DNA sequences for robust encryption: a strategy for IoT security enhancement

Sequências de DNA para encriptação robusta: uma estratégia para o aprimoramento da segurança em IoT

DOI: 10.54021/seesv5n1-067
Recebimento dos originais: 22/03/2024
Aceitação para publicação: 08/04/2024

Kadda Benyahia
PhD in Computer Science
Institution: Laboratory Technology of Communication, University of Tahar Moulay
Address: BP 138, cite ennasr, Saida, 20000, Algeria
E-mail: benyahiak@gmail.com / kadda.benyahia@univ-saida.dz

Abdelkader Khobzaoui
PhD in Computer Science
Institution: Mathematics Laboratory, Djillali LIABES University
Address: BP 89 Sidi Bel Abbes, 22000, Algérie
E-mail: akhobzaoui@yahoo.fr

Soumia Benbakreti
PhD in Applied Mathematics to Computer Science
Institution: Mathematics Laboratory. Djillali LIABES University
Address: BP 89 Sidi Bel Abbes, 22000, Algérie
E-mail: souben2223@gmail.com

ABSTRACT
As the Internet of Things (IoT) permeates our lives, connecting everything from household appliances to complex industrial systems, the imperative to secure these devices intensifies. Cryptography, as a cornerstone of digital security, plays a crucial role in safeguarding transmission channels from intrusions and misuse. Cryptography secures communications and data within IoT networks by ensuring three key functions: confidentiality, integrity, and authentication. DNA-based cryptography emerges as a promising innovation in the field of cybersecurity, particularly for the Internet of Things (IoT), where data and communication security is an escalating concern. Utilizing the unique properties of DNA, such as its massive storage capacity and biomolecular complexity, this approach introduces a novel dimension of security. This study introduces a balanced approach within DNA cryptography to enhance message security in Internet of Things (IoT) settings. It outlines a method for creating secure symmetric keys using DNA sequences, typically derived from human chromosomes, and then applies biological techniques like transcription and a biological Xor operation. This step is succeeded by a translation phase that utilizes an index table created from an initial key, making the process more complex.
Keywords: cryptography, DNA, encryption, decryption, security.

RESUMO
À medida que a Internet das Coisas (IoT) permeia nossas vidas, conectando tudo, desde eletrodomésticos a sistemas industriais complexos, a necessidade de proteger esses dispositivos se intensifica. A criptografia, como um pilar da segurança digital, desempenha um papel crucial na proteção dos canais de transmissão contra intrusões e abusos. A criptografia garante a comunicação e os dados dentro das redes IoT ao assegurar três funções-chave: confidencialidade, integridade e autenticação. A criptografia baseada em DNA surge como uma inovação promissora no campo da segurança cibernética, especialmente para a Internet das Coisas (IoT), onde a segurança dos dados e das comunicações é uma preocupação crescente. Utilizando as propriedades únicas do DNA, como sua capacidade massiva de armazenamento e sua complexidade biomolecular, essa abordagem introduz uma nova dimensão de segurança. Este estudo introduz uma abordagem equilibrada na criptografia de DNA para aumentar a segurança das mensagens em ambientes da Internet das Coisas (IoT). Ele descreve um método para a criação de chaves simétricas seguras usando sequências de DNA, geralmente derivadas de cromossomos humanos, e depois aplica técnicas biológicas como transcrição e uma operação de Xor biológico. Esse passo é seguido por uma fase de tradução que utiliza uma tabela de índice criada a partir de uma chave inicial, tornando o processo mais complexo.

Palavras-chave: criptografia, DNA, encriptação, decriptação, segurança.

1 INTRODUCTION
Nowadays, we hear more and more about the Internet of Things (IoT), it is the set of objects that can communicate wirelessly. In the IoT environment, several messages are sent by publishers, on transmission channels which will be read by subscribers via a server which is responsible for linking them.

With the increasing pace of Internet and networking technology, day by day, threats are also increasing for users, due to the large number of connected objects and the large flow of information between these objects. Therefore, to ensure that information reaches the intended sender and recipient, all the weaknesses of the security systems must be overcome.

Several research works have been carried out around data security in the IoT environment while seeking to meet the requirements of time and data flow, hence encryption is one of these solutions. [7,9,15]

The task of any cryptographic algorithm is to secure data for a very long time. The general principle of attacks is to determine the authentication key by
looking for significant or repetitive signs in the decryption output using different keys, which is a long and time-consuming process, the use of new technology using powerful computers to remedy this problem. And therefore, A security system can have many weaknesses, such as the size of the length of the encryption key and where the numbers are stored.

DNA cryptography solves such problems and gives hope for developing unbreakable algorithms. Data is secure either inside DNA or using DNA sequences to create encrypted text can only be decrypted if the correct key or sequence is known. The majority of research work has shown that based cryptography (DNA) is more resistant to cryptographic attacks in these situations.

In this paper, we present a symmetric cryptosystem, which consists of encrypting and decrypting DNA-based text. Our contribution consists in defining a new symmetric encryption / decryption approach while basing its approach on the biological characteristics of DNA.

Our system is based on the DNA sequences which in our case a human chromosome from which we generate the sub keys which will be used in the encryption phase. Some biological operations are added to the encryption process such as transcription and translation. A presentation of DNA and DNA-based cryptography is the subject of the following section and we present our approach in detail in the third section followed by a discussion.

2 LITERATURE SURVEY

Computer security generally targets: confidentiality, data integrity, authentication, non-repudiation [24]. Encryption is the art of hiding information, with the development of its science, methods have been developed to reverse it. Encryption is in turn divided into two main branches, the symmetric and asymmetric encryption systems. The symmetric encryption which is the subject of our approach is characterized by identical encryption and decryption keys (Figure 1) [7].
Deoxyribonucleic acid, or DNA, is the chemical component that makes up the chromosomes found in the nucleus of a cell. It is he who is responsible for the transmission of characters from one generation to the next generation. It is made of sugar (deoxyribose), phosphate and the following four bases: adenine (A), cytosine (C), guanine (G) and thymine (T). (Figure 2) [3].

There are different techniques of cryptography DNA that has been developed. In 1994, Adleman [2] laid the foundation of DNA informatics by providing
solutions to combinatorial problems using molecular computation Lipton [6],[20] extend the research of Adelman by solving another NP-complete problem called “satisfaction” by using DNA molecules in a test tube to encode the graph for 2-bit numbers and break one of the symmetric key algorithms used for cryptographic purposes known as DES (Data Encryption Standard). They performed biological operations on the DNA strands, they broke DES in just 4 months.

Based on the work of Adelman, in 1997 Ouyanag et al. [21] showed the effectiveness of DNA by generating solutions of NP-complete problems. DNA cryptographic approach based on the “one-time-pad” molecular theory and performed the encryption / decryption of the 2D image is developed by Chen [15].

Amine et al. [4] proposed a symmetric key based cryptographic DNA approach, where key sequences are obtained from the genetic database.

In 2008, Verma et al. [27] proposed a new paradigm for secure routing in ad hoc mobile networks (Manet) that uses the pseudo-DNA cryptography approach to secure networks. ad hoc. By transforming the message into a DNA format then it goes through an mRNA transcription phase that will be translated into proteins (translation) which is the result of encryption. This encrypted Text is sent through a secure channel to the recipient and a symmetric key with one-time pad is used at the endpoints (encryption and decryption).

The use of public-key encryption technique that uses DNA synthesis, digital coding of DNA, and PCR amplification to provide security during communication is the subject of the algorithm proposed by Cui et al. [8]. This encryption scheme has a great deal of confidentiality. In the same idea, Lai et al. [19] proposed a public key system crypto DNA that is based on microarrays. The chip technology in which the DNA chip is made with probes. One set of probes is used for the encryption process and another for the decryption process.

Kumar and Singh [18] proposed a new algorithm data based on DNA sequences. They explained this algorithm using a simple example of “HELLO” as plaintext and generate a 350-bit key that is 70 times longer than plain text and perform encryption and decryption on plain text using symmetric key encryption

Pramanik and Setua [23] proposed a novel technique of parallel DNA cryptography using the molecular structure of DNA and the hybridization technique that would certainly minimize the requirement the delay.
In 2013, Tornea and Borda [26] proposed DNA-based ciphering based on DNA indexing. DNA cryptography and deep learning are jointly used in DNA cryptography. Kalsi et al. [16] introduced a method for key generation based on the theory of natural selection using Genetic Algorithm with Needleman-Wunsch (NW) algorithm. They proposed a method for implementation of encryption and decryption based on DNA computing using biological operations transcription, translation, DNA sequencing and deep learning.

In 2018, Tiwari & kim [25] propose a new hybrid DNA-encoded Elliptic curve cryptography scheme that provides multilevel security to secure a IoT environment.

Benyahia et al [5] propose a symmetric cryptography system based on DNA called Stegano-DNA- which operates under two main modules: scrambling and encryption. In [22], the authors propose a symmetric crypto-system which consists in cutting the message to encrypt/decrypt in blocks of characters and use a symmetric key extracted from a chromosome for encryption and decryption. In studies [17], Khobzaoui et al presented a symmetric cryptosystem that involves segmenting the message to be encrypted/decrypted into character blocks and using a symmetric key derived from a chromosome for both encryption and decryption processes. Abdelkader et al [1] presented a novel two-level encryption/decryption scheme that merges the principles of cryptography and steganography is proposed.

The various studies reviewed have demonstrated diverse uses of DNA for information encryption while leveraging its advantages. In this article, we propose a cryptographic system based on DNA sequences for the generation of subkeys that will be used in the different phases of encryption.

The next section discusses our approach in detail.

3 PROPOSED CRYPTO SYSTEM

Our algorithm is symmetrical (with secret key) acting by block of 128 bits, the plain text is first cut up into blocks of 128 bits, each block will be encrypted using an encryption key of 128 bits extracted from a long DNA sequence which is usually a chromosome of the human gene. This key is also used in decryption.
3.1 ENCRYPTION PROCESS

Our algorithm is composed of several phases: an initial phase which consists in cutting the text of 16 characters which will then be coded in binary. The DNA transformation phase consists of converting the block into DNA base (A, T, C, G) and carrying out DNA transformation such as transcription and translation. In addition, it includes a sub-key generator which generates nbr (number of blocks) under 128-bit keys from a P position and a DNA sequence. The procedures for the encryption phase are illustrated in Figure 3.

3.1.1 Block Extraction and Coding

In this phase, the plain text goes through the segmenting or cutting into blocks step, blocks of 16 characters are created. The next step is to code these characters in binary (0,1) based on ASCII coding, the result of this step is blocks of 128-bit.

3.1.2 Nucleotide Base Coding

In DNA cryptography, the binary data coded digitally are coded by the combinations of the two states 0 or 1. The coding of DNA can be obtained using four types of bases such as adenine (A), Thymine (T), cytosine (C) and guanine (G). Table 1 show the digital coding of DNA.

<table>
<thead>
<tr>
<th>DNA Base</th>
<th>Binary Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>00</td>
</tr>
<tr>
<td>T</td>
<td>01</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>11</td>
</tr>
</tbody>
</table>

Source: Authors
3.1.3 Transcription

In biochemistry, the transcription process begins when an enzyme called RNA polymerase (pol RNA) attaches to the template DNA strand and begins to catalyze the production of complementary RNA, called mRNA. A single strand DNA copy is produced, ready for the translation process as shown in Table 2 and Figure 4.

<table>
<thead>
<tr>
<th>Original Nucleotide</th>
<th>Complementary Nucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>U</td>
</tr>
<tr>
<td>T</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
</tr>
</tbody>
</table>

Source: Authors
3.1.4 Generation of sub-keys

Our algorithm uses a key composed of two parts. The first part is a long integer p which allows to calculate the position pos in the sequence from which, we start the generation of sub keys using the formula (1)

\[ pos = p \mod \text{length(sequence)} \]  

(1)

where:

- \( p \) – a long integer generated by the emitter
- \( \text{sequence} \) – is the DNA sequence from which the subkey is extracted, which is usually a chromosome of the human gene
- \( pos \) – the starting position in the sequence that the extraction of sub-keys begins

The second part is the chromosome number (identifier) of the human gene, it is the identifier of the sequence which is generally a 6 to 8 character string in GenBank [11]

The 128 bits sub-keys is 64 are generated from the pos position in the sequence defined by the chromosome used (identifier)

3.1.5 Biological xor

It is an operation defined between nucleotid bases according to the following Table 3.
3.1.6 Generation of index table

From the key $p$ (already seen in section 3), a key $k$ is calculated by the formula:

$$ k = p \mod 256 $$

where:

$p$ – a long integer generated by the emitter

From the value of $k$ an index table of size 256 value is generated where each value of the table corresponds to a combination of A, U, G and C. Figure 5 shows the table generated for a value of $k = 121$. 

![Figure 5. Index table for $k=121$ (extract)](source)
3.1.7 Translation

This procedure is also a substitution which simulates the process of translation of the central dogma. Recall that translation is the translation of a strand of mRNA into proteins (a series of amino acids).

In this process, the ribosome (the molecule that performs translation) reads the first three bases (encoding) of the mRNA, and using the genetic code table, finds the appropriate amino acid, then the following three bases, so on until meeting a stop sequence which indicates the end of the translation. First, we changed the genetic code table to an index table generated in the previous section, such as each. In addition, we have increased the number of amino acids to 256 instead of 21. So we need 8 bits (or 4 bases) to code these 256 amino acids. Our table of the genetic generated according to the key. To apply a substitution: We read the first 4 bases (1 byte) of the block, find the value read, in the index table and replace the value in the block with the corresponding index in the index table. This process is repeated until the end of the block.

The result of this phase is our CypherText.

3.2 DECRYPTION PROCESS

The decryption algorithm is used to decrypt a block of numbers, which has been encrypted by the encryption algorithm. The encryption and decryption key being exactly the same. The decryption algorithm does the exact opposite of the encryption algorithm, all modules that exist in encryption are reversed in decryption. Figure 6 shows the decryption process. The key received is composed of two parts: A long integer p and the identifier of chromosome used.
4 SECURITY ANALYSIS

The security analysis of the proposed cryptosystem is presented in this section.

4.1 BRUTE-FORCE ATTACK

Our algorithm is sense to be executed on an IoT platform composed of several interconnected objects, The task of any cryptography algorithm is to secure data for a very long time which is the case in our algorithm. Firstly, we use sub-keys come from a genetic data base of human chromosome GenBank which is a
collection of all publicly available DNA sequences. It contains approximately 140,000,000 DNA sequence records which makes the technique of attack by attempted key testing longer, almost $2^{128}$ starting positions which makes the combination of sub-keys endlessly.

Secondly, the generation of the index table which is necessary in the translation phase requires a key from 0 to 256 which also makes the combination creation phase longer.

The second part in our keys is the id of the human chromosome that will be used in the generation of sub-keys. This identifier is 6 to 8 characters in length in GenBank [11]. Since it carries very important information and the number of chromosomes is limited, we suggest sending it by a route that uses encryption safer.

We can judge that our key makes our algorithm more robust and more resistant to different attacks.

4.2 KNOWN PLAINTEXT ATTACK

In a known plaintext attack, the intruder endeavors to obtain both the ciphertext and its matching plaintext. The objective is to deduce the secret keys and devise a method capable of decrypting subsequent messages. In the proposed scheme, a Biological XOR operation is conducted between the DNA bases of the plaintext and their corresponding counterparts in the secret key, resulting in a new DNA string. Subsequently, this DNA string is substituted with decimal values during the translation phase, following the index-table provided in Figure 5. Consequently, any two similar plaintexts are encrypted to produce two entirely distinct ciphertexts, thereby bolstering the system against such attacks.

4.3 CIPHERTEXT ONLY ATTACK

During a ciphertext-only attack, the assailant endeavors to obtain access to several encrypted messages exclusively. The primary objective of such an attack is to retrieve as many plaintext messages as possible or to ascertain the secret keys through guesswork. Once the encryption key is compromised, it becomes feasible to decrypt all messages encrypted using this key. In the proposed scheme, the index tables (Figure 5) utilized in the translation phase and the DNA sequences employed for sub-keys generation are generated anew for each data transmission.
Consequently, the DNA sequences produced for two similar ciphertexts during decryption are entirely distinct. Consequently, the plaintexts differ as well, thereby fortifying the proposed scheme against this type of attack.

4.4 DIFFERENTIAL CRYPTANALYSIS ATTACK

A differential cryptanalysis attack is regarded as a method for analyzing pairs of plaintext and corresponding ciphertext to compromise the key, or more precisely, to minimize the time needed to obtain it. The proposed scheme is safeguarded against this attack as a new secret key is generated using various attributes of the user (such as the starting position and DNA sequence) and a generated index table. Consequently, the same plaintext undergoes transformation into different ciphertexts during each transmission, thus reinforcing the security of the proposed scheme.

4.5 MAN-IN-THE-MIDDLE ATTACK

In a man-in-the-middle attack, the attacker intercepts communications between the sender and receiver, allowing them to eavesdrop on or modify the traffic between the two parties. Attackers may employ this tactic to obtain personal information, spy on individuals, disrupt communications, or corrupt data. In the proposed scheme, the encryption /decryption key is shared securely between the sender and receiver. Subsequently, the text is encrypted using this shared key before transmission to the intended recipient. Only the recipient possessing the key can decrypt the message. In this scenario, an attacker cannot simply intercept and alter communications between the sender and receiver without the capability to decrypt or re-encrypt the message. Thus, the proposed scheme effectively withstands such attacks.

5 PERFORMANCE ANALYSIS

Any cryptosystem presented should be evaluated to ensure its robustness and its efficiency in encrypting data, so different evaluation metrics have been presented for that purpose. This section aims to evaluate and test the performance of our algorithm. We wish, by using a set of short text, to illustrate the advantages
of using our encryption and decryption system in an IoT environment. Figure 7 presents the functional diagram of the test phase

Figure 7. Block diagram of the testing phase

5.1 EXPERIMENTAL SETUP

As an IoT object, we used in our experiments two Raspberry Pi 3.0, Quad Core 1.2GHz Broadcom BCM2837 64-bit CPU, 1GB RAM.

5.1.1 Dataset

The data used for evaluating the proposed cryptosystem is taken from the 20 Newsgroup dataset [28], which is the collection of newsgroups documents and it contains textual documents.

5.1.2 DNA sequences

Zea mays cultivar B73 chromosome2’ from NCBI[11], which contains 244442276 nucleotide bases.

5.2 RESULTS AND DISCUSSION

5.2.1 Evaluation of Execution Time

The encryption and decryption times are assessed across 5 sets of data extracted from the dataset, categorized according to their size. Encryption and decryption times are determined by conducting 20 experiments for each size.
classification. The average values obtained from the results are then considered, and these findings are displayed in Table 4.

**Table 4. Encryption/Decryption time depending on plain text size**

<table>
<thead>
<tr>
<th>Plain text size (Byte)</th>
<th>Encryption time (Ms)</th>
<th>Decryption time (Ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>14</td>
<td>8.87</td>
</tr>
<tr>
<td>28</td>
<td>31,048</td>
<td>17,639</td>
</tr>
<tr>
<td>56</td>
<td>44,074</td>
<td>27,837</td>
</tr>
<tr>
<td>112</td>
<td>56,583</td>
<td>28,121</td>
</tr>
<tr>
<td>224</td>
<td>72,511</td>
<td>45,888</td>
</tr>
</tbody>
</table>

Source: Authors

From the Figure 8, we note that the encryption (respectively decryption) time increases linearly with the increase in the size of the plaintext (respectively ciphertext), which confirms the theoretical results of the study of the complexity of the algorithm.

**Figure 8. Encryption/Decryption time depending on plain text size**

5.2.2 Key generation time

The key generation time is calculated for the proposed scheme (Table 5). Based on the One-Time Pad (OTP) method, the sender and receiver generate sub-keys from shared DNA sequence and start position.

Figure 9 shows the results of the key generation time.

**Table 5. Key generation time depending on plain text size**

<table>
<thead>
<tr>
<th>Plain text size (Byte)</th>
<th>Key generation time (Ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>6.23</td>
</tr>
<tr>
<td>28</td>
<td>6.84</td>
</tr>
<tr>
<td>56</td>
<td>7.39</td>
</tr>
<tr>
<td>112</td>
<td>8.72</td>
</tr>
<tr>
<td>224</td>
<td>9.10</td>
</tr>
</tbody>
</table>

Source: Authors
5.2.3 Avalanche effect

The avalanche effect is a cryptographic property that asserts that a minor alteration in the plaintext or the secret key causes a considerable change in the ciphertext. To assess this effect in the proposed scheme, 1 bit of the 128-bit key is flipped, and the outcomes are documented in Table 6.

<table>
<thead>
<tr>
<th>Plain text size (Byte)</th>
<th>Change in the ciphertext</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>73%</td>
</tr>
<tr>
<td>28</td>
<td>79%</td>
</tr>
<tr>
<td>56</td>
<td>81%</td>
</tr>
<tr>
<td>112</td>
<td>83%</td>
</tr>
<tr>
<td>224</td>
<td>85%</td>
</tr>
</tbody>
</table>

The results indicate that flipping 1 bit of the secret key results in a minimum change of 73% and a maximum change of 85% in the ciphertext. This demonstrates the robustness and efficiency of the proposed cryptosystem.

6 CONCLUSION

We presented a symmetric approach to cryptography based on DNA sequences. Our approach in its phase of generation of sub-keys based on human chromosomes which have characterized by their lengths, the representation under nucleotide bases, transcription and translation are the subject of the second phase of our process. The proposed cryptosystem supports security against various security attacks.
Experimental results demonstrate that the proposed system exhibits improved encryption, decryption, and key generation times. Furthermore, the avalanche effect is successfully achieved, whereby a minor alteration in the secret key leads to a significant change in the ciphertext. Consequently, the proposed DNA-based cryptosystem is more efficient and secure.

The proposed approach represents an exciting frontier for information security, offering potential benefits that are not only technological but also social and academic: it enhances the security of personal and professional data through the use of an encryption key derived from a DNA sequence, making keyless decryption extremely difficult and thus providing protection against cyberattacks.

As future work, we plan to optimize our approach to address the unique challenges posed by the IoT environment. This involves developing algorithms that require fewer resources for their execution, are suitable for IoT devices with limited processing and storage capabilities, and are energy-efficient.
REFERENCES


[23] PRAMANIK, Sabari; SETUA, Sanjit Kumar. DNA cryptography. In: 2012 7TH INTERNATIONAL CONFERENCE ON ELECTRICAL & COMPUTER ENGINEERING (ICECE), 2012, Dhaka, Bangladesh. 2012 7th International


