ADAPTIVE SAMPLING FOR STORAGE OF PROGRESSIVE IMAGES ON DNA

Xavier Pic*, Nimesh Pinnamaneni**, Raja Appuswamy*

* EURECOM, Data Science Department, Sophia Antipolis, France
** Helixworks Technologies, Ltd., MTU Campus, Bishopstown, Cork, Ireland

ABSTRACT

Traditional storage media cannot keep up with the rapidly growing demands in data storage. DNA molecules offer a dense, durable and low-energy alternative for storing data. Prior work on DNA storage has developed efficient coding and error control techniques to transform DNA into a reliable data storage medium. However, in order to be useful as general purpose archival medium, solutions must also be developed to support applications that need advanced access paths over data stored in DNA to enable search and selective retrieval.

In this paper, we focus on one such application–adaptive resolution selection over image collections stored in DNA. We introduce a DNA-based image storage system that can adaptively lower the retrieval cost of the image by selecting and decoding only the oligos containing a resolution-reduced version of the image. Each resolution layer is encoded into a set of oligos using the JPEG 2000 progressive codec and the JPEG DNA VM codec, a DNA-based coder that aims at retrieving a file with a high reliability. The novelty of the image storage system lies in its adaptive resolution selection process, based on the random access capabilities provided by the Adaptive Sampling functionality of the Nanopore sequencer. Adaptive Sampling allows the sequencing process to select specific resolution layers oligos of the image, that can then be decoded. Additional resolution layers can even be dynamically selected to refine the decoded image.

Index Terms— DNA data storage, JPEG 2000, progressive, random access, JPEG DNA VM

1. INTRODUCTION

The data storage industry is facing unprecedented challenges as the demand for storage continues to increase at an exponential rate. Conventional storage devices, have fundamental durability and density limitations that make long-term data storage infeasible. Synthetic DNA molecules have emerged as an alternative for overcoming these challenges [1], mainly due to their long lifespan, high density and low energy needs.

A classic solution for storing data onto DNA molecules is composed of both biochemical and computational processes. The two main biochemical processes are synthesis, to create DNA molecules with the desired sequences of nucleotides, and sequencing, to retrieve the sequence of nucleotides of those molecules. The main computational process is the DNA-adapted codec, that transforms data into quaternary sequences for synthesis, and back during sequencing. The codec also compresses the data to reduce synthesis and sequencing cost. Prior work on DNA storage has focused on optimizing both biochemical and computational processes with the goal of transforming DNA into a reliable digital storage medium despite errors that arise at various points in the reading and writing pipeline. However, current DNA storage solutions are not designed to support an important type of access mode that is triggered when images are stored in DNA–the adaptive selection of the decoded image resolution.

Adaptive resolution selection is used when one might choose to display a low resolution version of images on resource constrained devices, instead of their high resolution versions. Supporting adaptive resolution selection requires the use of efficient image coding techniques, and random access support from underlying storage media, to be able to retrieve selective portions of an image depending on resolution requirements. Prior work on DNA data storage has demonstrated the use of Polymerase Chain Reaction (PCR) for enabling such random access. However, another alternative, based on the Read Until function of the Nanopore sequencers, can help retrieve oligos with specific reference sequences, without doing any primer-specific PCR augmentation. Till date, Read Until has only been used to perform random access over arbitrary binary files; its utility in the context of image storage for features like adaptive selection of decoded resolution has remained unexplored.

In this paper, we bridge this gap by presenting a novel approach that brings together progressive image coding with Read Until functionality of Nanopore sequencers to support the adaptive resolution selection for image collections stored in DNA, at a low read cost compared to existing solutions.

2. CONTEXT

2.1. DNA-adapted coding

Research activities in the field of DNA data storage have been rapidly developing over the past years. This section provides an overview of a few pioneering approaches. The recent survey [2] makes detailed comparisons of various approaches that focus on storing generic, binary data using DNA. In 2012, Church et al. [1], introduced an approach to enable large-scale encoding and decoding of any bonary data into synthetic DNA molecules. The paper also identified the main coding constraints that have to be respected when encoding data into DNA molecules. In 2013, Goldman et al.[3] provided an en-

THIS WORK WAS FUNDED BY THE EUROPEAN UNION'S HORIZON RESEARCH AND INNOVATION PROGRAMME PROJECTS GLACIATION (GRANT NO. 101070141), AND ANR PEPR PROGRAM MOLECULARXIV

coder respecting some of these constraints. In 2015, Grass et al.[4] introduced the first error correction codes into a DNA data storage solution. This error correction mechanism aims at robustifying the whole storage process against the errors occurring because of the biochemical operations. Following this, other error correction solutions [5, 6, 7, 8, 9, 10, 11] appeared, by adding redundancy to binary data to detect and correct errors. Random access solutions that enable selective access to binary data stored in DNA to improve read cost have also been investigated [12, 13].

Recently, DNA-adapted image codecs also emerged. One of the first solutions, Dimopoulou et al.[14] developed a JPEG-based image coder adapted to DNA data storage. Similarly, Li et al.[15] introduced a DNA-adapted coding scheme based on JPEG. Pic el al.[16] used an improved DNA-adapted entropy coder to increase the performance of this JPEG-based DNA-adapted image coder. Lazzarotto et al.[17] developed the JPEG DNA VM software that encodes data with a system based on Raptor codes[18]. Learning-based image compression methods have also been studied [19, 20], where the authors use a variational autoencoder to compress the image into a latent space that is later encoded into DNA. In [21], the authors leverage pixel domain representation to reconstruct the images with better compression performance.

2.2. Random Access in DNA Data Storage

The majority of random access methods involves the incorporation of unique address sequences or physical handles that enable selective identification and isolation of target DNA strands. These addressing schemes must satisfy several critical requirements: orthogonality to prevent cross-reactivity, thermodynamic stability under operational conditions, and compatibility with downstream sequencing and decoding processes [12, 22, 23, 24, 25, 26].

2.2.1. PCR-based Random Access

PCR-based random access represents the first scalable approach to selective data retrieval in DNA storage systems. This method uses primer sequences appended to oligos, enabling the amplification of target files through PCR. In [12], the authors demonstrated the first large-scale implementation, successfully retrieving individual files from a pool containing over 13 million DNA oligos, encoding 200 megabytes of data. The system architecture requires careful primer design to ensure orthogonality and prevent non-specific amplification [12, 25, 27, 28].

PCR-based systems demonstrate excellent scalability, with theoretical capacity for 8.98×10^{21} addressable targets and a system capacity of 65.8 ZB [22]. However, several limitations constrain their practical implementation. The amplification process is inherently destructive, requiring complete consumption of the original DNA pool for each access operation [12]. Access times range from several hours due to thermal cycling requirements, and elaborate workflows such as emulsion PCR [12, 22] and micro-encapsulation techniques employed to overcome amplification bias that can skew representation of different sequences [29]. Additionally, using primer sequences, particularly nested or hierarchal primers [12, 22] within an oligo requires 20-60nt overhead, taking away from nucleotide allocation to a data payload in a 200nt oligo, which increases the cost of synthesis.

2.2.2. Silica Encapsulation with Surface Labelling

The silica encapsulation approach addresses several limitations of PCR-based systems by physically protecting DNA within impervious silica capsules which are surface-labeled with single-stranded DNA barcodes corresponding to addresses or file metadata [30]. Unlike PCR retrieval, FACS-based sorting of silica-encapsulated DNA allows for the non-destructive selection of target files without amplification, avoids bias, and decouples address labels from payload strands, improving net data density. The encapsulation process utilises sol-gel chemistry to create stable silica matrices around DNA molecules [30]. Experimental research demonstrated that DNA encapsulated in silica maintains structural integrity and remains sequenceable after accelerated aging equivalent to 2000 years in central Europe [4]. The silica matrix protects against degradation from UV radiation, oxidation, and hydrolysis [31].

Fluorescence-activated cell sorting (FACS) enables precise selection of target capsules from complex mixtures [30]. The system achieves selection sensitivity of one in 10^6 files per optical channel, with the capability to scale to $10^6 N$ files using common commercial FAS systems which offer up to $N = 17$ optical channels [30]. FACS operates by detecting oligonucleotide probes with fluorescent labels hybridised to barcode sequences on the surface of a silica capsule, allowing rapid identification and physical separation of target capsules while leaving the rest of the pool intact for future access [32].

Encapsulation preserves the integrity of the pool for repeated queries, and the silica matrix confers long-term environmental protection against hydrolysis, oxidation, and UV damage [30, 4]. However, this approach involves elaborate workflows for sol-gel encapsulation and surface labeling, introduces latency due to chemical deprotection steps required to release DNA from capsules before sequencing, and relies on sophisticated and costly instrumentation such as FACS.

In summary, prior work on random access in DNA storage has predominantly investigated the ability to selectively retrieve arbitrary binary files from an oligo pool. To our knowledge, no work has investigated the utility of these techiques for supporting more complex access patterns required for supporting adaptive resolution selection over image collections stored in DNA. We believe that issues mentioned earlier make both PCR-based and silica encapsulation-based techniques unsuitable as mechanisms for supporting such advanced access paths.

3. PROPOSED METHOD

This work proposes a new method enabling adaptive resolution selection that brings together progressive image coding and Adaptive Sampling capability of Nanopore sequencing.

3.1. Adaptive Sampling

Adaptive Sampling provided by Nanopore relies on the presence of user-provided reference sequences located at the beginning of oligos to accept or reject an oligo for sequencing. More specifically, as each DNA strand begins translocation in the Nanopore sequencer, buffered current signals are basecalled and aligned against these reference sequences; if a match is detected, sequencing proceeds, otherwise voltage reversal ejects the strand back into the pool. Roman Sokolovskii et al. [33] demonstrated that this approach can selectively enrich target subsets from a heterogeneous DNA pool, dynamically switch targets mid-run, and employ library recovery protocols so that non-target strands remain available for future retrievals.

Compared to PCR-based random access, adaptive sampling inherently avoids amplification bias and pool depletion, whereas adaptive sampling ejects non-matching strands intact, preserving library diversity and enabling reuse. Dynamic target switching during a run allows multiple files to be retrieved from a single aliquot without additional preparation steps [33]. Relative to silica encapsulation with FACS-based sorting, adaptive sampling removes the need for sol-gel encapsulation because selection occurs inline during sequencing via in-read alignment. Although adaptive sampling's efficiency diminishes for very rare targets in large pools without prior enrichment, and its need for long constructs requires assembly or similar workarounds, its non-destructive, dynamic, and amplification-free characteristics make it a compelling alternative or complement to PCR- and silica-based random access methods.

3.2. Progressive Decoding and Adaptive Sampling

Progressive decoding was used in the past to mitigate the constraints imposed while exchanging images on slow networks: while the data was being transmitted, the first layers could progressively be decoded, to obtain a degraded preview of the desired image. Viewed in the context of DNA storage, our idea is to support adaptive resolution selection by pairing progressive decoding with Adaptive Sampling. More specifically, prior work stores each image separately in one or more oligos. In contrast, we decompose each image into a series of resolution layers so that we can group and encode the data corresponding to each resolution layer in distinct sets of oligos. Oligos corresponding to each layer are then indexed using a distinct reference sequence that is drawn from a predesigned library of reference sequences. Thus, an image can be progressively retrieved from an oligo pool, by looking, in the dictionary of reference sequences, for the sequences of the different resolution layers of that specific image. The different layers of the image are retrieved starting with the lower resolutions. During the sequencing process, the image codec will successively input different reference sequences to the Nanopore sequencer, that will update the parameter of the running Read Until process. The current reference sequence that is input to the Read Until process will allow the sequencer to only sequence the oligos starting with this input, and reject back into the general oligo pool the other ones.

The novelty of our contribution does not only come from its progressive coding system, but also from the sequencing process that it is associated with: the coding system is designed to output oligos that can be selectively sequenced by the Nanopore sequencer in its adaptive sampling mode. In return, the adaptive sampling process allows us to integrate in our decoding workflow a PCR-free random access process, that aims at selecting the data to be decoded, hence reducing the reading cost of the whole system. Avoiding the use of several PCR processes to access the image data is crucial, not only because of its relatively lengthy process, but more specifically because it requires a human intervention at the biochemical level every time.

3.3. Encoding and Decoding Method

Our image coding solution is based on the JPEG2000 codec in its progressive coding mode. The codec outputs a bitstream organized in N_{levels} resolution layers of increasing sizes. This bitstream is sliced into a set of N_{levels} binary files, corresponding to the different layers. Each resolution binary file is encoded into a pool of DNA-adapted short payload oligos using the JPEG DNA VM codec. Each pool is then be separately synthesized and prepared for adaptive sampling.

Nanopore's Adaptive Sampling imposes some restrictions with respect to the oligo length. The core mechanism requires buffering and processing on the order of $\sim 400\,\mathrm{nt}$ to accommodate signal latency. Further, as mentioned earlier, the signal needs to be basecalled and aligned, adding to the latency before a keep-or-eject decision can be made [34]. This implies that sufficiently long DNA strands ($\geq$ 1000–1500 nt) are needed for reliable early alignment against short address motifs. However, current synthesis techniques impose upper limit on the oligo length and it is exponentially more expensive to synthesis longer oligos compared to shorter ones. We overcome this problem in our design by ligating shorter oligos to make longer ones using adapters as done in prior work [33].

The adaptive sampling method of Random Access relies on the presence of reference sequences, located at the beginning of oligos. These sequences are used as indexes for the following long payload, appended to the end of the reference (Fig. 1). The reference sequences used for each resolution are designed during the preparation of the oligos. A reference sequence dictionary is built to associate any resolution level of an image with the reference sequence that was used for the preparation of its pool of oligos. Once all the resolution levels of all the images in the dataset have been prepared, they can be merged into a single general pool of oligos. Each image resolution layer has its own unique reference sequence, resulting in a total of $N \times L$ reference sequences, where N is the number of images, and L the number of levels per image. The decoding process begins by providing to the Nanopore sequencer a succession of selected reference sequences. The image is then reconstructed from the sequenced oligos with

Fig. 1. Oligo structure for the Adaptive Sampling sequencing mode

the JPEG DNA VM and JPEG 2000 codecs. Should the quality of the reconstructed image be refined, additional reference sequences can be sent to the sequencer. The selected reference sequences correspond in the dictionary to the resolution layers to be decoded. If several images (k) should be decoded, the reference sequences of the selected layers of the k images should be provided to the sequencer in the same manner.

4. EXPERIMENTAL RESULTS

The proposed codec is primarily designed to decrease the cost necessary to retrieve an image from a pool of oligos. Since the focus of this paper was on the improvement of the reading cost, the experiments will focus on this aspect of the DNA data retrieval process, without considering any simulation of the synthesis and sequencing processes. The first step to measure this cost is to identify the procedure used for image retrieval. In DNA coding systems that do not implement any random access at the oligo level, the cost of reading an image is equal to the number of oligo reads necessary to read the whole dataset. For this type of coder, Equation (1) can represent the read-cost necessary for the retrieval of an image. On the other hand, an image retrieval process that integrates random access or progressive decoding in its core coding system can improve on the read cost in Equation (1). Firstly, if progressive decoding is enabled, the read cost can be limited to Equation (2), since only the first layers need to be decoded.

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

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

The $nucs(i,k)$ value represents the number of nucleotides to sequence to be able to decode a layer:

$$nucs(i,k) = coverage(i,k) \times number\_oligos(i,k) \quad (3)$$

With this, we can define a read-cost gain that measures the improvements provided by our progressive decoding approach:

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

4.1. Performance evaluation

The performance of any DNA encoding method can be evaluated with respect to a series of metrics such as RD-curves, reading cost and writing cost. As our work primarily focuses on progressive decoding (PD) , the main metric we focus on in our study is the reading cost. We especially study the evolution of the reading cost for a given encoded image, as we read through each resolution layer. The image is encoded into a series of resolution layers . The resolution layers each divide the size of the image by a factor of 2 in each dimension (4 in total). The results presented here were obtained by encoding 5 images of the kodak[1] dataset, each with 3 resolution levels. The oligos had data blocks of length 148. During synthesis, these data blocks will be ligated together and reference sequences specific to the image and resolution level will be added to each ligated molecule, for random access with Adaptive Sampling. A theoretical read-cost gain was observed (Table 1) that depicts the improvements provided by progressive decoding against a similar coding method that does not provide any adaptive read-cost reduction such as the selection of resolution.

In the results shown in Table 1, we measured the number of oligos necessary to decode each image until a certain resolution level, and averaged the results over all the images. In these conditions, the progressive decoder provides gains of up to 7.5×, when only the initial layer is targeted. This gain quickly decreases if more layers are targeted, and if all the layers are targeted, no gain can be leveraged from PD, because all oligos need to be sequenced. These gains, though, heavily depend on the dimensions of the chosen resolution layers: smaller layers will leverage better gains, at the cost of more distorted images.

The gains obtained are orthogonal to improvements in coding performance and read cost that can be achieved by adjusting the encoding options of both the JPEG DNA VM and JPEG2000. The JPEG DNA VM software parameters can be adjusted differently for each resolution layer, especially the redundancy, so that lower resolution layers are better protected against errors. For this reason, we can extract information from a comparison with the method from [35], which also leverages gains and uses these same codecs. Although the results of the compression process are not in the favor of our novel compression method, our contribution still shows significatn rate gains when using progressive coding. More importantly, these results do not take in consideration the whole biochemical process, that is very different between both methods. Our contribution here only requires the use of one PCR process at the beginning of the decoding workflow, while the paper in [35] bases its random access on PCR, hence requiring $1 + 2 \times k$ (with k the number of resolution levels to access) PCR runs for the other method. In other terms, the biochemical complexity $C_{contrib}$ of the proposed method should be of the form $C_{contrib} \approx \mathcal{O}(1)$, while the previous one is of the form $C \approx \mathcal{O}(k)$.

5. CONCLUSION

In this paper, we introduced a new approach that combines progressive image decoding with Adaptive Sampling to enable a new access path over image collections stored in DNA–adaptive resolution selective. Our theoretical performance analysis showed that our approach can provide substantial reduction in read cost, while simultaneously holding the promise of eliminating multiple PCR amplification rounds. In future work, we plan to perform wet lab experiments to evaluate our method with synthesis and sequencing.

[1] https://r0k.us/graphics/kodak/

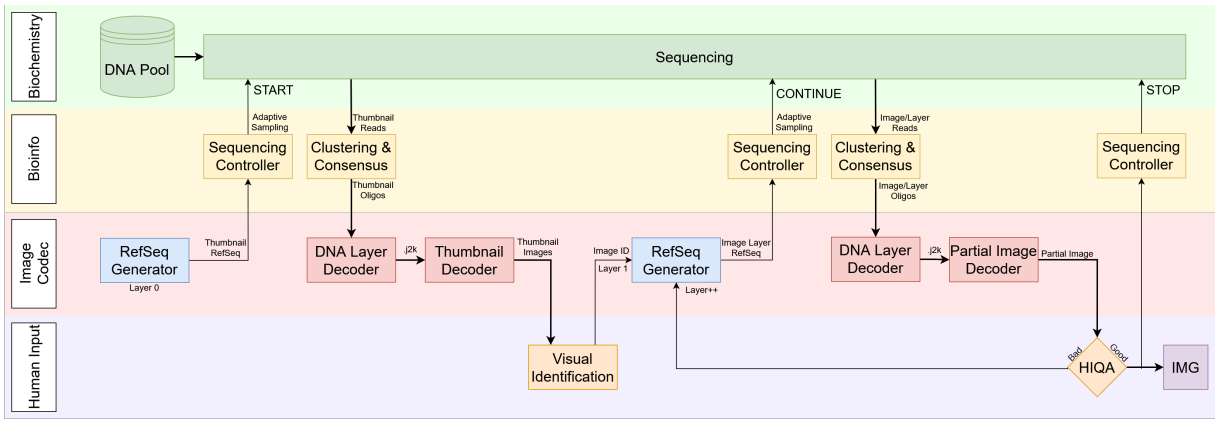

Fig. 2. Thumbnail-based decoding workflow of the Adaptive Sampling based DNA image codec

Layer	L_0	L_1	L_2
# Oligos	6764	8404	15542
Theoretical G_{pd} Read-cost gain	**21.64**	**17.81**	**10.41**
# Oligos	3812	5648	9090
Theoretical G_{pd} Read-cost gain	38.4	26.5	17.8

Table 1. Average read-cost gains G_{pd} for each target resolution level L_k, averaged over all the selected images. The top read cost gains are the results of this contribution, and the bottom are obtained with the method in [35], with identical parameters.

6. REFERENCES

[1] G. M. Church et al., "Next-generation digital information storage in DNA," Science, vol. 337, no. 6102, 2012.

[2] T. Heinis et al., "Survey of information encoding techniques for DNA," ACM Comput. Surv., vol. 56, Nov. 2023.

[3] N. Goldman et al., "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," Nature, 2013.

[4] R. N. Grass et al., "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," Angewandte Chemie International Edition, 2015.

[5] Y. Erlich and D. Zielinski, "DNA fountain enables a robust and efficient storage architecture," Science, 2017.

[6] S. M. H.T. Yazdi et al., "Portable and error-free DNA-based data storage," Scientific Reports, vol. 7, no. 1, pp. 1–6, 2017.

[7] M. Welzel et al., "DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in dna storage," Nature Communications, 2023.

[8] M. Blawat et al., "Forward error correction for DNA data storage," Procedia Computer Science, 2016.

[9] Ben C. et al., "GCNSA: DNA storage encoding with a graph convolutional network and self-attention," iScience, vol. 26, no. 3, pp. 106231, 2023.

[10] Ben Cao et al., "Adaptive coding for DNA storage with high storage density and low coverage," NPJ Systems Biology and Applications, vol. 8, 2022.

[11] O. Sabary et al., "Reconstruction algorithms for DNA-storage systems," Scientific Reports, vol. 14, 2024.

[12] L. Organick et al., "Random access in large-scale DNA data storage," Nature Biotechnology, 2018.

[13] E. Marinelli et al., "CMOSS: A reliable, motif-based columnar molecular storage system," in Proceedings of the 17th ACM International Systems and Storage Conference, 2024.

[14] M. Dimopoulou et al., "A JPEG-based image coding solution for data storage on DNA," EUSIPCO, 2021.

[15] B. Li et al., "IMG-DNA: approximate DNA storage for images," New York, NY, USA, 2021, SYSTOR '21.

[16] X. Pic and M. Antonini, "A constrained shannon-fano entropy coder for image storage in synthetic DNA," European Signal Processing Conference (EUSIPCO), 2022.

[17] D. Lazzarotto et al., "Technical description of the EPFL submission to the JPEG DNA CfP," arXiv preprint, 2023.

[18] A. Shokrollahi, "Raptor codes," IEEE Transactions on Information Theory, vol. 52, pp. 2551 – 2567, 06 2006.

[19] Y. Zheng et al., "DNA-QLC: an efficient and reliable image encoding scheme for dna storage," BMC Genomics, 2024.

[20] G. Franzese et al., "Generative DNA: Representation learning for DNA-based approximate image storage," in VCIP 2021, 2021.

[21] S. Seo et al., "Information density enhancement using lossy compression in DNA data storage.," 2024.

[22] K. J. Tomek et al., "Driving the scalability of DNA-based information storage systems," ACS Synthetic Biology, 2019.

[23] A. El-Shaikh et al., "High-scale random access on DNA storage systems," NAR Genomics and Bioinformatics, vol. 4, 2022.

[24] K. N. Lin et al., "Dynamic and scalable DNA-based information storage," Nature Communications, vol. 11, 2020.

[25] S. K. Subramanian et al., "A set of experimentally validated, mutually orthogonal primers for combinatorially specifying genetic components," Synthetic Biology (Oxford), 2018.

[26] G. Gowri et al., "Scalable design of orthogonal DNA barcode libraries," Nature Computational Science, vol. 4, 2024.

[27] Qiang Z. et al., "A novel constraint for thermodynamically designing DNA sequences," PLOS ONE, vol. 8, 2013.

[28] A. S. Booeshaghi et al., "Quantifying orthogonal barcodes for sequence census assays," Bioinformatics Advances, vol. 4, no. 1, 2023.

[29] B. B. Bögels et al., "DNA storage in thermoresponsive microcapsules for repeated random multiplexed data access," Nature Nanotechnology, vol. 18, no. 8, pp. 912–921, 2023.

[30] J. L. Banal et al., "Random access DNA memory using boolean search in an archival file storage system," Nature Materials, 2021.

[31] D. Kapusuz and C. Durucan, "Exploring encapsulation mechanism of DNA and mononucleotides in sol-gel derived silica," Journal of Biomaterials Applications, vol. 32, 2017.

[32] Bio-Rad, "Fluorescence-Activated Cell Sorting (FACS)," https://www.bio-rad.com/en-ie/feature/fluorescence-activated-cell-sorting.html, Accessed: 2025-06-20.

[33] R. Sokolovskii et al., "Adaptive sampling in nanopore sequencing for pcr-free random access in DNA data storage," bioRxiv, 2024.

[34] Oxford Nanopore Technologies, "Adaptive sampling," https://nanoporetech.com/document/adaptive-sampling#introduction-advanced, Accessed: 2025-06-20.

[35] X. Pic and R. Appuswamy, "Combining progressive image compression and random access in DNA data storage," 25th International Conference on Digital Signal Processing, 2025.