Storing data in the backbone of DNA

Sim Superville[1]

[1]Affiliation not available

May 5, 2020

DNA has stored 'data' for millions of years, in the form of the 4 bases; A, T, G, and C. Synthetic attempts to recreate this huge data storing capability have promised the same levels of storage density and stability, but are yet to compete with current optical/magnetic storage devices due to the high error rates and large costs. A novel approach has been found, storing the data in the backbone of native DNA instead of the bases themselves, this greatly reduces costs, while reducing the error rate to 0. This new technique is also capable of bitwise random-access memory, opening the door to the world of molecular computing.

In nature, DNA is found as a double-helix ladder compromising of billions of molecular building blocks known as bases (A, T, G, and C), coding instructions for proteins. The closely-packed nature of these nucleotides enable DNA to code for huge quantities of data; one gram of DNA is capable of holding 455 exabytes of data[1] (1 exabyte = 10^{18} bytes), enough storage for all of the data produced by every major tech company... with room to spare. Traditional data storage methods are reaching their physical limits, with hard-drives being restricted to 1 terabyte (10^{12}bytes) per square inch[2]. With the exponential growth of big data, alternative data storage methods such as DNA are starting to be explored, made possible due to the huge advances in DNA sequencing technology.

Current approaches of DNA synthesis-based storage[2–4]assign segments of binary code to specific nucleotide sequences, these are then synthesised *in vitro* using enzymes to bind the nucleotides together, iterating through many cycles to form the string of DNA. The information encoded in the DNA can then be retrieved via next generation sequencing (NGS) or third generation nanopore sequencing. Nanopore sequencing entails feeding the string of DNA through a tiny hole, much like a thread going through a needle, reading each nucleotide as it passes through. Advancements in sequencing technologies allow for large sequences to be read at a relatively low cost; the total cost of the first human genome was [?] £2.07 billion[2], whereas sequencing a full genome now costs around £1000[2].

However, despite the promise of synthesis-based DNA storage, issues arise during the synthesis step. The first issue is that adding one nucleotide per cycle takes hours to complete a full sequence, making it very slow and expensive compared to traditional optical and magnetic writing techniques. Secondly, DNA synthesis is naturally a very error prone process, with 1% of bases containing substitution (incorrect base attached) or indel (insertion or deletion of a base) errors[3]. This is a big issue when dealing with large volumes of DNA as errors are inevitable, this can have dire repercussions when dealing with compressed file formats. Finally, DNA synthesis machinery is limited to creating relatively short sequences, reducing the amount of readouts/ data, per fragment[5].

Tabatabaei et al. [5] propose a novel method to storing data using DNA, via a topological 'nicking' approach instead of direct synthesis. Therefore, the data is not stored in the nucleotides themselves, but the sugar-phosphate backbone of the DNA. This enables you to use a known sequence of DNA which you can map to

a reference genome, bypassing the aforementioned issues associated with synthesis-based storage.

> **" The very same medium that literally specifies who we are as individuals might also store our art, our culture, and our history as a species."[2]**

Enzymatic nicking involves using a highly-specific and versatile endonuclease enzyme (*Pf* Ago) to cut a single bond in the sugar-phosphate backbone (fig.1.a), forming fragments which can be recognised using NGS/nanopore sequencing. Figure 1. outlines the enzymatic nicking process and visualises the writing/reading aspect of nicking -related DNA storage.

The *Pf* Ago enzyme used by Tabatabaei *et al.* is a highly accurate artificial restriction enzyme[5] with a high turnover rate, enabling one enzyme to produce hundreds of nicks. Normally, *Pf* Ago cleaves both strands of DNA when it binds to the recognition site. However, this is unnecessary with regards to storage as it requires the presence of two guide DNA (gDNA) molecules in close proximity. Tabatabaei found that under the correct conditions (e.g. buffer and temperature) *Pf* Ago can target only one of the DNA strands, using a single gDNA, allowing efficient and precise nicking of that strand simultaneously, in under 40 minutes. This greatly reduces the writing latency expressed by other DNA-storage methods. The lack of errors when writing also means that high coverages are not required for accurate data readout, as the reads are easily mapped to the reference. By decreasing the coverage, costs associated with sequencing also decrease, increasing the cost-effectiveness of this technique. To prove that enzymatic nicking worked, the researchers[5]compressed and converted two files into ASCII format, and retrieved them with 100% accuracy (not possible with synthetic DNA storage). The first was a 0.4KB text file containing Lincoln's Gettysburg Address (272 words), and the second was a 14KB JPEG image of the Lincoln Memorial. Both were successfully retrieved using NGS, however solid-state nanopore sequencing could also have been used.

An issue with nanopore sequencing is the lack of translocation controls, making the nick reads noisy. To combat this, Tabatabaei *et al.*used *Pf* Ago to extend nicks to 'toehold' regions, consisting of two nicks placed closely together, forming a short segment of DNA. This is easily recognisable by solid-state nanopores, allowing for quick and clear reads.

Toeholds also allowed Tabatabaei *et al.* to achieve something synthetic DNA storage methods could not; Bitwise Random Access. This makes performing molecular computations possible, using toeholds as initiation sites to allow the controlled binding/release of DNA strands. This ability was then enhanced by the addition of fluorescent tags which enabled the visual recognition and retrieval of a specific toehold.

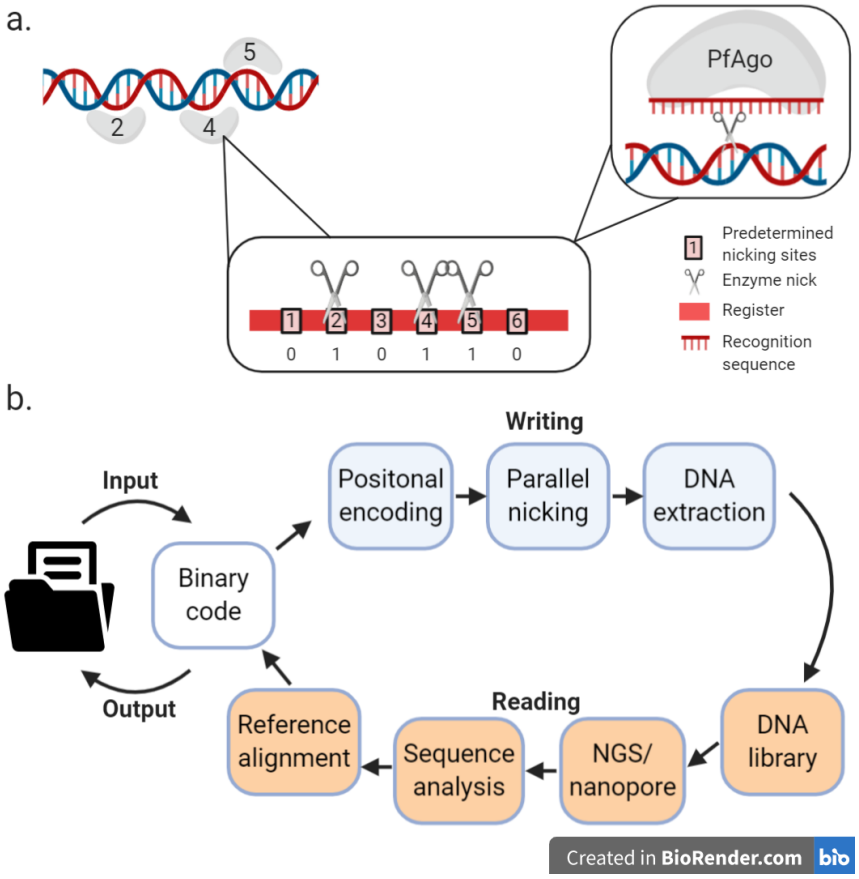

Figure 1 | Enzyme nicking based DNA storage. Tabatabaei *et al.* proposed the use of enzymatic nicking to write and read binary encoded data. **a.** Predetermined nick sites on the DNA register are either left un-nicked (0) or nicked (1) by the *Pf* Ago enzyme, transcribing the inputted binary code. Multiple enzymes can bind in parallel at the same time, enabling rapid transcription of the input data into the register. Each *Pf* Ago enzyme has a specific recognition sequence which ensures the correct bond is nicked, greatly reducing writing errors. **b.** The binary information from the input file is first encoded into predetermined nicking sites, where the sites representing a '1' are nicked. The native DNA is then extracted, and the resulting single-stranded DNA products of different lengths are sequenced. The sequence data is then analysed and mapped against the native DNA reference, allowing the position of the nicks to be identified via the size of the fragments. This is then translated back into binary code, allowing for recovery of the file. (figure created using Biorender®)

The tag is easily removed upon hybridization of the DNA, making the readout process non-destructive[5].

The computational process of enzymatic nicking can be further enhanced by the simultaneous use of multiple registers and sense-antisense recording. The order of the multiple registers is dictated in the genome, allowing for greater storage space in a retrievable format. Sense-antisense recording utilises parallel nicking, this further enhances the storage capabilities of this technique. It involves using *Pf* Ago to nick both DNA strands, converting the binary format to a ternary format, enabling more information to be encoded in the same space. In other words, for each nicking site, no nick signifies 0, a nick on the sense strand represents 1, and a nick on the anti-sense strand represents 2. This technique enables the storage of [?] 4 Exabytes of data per gram, which is somewhat shy of the 200 Ebytes/g capability of synthesis-based DNA storage methods, but nonetheless blows conventional data storage capacities out the water.

The computational potential of this work is what sets it aside from previous DNA-based storage methods.

Previously, modifying stored data in synthesised DNA involved sequencing the data encoded, altering it on a separate (traditional) computer, then writing it into a new DNA molecule. Whereas enzymatic nicking utilises strand displacement (made possible by toeholds), making in-memory parallel computations possible without the need to synthesise new DNA.

The major problem for this technique lies in its cost. It may be cheaper than DNA-synthesis strategies but is still a long way away from cost-efficient scaling which can rival its mechanical counterparts. Despite this, further optimizations in DNA technologies will see a sharp decrease in cost, much like the cost of sequencing the human genome. However, with that in mind, these future advances are not limited to DNA nicking. DNA-synthesis strategies will also improve[4](arguably at a faster rate than the DNA nicking approach), possibly reaching an error rate to rival that of nicking, making one of its key arguments redundant. Future approaches may also combine the two methods; writing data in synthetic DNA, then utilising the flexibility of nicking to encode metadata which is easily modified through ligation/toeholds. The future looks promising for DNA-storage, but only time will tell if it becomes another expensive novelty, or the solution to the world's big data crisis.

Simeon Superville

References

1. Aron, J. DNA in glass – the ultimate archive. *New Sci.***225** , 15 (2015).

2. Milenkovic, O., Gabrys, R., Kiah, H. M. & Yazdi, S. M. H. T. Exabytes in a Test Tube: With the right coding, DNA could archive our entire civilization. *IEEE Spectr.* **55** , 40–45 (2018).

3. Palluk, S. *et al.* De novo DNA synthesis using polymerasenucleotide conjugates. *Nat. Biotechnol.* **36** , 645–650 (2018).

4. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science (80-.).* **355** , 950–954 (2017).

5. Tabatabaei, S. K. *et al.* DNA Punch Cards: Encoding Data on Native DNA Sequences via Nicking. *bioRxiv* 672394 (2019). doi:10.1101/672394