Titre :

DNA CODING FOR IMAGE STORAGE USING IMAGE COMPRESSION TECHNIQUES

Auteurs :

Melpomeni Dimopoulou [1] dimopoulou@i3s.unice.fr
Marc Antonini [1] am@i3s.unice.fr
Pascal Barbry [2] barbry@ipmc.cnrs.fr
Raja Appuswamy [3] Raja.Appuswamy@eurecom.fr

[1] I3S, 2000, Route des Lucioles, 06900, Sophia Antipolis, France
[2] IPMC, 660 Route des Lucioles, 06560, Sophia Antipolis, France
[3] EURECOM, 450 Route des Chappes, CS 50193 - 06904 Sophia Antipolis, France

Résumé (*100-200 mots*) :

Living in the age of the digital media explosion the urge for finding new efficient methods of data storage increases significantly. Existing storage devices such as hard disks, flash, tape or even optical storage have limited durability in the range of 5 to 20 years. Recent studies have proven that the method of DNA data storage introduces a strong candidate to achieve data longevity. The DNA's biological properties permit the compression of a great amount of information into an extraordinary small volume while also promising efficient data storage for millions of years with no loss of information. This work proposes a new encoding scheme especially designed for the encoding of still images, extending the existing algorithms of DNA data storage by introducing image compression techniques.

Mots-clefs (*3 à 5 max.*) :

DNA data storage, wavelet decomposition, data compression, error correction

Contexte et état de l'art :

DNA data storage is the procedure of encoding any binary information into a quaternary code of A, T, C, G, the symbols that correspond to the 4 different types of nucleotides (nts), the DNA's main building blocks. The selection of the encoding algorithm is strongly restricted by the biological procedures involved in the encoding scheme in figure [1]. DNA synthesis is the procedure of chemically producing chunks of nucleotides (oligos) corresponding to the desired data. This process is almost error-free with an error probability lower than 0.1% for DNA sequences no longer than 100 nts. The DNA sequencing is the procedure of reading the stored oligos and recover the initial quaternary encoded sequence that has been formed into DNA. However, the biological procedure of sequencing is a significantly error-prone procedure. G. Church et al. in [1] have studied the main causes of the biological error while in [2] Goldman et. al proposed an efficient encoding procedure to prevent this kind of errors as much as possible and introduced error correction. In addition to error-correction algorithms, sequencing machines use the method of PCR amplification creating many copies of each oligo introducing redundancy to reduce the biological error as much as possible. One can imagine the procedure of PCR amplification like the classical repetition coding used for transmission over a noisy channel that may corrupt the transmission in various positions. As in repetition coding, the main idea of PCR amplification is to repeat the oligos several times hoping that the sequencing corrupts only a minority of these repetitions. Under this simple hypothesis, we assume that after all the copies of oligos are sequenced, one can apply the method of majority vote to distinguish the most representative oligos. In other words, under the main assumption that the oligos that appear more times than the others will most probably be error-free, during the reconstruction we select the oligos with the highest frequency of appearance.

Travail proposé (*Décrire l'objectif du travail et indiquer clairement le problème/défi technique étudié*) :

In this work we have built an encoder specifically designed according to the needs of the biological procedures of synthesis and sequencing. First, we compress the input image using image compression techniques. More precisely, we use the Discrete Wavelet Transform (DWT) to decompose the image into 3 levels producing 10 wavelet subbands. Then we encode each subband separately using uniform quantization for compression. As a next step we need to encode the quantized coefficients into a quaternary code of A, T, C, G respecting four basic restrictions imposed by the biological procedures. To begin with, the sequence should not contain repetitions of the same nucleotide more than 3 times (homopolymers). Additionally, the encoded sequence should not contain repetition of the same codewords and the percentage of G, and C should be lower or equal to the percentage of A and T. This means that the succession of symbols should be as random as possible so to prevent errors. Finally, as the synthesis error increases exponentially with the increase in the sequence length, the encoded sequence to be sent for synthesis should not exceed the 100nts. This last restriction imposes the need for cutting the encoded subband sequence into smaller chunks of 91 nts and adding a header section in each one of them that will contain information about their order in the initial sequence. After the encoding procedure the encoded oligos are biologically synthesized into synthetic DNA which is stored in a small tube depicted in image [1]. In the decoding part the stored oligos are being replicated in many copies before they are read with the sequencing procedure which further adds sequencing noise. As a last step we choose the most representative oligos and we reconstruct the initial image following the inverse procedure. In this first experiment we tested two different cases of reconstruction. One by using the most frequent oligos and one choosing the oligos randomly. This experiment produced 662 oligos compressing the original image at a bit rate of 1.96 nts/pixel (equivalently 4 bits/nt) and a PSNR of 32.51 dB. The results are presented in figures [2] and [3] respectively.

Résultats :

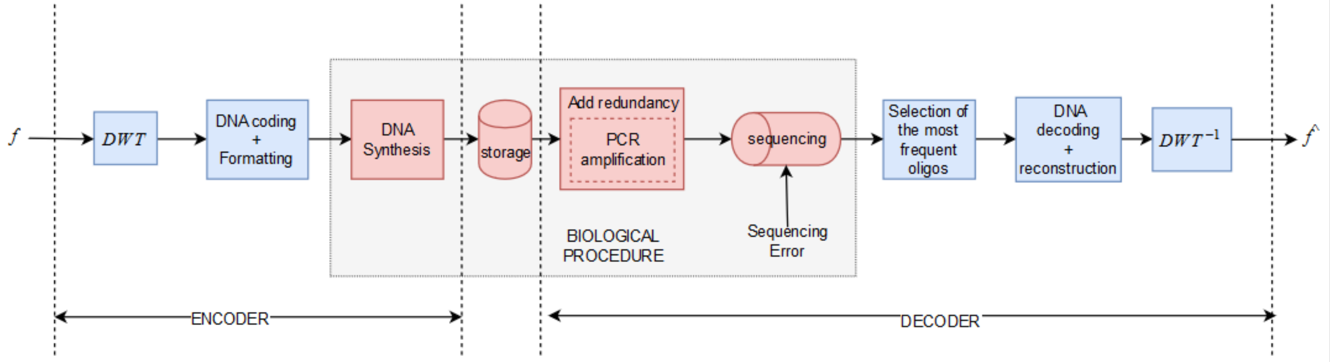

Figure 1: The encoding scheme. The initial image is being decomposed into wavelet subbands and each one of them is encoded into a quaternary code and cut into smaller chunks. The encoded sequence is synthesized into DNA and stored in a tube. PCR amplification produces many copies of each oligo creating redundancy. The copies are being read through the sequencer which introduces sequencing error. The most representative oligos are then selected to decode and reconstruct the encoded sequence and using inverse DWT an estimation of the initial image is produced.

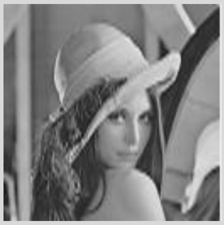

Figure 2 (left): The reconstructed image using the most frequent oligos. (No sequencing noise)

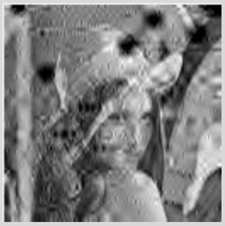

Figure 3 (left): The reconstructed image using random selection of oligos.

Image 1 (left): The storage tube containing 103 ng of synthetic DNA

Conclusion et perspectives :

Given the results of figures [2] and [3] one can clearly conclude that choosing the most frequent oligos from the different copies provided by the sequencing, it is more probable to achieve the best possible reconstruction of the image. In this experiment we have managed to get a reconstruction without any sequencing noise. This very first and simple attempt to introduce image compression techniques into DNA data storage has provided very promising results. Future experiments using error correction techniques and introducing more robustness to sequence error, will further improve the compression ratio and the quality of the current results.

Références (*5 max.*) :

[1] Church, G.M., Gao, Y. and Kosuri, S., 2012. Next-generation digital information storage in DNA. Science, p.1226355.

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