

DNA DATA STORAGE

¹Harshad Gawai , ²Vaibhav Shriramjwar, ³Nikita Mohod

Final Year Student, Computer Science & Engineering, Sipna College of Engineering & Technology, Amravati, Maharashtra, India, ^{1,2}Assistant Professor, Computer Science & Engineering, Sipna College of Engineering & Technology, Amravati, Maharashtra, India

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, SSDs, HDDs, Genes

1. INTRODUCTION

Today's 'Digital universe' forecasts that total data in 2017 is 16 zettabytes (10²¹). Companies are building Data Centers in thousands of acres of land areas. Facebook 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 petabytes (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.



Fig-1: The DNA Data Storage

2. BEGINNING OF THE IDEA

The idea and the general considerations about the possibility of recording, storage and retrieval of information on DNA molecules were originally made by Mikhail Neiman and published in 1964–65 in the journal-USSR, and the technology may therefore be referred to as MNeimONics, while the storage device may be known as MNeimON (Mikhail Neiman OligoNucleotides). In 2007 a device was created at the University of Arizona, using addressing molecules to encode mismatch sites within a DNA strand. On August 16, 2012, the journal Science published research by George Church and colleagues at Harvard University, in which DNA was encoded with digital

information that included an HTML draft of a 53,400 word book written by the lead researcher, eleven JPG images and one JavaScript program.

3. OVERVIEW OF THE DATA STORAGE SYSTEM

We imagine DNA digital data storage as the last level of a deep storing hierarchy, giving very dense and durable storage with access times of many hours to days. DNA synthesis and sequencing can be made arbitrarily parallel, making the necessary read and write bandwidths available. We now detail our proposal of a system for DNA based storage with random access support.

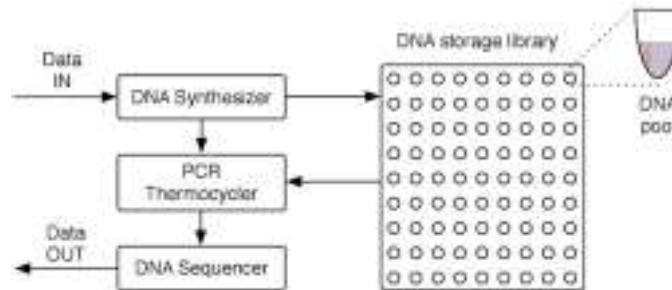


Fig-2: The Basic Overview

4. 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, SSDs 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.

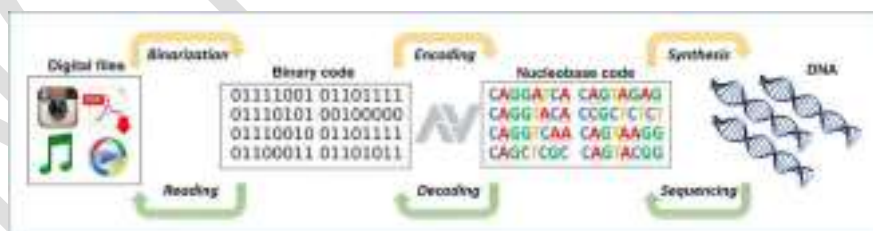


Fig-3: Working of DNA Digital Data

4.1 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.

4.2 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.

4.3 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 strands for another series at once. Now the created DNA is protected from damages that is caused by light and humidity. Therefore, it is dried and stored in cool place also blocking water and light.

4.4 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.

4.5 Sequencing

To read back the data, a machine called Srquencing Machine is used. It looks like the machines that are used for the present day analysis of Genomic DNA in differnet cells. As a result of this process, molecules are identified and a letter sequence is generated. This sequencing is done to obtain the final format of digital data.

4.6 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.

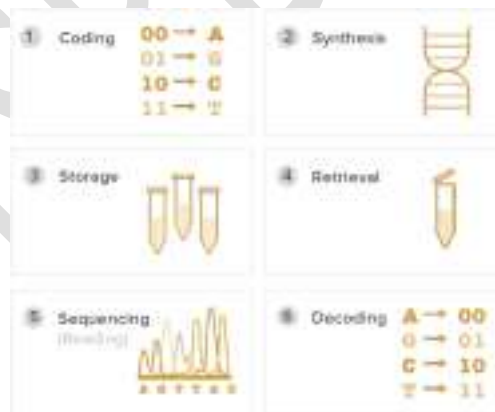


Fig-4: DNA Data Storage Process

- Digital Data to DNA process can be figured out from this diagrammatic procedure representation.
- Binary text file Base Encoding DNA Encoding DNA Fragmenting Indexing and Storing

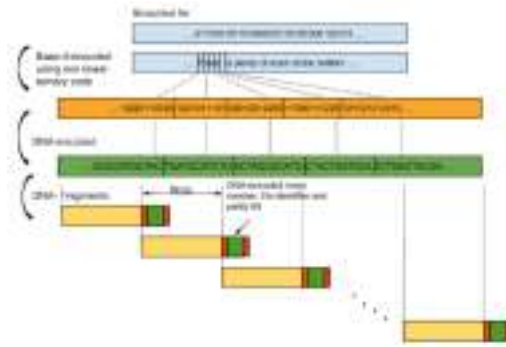


Fig-5: The Basic Overview

5. 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-6: Human DNA Storage vs Data Centers

- 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
 - 4) Higher Density DNA Digital Data Storage

6. PREPARATION OF DNA DATA STORAGE

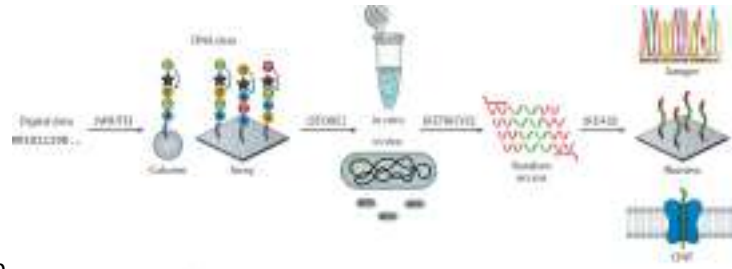
Twist Bioscience has its own innovative methods of making synthetic DNA and to get it ready for the processes of the storage mechanisms.



Fig-7: Silicon-based Synthesis

As shown in the diagram, 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.



Lp

Fig-8: Manufacture of Artificial Genes and DNA

7. 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-9: Microsoft's 1 Gram of DNA

STORAGE LIMITS
Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.

	Hard disk	Flash memory	Bacterial DNA
Read-write speed (us per bit)	~3,000-5,000	~100	<100
Data retention (years)	>10	>10	>100
Power usage (watts per gigabyte)	~0.04	~0.01-0.04	<10 ⁻⁶
Data density (bits per cm ³)	~10 ²⁴	~10 ²⁴	~10 ³¹

WEIGHT OF DNA NEEDED TO STORE WORLD'S DATA
1 kg

Fig-10: DNA Data Storage Limits

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

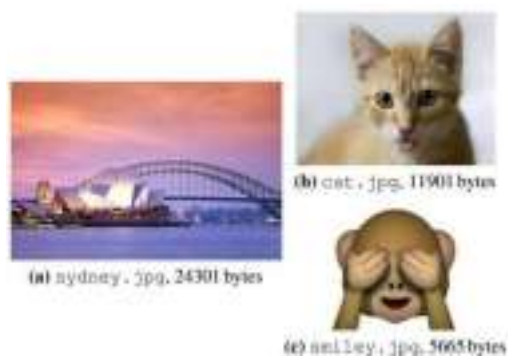


Fig-11: Goldman's DNA stored image files.

8. ADVANTAGES AND DISADVANTAGES

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

CONCLUSION

The DNA Data Storage technology is a tech-changing idea which reinvents the way of how we store and retrieve data from electron-ic devices. DNA is a natural solution to land problems. It can replace the huge data centers 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 enable us to store our memories for thousands of years, sharing knowledge to all the generations ahead.

REFERENCES

1. <https://twistbioscience.com/company/blog/twistbiosciencednastoragefountain2>
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/>
10. J. Gantz, D. Reinsel. “Extracting value from chaos”, International Data Corporation (IDC), Framingham (2011).
11. C. Bancroft, T. Bowler, B. Bloom, C. T. Clelland. “Long Term Storage of Information in DNA Science” (2001).
12. George M. Church, Yuan Gao, Sriram Kosuri. “Next Generation Digital Information Storage in DNA” (2012).
13. M. Burrows and D. J. Wheeler, “A block-sorting lossless data compression algorithm,” Digital System Research Center, USA, 1994.
14. G.C.Smith, C.C.Fiddes, J.P.Hawkins, and J.P.L.Cox, “Some possible codes for encrypting data in DNA”, 2003.
15. D. A. Huffman, “A method for the construction of minimum-redundancy codes,” 1952.
16. L. Adleman. Molecular computation of solutions to combinatorial problems. Science, 266(5187):1021–1024, 1994
17. M. E. Allentoft, M. Collins, D. Harker, J. Haile, C. L. Oskam, M. L. Hale, P. F. Campos, J. A. Samaniego, M. T. P. Gilbert, E. Willerslev, G. Zhang, R. P. Scofield, R. N. Holdaway, and M. Bunce. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils.
18. N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. Le Proust, B. Sipos, and E. Birney, “Towards practical, highcapacity, low-maintenance information storage in synthesized DNA,” 2013.
19. C. T. Clelland, et al., “Hiding messages in DNA microdots”, 1999.