Acad. Sci.*, vol.89, no. 12, pp.5381ñ5383, 1992.

[7] J.Chen, H.Li, K.Sun, and B.Kim, “How will bio-informatics impact signal processing?” *IEEE Signal Processing Mag*., vol. 20, no. 6, pp. 16-26, Nov. 2003.

[8] W.A. Cerminshuiren, "Data storage uaing DNA", in *10th Foresight Nanotechnology Conference*, 2002.

[9] A. Suyama. DNA chips- integrated chemical circuits for DNA diagnosis and DNA computers", 1998.