Vol.1, No.3, 181-184 (2009)
doi:10.4236/ns.2009.13023

Natural Science

Hiding data in DNA of living organisms

Shu-Hong Jiao[1], Robert Goutte[2]

[1]Information and Telecom Department, Harbin Engineering University, Harbin, China
[2]Universite De Lyon , Lab Creatis Lrmn, INSA Lyon, France; goutte@creatis.insa-lyon.fr

Received 12 June 2009; revised 16 July 2009; accepted 18 July 2009.

ABSTRACT

Recent research has considered DNA an interesting medium for long-term and ultra compact information storage and a stegomedium for hidden messages. Artificial components of DNA with encoded information can be added to the genome of living organisms, such as common bacteria. With this approach, a medium for very height densities information storage, watermarks for protection patents of genetically modified organisms (GMOs) and secure public keys for decrypting hidden information in steganocryptography, can be obtained. In this paper, we have selected a Bacillus subtilis gene (tatAD) and use the specific properties of silent mutations to obtain a biologically innocuous product. An adapted code for the message insertion in this gene is proposed.

Keywords: DNA; Genome; Coding; Steganography; Living Organism

1. INTRODUCTION

There has been growing interest in using DNA to store information, one the main reasons being the very high storage densities that can be achieved. The durability of DNA would make it particularly useful for preserving archival material over extensive periods of time (long-term storage). Message DNA has been used in computations in biologic mathematics, in steganography and as a mean of short-term trademarks. These different fields use the specific properties of genetic code and the possibilities to encode artificial information.

However, the artificial introduction of a new gene or the modification of an existing gene in the DNA of a living organism, can involve prejudicial deterioration of the behavior and the reproduction of this species. The solution suggested must thus notably allow a great computer security, but also the genetic consequences of their implementation.

2. THE GENETIC CODE

The genetic code is the biochemical basis of heredity and nearly universal in all organisms (eukaryotes or prokaryotes): humans, animals, plants, bacteria and viruses.

DNA (deoxyribonucleic acid) is a long molecule, with two strands rolled up in a double helix. Each strand is formed by sugar phosphate backbone, connected with single molecules called bases, which contain carbon, nitrogen and cyclical structures. The four bases are known as adenine A, thymine T, guanine G and cytosine C.Any strand of DNA adheres to its complementary strand, in which T substitute for A, G for C, and vice versa. The links between pairs of bases are responsible for binding together two stands together to form the double helix.

The genetic information in DNA is found in the ordered sequence of these four bases (the double helix structure of DNA introduces redundant information, resulting in complementary nitrogenous base links).

Unlike DNA, RNA (ribonucleic acid) is a single stranded molecule and does not form of double helix. The bases are the same that DNA, except U (uracil), which replaces T. The complementarily becomes:

$$U \leftrightarrow A \text{ and } G \leftrightarrow C$$

Messenger RNA (mRNA) carries all the genetic information to ribosome for proteins synthesis by the cell.

A codon is a triplet of three bases (T,C,A and G). With these four letters, 43=64 combinations are possible. With three exceptions, each codon encode for one of the 20 amino acids, used in the proteins synthesis (the three exceptions are TAA, TAG and TGA: codons STOP).

ATG correspond at methionine within the gene; at the beginning of the gene, ATG is also a signal START. Ribosome assembles individual amino acids into peptide chains. Peptides are short chains of amino acids that are linked together. If the number of amino acids in the chain reaches around ten or more, such substances are called polypeptides and large polypeptides are called proteins.

First Position	Second Position of Codou T	Second Position of Codou C	Second Position of Codou A	Second Position of Codou G	Third Position
T	TTT Phe [F] TTC Phe [F] TTA Leu [L] TTG Leu [L]	TCT Ser [S] TCC Ser [S] TCA Ser [S] TCG Ser [S]	TAT Tyr [Y] TAC Tyr [Y] TAA Ter [end] TAG Ter [end]	TGT Cys [C] TGC Cys [C] TGA Ter [end] TGG Trp [W]	T C A G
C	CTT Leu [L] CTC Leu [L] CTA Leu [L] CTG Leu [L]	CCT Pro [P] CCC Pro [P] CCA Pro [P] CCG Pro [P]	CAT His [H] CAC His [H] CAA Gln [Q] CAG Gln [Q]	CGT Arg [R] CGC Arg [R] CGA Arg [R] CGG Arg [R]	T C A G
A	ATT Ile [I] ATC Ile [I] ATA Ile [I] ATG Met [M]	ACT Thr [T] ACC Thr [T] ACA Thr [T] ACG Thr [T]	AAT Asn [N] AAC Asn [N] AAA Lys [K] AAG Lys [K]	AGT Ser [S] AGC Ser [S] AGA Arg [R] AGG Arg [R]	T C A G
G	GTT Val [V] GTC Val [V] GTA Val [V] GTG Val [V]	GCT Ala [A] GCC Ala [A] GCA Ala [A] GCG Ala [A]	GAT Asp [D] GAC Asp [D] GAA Glu [E] GAG Glu [E]	GGT Gly [G] GGC Gly [G] GGA Gly [G] GGG Gly [G]	T C A G

Figure 1.1. The genetic code (DNA).

First Position	Second Position of Codou T	Second Position of Codou C	Second Position of Codou A	Second Position of Codou G	Third Position
T	Phe [F] TT	Ser [S] TC	Tyr [Y] TA	Cys [C] TG	T C A G
C	Leu [L] CT	Pro [P] CC	His [H] CA	Arg [R] CG	T C A G
A	Ile [I] AT	Thr [T] AC	Asn [N] AA	Ser [S] AG	T C A G
G	Val [V] GT	Ala [A] GC	Asp [D] GA	Gly [G] GG	T C A G

1) In yellow: The eight blocks with a single amino acid.
2) Blocks repaired with only two first bases.

Figure 1.2. Simplified code.

3. DEGENERACY OF GENETIC CODE

Many codons are degenerate, or redundant, meaning two or more codons (synonymous codons) may code for the same amino acid. See **Figure 1.1.** Degenerate codons typically differ in their third position. (e.g.: GAA, GAC for glutamine). See **Figure 1.2.**

The degeneracy of the genetic code is what accounts for the existence of "silent mutations". A mutation occurs when a DNA gene is damaged or changed. A silent mutation is a mutation that does not alter the amino acid of a gene, usually because of codon ambiguity.

4. STATE OF ART, PROPOSED CODE AND SPECIFIC PROPERTIES

The first paper on hiding messages in DNA was published by T. Clelland and al (ref 1) in 1999 and involved the insertion of a brief message in a sample of human DNA. In 2001, these same authors published a paper showing of possibilities of long-term storage of information in DNA and used a classical codon encoding for the English alphabet. In 2003, C. Smith and al (ref 3) described a possible code for encrypting data.

In DNA and Boris Shimannovsky and al (ref 4) proposed an interesting and original arithmetic code. Pak Chung Wong and al (ref 5) presented new potential applications for DNA organic data memory.

The advantages of these approaches in production keys were presented by Masanori Arita (ref 6) and Tanaka and all, in 2005 (ref 7). Recently, Nozomu Yachie and al (ref 8) presented a very complete methodology, simple, flexible and robust of data storage based on sequence alignment of genomic DNA of living organism. D. Heider and al publish, in 2007, a program called DNA Crypt whose use is centred on the patent protection of genetically modified organisms (GMOs) (ref 9).

In this work we propose a method of coding that satisfies two conditions:

Limitation of changes in the gene marker

Possibility to transfer the encrypted message in a key, made with a polypeptide chain.

5. CHOICE OF SITE

Bacillus subtilis is a non-pathogenic bacterium, which form pores that can survive in extreme conditions. This bacterium is a model which has already been chosen as storage of information by several teams.

To illustrate our proposed code, we have selected a component of this genome: the gene tatAD (alternate name ycbz). See **Figure 1.3.** This gene is very short; its length is 210 base pairs, which correspond to 70 amino acids. It is a protein classified stable.

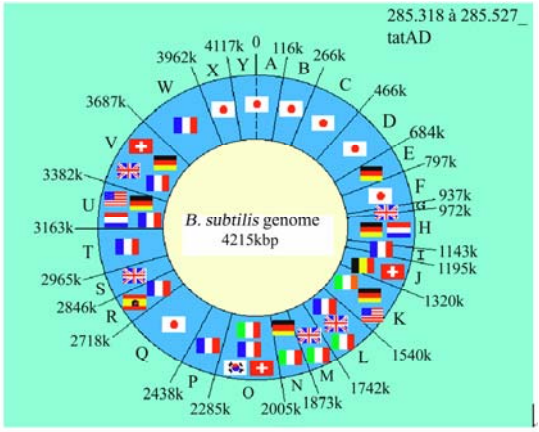

Figure 1.3. Bacillus subtilis.

The following table lists the genetic code of these 70 amino acids

ATG TTT TCA AAC ATT GGA ATA CCG GGC
TTG ATT CTC ATC TTC GTC
ATC GCC ATT ATT ATT TTT GGC CCT TCC AAG CTG CCG GAA ATC GGG
CGT GCC GCG AAA CGG ACA CTG CTG GAA TTT AAC AGC GCC ACA AAC
TCA CTT GTG TCT GGT GAT GAA AAA GAA GAG AAA TCA GCT GAG CTG
ACA GCG GTA AAG CAG GAC AAC AAC GCG GGC

A short list concern only the underlining codons: TCA GGA CCG GGC CTC ……, codons-owned blocks of the **Table 2**, in yellow mark.

6. IMPLEMENTATION

6.1. Encryption

By example, we consider the message "CODING". After translation with ASCII code (with 8 bits per character), this message requires 48 bits.

- C 0 1 0 0 0 0 0 1
- O 0 1 0 0 1 1 1 1
- D 0 1 0 0 0 1 0 0
- I 0 1 0 0 1 0 0 1
- N 0 1 0 0 1 1 1 0
- G 0 1 0 0 1 1 1 1

With the rate of 2 bits per codon, this message requires 24 codons. For his implementation, only 35 codons are selected (codons underlined).The beginning and the end of this message necessarily belongs to these 35 codons, so that only two 6 bits numbers (two times three codons) are needed for their localization.

If we assume that the beginning is located at the 9th codon, and is 48 bits long (24 codons), the end of the message will be located at 32 th codon.

Beginning at 9→ 00 10 01
End at 32→ 01 00 00

Six codons are needed for their localization.

/1 2 3/-//4 5 6//7 8///9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32///33 34 35.

The coding consists in replacing the 3rd base of these codons by A, C, G, T according this table:

A if 00 C if 01
G if 10 T if 11

Before coding we, in the first line have

ATG TTT TCA AAC ATT GGA ATA CCG GGC TTG ATT CTC ATC TTC GTC

After coding we obtain:

ATG TTT TCA AAC ATT GGG ATA CCC GGG TTG ATT CTA ATC TTC GTA

The first 24 codons selected on the short list are:

CCT TCC CTC CCG GGG CGT
GCC GCC CGG ACA CTG CTG
GCC ACA TCA CTT GTG TCT
GGT TCA GCT CTG ACA GCG

After coding, we obtain:

CCC TCA CTA CCT GGC CGA
GCT GCT CGC ACA CTC CTA
GCC ACA TCG CTC GTC TCA
GGT TCG GCC CTA ACC GCT

4 codons (underlined) are not modified by the coding. Only 25 codons of the original gene's 70 codons (tatAD) are modified. All the mutations are particular missense mutations: silent mutations; with its location, the encryption of this message requires 60 bits and 25 mutations. Thus, on average, each silent mutation can insert slightly more than 2 bits of hidden message.

6.2. Decryption

Decryption is particularly easy. After obtaining the beginning and the end of the message from the first 6 codons of the shortlist, it is sufficient to replace, in the sequence of codons corresponding to the message, the last base of each codon by its equivalent in bits:

(A→00, C→01, G→10, T→11)

In this example, a very short gene is used for demonstration purposes, but this method can be applied to longer genes, that are more than 10000 base pairs long (such as the gene srfAA de Bacillus subtilis) which then allows the insertion of messages 50 that are time longer (~2400 bits and therefore 300 ASCII characters).

7. MATERIALISATION OF THE KEY, COMPLEMENTS

7.1. Producing the Key

It is possible to replace the list of 30 codon, which can be synthesized and materialized in the form of a polypeptide, as in the preceding example, with a new list of 60 amino acids. Indeed, if we consider the beginning of the Short list (after coding):

TCA GGG CCC ……..We have the possibility to introduce a peptide chain:

Ser (TCA) Gly (GGG) Pro (CCC) ….

But it is not possible, with only this chain, to rediscover the different codons of the short list in reason of the redundancy of the genetic code.

Ser correspond to TCT TCC TCA TCG
Gly to GGT GGG GGA CGG and
Pro to CCT CCC CCA CCG

For resolve this difficulty we propose the following connections:

T C A G G G C C C
/\ /\ /\
T C X C A X' G G X G G X' C C X CCX'

If X =A and X'=C we obtain:

T C A C A C G G A G G C C C A C C C
Ser His Gly Gly Pro Pro

new peptidic chain. With the chain we can retrieve the three codons of the short list

For this, it is sufficient to conserve uniquely two bases/in each codon.

T C C A G G G G C C C C and to regroup:

T C G G C C
C A G G C C

T C A G G G C C C

for obtain the three original codons.

7.2. Complements

It is possible to associate this encryption method with a binary encryption algorithm (AES, RSA, Blowfish). In this case, the inserted message is not the clear version, but a secure version after using these specific algorithms.

7.3. Simplified Version

If, in spite of the precaution to using silent mutations during the identification of genetically modified organisms, residual biological changes still exist in the host organism, then a simplified version of the method discussed above can be used.

From the gene initially modified by the agrochemical firm (if this gene is accessible to the user), we propose to use this gene as stegomedium, to produce a custom key. This key can allow access to the reference of manufacturer. With this goal in mind, in the preceding example (with the message "CODING" of 48bits), we select 24 codons of the 70 existing in the tatAD gene.

Each codon selected is identified by its rank in the chain (7 bits per codons are necessary, here corresponding to a total of 24x7=168 bits). These 168 bits are the customized key (customizations results from the different possible orders in this gene's chain), The last base of the codon selected (A,T,C or G) allows access to the bits 01, 10 ,00, 11 that form the message. It is important to note that the marker gene is not altered by this operation. Information on origin and dates of manufacture concerning these seeds can be obtained. Security comes both the use of the key and the specific nature of the changes, made voluntarily by the manufacturer on the original gene, for obtain the desired biological effects.

8. CONCLUSIONS

The use of the degeneracy of the genetic code and, in particular, the silent mutations, produces coding that does not practically alter the properties of the inserted gene, nor the characteristics of the host genome (very important conditions when we working with the live organisms). Memorization of the key information and the production of the hidden message in the form of a physical polypeptide provide additional data transfer security, while the coding protocol is being implemented).

DNA is a storage medium extremely effective: it is compact and his signature is innocuousness and secrecy. In the spore form, Bacillus subtilis is resistant to extremes terrestrial and extraterrestrial environment during the interstellar travel, for example (ref 10). In the nearly future, this bioengineering method can be used in routine, for protection of the gene patents, digital copyrights and tool for the marking in biodiversity

REFERENCES

[1] C. C. Taylor, V. Risca, and C. Bancroft, (1999) Hiding messages in DNA microdots. Nature, **399**, 533-534.

[2] C. Bancroft, T. Bowler, B. Bloom, and C. C. Taylor (2001) Long-term storage of information in DNA. Science, **293(5536)**, 1763-1765.

[3] G. C. Smith, C. C. Fiddes, J. P. Hawkins, and J. P. L. Cox, (2003) Some possible codes for encrypting data in DNA. Biotechnology, **25**, 1125-1130.

[4] B. Shimanovsky, J. Feng, and M. Potkonjak, (2003) Hiding data in DNA, LNCS **2578**, 373-386.

[5] P. C. Wong, K. K. Wong, and H. Foote, (2003) Organic data memory using the DNA approach, Communication of the ACM, **46(1)**.

[6] M. Arita and Y. Ohashi, (2005) Secret signature inside genomic DNA biotechnol. Prog, **20 (5)**, 1605-1607.

[7] K.Tanaka, A. Okamoto, and I. Saito, (2005) Public-key system using DNA as a one way function for key distribution. Biosystems, **81(1)**, 25-29.

[8] N. Yachie, K. Sekiyma, J. Sugahara, Y. Ohashi, and M. Tomita, (2007) Alignment-based approach for durable data storage into living organisms biotechnol. Prog., **23**, 501-505.

[9] D. Heider and A. Barnekow, (2007) DNA-based watermarks using the DNA-Crypt algorithm, BMC Bioinformatica, **8**, 176.

[10] W. L. Nison, N. Munakata, G. Horneck, H. J. Melosh, and P. Setlow, (2000) Resistance of Bacillus endospores to extreme terrestrial and extraterrestrial environments. Micbiol. Mol. Bio. Rev., **64(3)**, 548-57.