ON THE SHANNON CAPACITY OF DNA DATA EMBEDDING

Félix Balado

School of Computer Science and Informatics, University College Dublin, Ireland

ABSTRACT
This paper firstly gives a brief overview of information embedding in deoxyribonucleic acid (DNA) sequences and its applications. DNA data embedding can be considered as a particular case of communications with or without side information, depending on the use of coding or noncoding DNA sequences, respectively. Although several DNA data embedding methods have been proposed over the last decade, it is still an open question to determine the maximum amount of information that can theoretically be embedded —that is, its Shannon capacity. This is the main question tackled in this paper.

Index Terms — Data hiding, DNA, Shannon capacity

1. INTRODUCTION
A number of methods have been proposed over the last ten years for mathematically embedding information within DNA, the molecule that constitutes the building block of life [1, 2, 4, 5, 6, 7]. However, important issues related to DNA data embedding are not yet elucidated yet. What stands out among them is the establishment of the upper limit on the amount of information that can be reliably embedded within DNA under a given error rate, which is just the Shannon capacity [8] of DNA data embedding. Since a DNA sequence is conceptually equivalent to a digital signal, DNA data embedding is —depending on the host sequence considered— either an instance of digital data hiding [9] or just a plain communications problem. As we will see, these facts can be exploited for capacity analysis.

2. PRELIMINARY CONCEPTS
Firstly, some basic concepts. The importance of the DNA molecule relies on it containing the instructions for the development and functioning of any living being. Chemically, DNA is formed by two nucleotide strands helicoidally twisted around each other, and mutually attached by means of two antiparallel base sequences. The only possible bases are the four molecules Adenine, Cytosine, Thymine, and Guanine, abbreviated A, C, T, and G, respectively. The interpretation of DNA as a one-dimensional digital signal is straightforward—one of the two antiparallel base sequences is enough to represent the information conveyed by a DNA molecule. For our purposes, it suffices to know that codons, which are the minimal “codewords” with biological meaning, are formed by triplets of consecutive bases in a sequence. In essence, the codons in some regions of a DNA sequence can be translated into amino acids. These amino acids are then sequentially assembled in chains which form proteins, the basic compounds of the chemistry of life. There are $4^3 = 64$ possible codons, since they are triplets of quaternary (4-ary) symbols. Crucially, there are only 20 possible amino acids, mapped to the 64 codons according to the equivalences in Table 1 (using the arbitrary mapping \{A, C, T, G\} ↔ \{0, 1, 2, 3\}).

2.1. DNA Data Embedding and Applications
If certain constraints are observed, DNA can also be used to convey additional arbitrary data [1, 2, 4, 5, 6, 7]. It is a key fact that information embedded within DNA will travel alongside each replication, whether it takes place in vivo or in vitro, that is, whether it happens inside or outside living organisms. There are two ways to achieve this information embedding:

1. By replacing or appending noncoding DNA (ncDNA) segments, which never get translated to proteins [1, 2, 3, 4]. This amounts to transmitting an arbitrary digital signal, as one can freely establish the host segments that carry the information.

2. By modifying coding DNA (cDNA) segments, which may get translated to proteins [5, 6, 7]. This amounts to transmitting a digital signal embedded within a genetic host under certain constraints. That is, a classic data hiding [9] problem in which the host sequence acts as side information at the encoder.

If information-carrying DNA is to remain functional, its translation to proteins must be the same as before the embedding operation. When modifying cDNA this can be achieved by exploiting codon equivalence. Although when modifying ncDNA it could seem that one does not have to worry (as it does not translate into proteins), recent investigations have argued that ncDNA might not really be “junk DNA” [10]. Hence ncDNA modification might alter regulatory regions whose biological task is yet unknown. Therefore the use of cDNA, whose workings are well understood, seems much more promising if both effectiveness and unobtrusiveness are to be achieved. The goal of DNA data embedding can be at least twofold:

• Tagging genetic material for tracking purposes. Reliable DNA embedding may allow new forms of genetic fingerprinting by attaching unique tags to differentiate among functionally identical genetic material. One potential application is tracking the spatial and temporal evolution of different instances of genes with identical protein translation. Another interesting application may be detecting mutations by solely relying on the embedded information. This is relevant in cases where there is not one single host genome to be used as a reference, such as in viral quasispecies.

Intellectual property protection of DNA sequences is also proposed by several authors [5, 6, 7]. Gene patents have proved to be commercially important [11], which is also illustrated by the existence of several DNA data embedding patents (see for instance [12]). The idea is to track illicit copies of genetic material to a leaking point by assigning different fingerprints to different licensees of functionally identical genetic material. An in-depth discussion of the ethical implications of these procedures is out of the scope of this paper.
of this paper, but note however that current genetic profiling techniques based on ncDNA features allow already to track individual genomes (although not identical genes directly).

Also, the effect of any data embedding method which does not alter the length of the host sequence amounts to that of writing to a digital memory. In particular, prior information is necessarily overwritten. Thus any protection granted by DNA data embedding can be easily thwarted—for a single sequence—by an active third party.

- Using genetic material as a massive and compact storage medium. Long-term storage of data in the DNA of living organisms, such as bacteria, has been actually implemented with real organisms by Wong et al. [2], Yachie et al. [4], and other groups of researchers.

2.2. Capacity and Robustness Under Mutations

Random mutations occur in most types of DNA replication, and this will affect information embedded in DNA sequences. Recent studies suggest that the single base substitution error rate in prokaryotic DNA in vivo may be in the range of $10^{-8}$ to $10^{-7}$ per replication [13]. A standard single base substitution error rate of $10^{-10}$ per replication for eukaryotic cells is cited in [7]. These figures are very low; however, single base substitution error rates due to replication by some particular polymerases can be as high as $10^{-3}$ to $10^{-1}$ per replication [13]. Furthermore, these rates refer to one single replication. After $R$ replications a constant mutation rate $q$ becomes $q^{R(10)} = 1 - (1 - q)^R$, and $q^{R(10)} \to 1$ as $R \to \infty$. For instance, Fu [14] has estimated an accumulated single base substitution error rate of $1.71 \times 10^{-2}$ per year in the genome of HIV. Because of these facts, and despite the discussion in [3], robust DNA data embedding is key, both in organisms with high generations-per-day ratios and in environments with high mutation rates, or simply when information must be kept intact over protracted periods of time. As the Shannon capacity embodies the concept of robustness—maximum resilience to random distortions of a communications system—it is determinative for DNA data embedding is very important.

2.3. Brief Literature Review

All prior art in DNA data embedding deals with the proposal of practical methods. Among the methods that deal with ncDNA [1, 2, 3, 4] the earliest one is by Clelland et al. [1], who rely on a primer as a key to locate inserted data. A similar method is used by Wong et al. [2]. Smith et al. [3] propose Huffman and repetition-based comma codes, and consider isothermal constraints. The method by Yachie et al. [4] uses repetition for robustness. Among the methods that address cDNA [5, 6, 7], the one by Shimanovsky et al. [5] is based on arithmetic coding. Modegi [6] proposes a reversible algorithm. The method by Heider and Barnekow [7] uses a Hamming error-correcting code for some robustness.

In summary, prior art does not answer, nor did it attempt to answer, the fundamental question posed here. Moreover, none of the methods above were developed according to optimal channel coding or data hiding principles able to furnish robustness guarantees. The closeness to optimality of the most relevant among these methods will be discussed in Section 4.

Notation. Calligraphic letters ($X$) denote sets. Boldface letters ($\mathbf{x}$) denote row vectors. Uppercase ($X, \mathbf{X}$) and lowercase ($x, \mathbf{x}$) letters denote random and deterministic variables, respectively. $p(X)$ is the probability mass function (pmf) of $X$, and $E[X]$ its expectation. $I(X; Y)$ is the mutual information between $X$ and $Y$, and $H(X)$ is the entropy of $X$.

3. LIMITS OF DNA DATA EMBEDDING

While embedding data in ncDNA is a standard communications problem (as we will see, with some special features), the more difficult task of embedding data in cDNA amounts to communications with side information at the encoder. For this reason, prior data hiding research [9] is very relevant. The building codewords in DNA are discrete; however the bulk of data hiding research has dealt with continuous-valued signals, and only to a lesser extent with discrete-valued signals [15, 16]. Standard data hiding using discrete binary host signals bears some resemblances but also some differences to cDNA data embedding. Assume that a discrete binary (2-ary) host $\mathbf{x} = \{x_1, \cdots, x_N\}$, with $x_i \in X = \{0, 1\}$, is modified to embed a message $m$ chosen from a set $M$ with cardinality $|M|$. We need to specify both an embedding function $e(\cdot) : \mathcal{X} \rightarrow \mathcal{X}^N$ and a decoding function $f(\cdot) : \mathcal{X}^N \rightarrow M$ such that: a) the watermarked signal $\mathbf{y} = e(\mathbf{x}, m)$ is “close” to $\mathbf{x}$, and b) decoding a distorted version $\mathbf{y} = \mathbf{y} + \mathbf{w}$ (with $\mathbf{w}$ in $X$ and using modulo-2 addition) is asymptotically correct. Closeness is measured by means of the Hamming distance $d_H(\mathbf{x}, \mathbf{y})$ (number of different same index elements between two vectors). For $\frac{1}{|X|}E[d_H(\mathbf{X}, \mathbf{X})] \leq d$ and Bernoulli($q$) distortion, Pradhan et al. [15] and Barron et al. [16] have determined the maximum rate $R_{\text{max}}$ (bits/host symbol) that can be embedded and decoded asymptotically without errors, when the components of $\mathbf{x}$ are independently drawn from a Bernoulli($\frac{1}{2}$).

Our goal in cDNA data embedding is the same, but both closeness and distortion have now genetic meaning. Some further definitions are necessary next. For a cDNA sequence the elements of $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$ are codons, that is, $x_i \in X$, with $X \triangleq (\mathcal{X}^3)$ a 64-ary alphabet derived from the 4-ary alphabet $\mathcal{X} \triangleq \{0, 1, 2, 3\}$. We indicate by $x^3$ the representation of $\mathbf{x}$ using bases, that is, a 3N-length vector with $x_i^3 \in \mathcal{X}^3$. We also denote by $x'_3 \equiv \alpha(x_i)$ the amino acid into which codon $x_i$ translates (see Table 1); similarly, $x' = \alpha(\mathbf{x}) = \{x_1', x_2', \cdots, x_N'\}$ is the unique amino acid sequence defined by $\mathbf{x}$. The multiplicity associated with an amino acid $x'$ is written as $\mu(x')$. We will assume that the components of $\mathbf{x}$ are inde-

<table>
<thead>
<tr>
<th>Amino acid, $x'$</th>
<th>Phe</th>
<th>Tyr</th>
<th>Cys</th>
<th>Ser</th>
<th>Leu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codons</td>
<td>222 202 232 212 220 200 233 102 100 113 132 012 012 312 332 002 000 022 023 322 302 300</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>221 201 231 211 223 203 101 103 112 131 011 311 331 001 003 021</td>
<td>321 301 303</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>210 122 230</td>
<td>111 130 010 310 330</td>
<td>020</td>
<td>320</td>
<td></td>
</tr>
<tr>
<td></td>
<td>213 121</td>
<td>110 133 013 313 333</td>
<td>323</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>032 120</td>
<td>030</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>031 123</td>
<td>033</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Equivalences between amino acids and codons. Note that $\sum_{x'} \mu(x') = 64$. Stp is loosely classed as an “amino acid”.
Fig. 1. Base mutation channel.

1. Payload Computation

We assume no mutations throughout this subsection, that is, \( \mathbf{Y} = \mathbf{Y} \).

**Noncoding DNA.** In this context may be simply called payload \((P)\).

**Coding DNA.** We wish to determine the maximum number of sequences \( \mathbf{Y}^{(m)} \) such that \( d_H(\alpha(\mathbf{Y}^{(m)}), \alpha(\mathbf{X})) = 0 \), for \( m = 1,2,\ldots,|\mathcal{M}| \). The amount sought is just \( |\mathcal{M}| = \prod_{i=1}^{\mathcal{M}} \mu(\alpha(x_i)) \). Equivalently, the payload embeddable in \( \mathbf{Y} \) is \( P_c = \log_2 \mu(\alpha(\mathbf{X})) = \frac{1}{|\mathcal{M}|} \sum_{i=1}^{\mathcal{M}} \log_2 \mu(\alpha(x_i)) \) bits/codon. In order to see this result on average, we can either use the random variable \( \mathbf{X} \) or else \( \mathbf{X}' = \alpha(\mathbf{X}) \). The average payload is then \( P_c = E[\log_2 \mu(\alpha(\mathbf{X}))] = E[\log_2 \mu(\mathbf{X}')] \) bits/codon. For example, if \( \mathbf{X} \) is uniform, \( P_{c_{\text{uni}}} = 1.7819 \) bits/codon. The pmf \( p(X') \), which is straightforward from the multiplicities in Table 1, will not be uniform in this case. Observe that just by enforcing codon equivalence, \( P_c \) decreases to below one third of \( 3P_{uc} \). A distribution \( p(X') \) that maximises \( P_c \) is any for which \( E[\mu(X')] = 0 \), that is, whose support only includes one or more of the amino acids Ser, Leu and Arg. The maximising pmf needs not be deterministic. This leads to the upper bound \( P_{c_{\text{ub}}} = \log_2 6 = 2.5850 \) bits/codon for any cdNA method, on any host—including those that are deterministic.

2. Capacity Computation

We assume next that \( \mathbf{Y} \) can randomly mutate to yield a new sequence \( \mathbf{Z} \). We will consider the symmetric “base mutation channel” in Fig. 1, whose transition probability matrix is \( \Pi \equiv [\pi_{i,j}] \) with \( \pi_{i,j} = \sigma_{j,i} = p(Z^B = j|X^B = i) \) for \( i,j = 1,2,3,4 \), and then \( \pi_{1,1} = \pi_{2,2} = \pi_{3,3} = \pi_{4,4} \). The probability of mutation, or mutation rate, is \( q = 1 - \pi_{i,i} = \sum_{i \neq j} \pi_{i,j}, \) for any arbitrary \( i \). We will also assume that mutations are mutually independent, which is a worst-case analysis. Although other mutations such as insertions and deletions can also occur, the study of the scenario described above is a necessary first step in the study of the maximum amount of information that can theoretically be embedded in DNA. This task requires resorting to the concept of Shannon capacity [8].

**Noncoding DNA.** In this case, in which \( x^B = y^B \), capacity is \( C_{uc} = \max I(Z^B;Y^B) \) bits/base, where the maximisation is over all input distributions \( p(Y^B) \). This is just the capacity of a standard \( M \)-ary symmetric channel, which, taking any arbitrary \( i \in \{1,\ldots,M\} \), is given by [17] \( C^{(M)}(\epsilon) \equiv \log_2 M + \sum_{k=1}^{\mathcal{M}} \pi_{i,k} \log_2 \pi_{i,k} \) bits/symbol. Here, \( M = 4 \) and \( C_{uc} = C^{(4)} = 2 + \sum_{k=1}^{\pi_{i,k} \log_2 \pi_{i,k}} \) bits/symbol, which is achieved for uniform \( Y^B \) [17]. Isothermal constraints, which imply that \( p(Y = 0) = p(Y = 2) = \epsilon \) (\( p(Y = 1) = \max(Y = 3) = 0 \)) with \( \epsilon \geq 0 \), are sometimes enforced to improve the polymerase chain reaction (PCR) [3]. If \( \pi_{i,j} = \xi \) for all \( i \neq j \), it can be shown that the maximum isothermal rate, \( R_{uc}^{\max}(\epsilon) \), is achieved when \( p(Y = 0) = p(Y = 2) = 1/2(1+\epsilon) \) and \( p(Y = 1) = p(Y = 3) = \epsilon(1+\epsilon). \) Of course, \( R_{uc}^{\max}(\epsilon) \leq C_{uc} \), with equality for \( \epsilon = 1 \). Also, \( R_{uc}^{\max}(\epsilon) = R_{uc}(1/\epsilon) \) and the minimum is for \( \epsilon = 0 \).

**Coding DNA.** In this scenario the host (i.e., side information) must necessarily be taken into account by the encoder. Capacity is then given by Gel’fand and Pinsker’s formula [18] \( C_c = \max I(Z;U) - I(X';U) \) bits/codon, where the maximisation is over all distributions \( p(Y, U|X') \) under the constraint \( d_H(\alpha(Y), X') = 0 \), with \( U \) an auxiliary random variable. As \( Y = e|X', U) \), and as the support of \( Y|x' \) must be the set of codons \( S_{uc} \) corresponding to the amino acid \( x' \)—in order to satisfy the constraint—the cardinality of \( U|x' \) must be exactly \( |S_{uc}| \). As \( U \) must also be a good source code for \( X' \) in order to make \( I(X'|U) \) small, the support of \( U|x' \) must actually be \( S_{uc} \). One can now establish \( Y|x' = U|x' \) without loss of generality. This discussion on \( U \) also implies that \( H(X'|U) = 0 \), since given a codon there will be no uncertainty on the amino acid represented, and therefore \( I(X'|U) = H(X') \). Since \( p(Y|U, X') \) is deterministic, we just have to determine next the maximising distribution \( p(U|X') \). See first that the codon mutation channel is symmetric with transition matrix \( \Gamma = \Pi \otimes \Pi \), where \( \otimes \) is the Kronecker product. If \( \mathbf{X} \) is uniform and we choose \( U|x' \) to be uniform for every \( x' \), then we achieve uniformity of \( U \). Since a uniform input maximises mutual information over a symmetric channel [17], the achievable rate in this case is

\[
R_{c_{\text{uni}}} = C^{(64)} - H(X') \text{ bits/codon.}
\]
case $\Gamma = \frac{1}{\mu_c} \mathbf{1}^T \mathbf{1}$, and so $\mathbf{Z}$ is uniform independently of $\mathbf{U}$. Then $H(\mathbf{Z}) = H(\mathbf{Z}|\mathbf{U})$ and $C_c|_{\mu_c=\frac{1}{4}} = C_{\text{steg}}|_{\mu_c=\frac{1}{4}} = 0$.

### 3.3. Steganographic Rate

A real DNA sequence has a specific *codon count bias* [6], expressed in its characteristic 21 empirical pmfs $p(X|x')$. This bias is suppressed in the maximisation strategies described above. However, one can exploit it for *steganographic purposes*: the original codon bias is preserved by pegging $p(U|x')$ to the codon bias of the host. In this way the information-carrying sequence conforms to Cachin’s criterion for steganography. The ensuing rate can be computed by evaluating Gel’fand and Pinsker’s formula; obviously, $R_{\text{steg}}^c \leq R_c$.

### 5. REFERENCES


