Biological One-way Functions

Qinghai Gao [1], Xiaowen Zhang [2], Michael Anshel [3]
gaoj@farmingdale.edu zhangx@mail.csi.cuny.edu csmma@cs.ccny.cuny.edu

[1] Dept. Security System, Farmingdale State College / SUNY, Farmingdale, NY11735
[2] Dept. Computer Science, College of Staten Island / CUNY, Staten Island, NY10314
[3] Dept. Computer Science, City College of New York / CUNY, New York, NY10031

Abstract: Biology has been a rich source of inspiration for computer security professionals. The central dogma of biology contains intronization cipher and substitution cipher with three many-to-one mappings. In cryptography, substitution cipher has been used long before human identifies DNA. However, it seems that no enough attention has been paid to the intronization cipher because it alone does not make a good cipher. In this paper, the intronization cipher with pseudo-random sequence and substitution cipher with modified genetic code are proposed to improve the security of the intronization cipher. Inverse mapping from protein to RNA and from ciphertext to plaintext are extremely hard; therefore we prefer to treat these ciphers as two biological one-way functions.

Key words:
CDB, BOWF, DNA, splicing, intronization, substitution.

1. Introduction

One-way functions play an essential role in information security. Two generic categories of one-way functions have been proposed in the literature, mathematical one-way function [1] and physical one-way function [2].

Biological systems have been a rich source of inspiration for computer security professionals [3]. Gray [4] named a biological computer system consisting of the following components: several thousand microprocessors → ribosomes; DRAM memory → DNA; program code organized into ~150,000 subroutines → genes; power supply →mitochondria. Genetic algorithm, which imitates the principles of biological evolution by applying three basic operations: selection, mating and mutation, has been used for cryptanalysis of ciphers, for development of cryptographic primitives, and for cryptographic protocol design [5]. Artificial neural network has been proposed for cryptanalysis [6], key-exchange protocol and stream cipher [7, 8]. Artificial immune system, which uses three immunological principles: negative selection, clonal selection, and immune network theory, has been proposed for virus detection [9, 10], intrusion detection [11], and steganography [12]. DNA-based biomolecular cryptography [13, 14] and steganography [15, 16] also attract attentions of researchers.

The Central Dogma of Biology (CDB) contains three many-to-one mappings within two ciphers: intronization cipher and substitution cipher. With our modification and enhancement to

the two ciphers, we obtain two one-way functions. We call them central dogma inspired biological one-way function, which is a descriptive term for information security researchers to capture the essence of the CDB.

The rest of the paper is organized as the following. In section 2, we analyze the encoding and decoding processes in the CDB and introduce biological one-way function (BOWF). Section 3 proposes a pseudo random sequence based intronization technique. Section 4 proposes a substitution cipher using modified genetic code. Section 5 concludes the paper and proposes future research.

2. Biological One-way Function

The Central Dogma of Biology, transcription of DNA to RNA and translation from RNA to protein can be described in Figure 1.

Figure 1 Many-to-one mappings in the Central Dogma of Biology

The protein production process shown in Figure 1 contains two ciphers with three many-to-one mappings.

Intronization Cipher

The first cipher is the intronization cipher which refers to the two independent mappings between RNA (plaintext) and DNA (ciphertext).

The first mapping from DNA to RNA (1:M) refers to the *splicing operation*. In eukaryotic cell, only less than ten percent of the entire DNA sequence is directly used for protein coding. That is to say, large amount of intron (non-coding) regions exists in DNA [17]. Modern biologists believe that intron (non-coding) regions have functions [18]. One common problem that biologists face is how to determine the splicing sites. Two phenomena, Alternative Splicing - same gene splices differently and produces different proteins, and Nested Gene - gene located within intron of another gene and transcribed in opposite direction, make it more difficult to find introns [19, 20]. Since each DNA sequence can produce many different RNA sequences, the process of finding and splicing out introns is the first **one-way transformation**.

The second mapping from RNA to DNA (L:1) refers to the *intronizing operation*. In biology, reverse transcriptase (aka, RNA-dependent DNA polymerase) transcribes single-stranded RNA into single-stranded DNA. Since many different DNA sequences can be generated from a single RNA sequence by inserting different introns, i.e., it is the second **one-way transformation**.

To secure a plaintext with the intronizing operation, we need to find a method that is easy to insert introns into the plaintext but difficult to remove them from the ciphertext. Section 3 introduces a pseudo random sequence based intronizing method. Note that it has some similarity to cryptographic key based steganography.

Substitution Cipher

The second cipher is the substitution cipher which refers to the mapping between mRNA and protein. As [21] points out, "*the genetic code is a substitution cipher, where codons are*

translated into amino acids. The substitution cipher has been known for about 50 years, but a logical origin of the cipher is still unknown. ... as a cipher, it is a molecular form of cryptography: meaning encoded in one molecular sequence and decoded into another."

In the Genetic Code (see Appendix), there are 62 codons (triplets of 4 letters A, U, G, C, 4x4x4) coding for 20 amino acids and the remaining 2 codons signaling the stop of translation. Generally there are 2, 4, or 6 codons coding for each amino acid. On average the ratio of RNA codon to amino acid is 3:1. Therefore, the number of possible RNA sequences for a given protein sequence of length $\boldsymbol{n}$ will be up-bounded to 3^n. Due to the redundancy of codons, in theory it is difficult to reversely translate a protein sequence into an mRNA sequence. Therefore, this mapping is the third **one-way transformation**.

Since RNA is a base 4 system and protein is a base 20 system, the mapping from RNA to protein is, in fact, a combination of substitution and fractionation.

One observation is that it is insecure to directly use the Genetic Code to encrypt plaintext of human languages because the 3-lettered codons generally start with two same letters in the same order. Therefore, on average reverse translation of a protein sequence could correctly reproduce about two thirds of its RNA sequence. With the redundancy of human languages, ciphertext can be decrypted easily. However, the substitution cipher can be made secure by artificially modifying the Genetic Code. Refer to Section 4 for more details. Since we are dealing with biology-inspired ciphers, the two concepts of biological key and biological attack are worth mentioning.

Key

Encoding and decoding biological information requires a key or a set of keys, which may include RNA molecules, proteins, and enzymes, among other things. In some cases environments also play some roles of a key. One example is that the eggs of crocodiles might hatch into male or female, depending on the environmental temperatures.

Attack

The weakest link expressed in the Central Dogma of Biology is the intermediate code, mRNA, which can be viewed as the plaintext. One example of a biological attack starting with RNA is the HIV virus. The genetic code of the HIV virus is RNA, which is unable to replicate outside of living host cells. After HIV infects human cells, the genetic information in its RNA is reversely transcribed into the DNA of human host cell and becomes embedded in the host DNA. Through replication of the host DNA, HIV viruses are reproduced.

3. An Intronization Cipher with Pseudo Random Sequence

As a cryptographic technique, intronization adds introns (aka, non-information carrying symbols) into plaintext to create ciphertext. Different methods can be used to insert introns. In this paper we propose using random sequence for intronization.

Such an example is given in Table 1. Row I and III are binary sequence generated by a pseudo-random number generator (PRNG). We can use a fast PRNG, like a LFSR-based generator. The generated sequence satisfies Golomb's principles [22]: roughly equal number of zeros and ones, and ones and zeros occurred in 'runs' as 1/2 of these runs will be one bit long, 1/4 will be two bits long, $1/2^i$ of these runs will be i-bit long. Here we let 0s be intron positions and 1s be exon positions. Given a plaintext DNA sequence (Step 1 is omitted), e.g.,

ATTGCGGATC, we can intronize it into the sequence shown in Row II. Each X can be A, T, G, or C.

In Table 1, Row IV is obtained with **two-step intronization**: the 1s and the 0s of row I are further divided into introns and exons.

Note that the sequence in Row II is independent of the sequence in Row IV, except that both are the ciphertexts for the same plaintext.

*Table 1 Intronization with random sequence**

From the example given in Table 1, we can see that one plaintext can have many ciphertexts due to different choices of X.

The security of the intronization operation partially depends on message expansion rate (i.e., the length of ciphertext divided by length of plaintext.), which is an undesirable results in terms of storage and processing. An ideal way of using intronization would be maximizing the security with limited message expansion rate. To control message expansion rate one of the methods proposed in [23] is called Exon Elimination.

4. A substitution cipher using modified genetic code

As mentioned in Section 2, it is insecure to directly use the Genetic Code to encrypt plaintext of human languages because the 3-lettered codons generally start with two same letters in the same order. Therefore, on average reverse translation of a protein sequence could correctly reproduce about two thirds of its RNA sequence.

One solution to the problem is to shuffle the natural codons among the 20 amino acids. For example, we can randomize and evenly redistribute the 62 codons among the 20 amino acids and assign 3 codons to each amino acid.

A second solution is to design variable-length genetic code, such as the Huffam (do you mean Huffman?) and Fano-Shannon code [24].

A third solution is to use longer (greater than 3) fixed-length codons in the genetic code. With 4-lettered codon we can randomly assign 22 codons to every amino acid since 4*4*4*4=256, 256/20=22 Remainder 16. Given a protein sequence of n amino acids (ciphertext) obtained from this scheme, the number of possible RNA sequences (plaintext) will be 22^n, where n does not have to be a very big number to be considered secure by modern standard.

Note that all proposed modifications make the inverse mapping from protein sequence to RNA sequence extremely difficult. Therefore it is better to use the modified substitution cipher as a one-way transformation rather than an encryption mechanism.

5. Conclusion and future research

Inspired by the universal information processing procedures in nature, we analyzed the encoding and decoding processes in the CDB and recognized three many-to-one mappings with intronization cipher and substitution cipher. Due to the conceptual one-way mappings, we call the ciphers the central dogma inspired biological one-way functions, which are believed to be descriptive for information security professionals.

We further developed the two ciphers to improve their security and proposed methods to apply them for information security. Specifically, we proposed to use DNA, RNA and protein alphabets to represent information, to use pseudo random sequence to intronize plaintext, and to use modified genetic code for information translation.

Future research will be on designing an encryption scheme in which a single ciphertext can be decrypted into different plaintexts based on key(s).

References

[1] Landau, S. (2006). Find Me a Hash. *Notices of the American Mathematical Society*, 53(3): 330-332.
[2] Pappu, R., Recht, B., Taylor, J., and Gershenfeld, N. (2002). Physical One-Way Functions. *Science*, 297:2026-2030.
[3] Williamson, M. (2002). Biologically Inspired Approaches to Computer Security. *HP Technical Report*: HPL-2002-131.
[4] Cray, S. (1996). An Imaginary Tour of a Biological Computer (Why Computer Professionals and Molecular Biologists Should Start Collaborating). *Remarks of Seymour Cray to the Shannon Center for Advanced Studies*, University of Virginia.
[5] Ibrahim, S., and Maarof, M. (2005). A Review on Biological Inspired Computation in Cryptology. *Jurnal Teknologi Maklumat*, 17 (1): 90-98.
[6] Ramzan, Z. (1998). On using neural networks to break cryptosystems. *Technical report*, Laboratory of Computer Science, MIT.
[7] Socek, D., and Culibrk, D. (2005). On the security of a clipped hopfield neural network-based cryptosystem. *Proc. of the 7th workshop on Multimedia and security*.
[8] Volkmer, M., and Wallner, S. (2005). Lightweight key exchange and stream cipher based solely on tree parity machines. *European Network of Excellence for Cryptology Workshop on RFID and Lightweight Crypto*, pp. 102-113.
[9] Goel, S., and Bush, S. (2003). Kolmogorov Complexity Estimates for Detection of Viruses in Biologically Inspired Security System: A Comparison with Traditional Approach. *Proc. in Adaptive and Resilient Computing Security Workshop*.
[10]Lee, H., Kim, W., and Hong, M. (2002). Artificial Immune System against Viral Attack. *Lecture Notes in Computer Science*, 3037: 499-506.
[11] Kim, J. (2002). Integrating Artificial Immune Algorithms for Intrusion Detection. *PhD Dissertation*, Department of Computer Science, University College London.
[12] Jackson, J., Gunsch, G., Claypoole, R., and Lamont, G. (2003). Blind Steganography Detection Using a Computational Immune System: A Work in Progress. *International Journal of Digital Evidence*, 4(1).
[13]Gehani, A., LaBean, T., and Reif, J. (2000). DNA-Based Cryptography. *5th DIMACS DNA-Based Computers*, American Mathematical Society, USA.
[14] Regoli, M. (2009). Bio—Cryptography: A Possible Coding Role for RNA Redundancy American Institute of Physics. *Conference Proc.*, 1101: 368-373.
[15] Risca, V. (2001). DNA-BASED STEGANOGRAPHY. Cryptologia, 1558-1586, 25(1): 37-49.
[16] Clelland, C., Risca, V., and Bancroft, C. (1999). Hiding messages in DNA microdots. *Nature,399:533-534.*
[17] Wu, A., and Lindsay, R. (1996). A Survey of Intron Research in Genetics. *Lecture Notes in Computer Science*, 1141: 101-110.
[18] Nowak, R. (1994). Mining treasures from 'junk DNA'. *Science*, 263(5147): 608 - 610.
[19] Irimia, M., Rukov, J., Penny, D., Vinther, J., Garcia-Fernandez, J., and Roy, S. (2008). Origin of introns by 'intronization' of exonic sequences. *Trends in Genetics*, 24(8): 378-81.
[20] A. Matlin, F. Clark, and C. Smith. Understanding alternative splicing: towards a cellular

[21] White, M. (2004). Geometric Structure of Codon Relationships. Published online.
Available at: http://www.codefun.com/.
[22] Golomb, S. (1981). *Shift Register Sequences*. Aegean Park Press.
[23] Gao, Q., Zhang, X., and Anshel, M. (2008). Introduction to Geometric Intronization as a Security Technique. *International Journal of Computer Science and Network Security*, 8(12): 19-25.
[24] Doig, A. (1997). Improving the Efficiency of the Genetic Code by Varying the Codon Length-The Perfect Genetic Code. *Journal of Theoretical Biology*, 188 (3): 355-360.

Appendix : Natural Genetic Code

Amino Acid	Acronym	RNA Codons	Ratio
Gly	G	GGU, GGC, GGA, GGG	4
Ala	A	GCU, GCC, GCA, GCG	4
Pro	P	CCU, CCC, CCA, CCG	4
Val	V	GUU, GUC, GUA, GUG	4
Leu	L	UUA, UUG, CUU, CUA, CUG, CUC	6
Ile	I	AUU, AUC	2
Ser	S	UCU, UCC, UCA, UCG, AGC, AGU	6
Thr	T	ACU, ACC, ACA, ACG	4
Asn	N	AAU, AAC	2
Cys	C	UGU, UGC	2
Met	M	AUG, AUA	2
Gln	Q	CAA, CAG	2
Asp	D	GAU, GAC	2
Glu	E	GAA, GAG	2
Lys	K	AAA, AAG	2
Arg	R	CGU, CGC, CGA, CGG, AGA, AGG	6
His	H	CAU, CAC	2
Phe	F	UUU, UUC	2
Tyr	Y	UAU, UAC	2
Trp	W	UGG, UGA**	2
STOP		UAG, UAA	2

*Different versions have slightly different codon assignments
** UGA is also a STOP codon.

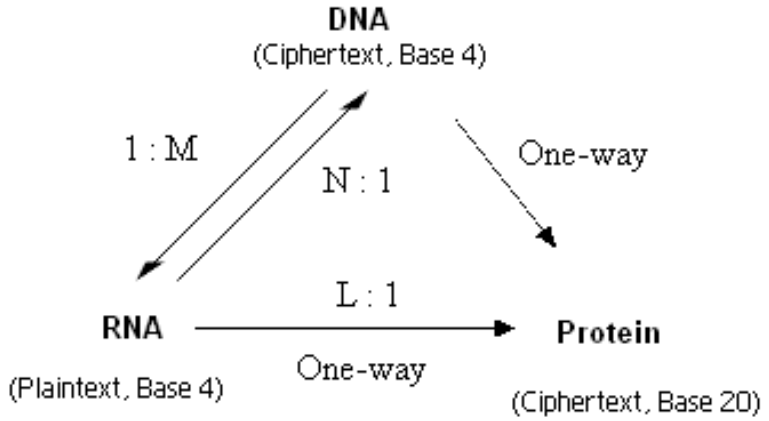

Figure 1 Many-to-one mappings in the Central Dogma of Biology

Table 1 Intronization with random sequence*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
I	0	0	1	0	1	1	0	0	0	1	0	1	1	1	1	0	0	1	1	0
II	X	X	A	X	T	T	X	X	X	G	X	C	G	G	A	X	X	T	C	X
III	0	1	0	1	1	0	1	0	0	1	0	1	1	0	1	0	1	0	0	1
IV	X	A	X	T	T	X	G	X	X	C	X	G	G	X	A	X	T	X	X	C

*X can be arbitrarily chosen from the alphabets (DNA or RNA)