Standard Scientific Research and Essays Vol 11(6): 178- 184, June, 2023 (ISSN: 2310-7502)
http://www.standresjournal.org/journals/SSRE

Review Research

DNA Data Storage: An Overview

***Nwakamma Mary Nkechinyere, Nwogu Uchenna Obioma and Oji-Dike Anthony Ikenna**

Department of Electrical and Electronics, Federal Polytechnic, Nekede, Imo State, Nigeria

*Corresponding Author E-mail**:** **marynwakamma@yahoo.com**, **marynwakamma@fpno.edu.ng**

Accepted 20, June, 2023

Abstract

In Today's Post 21st Century Global society where science and technology have evolved drastically, especially in the ICT sector, Deoxyribonucleic Acid (DNA) chips are deployed in data communication and networking applications that require smart and reliable storage to accommodate large data with less or no vulnerability to hacks, cyber compromise, or self-destruct. This Biotech device, the Deoxyribonucleic acid (DNA) module, is an appealing option for such a purpose due to its endurance, a higher degree of compaction, and similarity to the sequential code of 0’s and 1’s as found in a computer. This emerging field of DNA as means of data storage has the potential to transform science fiction into reality, where a device that can fit in our palms can accommodate the information of the entire world, as the latest research has revealed that just four grams of DNA could store the annual global digital information. DNA has all the properties to supersede the conventional hard disk, as it is capable of retaining ten times more data, has a thousand fold storage density, and consumes 108 times less power to store a similar amount of data. Although DNA has an enormous potential as data storage device of the future, multiple bottlenecks such as exorbitant costs, excruciatingly slow writing and reading mechanisms, and vulnerability to mutations need to be resolved.

Keywords: DNA, cyber compromise, data storage, sequential code, mutation

INTRODUCTION

Deoxyribonucleic acid, more commonly known as DNA, is a complex molecule that contains all of the information necessary to build and maintain an organism. All living things have DNA within their cells. In fact, nearly every cell in a multicellular organism possesses the full set of DNA required for that organism.

However, DNA does more than specify the structure and function of living things – it also serves as the primary unit of heredity in organism of all types. In other words, whenever organisms reproduce, a portion of their DNA is passed along to their offspring. This transmission of all or part of an organism’s DNA helps ensure a certain level of continuity from one generation to the next, while still allowing for slight changes that contribute to the diversity of life.

Digital Data Storage, DDS is the industry standard for storing computer DAT (Digital Audio Tapes). DSS tape drives utilize the same technology as a VCR to record information and are generally used to backup data on network servers.

DNA digital data storage is the process of encoding and decoding binary data to and from synthesized strands of DNA. Process of using DNA molecules as a storage medium, unlike the optical at magnetic forms of storage technologies present today.

DNA as a storage medium has enormous potential because of its high storage medium has enormous potential because of its high storage density, its practical use is currently severely limited because of its high cost and very slow read and wrote times.

In June 2019, scientists reported that all 16GB of text from Wikipedia's English-Language version have been encoded into synthetic DNA. In 2021, scientists reported that a custom DNA data writer had been developed which was capable of writing data into DNA at 18Mbps.

STATEMENT OF PROBLEM

Cost of Operation

Ethical Implication so far, most of the discussions in academic research on the challenges faced by DNA based data storage system have been mostly technical in nature. Reducing system latency by making it scalable, Reducing cost of operation and striving for full automation are some of the technical discussion that have been published.

Objective of the study

This research work, will be talking about DNA Digital Data Storage this will go down to the system over view of DNA, mainstream encoding and decoding methods of DNA, Data storage and the challenges of DNA during storage.

SCOPE OF STUDY

This work is divided into 4 stages

1. Will be general introduction which will deal with the structural construction of the work, the statement of problem objective of the study, the scope of the study.
2. Will be looking at the literature review, taking review of writers who had works or something relating to the topic.
3. Will be the main body of the work, where we will be talking about DNA Digital Data Storage, system overview, encodings and decoding methods of DNA data storage and the challenges of DNA during data storage.

LITERATURE REVIEWS

There have been numerous research on DNA, which this study seeks to explore. These include works that generally bother on DNA. DNA-based data storage systems with their distinct attributes opens up new space for HCL research in the context of how they could shape the way we perceive, interact and use data in the future. This article highlighted just a handful of the many opportunities that the new technology brings to the HCL community. Kim R. (2020) DNA as digital data storage: Opportunities and Challenges for HCL.

According to an article titled; DNA as a digital information storage device; Hope or Hype? the entire human race is driven by information, and in this technology-oriented era, information is power. Humans have a natural propensity for accessing more and more information in a little time and space, as possible. Storage of all the accumulated information for future reference is an inherent part of our intellectual evolution (Darshan et al., 2018).

Encoding and Decoding methods in the field of DNA storage, covering several primary encoding and decoding methods; direct mapping between 0-1 binary digital data and A-T-C-G quaternary DNA storage data in the early stages. Chentang et al. (2022) Mainstream encodings and decoding methods of DNA data storage.

DNA digital data

DNA storage breaks through the limitation of maximizing the storage medium of Silica-based materials (such as hard disk, optical disk and removable magnetic disk); compared with the existing digital data storage technologies, DNA storage technology has the advantages of high data density, long storage time, low energy consumption, convenience for carrying, concealed transportation, and multiple encryptions (De Silva and Ganegoda, 2016).

Furthermore, with the rapid development of biotechnology and information technology (BT and IT), DNA storage is expected to fundamentally change the pattern of data storage and transmission, further leading to revolutionary changes in various important areas of the National economy such as; manufacturing, internet industry, and national security (The DNA data storage alliance 2021) preserving our digital legacy: an introduction to DNA data storage.

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 solution might turn out to be one of the best archiving information in DNA molecules.

The prevailing long-term cold storage, which dates from the 1950s, wrote data to pizza-sized reels of magnetic tape. By comparison, DNA storage is potentially less expensive, more energy efficient and long-lasting.

SYSTEM OVERVIEW

Short form deoxyribonucleic Acid, DNA is a molecule that carries genetic information in living organisms. They are found in nature, but they can also be synthesized artificially. There are four types, known as bases: Adenine(A), Thymine (T), Cytosine (C), and Gunanine (G). Each DNA base can potentially hold up to two bits (binary digits) of electronic data. For example, a sequence of 00, 01, 10, and 11 can be coded as A, T, C, and G, respectively, as DNA.

Several attributes make DNA an ideal candidate for data storage. Firstly, the recovery and reading of DNA that is hundreds of thousands of years old (Ball, 2020), suggest that DNA's high stability, and the potential to offer high retention capability as a data storage medium. Moreover, DNA offer low energy for operation (Ceze et al., 2019) said that, making it an economically attractive alternative to existing forms of data storage.

DNA's strongest attribute, however, is its impressive storage capacity. Capable of retaining up to 1018 bytes of information per cubic millimeter, DNA provides density that is roughly six orders of magnitude higher than the currently available, densest storage Media (Rutten et al., 2018). To give a better sense of perspective, such density would allow all information produced around the world over a one-year period, to be stored in just 4g of DNA (Rutten et al., 2018).

Overall, these attributeswould help in making DNA-based storage systems to address the issue of exponentially growing digital data production in outpacing the growth of traditional digital data storage systems (e.g., tape or hard disk drives, Blu-ray, and flash).

THE PROCESS

DNA data will not be stored in binary digits (1∞0). Instead, they would be encoded into DNA nucleotide bases (A, C, G, T) and stored. These strands are then converted to binary digits when needed.

In terms of explaining how DNA fits into the overall process of the data storage system. It provides a generic framework, illustrating the basic steps that are involved. Digital data is firstly encoded into DNA sequences, which are subsequently artificially synthesized into real DNA strands. These are then stored/archived away in either of the two formats. The first format, which is more commonly researched, is in vitro (either as a frozen liquid or dried format). The second format involves the insertion of the DNA into a host cell (in vivo), such as E.coli bacteria, effectively creating a type of microbial "memory unit" (kim et al., 2019).

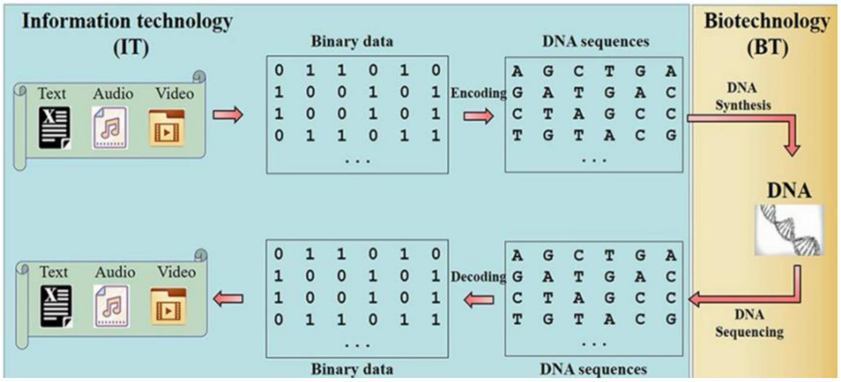
In order to request and retrieve back the data, an appropriate method for random access is carried out on the stored DNA batch (e.g through polymerase chain reaction (PCR) O, followed by the use of a sequencing machine to read the DNA, and then the final step of decoding the sequences back into an electronically compatible form.

MAINSTREAM ENCODING AND DECODING METHODS

Church Encoding and Decoding Method

In August 2012, the George Church group at Harvard University published a landmark paper in the research field of DNA storage and achieved DNA storage for 5.2MB data of HTML files, JPG images, and JavaScript programs (Church et al., 2012). They first proposed the “bit-base” mapping rule/codebook: one bit per base. Especially, each bit corresponds to a nucleotide, in which A or C represents), and T or G represents 1. In the codebook, they generated DNA sequences with more than 3 nthomopolymers were excluded. Here, the sequence structure of encoding DNA’s (159 nt) includes a 96 nt DNA fragment for data payload, a 19 nt fragment for address information, and two 22 nt fragments as PCR amplification primers on both ends. For encoding, we constructed a paired-end library for each 159 nt DNA fragment and then sequenced it on an Illumina HiSeq 2000 platform in 100-bp paired-end mode. Further, the sequenced pair-end reads were combined into a single contig by overlapping to reduce the systematic sequencing errors. However, due to systematic errors and without using error-correcting code during DNA storage, the data could not be recovered completely in spite of higher sequencing depth (average coverage: 3000x).

In 2016, Blawat improvised Church’s method using triple error-detecting/correcting codes to achieve the DNA storage and completely accurate recovery for 22MB data of a MPEG-compressed movie (Blawat et al 2016); the address information was horizontally protected by Bose-Chaudhuri- Hoequenghem codes (BCH) (63, 39) with the minimum Hamming distance of 9 (occupying 39 bits); the consecutive data blocks (223 bits for data payload) were vertically protected by Reed-Solomon codes (RS) (225, 223, 33); 16-bit Cyclic Redundancy Check (CRC) was used for error detection of address information and data payload (Figure 1)



0	1	1	0	1	0
1	0	0	1	0	1
1	0	0	1	0	1
0	1	1	0	1	1
...					

A	G	C	T	G	A
G	A	T	G	A	C
C	T	A	G	C	C
T	G	T	A	C	G
...					

0	1	1	0	1	0
1	0	0	1	0	1
1	0	0	1	0	1
0	1	1	0	1	1
...					

A	G	C	T	G	A
G	A	T	G	A	C
C	T	A	G	C	C
T	G	T	A	C	G
...					

Figure 1. Mainstream Encoding Process of DNA

Challenges of DNA data storage

Considering the long data retention time and high storage density of DNA, it is apparent to predict that DNA has tremendous potential to become the strongest competitor in the semiconductor industry. However, it will not be easy to achieve that potential. Researchers will need to carefully address and overcome several challenges to develop user-friendly DNA-based data storage devices. The whole process of encoding information in DNA and subsequent retrieval of the required information is far more likely to face tough completion from optical, magnetic, or quantum techniques in the foreseeable future. The mechanization of various molecular processes associated with DNA has failed to perfectly mimic the natural processes.

The presence of homopolymers, various sequencing errors, and errors due to lower access rates are some examples of this. A living cell has a precisely designed proofreading and DNA repair mechanism for the correction of various errors in the DNA. Although the error-free synthesis, amplification, and sequencing of DNA cannot be achieved yet, a breakthrough was made by Blawat (Blawat et al., 2016), who recently reported storage and successful error-free retrieval of 22MB of digital data in synthetic DNA, using a forward error correction scheme. Due to its sensitive structure, DNA is prone to mutations under extreme conditions; hence, the chances of data lateration cannot be ignored. At the same time, it is quite difficult to synthesize long sequences of DNA de novo. Moreover, while conventional and popular storage systems such as hard drives are more expeditious in erasing and rewriting data, this aspect has barely been dealt with for DNA data storage systems.

DNA-based data storage systems have been mostly technical in nature. Reducing system latency, making it scalable, reducing the cost of operation, and striving for full automation, are some of the technical discussions that have been published so far (Ceze et al., 2019).

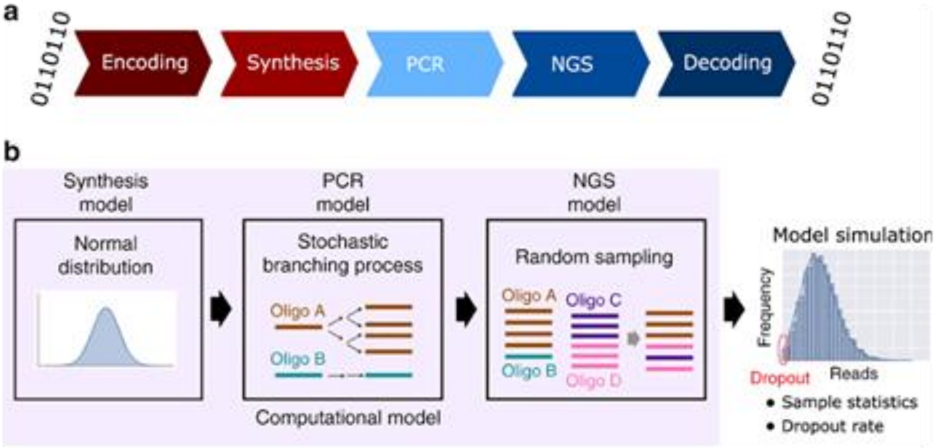


Figure 2. Flow Model of DNA Data Storage

On the other hand, discussions on the social implications of the technology have been lacking overall: What are the potential ethical issues surrounding our prospective interactions with the new technology, and what are the possible ways to address them?

Technology ownership, in the early years of market adoption, DNA-based data storage systems will most likely be available to the public as an online data storage service, rather than hardware that users can own and operate the physical steps involved. The machine(s) and reagents that enable different steps, such as data encoding and decoding, would simply be too suspensive, with some of the protocols involved also possibly patent-protected.

As such, the technology's likely initial owners, which would include Co-operations (Microsoft June 2020), government bodies (Defense Advanced Research Projects Agency (DARPA) June 2020]), and academia in partnership with either of these organizations, would have most of the control in gathering, storing, and using data.

Historically, handing personal data to such organizations has led to issues associated with data privacy breaches and data misuse in the past (Culnan 2003 and Lyon 2014). And crucially, there is currently no convincing guarantee that such issues cannot rise again with the usage of the new technology through third parties.

A possible long-term solution to these potential issues could involve designing better or alternative systems that are affordable and accessible for public use, thereby handing the agency of data management to the user on how and why certain data are to be achieved and retrieved.

A possible shorter-term solution, on the other hand, would involve gaining better awareness and insight into;

1) How DNA-based storage systems work,
2) Their potential pitfalls.
3) The overall business model of the storage services offered
4) Their associated customer rights

One of the main ingredients that is needed to create meaningful and productive dialogue in a research community is effective and clear communication of ideas and arguments. This may be achieved through verbal or written channels, or through artefact designs that project or embody the ideas and arguments.

As the concepts of alternative, biological data storage and retrieval systems, effective ways to communicate the technology between those who may be designing and/or studying such systems would be one of the challenges that should need to be overcome.

As an example, on a basic level, the process of encoding digital data into DNA, and decoding it back to binary bits, allow languages of molecular biology and computer programming to intertwine, which may pose technical difficulties for those that may be unfamiliar with either of the two disciplines.

CONCLUSION AND RECOMMENDATIONS

Conclusion

DNA molecules, with its special qualities, make for compelling media, for alternative data storage systems. Key technological developments over the recent years have allowed the prospect of DNA-based digital data storage systems (and services) to edge closer to the main stream usage. Despite these progressions, there has been little discussions within the research on the implications of this technology. This paper highlighted just a few prospective challenges presented by DNA-based data storage systems.

Overall, DNA-based data storage systems remain a largely open research topic for and the questions that have been raised here are by no means exhaustive. They are intended to serve as an opening for further in-depth discourse and exploration.

Recommendation

At present, the existing data storage technologies are increasingly unable to meet their requirements of the explosive growth of data. DNA storage is expected to develop into a major data storage form in the future. Among them, the development of efficient and accurate DNA storage encoding and decoding methods and related error-correction algorithms will play a very important and even decisive role in the transition from the laboratory to practical application, which may fundamentally change in the information industry in the future.

References

Ball J(2020). Ancient Horse Bone yield oldest DNA sequence. BBC News. https://www.bbc.co.uk/news/science-environment-23060993.

Blawat M, Gaedke K, Hutter I, Chen XM, Turczyk B, Inverso S, Pruitt BW, Church GM.2016). Forward error correction for DNA data storage.

Ceze L, Nivala J, Strauss K(2019). Molecular digital data storage using DNA.

Church GM, Gao Y, Kosuri S(2012) Next-generation digital data storage using DNA.

Culnan MJ, Bies R.J(2003). Consumer Privacy: balancing economic and justice considerations.

DARPA. (2020). Turning to Chemistry for New “Computing” Concepts.

DeSilva PY, Ganegoda GU(2016). New trends of digital data storage in DNA. Biomed. Res. Int.

Kim R, Poslad S(2019). The thing with E.coli: Highlighting opportunities and challenges of integrating bacteria in IoT and HCI.

Lyon D.(2014). Surveillance, snowden, and big data: capacities, consequences, Critique. Big Data Soc.
Microsoft.(15 June 2020) DNA Storage
Rutten MGTA, Vaandrager FW, Elemans JAAW, Nolte RJM(2018). Encoding Information into Polymers.
The DNA Data Storage Alliance (2017): Preserving our digital legacy: an introduction to DNA data storage.