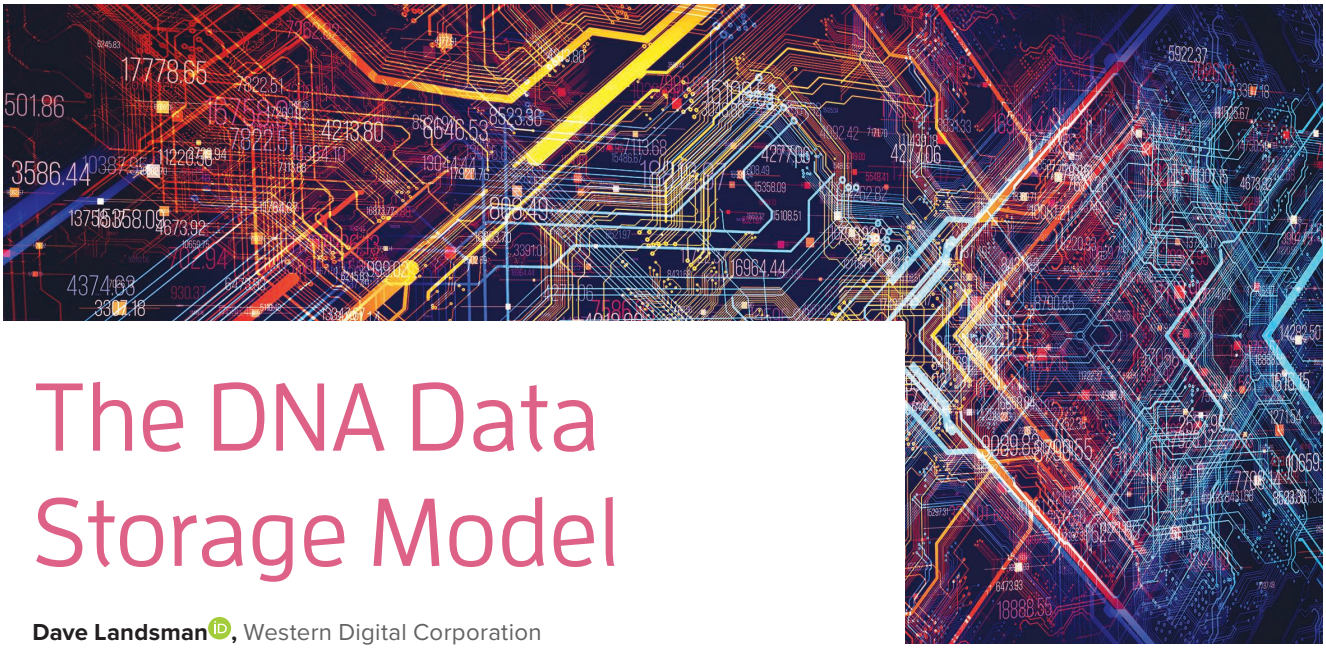

The DNA Data Storage Model

Dave Landsman, Western Digital Corporation
Karin Strauss, Microsoft Corporation

Reliably storing digital data in synthetic DNA fits naturally into the layered Open Systems Interconnect model and shares many parallels with reliably storing data using existing storage technologies and interfaces.

DNA data storage, or using synthetic DNA as a data storage medium, is being seriously considered as an archival storage solution due to its volumetric data density potential, data retention characteristics, sustainability, and potential for dramatically lower total cost of ownership versus existing storage technologies.

INTRODUCTION

The biotechnology industry has made DNA data storage possible today due to decades of investment in molecular-level technologies for medical and life sciences applications that now enable us to construct and read synthetic DNA, base by base. These fundamental capabilities make it possible to encode digital data into a sequence of bases (adenine, guanine, cytosine, and thymine, or AGCT), write that sequence as a set of corresponding DNA molecules (synthesis), store the molecules, prepare them for reading (retrieval), read them back as a sequence of bases (sequencing), and finally, decode the original digital data (Figure 1). To learn more about this process, see *Preserving Our Digital Legacy: An Introduction to DNA Data Storage*.[1]

Even though synthetic DNA as a data storage medium is similar to traditional storage media in many ways, it is worth highlighting some key differences.

First, in traditional storage, the media is premanufactured (e.g., SSDs use NAND cells, HDD, and tape use magnetic domains on a platter or strip, respectively) and written by modifying the state of the media. For such devices, capacity and throughput scaling require modifications to both media and write/read heads. In contrast, the most common DNA data storage method does not employ a premanufactured media substrate. (Note: Methods to attach DNA to a planar substrate, to serve as a "memory/storage cell", are being considered; this article does not cover these methods.) Instead, the storage media—DNA molecules—are manufactured during write operations. In this case, DNA's universal, fixed, and reader/writer

Digital Object Identifier 10.1109/MC.2023.3272188
Date of current version: 26 June 2023

0018-9162/23©2023IEEE

EDITOR TOM COUGHLIN
Coughlin Associates; tom@tomcoughlin.com

independent physical structure enables throughput improvements without media changes, or data migration when adopting new writer/reader generations.

Second, because DNA media is detached from any array-based substrate, protocol information such as object identifiers and segment indices must be embedded within each DNA molecule in the DNA archive, for locating objects, object segmentation, etc. Despite this overhead, DNA can achieve much higher volumetric density as detached media than as array-based media.

Third, DNA has unique error characteristics as a medium for end-to-end storage. For example, in addition to substitutions (akin to bit flips), insertions and deletions may also occur. Encoders try to avoid certain sequences[2,3,4] to reduce interference with the writing and recovery process.

Despite the uniqueness of synthetic DNA as a storage medium, there are many parallels with traditional data storage and storage interface mechanisms. In this article, we draw parallels between the DNA data storage model and the Open System Interconnection (OSI) model (Figure 2). Before we begin describing the DNA data storage layer model, we describe a few things about DNA.

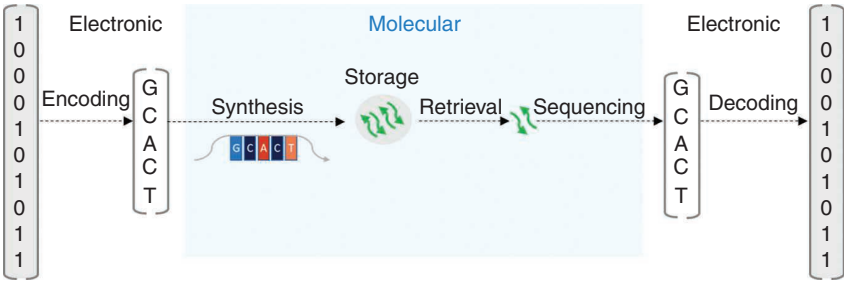

FIGURE 1. The DNA data storage system. Encoding and decoding are performed in the electronic domain and translate bits to bases and vice versa. Synthesis and sequencing are the interfaces to and from the molecular domain, creating and reading DNA sequences.

DNA MECHANICS IN BRIEF

We are most familiar with DNA as a "double-helix" (Figure 3), or dual-stranded DNA (dsDNA), where the base adenosine on one strand has a chemical binding affinity (complementarity) with the base thymine on the other strand, and the base cytosine has an affinity with the base guanine. Complementarity is used in nearly all of the techniques employed in DNA data storage, in the process called hybridization (Figure 4).

In organisms, cell division naturally replicates the genetic code. Cellular mechanisms separate the two strands of the original dsDNA into two single-stranded DNA strands (ssDNA) and create two new dsDNA molecules from them, effectively copying the cell's genetic information. Both ssDNA and dsDNA are used in different parts and applications of the DNA data storage pipeline.

Note that no organisms or cells are used for DNA data storage: synthetic DNA for data storage is constructed and manipulated through well-controlled chemical processes, covered in the section "The DNA Data Storage Physical Layer."

APPLICATION, PRESENTATION, AND SESSION LAYERS

The upper three layers of the OSI model map to the DNA data storage

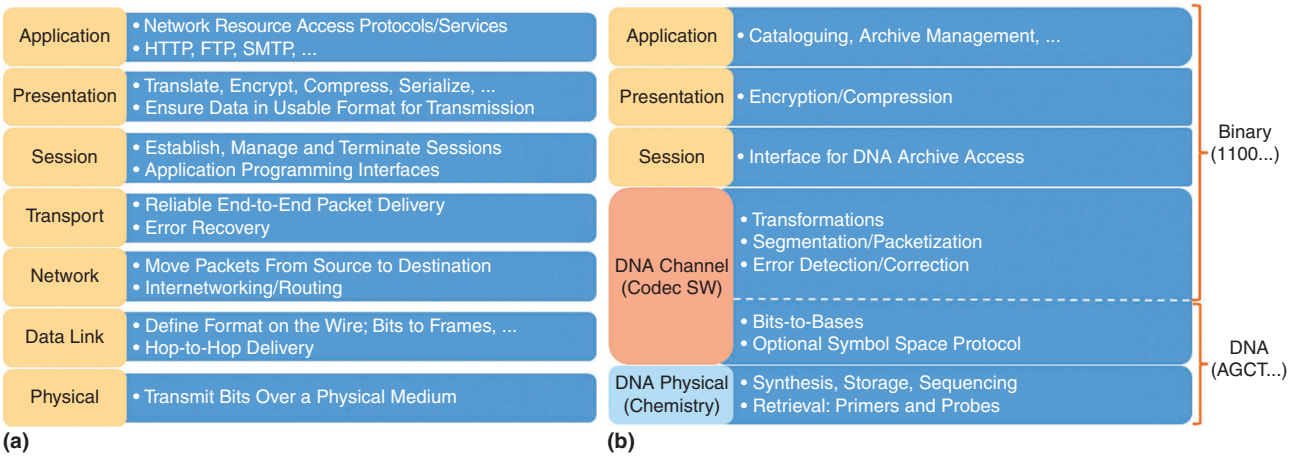

FIGURE 2. The layers of (a) the OSI model and (b) the DNA data storage model.

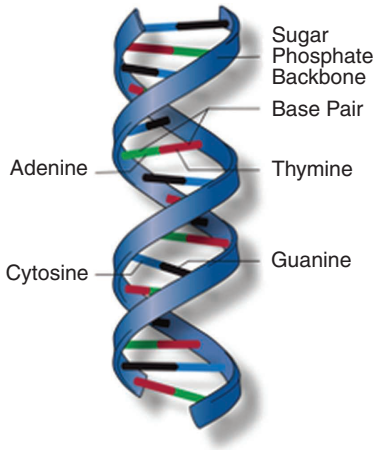

FIGURE 3. The DNA double helix. (Source: National Human Genome Research Institute; used with permission.)

model effectively unchanged in function; at this level of abstraction, DNA is simply another storage medium.

The application layer is the interface closest to users. For example, since DNA data storage is best matched to long-term storage, the application layer is likely to define how data are logically organized for archival purposes, including metadata describing the data in the archive.

The presentation layer maps naturally to the preparation of bitstreams as input to the lower layers. It may include transformations such as encryption and compression. Like with other media, such functions may be accelerated by special-purpose (silicon-based) hardware support.

The session layer provides access to the DNA data storage interface. Various implementations are possible, but the most commonly discussed is object based. In this implementation, the interface provided by the session layer is basically an object store with a key-value schema. It offers basic primitives such as read/write of individual objects or all objects, as well as primitives with more complex semantics, such as indexed search. These commands translate to logical and physical storage operations at lower layers.

THE DNA DATA STORAGE CHANNEL LAYER

The DNA channel layer takes as input a bitstream or other digital object from the session layer and processes those bits to ensure that the DNA sequences written and read by the physical layer can be successfully decoded, enabling recovery of the original digital source data. The DNA channel layer is implemented as a software codec.[2,3,4,5,6,7,11]

The DNA channel [Figure 5(b)] shares conceptual characteristics with a more traditional network/electrical channel [Figure 5(a)]. The DNA channel layer roughly incorporates the functionality of the transport through the data link layer in the OSI model. It receives a bitstream from the session layer, preparing it for handoff to the lower layers for the conceptual equivalent of "transmission," which, in the DNA case, means writing, storing, recovering and reading DNA molecules. As there is no physical "wire" in DNA data storage, the data link layer functionality is embedded in the DNA codec, instead of in the transmitter/receiver pair (Figure 5, dashed lines).

When the bitstream is presented to the DNA channel layer, the following types of operations are performed, not necessarily in this order.

Packetization: In a network or fabric, the key function of the transport and network layers is packetization, routing, and flow control across the link. Neither routing nor flow control are relevant here, but packetization is. The longest strand of synthetic DNA that can currently be constructed base by base with standard chemistry, while maintaining sufficient accuracy, is around 300 bases. Even if every base in a strand could be used for payload data, a typical strand would encode tens of bytes. Thus, like in traditional networking, where bitstreams must be broken into smaller pieces to fit limited-size packets, DNA data storage requires breaking data objects into many segments during encoding to fit limited-size strands. Reassembling these segments in the right order when sequencing requires adding indices to each segment before synthesis. The need for segment indices (and other protocol fields) introduces a tradeoff between the number of segments used to represent an object in a DNA pool (that is, maximum object size) and the number of bases used for the segment index (that is, index overhead).

Error correction: In constructing segments for storage, the DNA codec must add redundant information for error correction because random errors (base insertions, deletions, and substitutions) and erasures (missing sequences) may affect such segments through the DNA physical layer. Error correction can augment segments with additional data (inner code) or add segments that contain additional data (outer code). After the DNA is sequenced, the DNA codec uses this redundant information to recover lost sequences, correct remaining errors, and reliably deliver the original bitstream back to the upper layers.

Translation: To store digital data in DNA, the data must be translated from digital (0 or 1) to bases (A, T, C, or G). Although it is possible to simply map every two bits in a bitstream into one of the four possible bases, there are advantages to other mappings. For example, mappings of longer bitstreams to longer base sequences may better accommodate certain error correction methods and be more space efficient.

Transformations: With DNA, patterns of repeated bases

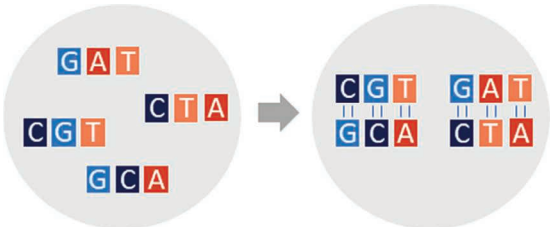

FIGURE 4. Hybridization: A process in which complementary bases bind to each other.

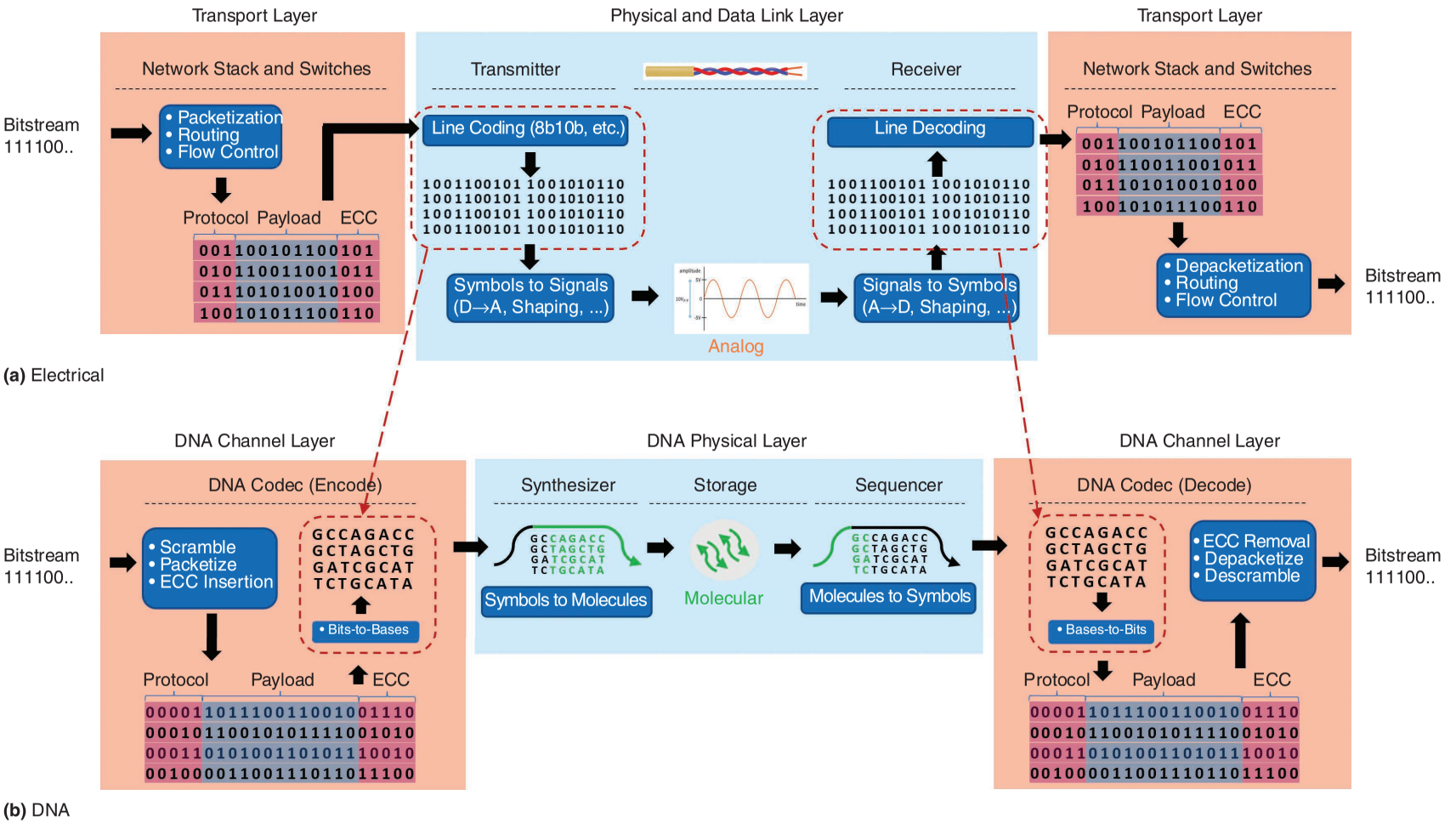

FIGURE 5. Comparing (a) the electrical channel model with (b) the DNA channel model.

(homopolymers), a high proportion of Gs and Cs (high GC content), and some other specific patterns can cause errors. For example, some sequencing techniques exhibit errors with long homopolymers, and high GC content may affect sequencing preparation protocols. The transformations in the DNA channel layer avoid transforming digital bitstreams into such problematic patterns of bases, thereby reducing downstream errors and associated error correction overhead, and improving overall data reliability. This is conceptually similar to the actions in a network/electrical channel to mitigate analog effects. For example, a serial link maintains close to an equal number of ones and zeroes on the wire.

Optional DNA space protocol: In addition to the actions above, the codec may insert additional base sequences into the already transformed digital data (process not shown in Figure 3). The purposes are use-case specific but generally involve random access to or search within the DNA archive. An example is the addition of an object ID, assigned by the session layer to sequences belonging to the same object. The section "Retrieval" discusses this further.

After these steps, the resulting DNA sequences are handed over to the DNA physical layer.

In summary, the DNA channel layer encodes an input bitstream so that the DNA sequences that are sent to the physical layer can be efficiently and successfully decoded after "transmission" (synthesis, storage, retrieval, and sequencing). In doing this, the DNA channel uses transformations and processing steps that are well structured and similar in nature to those used in existing storage systems. In the next section, we discuss the DNA physical layer.

THE DNA DATA STORAGE PHYSICAL LAYER

The DNA physical layer [Figure 2(b)] involves the basic chemistry of DNA for writing (synthesis), physically preserving (storage), and reading (sequencing) DNA molecules. It also involves the basic chemistry of DNA for retrieval, which is both a general preparation step for sequencing and also an integral step in how systems implement DNA storage operation requests from the upper layers. We will first briefly discuss synthesis, storage, and sequencing and then cover retrieval as it is central to the functioning of logical storage operations for DNA data storage.

In summary, the DNA channel layer encodes an input bitstream so that the DNA sequences that are sent to the physical layer can be effectively and successfully decoded after "transmission."

Synthesis

DNA synthesis has historically been implemented with large machines, using hoses, valves, and plastic containers known in the life sciences as well plates. Hoses and valves will not disappear; however, advances in the semiconductor industry have enabled silicon technology-based implementations that miniaturize the synthesis device and can achieve higher write throughput by synthesizing more sequences in parallel (thousands in well plates versus billions on chips). For example, chips akin to static RAM arrays have been used to synthesize sequences of DNA, one sequence per array position[8] (Figure 6).

Like other storage technologies, scaling relies on miniaturization of these positions so that more fit onto a chip, but instead of increasing capacity, more positions (synthesis sites) on a DNA synthesis chip increase write throughput, as more sequences can be synthesized simultaneously.

Sequencing

The two main methods of DNA sequencing considered for DNA data storage today are sequencing-by-synthesis (SBS) (Figure 7) and nanopore sequencing (Figure 8). SBS[20,21] indirectly identifies the bases in a source ssDNA strand by reconstructing a dsDNA strand from that source. As each complementary base is attached in building the dsDNA strand, the attachment event enables identification of the original source base, usually by optical means. Nanopore sequencing[22,23] detects bases directly (no dsDNA construction) by measuring disturbances in ionic current as a DNA strand is guided through a tiny opening (nanopore). Nanopores can be biological or, in development, solid state. A main tradeoff between the two methods is that SBS offers higher accuracy but typically requires a batched process that

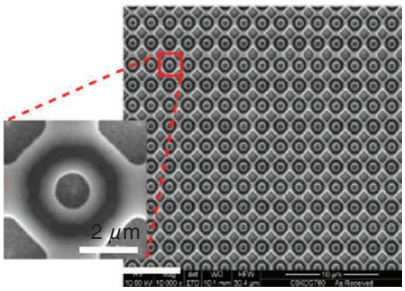

FIGURE 6. An electrochemical DNA synthesis array at 2 μm pitch.[8]

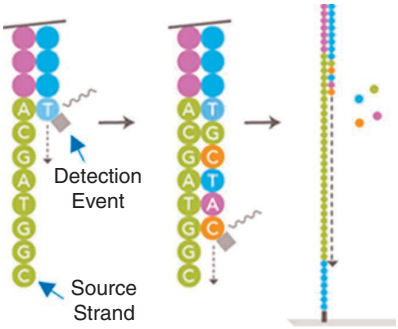

FIGURE 7. SBS sequencing. (Source: Illumina, Inc.; used with permission.)

results in high latency, while nanopore sequencing offers lower latency at typically lower accuracy. Both methods require physical retrieval and preparation of the molecules for reading that may add to the readout latency.

As with synthesis, further scaling in sequencing is needed to make DNA data storage practical, but the fundamentals for both synthesis and sequencing are now in place on semiconductor-based platforms, and scaling such platforms is something for which the technology industry has an impressive track record.

Storage

After being written and before being read, the DNA molecules need to be stored in an environment that prevents them from degrading.[9,10,11] While a wide variety of preservation methods are common in biotechnology broadly, DNA used for data storage is typically stored dry and in a chemically inert environment to increase storage density and stability. Significantly, with such methods, DNA for data storage shows potential for extremely long endurance at room temperature, supporting reliable long-term storage that is low cost and sustainable. We expect that large pools of DNA molecules (say, terabytes of storage) will be organized into even larger libraries where pools will be addressed using a physical coordinate system,[12] like tape libraries today.

Retrieval: Probes for storage operations

Retrieval defines how DNA molecules are extracted from a DNA archive and prepared for sequencing. The most basic operation is to read the entire archive, which is equivalent to reading all data in a tape from beginning to end. However, when a pool of DNA contains multiple objects, random access operations (for example, "seeking" the location of an object or searching for sets of objects) need to be performed on the pool.

There are a variety of methods to implement such operations, but two popular methods are PCR[3,13,14] and DNA pull-out with magnetic nanoparticles[15] (Figure 9).

PCR requires the session layer to assign a set of object IDs, which, at encoding time, the channel layer appends as a set of predefined DNA base sequences to both ends of every DNA sequence belonging to an object in the archive; these base sequences are referred to as target sites. At retrieval time, probes, that is, short DNA sequences that perfectly complement the target site of one or more objects, attach to the target molecules (recall that A pairs with T, and C with G) and, along with special enzymes called polymerases, kickstart the PCR process. Over multiple PCR cycles, the attached probes (called primers when used in PCR), enable the polymerases to copy only the target DNA molecules, making them more abundant than molecules representing other objects.

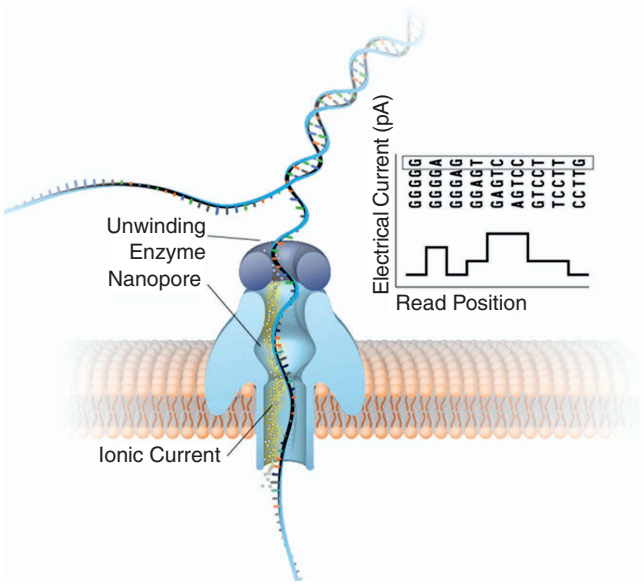

FIGURE 8. Nanopore sequencing. (Source: National Human Genome Research Institute, https://www.genome.gov; used with permission.)

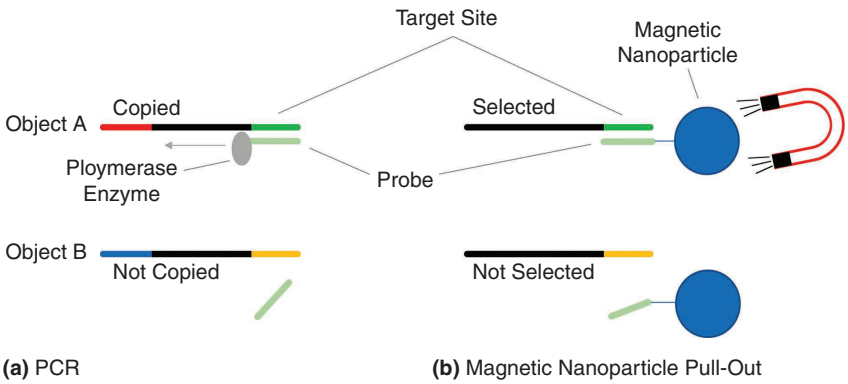

FIGURE 9. Two random access methods: (a) Probe attaches to the target site so that only the DNA sequences with the Object ID are copied during PCR. (b) Probe attaches to the target site of sequences with the Object ID, and magnetic nanoparticles used for molecular pull-out with a magnet.

The result is a pool in which most of the DNA molecules represent the object(s) to be read.

DNA pull-out with magnetic nanoparticles uses the same principle of DNA attachment but a different method to make the molecules representing the object of interest more numerous in a solution. With this process, the channel layer only needs to append the object IDs as target sites at one end of every DNA sequence belonging to an object. At retrieval time, the probes are attached to magnetic nanoparticles. As the probes bind to target DNA molecules, the target DNA molecules can then be separated from the rest of the DNA molecules in the pool with a magnet.

For either random access method, additional preparation steps such as further PCR steps and DNA cleanup may be required, depending on the sequencer used for reading the information out.

Tradeoffs between different chemical implementations of object random access along dimensions such as reliability, preparation, and reading overheads are an active area of research for DNA data storage,[16] as well further work on implementing even more advanced operations such as search,[17] content similarity search,[18] file preview,[19] etc.

This discussion has shown how, for the same read-object request from the session layer, different mechanisms can be used at the lower channel and physical layers, similar to how two NAND flash storage devices are implemented differently despite using the same command interface.

While DNA as a storage medium has fundamental differences from traditional storage, many of the data transformations and error processing considerations for DNA data storage have analogies to transmitting data through "traditional" network/storage electrical channels. It is our hope that framing DNA data storage implementation methods in the OSI layered storage model will help the nascent DNA data storage ecosystem evolve. C

ACKNOWLEDGMENT

We thank Kyle Tomek, John Hoffman, Damien Le Moal, and Luis Ceze for their valuable contributions and feedback on this manuscript. We also thank our respective research teams and collaborators for their substantial contributions to this field. Dave Landsman is the corresponding author.

REFERENCES

1. "Preserving our digital legacy: An introduction to DNA data storage," DNA Data Storage Alliance. [Online]. Available: https://dnastoragealliance.org/dev/wp-content/uploads/2021/06/DNA-Data-Storage-Alliance-An-Introduction-to-DNA-Data-Storage.pdf
2. N. Goldman et al., "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, vol. 494, no. 7435, pp. 77–80, Feb. 2013, doi: 10.1038/nature11875.
3. L. Organick et al., "Random access in large-scale DNA data storage," *Nature Biotechnol.* vol. 36, no. 3, pp. 242–248, Mar. 2018, doi: 10.1038/nbt.4079.
4. S. R. Srinivasavaradhan, S. Gopi, H. D. Pfister and S. Yekhanin, "Trellis BMA: Coded trace reconstruction on IDS channels for DNA storage," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2021, pp. 2453–2458, doi: 10.1109/ISIT45174.2021.9517821.
5. G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, vol. 337, no. 6102, p. 1628, Sep. 2012, doi: 10.1126/science.1226355.
6. W. H. Press, J. A. Hawkins, S. K. Jones Jr., J. M. Schaub, and I. J. Finkelstein, "HEDGES error-correcting code for DNA storage corrects indels and allows sequence constraints," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 31, pp. 18,489–18,496, Aug. 2020, doi: 10.1073/pnas.2004821117.
7. Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," *Science*, vol. 355, no. 6328, pp. 950–954, Mar. 2017, doi: 10.1126/science.aaj2038.
8. B. H. Nguyen et al., "Scaling DNA data storage with nanoscale electrode wells," *Sci. Adv.*, vol. 7, no. 48, Nov. 2021, Art. no. eabi6714, doi: 10.1126/sciadv.abi6714.
9. D. Coudy, M. Colotte, A. Luis, S. Tuffet, and J. Bonnet, "Long term conservation of DNA at ambient temperature. Implications for DNA data storage," *PLoS One*, vol. 16, no. 11, Nov. 2021, Art. no. e0259868, doi: 10.1371/journal.pone.0259868.
10. L. Organick et al., "An empirical comparison of preservation methods for synthetic DNA data storage," *Small Methods*, vol. 5, no. 5, May 2021, Art. no. 2001094, doi: 10.1002/smtd.202001094.
11. R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, and W. J. Stark, "Robust chemical preservation of digital information on DNA in silica with error-correcting codes," *Angewandte Chemie*, vol. 54, no. 8, pp. 2552–2555, Feb. 2015, doi: 10.1002/anie.201411378.
12. S. Newman et al., "High density DNA data storage library via dehydration with digital microfluidic retrieval," *Nature Commun.*, vol. 10, no. 1, Apr. 2019, Art. no. 1706, doi: 10.1038/s41467-019-09517-y.
13. S. M. Tabatabaei Yazdi, Y. Yuan, J. Ma, H. Zhao, and O. Milenkovic, "A rewritable, random-access DNA-based storage system," *Scientific Rep.*, vol. 5, Sep. 2015, Art. no. 14138, doi: 10.1038/srep14138.
14. C. Winston, L. Organick, D. Ward, L. Ceze, K. Strauss, and Y. J. Chen, "Combinatorial PCR method for efficient, selective oligo retrieval from complex oligo pools," *ACS Synthetic Biol.*, vol. 11, no. 5, pp. 1727–1734, Feb. 2022, doi: 10.1021/acssynbio.1c00482.
15. K. N. Lin, K. Volkel, J. M. Tuck, and A. J. Keung, "Dynamic and scalable DNA-based information storage," *Nature Commun.*, vol. 11, no. 1, Jun. 2020, Art. no. 2981, doi: 10.1038/s41467-020-16797-2.
16. K. J. Tomek et al., "Driving the scalability of DNA-based information

storage systems," *ACS Synthetic Biol.*, vol. 8, no. 6, pp. 1241–1248, May 2019, doi: 10.1021/acssynbio.9b00100.
17. J. L. Banal et al., "Random access DNA memory using Boolean search in an archival file storage system," *Nature Mater.*, vol. 20, no. 9, pp. 1272–1280, Sep. 2021, doi: 10.1038/s41563-021-01021-3.
18. C. Bee et al., "Molecular-level similarity search brings computing to DNA data storage," *Nature Commun.*, vol. 12, no. 1, Aug. 2021, Art. no. 4764, doi: 10.1038/s41467-021-24991-z.
19. K. J. Tomek, K. Volkel, E. W. Indermaur, J. M. Tuck, and A. J. Keung, "Promiscuous molecules for smarter file operations in DNA-based data storage," *Nature Commun.*, vol. 12, no. 1, Jun. 2021, Art. no. 3518, doi: 10.1038/s41467-021-23669-w.
20. C. Fuller et al., "The challenges of sequencing by synthesis," *Nature Biotechnol.*, vol. 27, pp. 1013–1023, Nov. 2009, doi: 10.1038/nbt.1585.
21. S. Goodwin, J. McPherson, and W. McCombie, "Coming of age: Ten years of next-generation sequencing technologies," *Nature Rev. Genetics*, vol. 17, pp. 333–351, Jun. 2016, doi: 10.1038/nrg.2016.49.
22. M. MacKenzie and C. Argyropoulos, "An introduction to nanopore sequencing: Past, present, and future considerations," *Micromachines*, vol. 14, no. 2, 2023, Art. no. 459, doi: 10.3390/mi14020459.
23. D. Branton et al., "The potential and challenges of nanopore sequencing," *Nature Biotechnol.*, vol. 26, pp. 1146–1153, Oct. 2008, doi: 10.1038/nbt.1495.

DAVE LANDSMAN is a distinguished engineer at Western Digital Research, Milpitas, CA 95035 USA. Contact him at dave.landsman@wdc.com.

KARIN STRAUSS is a senior principal research manager at Microsoft Research, Redmond, WA 98052 USA and an affiliate full professor at the University of Washington, Seattle, WA 98195 USA. Contact her at kstrauss@microsoft.com.

Over the Rainbow: 21st Century Security & Privacy Podcast

Tune in with security leaders of academia, industry, and government.

OVER THE RAINBOW
by IEEE Security & Privacy

Subscribe Today
www.computer.org/over-the-rainbow-podcast

Digital Object Identifier 10.1109/MC.2023.3277764