Biological Alphabets and DNA-based Cryptography

Qinghai Gao
Department of Security Systems, Farmingdale State College, SUNY
GaoQJ@farmingdale.edu

Abstract

The redundancy in English language makes cipher easier to attack. The base 4 system of DNA is lack of linguistic properties found in human languages and has higher expressive power per symbol. In this paper, we propose representing information using biological alphabets (including those of DNA, RNA and protein) to enhance the security of ciphertext, using DNA sequence for secure communication and key distribution, and using the chemical information of biological alphabets for steganography – Information Hiding.

Biological Alphabets

Human languages have redundancy. Shannon [1-2] estimated the entropy of written English to be 0.6 to 1.3 bits per character (bpc), based on how well people can predict successive characters in text. Cover and King [3] concluded 1.25 bpc. The redundancy makes cipher using these languages easier to attack.

In biology the genetic information existing in DNA is a base 4 system. Lanctot et al. [4] reported some experimental results on the entropy of DNA. The first experiment was focused on the coding and non-coding regions in E. coli to test the hypothesis that the non-coding regions have a role by showing that they may be more regular than coding regions, which would support the conjecture that non-coding regions in prokaryotes are not junk (About 90% of the genome of higher Eukaryotes is non-coding whereas 15% of the genome of E. coli is non-coding. Presently biologists have found many functions of non-coding regions). The results are followed:

- 1.85 bits/symbol for coding regions (4,090,525 bases)
- 1.80 bits/symbol for non-coding regions (640,039 bases)

The second experiment was conducted to test the hypothesis that highly expressed essential genes have lower entropy than normal genes in E. coli. The results are followed:

- 69 highly expressed essential genes: ~1.752 bits/symbol
- 244 normal genes: ~1.785 bits/symbol

Farach et al. [5] estimated the entropy of introns and exons from human DNA sequences are between ~1.8 bits/symbol and ~2 bits/symbol. Behr et al. [6] estimated and compared the entropies of the Bible and the United Nation Treaties across a series of written natural languages, including English, Spanish, French, Chinese, Korean, Arabic, Japanese and Russian. Their results show that all these languages have similar expressive power. All these research results tell us that DNA has higher expressing power than human language like English.

One question that has been asked was whether DNA is a Language. Based on that Zipf's law (The frequency *f* of each word in a text and its rank *r* are determined by formula $f \propto r^{-k}$ with k close to 1 for all languages), which is followed in every human language, Tsonis et al. [7] did DNA analysis and concluded that DNA does not follow the law. Therefore, DNA is different from human languages.

Tsonis et al. [7] did more statistical analysis by associating the 26 letters of English with the triplets according to the frequency for both coding and non-coding sequences, respectively. They constructed sentences with and without grammatical rules, then randomly shuffle the letters in these sentences to destroy possible structures, and found that shuffled sentences scored homologies with real DNA sequences that are identical to the scores of the non-shuffled sentences. Based on these results, they concluded that DNA sequences show NO linguistic properties. This conclusion is supported by the entropy values reported by other researchers [3-5].

These results tell us that mapping from English to DNA, RNA sequence, and protein sequence will remove some of the redundancy in natural languages and make frequency analysis based attack against some ciphers more difficult. With this finding we propose the following symmetric encryption algorithm.

Symmetric Encryption Algorithm

One generic encryption scheme can be as followed:

- Step 1: Obtain binary string from the to-be protected information
- Step 2: Map it to DNA sequence
- Step 3: Map DNA to RNA sequence (1-to-1 substitution, optional)
- Step 4: Map RNA to protein sequence (3-to-1 compression)

The encryption scheme is not safe if the natural genetic code is used because the redundancy of the natural genetic codons reveals about two thirds of the RNA sequences even the one-way mapping from RNA to protein is applied. In order to make it safer we propose the following two methods.

Method 1: More rounds

- Step 5: Map the protein sequence to ASCII (1-to-8 expansion)
- Step 6: Map ASCII to DNA sequence (8-to-4 compression)
- Step 7: Map DNA to protein (3-to-1 compression)

The second round would make the encryption safer with more substitution and diffusion. However, the message expansion rate will be the ratio of ciphertext length to plaintext length, which is 4/3 for every round.

Method 2: Changing the natural genetic code

As aforementioned, the natural genetic codons of the natural genetic code generally start with two same letters in the same order. On average, reverse translation of a protein sequence could correctly reproduce about two thirds of the RNA sequence. Therefore it is not secure to use it directly.

One solution to the problem is to shuffle the natural genetic codons among the 20 amino acids, for example, randomize and evenly redistribute the 64 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 designs already exist in literature, such as the Huffman and Fano-Shannon code [8], even though these new coding schemes may not be designed for the purpose of information protection but to check the efficiency of the natural genetic code.

A third solution is to use length-fixed longer (>3) codons. For example, with 4-lettered codon we obtain 256 different codons. We can assign 22 codons to every amino acid since 4*4*4*4=256, 256/20=22 Remainder 16.

Note than for each of these modifications some information needs to be added as part of the secret key. With these new designs the biological alphabets based encryption scheme will be much more secure.

DNA-based Cryptography

Adleman [9] started DNA computing in 1994 by solving a small instance of the Hamiltonian path problem in wet lab. His success attracted a great deal of attention [10-12] in the last few years of the 20th century. Gehani et al. [10] studied DNA-based cryptography and designed DNA One-Time Pad in 1999. Since 2000, the interests in DNA cryptography have been limited. One of the reasons is that the wet-lab manipulation of DNA molecules is difficult even for biochemist and biologist. Recently there are some renewed interests [13-15] in this area.

Inspired by the gene recognition with primer, the mRNA splicing, and the one amino acid mapping to multiple codons in biology, we propose some new methods that can be used for information protection.

Secure communication protocol with DNA primer

In biology, genes are recognized by primers (a short DNA sequence). These primers can be viewed as keys to select a sequence. Inspired by the process, we propose a secure communication protocol as the following.

Suppose Alice and Bob each have a copy of a code book, which contains many different DNA sequences. Each sequence can code one message out of a set of possible messages specifically pre-designed. When Alice needs to send Bob a message, she just sends Bob the primer. When Bob get the primer, he lets it anneal to one of the sequences in the shared codebook and decode the message. One requirement for the protocol is that the original codebook should be secure.

The advantage of the protocol lies in that *the real message was never transmitted*. Even the primer sequence is intercepted; the attacker has no chance to know what the real message is. One disadvantage of the protocol is that the sequences in the codebook have to be pre-fabricated.

DNA-based key distribution

Assume Alice and Bob share a secret DNA sequence codebook. Alice can design a sequence that is maximally match only one of the sequences in the codebook, and then send the designed sequence to Bob through public channel. When Bob receives the sequence, he will try to find the maximum match in his codebook.

Assume Bob found the following private sequence from the codebook:
Private:ACTTACCGGGACTGGTAATGGCCGTTAGGATTTGCCAAAGTTTGAACT
Public: ACGTACCGCTTGTGGCATCGGCAATCGATATTTGACTTC GTCCGAAGT

Bob would use the non-matching letters in the private sequence as the encryption key:
--T-----GGAC---T-AT---CG-TAGG-----C-AAA--TT---C-

Knowing the public string only, an attacker has no easy of finding the key except guessing.

DNA-based Steganography

Digital watermarking and image steganography often utilize the pixels of an image as information hiding media. One widely known example is to use the least significant bit of every pixel of an image to hide information. A few papers [16-19] also reported on how to use DNA to hide information. Here we can propose a new method using DNA/RNA sequences to steganographically hide information.

Fig. 1 shows the chemical structures of the bases for DNA and RNA, each of which contains large amount of information that could be used for steganography. Table 1 lists some of information that can be easily represented as small integers. Each row of the table can be used to carry the hidden information with DNA or RNA sequences.

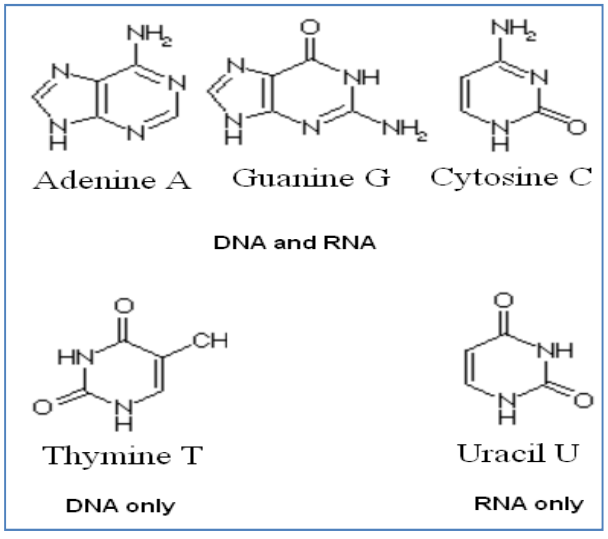

Fig. 1 Chemical structure of the bases of DNA and RNA

Table 1 A few properties of the bases usable for steganography

Bases		**A**	**G**	**C**	**T**	**U**
Formula		$C_5N_5H_4$	$C_5N_5H_4O$	$C_4N_3H_4O$	$C_5N_2H_5O_2$	$C_4N_2H_3O_2$
MW		134	150	109	125	111
Melting Point		360	360	320	316	335
Isolated E-pair		5	7	6	6	6
C	All	5	5	4	5	4
	4^{th}	3	4	2	2	3
	3^{rd}	2	1	2	2	1
N		5	5	3	2	2
H		4	4	4	5	3
O		0	1	1	2	2
N-H Bond		2	3	2	1	1
C-H Bond		2	1	2	4	2
C-N Bond		6	7	4	4	4
C-C Bond		1	1	0	2	1
C=N Bond		3	2	1	0	0
C=O Bond		0	1	1	2	2
Ring Single Bond		6	7	4	5	5
Ring Double Bond		4	3	2	1	1

To use the method two parties can decide on what property to use as information carrier. This new scheme is secure for two reasons: one is the attacker's ignorance; another is the large number of choices of properties.

Summary and future research

In this paper we propose a few new methods to protect information, including representing information using biological alphabets to enhance the security of traditional encryption, using DNA primer for secure communication and key distribution, and using the chemical information of DNA bases for steganography. Future research will be conducted on testing these schemes in broad circumstances.

Bibliography

[1] Shannon, C. (1951). "Prediction and entropy of printed English". *Bell Systems Technical Journal*, 30:50-64.
[2] Mahoney, M. (2000). "The Cost of Natural Language Modeling", *Ph.D. Dissertation*, Florida Institute of Technology.
[3] Cover, T. & King, R. (1978). "A convergent gambling estimate of the entropy of English", *IEEE Transactions on Information Theory*, 24(4):413-421.

[4] Lanctot, J., Li, M., & Yang, E. (2000). “Estimating DNA Sequence Entropy**”,** *Symposium on Discrete Algorithms*.
[5] Farach, M., Noordewier, M., Savari, S., Shepp, L., & Wyner, A. (1995). “On the entropy of DNA: algorithms and measurements based on memory and rapid convergence, *Symposium on Discrete Algorithms.*
[6] Behr, F., Fossum, V., & Mitzenmacher, M. (2002). “Estimating and Comparing Entropy across Written Natural Languages Using PPM Compression”, *Technical Report TR-12-02*, Harvard University.
[7] Tsonis, A., Elsner, J., &Tsoni, P. (1997). “Is DNA a language?” *Journal of Theoretical Biology*, 184(1): 25-29.
[8] 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.
[9] Adleman, L. (1994). "Molecular computation of solutions to combinatorial problem," *Science*, 266: 1021-1024.
[10] Gehani, A., LaBean, T., & Reif, J. (1999). "DNA-Based Cryptography". *Proc.5th DIMACS Workshop on DNA Based Computers*.
[11] Leier, A., Richter, C., Banzhaf, W., & Rauhe, H. (1999). “Cryptography with DNA binary strands”, *Biosystems*.
[12] Figureau, A., Soto, M., & Toha, J. (2000). "Biocryptography", *Medical Hypotheses,* 54(3): 394-6.
[13] Chen, J. (2003). “A DNA-based, Biomolecular Crytpography Design”. *Proc. 2003 International Symposium on Circuits and Systems.*
[14] Nixon, D. (2003). “DNA and DNA Computing in Security Practices – Is the Future in Our Genes”, *GSEC Assignment* Version 1.3, SANS Institute. Available at: http://www.1000projects.com/paperpresentations/cse/DNA/DNA_COMPUTING_1.DOC
[15] Xiao, G., Lu, M., Qin, L., & Lai, X. (2006). “New field of cryptography: DNA cryptography”, *Chinese Science Bulletin*, 51(12): 1413-1420.
[16] Blackman, D. (1993). *The Logic of Biochemical Sequencing*, CRC Press.
[17] Kessler, G. (2004). “An Overview of Steganography for the Computer Forensics Examiner”, *Forensic Science Communications*, 6(3). Available at: http://www.fbi.gov/hq/lab/fsc/backissu/july2004/research/2004_03_research01.htm
[18] Clelland, C., Risca, V., & Bancroft, C. (1999). "Hiding messages in DNA microdots." *Nature,* 399(6736):533-534.
[19] Richter, C., Leier, A., Banzhaf, W., & Rauhe, H. (2000). “Private and Public Key DNA steganography”, *Proc.6th DIMACS Workshop on DNA Based Computers,* University of Leiden, Leiden, The Netherlands.