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, 3, 4, 5, 6, 7]. However, important issues related to DNA data embedding are not 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 $\{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\} \leftrightarrow \{0, 1, 2, 3\}$).

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 09/RFP/CMS2212.

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

Amino acid, x'	Phe	Tyr	Cys	Ser	Leu	*Stp*	Trp	His	Gln	Pro	Arg	Thr	Ala	Gly	Asn	Lys	Ile	Met	Val	Asp	Glu
Codons	222	202	232	212	220	200	233	102	100	113	132	012	312	332	002	000	022	023	322	302	300
	221	201	231	211	223	203		101	103	112	131	011	311	331	001	003	021		321	301	303
				210	122	230				111	130	010	310	330			020		320		
				213	121					110	133	013	313	333					323		
				032	120						030										
				031	123						033										
Multiplicity, $\mu(x')$	2	2	2	6	6	3	1	2	2	4	6	4	4	4	2	2	3	1	4	2	2

Table 1. Equivalences between amino acids and codons. Note that $\sum_{x'} \mu(x') = 64$. *Stp* is loosely classed as an "amino acid".

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 DNA 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 media. 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 $q^{(R)} = 1-(1-q)^R$, and $q^{(R)} \to 1$ as $R \to \infty$. For instance, Fu [14] has estimated an accumulated single base substitution error rate of 1.71×10^{-2} over a 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— its determination 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 ($\mathcal{X}$) 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 $\overline{\mathbf{x}} = \{x_1, \cdots, x_N\}$, with $x_i \in \mathcal{X} = \{0, 1\}$, is modified to embed a message m chosen from a set $\mathcal{M}$ with cardinality $|\mathcal{M}|$. We need to specify both an embedding function $e(\cdot,\cdot) : \mathcal{X}^N \times \mathcal{M} \to \mathcal{X}^N$ and a decoding function $f(\cdot) : \mathcal{X}^N \to \mathcal{M}$ such that: a) the *watermarked* signal $\overline{\mathbf{y}} = e(\overline{\mathbf{x}}, m)$ is "close" to $\overline{\mathbf{x}}$; and b) decoding a distorted version $\overline{\mathbf{z}} = \overline{\mathbf{y}} + \overline{\mathbf{n}}$ (with $n_i \in \mathcal{X}$ and using modulo-2 addition) is asymptotically correct. Closeness is measured by means of the Hamming distance $d_H(\overline{\mathbf{y}}, \overline{\mathbf{x}})$ (number of different same index elements between two vectors). For $\frac{1}{N}E[d_H(\overline{\mathbf{Y}}, \overline{\mathbf{X}})] \leq d$ and Bernoulli(q) distortion, Pradhan *et al.* [15] and Barron *et al.* [16] have determined the maximum rate R^{unif} (in bits/host symbol) that can embedded and decoded asymptotically without errors, when the components of $\overline{\mathbf{x}}$ 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 $\overline{\mathbf{x}} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ are codons, that is, $\mathbf{x}_i \in \mathcal{X}$, with $\mathcal{X} \triangleq (\mathcal{X}^B)^3$ a 64-ary alphabet derived from the 4-ary alphabet $\mathcal{X}^B \triangleq \{0, 1, 2, 3\}$. We indicate by $\mathbf{x}^B$ the representation of $\overline{\mathbf{x}}$ using bases, that is, a $3N$-length vector with $x_i^B \in \mathcal{X}^B$. We also denote by $x'_i \triangleq \alpha(\mathbf{x}_i)$ the amino acid into which codon $\mathbf{x}_i$ translates (see Table 1); similarly, $\mathbf{x}' = \alpha(\overline{\mathbf{x}}) = \{x'_1, x'_2, \cdots, x'_N\}$ is the unique amino acid sequence defined by $\overline{\mathbf{x}}$. The multiplicity associated with an amino acid x' is written as $\mu(x')$. We will assume that the components of $\overline{\mathbf{x}}$ are inde-

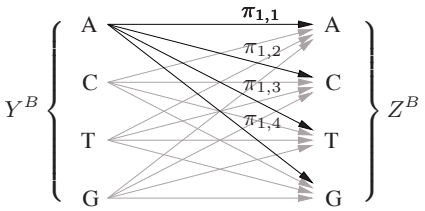

Fig. 1. Base mutation channel.

pendently drawn from a random variable with pmf $p(\mathbf{X})$. Although real DNA sequences do show statistical dependencies, note that independence can be approximated in practical methods by means of pseudorandom interleaving of $\overline{\mathbf{x}}$ followed by deinterleaving of $\overline{\mathbf{y}}$. Finally, $\mathbf{x}^B$ may also denote a ncDNA sequence, with length not necessarily a multiple of three.

The first issue for cDNA capacity analysis is that nonzero inequality constraints on the average Hamming distance, such as the ones used in [15, 16], are meaningless if one wants to carry through to $\overline{\mathbf{y}}$ the full biological functionality of $\overline{\mathbf{x}}$. Instead one must always establish the deterministic equality constraint $d_H(\alpha(\overline{\mathbf{y}}), \alpha(\overline{\mathbf{x}})) = 0$. The second and most important issue is that, since codon equivalence is not evenly spread over the amino acids ensemble, the embedding limits for cDNA hosts are not immediately obvious.

3.1. Payload Computation

We assume no mutations throughout this subsection, that is, $\overline{\mathbf{z}} = \overline{\mathbf{y}}$. In this context capacity may be simply called payload (P).

Noncoding DNA. In this case the analysis is trivial. As $\mathbf{x}^B = \mathbf{y}^B$, and as DNA bases constitute a 4-ary alphabet, one can always embed $P_{\rm nc} = \log_2 |\mathcal{X}^B| = 2$ bits/base.

Coding DNA. We wish to determine the maximum number of sequences $\overline{\mathbf{y}}^{(m)}$ such that $d_H(\alpha(\overline{\mathbf{y}}^{(m)}), \alpha(\overline{\mathbf{x}})) = 0$, for $m = 1, 2, \ldots, |\mathcal{M}|$. The amount sought is just $|\mathcal{M}| = \prod_{i=1}^{N} \mu(\alpha(\mathbf{x}_i))$. Equivalently, the payload embeddable in $\overline{\mathbf{x}}$ is $P_{\rm c} = \frac{1}{N} \log_2 |\mathcal{M}| = \frac{1}{N} \sum_{i=1}^{N} \log_2 \mu(\alpha(\mathbf{x}_i))$ bits/codon. In order to see this result on average, we can use either the random variable $\mathbf{X}$ or else $X' = \alpha(\mathbf{X})$. The average payload is then $\overline{P}_{\rm c} = E[\log_2 \mu(\alpha(\mathbf{X}))] = E[\log_2 \mu(X')]$ bits/codon. For example, if $\mathbf{X}$ is uniform, $\overline{P}_{\rm c}^{\rm unif} = 1.7819$ 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, $\overline{P}_{\rm c}$ decreases to below one third of $3\,P_{\rm nc}$. A distribution $p(X')$ that maximises $\overline{P}_{\rm c}$ is any for which $E[\mu(X')] = 6$, 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 $P_{\rm c}^{\rm ub} \triangleq \log_2 6 = 2.5850$ bits/codon for any cDNA method, on any host —including those that are deterministic.

3.2. Capacity Computation

We assume next that $\overline{\mathbf{y}}$ can randomly mutate to yield a new sequence $\overline{\mathbf{z}}$. We will consider the symmetric "base mutation channel" in Fig. 1, whose transition probability matrix is $\Pi \triangleq [\pi_{i,j}]$ with $\pi_{i,j} = \pi_{j,i} = p(Z^B = j-1 | Y^B = i-1)$ 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 $q \triangleq 1 - \pi_{i,i} = \sum_j \pi_{i,j \neq i}$, 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 $\mathbf{x}^B = \mathbf{y}^B$, capacity is $C_{\rm nc} = \max I(Z^B; Y^B)$ 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] $C^{(M)} \triangleq \log_2 M + \sum_{k=1}^{M} \pi_{i,k} \log_2 \pi_{i,k}$ bits/symbol. Here, $M = 4$ and $C_{\rm nc} = C^{(4)} = 2 + \sum_{k=1}^{4} \pi_{i,k} \log_2 \pi_{i,k}$ bits/base, which is achieved for uniform Y^B [17]. Isothermal constraints, which imply that $p(Y^B = 0) + p(Y^B = 2) = \varepsilon\ (p(Y^B = 1) + p(Y^B = 3))$ with $\varepsilon \geq 0$, are sometimes enforced to speed up 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, $R_{\rm nc}^{\rm iso}(\varepsilon)$, is achieved when $p(Y^B = 0) = p(Y^B = 2) = 1/2(1+\varepsilon)$ and $p(Y^B = 1) = p(Y^B = 3) = \varepsilon/2(1+\varepsilon)$. Of course, $R_{\rm nc}^{\rm iso}(\varepsilon) \leq C_{\rm nc}$, with equality for $\varepsilon = 1$. Also, $R_{\rm nc}^{\rm iso}(\varepsilon) = R_{\rm nc}^{\rm iso}(1/\varepsilon)$ and the minimum is for $\varepsilon = 0$.

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] $C_{\rm c} = \max I(\mathbf{Z}; \mathbf{U}) - I(X'; \mathbf{U})$ bits/codon, where the maximisation is over all distributions $p(\mathbf{Y}, \mathbf{U} | X')$ under the constraint $d_H(\alpha(\mathbf{y}), x') = 0$, with $\mathbf{U}$ an auxiliary random variable. As $\mathbf{Y} = e(X', \mathbf{U})$, and as the support of $\mathbf{Y}|x'$ must be the set of codons $\mathcal{S}_{x'}$ corresponding to the amino acid x' —in order to satisfy the constraint—, then the cardinality of $\mathbf{U}|x'$ must be exactly $|\mathcal{S}_{x'}|$. As $\mathbf{U}$ must also be a good source code for X' in order to make $I(X'; \mathbf{U})$ small, the support of $\mathbf{U}|x'$ must actually be $\mathcal{S}_{x'}$. One can now establish $\mathbf{Y}|x' = \mathbf{U}|x'$ without loss of generality. This discussion on $\mathbf{U}$ also implies that $H(X'|\mathbf{U}) = 0$, since given a codon there will no uncertainty on the amino acid represented, and therefore $I(X'; \mathbf{U}) = H(X')$. Since $p(\mathbf{Y}|\mathbf{U}, X')$ is deterministic, we just have to determine next the maximising distribution $p(\mathbf{U}|X')$. See first that the codon mutation channel is symmetric with transition matrix $\Gamma = \Pi \otimes \Pi \otimes \Pi$, where $\otimes$ is the Kronecker product. If $\mathbf{X}$ is uniform and we also choose $\mathbf{U}|x'$ to be uniform for every x', then we achieve uniformity of $\mathbf{U}$. Since a uniform input maximises mutual information over a symmetric channel [17], the achievable rate in this case is

$$R_{\rm c}^{\rm unif} = C^{(64)} - H(X') \text{ bits/codon.} \tag{1}$$

For any $p(\mathbf{X})$ —not just the uniform— $\mathbf{U}|x'$ must also be uniform in a maximising strategy, because for any given $H(X')$ this will maximise $H(\mathbf{Z})$. In these conditions we can numerically evaluate Gel'fand and Pinsker's formula to obtain $R_{\rm c}$. As we can also write $H(\mathbf{U}) = H(X') + \overline{P}_{\rm c}$, the achievable rate can also be expressed as $R_{\rm c} = \overline{P}_{\rm c} - H(\mathbf{U}|\mathbf{Z})$.

Of course $R_{\rm c} \leq C_{\rm c}$. For the capacity, the inequality $C_{\rm c} \leq \min(P_{\rm c}^{\rm ub}, 3\,C_{\rm nc})$ always holds. For $q = 0$ we have that $R_{\rm c} = \overline{P}_{\rm c}$ since $H(\mathbf{U}|\mathbf{Z}) = 0$, and then the upper bound to $C_{\rm c}$ is achieved with any of the strategies discussed in Sect. 3.1. If we still use one of these strategies for $q > 0$, we see that we must choose one with X' deterministic to minimise $H(\mathbf{U}|\mathbf{Z})$. The optimum is for the amino acid that maximises $H(\mathbf{Z})$ among the three candidates. We may surmise that the rate $R_{\rm c}^*$ associated to this distribution of X' is actually $C_{\rm c}$, although this is not proved here. When $\pi_{i,j} = \xi$ for $i \neq j$, the optimising strategy is $X' = \text{Ser}$. Furthermore, any strategy with deterministic X' reaches capacity for $q = 3/4$, as in this

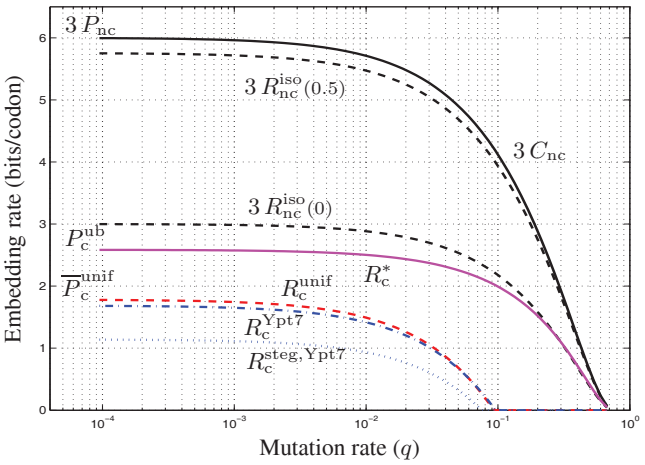

Fig. 2. Capacity and achievable rates for ncDNA and cDNA.

case $\Gamma = \frac{1}{64}\mathbf{1}^T\mathbf{1}$, and so $\mathbf{Z}$ is uniform independently of $\mathbf{U}$. Then $H(\mathbf{Z}) = H(\mathbf{Z}|\mathbf{U})$ and $C_{\rm c}|_{q=\frac{3}{4}} = C_{\rm nc}|_{q=\frac{3}{4}} = 0$.

3.3. Steganographic Rate

A real DNA sequence has a specific *codon count bias* [6], expressed in its characteristic 21 empirical pmfs $p(\mathbf{X}|x')$. 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 $p(\mathbf{U}|x')$ 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, $R_{\rm c}^{\rm steg} \leq R_{\rm c}$.

4. DISCUSSION

We assume here that $\pi_{i,j} = \xi$ for all $i \neq j$, and then $q = 3\xi$. Fig. 2 shows that the achievable rates can significantly decrease when the mutation rate increases. For cDNA the achievable rate can even be zero for $q < \frac{3}{4}$. The threshold where $R_{\rm c} = 0$ is dependent on $p(\mathbf{X})$. $R_{\rm c}^{\rm Ypt7}$ and $R_{\rm c}^{\rm steg,Ypt7}$ correspond to the Ypt7 sequence from yeast[1]. Also shown are two achievable rates with isothermal constraints $\varepsilon = 0$ and $\varepsilon = 0.5$. Let us evaluate next some results from the literature. Using ncDNA, Smith *et al.* [3] embed 5.8282 and 3.1610 bits/codon using Huffman and comma codes, respectively (the latter provides some resilience to deletion and insertion). As for cDNA, Shimanovsky *et al.* embed 1.6667 bits/codon in a test sequence [5]. On average their method will asymptotically achieve $\overline{P}_{\rm c}$, but it is too fragile for practical purposes: since it is based on arithmetic coding, mutation errors will propagate. On the other hand, a simple method based on the discussion in Section 3.1 will also achieve $\overline{P}_{\rm c}$ and be free of the error propagation issue. Modegi only embeds 0.1962 bits/codon [6] which he then exploits to recover the original codon sequence, albeit aided by external data.[2] Heider and Barnekow [7] embed 0.6408 bits/codon in the Ypt7 sequence under the very low mutation rates $q = 10^{-10}$ and $q = 10^{-7}$ (the only work to report q). In this nearly errorless scenario, the rate is even clearly below $R_{\rm c}^{\rm steg,Ypt7} \leq R_{\rm c}^{\rm Ypt7}$. To sum up, all of these methods implement rates far away from capacity and/or operate in the far low-error end of the achievable region, disregarding robustness or dealing with it suboptimally —by using repetition or Hamming codes.

[1]Data obtained from GenBank, accession number NC_001145.

[2]With no external aid and $q = 0$, on average one can revert cDNA embedding if $H(\mathbf{X}) \leq \overline{P}_{\rm c}$; for instance, this rules out uniform hosts. With ncDNA reversibility is always possible, as $H(X^B) \leq P_{\rm nc}$ always holds.

5. REFERENCES

[1] C. T. Clelland, V. Risca, and C. Bancroft, "Hiding messages in DNA microdots," *Nature*, vol. 399, no. 6736, pp. 533–534, June 1999.

[2] P. C. Wong, K. Wong, and H. Foote, "Organic data memory using the DNA approach," *Comms. of the ACM*, vol. 46, no. 1, pp. 95–98, January 2003.

[3] G. C. Smith, C. C. Fiddes, J. P. Hawkins, and J. P. Cox, "Some possible codes for encrypting data in DNA," *Biotech. Lett.*, vol. 25, no. 14, pp. 1125–1130, July 2003.

[4] N. Yachie, K. Sekiyama, J. Sugahara, Y. Ohashi, and M. Tomita, "Alignment-based approach for durable data storage into living organisms," *Biotechnol. Prog.*, vol. 23, no. 2, pp. 501–505, 2007.

[5] B. Shimanovsky, J. Feng, and M. Potkonjak, "Hiding data in DNA," in *Procs. of the 5th Intl. Workshop in Information Hiding*, Noordwijkerhout, The Netherlands, October 2002, pp. 373–386.

[6] T. Modegi, "Watermark embedding techniques for DNA sequences using codon usage bias features," in *16th Intl. Conf. on Genome Informatics*, Yokohama, Japan, December 2005.

[7] D. Heider and A. Barnekow, "DNA-based watermarks using the DNA-Crypt algorithm," *BMC Bioinformatics*, vol. 8, no. 176, February 2007.

[8] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, July and October 1948.

[9] P. Moulin and R. Koetter, "Data-hiding codes," *Proc. IEEE*, vol. 93, no. 12, pp. 2083–2126, December 2005.

[10] W. Wayt Gibbs, "The unseen genome: Gems among the junk," *Scientific American*, p. 53, November 2003.

[11] R. S. Eisenberg, "Structure and function in gene patenting," *Nature Genetics*, vol. 15, no. 2, pp. 125–130, 1997.

[12] C. Bancroft and C. T. Clelland, "DNA-based steganography," U.S. Patent 6,312,911, June 2001.

[13] T. A. Kunkel, "DNA replication fidelity," *J. Biol. Chem.*, vol. 279, no. 17, pp. 16895–16898, April 2004.

[14] Y. Fu, "Estimating mutation rate and generation time from longitudinal samples of DNA sequences," *Mol. Biol. and Evolution*, vol. 18, no. 4, pp. 620–626, 2001.

[15] S. S. Pradhan, J. Chou, and K. Ramchandran, "Duality between source coding and channel coding and its extension to the side information case," *IEEE Trans. on Inf. Theory*, vol. 49, no. 5, pp. 1181–1203, May 2003.

[16] R. J. Barron, B. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information and some applications," *IEEE Trans. on Inf. Theory*, vol. 49, no. 5, pp. 1159–1180, May 2003.

[17] R. B. Ash, *Information Theory*, Dover, New York, 1965.

[18] S. I. Gel'fand and M. S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory*, vol. 9, no. 1, pp. 19–31, 1980.