IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13660

DNA Data Storage

Surabhi M V[1], Jeevan K P[2], Koushik R [3]

Student, CSE, VTU CPGS, Mysore, India [1, 2, 3]

Abstract: Human Beings have always tried to simplify the way of storing data maintaining both security and speed of access. This decade (2011-2020) is focusing on improving data storage devices. New technologies like SSDs (Solid State Drives), technical upgrades in SATA or IDE HDDs (Hard Disk Drives), etc with Terra Bytes of storage capabilities have come to light in recent past. However, DNA Data Storage technology is the next generation of storage technique, which has a lots of storage capability. DNA Data Storage will reinvent the way of storing data. This paper discusses about this storage mechanism and emphasizes on the on-going re-search in this field.

Keywords: Data Storage, DNA, SDDs, HDDs, Genes.

I. INTRODUCTION

Today's 'Digital universe' forecasts that total data in 2017 is 16 zetta bytes (1021). Companies are building Data Centres in thousands of acres of land areas. Face book recently built a data center dedicated to one Exabyte of data storage. Similarly, many organizations and universities started working on data storage solutions for better storage duration, less cost, and less space occupying. DNA Storage Technology was first demonstrated in the 1980s and it became the largest project by 2010. The DNA Data Storage technology uses DNA (Deoxyribo Nucleic acid) for storage. This system of storage is more efficient compared to the present day's magnetic tapes and hard drive systems. One of the reasons why DNA is considered as a better storage system is that 215 pet bytes (215 million gigabytes) can be stored in just 1 gram of DNA. Here, the DNA can be any living being's DNA. In fact, a woolly mammoth's DNA was first studied for this purpose. DNA is 'apocalypse-proof' because even after global disasters, one thing that we can always preserve and store is DNA. However, storing and retrieving data from it is a slower process as of today, because it is still under development and might take a few years more to become commercially viable.

II. BEGINNING OF THE IDEA

Mikhail Samoilovich Neiman, a Russian physicist proposed the idea of the possibility of storing and retrieving information from DNA molecules. This technology was known as MNeimON (Mikhail Neiman Oligonucleotides).

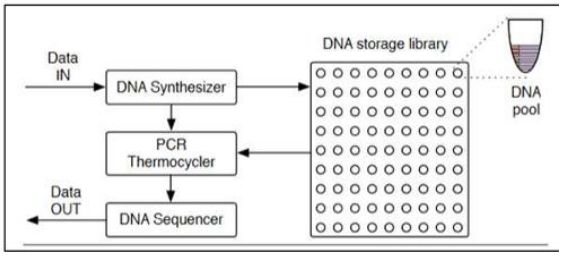


Fig. 1: The Basic Overview

III. HOW DOES IT WORK

At first place, DNA storage is ultra-compact as it can be stored safely for hundreds and thousands of years in a cool, dry place. It will not degrade easily like HDDs, SDDs and other memory devices. Oligonucleotide Synthesis machines are made to upload/store information in DNA and there are highly complex machines, which can retrieve the stored data called as DNA Sequencing Machines. Oligonucleotide synthesis is the chemical synthesis of relatively short fragments of nucleic acids with defined chemical structure. DNA molecules are long strands or sequences, which are made-up of nucleotides called Adenine (A), Cytosine (C), Thymine (T) and Guanine (G). Sequences of these nucleotides are made rather than creating sequences of 0s and 1s. The way it works is assigning digital data patterns (Binary form of data) to DNA nucleotides. The whole process starts with the simple concept of Preparing Bits to become Atoms.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

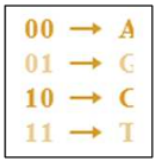
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13660

a. Encoding

BINARY CODES AS 00, 01, 10, 11 ARE REPRESENTED USING THE 4 NUCLEOTIDES A, T, G AND C. FOR EXAMPLE, 00 COULD BE EQUAL TO A, 01 TO C, 10 TO T AND 11 TO G. THEREFORE THE BINARY FOR OF DIGITAL DATA (01 11 10 00 11 11 10 11 01 00 01……) IS REPRESENTED BIOLOGICALLY AS C-G-T-A-G-G-T-G-C-A-C-…….THEREFORE THIS ORDER OF NUCLEOTIDES FORM A DNA STRAND. THIS IS HOW DIGITAL DATA IS ENCODED.



b. SYNTHESIS

The artificial DNA should be shorter because longer DNA is chemically harder to build. Digital data can be of large sizes, but a single DNA strand can only hold around 20 bytes. So data is broken into smaller pieces and an indicator is set to the sequence so that it will ensure all the pieces of data can stay in proper order. Hence, the data is synthesized.

c. Storage

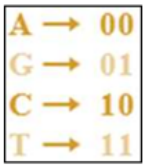
The chemical reactions used in synthesis are driven by a device, which takes the ATGC nucleotides, mixes them in a solution with some other chemicals to control reactions and order of the strands. This process also benefits us by creating backup by creating copies of each stands for another series at once. Now the created DNA is protected from damages that are caused by light and humidity. Therefore, it is dried and stored in cool place also blocking water and light.

d. Retrieval

The indicator installed during the synthesis of DNA is now used to retrieve the multiple strands of the DNA in a determined databased order.

e. Decoding

The letter sequence generated by the sequencing machine is now decoded back into an ordered sequence of 0s and 1s. As for today, DNA can be destroyed during this process, but as mentioned above, many copies of each sequence are made and they now come into play. If these backup copies are, also depleted, more duplicate copies can be made easily as DNA replication is also a natural process. In this system, the whole DNA is required to be analyzed even if we need to read or access only some part of the information in it. Therefore, some special Biochemistry methods are being developed and studied for accessing only required information at a much faster rate.



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13660

Digital Data to DNA process can be figured out from this diagrammatic procedure representation.

– Binary text file- Indexing and Storing-DNA Fragmenting - DNA Encoding - Base Encoding .

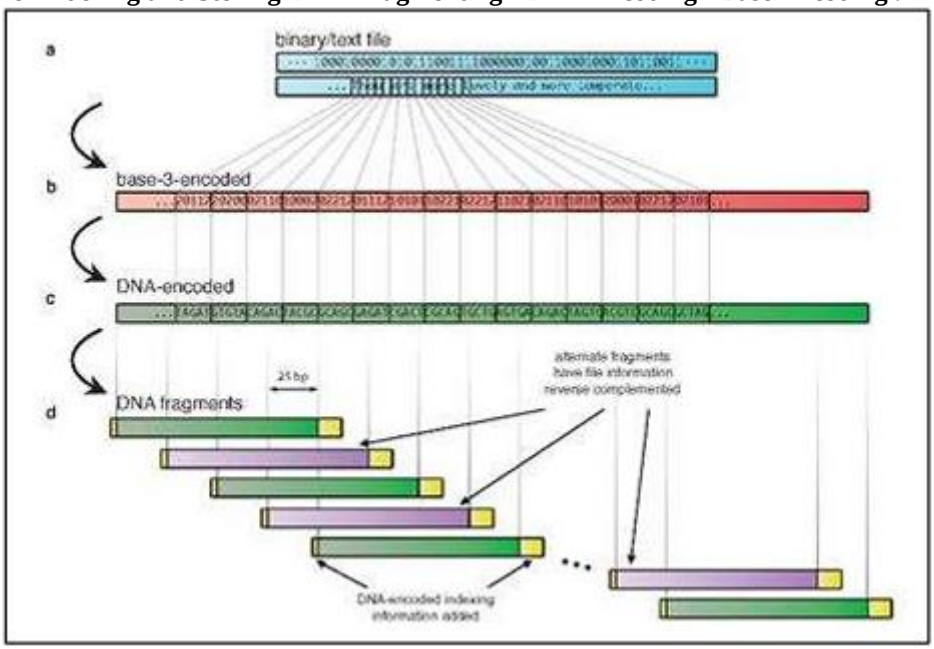


Fig 2 The Basic Overview

IV. CHALLENGES

DNA Data Storage technology is still being developed and experimented even today. The researchers of different universities, companies and organizations are trying to make the whole process completely automated. The process of building DNA and accessing it by reading it at a faster rate is also being improved on every step. But, as per today, the process is still relatively slow compared to Flash Drives. Many significant changes are to be made yet trying to improve and develop the systems rapidly. The main target of this technology now is to make it faster and cheaper.

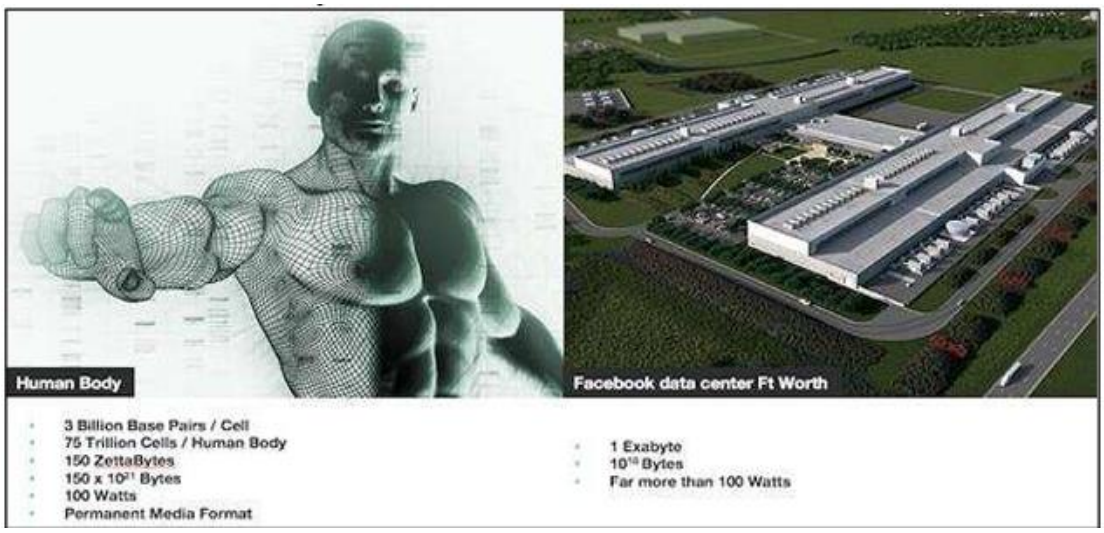


Fig 3 Human DNA Storage vs Data Centres

Twist Bioscience is a private company, which develops and manufactures synthetic DNA. Some of the most important products that the company researches and provides are

1) Oligo Pools
2) Genes and Gene Fragments
3) Therapeutic Antibody Design and Optimization Services

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13660

4) Higher Density DNA Digital Data Storage.

V. PREPARATION OF DNA FOR DATA STORAGE

Twist Bioscience has its own innovative methods of making synthetic DNA and to get it ready for the processes of the storage mechanisms.
Oligo Synthesis (making synthetic DNA) is done and Perfect Gene is prepared which is ready for storing and accessing DNA. Each of this perfect gene is capable of storing 1.8 kb of data. Billions and trillions of such genes together make up a DNA strand, which can store petabytes of data.

VI. MICROSOFT DNA RESEARCH

The Microsoft Corporation has started its research and experiments on DNA Data Storage too. In fact, it had the most successful start in this field of technology. Microsoft was able to store 200 megabytes of data related to literary and other articles into DNA. Microsoft also purchased 10 million strands of DNA (Oligonucleotides) from Twist Bioscience for research and implementation of this technology and to encode digital data. Microsoft is also the company to recover or retrieve 100% of the encoded data successfully. Based on its own study, the company also re-estimated that 1 cubic millimeter DNA can store 1 Exabyte (1 billion gigabytes) of data. Microsoft's digital data continuous to expand exponentially and therefore it is planning to use DNA Data Storage technology to replace its 1000's of acres of land occupied with data centers starting form one of them. It also has a plan to add DNA Data Storage to its cloud services for now. In the beginning, the speed of encoding and retrieval of data from DNA was only about 400 bytes/second. Now they are reaching speeds of 100 megabytes/second, achieving high and successful results. Microsoft had 36 Azure data centers and 8 being ready again, from which 1 data center is going to be a DNA-based data center. University of Washington is also working as a part of Microsoft.

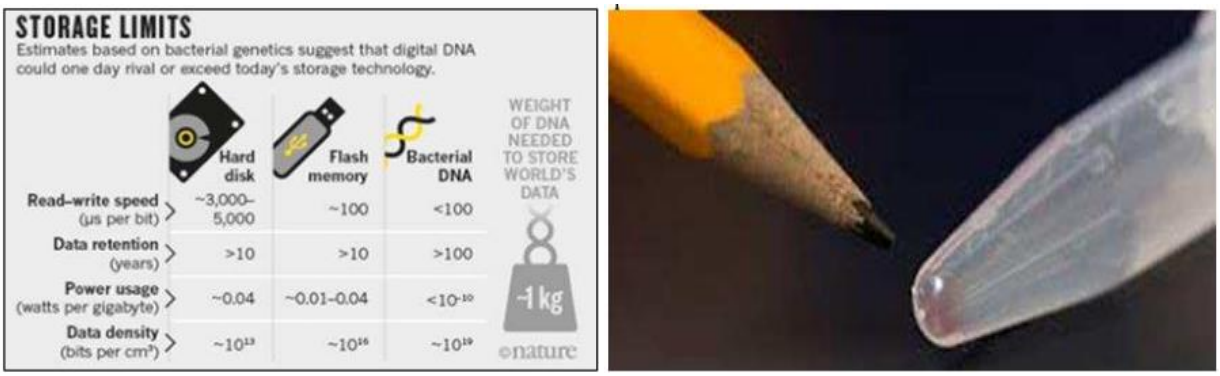


Fig 4 Microsoft's 1 Gram of DNA and DNA Data Storage Limits

VII. GOLDMAN ENCODING

The scientists applied the Goldman's encoding system, the XOR encoding to store 3 image files out of which 2 were successful and the other image file had 1 byte error due to bugged sequencing. These are the 3 image files that were originally used in this experiment. However, Goldman individually was able to store 154 sonnets of Shakespeare in ASCII, a PDF file, a color photograph and an mp3 file in DNA successfully.
1) Sydney.jpg , 24301 bytes
2) Cat.jpg, 11901 bytes
3) Smiley.jpg, 5665 bytes

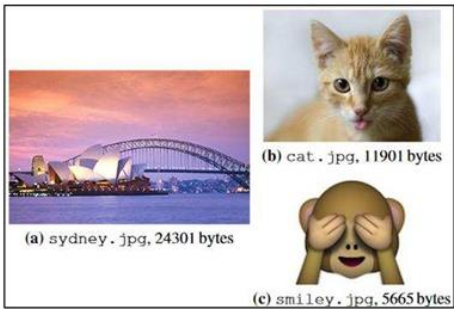


Fig 5 Goldman's DNA stored image files

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 6, June 2024

DOI: 10.17148/IJARCCE.2024.13660

ADVANTAGES	DISADVANTAGES
Highly durable, stable and easily synthesized	The process of copying and retrieving is a slow process
Needs easy maintenance and information is stored for thousands of years	Not yet fully developed and the DNA replicators can be of high costs.
Highly reliable and 2.2 petabytes of data can be stored in 1 gram of DNA.	Difficult to identify where the process went wrong at any point of the mechanism.

Fig 6 DNA Data Storage facts

CONCLUSION

The DNA Data Storage technology is a tech-changing idea which reinvents the way of how we store and retrieve data from electronic devices. DNA is a natural solution to land problems. It can replace the huge data centres of the world which are very expensive as they consume large amounts of electricity and other resources. Storing data in DNA is a natural way of data storage which saves us tones of resources and enables us to store our memories for thousands of years, sharing knowledge to all the generations ahead.

REFERENCES

[1] https://twistbioscience.com/company/blog/twistbiosiencednastoragefountain
[2] https://www.microsoft.com/en-us/research/project/dna-storage/
[3] https://twistbioscience.com/company/blog/could-dna-super-charge-the-digital-revolution
[4] https://www.the-scientist.com/?articles.view/articleNo/32494/title/DNA-Data-Storage/
[5] https://www.nature.com/news/how-dna-could-store-all-the-world-s-data-1.20496
[6] https://www.microsoft.com/en-us/research/wp-content/uploads/2016/04/dnastorage-asplos16.pdf
[7] https://twistbioscience.com/products/storage
[8] http://www.sciencemag.org/news/2017/03/dna-could-store-all-worlds-data-one-room
[9] https://www.technologyreview.com/s/607880/microsoft-has-a-plan-to-add-dna-data-storage-to-its-cloud/