Monograph - DNA Data Storage & Indexing

Rui Fernandes

***Abstract* –As the growth of the Internet becomes exponential so does the need for the storage and handling of the data that composes it. There are areas that dedicate constant research and advancements on perfecting and innovating data storage. One such more and more emerging area is the DNA-based information storage paradigm that brings the digital and biotechnological worlds together. This small monograph intends to bring a more digestible look into the evolution of this area based on a surface analysis of 3 papers.**

***Keywords* –DNA, Storage, Indexing, Data**

I. Introduction

Before going into what the methods and objectives of the three chosen papers of this are, it's important to contextualize a few concepts that surround this topic.

Firstly we must remember what is DNA and what does it do. DNA, or Deoxyribonucleic acid, is a molecule composed of two chains forming the universally known double helix. It is essential for all known living organisms as it carries all the genetic information needed for the functioning, growth and reproduction of it's given organism, it is, in essence, our code. DNA strands are composed of many simpler units called nucleotides in a chain, these are composed, amongst other components, of one of the four nucleobases:

- Cytosine [C].
- Guanine [G].
- Adenine [A].
- Thymine [T].

It's also important to note the existence of RNA which is made up of a single strand and it's Thymine components are substituted with Uracil [U] nucleobases.

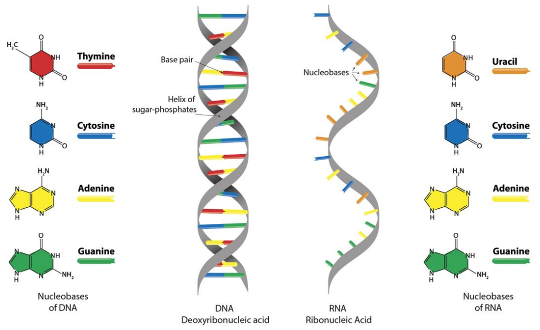

Fig. 1 - DNA and RNA composition

Next, it's relevant to understand that in the biotechnology and genetic studies and research along the decades, DNA analysis, synthesis and manipulation became widespread, accessible and cheaper. Nowadays it's the techniques are firmly established and continue to improve, it's possible to order DNA strands online from international companies that synthesize it per request.

But why is DNA relevant to digital data storage? Well, the ever growing digital world is in a struggle to store data in a compact and efficient way. While current main-stream technologies continue to improve, there are various alternatives surging, as the typical methods require materials that are limited and can be depleted, such as silicon. On a small scale it might not seem like a problem, a relatively small hard-drive can easily store one Terabyte or more. But in comparison, now looking at DNA, a single gram of DNA strands contains 215 Petabytes, for an even more hypothetical look at it, the entire internet was estimated to be 1 million Exabytes in 2014, this means that the entire internet would have fit inside 4.6 metric tons of DNA, which is less than an African Elephant.

Of course harnessing and using DNA with the intent of digital storage isn't that easy and isn't, currently, without many faults. But research keeps speeding up in this area, and this monograph will now focus on 3 research papers that focus on DNA storage and computing and show evolution in less than a decade.

II. Information storage in synthesized DNA

The first analysed research paper, from 2013, goes into the possible advantages of DNA as medium of storage for digital information, adding to the already mentioned the fact that DNA should have an exceptionally long lifespan in low-maintenance environments.

This paper functioned a lot as a proof of concept, to truly demonstrate at the time that it was possible to store digital information with DNA.

The methods were simple enough on a surface level, they chose 5 common computer files that totaled 739 kilobytes of information, then taking the bytes that comprised them, translated them to a Base 3 encoded format that could be directly transposed to DNA encoding.

A strong point to mention about the methods is the insistence that what was being produced was clearly synthetic DNA as it lacked homo-polymers, sequences of identical bases, since those are more prone to error.

The transposed DNA sequences still needed more steps before synthesis though, first they need to be broken up into many pieces and replicated as to generate overlap between the information for error correction. After, a simple component is added to each fragment to keep information of what file it belongs to and it's position in the chain.

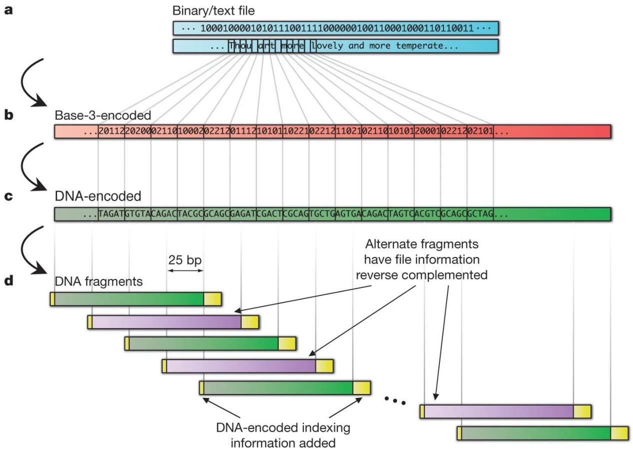
a
Binary/text file
··· 100010000101011100111100000010011000100011011001 1···
...Thou art more lovely and more temperate...
b
Base-3-encoded
...20112202000211010002022120111210101102210221211021021101010120001022120210 1...
c
DNA-encoded
...TAGATGTGTACAGACTACGCGCAGCGAGATCGACTCGCAGTGCTGAGTGACAGACTAGTCACGTCGCAGCGCTAG...
d
DNA fragments
25 bp
Alternate fragments have file information reverse complemented
DNA-encoded indexing information added

Fig. 2 - Methodology used in the first paper

The strands would then be synthesised and kept in an appropriate environment until the information was needed for access. At which point the reverse operation needed to happen, all the DNA strands would be reconstructed and sequenced, this is, recovering the DNA encoding. This whole operation is done *in silico*, a term in biology that refers to experiments done with a computer, with software.

From there it's simple enough to translate the sequence all the way back to a bytes file, and in fact all 5 files were reconstructed with 100% accuracy.

While it is a revolutionary process, it has many blatant problems working against it at this point. The first being that to get any information from the DNA constructed database, the whole database needs to be sequenced and reconstructed, this issue makes it impractical at best to scale it up to useful data amounts. The other notable issue is that the exposed process is reliant on huge overlap, a single partition of information is represented in 4 synthesised fragments of DNA, this once again would be a massive problem for scaling to bigger files, having all information occur 4 times.

III. Random access

The second paper, from 2018, improves on the now established methodology for the usage of DNA as storage for digital files. While it does improve a lot on error correction and having less necessity for overlap when compared to studies before it, the main take from this paper is the approach to implement Random access to the medium.

To solve the necessity for sequencing the whole database to get a single portion of information, the authors of the paper devised a way for random access to be applied to DNA storage, through PCR primers.

PCR, or Polymerase chain reaction, is a method widely used to rapidly make millions to billions of copies of a given DNA strand, and it relies on primers, small sequences to attach to the said strand. For the studies purpose these became useful as a sort of ID to the strands, a better one than before, as these can be more easily identifiable and reacted with.

The researchers made a library of unique primers that would be dissimilar from each other, amongst many other restrictions. Then by applying a single one per DNA strand that represented the digital files, the files would be more easily recoverable and then the single wanted one sequenced and decoded.

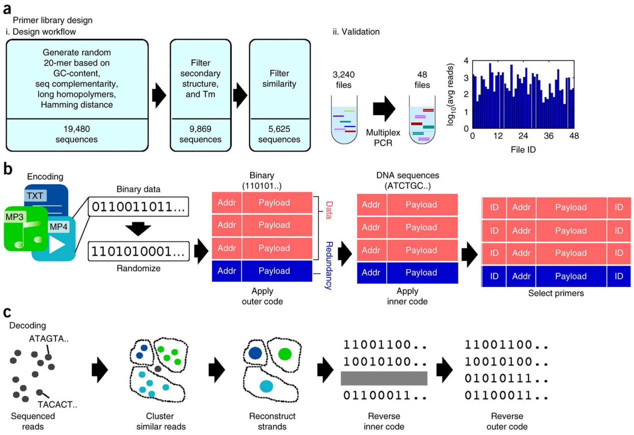

Fig. 3 - Methodology used in the second paper

IV. Indexing and Information Retrieval

Until now only DNA data storage methods were really covered, but the third paper takes DNA computing in a different direction, keeping files digital, but making the index and information retrieval process completely based on DNA.

The researchers operate with an image database that includes 1.5 million images, and don't intend to store them, but want to make a recommendation system that would recover similar images to a query image, while relying on DNA computing technologies.

As a surface level analysis, the images are computed, through machine learning, into a geometric space composed of feature vectors that represent document similarity by physical closeness.

Afterwards, these vectors are converted into DNA strands composed of both the feature vector and an ID much like the previous study that relies on PCR primers. These DNA strands are kept in a biological database ready to be queried.

To query the database, the input image is too converted to a feature vector and synthesised, with a biotin ending that will be used later for filtering.

The query strand is meant to interact with the database and form stable hybridised strands with the strands that represent images similar to it.

Finally the strands are recovered and then sequenced, post decoding the ids can be accessed and utilized in the digital database.

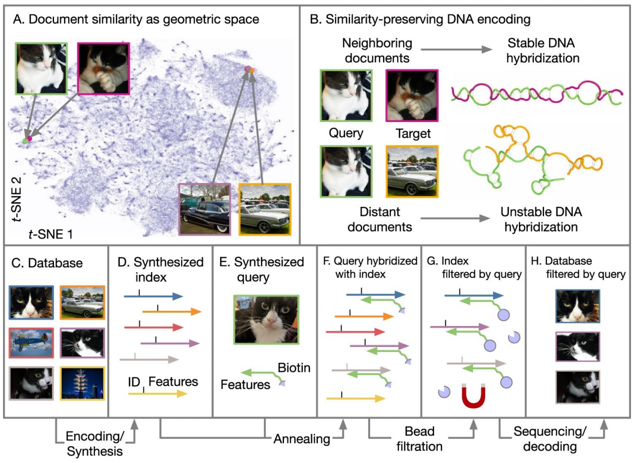

Fig. 4 - Methodology used in the third paper

V. Discussion

We can see that DNA is a relevant alternative to traditional data storage and indexing methods, and it is in constant evolution and improvement, so it could become a reality in years to come.

VI. References

[1] Goldman, N., Bertone, P., Chen, S. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature 494, 77–80 (2013). https://doi.org/10.1038/nature11875

[2] Organick, L., Ang, S., Chen, YJ. et al. Random access in large-scale DNA data storage. Nat Biotechnol 36, 242–248 (2018). https://doi.org/10.1038/nbt.4079

[3] Bee, C., Chen, YJ., Queen, M. et al. Molecular-level similarity search brings computing to DNA data storage. Nat Commun 12, 4764 (2021). https://doi.org/10.1038/s41467-021-24991-z

[4] Zhirnov V, Zadegan RM, Sandhu GS, Church GM, Hughes WL. Nucleic acid memory. Nat Mater. 2016 Apr;15(4):366-70. doi: 10.1038/nmat4594. PMID: 27005909; PMCID: PMC6361517.

[5] Polymerase chain reaction Wikipedia page https://en.wikipedia.org/wiki/Polymerase_chain_reaction

[6] DNA Wikipedia page https://en.wikipedia.org/wiki/DNA

[7] Forbes Article, DNA Data Storage Is About To Go Viral https://www.forbes.com/sites/johncumbers/2019/08/03/dna-data-storage-is-about-to-go-viral/

[8] Starry Article, How big is the internet? https://starry.com/blog/inside-the-internet/how-big-is-the-internet