On the reduction of the cost for encoding/decoding digital images stored on synthetic DNA

Melpomeni DIMOPOULOU[1], Eva GIL SAN ANTONIO[1], Marc ANTONINI[1], Pascal BARBRY[2], Raja APPUSWAMY[3]

[1]Université Côte d'Azur, CNRS, I3S laboratory, 2000, route des Lucioles, Les Algorithmes, bât.Euclide B 06900 Sophia Antipolis, France

[2]IPMC, 660 Route des Lucioles, 06560 Valbonne, France

[3]EURECOM, 450 Route des Chappes, CS 50193 - 06904 Biot Sophia Antipolis cedex, France

dimopoulou@i3s.unice.fr, gilsanan@i3s.unice.fr, am@i3s.unice.fr
barbry@ipmc.cnrs.fr, Raja.Appuswamy@eurecom.fr

Résumé – De nos jours, nous assistons à une explosion du volume des données numériques produites dans le monde entier. La question de leur archivage devient donc de plus en plus cruciale, tant pour pérenniser notre héritage scientifique et culturel que pour permettre leur réutilisation. Ainsi, la recherche de solutions pour un stockage efficace à long terme des données "froides", c'est-à-dire des données peu ou plus utilisées, suscite un intérêt de plus en plus grand. Des études récentes ont prouvé qu'en raison de ses propriétés biologiques, l'ADN peut être un excellent candidat pour le stockage d'informations numériques, permettant également une conservation des données sur le long terme. Cependant, les procédures biologiques de synthèse et de séquençage de l'ADN sont coûteuses, tout en introduisant d'importantes contraintes dans le processus de codage. Dans cet article, nous proposons un schéma de codage adapté au stockage d'images sur ADN, qui respecte les contraintes introduites par les procédures biologiques. D'autre part, la solution proposée permet de réduire les coûts introduits par la synthèse et le séquençage.

Abstract – Living in the age of data explosion, the research of solutions for efficient long term storage of the infrequently used "cold" data is becoming of great interest. Recent studies have proven that due to its biological properties, the DNA is a strong candidate for the storage of digital information allowing also data longevity. However the biological procedures of DNA synthesis and sequencing are expensive while also introducing important restrictions in the encoding process. In this work we present a new constrained encoding method for the robust encoding /decoding of images to be stored into DNA. Furthermore we study the possibility of fully retrieving the stored information using less sequencing samples and consequently reducing the sequencing cost.

1 Introduction

Digital evolution has caused an immersive increase in the amount of data that is being generated and stored. The digital universe is forecast to grow to over 160 zettabytes in 2025. At the same time studies show that after storage, 80% or more of this data might not be needed for months, years, decades, or maybe ever. The rising need for long-term storage for this kind of "cold" data is becoming of great interest. Existing storage systems suggest efficiency in capacity yet lacking in durability. Hard disks, flash, tape or even optical storage have limited lifespan in the range of 5 to 20 years. Interestingly, recent studies have proven that it was possible to use synthetic DNA for the storage of digital data, introducing a strong candidate to achieve data longevity. The DNA's biological properties allow the storage of a great amount of information into an extraordinary small volume while also promising lossless and efficient storage for centuries or even longer. DNA is a complex molecule corresponding to a succession of four types of nucleotides (nts), Adenine (A), Thymine (T), Guanine (G), Cytosine (C). It is this quaternary genetic code that inspired the idea of DNA data storage which suggests that any binary information can be encoded into a DNA sequence of A, T, C, G. The main challenge lies in the restrictions imposed by the biological procedures of DNA synthesis (writing) and sequencing (reading) which are involved in the encoding process and introduce significant errors in the encoded sequence while also being relatively costly (several dollars for writing and reading a small strand of nucleotides). However, encoding digital data onto DNA is not obvious, because when decoding, we have to face the problem of sequencing noise robustness. In [4] there has been a first attempt to store data into DNA while also providing a study of the main causes of biological error. In order to deal with errors previous works in [7] and [3] have suggested dividing the original file into overlapping segments so that each input bit is represented by multiple oligos. However, this procedure introduces extra redundancy and is poorly scalable. Other studies [2],[8] suggest the use of Reed-Solomon code in order to treat the erroneous sequences while in [6] a new robust method of encoding has been proposed to approach the Shannon capacity. Nevertheless, all the above works attempt to encode the data using a binary stream without taking into account the input data's characteristics. In this work we propose a new efficient and robust encoding algorithm especially designed for the DNA storage of

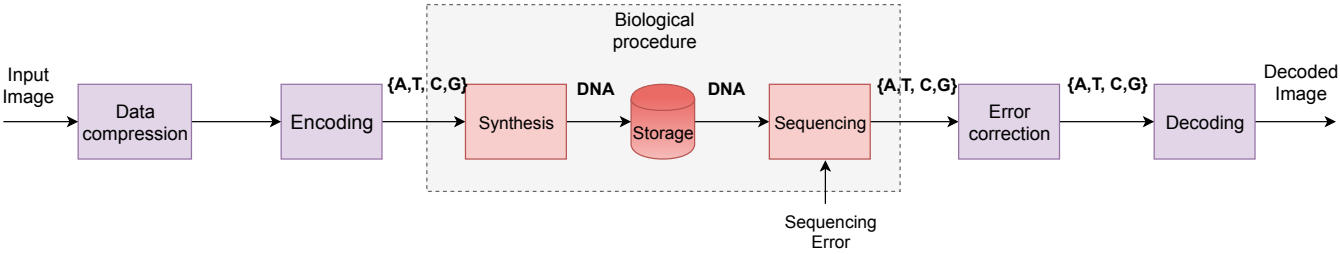

Figure 1 – The general encoding

digital images that we have previously compressed using image compression techniques. This allows to control the compression ratio by the way of an optimal nucleotide allocation, and consequently to control the DNA synthesis cost. Furthermore, we extended our study to minimize the DNA decoding cost by optimizing the amount of synthetic oligonucleotides necessary for a robust sequencing.
The paper is organized as follows. In section 2 we introduce the general encoding process, highlighting the biological restrictions and the proposed solution for creating a robust DNA code to synthesis and sequencing noise. In section 3.2 we propose a study for optimizing the amount of oligonucleotides used for decoding and show some results in section 3. Finally, section 4 concludes the paper and proposes some future works.

2 Encoding images on DNA

2.1 The proposed encoding workflow

The main goal of DNA data storage is the encoding of the input data using a quaternary code composed by the alphabet {A,T,C,G} to be stored into DNA. The general idea of our proposed encoding process is depicted in figure 1 and can be very roughly described by the following steps. Firstly, the input image has to be compressed using a lossy/near lossless image compression algorithm. More specifically we have used a Discrete Wavelet Transform (DWT) with 9-7 filters [1], quantizing each subband independently thanks to a uniform scalar quantizer. Then, the compressed image subbands are encoded into a sequence of A, T, C and G to be later synthesized into DNA. The encoding procedure we developed in our work is described in the next sections. Error in DNA synthesis and yield of production can be well controlled when the oligonucleotide size stays below 150 nts. This implies that the encoded sequence needs to be cut into smaller chunks before it is synthesized into DNA and stored into special small storage capsules. DNA sequencing is the process of retrieving the stored information by reading the content of the stored oligos. Unfortunately this procedure is very error prone causing errors like substitutions, insertions or deletions of nucleotides. In order to deal with such errors, before sequencing, the stored data is cloned into many copies using a biological process called Polymerace Chain Reaction (PCR) amplification. In addition to this, during the sequencing, next generation sequencers (NGS) like Illumina use the method of bridge amplification (BA) for reading the oligos. As exlained in [9], BA is a process similar to PCR, which allows nucleotide recognition while producing many copies of each oligo introducing extra redundancy that is necessary for the reduction of the sequencing error. As a result, the data provided by the sequencer is multiple copies of each input oligo that may contain errors. This implies the need of choosing the good oligo copies before we reconstruct the initial sequence, concatenating the selected oligos and decoding them to get back the stored image.

2.2 Biological restrictions

The encoding of digital data into DNA is highly restricted by constraints imposed by the biological procedures of the DNA synthesis and sequencing. To begin with, DNA synthesis is an error free process when the sequence to be synthesized into DNA does not exceed the length of 150 nts. As the sequences get longer the synthesis error increases exponentially. This yields the need for cutting the encoded data into smaller chuncks so to reduce the synthesis error. Consequently, in order to achieve reconstruction, special headers (noted H in figure 2) need to be inserted into each of the encoded chuncks, containing information about the position of these data into the initial encoded sequence. This formatting of the data is described by figure 2. The DNA sequencing is the most challenging part of the encoding procedure as it may introduce high percentage of errors like insertion, deletion or substitution of nucleotides. These errors can be reduced if the encoding algorithm respects some particular rules in order to avoid ill-cases which can lead to wrong recognition of nucleotides during the sequencing. The biological restrictions are the following:

- **Homopolymers:** Consecutive occurencies of the same nucleotides should be avoided.
- **G, C content:** The percentage of G and C in the oligos should be lower or equal to the one of A and T.
- **Pattern repetitions:** The codewords used to encode the oligos should not be repeated forming the same pattern throughout the oligo length.

2.3 A biologically constrained quaternary code

In [5] we proposed a quaternary encoding algorithm which takes into consideration all of the encoding restrictions described in 2.2. Let $Q(x) = f(\alpha(x))$ be the quantized values $\hat{x}^i \in \Sigma$

Figure 2 – Format of the oligos - S denotes the sense nucleotide which determines whether a strand is reverse complemented when sequenced. P is a parity check nucleotide while the ID is an identifier of the image so to be distinguished from other data that may be stored. H is a specific header, Offset specifies the chunck's data position in the encoded sequence and Payload contains the encoded chunks.

produced by the quantizer with $i \in \{1, ..., k\}$ and Σ the quantization codebook of size k. f is called the decoding function and $\alpha(x) = i$, $i \in \{1, ..., k\}$, the encoding function providing the index of the quantization levels. In order to generate a DNA code Γ we introduce two separate alphabets:

- $\mathcal{D}_1 = \{AT, AC, AG, TA, TC, TG, CA, CT, GA, GT\}$
- $\mathcal{D}_2 = \{A, T, C, G\}$

$\mathcal{D}_1$ is an alphabet composed by concatenations of two symbols from $\mathcal{D}_2$ selected in such a way that no homopolymers or high GC content is created. In order to encode the quantized sequence onto DNA we define the code Γ as the application: $\Gamma : \Sigma \rightarrow \mathcal{D}^\star$ where $\mathcal{D}^\star$ is a dictionary composed by $L \geq 2k$ codewords c_i of length l. We denote $\Gamma(\hat{x}^i) = c_i$ the codeword associated with the quantized value $\hat{x}^i \in \Sigma$. $\mathcal{D}^\star$ is constructed by all the possible concatenations of symbols from $\mathcal{D}_1$ and $\mathcal{D}_2$ according to the following rules:

- Codes with codewords of an even length (l even) are being constructed selecting $\frac{l}{2}$ doublets from $\mathcal{D}_1$,
- Codes with codewords of an odd length (l odd) are being constructed selecting $\frac{l-1}{2}$ doublets from $\mathcal{D}_1$ and a single symbol from $\mathcal{D}_2$.

The quantization using big values of quantization step-size q can lead to long repetitions of the same quantized value. The use of existing algorithms for the encoding of such a sequence into DNA would thus create pattern repetitions. In order to avoid patterns, our algorithm uses a pseudorandom mapping which associates a quantized value to more than one possible codewords. More precisely our algorithm maps the index of levels of quantization i to the codewords of $\mathcal{D}^*$ as described in figure 3. The code Γ is constructed so that each quantized value in Σ is mapped to a set of different non-empty quaternary codewords in $\mathcal{D}^\star$ following a one-to-many relation in such a way that it is uniquely decodable. Since we ensure $L \geq 2k$, the pseudorandom mapping can at least provides two possible codewords for one input symbol. More precisely, the mapping is described by the following steps (interested reader can find the complete algorithm in our paper [5]):

1. Build the corresponding code $\mathcal{D}^*$ of size L using all possible codewords of length l which can be built following the two rules described previously,
2. Compute the number of times m that k can be replicated into the total size L of the code $\mathcal{D}^*$: $m = \lfloor \frac{L}{k} \rfloor$,
3. The mapping of the quantized value $\hat{x}_i$ to a codeword c_i is given by: $\Gamma(\hat{x}^i) = \mathcal{D}^*(i + rand(0, m-1) * k)$.

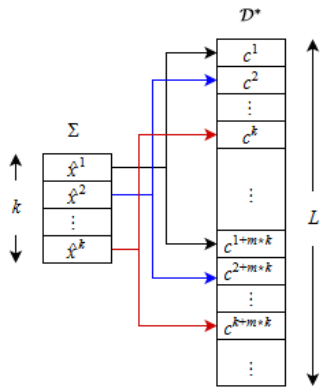

Figure 3 – Mapping the quantized values from codebook Σ to codewords of $\mathcal{D}^*$

3 Experimental results

3.1 DNA synthesis and sequencing

In this study we have carried out a real biological experiment for storing a Lena image of size 128 by 128 pixels into DNA. The choice of the size of the image was constrained by the high expenses of the biological procedures involved in the experiment. More specifically we have synthesized 662 oligos into DNA, of size 138 nucleotides (nts) each (including primers). This number of oligos corresponds to a compression ratio equal to 2.68 bits/nt. For the sequencing, we used the Illumina Next Seq machine, a sequencer which produces low percentage of sequencing error. This machine uses PCR amplification and Bridge Amplification which create many copies of the oligonucleotides in order to ensure an accurate sequencing (see section 2). However, although the robust code we proposed in section 2.3 can reduce the sequencing error, it will not eliminate it. This means that after the sequencing process, some oligonucleotides will contain errors due to wrong nucleotide recognition. As the Illumina sequencers do not introduce high noise percentage in the sequenced data we assume that only a minority of those copies will finally be affected by error. Thus, the most frequent oligonucleotides after sequencing are assumed to be the correct ones.

3.2 Subsampling the sequenced oligos

As mentioned in previous sections the DNA synthesis and sequencing are expensive and can cost several thousands of dollars depending on the size of the encoded data. On one hand, in order to reduce the synthesis cost, we compress the input image and control the coding rate thanks to a nucleotide-allocation algorithm. By doing so, one can select an optimal rate which shows no significant distortion in the visual result. On the other hand, the sequencing cost can be decreased by reducing PCR and BA cycles performed during the sequencing. This can be simulated by subsampling the data set of sequenced oligos provided by our experiment. In the initial experiment we discarded the oligos exceeding 91 nts as those oligos for the moment can not be decodable. This resulted to a total number of 28,876,259 sequenced oligos, from which the initial encoded image should

#samples = 1500
$PSNR = 13.05dB$
$SSIM = 0.59$

#samples = 2250
$PSNR = 28.12dB$
$SSIM = 0.95$

#samples = 2500
$PSNR = 40.79dB$
$SSIM = 0.97$

#samples = 3500
$PSNR = 47.53dB$
$SSIM = 0.99$

#samples = 5000
$PSNR = inf$
$SSIM = 1$

Figure 4 – Visual results after decoding at different subsampling rates. Lena 128x128 pixels

be reconstructed. In this experiment we have subsampled this initial number of oligos using different sampling sizes. For each case, the most frequent oligos from the subsampled data were selected as the most representative sequences and were used to decode and reconstruct the image.

Consequently, for the decoding, we have subsampled the sequenced data and reconstructed an image for each sampling size. The procedure was repeated twenty times for each sampling size and we computed the average of the Peak Signal to Noise Ratio (PSNR) as well as the percentage of exact matches when comparing the most frequent sequenced oligos with the encoded correct ones. Those results are presented in figure 5 and provide information about the quality of the reconstruction. In figure 4 we present the visual results for different sampling rates as also the corresponding values of PSNR and SSIM.

Interestingly enough we observe that we can achieve a perfect reconstruction by using only a small percentage of the sequenced oligos provided by our experiment. More precisely we can see in figure 5 that only using 5000 samples of the sequenced oligos we get 100% correctness when comparing to the 662 original synthesised oligos. Taking into account the initial number of amplified data (28,876,259 oligos) we conclude that only 0.0173% of those oligos are needed for perfect decoding. This can be confirmed by the evolution of the visual results in figure 4.

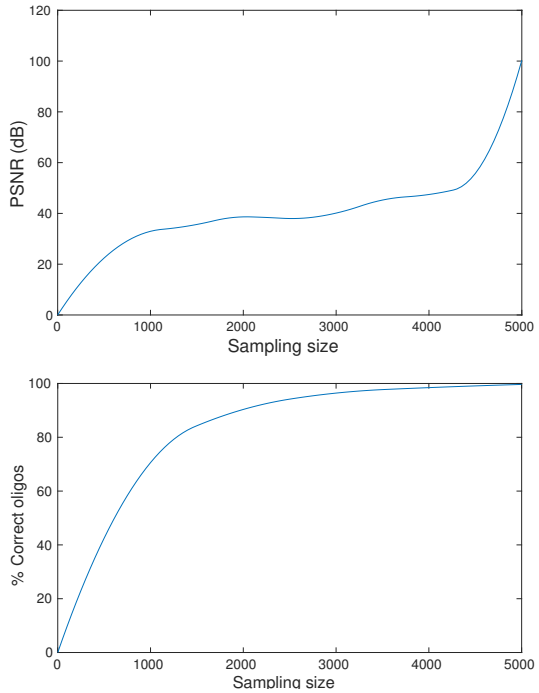

Figure 5 – The evolution PSNR and percentage of correct oligos for different sampling sizes

4 Conclusion

With this experiment we have shown that the cost of DNA synthesis can be controlled by compressing the input image to a given rate without causing significant visual distortion. Furthermore we have proven that the sequencing cost can be highly reduced as one can achieve a perfect reconstruction using a very small percentage of the sequenced data that has been used in our latest experiment. This fact is very important as we hope that lowering the expenses of such experiments will make the DNA data storage more popular to the public making a great step through in the field of data storage.

References

[1] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies. Image coding using wavelet transform. *IEEE Transactions on image processing*, 1(2):205–220, 1992.

[2] M. Blawat, K. Gaedke, I. Huetter, X.-M. Chen, B. Turczyk, S. Inverso, B. W. Pruitt, and G. M. Church. Forward error correction for dna data storage. *Procedia Computer Science*, 80:1011–1022, 2016.

[3] J. Bornholt, R. Lopez, D. M. Carmean, L. Ceze, G. Seelig, and K. Strauss. A dna-based archival storage system. *ACM SIGOPS Operating Systems Review*, 50(2):637–649, 2016.

[4] G. M. Church, Y. Gao, and S. Kosuri. Next-generation digital information storage in dna. *Science*, page 1226355, 2012.

[5] M. Dimopoulou, M. Antonini, P. Barbry, and R. Appuswamy. A biologically constrained encoding solution for long-term storage of images onto synthetic dna. In *Submitted to EUSIPCO*, 2019.

[6] Y. Erlich and D. Zielinski. Capacity-approaching dna storage. *bioRxiv*, page 074237, 2016.

[7] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized dna. *Nature*, 494(7435):77, 2013.

[8] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark. Robust chemical preservation of digital information on dna in silica with error-correcting codes. *Angewandte Chemie International Edition*, 54(8):2552–2555, 2015.

[9] I. Illumina. An introduction to next-generation sequencing technology. 2015.