Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XVIII, Issue I, Jan-June 2024

DNA: THE ULTIMATE DATA-STORAGE SOLUTION

Dr. Tripati Gupta[1], Ronak Khandelwal[2], Riya Sharma[3], Milan Verma[4]

[1]Associate Professor, Department of Mathematics, Jaipur Engineering College and Research Centre, Jaipur, Rajasthan, India

[2, 3, 4] Students Jaipur Engineering College and Research Centre, Jaipur, Rajasthan, India

ABSTRACT:

The global demand for data storage is currently outpacing the world's storage capabilities. DNA, the carrier of natural genetic information, offers a stable, resource- and energy-efficient and sustainable data storage solution. In this review, we summarize the fundamental theory, research history, and technical challenges of DNA storage. From a quantitative perspective, we evaluate the prospect of DNA, and organic polymers in general, as a novel class of data storage medium [1].

Keywords: DNA storage, information, Carbon coding

[1] INTRODUCTION

In a world flooded with data, figuring out where and how to store it efficiently and inexpensively becomes a larger problem every day. One of the most exotic solutions might turn out to be one of the best: archiving information in DNA molecules.

The prevailing long-term cold-storage method, which dates from the 1950s, writes data to pizza-sized reels of magnetic tape. By comparison, DNA storage is potentially less expensive, more energy-efficient and longer lasting. Studies show that DNA properly encapsulated with a salt remains stable for decades at room temperature and should last much longer in the controlled environs of a data centre. DNA doesn't require maintenance, and files stored in DNA are easily copied for negligible cost.

Even better, DNA can archive a staggering amount of information in an almost inconceivably small volume. Consider this: humanity will generate an estimated 33 zettabytes of data by 2025—that's 3.3 followed by 22 zeroes. DNA storage can squeeze all that information into a ping-pong ball, with room to spare. The 74 million bytes of information in the Library of Congress could be crammed into a DNA archive the size of a poppy seed—6,000 times over. Split the seed in half, and you could store all of Facebook's data.

Science fiction? Hardly. DNA storage technology exists today, but to make it viable, researchers have to clear a few daunting technological hurdles around integrating different technologies. As part of a major collaboration to do that work, A team at Los Alamos National Laboratory has developed a key enabling technology for molecular storage. The software, the Adaptive DNA Storage Codex (ADS Codex), translates data files from the binary language of zeroes and ones that computers understand into the four-letter code biology understands.

ADS Codex is a key part of the Intelligence Advanced Research Projects Activity (IARPA) Molecular Information Storage (MIST) program. MIST seeks to bring cheaper, bigger, longer-lasting storage to big-data operations in government and the private sector, with a short-term goal of writing one terabyte—a trillion bytes—and reading 10 terabytes within 24 hours at a cost of $1,000 [3].

FROM COMPUTER CODE TO GENETIC CODE

When most people think of DNA, they think of life, not computers. But DNA is itself a four-letter code for passing along information about an organism. DNA molecules are made from four types of bases, or nucleotides, each identified by a letter: adenine (A), thymine (T), guanine (G) and cytosine (C). They are the basis of all DNA code, providing the instruction manual for building every living thing on earth [3].

2. OVERVIEW

Research History

In 1953, Watson and Crick published one of the most fundamental articles in the history of biology in Nature, revealing the structure of DNA molecules as the carrier of genetic information . Since then, it has been recognized that the genetic information of an organism is stored in the linear sequence of the four bases in DNA. In just a decade, many researchers had proposed the concept of storing specific information in DNA . However, the concept failed to materialize because the techniques for synthesizing and sequencing DNA were still in their infancy.

In 1988, the artist Joe Davis made the first attempt to construct real DNA storage. He converted the pixel information of the image 'Microvenus' into a 0–1 sequence arranged in a 5 × 7 matrix, where 1 indicated a dark pixel and 0 indicated a bright one. This information was then encoded into a 28-base-pair (bp) long DNA molecule and inserted into Escherichia coli. After retrieval by DNA sequencing, the original image was successfully restored. In 1999,

Clelland proposed using a method based on 'DNA microdots' like steganography to store information in DNA molecules . Two years later, Bancroft proposed using DNA bases to directly encode English letters, in a way similar to encoding amino acid sequences in DNA.

These early attempts only stored less than tens of Bytes—a small amount of data with little scalability for practical usages. It was not until the first 10 years of the twenty-first century that the ground breaking work of Church and Goldman led to the return of DNA storage to mainstream interest . Church et al. successfully stored up to 659 KB of data in DNA molecules, while the maximal amount of stored data before this work was less than 1 KB . Goldman et al. stored even more data, reaching 739 KB. It is worth noting that the data stored in the two studies contained not only texts, but also images, sounds, PDFs, etc., which confirmed that DNA can store a wide variety of data types.

Church and Goldman's work led to a research fever of large-scale DNA storage. With increasingly complex compilation methods, the amounts of stored data gradually increased. By the end of 2018, the maximal amount of data stored in DNA exceeded 200 MB, which was stored in more than 13 million oligonucleotides . Along with the development of DNA synthesis and sequencing technologies, new DNA storage methods keep emerging, bringing DNA storage ever closer to practical applications [1].

This study presents a theoretical framework for utilizing DNA molecules to encode information, intending to circumvent expensive DNA synthesis and facilitate a more practicable approach to certain applications. Our proposed methodology suggests to leverage established biochemical techniques commonly employed in medical and biological research, such as CRISPR-Cas9 and gRNA reagents for labelling. Through comprehensive exploration, we establish upper bounds on achievement codes under specific conditions and introduce an efficient encoder-decoder pair optimized for maximal code size [2].

3. COST OF DNA STORAGE

Compared to traditional data storage methods, DNA storage has significantly lower storage maintenance costs. For example, if a data centre stores 10^9 G data on tape, it will require as much as $1 billion and hundreds of millions of kilowatts of electricity to build and maintain for 10 years. DNA storage can reduce all these expenses by 3 orders of magnitude. Nevertheless, the cost of DNA synthesis can be significant and it will become a limiting factor for DNA storage to commercialize. At the current cost of ∼$10−4/base and a coding density of 1 bit/base, a conservative estimate of the write cost is $800 million/TB, while tape costs about $16/TB . On the other hand, the read cost achieved by current sequencing technologies is orders of magnitude smaller, at ∼$0.01–1 million/TB . However, it is expected that the cost of DNA synthesis and sequencing will continue to decrease in the future, and new techniques and methods will be applied to DNA storage [1].

[4] THE AGE LIMIT OF DNA STORAGE

DNA molecules naturally decay with a characteristic half-life , leading to a gradual loss of stored information. The half-life of DNA highly correlates with temperature and the fragment length. For example, Allentoft concluded that a DNA molecule of 500 bp has a half-life of 30 years at 25◦C, which extends to 500 years for a fragment of 30 bp . Interestingly, fossils provide empirical evidence of DNA's stability over thousands of years . In this case, stability is significantly improved by low temperatures and waterproof environments . Indeed, at −5◦C, the half-life of the 30-bp mitochondrial DNA fragment in bone is predicted to be 158 000 years. Some studies have suggested that DNA can be placed in the extremely cold regions of Earth or even on Mars for millennium-long storage. Other studies have explored packaging materials for DNA molecules and have demonstrated impressive stability. Grass et al. encapsulated solid-state DNA molecules in silica and showed that they had better retention characteristics than pure solid-state DNA and DNA in liquid environments . Judging by first-order degradation kinetics, they concluded that it could survive for 2000 years at 9.4◦C or 2 million years at −18◦C, surpassing all potential quantitative data storage materials invented to date. It is reasonable to expect a long lifetime for data stored in DNA even at room temperature, which makes DNA storage especially suited for cold data with infrequent access. Further research may extend the lifetime of DNA storage over the duration of Human civilization with minimal maintenance [1].

[5] IMPLEMENTATION

DNA data storage could revolutionize data archiving, enabling , vast amounts of information to be stored in a tiny space. This could be crucial for long term preservation of historical records , scientific data and cultural heritage . Additionally , it could reduce the environmental impacts of traditional data centres , which consume significant energy and resources. For the future , DNA data storage could play a vital role in preserving our digital legacy for generations to come .It could also be used in fields like healthcare for storing vast amounts of genomic data or in space exploration for archiving information in extreme environments. It could be used In the field of data archiving for large organizations like libraries , museums, and research organizations .In the corporate world , companies deal with massive amounts of data ranging from customer information to financial records [6].

Journal of Analysis and Computation (JAC)
(An International Peer Reviewed Journal), www.ijaconline.com, ISSN 0973-2861
Volume XVIII, Issue I, Jan-June 2024

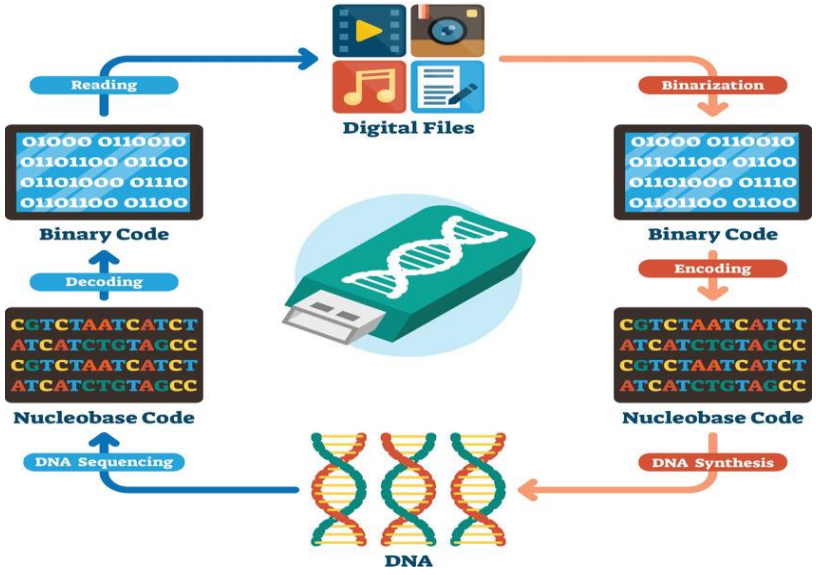

Figure: DNA based digital data storage [5]

[6] THE FUTURE OF DNA STORAGE

Prospects and challenges

Although DNA information storage has enormous application potential, many problems need to be addressed before its broader implementation. First, the cost of writing and reading information is still prohibitively high and the efficiency of storing data is too low. However, DNA synthesis and sequencing costs have been reduced by 10-million-fold over the past 30 years, and the trend will continue to meet the needs of practical DNA storage in the foreseeable future . It is predicted by the Molecular Information Storage Program that DNA synthesis cost will reduce to $10−10/bp by 2023 . At the same time, the read and write speeds have gradually increased. In their original study (2012), Church et al. concluded that DNA synthesis and sequencing technologies require improvements of 7–8 and 6 orders of magnitude, respectively, to compete with current information read and write speeds . The data presented by Goldman et al. show that the main contributor to the cost of DNA storage is synthesis and, based on their calculations, if the cost of synthesis is reduced by another 2 orders of magnitude (compared to 2013), DNA storage will outperform magnetic medium storage for decade-long data storage—a goal that could be achieved in just a few years . In 2017, Erlich et al. gave a cost of $3500 per MB—about a quarter of the cost estimated by Goldman et al, but they expected to use a more cost-effective approach for DNA synthesis as they developed a powerful error-correcting algorithm that tolerates base errors and losses. Very recently, Lee et al. showed a proof-of-

principle enzymatic DNA synthesis scheme, which did not achieve single-base precision, but was still sufficient for complete information retrieval and showed a strong cost advantage over traditional phosphoramidite synthesis . In addition, this synthesis scheme also supports a larger storage volume (~500 to several thousand bases per synthesis) at a higher speed. However, in their implementation, the amount of data stored was extremely limited (144 bits) and whether this approach can be scaled up remains to be tested. Advanced coding and decoding algorithms may ultimately lift the technical requirements on synthesis and sequencing and enable production-grade DNA storage. In addition, storage-specific read and write methods may be developed outside the current synthesis and sequencing frameworks. Writing by the massive assemblage of pre-made oligonucleotides in a way similar to movable-type printing, for example, has recently been claimed to reach a 1 TB/day storage speed.

Random access is another function necessary for information storage purposes. PCR is typically performed using specific primers to obtain selective information stored in DNA. For long-chain DNA storage, PCR with appropriate primers upstream and downstream of the desired information will suffice. However, for oligo DNA storage systems, the entire library needs to be sequenced and assembled before fragmental information can be acquired. Based on powerful error correction codes and algorithmic design, Organick et al. developed a framework to minimize the amount of sequencing required to obtain specific data in an oligo library . They managed to retrieve 35 files (with a total size >200 MB) independently without errors. According to their estimates, the method could be extended to an oligo library with a few TBs of storage capacity. It is worth mentioning that the work of Organick et al. is also an attempt to store the largest amount of data in DNA molecules so far (at the time of writing in 2019).

Finally, techniques to erase and rewrite information in DNA remain to be developed. Existing DNA storage methods support one-time storage only and thus are suitable for information that does not need to be modified, such as government documents and historical archives. However, the continuous development of synthetic biology has shown the possibility of solving this problem. Artificial gene circuits with stable DNA encoding functions have been designed . For example, using a ‘Set’ system of recombinant enzymes and a ‘Reset’ system of integrase and its excision partner, a controllable and rewritable switch could be implemented [1] .

Carbon-based storage

Thanks to the rapid development of DNA manipulation technologies, DNA has become a promising new storage medium. However, other types of polymers may also be used in the field of information storage. Most of them are organic polymers, which, together with DNA molecules, constitute a novel carbon-based storage system different from traditional silicon-based storage.

Like DNA, proteins are an indispensable class of molecules in living systems. Their heterogeneous composition shows potential usage for information storage. However, such attempts are currently focused on the state of the protein rather than its amino acid sequence.

For instance, a protein adopting two different states may encode 0 and 1, and information may be stored by switching and stabilizing the states by specific means. A typical example is a photo-switchable fluorescent protein, which changes colour when absorbing photons of a particular wavelength . Despite its high controllability, the information density is limited to 1 bit per molecule.

In theory, any heterogeneous polymer may serve the purpose of information storage as long as its component monomers can be handled with precision. Current attempts include DNA template guided incorporation of nucleic acid derivatives or small peptides into self-replicating biopolymers . In recent years, the discovery of six non-natural nucleic acids that are able to form stable DNA duplex structures and even carry on genetic information suggests their use for DNA storage . In addition to biopolymers, the synthesis of high-molecular-weight polymers such as polyamides and sequencing at the present time. For example, sequencing of synthetic polymers relies on more general analytical methods such as MS/MS and NMR. Interestingly, single-molecule nanopore sequencing is expected to be a powerful tool for reading information in synthetic polymers.

With more types of monomers able to be integrated, synthetic polymers may exhibit higher self-information and thus storage capacity. In addition, it may be more amenable to certain storage functions such as data erasure and rewriting. On a different scale, composite encoding has been applied to information storage. By using mixtures of nucleic acids or metabolites, one can potentially augment coding capacity in the continuous compositional space of components.

Taken together, synthetic polymers hold great promise for molecular information storage in non-living systems. With the development of sequence control and acquisition technologies, biological and synthetic polymers may form a new framework of carbon-based storage in the future and gradually replace traditional silicon-based storage systems in specialized or general applications [1].

[7] CONCLUSION

Modern societies generate huge amounts of data and the rate of their growth has multiplied in recent years. The need to store both currently generated data and those generated in the past using classical data storage methods are consuming huge financial outlay and physical space. It also entails high costs for the environment, with the introduction of new methods of data storage thus urgently required.

For a long time, people have paid attention to the high storage density and longevity of DNA. In this article, we have provided a brief overview of how information is encoded and stored in DNA. The continuous development of these methods leads to a reduction in the number of errors appearing in the encoding and decoding processes, extending the durability of DNA as a data carrier, and reducing the cost of its storage.

Despite the continued growth in the field of information storage on DNA, some challenges still remain. There is a need to refine the methods used for the fast and error-free synthesis of oligonucleotides, and in the long run, also of long DNA chains. The method used to read nucleotide sequences also must evolve towards greater credibility.

Despite the current obstacles, the prospects for implementing data storage on DNA are very promising. There are even new ideas related to the use of chemical analogues of DNA, such as TNA, with even higher possible storage densities [1].

REFERENCES

[1] Yiming Dong, Fajia Sun, Zhi Ping, Qi Ouyang, Long Qian ,National Science Review, Volume 7, Issue 6, June 2020

[2] Daniella Bar-Lev, Tuvi Etzion, Eitan Yaakobi, and Zohar Yakhini, Representing Information on DNA using Patterns Induced by Enzymatic Labeling

[3] Latchesar Ionkov and Bradley Settlemyer ,DNA: The Ultimate Data-Storage Solution , Issue 28, May 2021

[4] Tomasz Buko ,Nella Tuczko and Takao Ishikawa , DNA Data Storage , Issue 1, June ,2023

[5] Raphael Kim, Larissa Pschet, Conor Linehan, Chang Hee Lee , Archives in DNA: Workshop Exploring Implications of an Emerging Bio-Digital Technology through Design Fiction, Issue June, 2021

[6] George M. Church, Yuan Gao, Sriram Kosuri , Next-Generation Digital Information Storage in DNA , issue August , 2012