Capacity of DNA Data Embedding Under Substitution Mutations

Félix Balado, *Member, IEEE*

***Abstract*—A number of methods have been proposed over the last decade for encoding information using deoxyribonucleic acid (DNA), giving rise to the emerging area of DNA data embedding. Since a DNA sequence is conceptually equivalent to a sequence of quaternary symbols (bases), DNA data embedding (diversely called DNA watermarking or DNA steganography) can be seen as a digital communications problem where channel errors are analogous to mutations of DNA bases. Depending on the use of coding or noncoding DNA host sequences, which respectively denote DNA segments that can or cannot be translated into proteins, DNA data embedding is essentially a problem of communications with or without side information at the encoder. In this paper the Shannon capacity of DNA data embedding is obtained for the case in which DNA sequences are subject to substitution mutations modelled using the Kimura model from molecular evolution studies. Inferences are also drawn with respect to the biological implications of some of the results presented.**

***Index Terms*—Channel capacity, DNA, Genetic Communication, Watermarking, Evolution (biology)**

I. Introduction

THE last ten years have seen the proposal of numerous practical methods [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] for encoding nongenetic information using DNA molecules as a medium, both *in vitro* and *in vivo*. A high profile use of these techniques took place recently, when the J. Craig Venter Institute (JCVI) produced the first artificial bacteria which included a "watermark" containing authorship information [11], [12]. All of these proposals rely on the fact that DNA molecules —which encode genetic information in all living organisms, except for some viruses— are conceptually equivalent to sequences of quaternary symbols, which can be manipulated to store and retrieve arbitrary data by using the molecular biology analogues of "writing" (producing recombinant DNA) and "reading" (DNA sequencing). Therefore DNA data embedding is in essence an instance of digital communications, with the particularity that channel errors are analogous to mutations of DNA components. The two broad fields of application of DNA data embedding techniques are: 1) the use of DNA strands as self-replicating nano-memories with the ability to store huge amounts of data in an ultra-compact and energy-efficient way [2], [4], [7], [8]; and 2) security and tracking applications made possible by embedding nongenetic information in DNA: for instance, DNA watermarking [9], [11], [12], DNA steganography [1], [5] and DNA tagging [13]. The actual information embedded in security applications may include ownership, authorship and integrity data.

DNA data embedding techniques can be divided into two types, depending on whether the embedded information is hosted by coding or by noncoding parts of a genome —the ensemble of all the DNA of an organism. The coding parts of a genome are the genes that it contains, which are the basic units of genetic inheritance. The remaining parts of a genome are generically called noncoding DNA. Noncoding DNA is normally the host of choice in storage applications of DNA data embedding: it will be seen that more information can usually be embedded in noncoding sections, with less constraints and complications. However it is also important to consider genes (that is, coding DNA) as information hosts for several reasons. Firstly, embedding information within genes may be the only approach to mark or track them in a truly individualised way, which is of relevance in some security and tracking applications. One example is the potential use of DNA data embedding in gene patenting [5], [14], a procedure by which intellectual property of artificially engineered genes may be asserted. Patent information must reside within a gene —and not anywhere else— if that information is to travel with the gene when copied by an unauthorised party. It may also be desirable to include laboratory identification or integrity information in genes employed in gene therapies (human gene transfer), in which case the information must lie within genes [3]. Furthermore, Arita and Ohashi point out that it may be biologically safer to embed information within coding DNA rather than within noncoding DNA [5]. This may seem paradoxical, because noncoding DNA was once derogatorily labelled as "junk DNA" due to its supposed lack of biological purpose. However it is now increasingly recognised and understood that specific noncoding DNA sections can be involved in regulatory functions of gene activity. Therefore risks may lie ahead if information is embedded in noncoding regions which may hold as yet unknown functions.

DNA data embedding methods which target the genomes of self-replicating living organisms —such as the ones proposed and implemented by Wong et al. [4], Arita and Ohashi [5], Yachie et al. [7], Heider and Barnekow [14], and more recently Gibson et al. [12]— all face two fundamental questions: 1) how much information can be embedded in a given DNA segment of an organism?; and 2) how many generations of that organism can this information potentially survive intact? The

Manuscript received January 16, 2011; revised February 26, 2012, and July 7, 2012. This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 09/RFP/CMS2212. Preliminary parts of this work were presented at the SPIE Media Forensics and Security XII conference (January 2010), the IEEE ICASSP conference (March 2010), and the IET Irish Signals and Systems Conference (June 2012).

F. Balado is with the UCD School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland (e-mail: felix@ucd.ie).

methods just mentioned embed information at rates ranging from 2 bits per DNA base [4], [12] (using noncoding DNA) to 0.2136 bits per base [14] and 0.0926 bits per base [5] (using coding DNA), but will that information be retrievable without errors when mutations accumulate after a number of generations of the host organism? Could those rates possibly be higher for a given mutation rate? The answers are supplied by the upper limit on the amount of information that can be embedded within a given DNA molecule under a given mutation rate, and asymptotically decoded with no errors. This is the Shannon capacity [15] of DNA data embedding, which is the ultimate benchmark against which all actual DNA data embedding methods used in bioengineering applications must be measured.

In this paper the capacity of DNA data embedding under substitution mutations is determined. Substitution mutations are those that randomly switch bases in a DNA sequence, as may occur in the replication of an organism. These are modelled by means of a symmetric memoryless channel equivalent to the probabilistic model first proposed by Kimura [16] to study molecular evolution. Section II describes the framework, assumptions and model used. The capacity analysis, addressed in Section III, is straightforward when no side information is required by the encoder during the embedding process. This is the case for techniques that use noncoding DNA to host information, such as [4], [7] or [12]. On the other hand, the capacity analysis with side information at the encoder corresponds to techniques which use coding DNA, such as [3], [5], [9] or [14], whose encoder must strictly preserve gene expression. This analysis forms the main constituent of this paper. Where pertinent, biological implications of these information theoretical results are also discussed. In particular, the non side-informed scenario happens to be closely connected to previous studies by May [17], Battail [18], and other authors who have applied information theoretical concepts to molecular biology from an information transmission perspective.

II. Preliminary Concepts and Assumptions

Chemically, DNA is formed by two backbone strands helicoidally twisted around each other, and mutually attached by means of two nitrogenous *base* sequences. The four possible bases are the molecules adenine, cytosine, thymine and guanine, abbreviated to 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. The interpretation of DNA as a one-dimensional discrete digital signal is straightforward: any one of the two strands constitutes a digital sequence formed by symbols from a quaternary alphabet, which completely determines the DNA molecule.

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. In molecular biology this is called the 5'–3' direction, which refers to certain chemical feature points in a DNA backbone strand. The two strands in a DNA molecule are read in opposite directions; because of this, and because 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 chemical compounds called *amino acids* via transcription to the intermediary ribonucleic acid (RNA) molecule. RNA is similar to DNA but single-stranded and with uracil (abbreviated U) replacing thymine. 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 almost universal *genetic code*. The standard version[1] of the genetic code is shown in Table I and explained in more detail later.

The genome of an organism is the ensemble of all its DNA. Segments of a genome that can be translated into proteins by the genetic machinery through the process described above are called *protein-coding* DNA (pcDNA), or just coding DNA, whereas those segments that never get translated are called *noncoding* DNA (ncDNA). A *gene* is a pcDNA segment, or group of segments, which encodes one single protein, and which is flanked by certain start and stop codons (see Table I) plus other regulatory markers. 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.

The main assumptions made in this work are:

- *ncDNA can be freely appended or overwritten.* As briefly mentioned in the introduction this assumption is not always strictly true, because certain ncDNA regions act as promoters or enhancers for gene expression, or are transcribed into regulatory RNA (which is not translated into proteins). After conducting *in vivo* tests, Heider et al. have cautioned against embedding information in promoter ncDNA regions [10]. However the working hypothesis is valid in many other suitably chosen ncDNA regions, as proved by Wong et al. [4], Yachie et al. [7], [8] and Gibson et al. [12], who have successfully embedded information in the ncDNA of living organisms.
- *pcDNA can be freely modified as long as the genetic code is followed.* This is the classic standard assumption supporting the validity of the genetic code. This assumption was used *in silico* by Shimanovsky et al. [3] and *in vivo* by Arita and Ohashi [5] and Heider and Barnekow [14]. However, living organisms also exhibit preferred codon usage biases (also called codon usage statistics), which are characteristic distributions of the codons associated with a given amino acid. When pcDNA is modified, these codon biases can emerge completely changed even if the genetic code is strictly observed. This change might be detrimental for gene expression, for instance by extending gene translation times among other effects [19]. For this reason we will also analyse the embedding capacity

[1] http://www.ncbi.nlm.nih.gov/taxonomy (genetic codes)

x'	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	*Stp*
	GCA	AGA	AAC	GAC	TGC	CAA	GAA	GGA	CAC	ATA	CTA	AAA	ATG	TTC	CCA	AGC	ACA	TGG	TAC	GTA	TAA
	GCC	AGG	AAT	GAT	TGT	CAG	GAG	GGC	CAT	ATC	CTC	AAG		TTT	CCC	AGT	ACC		TAT	GTC	TAG
	GCT	CGA						GGT		ATT	CTT				CCT	TCA	ACT			GTT	TGA
$\mathcal{S}_{x'}$	GCG	CGC						GGG			CTG				CCG	TCC	ACG			GTG	
		CGT									TTA					TCT					
		CGG									TTG					TCG					
$\lvert\mathcal{S}_{x'}\rvert$	4	6	2	2	2	2	2	4	2	3	6	2	1	2	4	6	4	1	2	4	3

TABLE I
EQUIVALENCES BETWEEN AMINO ACIDS AND CODONS (STANDARD GENETIC CODE). START CODONS, WHICH DOUBLE AS REGULAR CODONS, ARE UNDERLINED.

of pcDNA with codon bias preservation constraints. To conclude, it is important to mention that only nonoverlapping genes will be considered, either on the same or on opposite strands. In any case overlapping genes are rare occurrences, except in very compact genomes.

Notation. Calligraphic letters ($\mathcal{X}$) denote sets; $|\mathcal{X}|$ is the cardinality of $\mathcal{X}$. Boldface letters ($\mathbf{x}$) denote row vectors, and $\mathbf{1}$ is an all-ones vector. If a Roman letter is used both in uppercase (X) and in lowercase (x), the two forms denote a random variable and a realisation of it, respectively. $p(X = x)$ is the probability mass function (pmf) or distribution of X; we will simply write $p(x)$ when the variable is clear from the context. For brevity, X may also denote the distribution of X if there is no ambiguity. $E[X]$ is the mathematical expectation of X, and $H(X)$ its entropy. Also, $h(q)$ is the entropy of a Bernoulli(q) random variable. $I(X;Y)$ is the mutual information between X and Y. Logarithms are base 2, unless explicitly stated otherwise. The Hamming distance between vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted by $d_H(\mathbf{x}, \mathbf{y})$.

A ncDNA sequence will be denoted by a vector $\underline{\mathbf{x}} = [x_1, x_2, \cdots, x_n]$, whose elements are consecutive bases from a base sequence. That is, $x_i \in \mathcal{X} \triangleq \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$, the 4-ary set of possible bases. A pcDNA sequence will be denoted by a vector of vectors $\overline{\mathbf{x}} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$, whose elements are consecutive codons[2] from one of the two antiparallel base sequences when the right reading frame among the three possibilities is chosen. Therefore, $\mathbf{x}_i \in \mathcal{X}^3$. The amino acid into which a codon $\mathbf{x}_i$ uniquely translates is denoted by $x'_i \triangleq \alpha(\mathbf{x}_i) \in \mathcal{X}'$, where $\mathcal{X}' \triangleq \{$Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, Val, *Stp*$\}$ is the set of all amino acids defined using the standard three-letter abbreviations of their names. $\mathbf{x}' = \alpha(\overline{\mathbf{x}}) = [x'_1, x'_2, \cdots, x'_n]$ denotes the unique amino acid sequence established by $\overline{\mathbf{x}}$, usually called the *primary structure* of the protein encoded by $\overline{\mathbf{x}}$.

The subset of synonymous codons associated with amino acid $x' \in \mathcal{X}'$, namely

$$\mathcal{S}_{x'} \triangleq \{\mathbf{x} \in \mathcal{X}^3 | \alpha(\mathbf{x}) = x'\},$$

is established by the genetic code in Table I. The ensemble of stop codons is collected under the special symbol *Stp*, and thus loosely classed as an "amino acid" for notational convenience. However *Stp* just indicates the end of a gene, and does not actually map to any amino acid. Three codons, underlined in Table I, also double as protein start translation commands when they appear at the start of a gene —the two corresponding to Leu only in eukaryotic organisms. The number $|\mathcal{S}_{x'}|$ of synonymous codons mapping to x' is called the *multiplicity* of x'. Due to the uniqueness of the mapping from codons to amino acids, observe that $\mathcal{S}_{x'} \cap \mathcal{S}_{y'} = \emptyset$ for $x' \neq y' \in \mathcal{X}'$, and that $\sum_{x' \in \mathcal{X}'} |\mathcal{S}_{x'}| = |\mathcal{X}|^3 = 64$ since $\cup_{x' \in \mathcal{X}'} \mathcal{S}_{x'} = \mathcal{X}^3$. Notice that, for most amino acids x', the third base (*wobble base*) tends to exhibit more variability than the other two bases in synonymous codons from $\mathcal{S}_{x'}$. Finally, a toy example of a pcDNA sequence may be $\overline{\mathbf{x}} = [[\mathrm{T}, \mathrm{A}, \mathrm{T}], [\mathrm{T}, \mathrm{G}, \mathrm{C}]]$, which would encode the amino acid sequence $\mathbf{x}' = \alpha(\overline{\mathbf{x}}) = [\mathrm{Tyr}, \mathrm{Cys}]$. The corresponding base sequence would be $\underline{\mathbf{x}} = [\mathrm{T}, \mathrm{A}, \mathrm{T}, \mathrm{T}, \mathrm{G}, \mathrm{C}]$.

A. Mutation channel model

As 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 may be termed "mutation channel" in this context. Herein we will only consider base substitution mutations (also called point mutations), which randomly switch letters within the DNA alphabet. From a communications viewpoint, apart from base substitutions, the most relevant types of mutations are random base insertions and deletions, which cause a synchronisation problem at the decoder. Insertions and deletions are jointly called *indels*, due to the ambiguity that sometimes exists about whether certain differences between DNA sequences are actually due to either type of mutation. The analysis will not consider indel mutations for the following reasons. On the one hand these mutations tend to have deleterious effects when they occur in pcDNA sections of an organism —especially if the number of such mutations is not a multiple of three, as this can derail protein translation— which arguably makes the survival of the organism less likely. Thus the arrival at the decoder of the mutated information-carrying host will also be less likely. Consequently a base substitutions only analysis is deemed a realistic approach in computing the capacity of pcDNA embedding. This is not the case when indels occur in ncDNA sections; however, the fact is that the exact Shannon capacity of a channel with insertions and deletions is currently unknown even in the simplest case of a binary channel [20]. For this reason, dealing with indel mutations is beyond the scope of this paper. In any case, the substitution mutations analysis still yields an upper bound to

[2]For simplicity, and without loss of generality, it is considered that consecutive codons are those corresponding to consecutive amino acids in the primary structure of a protein. In eukaryotic organisms (those whose cells contain a membrane-bound nucleus) the DNA sequence corresponding to a gene is in general divided into *introns* (noncoding sections, excised during transcription) and *exons* (protein-coding sections).

the capacity of ncDNA embedding, with respect to a more general mutation model.

It will be assumed in our analysis that mutations are mutually independent, which is a worst-case scenario in terms of capacity. As a result it is considered that the channel is memoryless and thus accepts a single-letter characterisation. Consequently we will drop vector element subindices whenever this is unambiguous and notationally convenient. The base substitution channel will be modelled by means of the two-parameter Kimura model of nucleotide substitution [16], which has been extensively used in molecular evolution studies. This consists of a 4×4 transition probability matrix $\Pi \triangleq [p(Z = z|Y = y)]$, where $z, y \in \mathcal{X}$, which presents the following structure:

$$\Pi = \begin{array}{c} \begin{matrix} \text{A} & \text{C} & \text{T} & \text{G} \end{matrix} \\ \begin{bmatrix} 1-q & \frac{\gamma}{3}q & \frac{\gamma}{3}q & (1-\frac{2\gamma}{3})q \\ \frac{\gamma}{3}q & 1-q & (1-\frac{2\gamma}{3})q & \frac{\gamma}{3}q \\ \frac{\gamma}{3}q & (1-\frac{2\gamma}{3})q & 1-q & \frac{\gamma}{3}q \\ (1-\frac{2\gamma}{3})q & \frac{\gamma}{3}q & \frac{\gamma}{3}q & 1-q \end{bmatrix} \end{array} \begin{matrix} \text{A} \\ \text{C} \\ \text{T} \\ \text{G} \end{matrix} \quad (1)$$

From this definition, the probability of base substitution mutation, or base substitution mutation rate, is

$$q = p(Z \neq y|Y = y) = \sum_{z \neq y} p(Z = z|Y = y), \quad (2)$$

for any $y \in \mathcal{X}$, whereas it must hold that $0 \leq \gamma \leq 3/2$ so that row probabilities add up to one. The particular structure of Π aims at reflecting the fact that DNA bases belong to one of two categories according to their chemical structure: purines, $\mathcal{R} \triangleq \{\text{A}, \text{G}\}$, or pyrimidines, $\mathcal{Y} \triangleq \{\text{C}, \text{T}\}$. Put simply, purines and pyrimidines are both cyclic compounds. Their main difference is that pyrimidines are single-ringed, whereas purines are double-ringed and one of their rings is in fact a pyrimidine. There are two types of base substitutions corresponding to these categories, which in biological nomenclature are:

- Base *transitions*: those that preserve the category which the base belongs to. In this case the model establishes that $p(Z = z|Y = y) = (1 - 2\gamma/3)q$ for $z \neq y$ when either both $z, y \in \mathcal{R}$ or both $z, y \in \mathcal{Y}$.
- Base *transversions*: those that switch the base category. In this case the model establishes that $p(Z = z|Y = y) = (\gamma/3)q$ for $z \neq y$ when $z \in \mathcal{Y}$ and $z \in \mathcal{R}$, or vice versa.

The channel model (1) can incorporate any given transition/transversion ratio ε by setting $\gamma = 3/(2(\varepsilon+1))$. Estimates of ε given in [21] for the DNA of different organisms range between 0.89 and 18.67, corresponding to γ between 0.07 and 0.79. This range of ε reflects the fact that base transitions are generally much more probable than base transversions due to the chemical structure similarity among compounds in the same category, that is, $\varepsilon > 1/2$ virtually always in every organism, and therefore $\gamma < 1$. However, many mutation estimation studies focus only on the determination of q (see for instance [22], [23]). If only the parameter q is known, one may assume the simplification $\gamma = 1$. This is known as the Jukes-Cantor model in molecular evolution studies. In this model all off-diagonal entries of Π are equal, that is, $p(Z = z|Y = y) = q/3$ for all $z \neq y$. Several observations are made for the capacity analysis with the Jukes-Cantor model throughout this paper.

The mutation model chosen implies a symmetric channel, since all rows (or columns) of Π contain the same four probabilities. Among all memoryless models used in molecular evolution, the Kimura model is the one with the highest number of parameters while still yielding a symmetric channel. It is well known that symmetric channels have a simple capacity analysis, which is exploited where possible. The most general time-reversible substitution mutations model may have up to 9 independent parameters, which in general yields a nonsymmetric channel. However, according to Li [24] mutation models with many parameters are not necessarily accurate, due to the estimation issues involved.

It will be assumed that the matrix Π models the base substitution mutations undergone by the genome of an organism during one generation. A Markov chain $Y \to Z_{(1)} \to Z_{(2)} \to \cdots \to Z_{(m)}$ can be used to model m generations of an organism at a given site (position) in a genome, whose corresponding base at generation $m = 0$ is represented by Y. The site can lie anywhere in the genome of an organism with asexual reproduction, or in the non-recombinant sections of an organism with sexual reproduction. The relationship between Y and $Z_{(m)}$ is given by the transition probability matrix $\Pi^m = [p(Z_{(m)} = z|Y = y)]$. As $\Pi = \Pi^T$ one can write $\Pi^m = \mathrm{V}\,\mathrm{D}^m\,\mathrm{V}^T$, with the eigenvalues of Π arranged in a diagonal matrix $\mathrm{D} \triangleq \mathrm{diag}(1,\, \lambda,\, \mu,\, \mu)$, where

$$\lambda \triangleq 1 - \frac{4\gamma}{3}q \quad (3)$$

$$\mu \triangleq 1 - 2\left(1 - \frac{\gamma}{3}\right)q, \quad (4)$$

and V is a matrix whose columns are the normalised eigenvectors of Π associated with the corresponding eigenvalues in D, that is

$$\mathrm{V} = \frac{1}{2}\begin{bmatrix} 1 & 1 & -\sqrt{2} & 0 \\ 1 & -1 & 0 & -\sqrt{2} \\ 1 & -1 & 0 & \sqrt{2} \\ 1 & 1 & \sqrt{2} & 0 \end{bmatrix}. \quad (5)$$

From the diagonalisation of Π^m it follows that the elements of its diagonal take all the value $\frac{1}{4}(1 + 2\mu^m + \lambda^m)$, the elements of its skew diagonal take all the value $\frac{1}{4}(1 - 2\mu^m + \lambda^m)$, and the rest of its entries are $\frac{1}{4}(1 - \lambda^m)$. Therefore any row (or column) of this matrix contains the same probabilities, as Π^m is also the transition matrix of a symmetric channel. From the diagonal elements one can see that the accumulated base substitution mutation rate after m generations is given by

$$q^{(m)} = p(Z_{(m)} \neq y|Y = y) = 1 - \frac{1}{4}\left(1 + 2\mu^m + \lambda^m\right). \quad (6)$$

It can be verified that $\lim_{m\to\infty} q^{(m)}|_{\gamma>0} = 3/4$, but $\lim_{m\to\infty} q^{(m)}|_{\gamma=0} = 1/2$. This is because $|\mu| < 1$ for any γ and $|\lambda| < 1$ when $\gamma > 0$, but $\lambda = 1$ when $\gamma = 0$. The behaviour of the particular case $\gamma = 0$ is connected to the fact that we must have both $q \in (0, 1]$ and $\gamma \in (0, 3/2]$ for the Markov chain to be aperiodic and irreducible, and thus possess a limiting stationary distribution. From the previous considerations, the limiting distribution —that is, the distribution of $Z_{(\infty)}$— is uniform, because $\lim_{m\to\infty} \Pi^m = \frac{1}{4}\mathbf{1}^T\mathbf{1}$. When $\gamma = 1$ and $q = 3/4$ then $\Pi = \frac{1}{4}\mathbf{1}^T\mathbf{1}$; hence $Z_{(m)}$ is always uniformly distributed for any m as in the limiting case.

Lastly, under the base substitution mutation model considered, codons undergo a mutation channel modelled by the 64×64 transition probability matrix

$$\mathbf{\Pi} = [p(\mathbf{Z} = \mathbf{z}|\mathbf{Y} = \mathbf{y})] = \Pi \otimes \Pi \otimes \Pi, \tag{7}$$

where $\otimes$ is the Kronecker product. This is because $p(\mathbf{Z} = \mathbf{z}|\mathbf{Y} = \mathbf{y}) = \prod_{i=1}^{3} p(Z_i = z_i|Y_i = y_i)$ according to our memoryless channel assumption. Trivially this channel is also symmetric. When m mutation stages are considered, Π^m replaces Π in (7), since $\mathbf{\Pi}^m = [p(\mathbf{Z}_{(m)} = \mathbf{z}|\mathbf{Y} = \mathbf{y})] = (\Pi \otimes \Pi \otimes \Pi)^m = \Pi^m \otimes \Pi^m \otimes \Pi^m$ [25].

III. Capacity Analysis

A. Noncoding DNA

In this section we consider the embedding capacity of noncoding DNA, which also establishes a basic upper bound to capacity for protein-coding DNA. As discussed in Section II, we are assuming that the information embedding operation can either overwrite or append a host ncDNA strand $\underline{\mathbf{x}}$, which amounts to freely choosing the input $\underline{\mathbf{y}}$ to the mutation channel. Therefore with ncDNA the embedding capacity is simply given by $C_{\text{nc}} \triangleq \max I(Z_{(m)}; Y)$ bits/base, where the maximisation is over all distributions of Y. This capacity is that of a symmetric channel for the mutation model considered. With this type of channel $H(Z_{(m)}|Y)$ is independent of the input, and uniformly distributed Y leads to uniformly distributed $Z_{(m)}$. Hence

$$C_{\text{nc}} = \log|\mathcal{X}| - H(Z_{(m)}|Y) \text{ bits/base}, \tag{8}$$

where $H(Z_{(m)}|Y)$ is the entropy of any row of Π^m, that is, $H(Z_{(m)}|Y) = -\sum_{z \in \mathcal{X}} p(Z_{(m)} = z|Y = y) \log p(Z_{(m)} = z|Y = y)$ for any $y \in \mathcal{X}$. C_{nc} is nonincreasing on m, since the information processing inequality implied by the Markov chain $Y \to Z_{(1)} \to Z_{(2)} \to \cdots \to Z_{(m)}$ implies that $I(Z_{(1)}; Y) \geq I(Z_{(2)}; Y) \geq \cdots \geq I(Z_{(m)}; Y)$. As long as the Markov chain is aperiodic and irreducible then $\lim_{m\to\infty} C_{\text{nc}} = 0$. The reason for this behaviour is that the limiting distribution is in this case independent of Y, which implies that $\lim_{m\to\infty} H(Z_{(m)}|Y) = H(Z_{(\infty)}) = \log|\mathcal{X}|$. It is interesting to note that, under aperiodicity and irreducibility of the Markov chain, this zero limiting capacity will also apply to models more involved than (1), such as those in which the channel matrix is parametrised by up to 9 independent values. Lastly, we also have $C_{\text{nc}}|_{\gamma=1,q=3/4} = 0$, since with these parameters $Z_{(m)}$ is always uniformly distributed.

As a function of γ the ncDNA capacity is bounded for every m and q as follows

$$C_{\text{nc}}|_{\gamma=1} \leq C_{\text{nc}} \leq C_{\text{nc}}|_{\gamma=0}. \tag{9}$$

Although it can be shown that these inequalities hold true in general, it is much simpler to prove them for the range of interest $\gamma \leq 1$ and $q \leq 1/2$. The latter condition implies that both $0 \leq \lambda \leq 1$ and $0 \leq \mu \leq 1$. For fixed m and q, the maximum (respectively, minimum) of C_{nc} over γ corresponds to the minimum (respectively, maximum) of the accumulated base mutation rate $q^{(m)}$. Differentiating (6) we obtain $\partial q^{(m)}/\partial\gamma = (mq/3)\left(\lambda^{m-1} - \mu^{m-1}\right)$. Therefore $q^{(m)}$ is monotonically increasing when $\gamma \leq 1$ (as this corresponds to $\lambda \geq \mu$), and then its maximum in that range occurs when $\gamma = 1$ and its minimum when $\gamma = 0$.

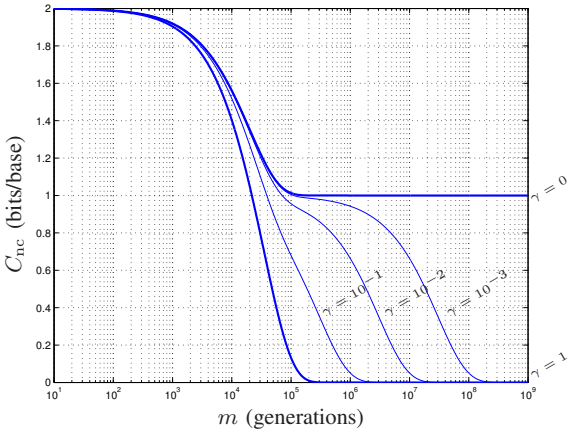

Fig. 1. Embedding capacity in ncDNA ($q = 10^{-5}$).

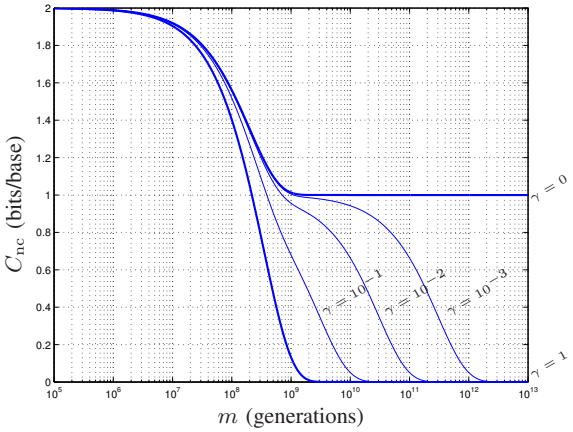

Fig. 2. Embedding capacity in ncDNA ($q = 10^{-9}$).

The upper bound can be written as

$$C_{\text{nc}}|_{\gamma=0} = 2 - h\left(\frac{1}{2} + \frac{1}{2}(1-2q)^m\right). \tag{10}$$

Notice that $\lim_{m\to\infty} C_{\text{nc}}|_{\gamma=0} = 1$, that is, the capacity limit is not zero when $\gamma = 0$ because then the Markov chain is reducible. $\gamma = 0$ cannot occur in reality since it would imply that transversion mutations are impossible. However it illustrates that the higher the transition/transversion ratio ε, the higher the capacity.

Figures 1 and 2 show C_{nc} for two different values of the base substitution mutation rate per replication, $q = 10^{-5}$ and $q = 10^{-9}$. These mutation rates have been arbitrarily chosen to illustrate the capacity values in representative cases of q corresponding to viruses [23] and microbes [22, Table 4], respectively. The results in the figures show the validity of the bounds (9) and the limiting behaviours discussed. From these figures one can also empirically see that, as a rule-of-thumb, capacity lies around $C_{\text{nc}} \approx 10^{-2}$ bits/base for $m \approx 6/(5\gamma q)$.

Finally, in Figure 3 we compare the ncDNA capacity computed with the Kimura model versus the same computation using an unconstrained estimate of Π^m given by Li [24,

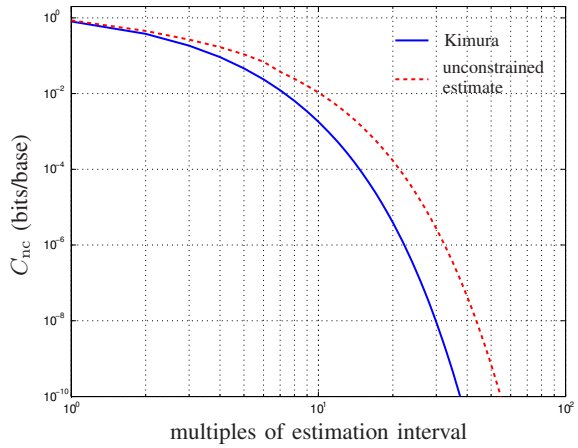

Fig. 3. Comparison of embedding capacity in ncDNA using the Kimura model versus an unconstrained estimate of the substitution matrix. $q = 0.2515$ and $\gamma = 0.7795$ are the averages for the estimate.

page 31] (for a mammalian ncDNA section and unspecified estimation interval m). The Blahut-Arimoto algorithm [26] has been used in this computation, since the channel modelled by the unconstrained estimate is nonsymmetric. We observe that the Kimura model suffices for approximating capacity, while simultaneously providing much more insight into the major factors influencing it when compared to an unconstrained model —in particular the influence of the transition-transversion parameter γ.

a) Biological interpretations: Expression (8) also gives the maximum mutual information between a DNA strand and its mutated version in nature. This has been termed the capacity of the genetic channel in studies applying information theory to molecular biology (for a critical review of the subject the reader is referred to [27]). Several authors have used the Jukes-Cantor model and the Kimura model with specific values of γ —apparently unaware of the prior use of these models in molecular evolution studies— in order to determine this capacity. The case $m = 1$ was numerically evaluated by May et al. [28], using values of q estimated from different organisms and the Jukes-Cantor and Kimura ($\gamma = 1/2$) models. Some authors have also considered the behaviour of capacity for $m > 1$. Gutfraind [29] discussed the basic effect of accumulated mutations on capacity (exponential decrease with m), but used a binary alphabet and the binary symmetric channel. Both Battail [18] and May [17] computed capacity under cascaded mutation stages using a quaternary alphabet and the Jukes-Cantor model. Battail obtained his results analytically —using a continuous-time approach rather than the discrete-time approach taken in this paper— and May obtained hers numerically. The results by Battail are essentially consistent with the ones presented here, but the results by May are not. However it would appear that this effect is due to the use of $m' = m \times g$ as the number of generations in [17], where $g \approx 3 \times 10^9$ base pairs is the length of the human genome.

In any case, none of the aforementioned approaches reflect the capacity increase afforded by a mutation model allowing $\gamma < 1$. It is possible that the trend towards higher capacity observed as $\gamma \to 0$ implies that evolution has favoured genetic building blocks which feature an asymmetric behaviour under mutations (that is, in the current genetic machinery pyrimidines versus purines instead of a hypothetical perfectly mutation-symmetric set of four bases for which $\gamma = 1$). If this assumption is correct, this information-theoretical induced "mutation-symmetry breaking" must have occurred early in evolutionary terms, since it is widely believed that the current genetic machinery evolved from a former "RNA world" [30] in which life would only have been based on the self-replicating and catalysing properties of RNA. In the RNA world there would not have been translation to proteins, and therefore no genetic code, and hence information was freely encoded using a 4-ary alphabet almost exactly like the one used in DNA. Note that uracil, which replaces thymine in RNA, is also a pyrimidine, that is, in the RNA world $\mathcal{Y} = \{\mathrm{U}, \mathrm{C}\}$. With these facts in mind, we may model the maximum transmissible information under mutations in the RNA world by relying on (8), and thus see that the mutation-symmetry breaking conjecture above applies to the evolution of RNA from hypothetical ancestor genetic building blocks. One must bear in mind that single-stranded molecules, such as RNA, are much more mutation-prone than double-stranded ones such as DNA[3]. Therefore smaller values of m would have sufficed for some type of mutation-symmetry breaking to be relevant in terms of information transmission at early stages of life.

B. Protein-Coding DNA

Unlike in the ncDNA case, embedding information in pcDNA is a coding problem with side information at the encoder. This side information is the exact knowledge by the encoder of the primary structure of the protein encoded by the host $\overline{\mathbf{x}}$, that is, $\mathbf{x}' = \alpha(\overline{\mathbf{x}})$. The primary structure $\mathbf{x}'$ determines the encoder state because it must hold for the information-carrying sequence $\overline{\mathbf{y}}$ that $\alpha(\mathbf{y}_i) = \alpha(\mathbf{x}_i) = x'_i$ for all i. This is equivalent to hiding data in a discrete host under an embedding constraint. Nevertheless, apart from the trivial difference of using a 4-ary instead of a typical 2-ary alphabet, several issues distinguish pcDNA data embedding as a special problem.

In order to illustrate these issues consider momentarily a typical data hiding scenario in which a discrete binary host, that is $\underline{\mathbf{x}} = [x_1, \cdots, x_n]$ with $x_i \in \mathcal{X} = \{0, 1\}$, is modified to embed a message b from a given alphabet. The watermarked signal $\underline{\mathbf{y}} = e(\underline{\mathbf{x}}, b)$ must be close to $\underline{\mathbf{x}}$, where closeness is usually measured by means of the Hamming distance $d_H(\underline{\mathbf{y}}, \underline{\mathbf{x}})$. Pradhan et al. [31] and Barron et al. [32] have determined the achievable rate in this scenario, assuming that the elements of $\underline{\mathbf{X}}$ are uniformly distributed, using the average distortion constraint $\frac{1}{n} E[d_H(\underline{\mathbf{Y}}, \underline{\mathbf{X}})] \leq d$, and supposing that $\underline{\mathbf{y}}$ undergoes a memoryless binary symmetric channel with crossover probability q. Their result is

$$R^{\mathrm{unif}} = \mathrm{u.c.e.}\{h(d) - h(q)\} \text{ bits/host symbol},$$

where $\mathrm{u.c.e}\{\cdot\}$ is the upper concave envelope of the argument (when constrained to nonnegative values). Similarly, our initial

[3]For instance, the genomes of RNA viruses such as HIV are known to exhibit as a whole base substitution mutation rates up to $q = 8 \times 10^{-5}$ [23], whereas the genomes of DNA-based organisms typically exhibit as a whole base substitution mutation rates 10^5 times lower.

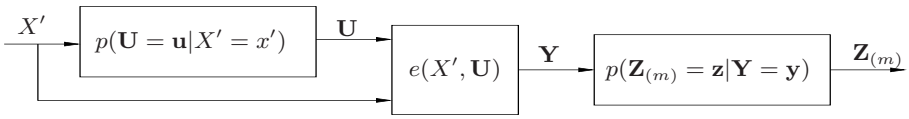

Fig. 4. Scheme of the side-informed pcDNA data embedding scenario.

goal for pcDNA data embedding is to obtain the achievable rate $R_{\rm pc}^{X'}$ for a fixed distribution of $X'=\alpha(\mathbf{X})$ under the symmetric channel discussed in Section II-A, in particular when $\mathbf{X}$ is uniformly distributed, which is denoted as $R_{\rm pc}^{\alpha({\rm unif})}$, similar to the analyses of Pradhan et al. [31] and Barron et al. [32]. Furthermore we will also determine capacity, namely, the maximum achievable rate over all distributions X' of the primary structure encoded by the host.

The first important difference in the pcDNA data embedding scenario is that average inequality constraints on the Hamming distance —such as the ones used in [31], [32]— are meaningless if one wants to carry through to $\overline{\mathbf{y}}$ the full biological functionality of $\overline{\mathbf{x}}$. Instead, since it must always hold that $\alpha(\overline{\mathbf{y}})=\alpha(\overline{\mathbf{x}})$, one must establish the deterministic constraint

$$d_H(\mathbf{y}',\mathbf{x}')=\sum_{i=1}^{n} d_H(y_i',x_i')=0. \tag{11}$$

This is equivalent to assuming that the primary structures of the proteins encoded by the information-carrying sequence and by the host sequence must be identical. This requires that $d_H(y_i',x_i')=0$ for all $i=1,\cdots,n$.

The second distinguishing feature of pcDNA data embedding is due to the variable support of the channel input variable. Whereas in discrete data hiding with binary host one always has $y_i\in\{0,1\}$ independently of x_i, in pcDNA data embedding we must have $\mathbf{y}_i\in\mathcal{S}_{\alpha(\mathbf{x}_i)}$ so that the constraint (11) can always be satisfied. Therefore the support of $\mathbf{y}_i$ is dependent on $\mathbf{x}_i$, as codon equivalence is not evenly spread over the ensemble of amino acids (see Table I).

1) Achievable Rate: Since side information at the encoder must be taken into account in the pcDNA case, then the achievable rate is given by Gel'fand and Pinsker's formula [33] $R_{\rm pc}^{X'}\triangleq\max\{I(\mathbf{Z}_{(m)};\mathbf{U})-I(X';\mathbf{U})\}$ bits/codon, where X' is the distribution of the primary structure that acts at the side information at the encoder, $\mathbf{U}$ an auxiliary random variable which must be determined for the problem at hand, and the maximisation is for nonnegative values of the functional on all distributions $p(\mathbf{y},\mathbf{u}|x')$ under the biological constraint $d_H(\alpha(\mathbf{y}),x')=0$. The intuition behind $\mathbf{U}$ is that this variable should simultaneously act as a good source code for representing the side information (by minimising $I(X';\mathbf{U})$ within the distortion constraint), and as a good channel code for conveying information over the mutation channel (by maximising $I(\mathbf{Z}_{(m)};\mathbf{U})$).

Gel'fand and Pinsker showed in [33] that in the maximisation problem above one may assume that the channel input $\mathbf{Y}$ is a deterministic function of the side information X' and the auxiliary variable $\mathbf{U}$, that is, $\mathbf{Y}=e(X',\mathbf{U})$. Consequently the relationship between the variables in the problem can be summarised by the diagram in Figure 4. Since the support of $\mathbf{Y}|x'$ can only be the set of codons $\mathcal{S}_{x'}$ —so that the biological constraint $d_H(\alpha(\mathbf{y}),x')=0$ can always be met— then the cardinality of the support set of $\mathbf{U}|x'$ must coincide with the multiplicity of x', that is, $|\mathcal{S}_{x'}|$. As $\mathbf{U}$ is an arbitrary auxiliary variable, we may choose the support set of $\mathbf{U}|x'$ to be $\mathcal{S}_{x'}$. This allows us to define the embedding function $\mathbf{Y}|x'=e(x',\mathbf{U})$ as $\mathbf{Y}|x'\triangleq\mathbf{U}|x'$ without loss of generality. Any of the $|\mathcal{S}_{x'}|!$ permutations of the elements of $\mathcal{S}_{x'}$ could actually be chosen to establish this function.

From these considerations it follows that $\mathbf{Y}=\mathbf{U}$. As a result it is assumed in the subsequent computations that $\mathbf{U}$ becomes the mutation channel input (equivalently, the encoder output). Noticing that $\mathcal{S}_{x'}\cap\mathcal{S}_{w'}=\emptyset$ for $x'\neq w'\in\mathcal{X}'$, the distribution of $\mathbf{U}$ can be put as $p(\mathbf{u})=p(\mathbf{u}|x')p(x')$ when $\mathbf{u}\in\mathcal{S}_{x'}$, because $p(\mathbf{v}|x')=0$ when $\mathbf{v}\notin\mathcal{S}_{x'}$. This discussion on $\mathbf{U}$ also implies that $H(X'|\mathbf{U})=0$, since given a codon $\mathbf{u}$ there is no uncertainty on the amino acid represented by it, and therefore $I(X';\mathbf{U})=H(X')$

Since $\mathbf{Y}|(x',\mathbf{u})$ is deterministic, the maximisation on $p(\mathbf{y},\mathbf{u}|x')$ amounts to a maximisation on $p(\mathbf{u}|x')$. Hence the achievable rate for a given distribution of X' can be expressed as

$$R_{\rm pc}^{X'}=\max_{p(\mathbf{u}|x')} I(\mathbf{Z}_{(m)};\mathbf{U})-H(X') \text{ bits/codon}. \tag{12}$$

As $H(\mathbf{Z}_{(m)}|\mathbf{U})$ only depends on the transition probabilities of the symmetric channel given by $\mathbf{\Pi}^m$, and as trivially $H(X')$ only depends on X', the optimisation in (12) amounts to the constrained maximisation of $H(\mathbf{Z}_{(m)})$ on $p(\mathbf{u}|x')$.

For the same reasons as the bounds (9) in Section III-A, the rate (12) is bounded as $R_{\rm pc}^{X'}|_{\gamma=1}\leq R_{\rm pc}^{X'}\leq R_{\rm pc}^{X'}|_{\gamma=0}$. Also, there are several cases where $R_{\rm pc}^{X'}$ can be analytically determined, which are discussed next. First of all, since $C_{\rm nc}|_{\gamma=1,q=3/4}=0$ then $R_{\rm pc}^{X'}|_{\gamma=1,q=3/4}=0$ for any X', because it must hold that $R_{\rm pc}^{X'}\leq 3\,C_{\rm nc}$. Therefore in this catastrophic case the choice of $p(\mathbf{u}|x')$ is irrelevant. Furthermore it can be shown that $p(\mathbf{u}|x')=1/|\mathcal{S}_{x'}|$ —that is, $\mathbf{U}|x'$ uniformly distributed— is the exact maximising strategy in two situations, which are discussed in the following two lemmas.

Lemma 1. *If $q=0$ then the achievable rate is*

$$R_{\rm pc}^{X'}|_{q=0}=E\left[\log|\mathcal{S}_{X'}|\right] \text{ bits/codon}. \tag{13}$$

Proof: Using the chain rule of the entropy one can write $H(\mathbf{U},X')=H(\mathbf{U})+H(X'|\mathbf{U})=H(X')+H(\mathbf{U}|X')$. Since $H(X'|\mathbf{U})=0$, and as $\mathbf{Z}_{(m)}=\mathbf{U}$ when $q=0$, then the achievable rate is given by $R_{\rm pc}^{X'}|_{q=0}=\max_{p(\mathbf{u}|x')}H(\mathbf{U})-H(X')=\max_{p(\mathbf{u}|x')}H(\mathbf{U}|X')$. As $H(\mathbf{U}|X')=\sum_{x'\in\mathcal{X}'}p(x')H(\mathbf{U}|x')$ is maximised when $H(\mathbf{U}|x')$ is maximum for all x', then $\mathbf{U}|x'$ must be uniformly distributed in all cases. Then $H(\mathbf{U}|x')=\log|\mathcal{S}_{x'}|$ and (13) follows. ■

Remark. Note that (13) is the embedding rate intuitively expected in the mutation-free case. For example, if $\mathbf{X}$ were uniformly distributed, which would yield $X' = \alpha(\mathbf{X})$ nonuniform with pmf $p(x') = |\mathcal{S}_{x'}|/|\mathcal{X}|^3$, then one would intuitively compute the rate as $R_{\mathrm{pc}}^{\alpha(\mathrm{unif})}|_{q=0} = \sum_{x'} \frac{|\mathcal{S}_{x'}|}{|\mathcal{X}|^3} \log |\mathcal{S}_{x'}| = 1.7819$ bits/codon,[4] since $|\mathcal{S}_{x'}|$ choices are available to the embedder when the host amino acid is x'. The rate when $\mathbf{X}$ is uniform can actually be obtained in closed form for every q using the following result.

Lemma 2. *If $\mathbf{X}$ is uniformly distributed then the achievable rate is*

$$R_{\mathrm{pc}}^{\alpha(\mathrm{unif})} = \widetilde{C}_{\mathrm{nc}} - H(X') \text{ bits/codon}, \tag{14}$$

where $\widetilde{C}_{\mathrm{nc}} \triangleq \max I(\mathbf{Z}_{(m)}; \mathbf{U})$ and this maximisation is unconstrained on $p(\mathbf{u})$, that is, $\widetilde{C}_{\mathrm{nc}}$ is the capacity of the symmetric codon mutation channel with transition probability matrix $\mathbf{\Pi}^m$.

Proof: Since $p(\mathbf{u}) = p(\mathbf{u}|x')p(x')$ when $\mathbf{u} \in \mathcal{S}_{x'}$, if $\mathbf{X}$ is uniformly distributed then $p(\mathbf{u}) = p(\mathbf{u}|x')|\mathcal{S}_{x'}|/|\mathcal{X}|^3$ when $\mathbf{u} \in \mathcal{S}_{x'}$. Therefore choosing $\mathbf{U}|x'$ to be uniformly distributed implies that $p(\mathbf{u}) = 1/|\mathcal{X}|^3$ for all $\mathbf{u}$. Since $\mathbf{\Pi}^m$ represents a symmetric channel and a uniform input maximises mutual information over such a channel, then $\widetilde{C}_{\mathrm{nc}} = \max_{p(\mathbf{u})} I(\mathbf{Z}_{(m)}; \mathbf{U})$ is achieved in (12). ∎

Remarks. Since $\widetilde{C}_{\mathrm{nc}}|_{q=0} = \log |\mathcal{X}|^3$, it can be seen that the particular case discussed in the previous remark can also be written as $R_{\mathrm{pc}}^{\alpha(\mathrm{unif})}|_{q=0} = \log |\mathcal{X}|^3 - H(X')$. An interesting insight is provided by the fact that the three parallel symmetric channels undergone by the three bases in a codon are mutually independent, and as a result one can use the equality $\widetilde{C}_{\mathrm{nc}} = 3\, C_{\mathrm{nc}}$ in (14). As $H(X')$ is the lower bound to the lossless source coding rate of X', expression (14) tells us something which is intuitively appealing but only exact when $\mathbf{X}$ is uniform: the pcDNA embedding rate equals three times the ncDNA embedding rate minus the rate needed to losslessly convey the primary structure encoded by the host to the decoder.

Unlike the case above, the distribution of $\mathbf{X}$ in real pcDNA sequences (that is, genes) is not uniform. To start with, there can only be a single *Stp* codon in a sequence that encodes a protein, which rules out uniformity of $\mathbf{X}$. For this reason we show next how to compute the achievable rate in the general case, which corresponds to an arbitrary host sequence encoding a primary structure distributed as X'. As in many other channel capacity problems, it does not seem possible to analytically derive a general optimum set of conditional pmf's $p(\mathbf{u}|x')$ in order to compute the maximum achievable rate (12). One can pose the analytical optimisation problem and see that it involves solving a nontrivial system of $|\mathcal{X}|^3 + |\mathcal{X}'|$ nonlinear equations and unknowns. However the numerical solution is straightforward by means of the Blahut-Arimoto algorithm [26] adapted to a side-informed setting. Such an algorithm has been described by Dupuis et al. [34] to determine the rate given by the general Gel'fand and Pinsker formula. However the Blahut-Arimoto algorithm can be given in a simpler way for this problem due to the peculiar form of (12), as discussed in Appendix A.

[4] $\alpha(\mathrm{unif})$ denotes the distribution of X' when $\mathbf{X}$ is uniformly distributed.

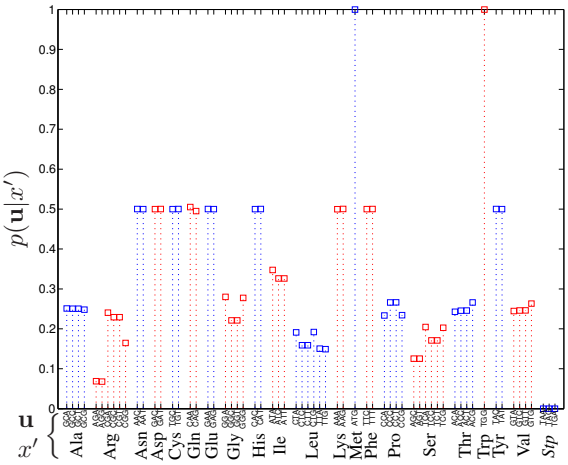

Fig. 5. Example of maximising $p(\mathbf{u}|x')$ distributions numerically obtained using the Blahut-Arimoto algorithm and $p(x')$ corresponding to gene Ypt7 from yeast (GenBank accession number NC_001145), employing $\gamma = 0.1$, $q = 10^{-2}$, $m = 10$. Conditional pmf's are depicted in alternating red and blue colours in order to facilitate plot reading.

Figure 5 shows an example of the optimal distributions numerically obtained for a real gene by means of this numerical method, using the high mutation rate $q = 10^{-2}$ in order to better visualise the results. A pattern can be observed in most of these distributions: among pairs of synonymous codons $\mathbf{u}, \mathbf{v} \in \mathcal{S}_{x'}$ it tends to happen that $p(\mathbf{u}|x') \approx p(\mathbf{v}|x')$ if $[u_1, u_2] = [v_1, v_2]$ and either $u_3, v_3 \in \mathcal{Y}$ or $u_3, v_3 \in \mathcal{R}$. This is due to the effect of mutations according to the Kimura model, since if two codons share their first two bases and their wobble bases are both either purines or pyrimidines, then they should behave symmetrically according to their information transmission properties. The departures from this behaviour are due to numerical effects caused by the finite number of iterations of the Blahut-Arimoto algorithm, whose convergence can be slow [35]. Finally it can be observed that for amino acids x' with $|\mathcal{S}_{x'}| = 6$ (Arg, Leu, and Ser) the two less typical codons from $\mathcal{S}_{x'}$ are assigned smaller likelihoods than the other four codons. This is due to their lower reliability as information transmission symbols, since these codons can lead to the wrong amino acid at the decoder with higher probability.

An analytical set of conditional distributions that yields rates close to the numerical maximum is discussed next. Although one cannot produce a uniform input $\mathbf{U}$ for distributions of X' other than the one in Lemma 2, the use of $p(\mathbf{u}|x') = 1/|\mathcal{S}_{x'}|$ for any x' turns out to yield rates generally close to the numerical maximum. Note from Figure 5 that the numerically obtained $p(\mathbf{u}|x')$ distributions do not differ excessively from the uniform for several amino acids x'. A justification for this is the following. Using the fact that conditioning cannot increase entropy, a suboptimal approach to the maximisation in (12) is given by maximising the lower bound $H(\mathbf{Z}_{(m)}|X') = \sum_{x' \in \mathcal{X}'} p(x') H(\mathbf{Z}_{(m)}|x') \leq H(\mathbf{Z}_{(m)})$. This requires maximising $H(\mathbf{Z}_{(m)}|x') = -\sum_{\mathbf{z} \in \mathcal{X}^3} p(\mathbf{z}|x') \log p(\mathbf{z}|x')$ for all x'. Observing from Table I that synonymous codons often

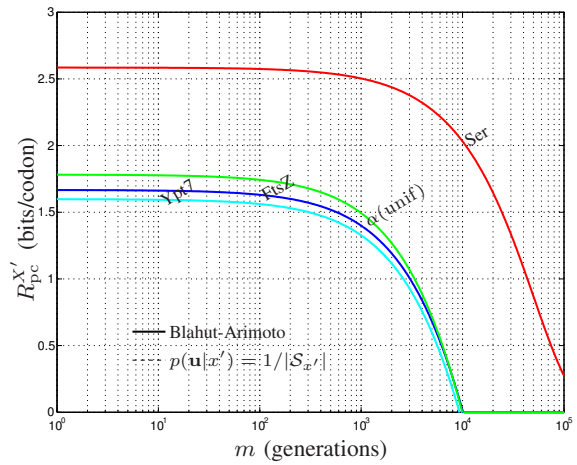

Fig. 6. Embedding rate in pcDNA for different distributions of X' ($\gamma = 1, q = 10^{-5}$).

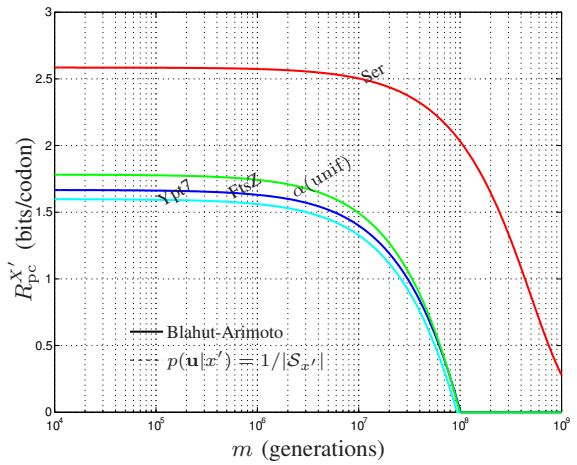

Fig. 7. Embedding rate in pcDNA for different distributions of X' ($\gamma = 1, q = 10^{-9}$).

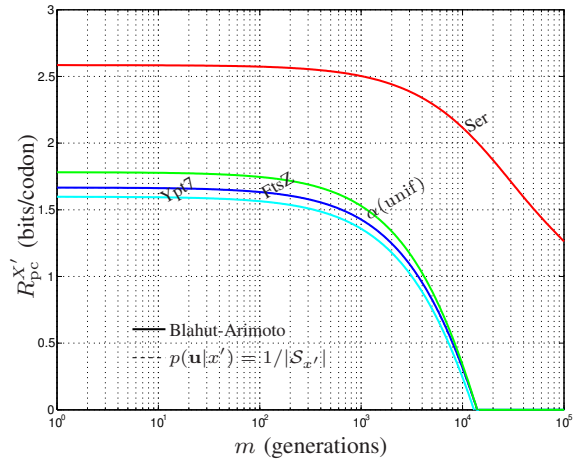

Fig. 8. Embedding rate in pcDNA for different distributions of X' ($\gamma = 0.1, q = 10^{-5}$).

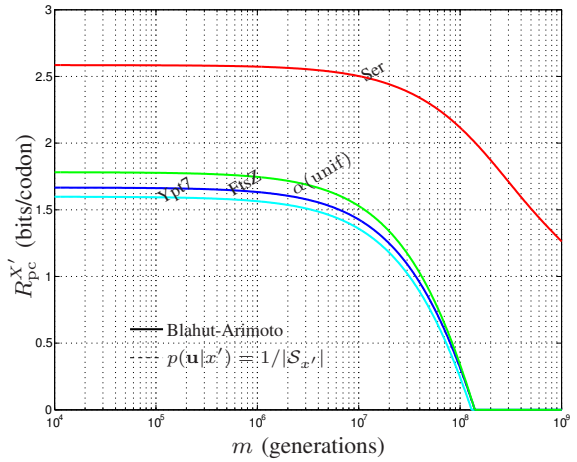

Fig. 9. Embedding rate in pcDNA for different distributions of X' ($\gamma = 0.1, q = 10^{-9}$).

share their two first bases, if $q \ll 1$ we can produce the approximations $p(\mathbf{z}|x') \approx 0$ when $\mathbf{z} \notin \mathcal{S}_{x'}$ and $p(\mathbf{z}|x') \approx \left(\sum_{\mathbf{v}\in\mathcal{S}_{x'}} p(\mathbf{z}|\mathbf{v})\right)^{-1} \sum_{\mathbf{u}\in\mathcal{S}_{x'}} p(\mathbf{z}|\mathbf{u})p(\mathbf{u}|x')$ when $\mathbf{z} \in \mathcal{S}_{x'}$. With this simplification, whenever $\sum_{\mathbf{u}\in\mathcal{S}_{x'}} p(\mathbf{z}|\mathbf{u})$ is constant for all $\mathbf{z} \in \mathcal{S}_{x'}$, choosing $p(\mathbf{u}|x')$ to be uniform implies that $p(\mathbf{z}|x')$ is also uniform, which maximises $H(\mathbf{Z}_{(m)}|x')$. It can be verified that this condition holds for all x' such that $|\mathcal{S}_{x'}| = 1, 2, 4$, which accounts for 16 out of the 21 elements in $\mathcal{X}'$.

Figures 6–9 present the achievable rates in pcDNA for several distributions of X', for the same two values of q in Section III-A and for $\gamma = 1$ (Jukes-Cantor model) and $\gamma = 0.1$ (realistic case). Two of the X' distributions were empirically obtained from the amino acid sequences encoded by two real genes: Ypt7 (*S. Cerevisiae*) and FtsZ (*B. Subtilis*), whose GenBank[5] accession numbers are NC_001145 and NC_000964, respectively. Also depicted are the rate (14) for $\mathbf{X}$ uniform and the rate for the deterministic distribution of X' with outcome Ser, which, as we will discuss in Section III-B2, yields capacity. The plots show that the difference between the results obtained with the Blahut-Arimoto algorithm and those obtained with $p(\mathbf{u}|x') = 1/|\mathcal{S}_{x'}|$ is negligible. In any case, all achievable rates are well below $3\,C_{\mathrm{nc}}$ bits/codon, the rate that can be attained when the embedder is unconstrained. The most important observed effect is that the rates for the two real genes are not far away from the rate for the uniform case, which has the advantage of being analytically determined. There is also a clear cutoff point beyond which no information can be reliably embedded within real genes, which, as an empirical rule-of-thumb, lies around $m \approx 1/(10q)$. Lastly the capacity results are higher for the realistic case with respect to the Jukes-Cantor model, similar to the results in Section III.

a) Codon bias preservation: A final consideration is that real genes exhibit *codon biases*, that is, distributions of $\mathbf{X}|x'$ characteristic to particular species. However the maximisation in (12) implies that, in general, the codon bias of the host sequence $\overline{\mathbf{x}}$ will not be maintained in the optimum information-carrying sequence $\overline{\mathbf{y}}$. Although both sequences will still translate into the very same protein (since $\alpha(\overline{\mathbf{y}}) = \alpha(\overline{\mathbf{x}})$ is enforced), if their codon biases are very different

[5]http://www.ncbi.nlm.nih.gov/genbank

this can have a marked influence on biologically relevant features such as gene translation times [19]. This subsection analyses the achievable rate when an additional codon bias preservation (cbp) constraint is enforced, as this can enhance the biocompatibility of pcDNA data embedding.

The basic observation to be made is the following: if one requires that the original codon bias of the host sequence be preserved in the information-carrying sequence, then one must peg the distribution of $\mathbf{U}|x'$ to the corresponding conditional distribution of the host sequence. With this special constraint, no maximisation on $p(\mathbf{u}|x')$ is required, or possible, and as $p(\mathbf{U}=\mathbf{u}) = p(\mathbf{X}=\mathbf{u})$ the corresponding rate is just $R_{\rm pc}^{\mathbf{X},\,\rm cbp} \triangleq I(\mathbf{Z}_{(m)};\mathbf{X}) - H(X') \leq R_{\rm pc}^{X'}$. The codon bias preservation constraint is equivalent to a steganographic constraint in data hiding problems, since the pmf of the host codons is preserved in the information-carrying sequence. This mirrors Cachin's criterion for perfect steganography [36]. A comparison of maximum rates and codon bias preservation rates for the same two real genes is given in Figure 10. Note that $R_{\rm pc}^{\rm unif,\,cbp} = R_{\rm pc}^{\alpha(\rm unif)}$ because of Lemma 2.

In reality preserving the codon bias of an actual host sequence implies preserving its *codon count*, that is, its codon histogram. Genes are usually short, typically in the order of hundreds of codons, and thus the degrees of freedom for preserving the codon count in the information-carrying gene may not be high. Consequently the information-theoretical analysis for codon bias preservation will usually be optimistic regarding the actual achievable rate when enforcing codon count preservation (ccp).

It is therefore interesting to analyse what can actually be achieved in the latter case. It is assumed in the following that $q=0$, that is to say, a mutation-free scenario. Let $n_{\mathbf{x}}$ be the number of times that codon $\mathbf{x}$ appears in an n-codon long gene $\overline{\mathbf{x}} = [\mathbf{x}_1, \mathbf{x}_2 \cdots, \mathbf{x}_n]$. Therefore $n = \sum_{\mathbf{x}\in\mathcal{X}^3} n_{\mathbf{x}}$. The number r of different DNA sequences with the same amino acid translation *and* the same codon count as $\overline{\mathbf{x}}$ is given by the following product of multinomial coefficients:

$$r = \prod_{x'\in\mathcal{X}'} \frac{\left(\sum_{\mathbf{x}\in\mathcal{S}_{x'}} n_{\mathbf{x}}\right)!}{\prod_{\mathbf{x}\in\mathcal{S}_{x'}} n_{\mathbf{x}}!} \tag{15}$$

This is because for each amino acid x' the multinomial coefficient gives all possible rearrangements of codons corresponding to x' at its $\sum_{\mathbf{x}\in\mathcal{S}_{x'}} n_{\mathbf{x}}$ positions in the sequence, which keeps the codon count fixed. Therefore the mutation-free codon count preservation rate for the finite sequence $\overline{\mathbf{x}}$ is $R_{\rm pc}^{\overline{\mathbf{x}},\,\rm ccp}|_{q=0} = (1/n)\log r$ bits/codon. If $\mathbf{X}$ represents the empirical distribution of codons in $\overline{\mathbf{x}}$ then one clearly has $R_{\rm pc}^{\overline{\mathbf{x}},\,\rm ccp}|_{q=0} \leq R_{\rm pc}^{\mathbf{X},\,\rm cbp}|_{q=0}$, and the rate on the right of the inequality can only be achieved asymptotically, as the length of the sequence distributed as $\mathbf{X}$ goes to infinity. Interestingly both rates are close for real genes. For instance, genes FtsZ and Ypt7 can be deemed to be short: 386 and 208 codons long, respectively. Despite this, $R_{\rm pc}^{\rm FtsZ,\,ccp}|_{q=0} = 1.3723$ bits/codon and $R_{\rm pc}^{\rm Ypt7,\,ccp}|_{q=0} = 1.2797$ bits/codon, which are just slightly lower than the asymptotic rates for the same genes, $R_{\rm pc}^{\rm FtsZ,\,cbp}|_{q=0} = 1.4201$ bits/codon and $R_{\rm pc}^{\rm Ypt7,\,cbp}|_{q=0} = 1.3179$ bits/codon.

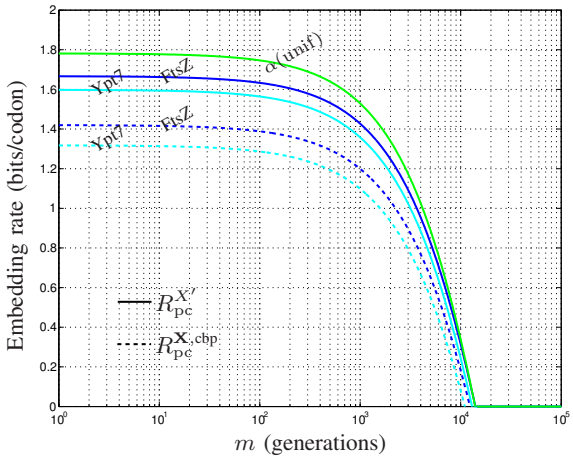

Fig. 10. Comparison of pcDNA embedding rate with and without codon bias preservation constraints, for different distributions of X' ($\gamma = 0.1, q = 10^{-5}$).

In any case, it is shown in Appendix B that for large n

$$R_{\rm pc}^{\overline{\mathbf{x}},\,\rm ccp}|_{q=0} \approx R_{\rm pc}^{\mathbf{X},\,\rm cbp}|_{q=0} = H(\mathbf{X}) - H(X') \text{ bits/codon}, \tag{16}$$

simply put, the combinatorial rate for codon count preservation coincides asymptotically with the Gel'fand and Pinsker's formula with codon bias preservation.

2) Capacity: To conclude, we undertake the computation of the capacity of pcDNA data embedding, which is the maximum achievable rate over all possible distributions of the primary structure encoded by a gene. This problem can be put as

$$C_{\rm pc} \triangleq \max_{p(x')} R_{\rm pc}^{X'} \text{ bits/codon}. \tag{17}$$

The explicit computation of $C_{\rm pc}$ is simple in two particular cases considered in Section III-B1 already:

- $\gamma = 1$ and $q = 3/4$: Trivially, $C_{\rm pc}|_{\gamma=1,q=3/4} = R_{\rm pc}^{X'}|_{\gamma=1,q=3/4} = 0$. However, the point that needs to be made is that although this is true for any X', only deterministic X' yields $H(X') = 0$ exactly, and so, by the continuity of the rate functional, this is the best strategy when approaching $q = 3/4$ from the left.
- $q = 0$: From Lemma 1, $C_{\rm pc}|_{q=0} = \max_{y'} \log|\mathcal{S}_{y'}|$. Since $|\mathcal{S}_{x'}| = 6$ is the maximum for all $x' \in \mathcal{W}' \triangleq \{Ser, Leu, Arg\}$, a distribution of X' that maximises (13) is any for which $\sum_{x'\in\mathcal{W}'} p(x') = 1$. Note that X' may be deterministic, but it does not have to be so. Capacity is then $C_{\rm pc}|_{q=0} = \log 6 = 2.5850$ bits/codon.

Remark. A trivial upper bound for any q is $C_{\rm pc} \leq C_{\rm pc}|_{q=0}$. Since $C_{\rm pc}|_{q=0} < 3\,C_{\rm nc}|_{q=0} = 6$, then side-informed pcDNA data embedding capacity will not be able to achieve non-side-informed ncDNA capacity for every mutation rate. This parallels the results in side-informed encoding with discrete hosts by Pradhan et al. [31] and Barron et al. [32], which have already been discussed. However it does not parallel the well-known "writing on dirty paper" result by Costa [37], which corresponds to continuous Gaussian host, mean squared error distortion, and additive independent Gaussian channel.

From our discussion above on C_{pc} for $q = 3/4$ (with $\gamma = 1$) and for $q = 0$, one might surmise that a primary structure pmf with support in $\mathcal{W}'$ could be capacity-achieving in all cases. The actual capacity-achieving distribution is given by the following theorem:

Theorem 1. *Capacity is achieved by the deterministic pmf of X' that maximises $H(\mathbf{Z}_{(m)})$.*

Proof: See Appendix C. ∎

Remarks. Denoting as ξ' the deterministic outcome of X', it can be numerically verified that $\xi' = \mathrm{Ser}$ maximises $H(\mathbf{Z}_{(m)})$ for all $\gamma \leq 1$, m, and q, and thus $C_{\mathrm{pc}} = R_{\mathrm{pc}}^{\mathrm{Ser}}$ in these conditions. Some examples showing this fact and the rates achievable with deterministic X' are presented in Figures 11–14, which are obtained for the same γ and q parameters as Figures 6–9. The maximum rates, corresponding to Ser in all cases, are highlighted. These plots also show that the rates obtained using the linearised approximation given in Appendix C are practically indistinguishable from the rates obtained using the Blahut-Arimoto algorithm. The approximation $p(\mathbf{u}|x') = 1/|\mathcal{S}_{x'}|$ is also good, but it worsens as γ decreases.

a) Biological interpretations: The results in this section are only concerned with artificial information embedding in pcDNA, and thus appear to have less obvious consequences in biological terms than the results concerning ncDNA. However an intriguing phenomenon which may allude to a biological meaning of these achievable rates is observed in Figures 13 and 14, that is, the range of γ in which the model is realistic. This phenomenon consists of a rate drop for two particular values of $\xi' \in \mathcal{X}'$ as $m \to \infty$ with respect to all other symbols with the same multiplicity $|\mathcal{S}_{\xi'}|$. These two values of ξ' correspond to the stop symbol (*Stp*) and to the amino acid Leu, two codons of which happen to have the twin function of start codons in eukaryotic organisms. Therefore, when considered as individual information transmission blocks, all stop codons and most of the start codons seem to be less suited than the rest of codons for carrying extra information (redundancy). One may surmise that these codons may have undergone special evolutionary pressures during the emergence of the genetic code, due to their very specific functions. One possibility may be that since the stop symbol cannot be repeated in a single gene, it has been under less pressure to perform as a genetic information carrier. A similar effect may have been at work in the evolution of the start codons corresponding to Leu. However a more interesting conjecture is that the narrower "genetic windows" for these special codons are due to natural selection broadly favouring an increase of the complexity of organisms over the generations, which by and large requires gene lengthening. Making these special codons less reliable, plus the effect of natural selection, may have been a mechanism for such progressive elongation of genes.

IV. Conclusion

This paper has provided an analysis of the Shannon capacity of DNA data embedding when mutations are described according to the Kimura model of molecular evolution. The analysis may be used to assess the optimality of a growing number of bioengineering procedures aimed at inserting arbitrary information in the genomes of living organisms, both within noncoding and protein-coding sections.

Some biological connections of the results given have also been discussed, leading to some unexpected insights into the mutation-asymmetry of the DNA bases and into the behaviour of key building blocks in the genetic code —start and stop codons. These results indicate that the Shannon capacity may hold clues to a deeper understanding of the genetic machinery, which is an exciting possibility.

Further research should consider insertion and deletion mutations (*indels*) in order to complete the treatment of the errors that a DNA data embedding algorithm may face. Although the exact capacity computation under indels is an unsolved problem in most digital communications settings, bounding strategies and approximations reliant on realignment methods which are widely used in bioinformatics may be useful for addressing this problem. Generalisations of the Kimura model with up to nine parameters may also be considered. However it has been shown by means of an unconstrained real estimate of the mutation channel that generalised models may not overly change the analysis herein.

Appendix

A. Blahut-Arimoto algorithm for pcDNA data embedding

Since the subtracting entropy in (12) is independent of $p(\mathbf{u}|x')$ we just need to maximise $I(\mathbf{Z}_{(m)}; \mathbf{U})$. In the standard Blahut-Arimoto algorithm one would alternatively maximise $p^*(\mathbf{u}|\mathbf{z})$ and $p^*(\mathbf{u})$ using $p(\mathbf{z}|\mathbf{u})$ (that is, $\mathbf{\Pi}^m$). However the distribution of $\mathbf{U}$ is constrained in (12) by the distribution $p(x')$ of the side information, and instead we have to maximise $p^*(\mathbf{u}|\mathbf{z})$ and $p^*(\mathbf{u}|x')$ alternatively. Writing the mutual information as

$$I(\mathbf{Z};\mathbf{U}) = \sum_{x' \in \mathcal{X}'} \sum_{\substack{\mathbf{z} \in \mathcal{X}^3 \\ \mathbf{u} \in \mathcal{S}_{x'}}} p(\mathbf{z}|\mathbf{u}) p(\mathbf{u}|x') p(x') \log \frac{p(\mathbf{u}|\mathbf{z})}{p(\mathbf{u}|x')p(x')}, \tag{18}$$

it is straightforward to see (following the same reasoning as in [26]) that the two iterative maximisation steps in this case are

- Step 1:

$$p^*(\mathbf{u}|\mathbf{z}) = \frac{p(\mathbf{z}|\mathbf{u})p(\mathbf{u}|x')p(x')}{\sum_{y'} \sum_{\mathbf{v} \in \mathcal{S}_{y'}} p(\mathbf{z}|\mathbf{v})p(\mathbf{v}|y')p(y')}, \tag{19}$$

  where $\mathbf{u} \in \mathcal{S}_{x'}$.
- Step 2:

$$p^*(\mathbf{u}|x') = \frac{\prod_{\mathbf{z}} p(\mathbf{u}|\mathbf{z})^{p(\mathbf{z}|\mathbf{u})}}{\sum_{\mathbf{v} \in \mathcal{S}_{x'}} \prod_{\mathbf{z}} p(\mathbf{v}|\mathbf{z})^{p(\mathbf{z}|\mathbf{v})}}. \tag{20}$$

Note that the first step is the same as in the standard Blahut-Arimoto algorithm, because $p(\mathbf{u}) = p(\mathbf{u}|x')p(x')$ if $\mathbf{u} \in \mathcal{S}_{x'}$. The second step is also essentially the same, but it is applied $|\mathcal{X}'|$ times to determine each input distribution $p(\mathbf{u}|x')$ instead of only once to determine the input distribution $p(\mathbf{u})$.

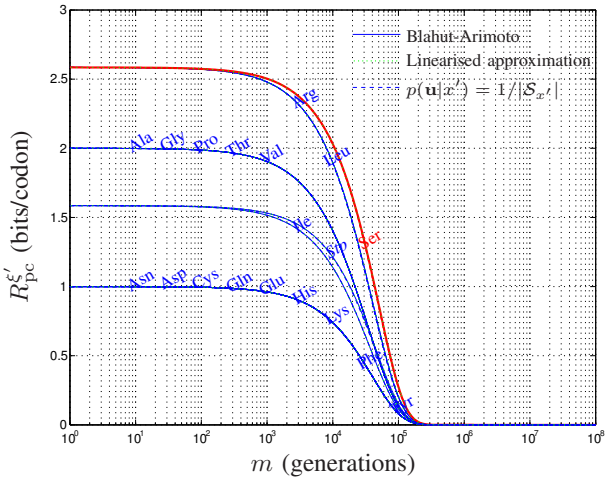

Fig. 11. Achievable pcDNA data embedding rates for deterministic X' ($\gamma = 1$, $q = 10^{-5}$).

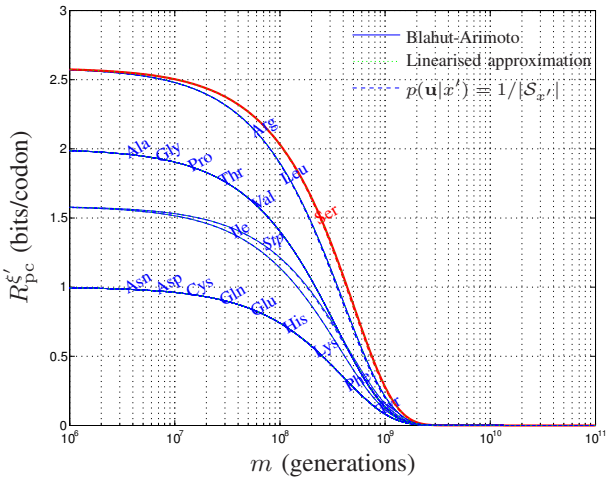

Fig. 12. Achievable pcDNA data embedding rates for deterministic X' ($\gamma = 1$, $q = 10^{-9}$).

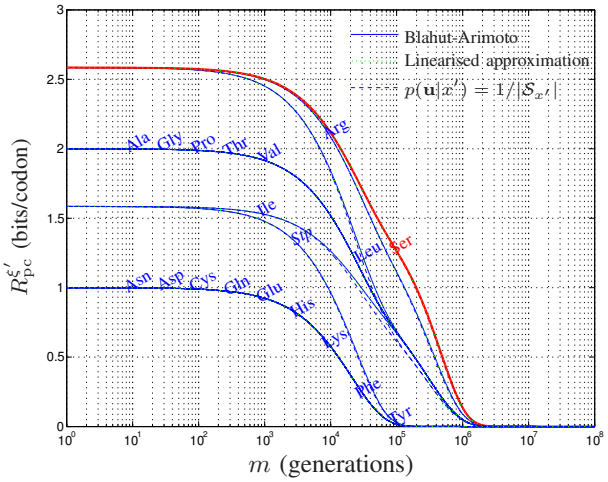

Fig. 13. Achievable pcDNA data embedding rates for deterministic X' ($\gamma = 0.1$, $q = 10^{-5}$).

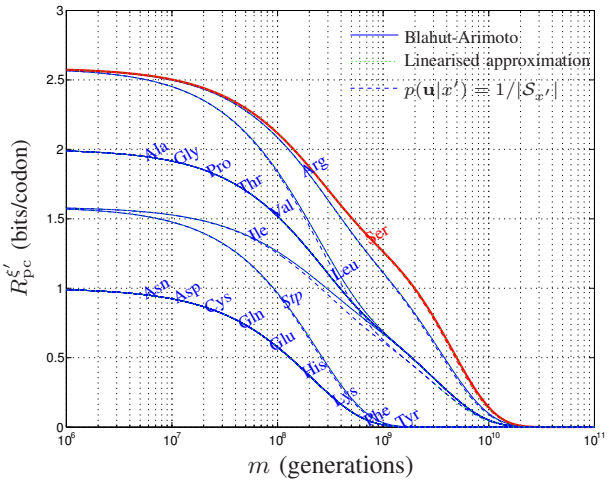

Fig. 14. Achievable pcDNA data embedding rates for deterministic X' ($\gamma = 0.1$, $q = 10^{-9}$).

B. Asymptotically achievable rate with codon count preservation and no mutations

The number of bits that can be embedded in the protein-coding sequence $\overline{\mathbf{x}}$ while preserving its codon count is $l = \log r$, with r given by (15), that is,

$$l = \sum_{x'} \log \left(\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}} \right)! - \sum_{x'} \sum_{\mathbf{x} \in \mathcal{S}_{x'}} \log n_{\mathbf{x}}! \qquad (21)$$

Assuming that n is large, this amount may be developed using Stirling's approximation $\log t! \approx t \log t - t$ (for natural logarithms). When using this approximation in (21) the summands without a logarithm cancel out and we get:

$$l \approx \sum_{x'} \left(\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}} \right) \log \left(\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}} \right) - \sum_{x'} \sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}} \log n_{\mathbf{x}}. \qquad (22)$$

Using the asymptotic expressions $p(x') \approx (1/n) \sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}}$ and $p(\mathbf{x}) \approx (1/n)\, n_{\mathbf{x}}$, it can be seen after some algebra that (22) can be in turn approximated by

$$l \approx n \left(\sum_{x'} p(x') \log p(x') - \sum_{x'} \sum_{\mathbf{x} \in \mathcal{S}_{x'}} p(\mathbf{x}) \log p(\mathbf{x}) \right). \qquad (23)$$

Therefore the rate $R_{\mathrm{pc}}^{\overline{\mathbf{x}},\,\mathrm{ccp}}|_{q=0} = l/n$ can be asymptotically approximated for large n as $R_{\mathrm{pc}}^{\overline{\mathbf{x}},\,\mathrm{ccp}}|_{q=0} \approx H(\mathbf{X}) - H(X')$ bits/codon.

C. Capacity-achieving strategy $p(x')$

In order to find the capacity-achieving strategy for the amino acid distribution one needs to solve

$$\frac{\partial}{\partial p(x')} \left[H(\mathbf{Z}_{(m)}) - H(X') + \nu \left(\sum_{y' \in \mathcal{X}'} p(y') - 1 \right) \right] = 0 \qquad (24)$$

for $x' \in \mathcal{X}'$, with ν a Lagrange multiplier. In what follows, $p(\mathbf{z}|x')$ denotes $p(\mathbf{Z}_{(m)} = \mathbf{z}|X' = x')$ for notational convenience. For simplicity, and without loss of generality, we assume natural logarithms in the optimisation. Using $\partial p(\mathbf{z})/\partial p(x') = p(\mathbf{z}|x')$, expression (24) becomes

$$\sum_{\mathbf{z}\in\mathcal{X}^3} p(\mathbf{z}|x') \log \left(\sum_{y'\in\mathcal{X}'} p(y')p(\mathbf{z}|y') \right) = \log p(x') + \nu, \quad (25)$$

for $x' \in \mathcal{X}'$. The solution remains unchanged if we multiply (25) across by $p(x')$. This operation allows us to see by inspection that any extreme of the Lagrangian inside the differential in (24) has to be deterministic, that is, $p(x') = 1$ for some $x' = \xi'$ and $p(x') = 0$ for $x' \neq \xi'$. Note that this is in agreement with the strategies for the cases $q = 0$ and $\gamma = 1$ with $q = 3/4$ discussed in Section III-B2. One may verify that a uniform distribution of X' cannot possibly solve (25) for all $x' \in \mathcal{X}'$, because $\sum_{\mathbf{z}} p(\mathbf{z}|x') \log \left(\sum_{y'} p(\mathbf{z}|y') \right)$ is not constant on x' unless $\gamma = 1$ and $q = 3/4$, in which case we have shown that the Shannon capacity is zero for any distribution.

According to the previous discussion, for any capacity-achieving solution it always holds that $H(X') = 0$. Therefore we just have to maximise $H(\mathbf{Z}_{(m)})$ over the ensemble of 21 deterministic distributions of X'. The computation of $R_{\mathrm{pc}}^{\xi'}$ and the maximising distribution $\mathbf{U}|\xi'$ can be accomplished using the Blahut-Arimoto algorithm, following the discussion in Section III-B1 on the optimal strategy for fixed $p(x')$. Note that $\xi' = \mathrm{Trp}$ and $\xi' = \mathrm{Met}$ can be completely ruled out, since $|\mathcal{S}_{\mathrm{Trp}}| = |\mathcal{S}_{\mathrm{Met}}| = 1$, and then only null rates are possible in these cases. We only need therefore to compute $R_{\mathrm{pc}}^{\xi'}$ for the remaining 19 amino acids and then choose the maximum. Also, $\xi' = \mathit{Stp}$ can only be considered hypothetically, since this symbol can only appear exactly once in a gene.

a) Approximation to maximising strategy for deterministic X': It is also possible to provide a closed-form approximation to the maximising distribution $\mathbf{U}|\xi'$, which is more accurate than just using the approximation $p(\mathbf{u}|\xi') = 1/|\mathcal{S}_{\xi'}|$ discussed in Section III-B1. It must firstly be observed that if X' is deterministic, then the side-informed setting amounts to a non-side informed setting with $|\mathcal{S}_{\xi'}|$ inputs and $|\mathcal{X}|^3$ outputs. This setting can be modelled by a transition probability matrix $\mathbf{\Lambda}$ whose rows are the rows of $\mathbf{\Pi}^m$ corresponding to the codons associated with ξ'. In general this channel will not be symmetric nor weakly symmetric, since although its rows are permutations of the same set of probabilities, its columns are not, and their sum is not constant either. However $H(\mathbf{Z}_{(m)}|\mathbf{U})$ is still independent of the distribution of $\mathbf{U}$, so we only need to maximise $H(\mathbf{Z}_{(m)})$ to find capacity. The corresponding conditions for the maximum are

$$\sum_{\mathbf{z}\in\mathcal{X}^3} p(\mathbf{z}|\mathbf{v}) \log p(\mathbf{z}) + 1 = \rho, \quad (26)$$

for $\mathbf{v} \in \mathcal{S}_{\xi'}$, and with ρ a Lagrange multiplier. Using $\log x \leq x - 1$ and $p(\mathbf{z}) = \sum_{\mathbf{u}\in\mathcal{S}_{\xi'}} p(\mathbf{z}|\mathbf{u})p(\mathbf{u}|\xi')$, one can write

$$\sum_{\mathbf{z}\in\mathcal{X}^3} p(\mathbf{z}|\mathbf{v}) \sum_{\mathbf{u}\in\mathcal{S}_{\xi'}} p(\mathbf{z}|\mathbf{u})p(\mathbf{u}|\xi') \leq \rho, \quad (27)$$

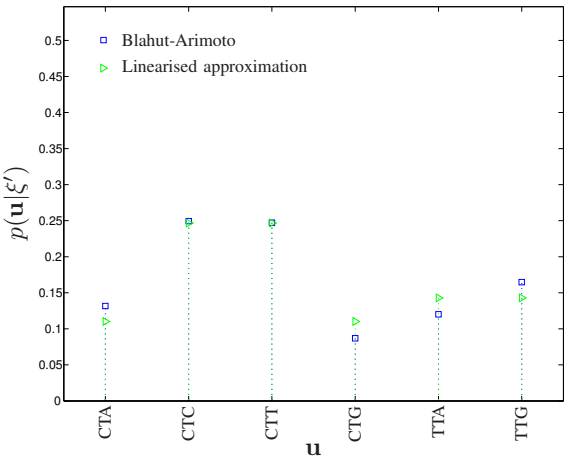

Fig. 15. Comparison of maximising $p(\mathbf{u}|\xi')$ distributions for the deterministic case $\xi' = \mathrm{Leu}$ ($\gamma = 0.1$, $q = 10^{-2}$, $m = 100$)

for $\mathbf{v} \in \mathcal{S}_{\xi'}$. Our approximation consists of solving $p(\mathbf{u}|\xi')$ by enforcing equality in (27) for all $\mathbf{v} \in \mathcal{S}_{\xi'}$. This yields the linear system

$$\boldsymbol{\pi} \left(\mathbf{\Lambda}\mathbf{\Lambda}^T \right) = \rho \mathbf{1}, \quad (28)$$

where the probabilities $p(\mathbf{u}|\xi')$, with $\mathbf{u} \in \mathcal{S}_{\xi'}$, are the elements of the $1 \times |\mathcal{S}_{\xi'}|$ vector $\boldsymbol{\pi}$ (arranged in the same codon order as the rows of $\mathbf{\Lambda}$), and $\mathbf{1}$ is an all-ones vector of size $1 \times |\mathcal{S}_{\xi'}|$. Since $\boldsymbol{\pi}$ must be a pmf, we may fix any arbitrary value of ρ, such as $\rho = 1$, and then normalise the solution $\widetilde{\boldsymbol{\pi}}$ to the resulting linear system, that is

$$\widetilde{\boldsymbol{\pi}} = \mathbf{1}(\mathbf{\Lambda}\mathbf{\Lambda}^T)^{-1}. \quad (29)$$

The matrix $\mathbf{\Lambda}\mathbf{\Lambda}^T$ is invertible if both $q \neq 1/(4\gamma/3)$ and $q \neq 1/(2(1-\gamma/3))$ because in this case the rows of $\mathbf{\Lambda}$ are linearly independent. This is due to the fact that under the two conditions above the rows of $\mathbf{\Pi}^m$ are linearly independent, since its eigenvalues are all the possible products of three eigenvalues of Π^m [25] and the conditions above guarantee that these are nonzero. A sufficient condition for the invertibility of $\mathbf{\Lambda}\mathbf{\Lambda}^T$ is $q < 1/2$, which spans most cases of interest.

Since the optimisation problem has been linearised, $\widetilde{\boldsymbol{\pi}}$ may contain negative values, but in practice these are relatively small. Setting these values to zero and normalising $\widetilde{\boldsymbol{\pi}}$, an approximation to the optimum distribution $p(\mathbf{u}|\xi')$ is obtained. An example of this approximation compared to the results of the Blahut-Arimoto algorithm is shown in Figure 15, where the high mutation rate $q = 10^{-2}$ is used for visualisation purposes.

Acknowledgement

The author is grateful to the anonymous reviewers for their detailed and constructive comments. He also wishes to thank D. Haughton, whose research motivated the codon count preservation analysis, and N. Barron for proofreading the final manuscript.

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] J. P. Cox, "Long-term data storage in DNA," *Trends in Biotechnology*, vol. 19, no. 7, pp. 247–250, July 2001.

[3] 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.

[4] 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.

[5] M. Arita and Y. Ohashi, "Secret signatures inside genomic DNA," *Biotechnol. Prog.*, vol. 20, no. 5, pp. 1605–1607, September-October 2004.

[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] 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, April 2007.

[8] N. Yachie, Y. Ohashi, and M. Tomita, "Stabilizing synthetic data in the DNA of living organisms," *Systems and Synthetic Biology*, vol. 2, no. 1–2, pp. 19–25, December 2008.

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

[10] D. Heider, M. Pyka, and A. Barnekow, "DNA watermarks in non-coding regulatory sequences," *BMC Research Notes*, vol. 2, no. 125, July 2009.

[11] D. G. Gibson, G. A. Benders, C. Andrews-Pfannkoch, E. A. Denisova, H. Baden-Tillson, J. Zaveri, T. B. Stockwell, A. Brownley, D. W. Thomas, M. A. Algire, C. Merryman, L. Young, V. N. Noskov, J. I. Glass, J. C. Venter, C. A. Hutchison, and H. O. Smith, "Complete chemical synthesis, assembly, and cloning of a mycoplasma genitalium genome," *Science*, vol. 319, no. 5867, pp. 1215–1220, 2008.

[12] D. G. Gibson, J. I. Glass, C. Lartigue, V. N. Noskov, R.-Y. Chuang, M. A. Algire, G. A. Benders, M. G. Montague, L. Ma, M. M. Moodie, C. Merryman, S. Vashee, R. Krishnakumar, N. Assad-Garcia, C. Andrews-Pfannkoch, E. A. Denisova, L. Young, Z.-Q. Qi, T. H. Segall-Shapiro, C. H. Calvey, P. P. Parmar, C. A. Hutchison, H. O. Smith, and J. C. Venter, "Creation of a bacterial cell controlled by a chemically synthesized genome," *Science*, vol. 329, no. 5987, pp. 52–56, 2010.

[13] D. C. Jupiter, T. A. Ficht, J. Samuel, Q.-M. Qin, and P. de Figueiredo, "DNA watermarking of infectious agents: Progress and prospects," *PLoS Pathogens*, vol. 6, no. 6, June 2010.

[14] D. Heider and A. Barnekow, "DNA watermarks: A proof of concept," *BMC Molecular Biology*, vol. 9, no. 40, April 2008.

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

[16] M. Kimura, "A simple method for estimating evolutionary rate in a finite population due to mutational production of neutral and nearly neutral base substitution through comparative studies of nucleotide sequences," *J. Molec. Biol.*, vol. 16, pp. 111–120, 1980.

[17] E. E. May, "Bits and bases: An analysis of genetic information paradigms," in *41st Asilomar Conference on Signals, Systems and Computers (ACSSC)*, Asilomar, USA, November 2007, pp. 165–169.

[18] G. Battail, "Information theory and error-correcting codes in genetics and biological evolution," in *Introduction to Biosemiotics*, M. Barbieri, Ed. Springer, 2007.

[19] Y. Lavner and D. Kotlar, "Codon bias as a factor in regulating expression via translation rate in the human genome," *Gene*, vol. 345, no. 1, pp. 127–138, 2005.

[20] M. Mitzenmacher, "A survey of results for deletion channels and related synchronization channels," *Probab. Surveys*, vol. 6, pp. 1–33, 2009.

[21] A. Purvis and L. Bromham, "Estimating the transition/transversion ratio from independent pairwise comparisons with an assumed phylogeny," *Journal of Molecular Evolution*, vol. 44, pp. 112–119, 1997.

[22] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow, "Rates of spontaneous mutation," *Genetics*, vol. 148, no. 4, pp. 1667–1686, 1998.

[23] 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.

[24] W. Li, *Molecular Evolution*. Sinauer Associates, 1997.

[25] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, 3rd ed. John Wiley & Sons, 1999.

[26] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. on Information Theory*, vol. 18, no. 4, pp. 460 – 473, Jul. 1972.

[27] F. Balado, "Genetic channel capacity revisited," in *Procs. of the 6th Int. ICST Conf. on Bio-Inspired Models of Networks, Information, and Computing Systems*, ser. LNICST, E. H. et al, Ed., vol. 103. York, UK: Springer, December 2011, pp. 85–98.

[28] E. E. May, M. D. Rintoul, A. M. Johnston, W. Hart, J. Watson, and R. Pryor, "Detection and reconstruction of error control codes for engineered and biological regulatory systems," Sandia National Laboratories, Tech. Rep., 2003.

[29] A. Gutfraind, "Error-tolerant coding and the genetic code," Master's thesis, University of Waterloo, 2006.

[30] W. Gilbert, "Origin of life: The RNA world," *Nature*, vol. 319, no. 6055, pp. 618–618, Feb 1986.

[31] 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.

[32] 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.

[33] 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.

[34] F. Dupuis, W. Yu, and F. Willems, "Blahut-Arimoto algorithms for computing channel capacity and rate-distortion with side information," in *Intl. Symposium on Information Theory (ISIT)*, June-July 2004, p. 179.

[35] Y. Yu, "Squeezing the Arimoto–Blahut algorithm for faster convergence," *IEEE Trans. on Information Theory*, vol. 56, no. 7, pp. 3149–3157, 2010.

[36] C. Cachin, "An information-theoretic model for steganography," in *Procs. of the Second International Workshop on Information Hiding*, ser. Lecture Notes in Computer Science, vol. 1525. Springer-Verlag, April 1998, pp. 306–318, revised version in Information and Computation, 2004.

[37] M. H. Costa, "Writing on dirty paper," *IEEE Trans. on Information Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

Félix Balado (M'03) graduated with an M.Eng. in Telecommunications Engineering from the University of Vigo (Spain) in 1996, and received a Ph.D. from the same institution in 2003 for his work in digital data hiding. He is currently a lecturer in University College Dublin (Ireland). His research interests lie in the areas of multimedia signal processing, data hiding, digital communications, and bioinformatics.