A MODIFIED WATERMARK SYNCHRONISATION CODE FOR ROBUST EMBEDDING OF DATA IN DNA

David Haughton and Félix Balado

School of Computer Science and Informatics
University College Dublin, Ireland

ABSTRACT

DNA data embedding is a newly emerging field aspiring to encode data in deoxyribonucleic acid (DNA). DNA is an inherently digital and noisy medium, undergoing substitution, insertion and deletion mutations. Hence, encoding information in DNA can be seen as a particular case of digital communications in which biological constraints must be observed. In this paper we propose a modification of Davey and MacKay's watermark synchronisation code (unrelated to digital watermarking) to create an encoding procedure more biocompatible with the host organism than previous methods. In addition, when combined with a low density parity check (LDPC) code, the method provides near-optimum error correction. We also obtain the theoretical embedding capacity of DNA under substitution mutations for the increased biocompatibility constraint. This result, along with an existing bound on capacity for insertion and deletion mutations, is compared to the proposed algorithm's performance by means of Monte Carlo simulations.

***Index Terms*—** DNA data embedding, decoder performance, LDPC, watermark code, capacity

1. INTRODUCTION

Recent advances in technology have enabled the use of DNA in novel ways, such as highly parallel computing and data storage. Numerous applications of DNA data embedding have been proposed in the past decade, such as the tracking of organisms' movements, the assertion of proprietary DNA sequences [1] and highly compact data storage [2]. One real world scenario, proposed in literature, where this field could directly apply is in determining the source of a biological containment following an outbreak [3]. A high profile instance of DNA data embedding was undertaken recently, the purpose of which was to encode, *in vitro* (using DNA in test tubes), an entire book in DNA sequences [4]. This method incorporated error correction in the form of repetition coding. While the experiment was not performed *in vivo* (using DNA within living organisms), the results are nonetheless very encouraging, as the information density (bits/mm^3) attained is several orders of magnitude higher than flash memory, even surpassing the demonstrated density of quantum memory.

Encoding data in living organisms is more challenging because it is undesirable for encoded regions to alter the biological processes of the host organism. *In vivo* data embedding involves encoding information within genes (protein coding DNA, or pcDNA) or within the rest of the genome (noncoding DNA, or ncDNA). Here we will consider the ncDNA data embedding scenario only. Several groups have successfully performed *in vivo* ncDNA information embedding already, such as Wong et al.[5] and Yachie et al.[6] using wild bacteria, or the JCVI using an artificially engineered synthetic bacterium [7]. Nevertheless, unlike in the method that we will present here, some key biocompatibility constraints for embedding information in ncDNA are not strictly enforced by any of these previous methods.

Also, DNA in the cell is subject to substitution and *indel* (insertion and deletion) mutations. From the point of view of DNA data embedding these are equivalent to a probabilistic "mutation channel" inducing random errors. As such data embedding algorithms should incorporate error correction specific to the DNA mutation channel. However despite the proposal of many DNA data encoding methods in the last ten years, none make use of optimal error correction, although Yachie et al.[6] do use repetition coding and Heider and Barnekow [8] have applied basic error correction codes such as Hamming. One particularly challenging error control problem is that *indel* mutations create desynchronisation errors. Given the nature of the mutations channel, LDPC codes in conjunction with modified watermark synchronisation codes proposed by Davey and MacKay [9] are ideally suited to this task.

To conclude, although the general limits of information embedding in noncoding DNA strands are known for some particular mutation channels [10], this is not the case when special constraints for increased biocompatibility, such as the ones considered here, are enforced. Consequently no practical method has been proposed to achieve these limits either. Here we also provide the Shannon capacity analysis for methods operating under this constraint

2. NOTATION AND FRAMEWORK

This section outlines the biological framework, in conjunction with the channel model and principle coding components. Throughout the paper random variables are represented by upper case italicised letters. Vectors are denoted by bold lower case letters and upper case calligraphic letters denote sets. The fundamental units in DNA are nucleotide bases, and the set of DNA bases is given by $\mathcal{X} \triangleq \{\text{A,C,T,G}\}$. A DNA molecule may be represented as an n length vector $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$, $x_i \in \mathcal{X}$. A DNA molecule is double-stranded, that is, composed by two sequences of identical length. One strand is completely determined by the other through the so-called Watson-Crick pairings C–G and A–T.

As already mentioned, there are two types of regions in the DNA of living organisms: protein coding DNA (pcDNA) and noncoding DNA (ncDNA). Although we are only concerned with data embedding in ncDNA it is necessary to review some facts about pcDNA in order to explain the increased biocompatibility constraint. pcDNA encodes genes as sequences of codons. Codons are base triplets,

This publication has emanated from research supported by Science Foundation Ireland under Grant Number 09/RFP/CMS2212.

each of which codes for an amino acid according to the genetic code [11]. Start codons are special codons which signal the genetic machinery to begin the translation of a gene into a protein (amino acid sequence). The standard set of start codons is [11]

$$\mathcal{S} = \{\text{ATG, CTG, TTG, CAT, CAG, CAA}\}. \quad (1)$$

A start codon can be found at any locus in a DNA molecule's two strands, and hence, there are six possible amino acid translation frames. It is also important to note that the strand complementary to $\mathbf{x}$ is read in the opposite direction, and thus termed antiparallel.

2.1. Encoding and Channel Models

In this section we outline the proposed DNA data embedding method, as summarised by Figure 1. Details of the individual components will be elucidated in Section 5. A binary message vector $\mathbf{m}$ to be encoded, is mapped to a quaternary vector, $\mathbf{b}$ using any trivial mapping $\{00, 01, 10, 11\} \mapsto \mathcal{X}$. Since $|\mathcal{X}| = 4$, all error correction methods discussed here operate over the finite field $\text{GF}(4)$. Next, an LDPC encoder generates the encoded vector $\mathbf{c}$ from $\mathbf{b}$. A biocompatible watermark encoder is then applied to enable the correction of *indel* mutations while simultaneously ensuring the biocompatibility of the encoded information. The output of this is $\mathbf{y}$, to be embedded in an organism in specific ncDNA sites known to have no biological function, as we will discuss in Section 3.

Next the organism (including $\mathbf{y}$) undergoes a probabilistic error channel which is modelled by a substitution mutation channel concatenated with an *indel* mutation channel. We will assume that all mutation events are mutually independent. The base substitution mutation model used here, called the Kimura model of molecular evolution [12], reflects that DNA bases belong to one of two chemically distinct groups: purines, $\mathcal{R} = \{\text{A, G}\}$ or pyrimidines $\mathcal{Y} = \{\text{C, T}\}$. Mutations within each of these two groups are more likely than mutations to the opposite group; these are known as transitions and transversions, respectively. For one generation of an organism, the transition probability matrix $\Pi = [p(Y|X)]$, with $X, Y \in \mathcal{X}$, of Kimura's model is

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

The two parameters of the model are q, the base substitution mutation rate per generation, and γ, which is obtained from the transitions/transversions rate ϵ as $\gamma = 3/(2(\epsilon + 1))$. The well known Jukes-Cantor model is a particular case of the Kimura model where $\gamma = 1$. The Kimura transition probability matrix after p generations is given by Π^p. Finally the rate of base *indel* mutations per generation and site is denoted by ρ.

The decoder receives a mutated version of $\mathbf{y}$, denoted $\mathbf{z}$, from a descendant of the original host organism. Between $\mathbf{y}$ and $\mathbf{z}$, p generations have elapsed. The watermark decoder resynchronises $\mathbf{z}$, which in effect means identifying *indels*, removing insertions and inserting a base for deletions. It also produces a maximum likelihood estimate of the substitution mutation channel parameters, $\hat{q}$ and $\hat{\gamma}$. These parameters are passed to the LDPC decoder, which decodes the output of the watermark decoder $\hat{\mathbf{c}}$ to produce $\hat{\mathbf{b}}$. The quaternary decoder then performs the reverse mapping of the quaternary encoder, mapping bases to information bits.

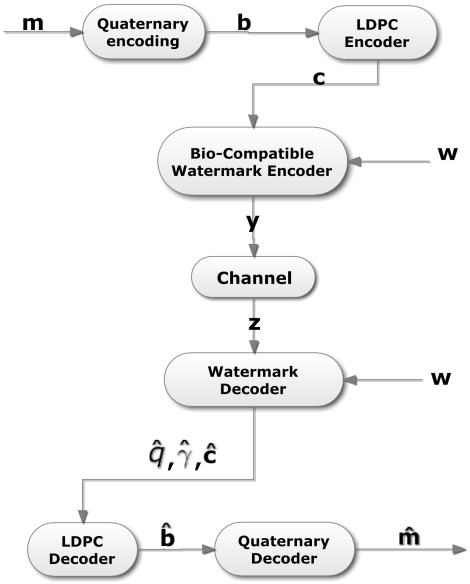

Fig. 1: Channel model with encoding and decoding elements.

3. CONSTRAINTS OF ncDNA EMBEDDING

When encoding data in ncDNA it is essential that the host's biological functionality remains unaffected. Although ncDNA has in the past been labelled as "junk DNA" due to its supposed lack of purpose when compared to gene-encoding pcDNA, recent research spearheaded by the ENCODE consortium [13] has demonstrated that up to 80% of ncDNA may be involved in regulatory functions of gene activity. However it is safe to assume that the remaining 20% of the ncDNA of a genome can be appended, inserted or overwritten without ill effects, as proven by Wong et al. [5], Yachie et al. [6, 14] and Gibson et al. [15], who have successfully embedded information in suitable regions of the ncDNA of living organisms. In truly redundant ncDNA regions, a further constraint to increase biocompatibility is that they must not be mistaken as protein coding regions by the cell's machinery. To enforce this, start codons must not appear in $\mathbf{y}$ or in its complementary antiparallel strand. The problem becomes increasingly difficult when we consider that the appearance of start codons must be prevented in each of the three reading frames, for both sequences. There are four dinucleotides, composed of the first two bases of each codon in $\mathcal{S}$, which indicate that a start codon may appear. This set is given by

$$\mathcal{B} = \{\text{AT, CT, TT, CA}\}. \quad (2)$$

A data embedding algorithm called BioCode ncDNA will now be presented. This method was previously proposed by us to ensure that start codons do not appear in an encoded sequence [16]. At the encoder, base y_i will be encoded at position i with reference to the trailing two bases already encoded, denoted by $\mathbf{d} = [y_{i-2}, y_{i-1}]$. It is only when $\mathbf{d} \in \mathcal{B}$ that the possible values of y_i must be restricted to a particular subset of $\mathcal{X}$. This subset contains all the bases for which a start codon will not appear both in the strand being encoded and in its antiparallel complimentary strand; otherwise any $y_i \in \mathcal{X}$ can be chosen by the encoder. During encoding, $\mathbf{d}$ can be monitored, and using a lookup of Table 1, bits from $\mathbf{m}$ can be encoded in such a way that start codons never appear. This procedure will be exploited in Section 5 to enforce the no-start-codons constraint.

$\mathbf{d}$	AT	CT	TT	CA	$\mathcal{X}^2 \setminus \mathcal{B}$
$\lvert\mathcal{S}_\mathbf{d}\rvert$	3	3	3	1	4
$\mathcal{S}_\mathbf{d}$	A	A	A	C	A
	T	T	T		T
	C	C	C		C
					G
	↑ Encode			↓ Decode	
$\mathcal{M}_\mathbf{d}$	0	0	0	–	00
	10	10	10		01
	11	11	11		10
					11

Table 1: Given the dinucleotide sequence $\mathbf{d}$ the next message base to be encoded is one belonging to the set $\mathcal{S}_\mathbf{d}$. Each bit message found in $\mathcal{M}_\mathbf{d}$ corresponds to a base in $\mathcal{S}_\mathbf{d}$.

4. CAPACITY OF ncDNA EMBEDDING WITH NO-START-CODONS CONSTRAINT

One way to determine the performance of a practical ncDNA method is to assess its distance from channel capacity, which is the upper bound on the maximum information rate that can be asymptotically transmitted without errors. In this section we establish the capacity of ncDNA with the no-start-codons constraint discussed in Section 3. In our computation only the substitution channel will be considered. The rationale behind this is that if the watermark code works perfectly this allows the LDPC decoder to "see" a substitution mutations channel only. It must be remarked that the exact capacity for even the simplest *indel* channel remains unknown [17]. In any case, the approach proposed here still yields an upper bound to capacity when *indels* are present.

To begin, the capacity the Kimura channel without the no-start-codons constraint is

$$C_{\mathrm{nc}} = 2 - H(Z_{(p)}|Y)\text{bits/base}, \tag{3}$$

where Y models a base at a given ncDNA site and $Z_{(p)}$ its mutated version after p generations of the host organism. This is just the capacity of a symmetric 4-ary channel, which is attained for uniform Y. The conditional entropy in (3) is easily computed using any row of Π^p. The case with the no-start-codons constraint, which we address next, is more complicated because a single-letter characterisation of the channel cannot be used (i.e. a base-by-base analysis). Firstly let us determine the achievable rate conditioned to $\mathbf{d} = [y_{i-2}, y_{i-1}]$, which we denote $R_i(\mathbf{d})$. If $\mathbf{d} \notin \mathcal{B}$, then there are no constraints on y_i, and the rate is then given by Equation 3. When $\mathbf{d} \in \mathcal{B}$ we have to consider two cases:

1. $\mathbf{d} = [\mathrm{C}, \mathrm{A}]$: observing Table 1 the rate in this case is 0, since only one base may be used to prevent a start codon.
2. Otherwise, the channel becomes nonsymmetric, with input belonging to $\{\mathrm{A}, \mathrm{T}, \mathrm{C}\}$ and output belonging to $\mathcal{X}$. In this case the maximising input distribution and maximum rate may be numerically obtained using the Blahut-Arimoto algorithm [18].

Combining these considerations, the average rate at step i is

$$R_i = \sum_{\mathbf{d}\in\mathcal{X}^2} R_i(\mathbf{d}) p(\mathbf{d}) \text{ bits/base}, \tag{4}$$

where $p(\mathbf{d})$ is the probability mass function (pmf) of the trailing two bases, $\mathbf{d}$. For the first step ($i = 3$) we may assume uniformity, that is, $p(y_1, y_2) = 1/16$. After that, using the 16 conditioned pmf's $p(y_i|y_{i-2}, y_{i-1})$ that yield the maximum rates at step i, we can compute $p(y_{i-1}, y_i) = \sum_{y_{i-2}\in\mathcal{X}} p(y_i|y_{i-2}, y_{i-1}) p(y_{i-2}, y_{i-1})$, which can be used to obtain R_{i+1} as in (4). The overall average rate is obtained in the limit as

$$\bar{R} = \lim_{n\mapsto\infty} \frac{1}{n} \sum_{i=1}^{n} R_i \text{ bits/base}. \tag{5}$$

5. ENCODING AND DECODING ALGORITHMS

Optimal coding strategies are needed to reach capacity. Such methods exist in digital communications for the equivalent of the substitution mutation channel, but have not yet been applied to the DNA data embedding problem. As for the *indel* mutation channel, whose capacity is unknown, one of the most effective resynchronisation error correction codes to date is the watermark code, proposed by Davey and MacKay over ten years ago [9]. These authors also proposed using an LDPC code concatenated with the watermark code to correct substitution mutations. Since LDPC codes are near-optimal under substitution channels, this combination is ideally suited to DNA data embedding, provided the biological constraints discussed in Section 3 are observed. This will be achieved by modifying the watermark code.

5.1. Encoding Algorithm

The LDPC code used in our experiments was created with reference to LDPC codes designed for wireless communications [19]. Parity check matrices were constructed randomly, with the condition that short cycles in the corresponding Tanner graph not be present. The resynchronisation method used here to correct *indel* mutations is a variant of the original watermark encoding algorithm [9]. While the watermark facilitates resynchronisation, it serves two other important functions for this encoding scheme: 1) to produce an estimate of the substitution mutation channel parameters; and 2) to prevent start codons from appearing.

Firstly a no-start-codon watermark $\mathbf{w}$ known to the encoder and decoder is created. To create $\mathbf{w}$ a pseudorandom binary vector is generated and then encoded in DNA using the BioCode ncDNA method in Section 3. The algorithm then iterates through $\mathbf{w}$ and identifies positions at which any base can be safely substituted in without creating a start codon. It is at these locations that $\mathbf{c}$ will be inserted. Given the base w_i, it is useful to define all possible codon offsets which this base is part of,

$$\mathcal{W}_i = \{[w_{i-2}, w_{i-1}, w_i], [w_{i-1}, w_i, w_{i+1}], [w_i, w_{i+1}, w_{i+2}]\}.$$

It is determined that base w_i is suitable for embedding only if $\mathcal{W}_i \nsubseteq \mathcal{S}$. Otherwise w_i has the potential to create a start codon if it is replaced by the wrong base. $\mathbf{w}$ may then be trimmed such that it only specifies enough embeddable positions to exactly contain $\mathbf{c}$. If this is done the position of substitutions does not have to be explicitly stated: the decoder can determine these positions by finding all possible locations that are suitable for embedding.

In the original watermark code proposal the forward-backward algorithm is used to infer *indel* sites. Our decoding algorithm, first proposed in [16], instead aligns $\mathbf{z}$ with $\mathbf{w}$ using the edit distance. This process of resynchronisation is suboptimal but still effective and more practical, since it uses hard decision decoding and thus does not require estimates of the *indel* rate. Two important factors

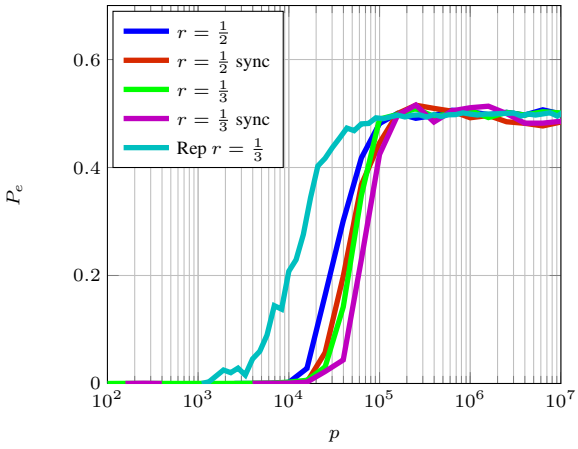

Fig. 2: The probability of bit error, P_e versus the number of generations p of a host organism for watermark codes (blue, green) compared to perfectly synchronised versions (red, purple). The rate, r in the legend above refers to the rate of the LDPC code. Also shown is the P_e of a repetition code for ncDNA embedding of rate $r = \frac{1}{3}$.

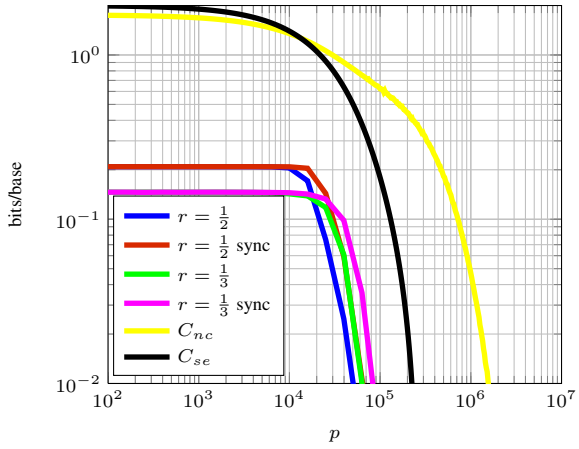

Fig. 3: The mutual information of watermark codes (blue, green) compared to perfectly synchronised versions (red, purple). Also shown is the bound presented in Section 4 for substitution mutation only (yellow). A capacity bound, proposed in literature [20], for the Jukes-Cantor model ($\gamma = 1$) with deletions (black) is also given.

must be incorporated into the scoring process: desynchronisation errors cannot occur in $\mathbf{w}$, and substitutions are expected in $\mathbf{z}$. These observations reduce the complexity of aligning the two sequences. This method of resolving *indels* is suitable considering that estimates from literature [21] show the ratio of *indels* to substitutions to be low.

5.2. Estimation of Channel Parameters

Once $\mathbf{z}$ has been resynchronised, the received watermark $\hat{\mathbf{w}}$ serves an additional purpose: the estimation of the parameters γ and q required for optimum decoding of the LPDC code. To obtain these estimates we undertake maximum likelihood estimation. Using the transition probability matrix we have $p(\mathbf{z}|\mathbf{y};\gamma,q) = \prod_{i=1}^{n} p(z_i|y_i;\gamma,q)$. Since there are only three different probabilities in the Kimura model, taking the natural logarithm we obtain

$$\log p(\mathbf{z}|\mathbf{y};\gamma,q) = n_{\mathrm{tv}} \log\left(\frac{\gamma}{3}\right) + n_{\mathrm{tr}} \log\left(1-\frac{2\gamma}{3}\right) + (n_{\mathrm{tv}} + n_{\mathrm{tr}}) \log\frac{q}{1-q} + n\log(1-q),$$

where n_{tv} and n_{tr} are the number of transversions and transitions that occur between $\mathbf{w}$ and $\hat{\mathbf{w}}$, and the watermark is of length n. Differentiating with respect to γ and q and equating to zero, we respectively obtain

$$\hat{\gamma} = \frac{3}{2(1+\hat{\varepsilon})} \quad , \quad \hat{q} = \frac{n_{\mathrm{tv}} + n_{\mathrm{tr}}}{n} \tag{6}$$

where $\hat{\varepsilon} = n_{\mathrm{tr}}/n_{\mathrm{tv}}$ is the empirical transitions/transversions ratio. It is also possible to obtain an estimate of p; however this is not trivial and $\hat{\Pi}^p$ can be accurately approximated as $\hat{\Pi}$ using $\hat{q}$ and $\hat{\gamma}$ alone.

6. RESULTS

Evaluation of the watermark code with LDPC was undertaken by means of Monte Carlo simulations, with the following parameters: $q = 10^{-5}$, $\gamma = 0.1$ and $\rho = 10^{-6}$. The value of ρ is considerably more extreme than estimated in literature [21], ($\rho = \frac{q}{10}$ was used over $\rho \approx \frac{q}{40}$). This was done purposefully to show how the watermark code performed under strenuous conditions.

Shown in Figure 2 is the probability of bit error (P_e) against the number of generations (p) of a host organism. The perfectly synchronised decoder has the position of *indels* revealed. The repetition code used BioCode ncDNA first to create a quaternary vector which was then repeated three times. Resynchronisation was performed using the heuristic multiple sequence alignment algorithm MUSCLE, followed by majority decoding on the aligned sequences. This resynchronisation procedure was previously proposed by us [22], however unlike here, it did not operate under the no-start-codon constraint. Figure 3 shows the mutual information of the watermark codes in addition to the bound presented in Section 4 and a bound proposed in prior work [20]. The results indicate that the watermark code is very efficient. The perfectly synchronised versions only perform slightly better than their blind counterparts. Furthermore the estimation of q and γ alone produces accurate results.

7. RELATION TO PRIOR WORK

The work presented here proposes a practical application of Davey and MacKay's watermark codes [9]. To the authors' knowledge, there only exists one other practical application for watermark codes [23]. The only previous DNA data embedding method capable of counteracting *indels* was proposed by Yachie et al. [6]. However their solution is based on repetition coding plus realignment, which is suboptimal with respect to the use of watermark codes and LDPC. Furthermore, the watermark code was modified in order ensure biocompatibility with the host organism by precluding the appearance of start codons in the encoding. The only previous DNA data embedding method that approximately implements this important constraint is the one by JCVI [7], but it just minimises the probability of start codons appearing without avoiding them altogether, and without using optimum error correction coding. In fact, we believe that ours is the first application of near-optimum error correction to the problem of DNA data embedding.

8. REFERENCES

[1] 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.

[2] A. Gehani, T.H. LaBean, and J.H. Reif, "DNA-based cryptography," in *5th DIMACS Workshop on DNA Based Computers*, June 1999.

[3] 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, June 2010.

[4] G. M. Church, Y. Gao, and S. Kosuri, "Next-Generation Digital Information Storage in DNA," *Science*, vol. 337, no. 6102, pp. 1628, Sept. 2012.

[5] 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.

[6] 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.

[7] C. A. Hutchison, M. G. Montague, and H. O. Smith, "Encoding text into nucleic acid sequences," US Patent 12/916,344, October 2010.

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

[9] M. C. Davey and D. J. C. MacKay, "Watermark codes: Reliable communication over insertion/deletion channels.," in *Proceedings 2000 IEEE International Symposium on Info. Theory*, 2000, p. 477.

[10] F. Balado, "Capacity of DNA data embedding under substitution mutations," *IEEE Trans. on Information Theory*, vol. 59, pp. 928–941, February 2013.

[11] National Center for Biotechnology Information (NCBI), "Taxonomy tools (genetic codes)," `http://www.ncbi.nlm.nih.gov/taxonomy`.

[12] 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.

[13] The ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, September 2012.

[14] 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.

[15] 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.

[16] D. Haughton and F. Balado, "BioCode: Two biologically compatible algorithms for embedding data in non-coding and coding regions of DNA," (under review), 2012.

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

[18] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. 18, pp. 460–473, 1972.

[19] R. Carrasco and M. Johnston, *Non-Binary Error Control Coding for Wireless Communication and Data Storage*, John Wiley & Sons, 2008.

[20] F. Balado, "On the embedding capacity of DNA strands under substitution, insertion, and deletion mutations," in *Procs. of the SPIE: Media Forensics and Security II*, San Jose, USA, 2010, vol. 7541.

[21] J. Q. Chen, Y. Wu, H. Yang, J. Bergelson, M. Kreitman, and D. Tian, "Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria," *J. Mol. Biol. Evol.*, vol. 26, no. 7, pp. 1523–1531, 2009.

[22] D. Haughton and F. Balado, "Repetition coding as an effective error correction code for information encoded in dna," *Bioinformatic and Bioengineering, IEEE International Symposium on*, vol. 0, pp. 253–260, 2011.

[23] D. J. Coumou and G. Sharma, "Watermark synchronization for feature-based embedding: Application to speech," in *IEEE Intl. Conf. on Multimedia and Expo (ICME)*, July 2006, pp. 849–852.