Gene Tagging and the Data Hiding Rate

Félix Balado and David Haughton

School of Computer Science and Informatics
University College Dublin

E-mail: felix@ucd.ie david.haughton@ucdconnect.ie

Abstract — **We analyze the maximum number of ways in which one can intrinsically tag a very particular kind of digital asset: a gene, which is just a DNA sequence that encodes a protein. We consider gene tagging under the most relevant biological constraints: protein encoding preservation with and without codon count preservation. We show that our finite and deterministic combinatorial results are asymptotically —as the length of the gene increases— particular cases of the stochastic Gel'fand and Pinsker capacity formula for communications with side information at the encoder, which lies at the foundations of data hiding theory. This is because gene tagging is a particular case of DNA watermarking.**

Keywords — **Gene tagging, DNA watermarking, combinatorial analysis, Gel'fand and Pinsker capacity.**

I Introduction

The advent of the biotechnological age has made it possible to tag genes —the molecular units of heredity in living organisms. Gene tagging refers to marking genes in unique, distinct ways, usually in order to enable tracking applications. The most fundamental strategy to tag a gene, which is essentially a digital sequence, is to modify it in such a way that it is possible to distinguish it later from the original gene (and from other differently tagged versions of it). This has sometimes been termed DNA watermarking, and a number of methods to digitally tag genes have been proposed over the last years (see [1, 2, 3]). DNA watermarking constraints are biological, and thus very different from the imperceptibility constraints typically used in watermarking of multimedia assets. The most important constraint is that the full biological functionality of the gene must always be preserved, so that the tagged gene may be reintroduced in a living being by means of recombinant DNA techniques and still remain fully operative. This has actually been achieved with living organisms by Arita and Ohashi [2] and by Heider and Barnekow [3].

In this paper we analyze the maximum number of ways in which a given gene can be tagged, relying on finite and deterministic combinatorial analyses. Furthermore, we show that the asymptotics of the information rates derived from these deterministic analyses coincide with the stochastic limiting rates furnished by the Gel'fand and Pinsker formula for the capacity of communications with side information at the encoder. This is not surprising given the fact that gene tagging is a special data hiding problem. This type of link between the asymptotics of deterministic combinatorics and probabilistic information-theoretical amounts is related to previous results, which we discuss towards the end of this paper.

II Basic Concepts and Notation

Calligraphic letters ($\mathcal{X}$) denote sets; $|\mathcal{X}|$ is the cardinality of $\mathcal{X}$. Boldface Roman letters ($\mathbf{x}$) denote row vectors, $\mathbf{x} = [x_1, \cdots, x_n]$. A Roman letter that appears in uppercase (X) and in lowercase (x) denotes a random variable and a realization of it, respectively. $p(X = x)$, or just $p(x)$ when unambiguous from the context, is the probability mass function or distribution of X. X can also refer to its distribution, depending on the context. $E(X)$ is the expectation of the random variable X and $H(X)$ its entropy. $I(X;Y) = H(X) - H(X|Y)$ is the mutual information between X and Y. Log-

arithms are base 2 throughout the paper, except when otherwise indicated. The indicator function is defined as $\mathbb{1}_{\{A\}} = 1$ if event A is true, and zero otherwise.

Our analysis will require some basic knowledge about the genetic machinery —in particular the definition of what a gene is. The DNA alphabet $\mathcal{X} \triangleq \{\mathrm{A}, \mathrm{C}, \mathrm{T}, \mathrm{G}\}$ is formed by the symbols corresponding to its four bases: adenine, cytosine, thymine, and guanine. A *codon* is formed by a triplet of consecutive bases in a genetic sequence. A *gene* is simply a sequence of codons —flanked by special start and stop markers— that can be translated into a sequence of amino acids by the genetic machinery, and then assembled in the same order imposed by the codon sequence to form a protein. The sequence of amino acids that the gene translates into is usually called its *primary structure*. Using their standard short names, the set of possible amino acids is

$$\begin{aligned}\mathcal{X}' \triangleq \{&\text{Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly,}\\ &\text{His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr,}\\ &\text{Trp, Tyr, Val, } \textit{Stp}\},\end{aligned} \tag{1}$$

and therefore $|\mathcal{X}'| = 21$. Every single codon $\mathbf{y} = [y_1, y_2, y_3] \in \mathcal{X}^3$ can be mapped to a unique amino acid or start/stop translation symbol, that is,

$$\xi(\mathbf{y}) = y' \in \mathcal{X}', \tag{2}$$

where the mapping $\xi(\cdot) : \mathcal{X}^3 \to \mathcal{X}'$ is established by the nearly-universal *genetic code* (easily found elsewhere, see for instance [4]), which partitions $\mathcal{X}^3$ into $|\mathcal{X}'|$ disjoint subsets of codons. The subset of synonymous codons associated to amino acid $y' \in \mathcal{X}'$ is $\mathcal{S}_{y'} \triangleq \{\mathbf{y} \in \mathcal{X}^3 | \xi(\mathbf{y}) = y'\}$. Note that the ensemble of stop codons is collected under the special label *Stp* in (1), and thus loosely classed as an "amino acid" for notational convenience. However *Stp* just indicates the end of gene translation and thus does not actually stand for any amino acid. Also Met (and two codons associated to Leu in eukaryotic cells) double as gene translation start symbols. We call the number of synonymous codons that map to amino acid y' the *multiplicity* of y'; this is just the cardinality of $\mathcal{S}_{y'}$, that is, $|\mathcal{S}_{y'}|$. It can be seen from the genetic code that multiplicities are uneven over the set of amino acids, as $|\mathcal{S}_{y'}| \in \{1, 2, 3, 4, 6\}$. Due to the uniqueness of the codon-to-amino acid mapping, $\mathcal{S}_{y'} \cap \mathcal{S}_{w'} = \emptyset$ for $y' \neq w' \in \mathcal{X}'$, and $\sum_{y' \in \mathcal{X}'} |\mathcal{S}_{y'}| = |\mathcal{X}|^3 = 64$ since $\cup_{y' \in \mathcal{X}'} \mathcal{S}_{y'} = \mathcal{X}^3$.

To sum up, the main notational conventions in what follows are that bold Roman letters (i.e. $\mathbf{y}$, $\mathbf{Y}$) are associated to codons, whereas primed Roman letters (i.e. y', Y') are associated to amino acids.

III Gene Tagging

From the previous introduction, it is clear that any particular gene can be written as a vector of codons $\overline{\mathbf{x}} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$, where $\mathbf{x}_i \in \mathcal{X}^3$ is the i-th codon in the gene. In this section we will obtain the tagging rates associated to $\overline{\mathbf{x}}$. A tagging rate is just the number of bits per codon needed to represent one of the possible tags. We will consider two relevant cases: 1) the basic case in which the tagged gene $\overline{\mathbf{y}}$ is just constrained to translate into the same protein as $\overline{\mathbf{x}}$; and 2) the codon count preservation case, in which $\overline{\mathbf{y}}$ additionally preserves a biologically meaningful feature of $\overline{\mathbf{x}}$ known as *codon bias*, that will be described in Section b). The codon count for a given codon $\mathbf{x} \in \mathcal{X}^3$ is $n_{\mathbf{x}} \triangleq \sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{x}_i = \mathbf{x}\}}$, that is, the number of times that it appears in gene $\overline{\mathbf{x}}$. Clearly, the sum of the codon counts for all possible codons has to be equal to the length of the gene, that is, $\sum_{\mathbf{x} \in \mathcal{X}^3} n_{\mathbf{x}} = n$.

a) Primary Structure Preservation Case

Since each codon can be replaced by any other synonymous codon without altering the translation of the gene into a protein, the number m of different DNA sequences $\overline{\mathbf{y}}$ with the same primary structure as $\overline{\mathbf{x}}$ is

$$\begin{aligned} m &= \prod_{i=1}^{n} |S_{\xi(\mathbf{x}_i)}| \\ &= \prod_{x' \in \mathcal{X}'} |S_{x'}|^{\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}}} \end{aligned} \tag{3}$$

This is because if a codon $\mathbf{x}$ represents the amino acid $x' = \xi(\mathbf{x})$ then we have $|\mathcal{S}_{x'}|$ alternatives to choose from. Using expression (3), the tagging rate for that gene can be expressed as $R = \frac{1}{n} \log m$ bits/codon, that is

$$R = \frac{1}{n} \sum_{x' \in \mathcal{X}'} \Big(\sum_{\mathbf{x} \in \mathcal{S}'_x} n_{\mathbf{x}} \Big) \log |\mathcal{S}_{x'}| \text{ bits/codon} \tag{4}$$

If we now approximate $p(x') \approx \frac{1}{n} \sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}}$ for large n, then the tagging rate (4) can in turn be approximated as $R \approx \widetilde{R} = \sum_{x' \in \mathcal{X}'} p(x') \log |\mathcal{S}_{x'}|$, that is, the asymptotic approximation is

$$\widetilde{R} = E(\log |\mathcal{S}_{X'}|) \text{ bits/codon.} \tag{5}$$

In the particular case in which codons are uniformly distributed, one has that $p(x') = |\mathcal{S}_{x'}|/|\mathcal{X}|^3$ and (5) becomes

$$\widetilde{R}^{\text{unif}} = H(\mathbf{X}) - H(X') \text{ bits/codon.} \tag{6}$$

That is, in this case the tagging rate can be approximated by the difference of discrete entropies between the codon and the amino acid distributions of the gene. We will see that this particular

result is parallel to the general result in the following section. However it must be noted that codons are never uniformly distributed in nature.

b) Codon Count Preservation Case

Tagging a gene with the preservation of its primary structure as the sole constraint —as we have done in the previous section, and as assumed in and [1], [2] and [3]— is risky from a biological point of view. Despite the fact that m tagged genes have the exact same protein translation as the original gene $\overline{\mathbf{x}}$, many of those possible $\overline{\mathbf{y}}$ sequences may have codon counts very different to that of $\overline{\mathbf{x}}$. In practice this means that the original codon bias, or equivalently the original distributions of codons per amino acid, $p(\mathbf{x}|x')$, will usually be different from the corresponding distributions in the tagged gene, $p(\mathbf{y}|x')$. This can have important effects for gene expression. For instance, it is known that the time that it takes for the genetic machinery to translate a gene into a protein is dependent on its codon bias [5], and this bias is variable and organism dependent. Therefore fully biocompatible gene tagging should preserve not only the primary structure of the original gene, but also its codon count.

The number m_c of different sequences $\overline{\mathbf{y}}$ which have the same amino acid translation *and* codon count as the original gene $\overline{\mathbf{x}}$ is given by the following product of multinomial coefficients:

$$m_c = \prod_{x' \in \mathcal{X}'} \binom{\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}}}{n_{\mathbf{x} \in \mathcal{S}_{x'}}} = \prod_{x' \in \mathcal{X}'} \frac{\left(\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}}\right)!}{\prod_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}}!} \tag{7}$$

This is because, for each amino acid x', the corresponding multinomial coefficient in (7) gives all possible rearrangements of codons corresponding to x' on the $\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}}$ positions at which this amino acid appears in the primary structure of the gene. Clearly, all such rearrangements are the only ways to preserve the original codon count.

As in Section a), the tagging rate can be expressed as $R_c = \frac{1}{n} \log m_c$ bits/codon. This expression can be straightforwardly developed as follows:

$$R_c = \frac{1}{n} \left\{ \sum_{x'} \log \frac{\left(\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}}\right)!}{\prod_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}}!} \right\} = \frac{1}{n} \left\{ \sum_{x'} \log \left(\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}} \right)! - \sum_{x'} \sum_{\mathbf{x} \in \mathcal{S}_{x'}} \log n_{\mathbf{x}}! \right\}. \tag{8}$$

Assuming that n is large, and for convenience switching momentarily to nats/codon instead of bits/codon, we can further develop expression (8) using Stirling's approximation $\log p! \approx p \log p - p$, which leads to

$$R_c \approx \widetilde{R}_c = \frac{1}{n} \left\{ \sum_{x'} \left(\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}} \right) \log \left(\sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}} \right) - \sum_{x'} \sum_{\mathbf{x} \in \mathcal{S}_{x'}} n_{\mathbf{x}} \log n_{\mathbf{x}} \right\}. \tag{9}$$

Using now the same approximation of $p(x')$ in Section a) and $p(\mathbf{x}) \approx \frac{1}{n} n_{\mathbf{x}}$, one gets after some manipulations that

$$\widetilde{R}_c = \left(\sum_{x'} p(x') \log p(x') - \sum_{x'} \sum_{\mathbf{x} \in \mathcal{S}_{x'}} p(\mathbf{x}) \log p(\mathbf{x}) \right). \tag{10}$$

Therefore, for any codon distribution, the tagging rate with codon count preservation can be approximated as

$$\widetilde{R}_c = H(\mathbf{X}) - H(X') \text{ bits/codon}. \tag{11}$$

Due to the additional constraint imposed, $m_c < m$ and then $R_c < R$ for any gene. However with the asymptotic approximations $\widetilde{R}_c = \widetilde{R}$ can be achieved if $\mathbf{X}$ is uniformly distributed, because in that case (6) takes the same form as (11).

IV Data Hiding Rate

Tagging a gene can be seen as akin to embedding a unique piece of information within that gene by exploiting its inherent redundancy, subject to the genetic constrains considered in Section III. As such, it is just a particular instance of data hiding, or information transmission with side information at the encoder, in which the tagging rate is the data hiding rate for this particular setting.

We will show in this section that the asymptotic results from Section III are just particular cases of the well-known Gel'fand and Pinsker capacity formula [6]. This result tells us that if X is some side information available at the encoder and Z is the channel output, then the maximum transmission rate is given by

$$C = \max_{p(u|x)} I(Z;U) - I(U;X) \text{ bits/symbol}, \tag{12}$$

where U is an auxiliary random variable that must be determined for each particular problem, and the channel input Y (the information-carrying signal) is a deterministic function of U and X, that is, $Y = e(U, X)$. Also Y must be close to X according to some problem-dependent distortion constraint.

In the gene tagging setting the side information at the encoder is an amino acid $X' = \xi(\mathbf{X})$ corresponding to a codon $\mathbf{X}$ of the original gene, which as we will see completely determines the channel

state. The auxiliary variable will be denoted as $\mathbf{U}$, because we will see later that it can be seen as taking codon values. The basic information-carrying signal is the tagged codon $\mathbf{Y}$. We do not need to consider a channel output, since we are not considering any distortions on the tagged gene in the combinatorial analysis, and therefore $I(\mathbf{Z};\mathbf{U}) = H(\mathbf{U})$. The fundamental distortion constraint is just the equality $\xi(\mathbf{Y}) = X'$, which guarantees that the tagged codon always stands for the same amino acid as the original codon. Since $\mathbf{Y} = e(\mathbf{U}, X')$, the cardinality of the support set of $\mathbf{U}|x'$ can only be $|\mathcal{S}_{x'}|$. Since $\mathbf{U}$ is an auxiliary variable one may arbitrarily choose the support set of $\mathbf{U}|x'$ to be actually $\mathcal{S}_{x'}$. One can then define the tagged codon variable as $\mathbf{Y}|x' = \mathbf{U}|x'$ without loss of generality (although there are actually $|S_{x'}|!$ equivalent choices for the encoder, from the achievable rate viewpoint), and thus guarantee that the genetic distortion constraint is fulfilled. Therefore the variable $\mathbf{Y}$, which represents the distribution of the tagged gene, is equal to $\mathbf{U}$ in what follows.

Focusing next our attention on the subtracting mutual information in (12) see that $H(X'|\mathbf{U}) = 0$, because given a codon there is no uncertainty about the amino acid that it represents. Then we may particularize (12) for gene tagging as

$$C = \max_{p(\mathbf{u}|x')} H(\mathbf{U}) - H(X') \text{ bits/codon}. \tag{13}$$

It must be remarked that the tagging rates in the combinatorial analyses (4) and (8) in Section III concern finite and deterministic genetic sequences. As such, they are always upper bounded by the rates given by the more general formula (13), that is, $R_c < R \leq C$. This is because this formula concerns stochastic rather than deterministic signals, and it is a limit which, in general, can only be attained asymptotically as the number of channel uses tends to infinity. However, as we will see next, the asymptotic approximations (5) and (11) of these combinatorial analyses are able to achieve C.

a) Primary Structure Preservation Case

In this case we just have to carry out the maximization in (13). Using the chain rule of the entropy we can write $H(\mathbf{U}, X') = H(\mathbf{U}) + H(X'|\mathbf{U}) = H(X') + H(\mathbf{U}|X')$. As $H(X'|\mathbf{U}) = 0$, this implies that (13) can be put as

$$C = \max_{p(\mathbf{u}|x')} H(\mathbf{U}|X'). \tag{14}$$

One can see now that the maximization of

$$H(\mathbf{U}|X') = \sum_{x' \in \mathcal{X}'} p(x') H(\mathbf{U}|x') \tag{15}$$

is achieved when $H(\mathbf{U}|x')$ is maximum for all x', which implies that $p(\mathbf{u}|x')$ be uniform for all x'.

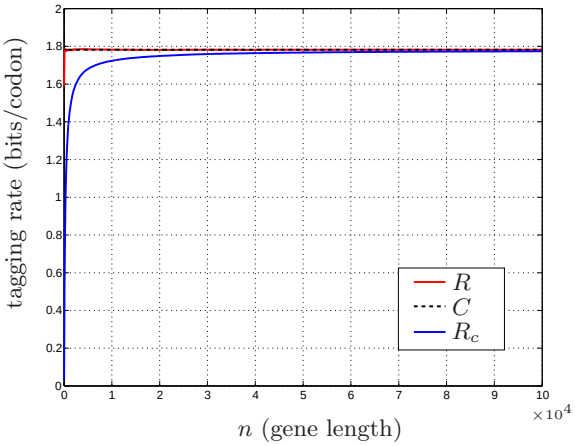

Fig. 1: Comparison of combinatorial rates (R, R_c) versus Gel'fand and Pinsker rate ($C = C_c$) for uniform host

Then $H(\mathbf{U}|x') = \log|\mathcal{S}_{x'}|$ and consequently

$$C = E(\log|\mathcal{S}_{X'}|) = \widetilde{R}. \tag{16}$$

b) Codon Count Preservation Case

Since (13) contemplates the preservation of the primary structure already, additionally preserving the codon count just amounts to enforcing $p(\mathbf{u}|x') = p(\mathbf{x}|x')$, that is, keeping the codon bias found in the original gene in the tagged gene. Therefore in this case no maximization of (13) is needed, or possible. Hence $\mathbf{U}$ has to be distributed as $\mathbf{X}$, and it follows that

$$C_c = H(\mathbf{X}) - H(X') = \widetilde{R}_c. \tag{17}$$

Due to the additional constraint $C_c \leq C$. As $\mathbf{Y} = \mathbf{U}$, note that the codon count preservation constraint is akin to a steganographic constraint in data hiding, that is, Cachin's criterion for perfect steganography [7].

V Discussion and Conclusions

Figure 1 shows a comparison of the combinatorial rates with and without codon count preservation against the Gel'fand and Pinsker capacity when $\mathbf{X}$ is uniformly distributed, for varying gene length n. In this case $C = C_c = 1.7819$ bits/codon. We can see that $R \approx C$ even for small n, whereas $R_c \to C$ as $n \to \infty$. Some results for real genes are also given in Table 1, showing that with real genes $\widetilde{R}_c \approx C_c$ and that preserving the codon bias does not necessarily entail an important tagging rate decrease with respect to the unconstrained rate C. The results in Table 1 use the empirical amino acid distribution and codon bias distributions of the corresponding genes, obtained from GenBank using the accession numbers provided.

It is interesting to mention that the connections that we have given between asymptotic combinatorial analyses and side-informed achievable rates

GenBank	Name	Species	$C = \widetilde{R}$	C_c	$\widetilde{R}_c$
NC_001137.3	RAD51	*S. Cerevisiae*	1.6651	1.6096	1.6092
NC_000913.2	ftsA	*E. Coli*	1.7487	1.7087	1.7086
NC_000964.3	aprE	*B. Subtilis*	1.6948	1.6490	1.6489

Table 1: Tagging rates (bits/codon) for some real genes

can actually be seen as an informal application of the method of types [8]. For instance, a well known such connection is the following one: take a Bernoulli random variable X and assume that n independent outcomes of this binary variable have resulted in k ones. The binomial coefficient gives the number m of different ways in which such a situation can happen, that is, $m = \binom{n}{k} = n!/(k!(n-k)!)$. Using the Stirling's approximation of the factorial we can describe these m possibilities using the rate $R = \frac{1}{n}\log m$ bits/outcome, that is

$$R \approx \widetilde{R} = -\frac{k}{n}\log\frac{k}{n} - \frac{(n-k)}{n}\log\frac{(n-k)}{n}. \qquad (18)$$

Now, as we did in Section III, for large n we can approximate $k/n \approx p(X=1)$ and $(n-k)/n \approx p(X=0)$, which yields the approximation $\widetilde{R} \approx H(X)$, or, equivalently, $\binom{n}{k} \approx 2^{nH(X)}$, that is, the combinatorial analysis can be asymptotically approximated using the entropy of the random variable that generates the deterministic outcomes. A more rigorous analysis actually shows that $\binom{n}{k} \leq 2^{nH(X)}$ [8].

To conclude, the work described here should mainly be seen as a theoretical contribution for the application of data hiding to an unusual kind of host. Our analysis might also be able to inform future practical tagging strategies in biotechnological applications. As an example, we may point out the recent suggestion by Jupiter *et al* of tagging infectious agents manipulated in biotechnological laboratories, in order to enable tracking and liability determination in case of leaks [9].

Acknowledgement

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

References

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

[2] M. Arita and Y. Ohashi. Secret signatures inside genomic DNA. *Biotechnol. Prog.*, 20(5):1605–1607, September-October 2004.

[3] D. Heider and A. Barnekow. DNA-based watermarks using the DNA-Crypt algorithm. *BMC Bioinformatics*, 8(176), February 2007.

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

[5] Y. Lavner and D. Kotlar. Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene*, 345(1):127–138, 2005.

[6] S. I. Gel'fand and M. S. Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, 1980.

[7] C. Cachin. An information-theoretic model for steganography. In *Procs. of the Second International Workshop on Information Hiding*, volume 1525 of *Lecture Notes in Computer Science*, pages 306–318. Springer-Verlag, April 1998.

[8] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

[9] Daniel C. Jupiter, Thomas A. Ficht, James Samuel, Qing-Ming Qin, and Paul de Figueiredo. DNA watermarking of infectious agents: Progress and prospects. *PLoS Pathogens*, 6(6), June 2010.