A SEQUENCING NOISE RESISTANT CODE MAPPING ALGORITHM FOR IMAGE STORAGE IN DNA

Melpomeni Dimopoulou, Eva Gil San Antonio, Marc Antonini

A SEQUENCING NOISE RESISTANT CODE MAPPING ALGORITHM FOR IMAGE STORAGE IN DNA

Melpomeni Dimopoulou, Eva Gil San Antonio, Marc Antonini
Université Côte d'Azur, CNRS, Laboratoire I3S

Abstract : *The continuous exponential increase in the generation of digital information is becoming inconsistent with the capacity and longevity limitations imposed by conventional storage devices which can't be reliable for more than 10-20 years. More precisely, 90% of the data on the internet has been only generated in the last 2 years while 80% of this information consists of "cold" data which is very rarely or never accessed but still needs to be stored in off-line back-up drives for security and compliance reasons. To ensure data reliability data centers are nowadays purging metric tons of hardware for the frequent replacement of those drives which is extremely expensive both in terms of money and energy. To handle this problem scientists have recently proposed the use of synthetic DNA as a means of digital data storage. This idea is inspired by the biological properties of the DNA molecule which contains all the necessary information for living organisms to survive, stored in a very limited volume such as the cells' nucleus. Furthermore, when stored under specific conditions, DNA can be decodable without loss of information for hundreds of years. In this work we propose an algorithm which optimally maps input quantization vectors to DNA codewords for the storage of quantized images into DNA.*

Keywords : DNA data storage, Image coding, Robust encoding, Vector Quantization.

1 Introduction

DNA data storage is a very promising yet challenging procedure as it requires the use of the delicate biological processes of DNA synthesis (writing) and sequencing (reading). More precisely DNA sequencing imposes some important restrictions in the encoding workflow as it can introduce errors in the decoded DNA sequence. In [2], there has been a first attempt to store data into DNA while also providing a study of the two main causes of this sequencing error. An additional restriction has been later included by a biological study in [7]. More precisely it has been noted that in order to reduce the sequencing error the encoding algorithms should respect the three following rules. 1) No repetitions of the same symbol more than 3 times (homopolymer rule). 2) The percentage of C and G should be lower or equal to the percentage of A and T. 3) Short repeated patterns should be avoided. In order to deal with errors, some studies in [1] and [6] suggest the use of Reed-Solomon codes in order to treat the erroneous sequences. In [5] we have proposed a new constrained fixed length algorithm for the encoding of quantized images into DNA. This work has been extended in the use of Vector Quantization (VQ) in [4]. However, respecting the rules imposed by the sequencing does not guarantee an error free decoding. This is because some sequencers like the Nanopore sequencer introduce very high error rates which can't be avoided. One can imagine the sequencing noise as the one introduced by noisy channels in telecommunications. In [3] there has been proposed an interesting algorithm for optimally assigning input values to binary words to achieve better resistance to the channel noise. Inspired by the work proposed in [3], we extend this algorithm to a quaternary representation for the mapping of input symbols to DNA codewords which aims to create an encoding that is more resistant to the sequencing noise. This algorithm uses a more sophisticated method for the mapping of input symbols to the different DNA codewords so that in case of an error the erroneous decoded symbol will be closer to the original one.

2 Proposed work

The work presented in this paper is an extension of the method proposed in [3]. The goal is finding an optimal mapping between input vectors obtained by a VQ algorithm and quaternary codewords so as to achieve resistance to sequencing errors. VQ is useful for the efficient compression of an image before it is stored into DNA to reduce the synthesis cost which can be relatively high. The purpose of the proposed algorithm is the mapping of close (in terms of Euclidean distance) quantization vectors $v_i, i \in [1, \ldots, M]$ from a codebook V to codewords from a code W which have a small Hamming distance. The idea behind this mapping lies in the fact that in case of an error during sequencing and assuming that the sequencing noise rate is small enough a correct codeword will be transformed to another one which will have a small Hamming distance with the correct one. The algorithm can be very roughly described by the following parts:

For each codeword : Create a sphere $H(w_i)$ containing the B_i codewords which have a Hamming distance of 1 compared to $w_i, i \in \{1, \ldots, L\}$. Define $B = max_i(B_i), i \in \{1, \ldots, L\}$.

- For each input vector v_i: Find a set $S(v_i)$ of B neighboring vectors v_l which are the closest to v_i in terms of Euclidean distance $d(v_i, v_l)$.
- For each input vector : Compute the empirical function $F(v_i) = \frac{p(v_i)}{\alpha^\beta(v_i)}$ where $p(v_i)$ is the probability of in the input sequence, and $\alpha(v_i) = \sum_{j|v_j \in S(v_j)} d(v_j, v_i)$ with $\beta \geq 0$ a trade-off parameter Progressively perform assignment of vectors v_i to codewords w_i such that vectors with a bigger $F(v_i)$ as well as its neighboring vectors $v_l \in S(v_i)$ will be assigned to the same

sphere of codewords $H(w_i)$ whenever possible as depicted in figure 1. If this is not possible assignment is performed such that vectors v_i are assigned to codewords with a small Hamming distance from the codewords already assigned to their neighboring vectors. The algorithm for this step is relatively complicated and thus for further information readers can refer to [3].

- Optimization of the first assignment:
  - Exchange the previously mapped codewords between each pair of vectors.
  - For each exchange check if the average distortion has decreased. If true keep this change, else keep the initial state of mapping.

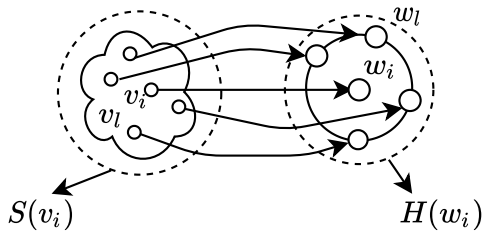

Figure 1: Assuming a vector v_i, the set $S(v_i)$ contains the B closest to v_i vectors v_l in terms of Euclidean distance. Then given a codeword w_i, the Hamming sphere $H(w_i)$ of radius 1 contains all possible codewords w_l with $i \in \{1, \ldots, B_q\}$ for which $d_H(w_i, w_l) = 1$. An optimal case of mapping would be the one where all vectors that belong to the same neighborhood $S(v_i)$ are assigned to the same sphere $H(w_i)$. However as this mapping is not possible for all the words $w_i \in \mathcal{C}^*$ we search for a solution that globally optimizes the assignment such that close codevectors are mapped to close codewords.

3 Results

For the experiments we quantized an image of 512x512 pixels using VQ with 100 vectors of length $n = 2$. We then decoded the image adding 3% of noise. This noise ratio is equal to the estimated percentage of noise added by the Nanopore sequencer. Figure 2 depicts the visual quality of the noisy decoding without optimal mapping while figure 3 represents the noisy decoding for the case where optimal mapping was used. The visual quality has improved significantly providing a gain of 3 dB in the PSNR.

Without optimal mapping PSNR=20.39 dB

With optimal mapping PSNR=23.4 dB

Figure 2: Visual result of substitution noise for the two different ways of mapping.

4 Conclusion and perspectives

In this work we proposed a new mapping algorithm for the optimal assignment of quantization vectors obtained by VQ quantization to DNA codewords. The obtained results are very promising encouraging us to further study and improve this approach.

References

[1] Meinolf Blawat, Klaus Gaedke, Ingo Huetter, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin W Pruitt, and George M Church. Forward error correction for DNA data storage. *Procedia Computer Science*, 80:1011–1022, 2016.

[2] George M Church, Yuan Gao, and Sriram Kosuri. Next-generation digital information storage in DNA. *Science*, page 1226355, 2012.

[3] JR Boisson De Marca, NS Jayant, et al. An algorithm for assigning binary indices to the codevectors of a multi-dimensional quantizer. In *1987 IEEE International Conference on Communications (ICC'87)*, pages 1128–1132. , 1987.

[4] Melpomeni Dimopoulou and Marc Antonini. Image storage in DNA using Vector Quantization. In *EUSIPCO 2020*, Amsterdam, Netherlands, January 2021.

[5] Melpomeni Dimopoulou, Marc Antonini, Pascal Barbry, and Raja Appuswamy. A biologically constrained encoding solution for long-term storage of images onto synthetic DNA. In *EUSIPCO 2019*, 2019.

[6] Robert N Grass, Reinhard Heckel, Michela Puddu, Daniela Paunescu, and Wendelin 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.

[7] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1):36, 2012.