IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9652

DNA Data Storage

Hanumantha Reddy[1], Prof. Prashanth K[2]

Student, Department of MCA, R V College of Engineering, Bengaluru, India[1]

Assistant Professor, Department of MCA, R V of College Engineering, Bengaluru, India[2]

Abstract: Data is stored in DNA. To enable random access to the data, a delimiter is utilised at the end of each file in this system. Use of a specialised Huffman tree will be used to encrypt the data. If necessary, it is possible to encode each file separately using a Huffman tree, which will both strengthen data security and compress the data. Any data fault that occurs while encoding is exclusive to that particular file. Data compression is accomplished via encoding with the Huffman tree. Since no one can decode it without the original tree, it offers security. There is a lot of specialised equipment required for DNA strand sequencing. DNA cannot therefore be read without the necessary tools. DNA contains a maximum of 2 nucleotide repetitions, excluding delimiters. It is possible to keep data for a considerable amount of time since DNA can preserve info for many years. This method allows for the packing of data and the provision of data security. Users can read many files simultaneously thanks to the possibility of parallel file reading. This method keeps two copies of the data. Consequently, its copy can be utilised to read data in the event of data damage. With minimal computational cost, this method can be utilised to store large amounts of data in a very tiny amount of space. This approach is scalable and works well for storing massive files. Additionally, making many copies is simple. Information can be kept in large data or archive systems using this technique.

Keywords: Digital data, data storage, encoding, synthesis, retrieval, decoding, and sequencing

I. INTRODUCTION

Standard hard drives are susceptible to mechanical breakdowns, extreme temperatures, dampness, and magnetic field exposure. Although solid state drives perform better than hard drives, they tend to lose information if left idle for longer than a few months. Therefore, the creation of a storage mechanism that successfully addresses the aforementioned shortcomings has been the focus of researchers' attention. Scientists focused on employing deoxyribonucleic acid (DNA) as a storage medium after considering how fossilised bones maintain genetic data for aeons. Incredible amounts of storage can be found in DNA. According to Castillo, every piece of information on the Internet could be found in a space no larger than one cubic inch.

Adenine, guanine, cytosine, and thymine (A, G, C, and T) base pairs in DNA can be used to store information in binary code rather than the computer's usage of 1s and 0s for data storage. This is why DNA is seen as the best medium in this regard. Given that a single nucleotide can hold two bits of information, DNA is thought to be the ideal high capacity storage media in this regard. As a result, 1 gramme of single stranded DNA (ssDNA) may encode data. Just 4 kilos of DNA can contain all the information created worldwide in a single year. DNA offers a large amount of memory because its structure is three-dimensional (3D). By drying and safeguarding from oxygen and water, DNA provides legible and trustworthy information for millennia, which can be extended to almost infinity.

The technologies available now won't be sufficient to handle these problems. With storage densities orders of magnitude higher than the most effective techniques used now, DNA offers a plentiful, long-lasting, and stable data storage solution. For comparison, 1 kg of DNA has the potential to contain all of the info currently available on the planet. Although the science underpinning DNA data storage has been established, there are currently few commercial applications for it. This is due to the fact that DNA technology, instead of being developed for data storage, was created to enable applications in the life sciences business.

II. LITERATURE SURVEY

Mullin [1] emphasised the advantages of DNA as a data storing technique as well as two significant drawbacks. Although it might be anticipated that this will get easier, the initial data recovery from the genome is a very time-consuming operation. Another consideration is the price, as such technology may be very alluring and hence more expensive.

Keown et al. [2] looked at the unmet capacity of current storage medium as well as the exponentially growing demand for data storage. The prospect of using DNA to archive data is appealing for meeting demand. The architecture for a

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9652

DNA-based archival storage system was proposed in this study. It is a key-value store that makes use of standard biochemical methods to offer random access.

The analogy between digital data and genetic data is described by Choi et al. in [3]. The genetic information is initially stored as molecular polymers, while digital information is initially stored as binary digits (bits, i.e. 0 and 1). These polymers are made up of four bases—A, C, T, and G—each pair of which corresponds to a small amount of data.

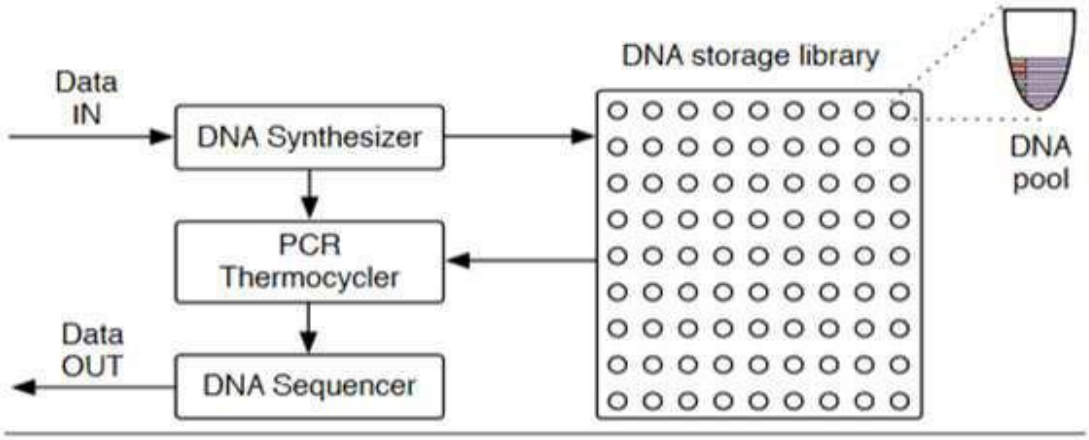
The first DNA-based storage system, which allows for random access to data blocks and rewriting of data stored at arbitrary points within the blocks, was described by Sandle et al. in their paper published in Nature [4]. Their approach is built on innovative DNA editing technologies and restricted coding techniques that guarantee data accuracy, precision, and sensitivity of access while also offering extraordinarily high data store capacity.

According to Golgman et al. [5], maintaining DNA-based storage just requires a cold, dry environment, which is true for any biological or chemical substance. Other characteristics were also covered, such as how effective copying is, which makes it great for backups and transportation. As a result, DNA storage can be described as a highly promising, useful, and affordable option for archiving material that is rarely accessed because the retrieval speed is slow.

Various codes are employed to encrypt data into DNA, and this page examines and describes the various methods. The use of DNA molecules as secret writing mediums and safe data storage are also covered. Also covered are big data analytics and storage, as well as how DNA computing has helped with challenging computational issues. This study outlines the limitations of current encoding techniques and suggests ways to get around such restrictions.

III. HOW IT WORKS

DNA storage is incredibly small since it can be kept safely for tens of thousands of years in a dry, cool environment. Unlike HDDs, SDDs, and other memory devices, it won't degrade quickly. Oligonucleotide Synthesis machines are designed to upload and store information in DNA, and DNA Sequencing Machines are extremely complicated machines that can recover the stored data. The chemical synthesis of relatively small nucleic acid fragments with a clear chemical structure is known as oligonucleotide synthesis. Adenine (A), Cytosine (C), Thymine (T), and Guanine are the four nucleotides that make up the lengthy strands or sequences that make up DNA molecules (G). These nucleotides are made rather than generating sequences of 0s and 1s. It operates by associating digital data patterns (in binary form) with DNA nucleotides. Preparing Bits to Become Atoms is the fundamental idea at the beginning of the entire process.



Encoding

The four nucleotides A, T, G, and C are used to represent binary codes such as 00, 01, 10, and 11. For instance, the values of 00 and 01 may be A, C, T, and G respectively. Therefore, the biological representation of the binary form of digital data (01 11 10 00 11 11 10 11 01 00 01.....) is C-G-T-A-G-G-T-G-C-A-C-. As a result, the nucleotides in this sequence make up a DNA strand. Digital data is encoded in this manner.

Synthesis

Because longer DNA is chemically more difficult to construct, the artificial DNA should be shorter. A single DNA strand can only carry about 20 bytes, but digital data can be of very huge amounts. Therefore, data is divided into smaller chunks, and a sequence indicator is set up to make sure the pieces of data stay in the right order. As a result, the data is combined.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 7.105 ⁜ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9652

Storage

The ATGC nucleotides are mixed in a solution with additional chemicals to regulate reactions and the order of the strands, which in turn drives the chemical reactions employed in synthesis. We also gain from this approach by generating backup copies of each strand for use in multiple series at once. The newly synthesised DNA is now shielded from light and humidity-related deterioration. It is therefore dried and kept in a cold environment while also filtering off light and water.

Retrieval

The numerous DNA strands are now retrieved in a predetermined database order using the indicator put in place during DNA synthesis.

Sequencing

A device known as the Sequencing Machine is utilised to read back the data. It resembles the devices used today to analyse the genomic DNA found in various cells. This procedure results in the identification of molecules and the production of a letter sequence. To achieve the final format of digital data, this sequencing is carried out.

Decoding

The sequencing machine's created letter sequence is now reverse-coded into an ordered series of 0s and 1s. As of now, DNA can be damaged during this process, but as was already indicated, numerous copies of each sequence are created, and they are already in use. As DNA replication is a natural process, further duplicate copies can be created readily if these backup copies are also exhausted. In this approach, the entire DNA must be examined, even if we only need to read or access a portion of the information it contains. As a result, unique biochemistry approaches are being developed and researched to obtain only the necessary data much more quickly.

IV. MICROSOFT DNA RESEARCH

Additionally, the Microsoft Corporation has begun its own research and testing on DNA Data Storage. In fact, it got off to the most fruitful start in this technological industry. Microsoft was able to fit 200 gigabytes worth of information on books and other articles into DNA. Microsoft also bought 10 million oligonucleotides (strands of DNA) from Twist Bioscience to use in research, development, and data encoding. Microsoft is the only company to successfully recover or extract all of the encrypted data. The business revised its estimate that 1 cubic millimetre of DNA can hold 1 Exabyte (1 billion gigabytes) of data based on its own research. [9]

Microsoft's digital data continues to grow dramatically, thus the company plans to replace the thousands of acres of land used for data centres starting with one of them using DNA Data Storage technology. Additionally, it has a plan to temporarily offer DNA Data Storage to its cloud services. At first, DNA could only encode and decode data at a rate of roughly 400 bytes per second. Now that they are hitting rates of 100 megabytes per second, they are producing excellent and fruitful results. Microsoft has 36 Azure data centres, and 8 of those are now operational, 1 of which will be a DNA-based data centre. Microsoft collaborates with University of Washington on many projects.

V. CONCLUSION

DNA Data Storage technology reimagines how we store and retrieve data from electronic devices. Land issues can be naturally solved by DNA. It can take the place of the massive data centres found all over the world, which are quite expensive because they use up a lot of resources like power. Data storage in DNA is a natural method that uses fewer resources and allows us to preserve memories for thousands of years while passing on knowledge to future generations. This approach is scalable and works well for storing massive files. Additionally, making many copies is simple. Information can be kept in large data or archive systems using this technique. In the distant future, DNA-based storage methods could be used to store data in a secure manner for a long time and could solve the issue of limited space instead of utilising traditional storage devices, which have a limited capacity to store data.

REFERENCES

[1] Mullin, E. July 12, 2017. Scientists used CRISPR to Put A GIF Inside a Living Organism's DNA. MIT Technology Review. Retrievable at: https://www. technologyreview.com/s/608268/scientistsusedcrispr-to-put-a-gif-inside-living-dna/

DOI: 10.17148/IARJSET.2022.9652

[2] Keown, A. June 26, 2018. Boston's Catalog Secures $9 Million in Funding to Advance DNA Data Storage Technology. Biospace. Retrievable at: https://www.biospace.com/article/bostons-catalog-secures9-million-in-funding-toadvance-dna-data-storage-technology

[3] Choi, Y., Ryu, T., Lee, A. C., Choi, H., Lee, H., Park, J., ... & Kwon, S. (2019). High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. Scientific reports, 9(1), 6582.

[4] Sandle, T. July 16, 2017. Major tech companies see DNA storage as the future. Digital Journal. Retrievable at: http://www.digitaljournal. com/tech-and-science/science/microsoftexploring-dna-storagesolution/article/497735

[5] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E. M., Sipos, B., & Birney, E. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature, 494(7435), 77.

[6] Blawat, M., Gaedke, K., Huetter, I., Chen, X. M., Turczyk, B., Inverso, S., ... & Church, G. M. (2016). Forward error correction for DNA data storage. Procedia Computer Science, 80, 1011-1022. [7] Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-generation digital information storage in DNA. Science, 337(6102), 1628-1628.

[7] Stock, M. March 22, 2016. DNA data storage could last thousands of years. Reuters. Retrievable at: https://www.reuters.com/article/us-dna-storage/dna-data-storage-couldlast-thousands-of-yearsidUSKCN0WO1DX
DNA DATA STORAGE Dept. of MCA, RVCE Page 16 2022

[8] Gudeman, K. July 17, 2018. Illinois receives $1.5 million to lower cost, improve viability of DNA data storage. Retrievable at: https://csl. illinois.edu/news/illinois-receives-15-millionlower-cost-improveviability-dna-data-storage

[9] Mayer, C., McInroy, G. R., Murat, P., Van Delft, P., & Balasubramanian, S. (2016). An Epigenetics-Inspired DNA-Based Data Storage System. Angewandte Chemie International Edition, 55(37), 11144-11148.

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

[11] Fritz, M. H. Y., Leinonen, R., Cochrane, G., & Birney, E. (2011). Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Research, 21(5), 734-740.

[12] Newman, S., Stephenson, A.P., Willsey, M., Nguyen, B. H., Takahashi, C. N., Strauss, K., & Ceze, L. (2019). High-density DNA data storage library via dehydration with digital microfluidic retrieval. Nature communications, 10(1), 1706.

[13] Yong, E. Mar 2, 2017. This Speck of DNA Contains a Movie, a Computer Virus, and an Amazon Gift Card. The Atlantic. Retrievable at: https://www.theatlantic. com/science/archive/2017/03/thisspeckof-dna-contains-a-movie-a-computervirus-and-an-amazon-gift-card/518373/

[14] Siddaramappa, V., & Ramesh, K. B. (2019). DNA-Based XOR operation (DNAX) for data security using DNA as a storage medium. An Integrated Intelligent Computing, Communication and Security (pp. 343-351). Springer, Singapore.

[15] Heckel, R., Shomorony, I., Ramchandran, K., & David, N. C. (2017, June). Fundamental limits of DNA storage systems. In 2017 IEEE International Symposium on Information Theory (ISIT) (pp. 3130-3134). IEEE

