Yunpeng ZHANG[1,2] Xianwei ZHANG[1] Bochen FU[1*] Taigang HE[2]

College of Software and Microelectronics, Northwestern Polytechnical University(1), Imperial College London (2)

Construct DNA Symmetric Encryption Based on the Problem of Fragment stitching

Abstract. *DNA cryptography is a new born information security field and have been the international frontier of cryptography emerged with the constantly invasion of traditional cryptography and improvement of research on DNA computing and bio-technology. The security of DNA cryptography is based on difficult biological problems which are unrelated with computing power and thus it is immune to current computer and even future quantum computers. In this paper, we study the difficult biological problem which is the key and basic part in designing of DNA encryption. After analysis of encoding of DNA and mechanism of symmetric encryption, we design a symmetric system using the technology of DNA digital coding and DNA fragment assembly. By implanting DNA secret fragment into the numerous plaintext ones, we destroy the key overlap phase of DNA fragment assembly, thus the normal assembly process cannot be accomplished. In this way, the safety of the encryption scheme is achieved.*

Streszczenie. *W artykule omówiono zagadnienie kodowania DNA i związanych z tym skomplikowanych procesów biologicznych, w odniesieniu do komercyjnych zastosowań. Z uwagi na duży postęp w badaniach nad DNA oraz rozwój klasycznej kryptografii, kodowanie DNA stało się nową, dostępną kategorią. Zaproponowana metoda implementacji tworzy bezpieczny schemat kodowania. (**Tworzenie kodowania symetrycznego DNA w oparciu o zagadnienie zszywania fragmentów**).*

Keywords: DNA cryptography, difficult biological problems, symmetric encryption
Słowa kluczowe: kryptografia DNA, problem biologiczne, kodowanie symetryczne.

Introduction

At present, the tasks of information security have become more and more important [1]. Cryptography is the most important component and infrastructure of communications security and computer security [2]. There is irrelevant between cryptography and molecular biology originally, but with the in-depth study of the modern biotechnology and DNA computing [3], these two disciplines began to work closely together.

DNA Technology

The fragment of DNA which has the genetic message is known as Genes, the Base two pyramiding and two purenesses. They are Cytosine, Thymine, Adenine and Guanine(C, T, A, G). DNA cryptography is a subject of study how to use the DNA as an information carrier and use the modern biotechnology as measures [4] to transfer the ciphertext into plaintext. Thus, biotechnology plays an important role in the field of DNA cryptography. The most popular biotechnologies are Polymerase Chain Reaction [5], Electrophoresis [6], DNA fragment assembly [7], and DNA chip technology [8]. When we do some research in this paper, we will use DNA fragment stitching software--DNA Baser Sequence Assembler. It is used for splicing DNA fragments.

Biological Problems

(1) Biological Problem

In literature [9], the author proposed a biological problem -- It is difficult to make a completely accurate sequencing decipher by obtain a unknown mixed DNA (PNA) probes, which only has different arranged of its nucleotide, on DNA chips(microarrays). Then the author had a discussion for this problem. And he proposed a non-deterministic symmetric encryption system -- DANSC [9] based on this problem. Generally speaking, the biological problem in literature [9] depends on the sequencing technology which is still in the primary stage and has its own weaknesses. And this will make a hidden danger when we build the encryption scheme.

DNA fragment assembly is a technology which needs to reconstruct a large number of DNA fragments into the original long chain of DNA. There have no exact algorithm can solve this problem [7]. As literature [7] stated, DNA fragment assembly can be divided into three phases: Overlap, Layout and Consensus. Overlapping is the basic and the key links of fragments splicing. At this stage, we comparing all of these fragments and finding out the best/long overlapping fragments of the began of some fragments and the end of another fragments. Obviously, if the stage cannot be finished properly, the following two stages also cannot be started. So it is extremely difficult to complete the assembly fragments.

(2) Demonstrate of the Difficulty Problem

We assume that a long-chain DNA sequence is ATCGA. Cutting this DNA randomly after copying it for several times, and then we can obtain a large number of fragments. And these contain three kinds of fragment as follow: ATC (use □ to express), TCG (use □ to express), and CGA (use □ to express). Obviously, the overlapping part of □ and □ is TC and CG for □ and □. The situation of the normal Layout of these fragments through identifying overlap is shown in Fig.1 (a), and we can obtain ATCGA -- the original long-chain DNA successfully. Now we consider this situation: Adding some additional fragments between the two sides of the fragment. Shown as Fig.1 (b). We add an X to the left side of □ and a Y to the right side of □, and then we adding an X to the left side and a Y to the right side of □. Obviously, these three new fragments we obtained have lost the character of overlap. And we cannot complete the layout. Then we can also consider adding some fragments in the middle of the fragment (shown as Fig.1(c)), similarly, we cannot complete the layout.

Fig.1. the overlap and layout of DNA fragments assembly: (a) normal (b) adding at both sides (c) adding in the middle

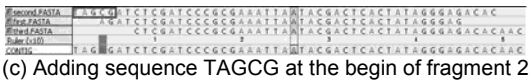
Now we use DNA fragments assembly software -- DNA Baser Sequence Assembler to have a simulation:

It is can be found through the above chart: Fig.2 (a) is the three fragments overlap in a normal situation, and finish overlapping successfully. Fig.2 (b) is shown that we deal with fragment 1 at its end and it has error in the stage of overlap. But we finally obtain the initial long-chain DNA by

using the error correction of the software. Fig.2(c) show that we ultimately failed to get the initial long-chain DNA, because there are some errors appears in overlap stage.

(a) Normal

(b) Adding sequence TAGCG at the end of fragment 1



(c) Adding sequence TAGCG at the begin of fragment 2

Fig.2. the result of fragment overlaps for all situations

The Coding Rules

The molecular weight of the four bases is: $C = 111.10$, $T = 126.12$, $A = 135.13$, $G = 151.13$. So the encoding format is 0123/CTAG by the size of molecular weight.

Coincidentally, according to digital computing, two pairs of the complementary figures are equal; it is means that $0 + 3 = 1 + 2$. And the value of the molecular weight of two pairs of the complementary are equal, it means $C + G = 262.23$, $A + T = 261.25$. In the digital encoding of binary of the bases, the number in first location known as the encoding bits of structure. When the first number is 1, we encoding purine bases. When the first number is 0, we encoding pyramiding bases. The last location is the encoding bits of functional gene. When the last number is 1, it representing keno group. And when the last number is 0, we encoding amino group. The combination of C and G is the complementary combination of 0 (00) and 3 (11), there are three hydrogen chain totally, belongs to the strong hydrogen bonding. The combination of T and A is the complementary combination of 1 (01) and 2 (10), and there are two hydrogen chain totally, belongs to the weak hydrogen bonding.

Construct DNA Symmetric Encryption

(1) Symmetric Encryption

Symmetric encryption also known as symmetric encryption algorithms or private key cryptographic algorithm. The private key cryptography can be divided into two major categories by encryption mode. And the two parts is: sequence cryptography and group cryptography. The advantage of private key cryptography is its fast speed of encryption. And it also has disadvantaged that is: (1) Distributing private key costs a lot and is very complex. (2) It cannot be used for digital signatures.

(2) Encryption Scheme

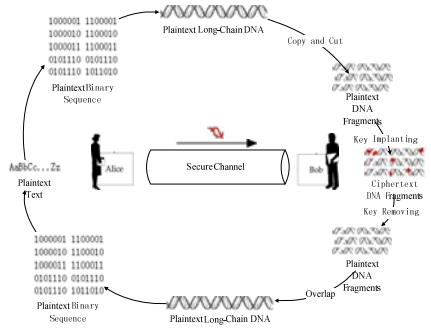


Fig.3. the whole process of the encryption scheme

Fig.3 has show the whole process of the encryption scheme. Now we will description the process of the encryption scheme step by step: (1) Alice will need to convert the plaintext into DNA chain (shown in Fig.4): through the software of the encoding technology of the third-party, first she needs to convert the plaintext into binary code sequence M'. And then translate M' into DNA chain. Thus, the production of plaintext DNA chain is finished. (2) Assuming there are N fragments after cutting of the plaintext DNA chain, the number of nucleotide bases of the key is KL, the copy times of the long-chain is M, the number of the fragment of KL consecutive bases of the plaintext DNA chain is L. Because the cutting is absolutely random. So after that, the number of the fragment which has L internality in large number of fragments is $n = [2/(KL+1)]*M$. Then we need the number of the short-chain key is $num = [N/n]$. Based on the analysis above, Alice needs the short-chain of the number is num and the length is KL as a key (we should ensure that there are no scission in every plaintext DNA long-chain). And then send the key to Bob through a secure channel. (3) After we copy the long-chain DNA for several times, we need cutting it into many fragments randomly. And embedding the short-chain key into them. Then starting the encryption process (Fig.5): First we pick up n DNA fragments from us obtained. And embedding the first short-chain key into each of them. After that we pick up another n DNA fragments from the remaining. And embedding the second short-chain key into each of them. And so on, since the last $n'(n' \leq n)$ that is remaining, and we embedding the num-th short-chain key into each of them. When finish the embedding for the entire DNA fragments, Alice will send them to Bob. (4) When the receiver Bob obtain the ciphertext, he can easily find out the DNA fragments which are embed the key by the help of the key. After remove these keys, he will start DNA fragment assembly to obtain the plaintext DNA long-chain. Then translating it into binary sequence, and through some software, he will obtain the plaintext.

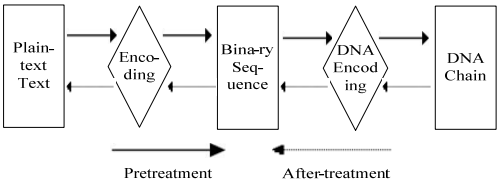


Fig.4. Data Pre/Post-Processing Flow

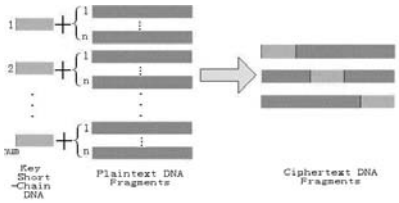


Fig.5. the Process of the Encryption Schematic

(3) Security Analysis

It is easy for the receiver to obtain the plaintext DNA long-chain if he has the key. But for the attacker, he doesn't know the keys, how is the result after the DNA fragment assembly? Restructuring the key DNA fragments and if we obtained the restoring DNA chain:

AGTTATTTAGATGTCTAGCGCCTATGTGCGCAGCCACCTATTTACTTACGTAACCACCTA
GAGAGCCTATGGGTAGATATTCACCTAACTACTTGTATAATTAGATATGCGCCTACGCCC
ACCTGCGTATTTGTATATTTGCATACTTAGCCACCTAACTTAGCGCCTATGCTCTAGATAT

Fig.6. Obtained DNA Chain Using Key DNA Fragments Assembly

By the use of comparison program of DNA chain, we found that the DNA chain in Fig.7 has the similarity of 8.0% with the plaintext DNA long-chain. And the similar similarity is shown in Fig.7 (Part of the result.).

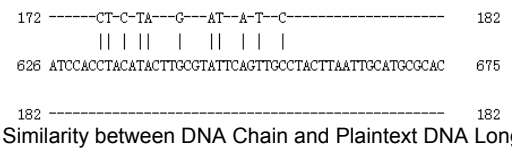


Fig.7. Similarity between DNA Chain and Plaintext DNA Long-chain

After study the consequence, we found that there are some errors in the stage of overlap and layout, and the typical ones are shown in Fig.8:

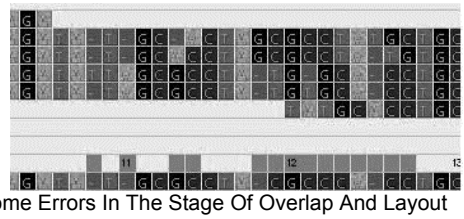


Fig.8. Some Errors In The Stage Of Overlap And Layout

The plaintext of which the attacker is recovered is shown in Fig.9:

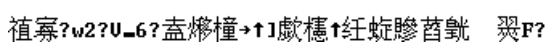


Fig.9. the Plaintext Recover by Attacker

Suppose the ciphertext DNA fragments were obtained by an attacker. He may use the following method to try to obtain the plaintext information: Attack Option One: Sequencing all the DNA fragments. In this encryption scheme, the final ciphertext is a lot of DNA short-chain which are being treated. And the original overlap part has been destroyed which is a necessary condition when we want to have the DNA fragment assembly successfully. Therefore, if we do not know the keys, in this situation, we cannot obtain the plaintext DNA long-chain by DNA fragment assembly for the ciphertext. Experiments show that what they obtained is only a short part of the plaintext DNA chain. Attack Option Two: Exhaustive keys. Suppose the minimum fragment's length is M after the cutting, the shortest length of the possible key DNA chain is N. So the length of the key of the explore attack is in the range between N to M. theoretically, there are $4N+4N+1+...+4M-1+4M=[(4M-N-1)4N]/3$ possible combinations totally. Suppose the length of the fragment is 40 – 100 after cutting, the length of key is 5. So theoretically, the total possible combinations may be $[(435-1)45]/3 \approx 3 \times 1023$. Therefore trying to get the key short-chain is extremely difficult. Attack Option Three: Making an attack when known part of the plaintext. The encryption scheme in this paper can translate the same plaintext into the same DNA sequence for each time in the pre-processing stage. So it is arousing the defects in the encryption scheme in this paper. And we need more work on it. Overall, the encryption scheme with the biological difficulty has improved the security level. But still cannot resist known plaintext attacks very well.

Conclusions

The authors study several difficult biological problems which is the key and basic part in designing of DNA encryption, and Interpretat our own difficult problem which is that it is difficult to assembly when overlaps are destroyed and prove its difficulty with examples in the paper. The second step is analysis of encoding of DNA and mechanism of symmetric encryption; the authors design a symmetric system using the technology of DNA digital coding and DNA fragment assembly. The security analysis proves that the scheme has high confidential strength. In the future, the authors will research some issues deeply in DNA cryptography research field that is rapid development.

Acknowledgments

This work is supported by Science and Technology Development Project of Shaanxi Province Project (2010K06-22g).

REFERENCES

[1] Xiong Fuqin, Cryptography Technology and Application, Science, 2010 (10)
[2] Stallings, William, Cryptography and Network Security [M].4thEd, Prentice Hall,2005,6.
[3] C Popovici ,"Aspects of DNA Cryptography", Annals of the University of Craiova Mathematics and Computer Science Series, Vol37, No3 (2010)
[4] Beenish Anam, Kazi Sakib, Md. Alamgir Hossain, Keshav Dahal, "*Review on the Advancements of DNA Cryptography*", eprint arXiv:1010.0186, 10/2010
[5] http://en.wikipedia.org/wiki/Polymerase_chain_reaction, 2011.1
[6] http://baike.baidu.com/view/25110.htm, 2011.1
[7] Luque, G., & Alba, E. (2005). *Metaheuristics for the DNA Fragment Assembly Problem.* International Journal of Computational Intelligence Research, 1(2), 98–108.
[8] http://www.docin.com/p-24661456.html, 2011.3
[9] Lu Mingxin, Lai Xuejia, Xiao Guozhen, *Symmetric Encryption Method of DNA Technology-based [J].* Science E: Information Science, 2007, 37（2）:175-182.

***Authors**: Prof. YunPeng Zhang, College of Software and Microelectronics, Northwestern Polytechnical University, 710072 Xi'an, China, E-mail: poweryp@163.com. Xianwei Zhang, 2007303118@163.com; Bochen Fu, fuxingfufxy@126.com; Taigang HE, t.he@imperial.ac.uk*