Hierarchical Encoding of JPEG2000-compressed Images for DNA data storage

Xavier Pic [1] Raja Appuswamy [1]

[1]Data Science Department, EURECOM, 450 Rte des Chappes, 06410 Biot, France

DNA and the data storage problematic

Current data storage media face challenges

- Demand is too high for the offer (with the current trend, data will have to be removed from storage)
- Limited life time (a magnetic tape last for about 10 years)
- High energy costs (data centers are consuming a large amount of energy
- Low storage density

DNA appears as a very promising candidate:

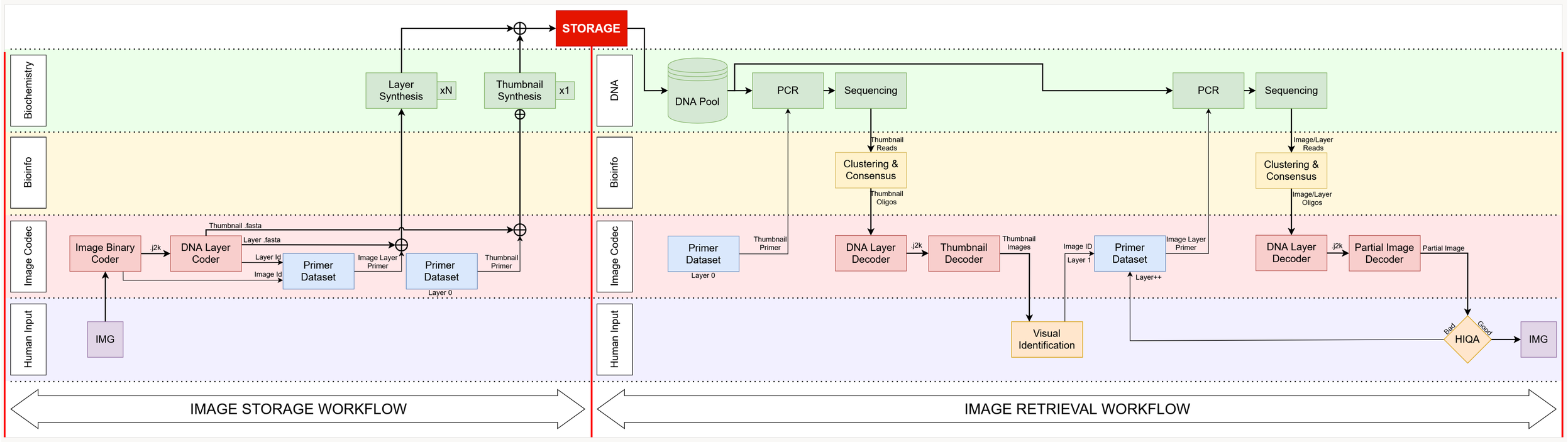
- Very high storage density (1EB/mm3)
- Very **low energy consumption** (a molecule can be stored without energy consumption)
- Very **long lifespan**

Encoding workflow

- **Image encoding:** The image is encoded with JPEG 2000 in progressive mode.
- **Bitstream Slicing:** The bistream is sliced into smaller binary files corresponding to each resolution layer.
- **DNA Coding:** The slices are encoded into DNA using the JPEG DNA VM software.
- **Primer preparation:** Each DNA file is concatenated with two pairs of primers, one indicating the image, the other the resolution layer.
- **Synthesis:** Each file is synthesized seperately.
- **Storage:** The oligos are stored in a common pool.

Figure 1. Structure of the oligos

Progressive image storage and retrieval workflow



Performance metrics: three read costs

This contribution aims at **reducing the read cost** necessary to retrieve an image from a dataset, that can be expressed in three manners in our case:

$$R_c(I,K) = \frac{\sum_{i=0}^{N_{images}} \sum_{k=0}^{N_{levels}} nucs(i,k)}{input\_image\_pixels} \quad (1)$$

$$R_{c\_pd}(I,K) = \frac{\sum_{i=0}^{N_{images}} \sum_{k=0}^{K} nucs(i,k)}{input\_image\_pixels} \quad (2)$$

$$R_{c\_ra}(I,K) = \frac{\sum_{i=0}^{N_{images}} nucs(i,0) + \sum_{k=1}^{K} nucs(I,k)}{input\_image\_pixels} \quad (3)$$

1. **Basic read cost**: All oligos are fully sequenced and decoded.
2. **Progressive read cost**: All oligos are sequenced and decoded up to a selected resolution.
3. **Progressive read cost with random access**: The oligos to be sequenced and decoded are: First, all the thumbnail oligos and then, only the layer oligos of the desired image.

Gain definition

We can define a read-cost gain as in two different ways, depending on whether or not Random Access is enabled (in both cases, Progressive decoding is used) as:

$$G_{pd}(I,K) = \frac{R_c(I,K)}{R_{c\_pd}(I,K)} \quad (4)$$

$$G_{ra}(I,K) = \frac{R_c(I,K)}{R_{c\_ra}(I,K)} \quad (5)$$

Progressive decoding

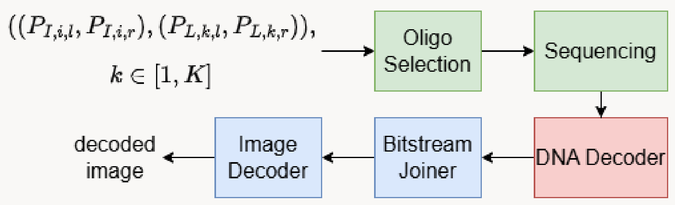


Figure 2. Progressive decoding process

The oligos related to each resolution are successively selected trough PCR, with the help of the layer primers. The decoding process then decodes this specific layer and refines the decoded image with it.

Oligo retrieval process

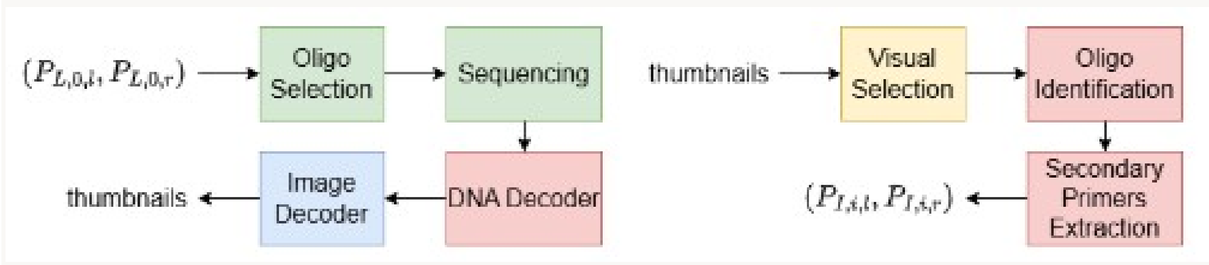


Figure 3. Random Access proposed solution through Thumbnail Retrieval

The only set of primers to be know before the decoding process are the resolution layer primers. The process is as follows:

- **Retrieval of the thumbnails:**
  - Visually identify the correct thumbnail
  - Deduce the primer identifying the desired image
  - Pull and augment all the oligos with this primer through a PCR run
- **Retrieval of the layer oligos:** With the help of the preset primers from each layer, the oligos related to each layer are progressively augmented, and sequenced

Summary

We designed a progressive image coder adapted to DNA, that helps tackling one of the main challenges inherent to DNA data storage: the sequencing costs. We evaluated the performance of such a coder and compared it to versions that were not using the random access capabilities of our solution. We obtained substantial gains on the read cost.

Results

The proposed method leverages **gains** on the read-cost both through **progressive decoding** and **random access**.
The proposed soluttion can improve the sequencing and decoding time by multiples, especially with the random acess system enabled.

	Layer	L_0	L_1	L_2	L_3	L_4
Progressive Decoding	# Oligos	2878	8539	25288	69358	151958
	Theoretical G_{pd} Read-cost gain	52.8	17.8	6.01	2.19	1
Random Access & Progressive Decoding	# Oligos	2878	3114	3812	5648	9090
	Theoretical G_{ra} Read-cost gain	52.8	48.8	39.9	27.1	17.0

Table 1. Theoretical and observed average read-cost gains G_{pd} and G_{ra} for each target resolution level, averaged over all the images of the kodak dataset. The level L_0 refers to the thumbnail while the level L_4 refers to the full image.

References

Taubman, D. S. and M. W. Marcellin (n.d.). "JPEG2000 Image Compression Fundamentals, Standards and Practice". In: *Springer* ().

Lazzarotto, D., J. E. Ramos, M. Testolina, and T. Ebrahimi (2023). "Technical description of the EPFL submission to the JPEG DNA CfP". In: *arXiv preprint*. DOI: arXiv:2312.00560.

Grass, R. N., R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark (2015). "Robust chemical preservation of digital information on DNA in silica with error-correcting codes". In: *Angewandte Chemie International Edition*.

Church, G. M., Y. Gao, and S. Kosuri (2012). "Next-Generation Digital Information Storage in DNA". In: *Science* 337.6102, pp. 1628–1628.