SECRET WRITING BY DNA HYBRIDIZATION

Monica E. BORDA Olga TORNEA Tatiana HODOROGEA
Technical University of Cluj-Napoca, Communications Department
Gh. Baritiu str. 25-27, fax: 004-64-401575, Romania, E-mail: Monica.Borda@com.utcluj.ro

Abstract: This paper presents an original algorithm of secret writing by DNA hybridization, based on existing ideas described in the literature. In order to perform highly secure communication, the cryptographic, steganographic and biomolecular computation strength is used. The proposed algorithm is exemplified using matlab bioinformatics toolbox.

Key words: *Secret writing, DNA hybridization, PCR, DNA cryptography, steganography, biomolecular computation.*

I. INTRODUCTION

Secret writing was from ancient times and still is nowadays a way of protecting information from adversaries. Cryptography [7] and steganography [8] are the most widely used techniques which implement the secret writing. The first one manipulates the information to be not understandable (ciphering) and the second one hides its very existence.

DNA cryptography is a new born cryptographic field, based on Adleman's research of DNA computing [1]. Its starting point is Viviana Risca's computer science project on DNA steganography, winner of "Junior Nobel Price", 1999-2000 edition [2], which is proposing hiding messages in DNA microdots.

The huge advantages that DNA structure offers for efficient parallel molecular computation and its enormous storing capabilities, made from this research field a very attractive one for future applications, despite today limitations (expansive and time consuming). In [3] two procedures of one-time-pad (OTP) encryption schemes are presented: substitution and XOR. Reference [4] is proposing DNA hybridization as a central mechanism for microprocessor-controlled parallel robotic workstation.

This paper investigates a variety of bioinformatics methods and proposes an algorithm for encrypting and hiding data in real or artificial DNA digital form.

Section II is about basics of DNA including hybridization and PCR (Polymerase Chain Reaction), the technique used to amplify specific regions of a DNA strand [9]. In section III the principle of OTP and its generation as single strand DNA (ssDNA) encryption key is presented. Section IV is dedicated to DNA message encryption, while section V describes the message hiding, based on the idea presented in [2]. The algorithm of message recovery is presented in section VI. Final conclusions and bibliography are ending the paper. The proposed algorithm is illustrated using matlab bioinformatics toolbox facilities.

II. BASIC OF DNA AND PCR

Deoxyribonucleic acid (DNA), the major support of genetic information (*genetic blueprint*) of all living cells, is composed of two long strands of nucleotides, each containing one of four bases (A – adenine, G – guanine, C – cytosine, T – thymine), a deoxyribose sugar and a phosphate group. The DNA strands exhibit chemical polarity, meaning that it has different chemical groups on each end of a molecule (5' – top end and 3' – bottom end) [9].

DNA molecule has double-stranded structure obtained by two single-stranded DNA chains hydrogen bonded together between bases (A-T and G-C). The double-helix structure is configured by two single antiparallel strands (*Figure 1*).

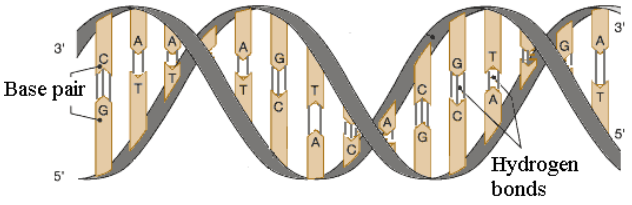

Figure 1.DNA Structure

DNA strands can be synthesized chemically using a machine known as DNA synthesizer. Single-stranded chains obtained with a DNA synthesizer are known as synthetic *oligonucleotides*; they are usually of 10-100 nucleotides in length. DNA strands that bond to each other through A - T and G – C bonds are known as complementary strands.

In this paper we'll refer to the individual strands as *single-stranded DNA (ssDNA)* and to the double helix as *double-stranded DNA (dsDNA)*. Individual, ssDNA can, under certain conditions, form dsDNA with other strands which are sufficiently complementary. This process is called *hybridization* because the double-stranded molecules are hybrids of strand which comes from different sources. The initial phase of hybridization is a slow process because of

Manuscript received May 4, 2009; revised June 24, 2009

the random chance that a region of two complementary strands will come together to form a short sequence of correct base pairs (*Figure 2*). This initial pairing is followed by a rapid matching of the remaining complementary bases [6].

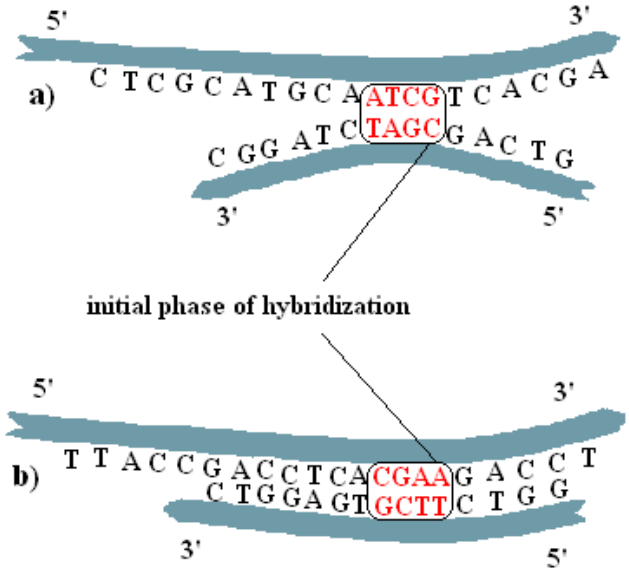

Figure 2.Hybridization a) Base-pairing can not go further because sequences have different bases, pairing is unstable and the strands come apart. b) Base-pairing continues because sequences are complementary.

Polymerase chain reaction (PCR) is biochemistry and molecular biology technique used to exponentially amplify certain regions of DNA, via enzymatic replication, starting from a DNA template containing the region of the DNA fragment to be amplified, the generated DNA is used as template for replication, which explains the exponential amplification.

The enzymatic reaction (polymerase), which replicates the DNA template is a reaction governed by stable temperature. It consists of repeating 20-35 times a cycle. A cycle has the following steps:

- *initialization:*
  - dsDNA template contains the region of the DNA fragment to be amplified
  - primers (oligonucleotides) which are complementary to the DNA regions at 5' and 3' ends of DNA segment which will be amplified
  - denaturation of the DNA template by melting at 94-98°C: the hydrogen bonds between complementary bases of the dsDNA template are disrupted, meaning that 2 ssDNA strands are obtained
- *annealing*: after the temperature is lowered primers can attach to the ssDNA template. Stable bonds are created when the primer sequence is complementary to the template sequence and a short section of dsDNA is obtained. At this moment polymerase enzyme attaches and starts DNA synthesis. (50-64 °C for 20-40 s)
- *extension (elongation)*: during 4-120 s at 72 °C, the DNA polymerase synthesizes new DNA strands complementary to the DNA template (*Figure 3*)

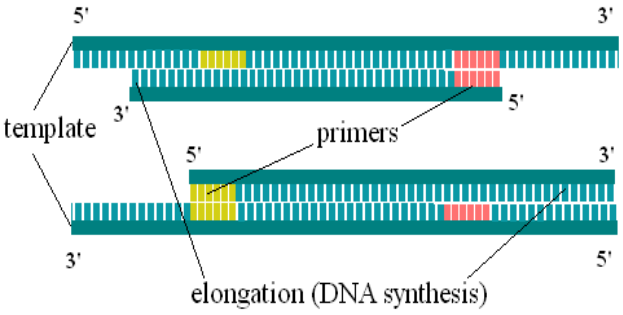

Figure3. Amplifying process in PCR technique

Each cycle multiply twice the specific target sequences (*Figure 4*).

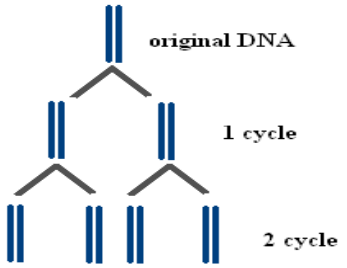

Figure 4.PCR which cycles 2 times

III. ONE-TIME-PAD OR ssDNA AS THE ENCRYPTION KEY

One-time-pad encryption uses a large non-repeating set of truly random key letters. Each pad is used exactly once, for only one message. The sender encrypts the message and then destroys the used pad. The receiver has an identical pad and uses it for decryption. The receiver destroys the same pad after decrypting the message. New message means new key letters. Cryptosystems which use a secret random one-time-pad are known to be perfectly secure [7].

In the proposed algorithm the one-time-pad is a ssDNA. Transmitter and receiver will have a great set of such non-repeating strands and each ssDNA (pad) from the set will be used just once and destroyed after that. A large set of unique DNA strands can be easily assembled using the artificial, random generated DNA strands or fragmented and denatured real DNA chromosomes from any organism. Use of the original DNA chromosomes is not recommended because it can be recognized by an adversary from encrypted message and OTP must be secret.

An illustration of the algorithm generating OTP is made using matlab bioinformatics toolbox; a random ssDNA of 220 bases long was generated:

TATGAGTTTGCCGAGACCTCGTCGATCTCTAAGATC
ACAAATGGCCTTCTAGGCCGTACACTGTACCCTACT
ACAAAAGTCTTAGAATAATGATCAGTCGGATTAAC
TGGCTTGACGAGGATAAGCCTTCATAAGAAAGAGA
GGGCTACTTATTTGTCCACCCACAGTCGGAACCTTC
TCTTGGTACACATACAGCGCAAGGACGCAGTTTTT
CAATGAC.

The length of the OTP depends on the length of the message. It must be 10 times longer than the binary message, because 1 message bit will be encoded in 10 nucleotides. Considering this, a set of ssDNA will be synthesized proportional to the approximate length of the intended messages. We exemplify below encryption of a message 21 bits long and decide to create an OTP of corresponding length.

IV. MESSAGE ENCRYPTION

Based on the idea found in [4], a DNA encryption algorithm is presented. The generated OTP ssDNA key is used to encode a message. The technique to hide the encrypted message is based on Viviana Risca's idea [2]. The DNA pad is scanned in reverse order from the sequence end towards the beginning and analyzed 10 bases at once.

The actual plaintext message is transformed and used in binary form. For example, the original message: "ZOO" in ASCII cod will be: "90 79 79" and in binary form: "101101010011111001111". The message bits are interpreted as follows:

"0" - neither operation is performed;
"1" – a strand, 10 bases long, complementary to ssDNA is created;

Hybridized segments on ssDNA are binary ones, while unchanged single stranded portions are binary zeros. This is why encryption on the ssDNA is performed starting from the end of the one-time-pad sequence. Binary message can be shorter then ssDNA and if encryption is done from the beginning to the end of the ssDNA pad then it will be no evidence about the last encrypted bits in case they are zero.

The encrypted message is a set of segments complementary to ssDNA, which encode binary ones from the plaintext message. For example, encryption of the plaintext message "ZOO" with the bits established above is:

AAAGTTACTG, ATGTCGCGTT, AACCATGTGT, GGGTGTCAGC, CTCCCGATGA, GAACTGCTCC, CTAATTGACC, ACTAGTCAGC, GAATCTTATT, GATGTTTTCA, TACCGGAAGA, TTCTAGTGTT, CAGCTAGAGA, GGCTCTGGAG.

Each binary one from the message has a corresponding oligonucleotide sequence which is a part of the ciphertext and is complementary to a certain place on ssDNA. For example, the sequence "AAAGTTACTG" encrypts the first bit "1" from the message and is complementary to the last 10 bases "TTTCAATGAC" from the ssDNA. The next 10 bases, counting from the end of ssDNA will not have complementary segment because the next bit from the message is "0".

V. HIDING THE ENCRYPTED MESSAGE

The encrypted message will be transmitted in a compact form and hidden using DNA steganography presented in [2] such that for an adversary it will be hard to find and extract the intended message. Viviana Risca's successful experiment was hiding DNA-encoded message in a microdot using as a background fragmented and denatured human DNA.

The message encrypted segments, presented above are bound together in a single strand or several strands, if a long message is encrypted using a special ligase protein and the complementary strand as a template *(Figure 5).*

Figure 5.Binding process between two segments

Message encrypted strand is placed between two PCR primer sequences *(Figure 6a)* and hidden between many other similar structures with the same length and letters probability appearance. Intended message from *Figure 6a* has the double stranded form and is added to the synthesized dsDNA. Without knowing primers which act as start and stop sequences of the secret message, the dsDNA will look the same (*Figure 6b).*

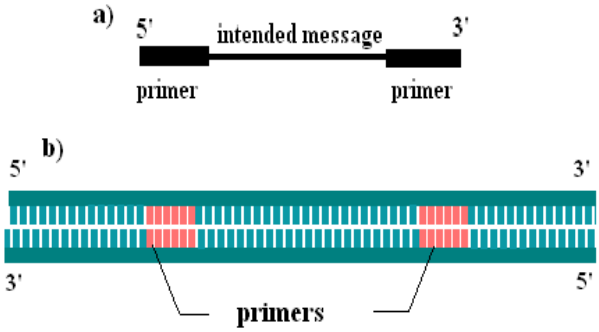

Figure 6. a) Structure of hidden DNA message b) dsDNA containing the message which can be recognized only knowing forward and reverse primers

As all nucleotides bases from the strands will look the same, an adversary can find the intended message only knowing the primer sequences in order to amplify the DNA and read the message. If 20-bases primers were used at message sequence assembling, separate amplifications with 4^{20} different primer pairs would be required for message recovery, followed by analysis of the obtained products. So it will be difficult to read the message without knowing the specific primer sequences.

Intended recipient can read the message knowing the right primers in order to amplify the message by PCR and using gel electrophoresis technique to read it.

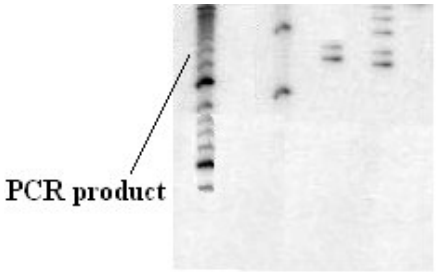

Figure 7. Gel electrophoresis analysis

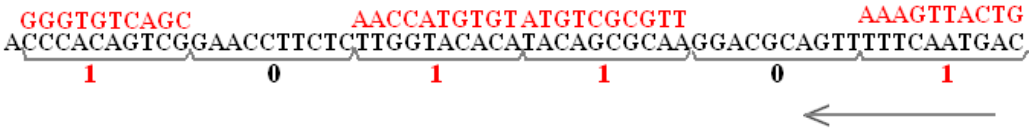

Figure 8.A fragment from hybridization between one-time-pad (ssDNA) and encrypted message

Gel acts like a filter sorting DNA strands by their length or type. Electrical current make DNA strands to move. Strands of the same length will move at the same place and form a group such that PCR product will be a noticeable arrow at gel analysis like in *Figure 7.*

VI. MESSAGE RECOVERY

Message recovering is possible for someone who knows the medium containing the message, primer sequences and the one-time-pad used for encryption. Decryption is done in few steps:

- PCR performing using the right primers to amplify the secret message (*Figure 3,4*)
- Analysis of the PCR product by gel electrophoresis in order to sort DNA strands and find the encrypted message (*Figure 7*) which is a DNA strand.
- Cleavage of the obtained strand in original segments 10 bases long using restriction enzymes to cut the strand.
- Hybridization process between achieved segments and the same ssDNA used at message encryption (one-time-pad) (*Figure 8*).
- Reading the message: hybridized segments on ssDNA are binary ones (1) and unchanged single stranded portions binary zeros (0).
- One-time-pad destruction.

For example, reading of the first "101101" bits from the recovered message "ZOO" is like in *Figure 8*. Reading is performed in reverse order. In case the message is shorter than the OTP, remaining portion is read as binary zeros.

DNA hybridization is a self-assembling process, its beginning is slowly due to the random chance that two complementary strands will come together to form a short sequence of correct base pairs, but its follow-up is a rapid matching of the remaining complementary bases. This important property can be exploited in searching algorithms and parallel computation techniques.

CONCLUSIONS

Based on the ideas presented in [1], [2], [3], and [4] an original DNA cryptographic algorithm was performed. We use one-time-pad principle and DNA hybridization for message encryption and then hide the results using steganography and bioinformatics tools. DNA hybridization is a self-assembling process and in this algorithm we pointed out the advantages of parallel computation on a huge scale offered by its properties. This algorithm was implemented in matlab using bioinformatics toolbox. Laboratory implementations are possible (microarray technology [9]), but are still expansive and time consuming. Despite of this, simple and effective algorithms are necessary in order to bring DNA computing on digital level and use it on large scale.

ACKNOWLEDGMENT

This paper was funded by the Romanian Agency UEFISCU within the PN II, IDEI no. 909/2007 research grant.

REFERENCES

[1] L. M. Adleman, Molecular computation of solution to combinatorial problems, *Science*, 266:1021-1024, November 1994.
[2] C. T. Taylor, V. Risca, and C. Bancroft, "Hiding messages in DNA microdots" *Nature*, 399:533-534, 1999.
[3] A. Gehani, T. LaBean, and J. Reif, "DNA-Based Cryptography", *Lecture Notes in Computer Science*, Springer, February 2004.
[4] S. Roweis, E. Winfree, R. Burgoyne, N. V. Chelyapov, M. F. Goodman, P. W. K. Rothemund, and L. M. Adleman, "A sticker based architecture for DNA computation" *In Proceedings of the Second Annual Meeting on DNA Based Computers*, volume 44 of DIMACS: *Series in Discrete Mathematics and Theoretical Computer Science*, pages 1-30, May 1996.
[5] C. R. Calladine, H. R. Drew, B. F. Luisi, A. A. Travers, "Understanding DNA The Molecule & How It Works", *Academic Press*, April 2004.
[6] D. L. Hartl, E. W. Jones, "Genetics: Analysis of Genes and Genomes", *Jones and Bartlett Publishers*, 2004.
[7] B. Schneier, "Applied Cryptography: Protocols, Algorithms, and Source Code in C", John Wiley & Sons, Inc, 1996.
[8] S. Katzenbeisser, F. Petitcols, Information hiding techniques for steganography and digital watermarking, *Artech House*, 2000.
[9] M. Schena, "Microarray Analysis", *Wiley-Liss*, July 2003.