IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of current version 20/12/2022.

Synthetic DNA as a Long Term Molecular Memory Solution

M. Garcia
Department of Electronic and Electrical Engineering, University of Bath, Bath BA2 7AY, United Kingdom

ABSTRACT Data produced globally keeps increasing yearly with conventional storage devices struggling to keep up. Synthetic DNA molecular memory offers an alternative sustainable solution at an order of six improvement on the best current long term storage technology. Advances in storage, sequencing, encoding, and synthesis offer an informationally dense, naturally redundant, and high longevity. With read speeds showing little in improvement, researchers have instead focused on bridging the gap between current and viable synthetic costs. This has been achieved through array-based synthesis. However, direct access of this technology is limited by the need for specialised protection from environmental degradation. The largest barrier to adoption has been shown to be the lack of automation and scalability. Despite these challenges, DNA long-term memory has renewed the exciting potential of biomolecular computing and 'living' circuitry. Until scalability becomes viable the research progression can only remain state of the art/developing. It is this researcher's opinion that chaperone proteins may hold the key to solving this.

KEYWORDS Synthetic DNA; Long Term Memory; Oligonucleotides

I. INTRODUCTION

With the ever-accelerating rate of technological progress and data accumulation comes the need to store such unprecedented amounts of data. Various memory solutions have been developed with the leading technologies of solid-state, optical, and magnetic making up most computerised memory. Current non-volatile memory is beginning to reach a miniaturisation limit. This physical size limit has begun to impose a limit on the information density that these classical solutions can achieve. Recent magnetic prototypes have demonstrated a density of 31 GB/cm2[1]. What is needed to keep up with this Sisyphean battle of data production are alternative methods that allow for sustainable memory devices. To be sustainable Lowering the energy cost, increasing durability, and increasing storage capacity are essential. In looking to achieve this gargantuan task, mother nature has developed her own solution: DNA memory. DNA is a self-replicating molecule that fitting the above criteria, containing all the information required to produce living organisms. The development of synthetic DNA memory allows scientists to make use of these properties to provide a sustainable data storage solution. DNA is a natural informational memory medium found in vivo (within living cells). Synthetic DNA generates artificial oligomers that mimic the encoding naturally found within DNA and RNA through sequencing of oligonucleotides.

DNA has benefited from evolutionary design, specifically in its longevity, naturally redundant, and density.

Longevity – DNA under ideal conditions is considered of lasting millions of years[2] before degrading due to the breakdown of bonds within the molecules. This is however not indicative of the informational degradation; the percentage of recoverable data is more dependent on physical redundancies and encoding method. When scientists unravel synthetic DNA, to combat degradation vectors the physical state of storage for the DNA offer a buffer to act as a protective material targeted at reducing the occurrence of mechanisms of damage.

Natural Redundancy – DNA inherently tries to replicate itself within living cells making it optimised for resource inexpensive and fast replication. DNA can be manipulated in a biosimilar method through using polymerase chain reactions (PCR) for use in vitro allowing for mass production of a single synthetic oligomers.

Density – DNA exhibits an informational density of 455 EB/g[3] making it 106 times more dense than current macro-storage technology.

II. PROCESS TO ACHIEVE SYNTHETIC DNA MEMORY SYSTEM

There are a series of steps to producing a DNA storage device. First, digital information such as files which are converted into an oligonucleotide representation that are synthesised into oligos (small DNA molecules) enzymatically or chemically and transferred into DNA pools (in vitro or in vivo). When ready to read the correct oligos are identified to be sequenced and then decoded to recreate the original data.

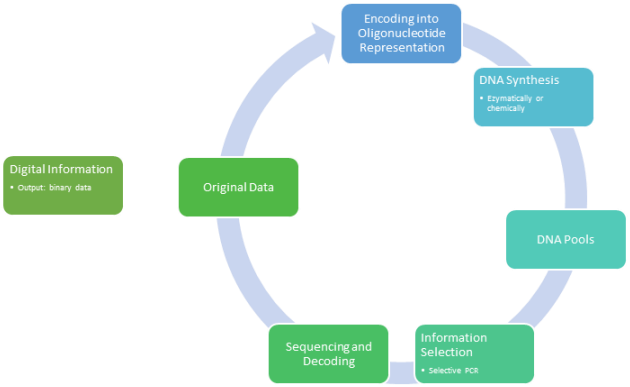

FIGURE 1. **Process to achieve synthetic DNA memory system**

A. ENCODING AND SYNTHESIS

DNA has its own natural encoding protocol consisting of four nucleotides that can be used to express a quaternary bit each, with a letter corresponding to the specified nucleotide.

- A – Adenine
- T – Thymine
- C – Cytosine
- G – Guanine

The synthetic DNA storage techniques aim to convert the binary information we wish to store into an arrangement of nucleotides in a strand of DNA. Each sub-sequence of four nucleotides represents a byte of data. Informational density can be increased introducing additional composite bases. Anvay et al.[4] introduced conventional composite bases M (50% G and 50% T) and K (50% A and 50% C) this improved the storage density of by 24% on the previous 4 nucleotide base alphabet. While this design can store more data per unit length, it needs a greater sequencing depth and more DNA copies to read the stored data. Thus, the reading operation becomes more resource costly. Hoshika et al.[5] synthesized eight new nucleotides (S, B, J, V, K, X, Z, and P) by modifying the chemical groups of natural nucleotides. However, this effort has been criticised as the introduction of orthogonal nucleotide pairs did not take into account the change in physiochemical DNA characteristics. This lead to decreasing read speeds and accuracy of current sequencing and retrieval technologies.

To this end, an alternative method for generating nucleotides was proposed by Chaput et al.[6] which involved modifying the fundamental nucleotides sugars. The produced nucleotides are called Xeno-nucleic acids which have a unique property of not hybridising with natural DNA. This is particularly important as this improves the informational density without risk of contamination when placed in vivo. It is important to note that this technology is limited by being difficult to sequence greatly reducing reading speed.

In practice there are errors[7] can be produced in any synthesis and sequencing step causing the need for extra design considerations. Errors have been mitigated using error correcting codes that have been attached to the end of the payload (the desired sequence to be written) to be synthesised into the oligos seen in Figure 1.

FIGURE 2. **Encoded sequence for synthesis**

DNA synthesis allows for a specified nucleotide sequence to be written and produced through either enzymically or chemically, resulting in the formation of oligos. Synthesising oligos enzymatically was first successful in 2010[8] with the use of terminal deoxynucleotidyl transferase. On the other hand, the chemical method uses phosphonamidite techniques. This synthesis technique can either be column based (traditional) or array based. Column based is well established however Array based is preferred to the well-established column based when producing many DNA strands because it has a higher throughput and lower resource cost[9]. This process allows multiple sequences being synthesised in parallel and leads to a lower reagent consumption during production. In addition, chemical synthesis can only achieve a produced oligo length of 300[9] or less. Above 300 in length and synthetic yield greatly decreases as synthetic errors are introduced, impeding the synthetic process. A longer oligo strand allows less of the strand to be devoted to indexing and ECC, boosting the information density of the system. However, research has shown a greater environmental resilience with small oligo length.

To get more accurate data recovery, the binary data is fragmented, and a relevant index is attached to the oligo sequence. In addition, newer methods have shown that storing the data as overlapping fragments in separate oligos may be a preferable method for large capacity storage. These indexing methods, combined with error correction such as Reed-Solomon codes[10], are added to either side of the payload to greatly increase the reliability of the memory system.

B. STORAGE

The high cost of synthesis and sequencing currently limits the long term data storage within synthetic DNA memory. Storage conditions are best optimised for long term stability of the synthetic DNA. The most relevant degradation modes for long-term storage application include oxidation, alkylation, hydrolysis, and UV radiation. Dehydration has been shown to reduce the molecular damage caused by hydrolysis on the phosphate backbone with efficacy at a storage time. The company DNA Stable has successful embedded DNA absorbed on to FTA filter cards into lyophilized powder stored silk matrices. The silk matrix provides a UV blocking effect. This increased the recoverable information from 20% unprotected to 80. Furthermore, salt has also been used as a stabilizer to maintain high loading of (>20 wt%)[11] DNA while still being reasonable accessible. This reduces the impact of humidity causing degradation. Above all, the DNA storage field's leading approach consists of using inorganic matrices of iron oxide, and or silica encapsulation[12] to scaffold the synthetic DNA. Grass et al. estimated that silica encapsulation could maintain DNA for 20–90 years at room temperature, 2000 years at 9.4 °C, and over 2 million years at −18 °C[13]. While effective, silica encapsulation has a few potential limitations. Firstly, the DNA must be unencapsulated to allow retrieval. When the read operation has occurred, the synthetic DNA must be encapsulated again. This increases the read and write time. Secondly, encapsulation inherently reduces the informational density. The best encapsulated solution to date is the layer-by-layer design with alternating DNA and cationic polyethyleneimine with a silica final encapsulation. This only sacrifices 1-2 orders of magnitude of informational density which makes it 4 orders of magnitude more dense than conventional storage and offering good robustness.

C. SEQUENCING

There are also errors that occur in the sequencing step in the sequencing of DNA are due to long repeated bases (especially homopolymers) and high GC content. Homopolymers, with more than six repeated nucleotides, have been shown to increase the error probability during deletion and insertion operations. Repeated GC content cause the formation of guanine tetraplex structures[14]. These structures can greatly increase the thermal stability but at the cost of sequencing instability. Oxford Nanopore Technologies[15] lead the charge in encoding technologies which can read more than 1000 base pairs per second. This was leveraged by Chen et al.[16] which attached short DNA hairpins to double-stranded stem lengths. However, with these hairpins, nanopore defects and fluctuations, noise and uncontrolled reading speed can all lead to wrongful addition of nucleotides. This indicates that rigorous error correction is necessary to improve this technology for DNA storage. Only by compensating for elimination errors and sequencing losses can this memory system be viable for long-term storage.

IV. FUTURE WORK

DNA is a sustainable memory storage system with a higher informational density, longer retention time, and lower power consumption. Although this memory storage methods have potential, there are many challenges that must be addressed for technology adoption. Future development is dependent on the development of the four major processes in the memory operation: encoding, synthesis, storage, and sequencing.

The first challenge involves the slow read speeds. The time taken to read data stored on synthetic DNA is expected to stay high. Long term storage can tolerate slower read speeds and has the added advantage of lower power consumptions due to long periods where no power is required. Therefore, research has begun to focus on maximising through put of synthesis and sequencing. Currently it remains challenging to enhance synthesis scale to that of conventional media when assessing decoding. It is likely that development of enzymatic synthesis can bridge this gap as the increase in speed, length and accuracy will reduce synthesis cost. Synthetic cost is currently the greatest barrier for wider adoption[17]. High throughput sequencing technology has continued to make significant advancements. Ultimately, significant technical improvements to synthesis and sequencing must be achieved to reduce the cost of memory storage of synthetic data to an adoptable level.

The second challenge is involved with the physical storage and preservation of synthetic DNA molecules. Factors such as humidity, high temperatures, and UV light are the greatest contributors[18]. Encapsulation has shown the best results however this method prevents direct access to the data. This has led to the continued search for alternative preservation methods that balances the stability of the device and the access to the synthetic DNA for greater read speeds.

The third challenge involves the lack of full automation and scalability. The libraries for synthetic seem to be a keen interest for some researchers as a potential path to achieve full automation without a significant decrease in informational density. Most library solutions[19] can now achieve full automations but struggle to remain stable and reliable for large scale storage[20]. Current systems are too reliant on human interaction within memory operation greatly effecting the scale these systems can operate at. This challenge best serves to solve the roadblocks to industrialisation.

V. CONCLUSION

Researchers have found that at the current rate of development for conventional memory systems due to density limit and will not be able to sustain the growth in the amount of data produced. This has caused many to look for new technology capable of accepting the additional storage

requirements. Discussed in this literature review a potentially complementary molecular memory alternative of Synthetic DNA memory systems. This memory system's characteristics of long retention time, naturally redundant, and extremely informationally dense. It is this researcher's opinion that the use of chaperone proteins[21] can be used as a biological mechanism for error correction of erroneous sequences. There appears to an underutilisation of bio-inspired mechanisms currently within research and offers a potential route to a more robust system. The evolution of synthetic DNA memory storage can lead to biomolecular computing and "living computer" offering a special self-repairing computational system.

REFERENCES

[1] Sperlea, T., Heider, D., and Hattab, G., 2022. A theoretical basis for bioindication in complex ecosystems. Ecological Indicators, 140, p.109050.

[2] Callaway, E., 2021. Million-year-old mammoth genomes shatter record for oldest ancient DNA. Nature, 590(7847), pp.537–538.

[3] Kannadasan, R., Saleembasha, M.S., and ArnoldEmerson, I., 2015. Survey on molecular cryptographic network DNA (MCND) using Big Data. Procedia Computer Science, 50, pp.3–9.

[4] Anavy, L., Vaknin, I., Atar, O., Amit, R., and Yakhini, Z., 2019. Data storage in DNA with fewer synthesis cycles using composite DNA letters. Nature Biotechnology, 37(10), pp.1229–1236.

[5] Hoshika, S., Leal, N.A., Kim, M.-J., Kim, M.-S., Karalkar, N.B., Kim, H.-J., Bates, A.M., Watkins, N.E., SantaLucia, H.A., Meyer, A.J., DasGupta, S., Piccirilli, J.A., Ellington, A.D., SantaLucia, J., Georgiadis, M.M., and Benner, S.A., 2019. Hachimoji DNA and RNA: A genetic system with eight building blocks. Science, 363(6429), pp.884–887.

[6] Chaput, J.C., Herdewijn, P., and Hollenstein, M., 2020. Orthogonal Genetic Systems. ChemBioChem, 21(10), pp.1408–1411.

[7] Xu, C., Zhao, C., Ma, B., and Liu, H., 2021. Uncertainties in synthetic DNA-based data storage. Nucleic Acids Research, 49(10), pp.5451–5469.

[8] LeProust, E.M., Peck, B.J., Spirin, K., McCuen, H.B., Moore, B., Namsaraev, E., and Caruthers, M.H., 2010. Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. Nucleic Acids Research, 38(8), pp.2522–2540.

[9] Kosuri, S., and Church, G.M., 2014. Large-scale de novo DNA synthesis: Technologies and applications. Nature Methods, 11(5), pp.499–507.

[10] Guruswami, V., and Wootters, M., 2016. Repairing reed-solomon codes. Proceedings of the forty-eighth annual ACM symposium on Theory of Computing.

[11] Liu, Y., Zheng, Z., Gong, H., Liu, M., Guo, S., Li, G., Wang, X., and Kaplan, D.L., 2017. DNA preservation in Silk. Biomaterials Science, 5(7), pp.1279–1292.

[12] Koch, J., Gantenbein, S., Masania, K., Stark, W.J., Erlich, Y., and Grass, R.N., 2019. A DNA-of-things storage architecture to create materials with embedded memory. Nature Biotechnology, 38(1), pp.39–43.

[13] Grass, R.N., Heckel, R., Puddu, M., Paunescu, D., and Stark, W.J., 2015. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angewandte Chemie International Edition, 54(8), pp.2552–2555.

[14] Chalikian, T.V., Liu, L., and Macgregor, Jr., R.B., 2020. Duplex-TETRAPLEX equilibria in guanine- and cytosine-rich DNA. Biophysical Chemistry, 267, p.106473.

[15] Laver, T., Harrison, J., O'Neill, P.A., Moore, K., Farbos, A., Paszkiewicz, K., and Studholme, D.J., 2015. Assessing the performance of the Oxford Nanopore Technologies Minion. Biomolecular Detection and Quantification, 3, pp.1–8.

[16] Chen, W.D., Kohll, A.X., Nguyen, B.H., Koch, J., Heckel, R., Stark, W.J., Ceze, L., Strauss, K., and Grass, R.N., 2019. Combining data longevity with high storage capacity—layer-by-layer DNA encapsulated in magnetic nanoparticles. Advanced Functional Materials, 29(28), p.1901672.

[17] Hughes, R.A., and Ellington, A.D., 2017. Synthetic DNA synthesis and assembly: Putting the synthetic in Synthetic Biology. Cold Spring Harbor Perspectives in Biology, 9(1).

[18] Tan, X., Ge, L., Zhang, T., and Lu, Z., 2021. Preservation of DNA for data storage. Russian Chemical Reviews, 90(2), pp.280–291.

[19] Sabary, O., Orlev, Y., Shafir, R., Anavy, L., Yaakobi, E., and Yakhini, Z., 2020. SOLQC: Synthetic Oligo Library Quality Control Tool. Bioinformatics, 37(5), pp.720–722.

[20] Takahashi, C.N., Nguyen, B.H., Strauss, K., and Ceze, L., 2019. Demonstration of end-to-end automation of DNA data storage. Scientific Reports, 9(1).

[21] Backe, S.J., Sager, R.A., Woodford, M.R., Makedon, A.M., and Mollapour, M., 2020. Post-translational modifications of hsp90 and translating the chaperone code. Journal of Biological Chemistry, 295(32), pp.11099–11117.