Robust image coding on synthetic DNA: Reducing sequencing noise with inpainting

Eva Gil San Antonio Mattia Piretti Melpomeni Dimopoulou Marc Antonini

Université Côte d'Azur, CNRS, I3s, France

THE CHALLENGE:

- Data explosion → 80% of the worldwide data is rarely accessed ("cold")
- Conventional storage devices have a limited life-span (15-20 years)

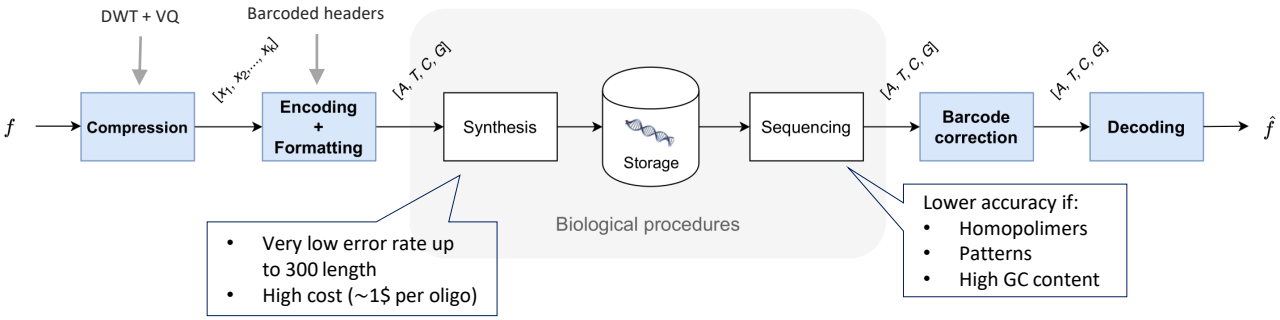
DID YOU KNOW THAT DNA…?

- Allows the storage of information at a high density ($500\ GB/mm^3$)
- Allows to preserve any information for hundreds of years without any loss if stored under the optimal conditions.

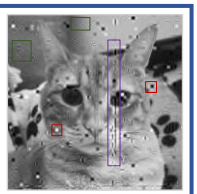
OUR CONTRIBUTION

Despite being a promising solution, **DNA data storage** faces two major obstacles: the large cost of synthesis and the high error rate introduced during **nanopore sequencing**. While most of the works focus on adding redundancy for error correction, this work combines noise resistance to minimize the impact of the errors in the decoded data and post-processing based on **Texture Synthesis** to further improve the quality of the decoding.

Proposed workflow for DNA data storage:



UNLIKE IN MOST INPAINTING SCENARIOS…



- DNA noise involves substitutions, insertions and deletions of nucletides
- Noise applied on each subband separately → Heterogeneous damage
- Large spots (Low freq. subbands), patterns (High freq. subbands) and lines (lost DNA strands due to errors in the header)

AUTOMATIC DETECTION OF THE DAMAGED AREAS

2-step algorithm (performed in each subband separately):

1) Detection of errors in single pixels (substitutions) → *Deviation of each pixel and its neighbours*

τ_1 : phase 1 threshold
N_p: neighborhood of the pixel p
I: damaged image
O_1: Output mask phase 1

$$\forall\, p_{(x,y)} \in I, S_p = \sigma(N_p)$$

$$\forall p_{(x,y)} \in I, \text{ if } \frac{\sqrt{(p_{(x,y)} - \overline{N_p})^2}}{S_p} \begin{cases} \geq \tau_1, O_{1(x,y)} = 1 \\ \leq \tau_1, O_{1(x,y)} = 0 \end{cases}$$

2) Detection of damaged neighbourhoods (indels) → *Internal variance of the neighbourhoods*

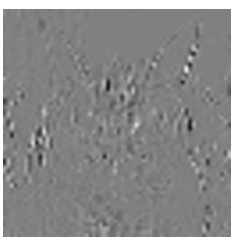
τ_2 : phase 2 threshold
$S = \sigma(I)$
O_2: Output mask phase 2

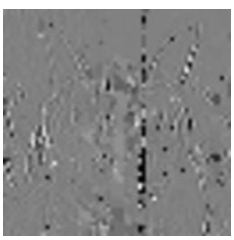
$$\forall p_{(x,y)} \in I, \text{ if } \frac{S_p}{S} \begin{cases} \geq \tau_2, O_{2(x,y)} = 1 \\ \leq \tau_2, O_{2(x,y)} = 0 \end{cases}$$

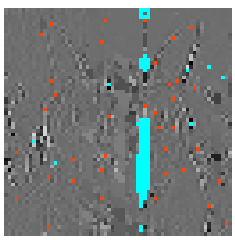
The result of the 2-step algorithm is a binary mask. The detected damaged regions are then corrected using Texture Synthesis*

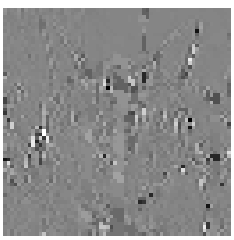
*A. Criminisi, P. Perez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," IEEE Transactions on Image Processing, vol. 13, no. 9, pp. 1200–1212, 2004.

Visual results on a DWT subband


Original subband


Noisy subband after sequencing noise


Detected damaged areas (1st step in red, 2nd step in blue),


Inpainted subband

THE RESULTS

Simulation of sequencing noise*

- Substitution and indel rates:
  - 2.3% deletions
  - 1.01% insertions
  - 1.5% substitutions
- 80% of the noise concentrated in the first and last 20nt of each oligo
- 200 noisy copies of each input oligo
- Headers encoded using barcodes
- Consensus based on majority voting

*Nanopore noise rates adapted from bibliography


Quantized image without sequencing noise
Compression ratio = 4.9708 bits/nt
PSNR = 48.12 dB
SSIM = 0.991


Visual impact of the sequencing noise in the decoded image
PSNR = 36.2 dB
SSIM = 0.92


Post-processed image
PSNR = 38.7 dB
SSIM = 0.94

CONCLUSIONS

While most works up to the date focus on Illumina sequencing due to its higher accuracy, we introduce nanopore sequencing in our workflow, attempting to speed up the process as well as reduce its cost. The results of this study are very promising given the high error-rates imposed by the nanopore sequencers showing significant visual improvement. However, since this study provides results on simulated nanopore noise, a wet-lab experiment is a priority future step to verify in practice the efficiency of the proposed encoding.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863320.