Anthony P. Machado, Gaby G. Dagher, and Eddie C. Davis

Hiding Data in Cellular DNA: Contextualizing Diverse Encoding Schemes

Abstract: DNA, the macromolecule used by organisms to store and transmit genomic data, has attracted the attention of privacy researchers as a channel for secure data transfer. DNA's small size and abundance in nature makes it an ideal steganographic medium for hiding messages. Already, artificially synthesized DNA has been used to store text, audio, and images. Encoded cellular DNA is not far behind, with much research being done on ways to safely embed data without harming the cell.

In this survey we provide the first systematic comparison of cellular encoding schemes proposed in the literature. Different DNA regions in the cell have their own unique bio-restrictions that must be satisfied for DNA storage. Drawing from a wide array of schemes, we compare the novel techniques used to meet these bio-restrictions. This contextualization of the research creates a bigger picture that can help guide the design of future schemes. We also survey the compression methods and error detection techniques used by the encoding schemes, and their effect on error rate and bits-per-base density. Finally, we propose future directions for research in untapped cellular regions such as mitochondrial DNA and we offer novel insights into the potential for epigenetic encoding with methylation and histones.

Keywords: keywords, keywords

DOI Editor to enter DOI
Received ..; revised ..; accepted ...

1 Introduction

Biosteganography is an emerging field in privacy research that combines techniques from genetic engineering, bioinformatics, cryptography and forensics to secretly transfer data within a living cell's DNA [1]. Data ranging from simple text messages to audio recordings and color images can be inserted into cellular DNA for secure transmission. Because DNA is information dense and occurs abundantly in nature, it makes an ideal medium for sending messages secretly. Basic steganographic principles dictate that if an attacker doesn't know where to find a confidential message in the first place, the message is far more secure [2]. Locating encoded DNA would be harder than finding a needle in a haystack. A needle, at least, can be seen by the human eye.

Traditional digital storage systems such as hard drives and SD cards have detectable emanations, which makes them vulnerable to side attacks [3]. Modified DNA, however, has no measurable emanations. If an agent needed to carry a confidential message through a tight security checkpoint, any form of electronic storage could be easily detected, while physical recordings like paper or tape could be visibly located by the security guard. A DNA encoded message, however, located in the cells of the agent's thumb, or in bacteria under the agent's nail, could be brought through without detection. The receiver of the message would need to know how to decode the data, and have the lab equipment to amplify and sequence the DNA in order to retrieve the message.

Inserting hidden data into DNA is also crucial for watermarking patented genes. Advances in gene editing technology have led to a rise in genetically modified organisms being developed in a variety of sectors such as agriculture, healthcare, and energy. For added security of these novel genomes, many companies have begun embedding hidden watermarks in their DNA so that ownership can be established in the case of theft. For this purpose too, secrecy is essential. The watermark must be embedded in such a way that the thief cannot find it, else the thief would simply remove it. Thus, biosteganography is an overlapping goal both of individuals wanting to send data privately and companies wanting to watermark genetic inventions.

Many encoding schemes have been proposed to accomplish data embedding in cellular DNA.

In this survey, we will provide the first systematic analysis of these diverse encoding schemes, in which:

- we define the unique bio-restrictions associated with cellular encoding and compare how the proposed encoding schemes meet these restrictions;
- we contrast the error rates and bits-per-nucleotide densities of each of the schemes, highlighting the effects of error correction and compression techniques on both; and

Anthony P. Machado: Boise State University, E-mail: anthony-machado@u.boisestate.edu
Gaby G. Dagher: Boise State University, E-mail: gabydagher@boisestate.edu

– we propose new directions for research incorporating novel epigenetic techniques.

The rest of the survey is organized as follows. In Section 2 we provide a background on genetics, explain techniques in artificial DNA encoding, and clarify the biological restrictions involved in cellular encoding. A systematic comparison of the encoding schemes for coding and non-coding regions is presented in Section 3, in which we look at how each encoding scheme meets cellular restrictions and the error rate and bits-per-nucleotide density it offers. Finally, in Section we propose new directions for future encoding schemes that use epigenetics and alternative genetic regions.

2 Background

2.1 DNA

Every living cell contains DNA molecules encoded with instructions for making the proteins necessary for the cell to function. DNA takes the form of a double helix made of two antiparallel strands. Each strand is composed of a sequence of 4 nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). The purines adenine and guanine, and the pyrimidines cytosine and thymine hydrogen bond with each other across the double helix. Every three nucleotides forms a codon that the cell reads and processes to create an amino acid, the building block of a protein [4].

There are only 20 amino acids, despite the fact that 4^3 = 64 possible codon variations exist. This is because many amino acids are correlated with up to five redundant codons. Also, three codons are used specifically as stop signals to indicate the end of a protein chain. The redundancy in codon to amino acid mapping is called codon degeneracy [5]. Codon degeneracy gives an evolutionary advantage to the cell by allowing certain mutations to occur in a codon while preserving the codon's functionality.

Amino Acids	Codons					
Ala	GCT	GCC	GCA	GCG		
Arg	CGT	CGC	CGA	CGG	AGA	AGG
Asn	AAT	AAC				
Asp	GAT	GAC				
Cys	TGT	TGC				
Gln	CAA	CAG				
Glu	GAA	GAG				
Gly	GGT	GGC	GGA	GGG		
His	CAT	CAC				
Ile	ATT	ATC	ATA			
Leu	TTA	TTG	CTT	CTC	CTA	CTG
Lys	AAA	AAG				
Met	ATG					
Phe	TTT	TTC				
Pro	CCT	CCC	CCA	CCG		
Ser	TCT	TCC	TCA	TCG	AGT	AGC
Thr	ACT	ACC	ACA	ACG		
Trp	TGG					
Tyr	TAT	TAC				
Val	GTT	GTC	GTA	GTG		
START	ATG					
STOP	TAA	TGA	TAG			

2.2 Artificial Encoding

Artificial DNA encoding was the

DNA was first encoded with a secret message in 1999 by Clelland et al [6]. Inspired by the tiny, concealed "microdot" messages used in World War II, artificial DNA strands were constructed using a simple substitution cipher, then mixed with human DNA and pipetted onto a printed period on filter paper. The encoded DNA was later recovered from the dot and sequenced to successfully read the secret message: "JUNE 6 INVASION: NORMANDY".

2.3 Cellular Encoding Restrictions

Cellular DNA encoding must not harm the carrier organism, either by removing cellular functionality or adding mutative behavior. To avoid this, encoding schemes must ensure that modified DNA strands remain biologically equivalent to their wild-type form. This section defines the bio-restrictions that exist for two distinct areas of cellular DNA, protein coding DNA (pcDNA) and noncoding DNA (ncDNA). As our knowledge of genetics continues to expand, more particular restrictions may become known.

2.3.1 pcDNA Constraints

The protein coding region of DNA contains the codons that are translated to amino acids, which are concatenated into proteins. Any data insertions in this area must meet the following constraints.

Protein Preservation The structure of the protein coded by the region must remain unchanged. Nucleotide insertions and modifications must not alter the codons in such a way that would change the original amino acid sequence.

Codon Bias Preservation Individual cells have specific ratios of cytoplasmic tRNA associated with their genomic codons, which can be disrupted if the codon balance is changed, and have a negative effect on the cell. Therefore, it is important that codon bias is preserved.

2.3.2 ncDNA Constraints

The noncoding region of DNA is often called "junk DNA" for its apparent lack of use in the cell. Because they appear to be non-functional, data can be embedded into these regions if the following constraints are met.

Truly nonfunctional region. When ncDNA was first discovered it was assumed to have no role in cell functionality, but recent studies have shown that up to 80% of ncDNA may have biochemical functions in the cell, despite not coding for proteins [7]. Therefore, it is first imperative that the individual who wishes to insert a message into ncDNA verify that they are encoding their message in the 20% that has no biochemical use.

No start codons. When a cell’s genetic machinery locates a start codon, it can begin the transcription process. To prevent unwanted transcription from happening in ncDNA with embedded data, it is important to make sure the encoded nucleotides to not create a start codon. When a DNA string is being transcribed, three-nucleotide codons can be read in six different reading frames. Therefore, there should not be a start codon in any of the six frames. The most common start codon is AUG, though some alternative start codons can also exist, particularly in bacteria [8]. If a cell contains alternative start codons, the encoding scheme should avoid all of them.

No homopolymers. A DNA homopolymer is a region where the same nucleotide is repeated multiple times. Too many repeats can cause errors during DNA replication through polymerase slippage [9]. These replication errors could quickly distort the inserted message and possibly damage the cell after a few generations. For this reason any ncDNA encoding scheme should not include homopolymers greater than length 3.

3 Compression and Error Correction

Smith et al. [10] were the first to suggest a data compression technique in DNA encoding, specifically the Huffman Code. The Huffman Code is a form of lossless data compression that forms a symbol table using fewer bits to encode more common characters [11]. Smith et al created a Huffman Code table mapping letters of the alphabet to nucleotide strings, where the most common English letter ’e’ was mapped to the nucleotide string "T", and the least common english letter ’z’ was mapped to a longer nucleotide string "CCCTG". This achieved an average encoding length of 2.2 bases per letter. The mapping is unambiguous, making only one possible interpretation of each message.

Comma encoding [12] specifies that encoded words be separated by a single nucleotide (i.e., G). The remaining four (or five in Smith et al.), are composed of the other three nucleotides. Additional constraints are that only three A-T pairs are allowed, and two G-C always on the top strand, so that the DNA molecules will have isothermal melting temperatures. This technique provides particularly good detection of insertions and deletions, unfortunately it is also space inefficient.

The alternating code, also from Smith et al. [10] consists of 64, 6 base pair codons, with nucleotides alternating between an A or G (purines) at odd positions, and C or T (pyrimidines) at evens (e.g., RYRYRY…, YRYRYR). Like the comma encoding, the alternating code results in isothermically stable molecules with a 1:1 ratio of A-T to G-C pairs. While more space efficient, this technique is not as proficient at error detection, as only 67% of codons result in nonsense codons after mutation, compared to 83% for the comma code.

Contrast mapping is an encoding scheme developed by Mousa et al. [13]. The binary message is divided into 6-bit groups, each converted to decimal (base 10). Pairs of consecutive values (x, y) are converted into (x', y') with the the following linear transformations: $x' = 2x - y$, $y' = 2y - x$. Values are limited to the subdomain: $0 \leq 2x - y \leq L$, $0 \leq$ $2y$ - $x \leq L$. Values are decoded with the following equations: $x = [3x' + 3y'], y = [3x' + 3y']$. This technique is sufficiently flexible to be applied to DNA or image steganography.

Others to write up...

1) Towards practical, high-capacity, low-maintenance information storage in synthesized DNA

2) Rewritable, Random-Access DNA-Based Storage System

3) Genomically encoded analog memory with precise in vivo DNA writing in living cell populations

4) HyDEn: A Hybrid Steganocryptographic Approach for Data Encryption Using Randomized Error-Correcting DNA Codes

5) Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes

4 Encoding Schemes

Encoding schemes for cellular DNA can involve several components:

1. Encryption algorithm
2. Mapping table
3. Compression
4. Error correction
5. Fake data embedding

The following encoding schemes use some or all of these elements in their design.

4.1 pcDNA Encoding

Shimanovsky et al [10] were the first to propose using codon degeneracy to encode data in pcDNA. By switching codons between their redundant forms with the modification of one nucleotide, data was inserted without altering protein translation. Arita and Ohashi [14] implemented this technique in a living cell. They did this using site-directed mutagenesis of wobble codons in the *ftsZ* gene of *Bacillus subtilis*. The university name "KEIO" was inserted by modifying the redundant nucleotides of the codons downstream of the *ftsZ* start codon. An unmodified codon represented value 0 while a codon with a wobble nucleotide changed to any of its non-wild type redundant forms represented value 1. Messages were translated using a 6-bit mapping table, with the first 5 bits corresponding to an English alphabet letter or basic punctuation, and the last used as a parity bit for error correction.

Heider et al. developed the DNA-Crypt [15] algorithm for creating DNA watermarks for marking genetically modified organisms. It is similar to the work of Arita et al. in that the encoding targets the wobble base pair in the genetic code. An encryption function E maps the plaintext (binary data) X to the ciphertext (genetic data) Y, such that $X \in \{0, 1\}$ and $Y \in \{\text{A, C, G, T}\}$. Two bits are encoded per base, or one byte for four bases. Error correction is achieved with a fuzzy controller that selects one of two algorithms, either the 8/4 Hamming code for mutations that differ in only bit (e.g., 00 to 01) or the WDH-code for those that differ by multiple bits. This encoding scheme allowed implementations of several cryptographic algorithms, including One-Time Pad, AES, Blowfish, and RSA. The accuracy was tested using the GTPase encoding Ypt gene in $S.cerevisiae$.

Haughton and Balado proposed BioCode [16], a pair of encoding algorithms, one for ncDNA, and one for pcDNA. The ncDNA algorithm expands upon DNA-Crypt by observing the no start codons in restriction. This is accomplished by defining a set of dinucleotides D = { AT, CT, TT, CA } that covers the possible eukaryotic start codons on either DNA strand. The trailing dinucleotide d is continually checked for membership in D. If found, d is replaced with a lookup table formed from a graduated mapping of the message space M_d to set S_d. The pcDNA encoding technique is similar, but enforces the additional constraint of Binary Codon Equivalency (BCE). This requires that the cardinality of the codon set ($|S_d|$) be varied during the embedding process to allow the usage of a static lookup table.

In order to better preserve codon bias, Lee developed a discrete wavelet transform (DWT) technique for pcDNA encoding [17]. The target coding sequence was divided into subsequences in which every codon was given a numerical code associated with amino acid histogram rankings. DWT coefficients were calculated for synonymous codons and the optimal subsequence was found, which was then replaced with the encoded subsequence. A nonlinear congruential-pseudorandom number generator then created the watermark and picked the location in the DWT domain for embedding.

4.2 ncDNA and Plasmid Encoding

The development of encoding schemes for ncDNA and plasmids have been strongly correlated, due to two regions having nearly identical bio-restrictions. The main difference being that plasmid encoding is also confined by length. However, due to their similarities and the parallel advancement of their encoding schemes, we will consider them both in this section.

Wong et al. [18] were the first to encode data into plasmids. Letters and punctuation were mapped to nucleotide triplets, then short text snippets of 19 to 33 characters were inserted into the plasmids of Deinococcus radiodurans, a bacteria that is very resilient to extreme conditions. The encoded section of the plasmid was flanked by sentinel sequences 20 base pairs long containing stop codons. This was done to prevent the bacteria from transcribing the message while reading the adjacent functional parts of the plasmid. After insertion of the plasmids into the host bacteria, they were incorporated into their genomes. No error correction, encryption, or compression techniques were used.

Yachie et al. [19] came up with the idea of using multiple alignment in lieu of error detection or correction techniques. With this scheme, messages were composed using the Keyboard Scan Code Set2, which contains all keyboard inputs, and converted these messages to binary. The encryption keys mapped four bits of binary code to two nucleotides. This type of mapping created four reading frames in the binary message (C1 - C4), each of which was converted to DNA and inserted adjacently into the plasmid, creating redundancy that allowed for multiple alignment when decoding.

Repetition coding was used again, this time for ncdDNA, by Haughton and Balado [20] who used two variant approaches. Both approaches required finding subsequences in the host ncDNA equal to the length of the encoded message, and replacing them with the message, as opposed to simply inserting the message and expanding the DNA strand. With the first approach, $RAlign$, the redundant subsequences were each prepended and appended with unique 24-bp long markers to aid alignment, and replaced specific subsequences in the genome that were known by both the sender and receiver. With the second approach, $HTAlign$, the prepended and appended markers were identical, which means the receiver does not need to know the location of the replaced subsequences, but can find the repeated messages by searching for the markers alone.

Wanting to improve on the very low bits per base density of previous schemes, Ailenberg and Rotstein suggested an improved Huffman code that could decrease the sequence sizes by up to 40% [21] compared to other schemes. Keyboard characters were mapped to variable length sequences, with the more common characters assigned to the shorter sequences. To help maintain C-G balance, CG-rich codons were pushed downwards on the frequency table. To illustrate the density that could be achieved, text and musical notes for "Mary Had a Little Lamb", along with an image of a lamb made with geometric shapes, were encoded into a mere 844-bp DNA fragment. It was estimated that this encoding scheme would average 3.5 bases-per-character, which was a noted improvement over previous schemes. No error correction was used.

Heider et al. [22] were interested in the possibility of encoding data in non-protein coding but biochemically functional DNA. 2-3 character watermarks were inserted into non-coding promoter regions and a regulatory RNA region in $Escherichiacoli$. The introduction of the watermarks were shown to disrupt the reading of the genes associated with the promoters and changed the secondary structure of the regulatory RNA molecule. This research was significant in showing that biochemically functional ncDNA is a poor region for encoding.

5 Future Directions

There are several other regions of DNA that have seen little research in biosteganography, but have substantial potential for message encoding and watermarking.

5.1 Mitochondrial DNA

Heider et al. [25] suggested using mitochondria for secret data encoding. Mitochondria are organelles in the cell that contain a small amount of DNA that encodes translation machinery and oxidative chain components [26]. Heider et al. identified an mtDNA section 1,541 base pairs long that contains no active gene regions and suggested that it would be an ideal location for embedding data with synonymous codons. They created a program called $ProjectMito$, derived from $DNA - Crypt$, that encrypts binary files and modifies them with the Hamming code for error correction, before translating the message into an mtDNA sequence. $ProjectMito$ can also be used for decryption after the DNA is sequenced.

6 Conclusions

References

[1] T. Brunet, “Aims and methods of biosteganography,” *Journal of Biotechnology*, vol. 226, pp. 56–64, 2016.

[2] D. Artz, “Digital steganography: hiding data within data,” *IEEE Internet Computing*, vol. 5(3), pp. 75–80, 2001.

[3] H. Tanaka, “Evaluation of information leakage via electromagnetic emanation and effectiveness of tempest,” *IEICE Transactions on Information and Systems*, vol. 91(5), pp. 1439–1446, 2008.

[4] J. Watson and F. Crick, “Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid,” *Nature*, vol. 171 (4356), pp. 737–738, 1953.

Table 1. Table 1

Method			Bio-Restrictions			Error Detection	Error Correction	Compression	Encryption	Blind Decoding
	No Homopolymers	No Start Codons	Preserves Functionality	Preserves Codon Bias	Balances C-G					
PROTEIN CODING DNA										
Secret Signature [14]	●	●	●			●				
DNA-Crypt [15]	●	●	●			●	●		●	
BioCode [16]	●	●	●	●	●	●	●			
DWT Based [17]	●	●	●	●	●	●	●			
NONCODING DNA and PLASMIDS										
Organic Memory [18]			●	●						
Alignment-Based [19]			●	●		○	○			
Improved Huffman [21]			●	●	○			●		
Regulatory Watermarks [22]										
RAlign [20]			●	●		○	○			
HTAlign [20]			●	●		○	○			
DNA Barcodes [23]	●		●	●		●	●			
DNA-Courier Attack [24]			●	●	●				●	

[5] J. Watson, T. Baker, S. Bell, A. Gann, M. Levine, and R. Losich, *Molecular Biology of the Gene, 6th Edition*. Pearson, 2008.

[6] C. Clelland, V. Risca, and C. Bancroft, “Hiding messages in DNA microdots,” *Nature*, vol. 399, pp. 533–534, 1999.

[7] E. P. Consortium *et al.*, “An integrated encyclopedia of dna elements in the human genome,” *Nature*, vol. 489(7414), pp. 57–74, 2012.

[8] F. e. a. Blattner, “The complete genome sequence of escherichia coli k-12,” *Science*, vol. 277:5331, pp. 1453–1462, 1997.

[9] E. Viguera, D. Conceill, and S. Ehrlich, “Replication slippage involves dna polymerase pausing and dissociation,” *The Embo Journal*, vol. 20(10), pp. 2587–2595, 2001.

[10] G. Smith, C. Fiddles, J. Hawkins, and J. Cox, “Some possible codes for encrypting data in DNA, volume = 25, pages = 1125-1130, year = 2003,” *Biotechnology Letters*.

[11] D. Huffman, “A method for the construction of minimum-redundancy codes,” *Proceedings of the IRE*, vol. 40(9), pp. 1098–1101, 1952.

[12] W. S. R. V. E. H. S. e. a. Brenner, S., “In vitro cloning of complex mixtures of dna on microbeads: physical separation of differentially expressed cdnas,” *Proceedings of the National Academy of Sciences*, vol. 97(4), pp. 1665–1670, 2000.

[13] M. K. A.-W. W. . H. M. M. Mousa, H., “Data hiding based on contrast mapping using dna medium,” *International Arab Journal of Information Technolology*, vol. 8(2), pp. 147–154, 2011.

[14] M. Arita and O. Yoshiaki, “Secret signatures inside genomic dna,” *Biotechnology Progress*, vol. 20, pp. 1605–1607, 2004.

[15] D. Heider and A. Barnekow, “Dna-based watermarks using the dna-crypt algorithm,” *BMC bioinformatics*, vol. 8(1), 2007.

[16] D. Haughton and F. Balado, “Biocode: Two biologically compatible algorithms for embedding data in non-coding and coding regions of dna,” *BMC bioinformatics*, vol. 14(1), p. 1, 2013.

[17] S. Lee, “Dwt based coding dna watermarking for dna copyright protection,” *Information Sciences*, vol. 273, pp. 263–286, 2014.

[18] P. Wong, K. Wong, and H. Foote, “Organic data memory using the dna approach,” *Communications of the ACM*, vol. 46:1, pp. 95–98, 2003.

[19] N. Yachie, K. Sekiyama, J. Sugahard, Y. Ohashi, and M. Tomita, “Alignment-based approach for durable data storage into living organisms,” *Biotechnology Progress*, vol. 23, pp. 501–505, 2007.

[20] D. Haughton and F. Balado, “Repetition coding as an effective error correction code for information encoded in dna,” *Internation Conference on Bioinformatics and Bioengineering*, 2011.

[21] M. Ailenberg and O. Rotstein, “An improved huffman coding method for archiving text, images, and music characters in dna,” *BioTechniques*, vol. 47, pp. 747–754, 2009.

[22] D. Heider, M. Pyka, and A. Barnekow, “Dna watermarks in non-coding regulatory sequences,” *BMC Research Notes*, vol. 2:123, 2009.

[23] D. Kracht and S. Schober, “Insertion and deletion correcting dna barcodes based watermarks,” *BMC Bioinformatics*, vol. 16:50, 2015.

[24] J. Chun, H. Lee, and J. Yoon, “Passing go with DNA sequencing: Delivering messages in a covert transgenic channel,” *IEEE CS Security and Privacy Workshop*, vol. 14:121, 2013.

[25] D. Heider, D. Kessler, and A. Barnekow, “Watermarking sexually reproducing diploid organisms,” *Bioinformatics*, vol. 24:17, pp. 1961–1962, 2008.

[26] R. Garesse and C. Vallejo, “Animal mitochondrial biogenesis and function: a regulatory cross-talk between two genomes,” *Gene*, vol. 263, pp. 1–16, 2001.

Table 2. Comparative evaluation of encoding schemes

	Data Type		Privacy-Preserving Domain				Hosting Environment					Security	
	Set-Valued	Other	Non-Interactive		Interactive		Single	Two		Multiple		Threat Model †	Public Verifiability
Approach			Differential Privacy	Syntactic Privacy	Differential Privacy	Syntactic Privacy		Horiz.	Vert.	Horiz.	Vert.		
Terrovitis *et al.*, He and Naughton	●			●			●						
Chen *et al.*	●		●				●						
Jiang and Clifton		●		●					●			○	
Jurczyk and Xiong		●		●						●		○	
Alhadidi *et al.*		●	●					●				○	
Mohammed *et al.* (DistDiffGen)		●	●						●			○	
Mohammed *et al.* (TIPS)		●		●					●			●	
Bhaskar *et al.*, Li *et al.*	●				●		●						
Wong *et al.*	●						●					○	
Giannotti *et al.*	●					●	●					○	
Kantarcioglu and Clifton	●									●		○	
Zhang *et al.*	●							●				○	
Wahab *et al.*	●				●			●		●		○	
Dwork *et al.*		●			●					●		●	
Narayan and Haeberlen		●			●			●		●		○	
Our proposed solution	●		●					●	●	●	●	●	●

† In this column, ○ denotes semi-honest threat model whereas ● denotes malicious threat model.