DNA DATA STORAGE SYSTEMS: AN EXTENDED REVIEW OF ARCHITECTURES, WORKFLOWS, CHALLENGES, AND RESEARCH DIRECTIONS

Ms. Amritha V, Ms. Divya P
MCA Scholar, Assistant Professor
Department of MCA,
Nehru College of Engineering and Research Centre, Thrissur, India

Abstract : DNA data storage is an emerging method for long-term digital archiving, where binary files are represented as nucleotide sequences and encoded as synthetic DNA strands, which are then stored as physical molecules, to be sequenced and computationally decoded later. This is an extended review of DNA storage as a complete system stack and an integration of research on encoding, synthesis, storage, sequencing, and reconstruction. We describe the widely cited milestones, the importance of indexing and constraint-aware sequence design, and redundancy and error correction schemes necessary for handling substitutions, insertions, deletions, and strand dropout errors. The article also covers random access, automation, and security implications for offline archives and system evaluation metrics for benchmarking. Finally, we conclude with application relevance to cold archives and research directions for faster adoption**.**

IndexTerms **- DNA data storage, archival storage, synthesis, sequencing, encoding, constraints, error correction, random access, automation, benchmarking.**

INTRODUCTION

The amount of data is increasing because of mobile devices, high-definition media, cloud services, enterprise platforms, scientific experiments and AI model training. In cases data is not used again after it is first used.. It still has to be kept for a long time because of rules and regulations audits and future references. When you have a lot of data it can be expensive to keep it for a time because storage hardware and formats change and physical devices get old and stop working.

Traditional ways of storing data like magnetic tape are good for storing cold data but they still need to be managed stored properly and updated from time to time. Big archives also have to think about how energy they use and how much work it takes to run them. These things make people want to find ways to store data that can keep it safe for a long time without needing much maintenance and using less energy.

DNA is a molecule that can store information and has shown it can last a long time if it is stored properly. DNA data storage uses this idea by changing data into sequences of A, C, G and T. A DNA storage system has a few parts: a digital codec that gets the data ready and encodes it a write stage that makes DNA strands a storage stage that keeps the strands safe and a read stage that figures out the order of the strands and puts the original file back together. Because making and reading DNA strands can be messy and some strands might be missing it is very important to have a system, for finding and fixing mistakes so that the data can be recovered reliably. DNA data storage is a way of storing digital data using DNA and it is being researched as a way to store digital data for a long time.

REVIEW METHODOLOGY

This is a narrative review that organizes the literature in the context of a DNA storage pipeline. The review will cover system overview and vocabulary, encoding and sequence requirements, physical writing and storage, sequencing and decoding, and random access retrieval. It will then cover the widely cited milestones that shaped the literature and provide a comparison with traditional storage media. Finally, it will synthesize the challenges and future research.
The review will focus on system-level ideas rather than the mathematical specifics. The major ideas that are evaluated in the literature include: the ability to recover from a realistic error model, strand dropout robustness, redundancy overhead, read and write times, and the dominant costs of synthesis and sequencing. These ideas will be covered with explanations that are suitable for a computer science and information technology audience.

TERMINOLOGY AND SYSTEM OVERVIEW

A. Core Terminology

In the literature of DNA storage, an oligonucleotide (or oligo) is a small artificial DNA molecule that serves as the building block of storage. A DNA library is a collection of many oligos that together hold one or more files stored in the DNA storage system. Sequencing yields reads, which are noisy measurements of oligos. Dropout corresponds to missing oligos that are not in a retrieved

set or are not measured during sequencing. A codec is the software system that deals with encoding, indexing, redundancy, and decoding.

B. Storage Stack View

A storage stack with three layers is the most appropriate way to understand DNA storage. The digital layer is responsible for file preparation, segmentation into blocks, indexing, constraint-aware encoding, and redundancy generation. The wet-lab layer is responsible for synthesis (writing) and preservation (storage). The read and compute layer is responsible for sequencing, read processing, reconstruction, and decoding with integrity validation. This framework of layers is very useful in understanding why DNA storage is more than a biology issue. It is an engineering system where software design mitigates physical noise.

LITERATURE REVIEW AND MILESTONES

DNA data storage progressed from the idea phase to proof-of-concept in the lab in the past decade. Church, Gao, and Kosuri (2012) showed the feasibility of DNA data storage by encoding a multi-megabit dataset in DNA and successfully decoding it after sequencing. Goldman et al. (2013) made the idea more practical by optimizing fragment design and using redundancy, allowing for error-free data recovery for larger datasets. Erlich and Zielinski (2017) proposed DNA Fountain code, a more resilient method that can withstand strand loss while remaining highly efficient by incorporating fountain code concepts. Takahashi et al. (2019) showed an end-to-end automated system, emphasizing the need for automation in such systems.

However, aside from these achievements, the literature is increasingly concerned with system-level engineering. Open issues include: more stringent sequence constraints for improved synthesis success rates, more efficient reconstruction pipelines that can handle insertions and deletions, addressing schemes for selective data retrieval from large data pools, and end-to-end comparisons of different pipelines for fair evaluation.

ENCODING AND CONSTRAINT-AWARE DESIGN

A. From Bits to Bases

Conceptual representation of DNA encoding: A conceptual representation of DNA encoding is the mapping of binary data to the four nucleotides A, C, G, and T. Although a straightforward 2-bit encoding is feasible, more complex encoding is used in DNA storage because the physical processes of DNA writing and reading are sensitive to sequence patterns. Encoding trade-offs error correction and redundancy overhead. Encoding should generate sequences that can be synthesized and sequenced with a low error rate, while maintaining redundancy overhead.

B. Indexing and Metadata

Indexing is a crucial step as the reads do not retain the order in which the fragments are stored. Every oligo has an index that corresponds to its location or block. Indexing helps in grouping and assembly during the decoding process. Metadata also encompasses the version of the codec and parameters such as the size of the block and redundancy rate, which become relevant when the archive is maintained for long periods and requires correct decoding.

C. Sequence Constraints

Constraint-aware design seeks to prevent patterns that will increase errors or lower yield. For instance, long homopolymers may be challenging for certain sequencing technologies, while high GC content may lower synthesis accuracy. Most encoders impose constraints on run length, ensure that the GC content is within a target range, and prevent motifs that form high secondary structure. Such constraints affect the storage efficiency of certain codecs.

REDUNDANCY AND ERROR CORRECTION

A. Error Sources

The errors in DNA storage occur during synthesis, storage handling, amplification, and sequencing. Substitutions involve the replacement of a base with another. Insertions and deletions involve adding or deleting bases, leading to a shift in alignment. Dropout involves the absence of some oligos in the retrieved sample or their absence in the sequencing reads. These are different from the errors in conventional storage, where errors are modeled as bit flips.

B. Multi-Level Protection

Good DNA storage usually involves protection at multiple scales. On the level of individual oligos, checksums and error-correcting codes can be used to spot and correct errors at the base level. On the level of a collection of oligos, redundancy can be used to tolerate dropout. Fountain codes, like DNA Fountain, are designed to generate a large number of encoded packets such that the original data can be recovered after obtaining a sufficient number of valid strands, even if some of them are missing. The aim is to reach a good trade-off between redundancy and cost.

WRITE, STORE, READ, AND DECODE PIPELINE

A. Writing by Synthesis

The write stage synthesizes oligos that stand for the encoded fragments. Currently, synthesis is a major cost factor and a significant barrier to widespread adoption. Write latency is also high since synthesis is much slower than electronic writing. Quality control and yield impact the number of oligos that can be used for decoding, which is affected by redundancy needs.

B. Storage Conditions and Stability

In archival storage, the integrity of the DNA over the years is of utmost importance. DNA can be stored in a dry environment or in a protective matrix to shield it from the effects of moisture, heat, and other environmental factors. In practical implementations, there is also a need for robust labeling and cataloging of the DNA samples since DNA is a tangible material. The information regarding the contents of a sample and how to interpret it needs to be stored in a reliable manner.

C. Reading by Sequencing and Reconstruction

The read stage involves the processing of DNA to generate a massive number of reads. The decoding pipeline may involve the removal of low-quality reads, the grouping of reads based on indices or addresses, and the construction of consensus sequences to

eliminate noise. Error correction is applied to fix errors, while redundancy is applied to fill missing fragments. The decoded blocks are then assembled to form the original file, which is verified through integrity checks.

RANDOM ACCESS, AUTOMATION, AND SECURITY

A. Random Access and Selective Retrieval

A useful archive should be able to retrieve a chosen file or record without having to sequence the whole collection. Random access techniques usually involve addressing systems that enable selective amplification or selection of target strands. Although this makes the archive more practical, it also imposes other requirements, such as minimizing cross-talk between addresses, preventing amplification bias, and maintaining efficient retrieval even with a large library.

B. Automation and Integrated Pipelines

For instance, many DNA storage techniques require manual processes, which may limit their repeatability. However, this may be addressed to a certain level by introducing robotics and automated systems to improve accuracy and consistency. Another aspect is that automated systems may aid in standardizing sample processing, which is essential for DNA storage to be a viable operational tier.

C. Security and Governance

The storage of DNA can also be kept offline as a physical archive, which reduces the threat of network-based attacks. However, this also brings about the issue of physical security, which involves handling the DNA securely, chain of custody, and access to the DNA sample. Encryption can be used to ensure confidentiality at the digital level, whereas operational security can be used to ensure that there is no loss of the sample.

COMPARATIVE ANALYSIS WITH CONVENTIONAL STORAGE

The best analogy for DNA's storage is tape, rather than a low-latency device such as SSDs. DNA provides many desirable characteristics for a deep archive, including very high density and zero power consumption when stored. By contrast, HDDs and SSDs offer excellent performance but require ongoing infrastructure and replacement cycles. Tape is a well-understood archival solution, but refresh cycles are needed, and care is necessary for its use.

Medium	Typical use	Power while stored	Access latency	Main limitation
DNA (emerging)	Deep archive, long retention	No	High (lab workflow)	Write/read cost, latency
Tape	Cold archive	Low when offline	Medium	Handling, refresh
HDD	General storage	Yes	Low	Wear, migration
SSD	Fast storage	Yes	Very low	Cost, endurance

APPLICATIONS

This is particularly well-suited for cold archives where information is valuable and must be preserved for long durations of time. Government record retention, compliance archives, record audit trails, etc., are some examples of usage scenarios for DNA storage. Cultural heritage preservation is another good use case for DNA storage because libraries and museums are already looking to preserve information for long durations with minimal risk.

There are also scientific archives that produce large volumes of information which must be preserved for long durations to support reproducibility and long-term analysis. DNA storage is considered for a deep archival use case for scientific archives because access is infrequent after initial processing. Safe offline storage is also a good use case for high-value backups when operations are strong.

CHALLENGES AND LIMITATIONS

Nonetheless, it is important to understand that DNA storage comes with certain limitations that need to be addressed. The first limitation is the cost. This is because the write process is dominated by DNA synthesis, while the read process is dominated by both sequencing and decoding. Another limitation is the latency. This is because the process of both reading and writing involves the laboratory process. Lastly, there is the limitation of error handling. This includes issues related to insertions, deletions, and dropout. In addition, the process of random access comes with certain limitations, including address bias. The operational limitations include issues related to tracking, metadata, and quality. This is why DNA storage is best suited for deep archives.

FUTURE DIRECTIONS

Some areas for research to help accelerate the deployment of DNA storage include reducing synthesis costs and increasing synthesis throughput, enhancing sequencing processes, and developing codecs that are aware of error rates to minimize redundancy without compromising on accuracy. Better reconstruction techniques can minimize the amount of sequencing required, thereby reducing costs for reading from a DNA storage device.

Scalability for random access is another important area for DNA storage, which allows for random access to large storage libraries.

CONCLUSION

In this way, the above review article has portrayed DNA data storage systems as a complete DNA storage stack for archival preservation. This is a complete review that shows how DNA data storage systems have progressed from feasibility to robust coding to automation, with more focus on the system level. DNA data storage has a good potential for archival preservation due to its density and no power requirement during storage. For DNA data storage to be realized, there needs to be a reduction in cost, improvement in read latency, enhancement in random access, and standardization. This can be achieved through progress in biotechnology and storage engineering.

REFERENCES

[1] G. M. Church, Y. Gao, and S. Kosuri, "Next-Generation Digital Information Storage in DNA," Science, vol. 337, no. 6102, p. 1628, 2012, doi: 10.1126/science.1226355.
[2] N. Goldman et al., "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," Nature, vol. 494, no. 7435, pp. 77–80, 2013, doi: 10.1038/nature11875.
[3] Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," Science, vol. 355, no. 6328, pp. 950–954, 2017, doi: 10.1126/science.aaj2038.
[4] C. N. Takahashi, B. H. Nguyen, K. Strauss, and L. Ceze, "Demonstration of End-to-End Automation of DNA Data Storage," Scientific Reports, vol. 9, art. no. 4998, 2019, doi: 10.1038/s41598-019-41228-8.
[5] Storage Networking Industry Association (SNIA), "DNA Data Storage Technology Review," Version 1.0, 2025.
[6] Storage Networking Industry Association (SNIA), "DNA Data Storage Codecs: Examples, Requirements, and Metrics," 2025.
[7] M. Welzel *et al.*, "DNA-Aeon provides flexible arithmetic coding for constraint adherence and error correction in DNA storage," *Nature Communications*, vol. 14, 2023, doi: 10.1038/s41467-023-36297-3.
[8] L. C. Meiser et al., "Synthetic DNA applications in information technology," Nature Communications, vol. 13, 2022, doi: 10.1038/s41467-021-27846-9.
[9] S. Yang et al., "DNA as a universal chemical substrate for computing and data storage," Nature Reviews Chemistry, vol. 8, no. 3, pp. 179–194, 2024, doi: 10.1038/s41570-024-00576-4.
[10] I. Shomorony and R. Heckel, "Information-Theoretic Foundations of DNA Data Storage," Foundations and Trends in Communications and Information Theory, vol. 19, no. 1, pp. 1–106, 2022, doi: 10.1561/0100000117.
[11] A. Doricchi et al., "Emerging Approaches to DNA Data Storage: Challenges and Prospects," ACS Nano, vol. 16, no. 11, pp. 17552–17571, 2022, doi: 10.1021/acsnano.2c06748.
[12] M. H. Raza, S. Desai, S. Aravamudhan, and R. Zadegan, "An outlook on the current challenges and opportunities in DNA data storage," Biotechnology Advances, vol. 66, 2023, doi: 10.1016/j.biotechadv.2023.108155.
[13] Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," Science, vol. 355, no. 6328, pp. 950–954, 2017, doi: 10.1126/science.aaj2038.
[14] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," Science, vol. 337, no. 6102, p. 1628, 2012, doi: 10.1126/science.1226355.
[15] N. Goldman et al., "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," Nature, vol. 494, pp. 77–80, 2013, doi: 10.1038/nature11875.
[16] W. H. Press et al., "HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints," Proceedings of the National Academy of Sciences, vol. 117, no. 31, pp. 18489–18496, 2020, doi: 10.1073/pnas.2004821117.
[17] J. Bornholt et al., "A DNA-Based Archival Storage System," in Proceedings of ASPLOS, ACM, 2016, doi: 10.1145/2872362.2872397.
[18] R. N. Grass et al., "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," Angewandte Chemie International Edition, vol. 54, no. 8, pp. 2552–2555, 2015, doi: 10.1002/anie.201411378.
[19] L. Organick et al., "Random access in large-scale DNA data storage," Nature Biotechnology, vol. 36, pp. 242–248, 2018, doi: 10.1038/nbt.4079.
[20] S. M. H. T. Yazdi et al., "A rewritable, random-access DNA-based storage system," Scientific Reports, vol. 5, 2015, doi: 10.1038/srep14138.
[21] S. Newman et al., "High-density DNA data storage library via dehydration with digital microfluidic retrieval," Nature Communications, vol. 10, 2019, doi: 10.1038/s41467-019-09517-y.