

Provided by the author(s) and University College Dublin Library in accordance with publisher policies. Please cite the published version when available.

|                                     |   |
|-------------------------------------|---|
| Title                               | On the embedding capacity of DNA strands under insertion, deletion and substitution mutations |
| Author(s)                           | Balado, Félix   |
| Publication Date                    | 2010-01   |
| Publication information             | Memon, N. D. et al. (eds.). Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol. 7541       |
| Publisher                           | SPIE--The International Society for Optical Engineering                                       |
| Link to publisher's version         | <a href="http://dx.doi.org/10.1117/12.838537">http://dx.doi.org/10.1117/12.838537</a>         |
| This item's record/more information | <a href="http://hdl.handle.net/10197/1863">http://hdl.handle.net/10197/1863</a>               |

Downloaded 2012-05-16T20:36:53Z

Some rights reserved. For more information, please see the item record link above.



# On the Embedding Capacity of DNA Strands under Substitution, Insertion, and Deletion Mutations

Félix Balado

School of Computer Science & Informatics  
University College Dublin, Belfield Campus, Dublin, Ireland

## ABSTRACT

A number of methods have been proposed over the last decade for embedding information within deoxyribonucleic acid (DNA). Since a DNA sequence is conceptually equivalent to a unidimensional digital signal, DNA data embedding (diversely called DNA watermarking or DNA steganography) can be seen either as a traditional communications problem or as an instance of communications with side information at the encoder, similar to data hiding. These two cases correspond to the use of noncoding or coding DNA hosts, which, respectively, denote DNA segments that cannot or can be translated into proteins. A limitation of existing DNA data embedding methods is that none of them have been designed according to optimal coding principles. It is not possible either to evaluate how close to optimality these methods are without determining the Shannon capacity of DNA data embedding. This is the main topic studied in this paper, where we consider that DNA sequences may be subject to substitution, insertion, and deletion mutations.

**Keywords:** DNA data embedding, data hiding, steganography, Shannon capacity.

## 1. INTRODUCTION

The last ten years have witnessed the proposal of numerous methods for embedding nongenetic information within DNA, the molecule that encodes genetic information in all living organisms.<sup>1-14</sup> A DNA sequence is conceptually equivalent to a digital signal, and due to this fact DNA data embedding is in essence an instance of digital communications. Therefore many techniques from this area are directly applicable to DNA data embedding. Furthermore, side information at the encoder has to be taken into account in the important scenario in which gene-encoding DNA strands are used as hosts. This makes this problem a very particular case of digital data hiding already, but we will see that true steganographic constraints can also be considered.

Most theoretical issues related to DNA data embedding are not yet elucidated. Perhaps the most fundamental 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 would allow to assess optimality of practical DNA data embedding methods with respect to their information-carrying abilities. In a biological context errors are tantamount to mutations, and therefore we will talk herein of “mutation channels” undergone by an information-carrying host DNA strand, thus drawing the parallel with a communications channel that introduces random errors. In this paper we will address the computation of the Shannon capacity<sup>15</sup> of DNA data embedding under several important mutation channels: substitutions, insertions, and deletions. As we will see, the first of these channels is well known from communications, and therefore straightforwardly extended to the problem at hand when no side information is used by the encoder. Side-informed scenarios can also be handled using standard data hiding theory. Insertion and deletion channels are less well understood in communications. However, it is possible to exploit biological constraints in order to obtain meaningful bounds to capacity in this case, relying on the erasure symmetric channel.

---

Further author information:

E-mail: [felix@ucd.ie](mailto:felix@ucd.ie), Telephone: +353 1 716 2927

## 2. PRELIMINARY CONCEPTS

The importance of the DNA molecule relies on it containing the instructions for the development and functioning of all living beings. Chemically, DNA is formed by two backbone strands helicoidally twisted around each other, and mutually attached by means of two *base* sequences. The four possible bases are the molecules adenine, cytosine, thymine, and guanine, abbreviated A, C, T and G, respectively. Only the pairings A-T and C-G can exist between the two strands, which is why each of the two base sequences is completely determined by the other, and also why the length of a DNA molecule is measured in base pairs (bp). According to this brief description, the interpretation of DNA as a one-dimensional discrete digital signal is straightforward: any of the two base sequences constitutes a digital signal formed by quaternary (4-ary) symbols.

As regards biological meaning, it suffices to know that *codons* —the minimal biological “codewords”— are formed by triplets of consecutive bases in a base sequence. Given any three consecutive bases there is no ambiguity in the codon they stand for, since there is only one direction in which a base sequence can be read. This is called the 5'-3' direction, in reference to certain feature points in a DNA backbone strand. The two base sequences in a DNA molecule are read in opposite directions, and because of this and of their complementarity they are termed antiparallel. Groups of consecutive codons in some special regions of a DNA sequence can be translated into a series of compounds called *amino acids* via their transcription to the intermediary ribonucleic acid (RNA) molecule. RNA is similar to DNA but single-stranded, and has the same information content as the codon sequence to be translated—the most important difference is that thymine is replaced by uracil in RNA's 4-ary alphabet. Amino acids are sequentially assembled in the same order imposed by the codon sequence. The result of this assembling process are proteins, which are the basic compounds of the chemistry of life. There are  $4^3 = 64$  possible codons, since they are triplets of 4-ary symbols. Crucially, there are only 20 possible amino acids, mapped to the 64 codons according to the so-called *genetic code* in Table 1 (which uses the arbitrary mapping  $\{A, C, T, G\} \leftrightarrow \{0, 1, 2, 3\}$ ). This redundancy implies that the genetic code effectively implements a biological error-correcting code in the translation of a base sequence to proteins.

The genome of an organism is the ensemble of all its DNA. Segments of a genome that can be translated into proteins through the process described above are called *coding* DNA (cDNA), whereas those segments that never get translated are called *noncoding* DNA (ncDNA). A *gene* is a cDNA segment, or group of segments, which encodes one single protein\*, and which is flanked by certain start and stop codons (see Table 1) plus other markers. cDNA segments constitute 1.2% of the total in the human genome, but 70% in *S. Cerevisiae* (baker's yeast).<sup>16</sup> Finally, for each base sequence there are three different reading frames which determine three different codon sequences. The correct reading frame is marked by the position of a start codon.

**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  $X$ . Also,  $H(q)$  is the entropy of a Bernoulli( $q$ ) random variable. The Hamming distance between vectors  $\mathbf{x}$  and  $\mathbf{y}$ , which gives the number of different same index elements between the two vectors, is denoted by  $d_H(\mathbf{x}, \mathbf{y})$ . All logarithms are base 2, unless otherwise indicated.

We will denote a cDNA sequence by a vector  $\bar{\mathbf{x}} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  whose elements are consecutive codons from one of the two antiparallel base sequences, assuming a suitable reading frame. That is,  $\mathbf{x}_i \in \mathcal{X} \triangleq \mathcal{X}^3$ , with  $\mathcal{X} \triangleq \{0, 1, 2, 3\}$ . We indicate by  $\mathbf{x}^B$  the representation of  $\bar{\mathbf{x}}$  using bases, that is,  $\mathbf{x}^B$  is a  $3N$ -length sequence with  $x_i^B \in \mathcal{X}$ . We also denote by  $x'_i \triangleq \alpha(\mathbf{x}_i)$  the amino acid into which a codon  $\mathbf{x}_i$  translates (see genetic code in Table 1); similarly,  $\mathbf{x}' = \alpha(\bar{\mathbf{x}}) = \{x'_1, x'_2, \dots, x'_N\}$  is the unique amino acid sequence defined by  $\bar{\mathbf{x}}$ . If  $\bar{\mathbf{x}}$  represents a gene,  $\alpha(\bar{\mathbf{x}})$  is the so-called primary structure of the protein it encodes. The multiplicity associated with an amino acid  $x'$  is written as  $\mu(x')$ . A ncDNA sequence will be denoted by an  $\mathbf{x}^B$  vector of arbitrary length, that is, not necessarily a multiple of 3. A cDNA sequence could be for instance  $\bar{\mathbf{x}} = \{\{2, 0, 2\}, \{2, 3, 1\}\}$ , encoding the amino acid sequence  $\alpha(\bar{\mathbf{x}}) = \{\text{Tyr}, \text{Cys}\}$ . In this case  $\mathbf{x}^B = \{2, 0, 2, 2, 3, 1\}$ .

---

\*We will assume that the DNA segments corresponding to a gene are always contiguous, or *spliced*. In eukaryotic organisms (fungi, plants, animals, ...) cDNA segments corresponding to the same gene (called *exons*) are frequently interspersed with ncDNA segments (called *introns*). Splicing is the genetic process by which introns are removed and exons are sequentially joined, before their translation into a protein. Alternative splicings are sometimes possible. Genes of more primitive prokaryotic organisms, such as bacteria, have no divisions into introns and exons and need no splicing.

| Amino acid, $x'$        | Phe | Tyr | Cys | Ser | Leu | <i>Stp</i> | 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 | 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 (genetic code). *Stp*, the ensemble of stop codons, is loosely classed as an “amino acid”; 023 (in eukaryotes or prokaryotes) and also 123 or 223 (in prokaryotes) may work as start codons. Note that  $\sum_{x'} \mu(x') = 64$ .

## 2.1 A Short Introduction to DNA Data Embedding

With the coming of age of genomics, researchers realised in the late 1990s that DNA can also be used to convey arbitrary data, and not only genetic information. The two key techniques that physically enable DNA data embedding are recombinant DNA methods (used to “write” DNA, by means of the creation and insertion of artificial DNA in a genome), and DNA sequencing (used to “read” DNA). The polymerase chain reaction (PCR) technique plays a foremost role in sequencing. Here we will not focus on such physical aspects, but rather on the information-theoretical properties of DNA data embedding algorithms when analysed as mathematical functions that embed information in (and retrieve information from) DNA. All prior art in DNA data embedding—which is briefly reviewed next—deals with the proposal of practical methods. These can be divided into two groups:

- Methods which encode information using DNA molecules, but do not aim at manipulating or modifying the DNA of living beings.<sup>1–3,7</sup> Clelland *et al.*<sup>1</sup> were the first to propose what they called DNA-based steganographic methods, and the first to implement and demonstrate their proposal *in vitro*. Their use of “steganography” is unrelated to the currently most widespread meaning of this term (that is, statistical undetectability according to Cachin’s criterion<sup>17</sup>), and refers instead to physically concealing information in such a way that PCR and a secret key are necessary for its retrieval. From a communications perspective, the generation of information-carrying DNA sequences is equivalent in this case to establishing a function  $e(\cdot) : \mathcal{M} \rightarrow \mathcal{X}^L$ , that generates a 4-ary signal  $\mathbf{y}^B = e(m)$  from a message  $m$  chosen from a finite set  $\mathcal{M}$ .
- Methods which aim at unobtrusively embedding information in the DNA of living beings, so that such information will travel alongside each cell replication without affecting whatsoever the biological properties of the organism.<sup>†</sup> This concept was first proposed in 2001 by Cox,<sup>4</sup> who suggested to use spores of the *B. Subtilis* bacteria or the *S. Cerevisiae* fungus to host information. There are two distinct approaches to embedding data within the DNA of living beings:
  1. Replacement or appendage of ncDNA segments, which never get translated to proteins.<sup>6,11,14</sup> Similarly to the aforementioned group of methods not aimed at living beings, this amounts to generating an ad-hoc information-carrying sequence  $\mathbf{y}^B$  which is then used to replace or append a noncoding segment of a host genome. ncDNA can therefore be seen as always non-side-informed, in communications terms. This approach was first implemented *in vivo* by Wong *et al.*<sup>6</sup> in 2003 using two bacterium species (*E. Coli* and *D. Radiodurans*). Since ncDNA does not contain genes, this procedure relies on the hypothesis that ncDNA modifications will not alter the protein profile of an organism. Nonetheless recent investigations show that many parts of ncDNA are not really “junk”<sup>19</sup> (as it has sometimes been dubbed), due to their involvement in regulating important functions of the genetic machinery. Although ncDNA data embedding has been successfully implemented *in vivo* both using bacteria<sup>6,11,14</sup> and fungi,<sup>14</sup> extreme care is needed in order to guarantee that ncDNA modification or appendage does not alter some biological functionality—worryingly enough, perhaps yet unknown.

<sup>†</sup>Information transmission down subsequent generations applies to the whole genome of asexual organisms such as bacteria. In organisms with sex—such as fungi, plants, or animals—it only applies to two parts of the genome: the non-recombinant part of the Y-chromosome in males, and the mitochondrial DNA in both sexes.<sup>18</sup>

Heider and Barnekow<sup>14</sup> have actually shown that noncoding regions acting as gene promoters or regulatory regions—which although not translated into proteins can be transcribed into RNA—are not suitable in general for data embedding. They consequently state that ncDNA data embedding must be examined on a case-by-case basis.

2. Modification of cDNA segments, which may get translated to proteins.<sup>5, 9, 10, 12, 20</sup> This situation is substantially different to the ones previously discussed, since in this case a host sequence  $\bar{\mathbf{x}}$  has to be carefully modified to embed a message  $m$ . The embedding function is now of the form  $e(\cdot, \cdot) : \mathcal{X}^N \times \mathcal{M} \rightarrow \mathcal{X}^N$ , and so the information-carrying signal is obtained as  $\bar{\mathbf{y}} = e(\bar{\mathbf{x}}, m)$ . Notice that this function is completely equivalent to an embedding function in data hiding. The fundamental issue is to guarantee that  $\bar{\mathbf{y}}$  and  $\bar{\mathbf{x}}$  be biologically identical, that is,  $\alpha(\bar{\mathbf{y}}) = \alpha(\bar{\mathbf{x}})$ , so that both sequences represent the same amino acid sequence. This can be achieved by exploiting the codon equivalence implicit in the genetic code (Table 1). An important implication of this constraint is that all cDNA data embedding methods must be side informed, since it is not possible to build an information-carrying “watermark”  $\bar{\mathbf{w}} \in \mathcal{X}^N$  independently of  $\bar{\mathbf{x}}$  if we wish to enforce that  $\bar{\mathbf{y}} = \bar{\mathbf{x}} \oplus \bar{\mathbf{w}}$  (using modulo-4 addition) be biologically equivalent to  $\bar{\mathbf{x}}$ . See as well that the genetic constraint plays a similar role as the one played by a perceptual distance constraint in multimedia data hiding. The main difference is that in DNA data embedding we have to deal with a deterministic equality constraint, as opposed to the typical statistical inequality constraint used in multimedia data hiding.

Considering the previously mentioned concerns about the risks lurking behind ncDNA modification, cDNA data embedding currently seems to be a more systematic and safer approach. This is also argued by Shimanovsky *et al.*,<sup>5</sup> who were the first to propose cDNA data embedding in 2002, and also the first to demonstrate a practical method *in silico* along these lines. Arita and Ohashi<sup>9</sup> were the first ones to implement a cDNA data embedding strategy *in vivo*, targeting the *B. Subtilis* bacterium, and Heider and Barnekow<sup>20</sup> the first to use cDNA of a living eukaryotic organism (*S. Cerevisiae*) as a host.

## 2.2 Applications

The goals of DNA data embedding can be at least twofold:

- Tagging genetic material for tracking purposes. Reliable DNA embedding offers the possibility of attaching unique fingerprints to differentiate large numbers of otherwise functionally identical genetic strands. Current genetic profiling techniques are passive and usually rely on ncDNA features, but DNA data embedding makes active fingerprinting of genes possible. One potential application is tracking both temporal and spatial pathways followed by individual organisms whose protein expression profile is identical. Another interesting application may be detecting or estimating mutations in organism populations by solely relying on embedded information. DNA data embedding becomes especially relevant to this case when there is not one unique host DNA sequence to use as a reference, such as in viral quasispecies.

Many authors have argued that another tracking application stems from the fact that DNA sequences represent valuable intellectual property.<sup>5, 6, 9, 10, 12</sup> Gene patents have proved to be commercially important,<sup>21</sup> which is also illustrated by the existence of several DNA data embedding patents (see for instance<sup>22</sup>). 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. However, it must be noted that the effect of any DNA data embedding method amounts to that of writing to a digital memory. In particular, prior information embedded on a given DNA sequence is necessarily overwritten, unless appending is used. Thus any protection granted by DNA data embedding can in principle be easily thwarted—for a given sequence—by an active third party. Therefore these strategies can only rely on the current difficulty to physically undertake such attacks on a large scale. Because of this, DNA data embedding can only probably be used in this scenario as a deterrent.

- Using genetic material as a massive and compact storage media. Long-term storage of data in the DNA of living organisms, such as bacteria, was first proposed by Cox.<sup>4</sup> DNA persistent memories may currently seem impractical due to their ridiculously low read/write throughput. However this will in all likelihood

radically improve in the future, as new fast sequencing techniques such as nanopore sequencing are being developed.<sup>23</sup> What is clear is that DNA is the oldest standard digital format, extremely compact, and unlikely to change in the times to come.<sup>4</sup>

Finally, the emerging area of DNA computation, in which DNA is used to realise massively parallel computations, always requires to read and write information using DNA. It is conceivable that future applications will require using the cDNA of living beings and then make use of DNA data embedding techniques.

Safety, ethical, and privacy considerations must of course be taken into account when introducing synthetic information into living beings. Safety concerns in DNA data embedding are the same as in any procedure based on recombinant DNA. We have argued that DNA data embedding is safe as long as it is limited to cDNA, which has been successfully tested *in vivo* by several groups of researchers. As regards ethical and privacy issues, it must be noted that genetic profiling based on ncDNA features allows to pinpoint and track individual genomes already, although not in all the scenarios discussed above using DNA data embedding.

### 3. LIMITS OF DNA DATA EMBEDDING

Any DNA data embedding method requires an associated decoding function  $d(\cdot) : \mathcal{X}^N \rightarrow \mathcal{M}$  in order to retrieve the information embedded in a given DNA sequence. However random mutations can occur in any real scenario, which may cause a mutated sequence  $\bar{\mathbf{z}}$  to be present at the decoder instead of the original information-carrying sequence  $\bar{\mathbf{y}}$ . In this case, the decoded message  $\hat{m} = d(\bar{\mathbf{z}})$  may differ too from the original message  $m = d(\bar{\mathbf{y}})$ , that is, mutations affecting the information-carrying sequence can lead to decoding errors in the embedded information.

Mutations happen in the genomes of organisms both during their lifetime and in their replication processes. Some of these errors cannot be corrected by the natural DNA self-repair mechanisms (basically reliant on the complementarity of the two strands), and hence they accumulate in each new generation as long as they are not deleterious for survival. The DNA molecule itself, although very stable, is also subject to degradation over time if left to its own devices outside the replication cycle —such as in a laboratory sample, or in a dead organism.

A recent study by Kunkel<sup>24</sup> suggests that single base substitution error rates in prokaryotic DNA *in vivo* may be in the range of  $10^{-8}$  to  $10^{-7}$  per replication. A standard single base substitution error rate of  $10^{-10}$  per replication in eukaryotes is cited in.<sup>12</sup> 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  $10^{-1}$  according to Kunkel.<sup>24</sup> Furthermore, these rates refer to one single replication, but, as discussed, mutations accumulate over time in living organisms. After  $R$  replications a constant mutation rate  $q$  becomes  $q^{(R)} = 1 - (1 - q)^R$ , and  $q^{(R)} \rightarrow 1$  as  $R \rightarrow \infty$ . For instance, Fu<sup>25</sup> has estimated an accumulated single base substitution error rate of  $1.71 \times 10^{-2}$  over a year in the genome of HIV (which is an RNA virus and therefore particularly unstable). Mutations other than substitutions are not less important: the study by Kunkel reveals that single base deletion error rates in DNA synthesis are in the range  $10^{-5}$  to  $10^{-1}$  per replication.<sup>24</sup>

For all these reasons robustness in DNA data embedding is paramount both in organisms with high generations-per-day ratios and in environments with high mutation rates, or simply when the information embedded must be kept intact over protracted periods of time.

#### 3.1 Mutation Channels

As we have mentioned in the introduction, an information-carrying DNA molecule undergoing mutations can be readily seen as a digital signal undergoing a noisy communications channel, which we may term “mutation channel”. We will consider herein three important types of mutations: substitutions, insertions and deletions. We will assume that mutations are mutually independent, which is a worst-case analysis in terms of capacity.

In order to model substitution mutations, that is, those that randomly flip letters from the DNA alphabet, we will consider the symmetric base mutation channel in Figure 1, whose transition probability matrix is  $\Pi \triangleq [\pi_{i,j}]$  with  $\pi_{i,j} = p(Z^B = j-1 | Y^B = i-1) = \frac{q}{3}$  for  $i, j = 1, 2, 3, 4$  with  $i \neq j$ , and  $\pi_{i,i} = p(Z^B = i-1 | Y^B = i-1) = 1 - q$  for  $i = 1, 2, 3, 4$ . Therefore  $q$  is the probability of base substitution, or base substitution mutation rate.

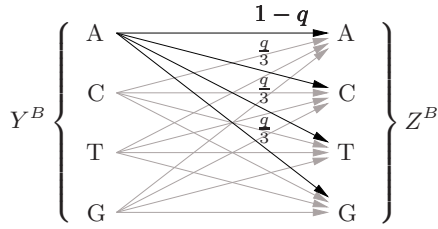


Figure 1. Base substitution mutation channel.

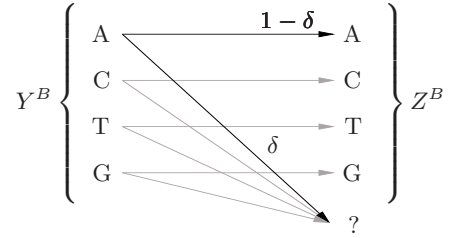


Figure 2. Base erasure mutation channel.

Insertions and deletions of bases in a sequence are commonly referred to in bioinformatics by the *indels* portmanteau.<sup>16</sup> Mutations caused by indels are generically called frameshift mutations. This is because one single indel will alter the codon reading frame after its occurrence, and therefore subsequent codons will be read with the wrong framing alignment during their translation to amino acids. For this reason a single indel will usually be deleterious for the translation of a gene into a protein. However if the difference between the number of insertions and deletions happens to be a multiple of three, only the portion of the mutated sequence between the first and the last indel will be misaligned with respect to the reading frame, which may not overly affect the functionality of the gene. Note however that a gene broken by a mutation does not always hamper the survival of the individual, and then any information carried by it may still be passed on.

For data embedding purposes, it is important to realise that insertions are just a synchronisation concern, whereas deletions also effectively erase information. If we could avail of a clairvoyant way to resynchronise the positions of an information-carrying sequence subject to frameshift mutations, then inserted bases could simply be ignored, but deleted bases would always be irrecoverable. We will exploit this fact to analyse the properties of the indel mutation channel by means the erasure mutation channel depicted in Figure 2, in which the base erasure probability  $\delta = p(Z^B = ?)$  is set to the base deletion probability. We will also consider the quaternary symmetric erasure mutation channel, in order to tackle the combined effect of substitutions and indels.

### 3.2 Capacity Analysis

A number of prior DNA data embedding methods have attempted to provide robustness against mutations by means of techniques from digital communications. For instance, Arita and Ohashi<sup>9</sup> use parity bits for error detection. Arita<sup>8</sup> combines a template with error-correction codes for adding robustness against substitution errors to comma-free codes—which offer already some robustness against indels. The method by Yachie *et al.*<sup>11</sup> uses repetition to counteract both substitutions and indel mutations. Finally Heider and Barnekow<sup>12</sup> use a Hamming error-correcting code in their DNA-Crypt algorithm.

Nevertheless none of the above methods is known to be optimum in the sense of achieving Shannon’s capacity—maximum asymptotically errorless transmission rate of a given communications system—, which is in fact yet unknown in most relevant DNA data embedding scenarios. The computation of capacity in channels such as the ones described in Section 3.1 is as important as in any other communications setting, since it allows both to gauge the ultimate limits that can be achieved with practical DNA data embedding methods, and to determine their optimality. The remainder of this paper is devoted to studying this problem.

In the ncDNA case we will assume that we can choose  $\mathbf{y}^B$  independently of  $\mathbf{x}^B$ , since we are assuming that we can append or replace the host at will without ensuing biological effects. Although we have discussed that this is not completely true, this assumption will allow us to obtain upper bounds on the achievable embedding rates. In the cDNA case we will assume that we can only choose among sequences that are biologically equivalent to the host  $\bar{\mathbf{x}}$  according to the genetic code, that is, those  $\bar{\mathbf{y}}$  for which  $\alpha(\bar{\mathbf{y}}) = \alpha(\bar{\mathbf{x}})$ . In the following we will assume that the elements of  $\bar{\mathbf{x}}$  are independently 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  $\bar{\mathbf{x}}$  followed by deinterleaving of  $\bar{\mathbf{y}}$ .

#### 3.2.1 Payload Computation

In order to see the absolute limits to DNA data embedding is interesting to briefly analyse the particular situation in which there are no mutations whatsoever, that is,  $\bar{\mathbf{z}} = \bar{\mathbf{y}}$ . In this context we may refer to capacity

as information payload ( $P$ ), or just payload.

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

**Coding DNA.** Given a particular host  $\bar{\mathbf{x}}$ , we wish to determine the maximum number of  $\bar{\mathbf{y}}^{(m)}$  sequences such that  $\alpha(\bar{\mathbf{y}}^{(m)}) = \alpha(\bar{\mathbf{x}})$ , for  $m = 1, 2, \dots, |\mathcal{M}|$ . The amount sought is just  $|\mathcal{M}| = \prod_{i=1}^N \mu(\alpha(\mathbf{x}_i))$ . Equivalently, the payload embeddable in  $\bar{\mathbf{x}}$  is  $P_c = \frac{1}{N} \log |\mathcal{M}| = \frac{1}{N} \sum_{i=1}^N \log \mu(\alpha(\mathbf{x}_i))$  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

$$\bar{P}_c = E[\log \mu(\alpha(\mathbf{X}))] = E[\log \mu(X')] \text{ bits/codon.} \quad (1)$$

For example, if  $\mathbf{X}$  is uniform then  $\bar{P}_c^{\text{unif}} = 1.7819$  bits/codon. Note that 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, the average payload has decreased in this case to below one third of  $3P_{\text{nc}}$ .

A distribution  $p(X')$  that maximises  $\bar{P}_c$  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

$$P_c^{\text{ub}} \triangleq \log 6 = 2.5850 \text{ bits/codon,} \quad (2)$$

for any cDNA data embedding method, using any host sequence. An important consequence of this bound is that, since  $P_c^{\text{ub}} < 3P_{\text{nc}}$ , side-informed cDNA data embedding capacity will not be able to achieve non-side-informed ncDNA capacity for all mutation rates. This is in parallel to other results in data hiding with discrete hosts,<sup>26,27</sup> and unlike the well-known result by Costa for continuous Gaussian hosts.<sup>28</sup>

### 3.2.2 Capacity Computation with Substitution Mutations

We assume in this section that  $\bar{\mathbf{y}}$  can randomly mutate to yield  $\bar{\mathbf{z}}$ , assuming the substitution mutation channel in Figure 1. In order to establish the amount of information that can theoretically be embedded in this scenario it is necessary to resort to the aforementioned concept of Shannon capacity.<sup>15</sup>

**Noncoding DNA.** In this case, in which we can freely choose the input  $\mathbf{y}^B$  to the mutation channel, capacity is  $C_{\text{nc}} = \max I(Z^B; Y^B)$  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 is given by  $C_s^{(M)} \triangleq \log M - H(q) - q \log(M-1)$  bits/symbol,<sup>29</sup> and which is achieved for a uniform input. Here,  $M = 4$  and then

$$C_{\text{nc}} = C_s^{(4)} = 2 + (1-q) \log(1-q) + q \log \frac{q}{3} \text{ bits/base.} \quad (3)$$

*Steganographic rate.* The uniform distribution of  $Y^B$  required to reach capacity will be different in general to the distribution  $p(X^B)$  of the ncDNA host sequence which is replaced or appended. We recall that Cachin's criterion for perfect steganography<sup>17</sup> requires that the distribution of the information-carrying signal be identical to that of the host. However it is also possible to obtain a steganographic achievable rate for any given host by computing  $R_{\text{nc}}^{\text{steg}} = I(Z^B; Y^B)$  with  $p(Y^B) = p(X^B)$ . Notice that this approach assumes independence of the host bases, which is not realistic in general. Dependencies between bases should be taken into account (for instance, using Markov chain models) if a more accurate result is wanted. For this reason,  $R_{\text{nc}}^{\text{steg}}$  is an upper bound to the actual steganographic rate of a ncDNA sequence.

*Isothermal rate.* In ncDNA data embedding scenarios such as DNA computing additional isothermal constraints can be enforced on the information-carrying signal in order to speed up PCR-based replication.<sup>7</sup> Isothermality means that the ratio between the content of A and T bases in the DNA sequence must be in a certain fixed proportion  $\varepsilon > 0$  with respect to the content of C and G bases in order to foster pairings between two different strands. In other words,

$$p(Y^B = 0) + p(Y^B = 2) = \varepsilon (p(Y^B = 1) + p(Y^B = 3)). \quad (4)$$

It is shown in Appendix A that the maximum isothermal rate  $R_{\text{nc}}^{\text{iso}}(\varepsilon)$  is achieved for the following  $p(Y^B)$ :

$$p(Y^B = 0) = p(Y^B = 2) = \frac{1}{2(1 + \varepsilon)}, \quad p(Y^B = 1) = p(Y^B = 3) = \frac{\varepsilon}{2(1 + \varepsilon)}. \quad (5)$$

Of course,  $R_{\text{nc}}^{\text{iso}}(\varepsilon) \leq C_{\text{nc}}$ , with equality for  $\varepsilon = 1$  since this yields a uniform input distribution. Also,  $R_{\text{nc}}^{\text{iso}}(\varepsilon) = R_{\text{nc}}^{\text{iso}}(1/\varepsilon)$  and the minimum is found for  $\varepsilon = 0$ .

**Coding DNA.** As we have discussed, the relevance of data hiding research to this problem is clear. The bulk of data hiding research has dealt with continuous-valued signals, and only to a lesser extent with discrete-valued signals such as DNA. Standard data hiding using discrete binary (2-ary) hosts bears some resemblances but also some important differences to cDNA data embedding. To illustrate this assume a discrete 2-ary host  $\bar{\mathbf{x}} = \{x_1, \dots, x_N\}$ , with  $x_i \in \mathcal{X} = \{0, 1\}$ , which is modified to embed a message  $m$ . The watermarked signal  $\bar{\mathbf{y}} = e(\bar{\mathbf{x}}, m)$  must be “close” to  $\bar{\mathbf{x}}$ , where closeness is usually measured by means of the Hamming distance  $d_H(\bar{\mathbf{y}}, \bar{\mathbf{x}})$ . When  $\bar{\mathbf{y}}$  is subject to distortions the decoder receives  $\bar{\mathbf{z}} = \bar{\mathbf{y}} \oplus \bar{\mathbf{n}}$  (with  $n_i \in \mathcal{X}$  and using modulo-2 addition). Pradhan *et al.*<sup>26</sup> and Barron *et al.*<sup>27</sup> have determined the maximum rate  $R^{\text{unif}}$  (in bits/host symbol) that can be embedded and decoded asymptotically without errors when the elements of the host random variable  $\bar{\mathbf{X}}$  are uniformly distributed and mutually independent. Their result is for the embedding constraint  $\frac{1}{N}E[d_H(\bar{\mathbf{Y}}, \bar{\mathbf{X}})] \leq d$  and Bernoulli( $q$ ) channel distortion (that is, when  $\bar{\mathbf{z}}$  is the output of a binary symmetric channel with input  $\bar{\mathbf{y}}$  and crossover probability  $q$ ).

Our goal for cDNA data embedding, in which discrete 4-ary hosts are used, is the same as above. However the first issue for capacity analysis is that nonzero inequality constraints on the average Hamming distance —such as the ones used in<sup>26,27</sup>— are meaningless if one wants to carry through to  $\bar{\mathbf{y}}$  the full biological functionality of  $\bar{\mathbf{x}}$ . Instead, one must always establish the null equality constraint  $E[d_H(\alpha(\bar{\mathbf{Y}}), \alpha(\bar{\mathbf{X}}))] = 0$ , which amounts to the deterministic constraint  $d_H(\alpha(\bar{\mathbf{y}}), \alpha(\bar{\mathbf{x}})) = 0$ . The second and most important issue is that, since codon equivalence is not evenly spread over the ensemble of amino acids, the embedding limits for cDNA hosts are not immediately obvious.

Since side information must be taken into account, capacity is then given in this scenario by Gel’fand and Pinsker’s formula<sup>30</sup>  $C_c = \max I(\mathbf{Z}; \mathbf{U}) - I(X'; \mathbf{U})$  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 biological 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 be no uncertainty on the amino acid represented, and therefore  $I(X'; \mathbf{U}) = H(X')$ . Since  $\mathbf{Y}|\mathbf{U}, X'$  is deterministic, we just have to determine next the maximising distribution  $p(\mathbf{U}|X')$ . See first that the  $64 \times 64$  codon mutation channel has transition matrix  $\Gamma \triangleq \Pi \otimes \Pi \otimes \Pi$ , where  $\otimes$  is the Kronecker product, and that therefore is symmetric if the base mutation channel is symmetric too. 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,<sup>29</sup> the achievable rate in this case is

$$R_c^{\text{unif}} = C_s^{(64)} - H(X') \text{ bits/codon}. \quad (6)$$

Since the three parallel symmetric channels undergone by the three bases in a codon are independent, we can also use the equality  $C_s^{(64)} = 3C_s^{(4)}$  in (6). Therefore (6) is telling us an intuitive fact: the cDNA embedding rate is the same as three times the ncDNA embedding rate minus the information needed to convey the amino acid representation of the host, since  $H(X')$  is the lower bound to any lossless source encoding of this variable.

In general real cDNA sequences are not uniform, and then a closed-form formula such as (6) is not always possible. However the achievable rate  $R_c$  with nonuniform hosts can still be obtained by numerical evaluation of Gel’fand and Pinsker’s formula. For a given host distribution, the maximum achievable rate is attained when  $\mathbf{U}|x'$  is uniform, because for any fixed  $H(X')$  this will always maximise  $H(\mathbf{Z})$ . In these conditions  $H(\mathbf{U}) = H(X') + \overline{P}_c$ , and then the achievable rate is  $R_c = \overline{P}_c - H(\mathbf{U}|\mathbf{Z})$ . With this formula we can see that, as expected,  $R_c = \overline{P}_c$  when  $q = 0$ , because in this case  $H(\mathbf{U}|\mathbf{Z}) = 0$ .

The computation of capacity ( $C_c$ ) requires maximising Gel'fand and Pinsker's formula over all  $p(X')$ . Of course,  $R_c \leq C_c$  for any host, but in any case it is also clear that  $C_c \leq \min(P_c^{\text{ub}}, 3C_{\text{nc}})$ . With this inequality it is easy to find the capacity-achieving strategies at the extreme values of  $q$ . As  $R_c = \bar{P}_c$  when  $q = 0$ , we have that in this case the upper bound on  $C_c$  is achieved by any of the distributions discussed at the end of Sect. 3.2.1. Furthermore, for  $q = 3/4$  any strategy with deterministic  $X'$  reaches capacity as well, as in this 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  $C_c|_{q=3/4} = C_{\text{nc}}|_{q=3/4} = 0$ .

*Steganographic rate.* Real cDNA sequences exhibit a specific *codon count bias* or *codon bias*,<sup>10,31</sup> which is defined by the characteristic 19 empirical pmfs  $p(\mathbf{X}|x')$  associated to a sequence (note that two out of 21 such pmfs are deterministic). When computing  $R_c$  as above we are disregarding this codon bias in order to maximise the rate of the information-carrying sequence. For steganographic purposes, it is also possible to preserve the codon bias in the output by simply pegging  $p(\mathbf{U}|x')$  to the codon bias of  $\mathbf{X}$ . If we evaluate Gel'fand and Pinsker formula in this case we obtain the steganographic rate  $R_c^{\text{steg}}$ . Obviously,  $R_c^{\text{steg}} \leq R_c$ , with equality for any host in which the 19 empirical pmfs  $p(\mathbf{X}|x')$  are uniform. One particular case in which this is so is when  $\mathbf{X}$  is uniform. Similarly to the ncDNA case, we have assumed codon independence, and then  $R_c^{\text{steg}}$  will be an upper bound with respect to more refined host models. Finally, note that if the host is isothermal<sup>7</sup> this is preserved with the steganographic approach. Enforcing isothermality on  $\mathbf{Y}$  when the host is not isothermal might not always be possible, and leads to a different kind of problem which is not studied here.

### 3.2.3 Capacity Bounds with Indel Mutations

In this section we will consider a channel that includes both substitutions and indels. The capacity of indel channels is poorly understood in communications: no closed-form expressions are known even in the simplest scenarios. However the capacity of an indel mutation channel can always be seen as trivially upper bounded by the erasure channel, since, recalling the discussion on clairvoyant resynchronisation in Section 3.1, the performance of any transmission scheme over an indel channel cannot be better than its performance over an erasure channel with erasure probability equal to the deletion probability of the first channel. The real issue which this approach is to determine how tight the erasure channel upper bound is, or equivalently, how realistically the clairvoyant resynchronisation scheme described in Section 3.1 can be mimicked in practice.

In communications, indels have to be dealt with by means of pilot signals such as watermark codes<sup>32</sup> (no relationship with watermarking), which in general decrease the achievable rate and hence the tightness of a bound based on the erasure channel. This also applies to ncDNA data embedding. Nevertheless, in the context of cDNA data embedding resynchronisation can be essentially achieved without overly decreasing the embedding rate. This is because the amino acid sequence—which is preserved in cDNA after embedding—constitutes an implicit synchronisation signal. Relying on this information, it can be shown that it is easy to resynchronise a frame-shifted information-carrying cDNA sequence codon-wise by means of a modified version of the Needleman-Wunsch algorithm,<sup>33</sup> which is used in bioinformatics for globally aligning two sequences.

**Noncoding DNA.** It is easy to show that the capacity of an  $M$ -ary erasure channel is  $C_e^{(M)} \triangleq (1 - \delta) \log M$  bits/symbol, which is achieved with a uniform input. The *symmetric erasure channel* with parameters  $q$  and  $\delta$ , which we will use to deal simultaneously with both substitutions and indels, can be seen as the cascade of a symmetric channel with an erasure channel. For the symmetric channel in this cascade we have that the crossover probability is  $q' = q/(1 - \delta)$ . This decomposition makes it clear that the capacity per symbol bit of the cascade is the product of the capacities per symbol bit of the cascaded channels, that is,  $\frac{C_{\text{se}}^{(M)}}{\log M} = \frac{C_s^{(M)}}{\log M} \cdot \frac{C_e^{(M)}}{\log M}$  bits/symbol bit. This is because a uniform input to the symmetric channel, which maximises its achievable rate, leads to a uniform output, which is in turn the input to the erasure channel and also capacity-achieving. Therefore,

$$C_{\text{se}}^{(M)} = \left( \log M - H\left(\frac{q}{1 - \delta}\right) - \frac{q}{1 - \delta} \log(M - 1) \right) \cdot (1 - \delta) \text{ bits/symbol.} \quad (7)$$

Setting  $M = 4$  we obtain the upper bound  $\bar{C}_{\text{nc}} = C_{\text{se}}^{(4)}$  bits/base for ncDNA data embedding with substitutions and indel mutations, which we have discussed that can be loose.

**Coding DNA.** Following the same arguments as in Section 3.2.2 we can see that for uniform  $\mathbf{X}$  the upper bound for the channel with substitutions and indels is  $\bar{R}_c^{\text{unif}} = 3\bar{C}_{\text{nc}} - H(X')$  bits/codon. The bound should be

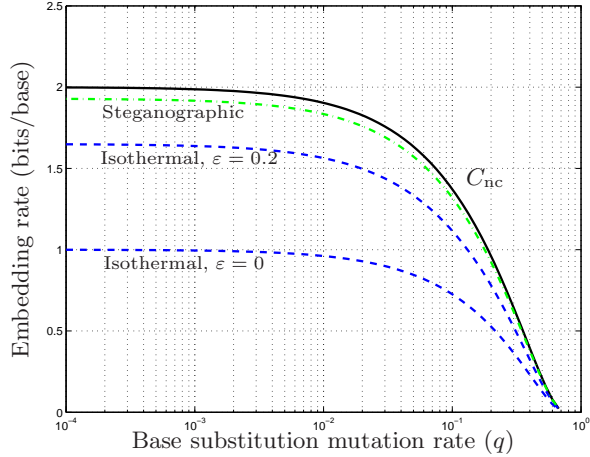


Figure 3. Achievable rates for ncDNA ( $R_{nc}$ ).

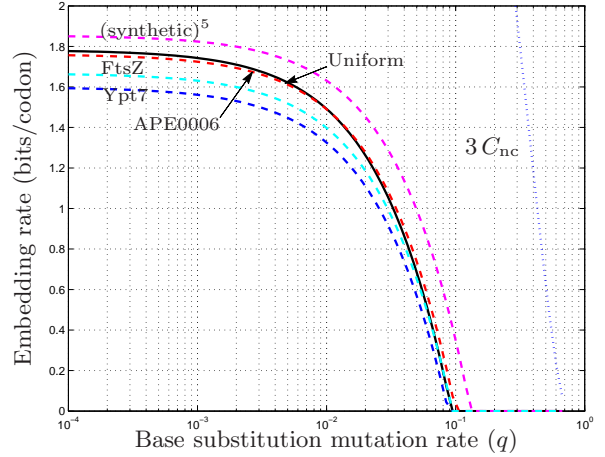


Figure 4. Achievable rates for cDNA ( $R_c$ ).

computed numerically for arbitrary pmfs. An interesting consequence of this bound, which is found by setting  $q = 0$ , is that no practical method is possible with uniform host if  $6(1 - \delta) - H(X') \leq 0$ , that is, if  $\delta \geq 0.2970$ .

#### 4. COMPARISONS WITH PRACTICAL METHODS

Figure 3 shows rates for ncDNA with base substitution mutations. In order to put these plots in perspective, see for instance that the Huffman and comma codes proposed by Smith *et al.*<sup>7</sup> for ncDNA give embedding rates of 1.9428 and 1.0537 bits/base, respectively. Huffman codes are, as expected, close to  $P_{nc}$ , but offer no protection against errors (for  $q > 0$ ). The comma codes used in<sup>7</sup> are isothermal with  $\varepsilon = 1$ ; while their rate is much lower than  $R_{nc}^{iso}(1) = C_{nc}$ , they also provide some resilience to indels. Lastly the steganographic rate for an ncDNA segment from *L. Lactis* is also shown (GenBank accession number M87483, 515–870).

| Author                                 | Rate (bits/codon) | Gene(s)     | Organism             | Environment      |
|--|-------------------|-------------|----------------------|------------------|
| Shimanovsky <i>et al.</i> <sup>5</sup> | 1.6667            | (synthetic) | —                    | <i>in silico</i> |
| Arita and Ohashi <sup>9</sup>          | 0.2778            | FtsZ        | <i>B. Subtilis</i>   | <i>in vivo</i>   |
| Modegi <sup>10</sup>                   | 0.1962            | APE0005,6   | <i>A. Pernix K1</i>  | <i>in silico</i> |
| Heider and Barnekow <sup>12,20</sup>   | 0.6408            | Ypt7, Vam7  | <i>S. Cerevisiae</i> | <i>in vivo</i>   |

Table 2. Embedding rates implemented in some practical cDNA methods

Figure 4 shows the achievable rates with cDNA using a uniformly distributed host and the genes employed in the practical cDNA methods listed in Table 2<sup>‡</sup>. We observe that null achievable rates may exist in this case for  $q < \frac{3}{4}$ . The threshold where  $R_c = 0$  is dependent on  $p(\mathbf{X})$ . The method by Shimanovsky *et al.*<sup>5</sup> will asymptotically achieve  $\bar{P}_c$  on average (i.e.  $R_c$  with  $q = 0$ ), but since it is based on arithmetic coding it suffers from error propagation for  $q > 0$ . The method by Modegi<sup>10</sup> has a very low embedding rate when compared to the achievable rate in Figure 4; although this could be due to its reversibility properties, these rely on external data. Heider and Barnekow<sup>12</sup> consider the very low  $q = 10^{-10}$  and  $q = 10^{-7}$ , but their DNA-Crypt method implements a much lower embedding rate than the one in theory achievable. With respect to the steganographic rates in Figure 5, note that they are a lot less clustered around the uniform rate than in Figure 4, since achieving uniformity of  $p(\mathbf{U}|x')$  is the maximising strategy used to obtain  $R_c$ .

Figure 6 shows the upper bounds obtained in Section 3.2.3. We will compare them to the results by Yachie *et al.*,<sup>11</sup> who are the only to give performance figures (obtained *in silico*) for deletion and substitution errors. Their ncDNA method implements *in vivo* the rates  $\frac{6}{n}$  bits/codon for  $n = 1, 2, 3, 4$  (with repetition coding), using *B. Subtilis* as the host. Errorless decoding is not achievable for this method with  $n = 1$ , but it is with  $n = 4$  for  $q \leq 0.1$  and  $\delta \leq 0.1$ . The corresponding rate, 1.5 bits/codon, is much lower than the respective upper bound in Figure 6, although we have discussed that this can be loose.

<sup>‡</sup>Data obtained from GenBank; accession numbers: NC\_001145 (Ypt7), NC\_000854 (APE0006), NC\_000964 (FtsZ).

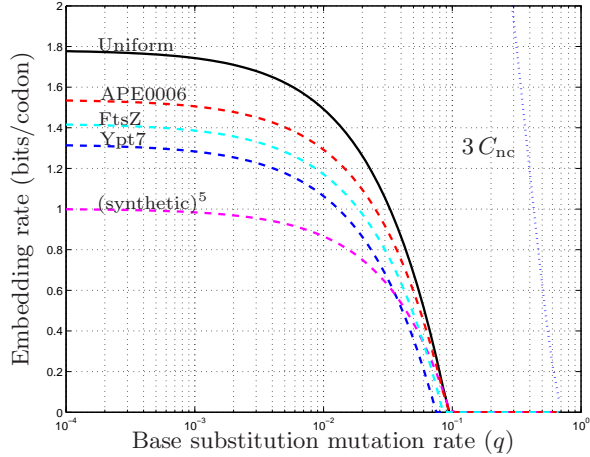


Figure 5. Steganographic rates for cDNA ( $R_c^{\text{steg}}$ ).

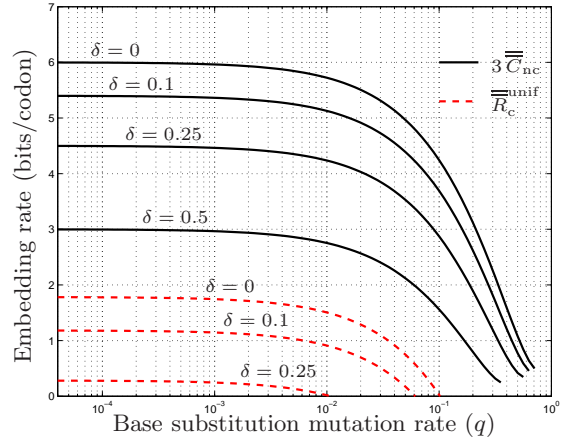


Figure 6. Upper bounds with substitutions and indels.

## APPENDIX A. ACHIEVABLE RATE FOR ncDNA WITH ISOTHERMALITY

As in a symmetric channel  $H(Z^B|Y^B)$  is independent of the input distribution, the maximum achievable rate is obtained by finding the pmf  $p(Y^B)$  that maximises  $H(Z^B)$  subject to the isothermal constraint. To this end we note first that  $p(Z^B = k) = (1 - q)p(Y^B = k) + \frac{q}{3}(1 - p(Y^B = k))$ . Using this expression we can maximise the entropy of  $Z^B$  with respect to  $p(Y^B)$ , subject to this solution being a pmf for which (4) holds. To this purpose we just build the Lagrangian functional for this problem, which is

$$\Phi \triangleq - \sum_{k=0}^3 \left( \left(1 - \frac{4q}{3}\right) \alpha_k + \frac{q}{3} \right) \log \left( \left(1 - \frac{4q}{3}\right) \alpha_k + \frac{q}{3} \right) + \lambda_1 \left( \sum_{k=0}^3 \alpha_k - 1 \right) + \lambda_2 (\alpha_0 + \alpha_2 - \varepsilon(\alpha_1 + \alpha_3)), \quad (8)$$

where  $\lambda_1$  and  $\lambda_2$  are Lagrange multipliers and  $\alpha_k \triangleq p(Y^B = k)$ . Differentiating (8) with respect to  $\alpha_k$  for  $k = 0, 1, 2, 3$  and equating to zero these expressions, it is readily seen that for the optimum it must hold that  $\alpha_0 = \alpha_2$  and  $\alpha_1 = \alpha_3$ . The maximising  $p(Y^B)$  in (5) can be obtained next by solving the constraint equations  $2(\alpha_0 + \alpha_1) = 1$  and  $\alpha_0 - \varepsilon \alpha_1 = 0$ .

## ACKNOWLEDGMENTS

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

## REFERENCES

- [1] Clelland, C. T., Risca, V., and Bancroft, C., “Hiding messages in DNA microdots,” *Nature* **399**, 533–534 (June 1999).
- [2] Gehani, A., LaBean, T., and Reif, J., “DNA-based cryptography,” in [5th DIMACS Workshop on DNA Based Computers], (June 1999).
- [3] Leier, A., Richter, C., Banzhaf, W., and Rauhe, H., “Cryptography with DNA binary strands,” *Bio Systems* **57**, 13–22 (June 2000).
- [4] Cox, J. P., “Long-term data storage in DNA,” *Trends in Biotechnology* **19**, 247–250 (July 2001).
- [5] Shimanovsky, B., Feng, J., and Potkonjak, M., “Hiding data in DNA,” in [Procs. of the 5th Intl. Workshop in Information Hiding], 373–386 (October 2002).
- [6] Wong, P. C., Wong, K., and Foote, H., “Organic data memory using the DNA approach,” *Comms. of the ACM* **46**, 95–98 (January 2003).
- [7] Smith, G. C., Fiddes, C. C., Hawkins, J. P., and Cox, J. P., “Some possible codes for encrypting data in DNA,” *Biotech. Lett.* **25**, 1125–1130 (July 2003).

- [8] Arita, M., "Comma-free design for DNA words," *Communications of the ACM* **47**, 99–100 (May 2004).
- [9] Arita, M. and Ohashi, Y., "Secret signatures inside genomic DNA," *Biotechnol. Prog.* **20**, 1605–1607 (September-October 2004).
- [10] Modegi, T., "Watermark embedding techniques for DNA sequences using codon usage bias features," in [16th Intl. Conf. on Genome Informatics], (December 2005).
- [11] Yachie, N., Sekiyama, K., Sugahara, J., Ohashi, Y., and Tomita, M., "Alignment-based approach for durable data storage into living organisms," *Biotechnol. Prog.* **23**, 501–505 (April 2007).
- [12] Heider, D. and Barnekow, A., "DNA-based watermarks using the DNA-Crypt algorithm," *BMC Bioinformatics* **8** (February 2007).
- [13] Chang, C.-C., Lu, T.-C., Chang, Y.-F., and Lee, C.-T., "Reversible data hiding schemes for deoxyribonucleic acid (DNA) medium," *International Journal of Innovative Computing, Information and Control* **3**, 1145–1160 (October 2007).
- [14] Heider, D., Pyka, M., and Barnekow, A., "DNA watermarks in non-coding regulatory sequences," *BMC Research Notes* **2** (July 2009).
- [15] Shannon, C. E., "A mathematical theory of communication," *Bell System Technical Journal* **27**, 379–423 and 623–656 (July and October 1948).
- [16] Deonier, R. C., Tavaré, S., and Waterman, M. S., [*Computational Genome Analysis: An Introduction*], Springer (2005).
- [17] Cachin, C., "An information-theoretic model for steganography," in [*Procs. of the Second International Workshop on Information Hiding*], *Lecture Notes in Computer Science* **1525**, 306–318, Springer-Verlag (April 1998). Revised version in *Information and Computation*, 2004.
- [18] Heider, D., Kessler, D., and Barnekow, A., "Watermarking sexually reproducing diploid organisms," *Bioinformatics* **24**, 1961–1962 (July 2008).
- [19] Gibbs, W. W., "The unseen genome: Gems among the junk," *Scientific American*, 53 (November 2003).
- [20] Heider, D. and Barnekow, A., "DNA watermarks: A proof of concept," *BMC Molecular Biology* **9** (April 2008).
- [21] Eisenberg, R. S., "Structure and function in gene patenting," *Nature Genetics* **15**(2), 125–130 (1997).
- [22] Bancroft, C. and Clelland, C. T., "DNA-based steganography." U.S. Patent 6,312,911 (June 2001).
- [23] Patel, P., "Advance in nanopore gene sequencing," *Spectrum, IEEE* **46**, 14 (June 2009).
- [24] Kunkel, T. A., "DNA replication fidelity," *J. Biol. Chem.* **279**, 16895–16898 (April 2004).
- [25] Fu, Y., "Estimating mutation rate and generation time from longitudinal samples of DNA sequences," *Mol. Biol. and Evolution* **18**(4), 620–626 (2001).
- [26] Pradhan, S. S., Chou, J., and Ramchandran, K., "Duality between source coding and channel coding and its extension to the side information case," *IEEE Trans. on Inf. Theory* **49**, 1181–1203 (May 2003).
- [27] Barron, R. J., Chen, B., and Wornell, G. W., "The duality between information embedding and source coding with side information and some applications," *IEEE Trans. on Inf. Theory* **49**, 1159–1180 (May 2003).
- [28] Costa, M. H., "Writing on dirty paper," *IEEE Trans. on Information Theory* **29**, 439–441 (May 1983).
- [29] Ash, R. B., [*Information Theory*], Dover, New York (1965).
- [30] Gel'fand, S. I. and Pinsker, M. S., "Coding for channel with random parameters," *Problems of Control and Information Theory* **9**(1), 19–31 (1980).
- [31] Arita, M., "Writing information into DNA," in [*Aspects of Molecular Computing*], *Lecture Notes in Computer Science* **2950**, 211–222, Springer (February 2004).
- [32] Davey, M. and Mackay, D., "Reliable communication over channels with insertions, deletions, and substitutions," *IEEE Transactions on Information Theory* **47**, 687–698 (Feb 2001).
- [33] Needleman, S. B. and Wunsch, C. D., "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology* **48**, 443–453 (March 1970).