Hiding Color Images in DNA Sequences

Marc B. Beck*, Roman V. Yampolskiy

Cybersecurity Lab, Department of Computer Engineering and Computer Science, Speed School of Engineering, University of Louisville, Louisville, KY 40292
{mbbeck05,roman.yampolskiy}@Louisville.edu

Abstract

With recent advances in genetic engineering it has become possible to embed artificial DNA strands into the living cells of organisms. With DNA having a great capability for data storage, many methods have been developed to insert artificial information into a DNA sequence. However, most of these methods focus on the encoding of text data and little research has been done regarding the encoding of other media. Few methods have been researched to encode images and most of those are only for black-and-white images. We are proposing an algorithm to insert and extract color images in the form of bitmap files into a DNA sequence in the form of a FASTA file. Results from our experiments show that the proposed method is significantly more efficient than previous approaches.

Introduction

Deoxyribose Nucleic Acid (DNA), the molecule that carries the hereditary information for every living organism, is a double helix with two anti-parallel strands. These strands contain four different nucleotides which are distinguished by the bases adenine (A), cytosine (C), guanine (G), and thiamine (T). Using combinations of those four nucleotides, DNA has the potential to store vast amounts of data within genomes, the length of which can range up to several billion bases (Anam, Sakib, Hossain, & Dahal, 2010).

Certain regions within a genomic sequence translate into genes that produce proteins consisting of amino acids. Marshall Nirenberg (Martin, Matthaei, Jones, & Nirenberg, 1962) first discovered the genetic code, which dictates how the combinations of the four bases of DNA are translated into twenty amino acids.

A sequence of three nucleotides that determines which amino acid will be incorporated next during protein synthesis is called a codon. The four nucleotides can be combined into 4^3=64 unique codons. Each codon, except the three STOP codons TAA, TAG, and TGA (Brenner, Stretton, & Kaplan, 1965; Martin et al., 1962) encodes for one of the twenty amino acids. This means that multiple codons encode the same amino acid, allowing for degeneracy.

It has become possible through recent advances in genetic engineering to insert artificial DNA sequences into the living cells of organisms (Gibson et al., 2010). This allows the insertion of information into the DNA strands of organisms such as bacteria for applications like data storage, watermarking, or communication of secret messages.

DNA as Storage Medium

Researchers are investigating DNA as an ultra-compact, long-term data storage medium (Church, Gao, & Kosuri, 2012; Goldman et al., 2013; Wong, Wong, & Foote, 2003) as well as a stegomedium for hiding information (Smith, Fiddes, Hawkins, & Cox, 2003). Messages are expressed as a series of As, Cs, Gs, and Ts in DNA code as opposed to ones and zeroes in binary code. In order to encode messages in DNA, researchers have developed various algorithms that can either insert a message into an existing DNA sequence (Jiao, 2009), or disguise it as a new one. These artificial DNA strands can be inserted into the genomes of living organisms, which has been proven possible by Venter (Gibson et al., 2010) and others (Jiao, 2009), (Arita, 2004; Brenner et al., 2000; Heider & Barnekow, 2008; Jiao & Goutte, 2008; Yachie, Sekiyama, Sugahara, Ohashi, & Tomita, 2007)].

The first cell with a synthetic genome was created in 2010 by Craig Venter, who led the private effort to sequence the human genome. Venter and his team at the J. Craig Venter Institute (JCVI) modified a computer file of the DNA sequence of the bacterium Mycoplasma mycoides. They then produced physical DNA from this sequence, which they inserted into a cell. This cell reproduced under control of the artificial DNA to create a new bacterium [4].

High density, redundancy, and a long life expectancy are some of the many advantages of using DNA as a data storage medium.

DNA Steganography

Steganography is the science of transmitting a message hidden inside a cover medium in a way that no one other than the sender and the intended recipient suspect its existence. The goal of steganography is to avoid suspicion to the existence of the message, while cryptography merely aims at making a message unreadable (Wong et al., 2003). Its properties as a data storage medium also make DNA a good medium for steganography.

Noncoding genomic regions may seem to be an obvious choice of locations for inserting messages. However, the biological purpose of these regions is not yet fully understood [11] and tampering with them might possibly cause the death of the organism.

Arita et al. (Arita, 2004) suggested to encode the message in the protein coding regions of genes instead. With 20 amino acids and one stop symbol using a total of 64 possible codons (Arita & Ohashi, 2004), there is redundancy where often two or more codons code for the same amino acid. This redundancy can be used to embed messages, since many of these redundant, or synonymous, codons typically differ in their third position, also known as the wobble base (Brenner et al., 1965). This means that changing the wobble base to hide messages will not affect the coded amino acid.

Coding Schemes for Hiding Text in DNA

A code is a cryptographic rule that determines which symbol from a target alphabet uniquely represents which symbol from a source alphabet. In DNA Steganography, the source alphabet is made of alphanumeric characters the case of text information, or color values of pixels in the case of images. The target alphabet consists of various combinations of the initials of the four nucleotides.

Which symbol in the target alphabet is chosen to represent which symbol in the source alphabet is determined by a set of rules called a coding scheme.

We have already compared several existing coding schemes in an earlier publication (Beck, Rouchka, & Yampolskiy, 2013). Simple substitution ciphers are the most common ones [7, (Clelland, Risca, & Bancroft, 1999)]. Other, more sophisticated coding schemes exist, which are either geared toward error detection and correction [(Arita & Ohashi, 2004), (Heider & Barnekow, 2007)], or optimization of the available capacity to hide messages (Huffman, 1952), (Ailenberg & Rotstein, 2009).

Insertion of Media other than Text

Most research in DNA steganography focuses on hiding text and only very little research has been done so far on hiding other media in a DNA sequence. Goldman et al. (Goldman et al., 2013) describe encoding five files of various types in a DNA sequence. These files include a JPEG 2000 image and a speech in MP3 format. The coding scheme they used utilizes several intermediate steps. First, the image file and the sound file are translated into binary. Then, the text file and the binary data from the other files is translated into a base-3 code and finally into sequences of DNA bases.

Davis (Davis, 1996) describes a method of encoding the black-and-white image of a relatively simple shape (5 by 7 bitmap) into a 35 bit binary sequence, which was then compressed. His approach compares the molecular weights of the bases to obtain an incremental reference. Starting with the smallest base, Cytosine, Davis assigns numbers to the bases in ascending order. This results in C = 1, T =2, A =3, and G =4. This method compresses the binary digits of the bit-mapped image into fewer DNA base symbols by using each base to indicate how many times each binary value (0 or 1) is to be repeated before changing to the respective other value. This technique is widely used in data compression. This can be represented as shown in Table 1. Using this coding method, the thirty-five-bit black-and-white image is translated to only eighteen DNA bases: CCCCCCAACGCGCGCGCT

These can be decoded to yield one of the two following binary sequences:

10101011100010000100001000010000100

or

01010100011101111011110111101111011

This depends on if either a 1 or a 0 is chosen to start the decoding sequence. Transforming either of the two sequences into the correct five-by-seven matrix will produce the image. Since the example used by Davis is bilaterally symmetrical, more than one of several possible five-by seven matrices will in this case result in producing the correct bitmap [22].

Table 1. Coding scheme used by Davis [22]

Base	Bit sequence
C	1 or 0
T	11 or 00
A	111 or 000
G	1111 or 0000

Ailenberg and Rotstein (Ailenberg & Rotstein, 2009) have developed a coding scheme to encode an image that is composed of shapes and their coordinates.

This way of encoding an image is not very efficient. A more feasible approach has been described by Yokoo and Oshima (Yokoo & Oshima, 1978). This approach suggests to arrange the 3-base codons of a DNA sequence in a two dimensional array and then translate one base of each codon into either black or white, with G and C being black and A and T being white, or vice versa. This is done for each base of all the codons, which would result in three separate images.

Hennings and Kettelberger (Hennings & Kettelberger, 2004) have developed a method to generate music by decoding and transcribing genetic information within a DNA sequence into a music signal having melody and harmony.

Methodology

We have developed a very similar coding scheme to the one described by Yokoo and Oshima [23], with the difference that we use all three bases of each codon for encoding color information instead of creating three separate images. Our approach will determine the width and height of the array used for creating the image using the two closest factors of the number of codons. This will result in a picture that is as close to a square in shape as possible.

The DNA sequence is arranged in a two dimensional array the same way as described by Yokoo and Oshima (Yokoo & Oshima, 1978), but in our case the first base of each codon is used to encode the red portion, the second base for the green portion, and the third for the blue portion of each pixel. DNA bases are translated into RGB values using the following coding tables:

Table 2. Translation of DNA bases to RGB values

Base	RGB
A	0
C	64
G	128
T	255

Table 3. Translation of RGB values into DNA bases

RGB	Base
0-63	A
64-127	C
128-191	G
192-255	T

Results

Each codon encodes one pixel and the coordinates of the codon in the array will be the coordinates of the pixel in the resulting bitmap. The following example shows each step of the encoding process:

DNA sequence:
ATA TAA TAA TAA TTA AAT AAA TTT AAA ATA AAT TTT GAG TTT ATA AAT AAA TTT AAA ATA TAA TTA TTA TTA AAT

DNA sequence as two dimensional array:

ATA	TAA	TAA	TAA	TTA
AAT	AAA	TTT	AAA	ATA
AAT	TTT	GAG	TTT	ATA
AAT	AAA	TTT	AAA	ATA
TAA	TTA	TTA	TTA	AAT

The array is created by taking the square root of the number of codons in the DNA sequence. The result is rounded up to give the width and rounded down to give the height of the image. The two numbers are multiplied and if the result is less than the number of codons, the smaller number is increased by 1. This will result in an array that is large enough for all codons, in some cases slightly larger. The extra space will be filled with white pixels in the resulting image.

Table 4. After translation into RGB:

0,255,0	255,0,0	255,0,0	255,0,0	255,255,0
0,0,255	0,0,0	255,255,255	0,0,0	0,255,0
0,0,255	255,255,255	127,0,127	255,255,255	0,255,0
0,0,255	0,0,0	255,255,255	0,0,0	0,255,0
255,0,0	255,255,0	255,255,0	255,255,0	0,0,255

This method allows the encoding of 64 colors and ensures that the encoding of all the most common colors such as red, green, blue, yellow, magenta, orange, grey, black, and white is possible.

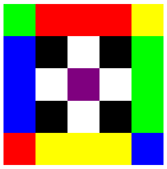

Fig. 1. Resulting image (enlarged by factor 16)

Future research and conclusion

The use of only 64 colors obviously leads to the loss of color information. Also, with the current algorithm the program assumes that the width and height of an image are as similar (a square, or approximately a square) as possible. For example, a 120x40 pixel image would be decoded as a 60x80 pixel image. A possible solution would be to encode the dimensions of the image as well. Our method is simpler and more storage space efficient than the one described by Goldman [6], but as a tradeoff can encode fewer colors. It is also more specialized toward images, while Goldman's approach is geared toward a variety of data types. Further research could lead to the development of algorithms to detect, extract and decode images that have been hidden in DNA sequences. These methods could be used for forensic purposes. Similar algorithms have already been developed for text-based DNA Steganalysis (Beck, Desoky, Rouchka, & Yampolskiy, 2014).

References

[1] Ailenberg, M., & Rotstein, O. (2009). An improved Huffman coding method for archiving text, images, and music characters in DNA. Biotechniques, 47(3), 747-754. doi: 10.2144/000113218

[2] Anam, B., Sakib, K., Hossain, A., & Dahal, K. (2010). Review on the Advancements of DNA Cryptography. Paper presented at the International conference on Software, Knowledge, Information Management and Application, Paro, Bhutan.

[3] Arita, M. (2004). Comma-free design for DNA words. Communications of the ACM, 47(5), 99. doi: 10.1145/986213.986244

[4] Arita, M., & Ohashi, Y. (2004). Secret Signatures Inside Genomic DNA. Biotechnology Progress, 20(5).

[5] Beck, M. B., Desoky, A. H., Rouchka, E. C., & Yampolskiy, R. V. (2014). Decoding Methods for DNA Steganalysis. Paper presented at the Paper presented at the 6th International Conference on Bioinformatics and Computational Biology (BICoB 2014), Las Vegas, Nevada, USA.

[6] Beck, M. B., Rouchka, E. C., & Yampolskiy, R. V. (2013). Finding Data in DNA: Computer Forensic Investigations of Living Organisms. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunication Engineering, 114(2013), 204-219.

[7] Brenner, S., Stretton, A. O., & Kaplan, S. (1965). Genetic Code: The 'Nonsense' Triplets for Chain Termination and their Suppression. Nature, 05 June 1965; 206(988), 994-998.

[8] Brenner, S., Williams, S. R., Vermaas, E. H., Storck, T., Moon, K., & McCollum, C. (2000). In vitro cloning of complex mixtures of DNA on microbeads: Physical separation of differentially expressed cDNAs. Proceedings of the National Academy of Sciences of the United States of America, 97(4), 1665-1670.

[9] Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-Generation Digital Information Storage in DNA. Science 28 September 2012, 337(6102),1628.doi: 10.1126/science.293.5536.1763c

[10] Clelland, C. T., Risca, V., & Bancroft, C. (1999). Hiding messages in DNA microdots.pdf. Nature, 399(10), 533-534.

[11] Davis, J. (1996). Microvenus. Art Journal, 55(1), 70-74.

[12] Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R. Y., Algire, M. A., . . . Venter, J. C. (2010). Creation of a bacterial cell controlled by a chemically synthesized genome. Science, 329(5987), 52-56. doi: 10.1126/science.1190719

[13] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., & Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature, 494(7435), 77-80. doi: 10.1038/nature11875

[14] Heider, D., & Barnekow, A. (2007). DNA-based watermarks using the DNA-Crypt algorithm. BMC Bioinformatics, 8, 176. doi: 10.1186/1471-2105-8-176

[15] Heider, D., & Barnekow, A. (2008). DNA watermarks: a proof of concept. BMC Mol Biol, 9, 40. doi: 10.1186/1471-2199-9-40

[16] Hennings, M. R., & Kettelberger, D. M. (2004). United States of America Patent No.: U. S. P. T. Office.

[17] Huffman, D. A. (1952). A Method for the Construction of Minimum-Redundancy Codes. Proceedings of the IRE, 40(9), 1098 - 1101. doi: 10.1109/JRPROC.1952.273898

[18] Jiao, S.-H. (2009). Hiding data in DNA of living organisms. Natural Science, 01(03), 181-184. doi: 10.4236/ns.2009.13023

[19] Jiao, S.-H., & Goutte, R. (2008). Code for Encryption Hiding Data into Genomic DNA. Paper presented at the 9th International Conference on Signal Processing, Beijing, China.

[20] Martin, R. G., Matthaei, J. H., Jones, O. W., & Nirenberg, M. W. (1962). Ribonucleotide composition of the genetic code. Biochemical and biophysical research communications, 6(6), 410-414.

[21] Smith, G. C., Fiddes, C. C., Hawkins, J. P., & Cox, J. P. L. (2003). Some possible codes for encrypting data in DNA. Biotechnology Letters, 25(14), 1125-1130.

[22] Wong, P. C., Wong, K.-K., & Foote, H. (2003). ORGANIC DATA MEMORY Using the DNA Approach. Communications of the ACM, 46(1), 95-98.

[23] Yachie, N., Sekiyama, K., Sugahara, J., Ohashi, Y., & Tomita, M. (2007). Alignment-Based Approach for Durable Data Storage into Living Organisms. Biotechnology Progress 2007 Mar-Ap, 23(2), 501-505.

[24] Yokoo, H., & Oshima, T. (1978). Is Bacteriophage X174 DNA a Message from Extraterrestrial Intelligence? Icarus, 38(1).