DNA Cryptosystem Using a Simple Replacement

Oleksandr Sievierinov
Information Security Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
oleksandr.sievierinov@nure.ua

Andrey Evheniev
Information Security Department
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
evheniev@gmail.com

Криптосистема DNA з Використанням Простої Заміни

Олександр Сєвєрінов
Кафедра інформаційної безпеки
Харківський національний університет радіоелектроніки
Харків, Україна
oleksandr.sievierinov@nure.ua

Андрій Євгенієв
Кафедра інформаційної безпеки
Харківський національний університет радіоелектроніки
Харків, Україна
evheniev@gmail.com

***Abstract*—The paper considers the main purposes and possibilities of using DNA in such areas of information security as cryptography and steganography. The main opportunities for implementing and supporting systems based on DNA transformations, as well as the use of DNA containers as storage for a large amount of data. A cryptographic system is proposed simply for replacing with the use of DNA components.**

***Анотація*—У роботі розглядаються основні цілі та можливості використання DNA в таких сферах інформаційної безпеки, як криптографія та стеганографія. Основні можливості для впровадження та підтримки систем на основі перетворень DNA, а також використання контейнерів DNA як сховища великого обсягу даних. Криптографічна система пропонується просто для заміни з використанням компонентів DNA.**

Keywords—DNA, cryptography, information security

Ключові слова—DNA, криптографія, інформаційна безпека

I. INTRODUCTION

The dimensions of the "digital universe" will exceed 16 zettabytes by 2017. A significant portion of this data is stored in the form of archives. For example, Facebook recently built a separate data center for the cold storage of 1 data exabyte. The same amount of information can fit into 1 mm3 of DNA. The preservation of data in DNA takes place in three stages: the conversion of digital data into a sequence of DNA nucleotides, the synthesis of DNA molecules and, directly, the storage of data. To count the data, it is necessary to select the required sequence from the DNA molecule and convert it to its original form. It should be noted that there are difficulties in working with DNA-storages, for example, there are questions about the cost of data encryption, but researchers are sure - as medical technologies develop, it will decrease. This is what happens. Time for synthesis and sequencing decreases exponentially, and the growth of their effectiveness follows the law [1].

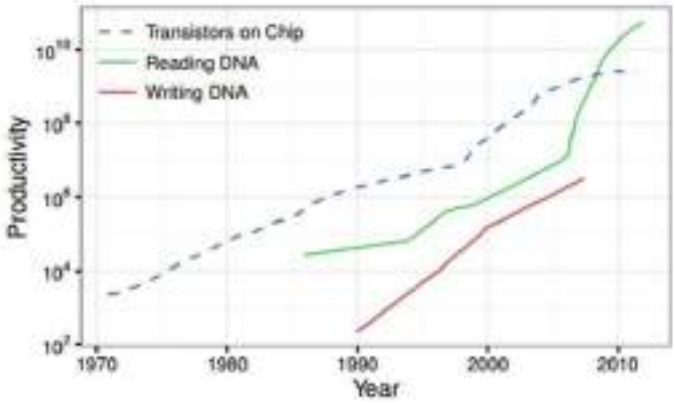

Fig. 1 - Productivity in relation to the year

The DNA molecule stores information in a quadruple number system, according to the number of nucleotides (0 = A, 1 = T, 2 = C, 3 = G). It is a compact container with a recording density that is thousands of times larger than that of existing media. However, in order for the technology to pass from scientific testing to commercial use, it is necessary to solve a number of problems. One is the specificity of digital information, in which the same bits can be repeated many times (CCCCCCCCCCCCCC). If you repeatedly repeat the same nucleotide in a DNA molecule, this adversely affects the stability of the cluster and information can be lost even if you use excessive duplication and error correction.

Researchers from the European Institute of Bioinformatics have published a paper describing how to significantly improve the stability of DNA. Simply, they propose to abandon the quaternary system (Base-4) in favor of the ternary (Base-3), and the fourth nucleotide to use for official purposes [1].

During the experiment, the researchers recorded almost a megabyte of information in the DNA, including all 154 Shakespeare sonnets in .txt format, a video recording of a 26-second Martin Luther King speech, a cover of the journal Bioinformatics Institute in .jpeg format, a scientific paper describing the structure of DNA in .pdf format, as well as another file describing the encoding process. In total, all fit in 739 kilobytes [2, p. 205].

When switching from Base-4 to Base-3, we lose 25% of the information capacity, but even in this version, scientists report an information recording density of petabytes per gram of biological material. The experiment showed the reliability of reading the information 100%. Theoretically, this scheme is capable of scaling up to the limits of the volume of all existing digital information, the authors of the study write[3].

II. USING DNA TO PROTECT DATA

DNA cryptography is a new field based on studies in computational DNA and new technologies, such as PCR (Polymerase Chain Reaction). DNA computing has a high level of computational ability and is capable of storing a huge amount of data. The DNA gram contains 1021 DNA bases, which is equivalent to 108 terabytes of data. In DNA cryptography, we use existing biological information from publicly available DNA databases to encode plaintext [3, p. 415].

A cryptographic process can use different methods. The methods of single-use notebooks describe the most effective security algorithms. In the case of a one-time notepad, open text is combined with a secret random key or our notebook, which is used only once. The notepad is combined with plain text, using a typical modular addition or XOR operation. The speed of the algorithm should be quite high. Biological Background: DNA is an abbreviation for deoxyribonucleic acid, which is the germplasm of all life styles [2, p. 217].

DNA is a kind of biological macromolecule and consists of nucleotides. Each nucleotide contains one base and there are four types of bases: adenine (A) and thymine (T) or cytosine (C) and guanine (G) corresponding to four types of nucleotides. Single-stranded DNA is constructed with orientation: one end is called 5 ', and the other end is called 3'. Usually, DNA exists as double-stranded molecules in nature. Two complementary strands of DNA are held together, forming a double helix structure. The structure of the double helix was discovered by Watson and Crick; Thus, the complementary structure is called the Watson Crick complementarity.

Their discovery is one of the greatest scientific discoveries of the 20th century. The development of DNA cryptography benefits from the development of DNA calculations (also called molecular calculations or biological calculations). DNA contains four types of nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). A DNA strand is a linear sequence of these nucleotides. Thus, we have four code (A, C, G, and T), so an obvious approach to storing binary data is to encode them in a quaternary number system, for example, 0 = A, 1 = C, 2 = G, and 3 = T, based on the data in Appendix A, the implemented encryption algorithm based on RSA using nucleotides is presented [4].

However, it should be borne in mind that synthesis and sequencing are prone to errors. The probability of errors can be reduced by encoding binary information not in a quaternary but in a ternarynumber system, as shown in the figure below. To avoid inefficient conversion of the original binary data to a ternary number system, the Huffman code is used. An example of such an input transformation in Figure 2.

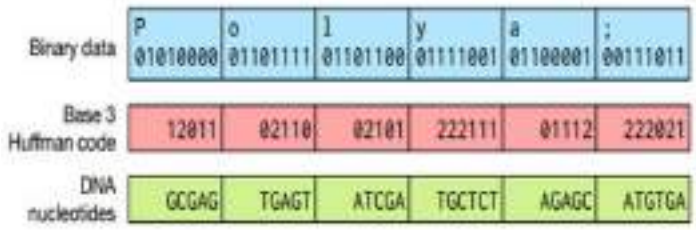

Fig. 2 - The process of converting a word to DNA nucleotides

Each of the three digits correlates with the DNA nucleotide in accordance with Figure 3, with the nucleotides in the chain being not repeated, which leads to a decrease in the sequencing error.

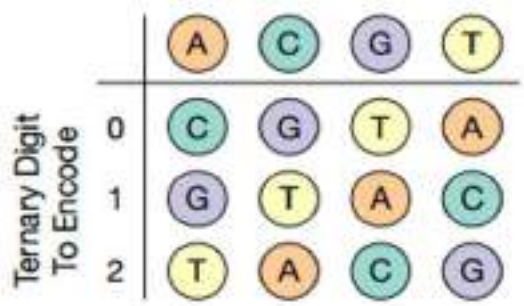

Fig. 3 - Coding of nucleotides

To ensure the possibility of random access to data, scientists organized the translation of keys into unique primer sequences. Primers are short synthetic threads that determine the beginning and the end of the region that needs to be amplified. Primers provide random access by polymerase chain reaction, which generates multiple copies of DNA. The circuits of a particular object have a common primer, and different chains with the same primer differ by address. "By controlling the sequences that are used as primers for polymerase chain reaction (PCR), we can specify which strands in the solution will undergo amplification. In order to consider the value of the key in the solution, we simply perform PCR using the primer corresponding to this key, "the scientists say [5].

III. DNA CRYPTOSYSTEM USING REPLACEMENT

A one-time substitution system uses a group of plain text binary messages and a table group that defines a random mapping to encrypted text. The input chain has length n and is divided into unencrypted words of fixed length. The table displays all possible lines of unencrypted text with a fixed length in the corresponding lines of the encrypted text, so that there is a unique inverse mapping.

Encryption occurs by replacing each word of the plaintext of the DNA with the corresponding encrypted DNA word. The mapping is realized using a long DNA notebook consisting of a number of segments, each of which points to a single word with plain text for encoding word combinations. A word with plain text acts as a primer binding, which then lengthens. This leads to the formation of a text pair of plaintext and ciphertext.

An ideal one-time library will contain a huge number of notebooks, and each of them will provide a completely unique, random display of unprotected words in encrypted words.

A repeating block consists of: one word of the sequence C_i, from a set of words corresponding to the code or codebook and the prefix P_i. Note that P_i includes a unique subsequence that prevents the attacks of frequency analysis by comparing several instances of the same plaintext message with different encrypted words. In addition, this prefix may optionally be used to encode the position of a word in a message [6].

Each pair of sequences i uniquely connects a text word with a ciphertext word.

The sequence of stoppers prohibits the expansion of the growing DNA chain beyond the bounds of the pairwise encrypted word. Using this theme, a library of unique codebook chains is created. Each separate chain from this codebook library specifies a specific unique set of pair words.

A one-time notebook consists of a DNA chain of length n, containing

$$d = \frac{n}{L_1+L_2+L_3}$$

copies of a repeating pattern: an encrypted word of length L_2, a word of open text of length $L_1$and a length stop sequence L_3.

Note that the length of the word grows logarithmically along the entire length of the notebook. Definitely

$$L_1 = c_1 \log_2 n;\ L_2 = c_1 \log_2 n \text{ and } L_3 = c_3,$$

where c_1, c_2, c_3 are fixed integer constants and $c_1, c_2, c_3 > 1$. Each repeater specifies a single match pair, and no word from the codebook or text word will be used more than once in any notebook.

Therefore, given the encrypted word C_i, we are sure that it displays only one word with plaintext P and vice versa. The stopper sequence acts as a "punctuation" between repeating links, so that the DNA polymerase can not continue copying the matrix filament (notepad). Stopper sequences consist of a sequence of identical nucleotides that act to stop the copying of the chain by DNA polymerase, given the lack of complementary nucleotide triphosphate in the tube. For example, the TTTT sequence will act as a stopping point if the polymerization mixture lacks its complementary base [7].

The experimental feasibility depends on the following factors: the size of the lexicon, which is the number of pairs of words of plaintext-encrypted text, the size of each word, the number of disposable DNA, the notebooks that can be created in the synthesis cycle, and the length of each message that must be encrypted. If the lexicon used consisted of English words, its size would be in the range of 10,000 to 25,000 word pairs. If, for experimental reasons, a smaller lexicon is required, then the words used can be a smaller set, such as ASCII characters, which leads to the size of the dictionary 128.

IV. CONCLUSIONS

In the future, such systems will potentially save a huge amount of data on microscopic media. Imagine a "flash drive" of 100 mm3, capable of storing about 100,000 PBB of data. However, meanwhile, the biggest obstacle to the introduction of such technologies is time. Decoding and reading the DNA molecule takes many hours. Therefore, this kind of storage is hardly suitable for the content of frequently used data, but it can turn our notion of long-term storage in data centers.

REFERENCES

[1] DNA Data cryptography [Electronic resource] // URL: https://www.slideshare.net/mayukhmaitra/dna-cryptography.

[2] G. Baumslag, B.Fine, and X.Xu, Cryptosystems Using Linear Groups Appl. Alg. in Engineering, Communication and Computing 17, 2006.

[3] A.G. Myasnikov, V.Shpilrain and A. Ushakov, Group-Based Cryptography, CRM Barcelona, 415, 2007.

[4] K.Ko, J.Lee, J.H. Cheon, J.W. Han, J.Kang, C.Park, New Public-Key Cryptosystem Using Braid Groups , Advances in Cryptology - CRYPTO 2000 Santa Barbara CA , Lecture Notes in Computer Science, Springer 1880, 2000, 166-183.

[5] V.Shpilrain and A. Ushakov, The Conjugacy Search Problem in Public Key Cryptography; Unnecessary and Insufficient Applicable Algebra in Engineering, Communication and computing, 17, 2006 285-289.

[6] Cerm´ak, J. “Digital generators of chaos,” Phys. Lett. 525, 1996, 151–160.

[7] K.Ko, J.Lee, J.H. Cheon, J.W. Han, J.Kang, C.Park, New Public-Key Cryptosystem Using Braid Groups , Advances in Cryptology - CRYPTO 2000 Santa Barbara CA , Lecture Notes in Computer Science, Springer 1880, 2000, 166-183.