DNA based storage: Introduction, Characteristics, Applications and Challenges

Deepak Kumar Sharma, Shiv Kumar, Amit Kumar
Division of Information Technology
Netaji Subhas University of Technology
(formerly NSIT, University of Delhi)
New Delhi, India
dk.sharma1982@yahoo.com, shivk.it.16@nsit.net.in, amitk.it.16@nsit.net.in

***Abstract*— Over the years, as humans have made progress, data has come to the forefront and has become one of the principal elements of life. No matter the field, all aspects of life are now dependent on data in one way or the other. Be it hospitals or financial institutions; sports teams or researchers, all operate on some form of data during their functioning. This ever increasing dependency on data further leads to the need for its storage. The capability of present storage mechanisms is not able to keep up with the exponentially increasing demand. This along with other factors such as high setup costs, high maintenance charges, security and accessibility are pushing towards an alternative avenue of storage. DNA or the code of life is very similar to the binary based data systems that we operate on, hence is being looked at, as the alternative to conventional methods. This field has seen massive amounts of developments in the recent past and is finding a strong footing. Its theoretical capability to store all the data ever created in a finger sized device is one of the many factors, which makes it such an interesting field to study and know about. This paper describes how this domain of storage system(s) basically functions, the work done in this field in the past, its advantages and limitations along with the challenges that this domain needs to overcome to become practically viable bringing a paradigm shift in computing.**

Keywords*— *DNA, storage devices, storage crisis, random access, encoding, privacy, accessibility, digital data, coding theory

I. INTRODUCTION

In this era of information, where data is rightfully treated as one of the principal elements of life, one thing that cannot be ignored is the explosion in the rate of its creation and exchange. Fields like *big data*, *data mining*, *human computer interaction*, *genomics*, *social interaction* etc. are becoming increasingly popular around the world and are yielding enormous amounts of data and information every second. Enterprises are using the ability and insight provided by domains like *Artificial Intelligence* and *Machine Learning* to gain significant amounts of knowledge, about the created data in order to produce better products and services, creating even more data in the process. From *autonomous cars* to *smart home devices*, our world is changing, and fast, transforming our understanding of everything around us and data is at the centre of it all.

The realisation that over 90 percent of the world's data ever created has come up in the last 2 years of civilisation[1] alone, is overwhelming. Humans are estimated to be creating around 20 quintillion bits of data everyday, which is astounding[1]. Also the fact that the rate at which data is being produced is speculated to increase exponentially in the years ahead is quite hard to digest [1].

Based on current capabilities and technologies, the storage systems are only capable of storing 18% of the world's data and this amount will hit rock bottom at around 3% by 2030 and reduce to 0.5% by 2040.[2]

Apart from the worries of analysing, understanding and finding patterns in such big, unstructured and complex amounts of data, there lies one, seemingly trivial but important issue of its storage. Present approach of data storage which ranges from portable hard drives and flash drives to cloud based data centres just won't be able to keep up with the increasing demand for storage space. Plus, not only do these methods require real estate and infrastructure which is indeed very limited, but they also add to the costs that come in the form of initial setup, electricity, manufacturing and maintenance. Another issue is that of security; even with multiple layers of security, if the data is on the internet, it can be hacked and retrieved. The severity of such a threat, cannot be taken lightly. With it also comes the issue of accessibility, i.e. if the data size is very large, then it can't be carried in flash drives and external hard drives, and the user must be connected to the internet to access the data present in some remote cloud based service and, hence if the user is offline then carrying big data becomes a big problem. These solutions also run the danger of being corrupted when brought in touch with certain external environments, which brings in contention their ability to act as archival systems for very long periods of time.

Silicon along with other elements which are non biodegradable, are also a concern to increasing pollution in the environment. Hence, it can be said, that it is time that we look for other alternative solutions to the ones presently employed for data storage. Otherwise, we are headed towards a crisis for which we are heavily underarmed.

One alternative that will be discussed in this paper is that of *deoxyribonucleic acid (DNA)* based storage[3], i.e. *nature's own storage medium.* They are of interest in this area due to some of their properties which include but are not limited to being highly-dense, sturdy and long lasting.

Just one gram of DNA is said to store about 455 exabytes of data[4], solving the real-estate and infrastructure side of problems and since that much amount of data can be localised to that little space, it allows users to carry their own data centres along with them wherever they go, providing them with better accessibility as well as returning physical ownership of their data. This also solves the problem of lack of privacy in conventional storage mediums.

Another aspect to be considered here is, the power required while working with DNA. It is very little as compared to conventional data storage solutions which require huge amounts of energy to maintain.[5]

It has also been seen that scientists have been able to extract information from DNA, known to be thousands of years old[6,7], hence making this form of storage very robust and immune to corruption . Generating copies of DNA is also possible using *Polymerase Chain Reaction* techniques, hence exchange of information does not get hindered if the transition takes place to this medium of storage. Also, since DNA is bio-degradable, it poses less of a threat to the environment as compared to its conventional counter-parts, which increases the life of the planet humans inhabit.

In the present paper, the structure of DNA and the concept of memory hierarchy are explained in brief in the second section which is used in section three to understand and review the work already done in this field. This is used in section four to understand the issues and challenges that DNA based storage faces in order to become the standard of storage around the world. We conclude by understanding the current situation and the future possibilities in the field.

II. BACKGROUND

DNA or *deoxyribonucleic acid*[3] is an extremely long chain of molecules that contains all the information necessary for the functioning of any living cell. It is the basic unit of genome which is *an organism's complete set of genetic instructions,* and hence considered the fundamental unit of life. Its structure (as shown in figure 1.) is said to be in the form of two strands made of molecules called nucleotides, intertwining to form a structure referred to as the double helix structure.

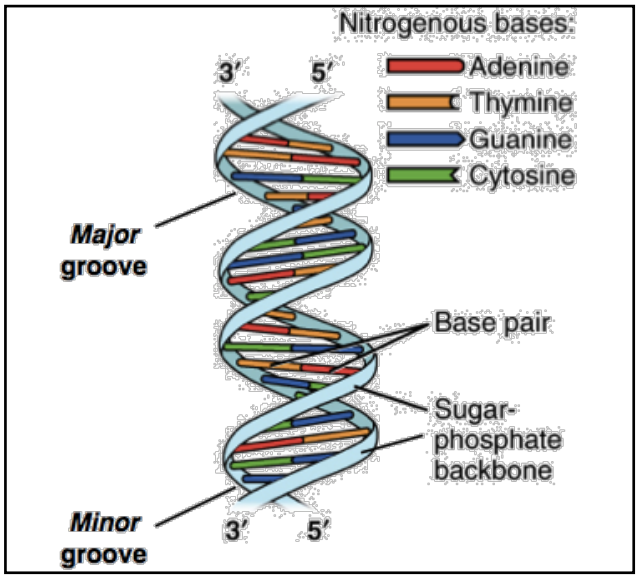

Figure 1. Molecular structure of DNA [3]

The nucleotides which make up these strands are themselves made up of some fundamental blocks, that are:

one of four possible bases that include, Adenine (A), Cytosine (C) , Guanine (G) and Thymine (T) , a phosphate group and the five carbon sugar.

Each base has a corresponding complimentary base and it forms a pair with only that complimentary base on the opposite strand, through hydrogen covalent bonds and together these complimentary pairs are called base pairs (*thymine compliments adenine using two H bonds and cytosine compliments guanine using three H bonds*). The double helix structure is very stable due to its regular shape and stacking of layers. The different arrangement of these base pairs gives different properties to all living beings. From eye colour to better agility, from regeneration properties to genetic diseases, everything is governed by the arrangement of these pairs. Different combinations of these base pairs hence, although very simple in concept leads to millions of possibilities which is similar to how the binary codes employed in the computational systems of today, use boolean logic to construct systems of such high complexity. The bases sequence can be seen analogous to a base 4 system i.e. 4 possible fundamental units to form complex combinations. This similarity with binary system is one of the prime reasons for the consideration of DNA as an alternative to current methods of data storage, and hence is also called the *code of life.*

Since we have established a direct relation between the binary form of data we are familiar with and the nucleotides based DNA, we need to understand how we can convert the binary form of data to nucleotide form of data.

There are many ways to establish this, but the basic idea is to assign codes to DNA nucleotides, encoding it in a way. For example, 00 be coded as A, 01 be coded as C, 10 coded as T and 11 coded as G. Now, for example, to store an image file, we encode it into a string of 0s and 1s. Lets say the first 8 bits are as follows 00111001. On breaking them into pairs, in sequence and then encoding as mentioned above, we get A-G-T-C. That's how we need to join the nucleotides to form a single strand of DNA to represent equivalent binary data.

So we can see the nucleotides equivalent to a 4 base system and accordingly encode data as is done by combining 0s and 1s to represent complex forms of data. The important thing post the theoretical conversion of this binary form of data into corresponding base(s) compatible data is to be able to read and write it accurately. That is where processes such as DNA synthesis and DNA sequencing come in.

DNA synthesis [8] is the natural or artificial process of synthesising strands of DNA. This method is used to create synthetic strands of DNA, and used to write the base(s) form of data obtained from binary data. First the single strands are assembled with the combination of the code to be encoded in the form of A,C,G and T, the sequence of which is obtained from the process as explained above; to form oligonucleotides (*short DNA molecules)* and later pairing them up correctly, using techniques such as oligonucleotide synthesis and polymerase chain reaction to form DNA strands. Further explanation of which is available here. *(provide reference).*

DNA sequencing [9] is the process of extracting the sequence of nucleotides from the strands of DNA, and is of prime importance in the process of retrieving the encoded data. There are many methods employed to sequence long strands of DNA with high efficiency and precision. More information on the methods used currently and ones in development can be found here.(provide reference)

Apart from the basic knowledge of DNA, another aspect to look into is of memory hierarchy which is an arrangement of different storage modules based on response time. Based on their response time and costs associated with them, different storage modules serve different purposes in a computer system. Processor registers are on top of this hierarchy yielding fastest response time followed by cache and RAM. The slowest response time are of tape based memories which are used to back up data.

How DNA based storage, if needed, can compliment the current system of memory modules is another important topic of discussion based on its access time and costing.

III. REVIEW OF PREVIOUS WORK

The use of DNA in computational systems to solve problems like Hamiltonian graph, was first proposed by L.M. Adleman [10]. Clell, Risca and Bancroft devised an approach of storing information in DNA, by encoding it into the *four* bases as stated before [11]. The inspiration for this approach was derived from the second world war where a steganography technique based around a microdot containing a downscaled picture of a letter, was used to communicate secret data. The researchers applied this approach on a DNA scale and hid data in a DNA strand. They synthesised the encoded strand sandwiched between polymerase chain reaction primer sequences and also used an encryption key.The encryption key as showcased in figure 2, was the mapping between the combination of three base(s) and the corresponding numerals and alphabets. This encoded strand was then surrounded by more than hundred times size worth of human DNA molecules meant to conceal the data. This provided the encoded message much required privacy and security. Only, a person with access to the PCR primer sequences and the encryption key would be able to decipher the code.

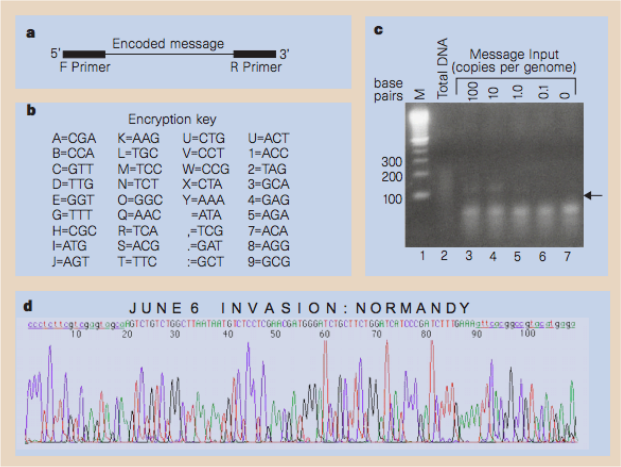

Figure 2. Genomic Steganography, [11]

An interesting and important observation from this study was that no matter the human DNA, if the primer sequences and encryption key were known then the data could be extracted. This research also established the strength of DNA storage as a much more secure and private storage medium when compared to its conventional counterparts in the field of steganography.

Bancroft et. al[12] followed a similar mechanism. They classified DNA into two types for this process, i.e. information DNAs (iDNA) which contained the information and the poly primer key (PPK) (*which is key to extracting information stored in iDNAs*). Each iDNA strand contained some basic elements, i.e. unique information segment, small common spacer indicating the start of the stored information, unique sequencing primer and common flanking forward and reverse amplification primers. They came up with an experiment which included amplification using polymerase chain reaction to extract data. They encoded and accurately extracted the opening lines of Charles Dicken's *'A Tale of Two Cities'*. Although, this study did not take into account the response of this storage media in external conditions since it was studied under strict laboratory conditions, this experiment proved the possibility of development of dedicated DNA based storage systems, which was a huge step in paving the way for future work to be done in this field.

The next major contribution[13] to this field was made by Wong et al , based on the problems in the structure of DNA strands that could break at both ends, which could lead to loss of information. They suggested that to prevent DNA from harsh conditions, a dependable medium to store the encoded strands and synthesised gene sequences were required. They employed a vector containing encoded data which is able to grow and accumulate to ensure longevity of encoded data, for this purpose. They used agents with high rapid regeneration rate and tolerance towards radiation as well as vacuum, such as Escherichia coli and Deinococcus radiodurans. This research was important as it showcased the protection mechanism needed to be employed for the protection of encoded data in this storage mechanism from harsh external conditions. This further proved the competence and capability of DNA based storage devices to be used for archival purposes given its density and longevity.

Up until 2012, the largest project to encode data in DNA was of 7920 bits. Then in [14], George M. Church et al. published a landmark paper highlighting their work on converting "*an html coded draft of a book that included 53,426 words, 11 JPG images, and one JavaScript program into a 5.27-megabit bitstream"* into encoded DNA. They were able to recover "*all data blocks with a total of 10 bit errors out of a total 5.27 million*" which was remarkably unprecedented. Their method had five advantages over the techniques employed in the past. The most important ones are highlighted here; they used one bit to encode each base instead of the formerly used two which allowed them to encode data in many ways avoiding sequences that might be difficult to read or write. They had split the bit stream into address blocks eliminating the need for long DNA strands that were difficult to assemble at that scale.The advances in DNA synthesis and sequencing technology (which had been a major hindrance for research in this area in the past years) had allowed them to encode and decode data in large amounts in almost hundred thousand times less the cost as compared to first generation encodings.

Nick Goldman and his team at EMBL-European Bioinformatics Institute[15] created a way to encode

"*739 kilobytes of hard-disk storage, synthesized DNA, sequenced it and reconstructed the original files with 100% accuracy."* Their data comprised of "*Shakespeare's Sonnets in ASCII format, Watson and Crick's 1954 Classic Paper, a*

medium resolution colour photograph, a snippet of the audio from Martin Luther King's famous speech in MP3 format and a Huffman code (again in ASCII text scheme) for an aggregate data size of 757,051 bytes."

Their analysis had showcased that DNA based storage systems could be scaled to handle global storage requirements and could become a realistic alternative for digital archiving, (which places the DNA based storage medium at the bottom of the storage hierarchy). This research came out as a major breakthrough for DNA based storage devices. It indicated that if the costs related to DNA synthesis were to reduce at the pace that they were, their scheme could become cost effective within a decade.

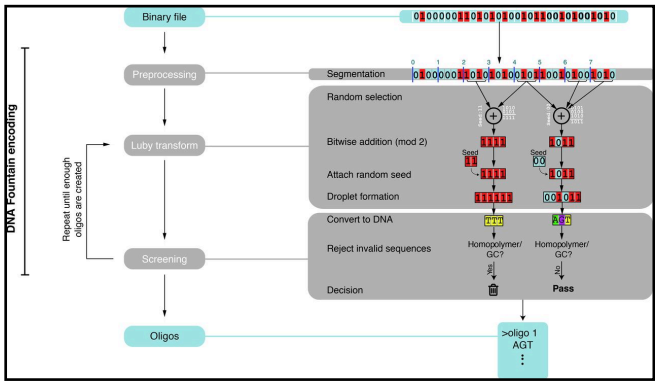

Figure 3. DNA Fountain encoding, [16]

Later in [16], researchers at Columbia University by the name of Dina Zielinski and Yaniv Erlich along with New York Genome Centre published a new technique called DNA fountain, using which they were able to store a complete Operating Systems and related files, that had a data storage density of 215 petabytes per gram of DNA which was approximately 85% of the theoretical limit of the Shanon capacity, orders of magnitude higher than previous studies and attempts. Figure 3 summarises the process, employed in this research. The issue with this technique was that it was very costly (around $3500 per MB to synthesise) and not feasible for large scale use.

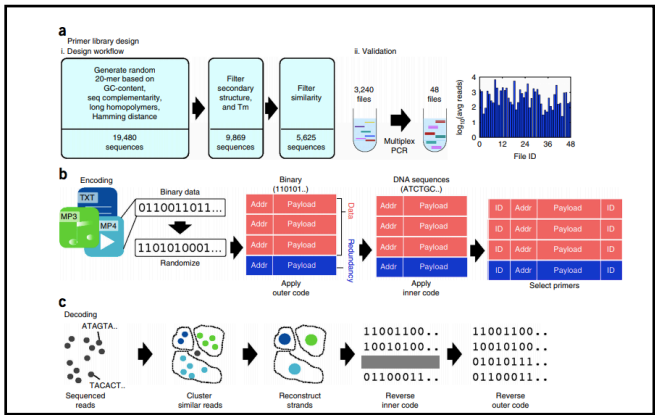

Figure 4. Design of random access primers and coding algorithm, [17]

A problem in information retrieval had been the need to sequence the entire pool of DNA data, even if a small subset was required which was inefficient and was a major contributing factor to the cost associated with DNA based storage. In [17], Microsoft in association with University of Washington demonstrated storage and retrieval of an unprecedented 200 MB of data across 35 distinct files. As depicted in figure 4, they used a large collection of primers to enable individual recovery of data. This research tackled the problem of sequential access of data in DNA based storage systems by showcasing random access and proved the viability of large scale DNA based storage systems for storage and retrieval.

IV. ISSUES AND CHALLENGES OF DNA BASED STORAGE SYSTEMS

Considering the work done up until now, it is very probable that DNA based storage systems will be the dominant technology around the world one day. But there are some issues and challenges that need to be handled in order for that to happen. The biggest challenge for DNA based storage remains the cost associated with the synthesis and sequencing of information on scale. It is imperative for this cost to be made much more affordable in order for this technology to be adopted by industrial ventures in the near future. With the boom in data creation, we need such affordability to come as soon as possible. Large strides have been taken in the last few years and if the growth continues at the same rate, then there is no reason why this technology can't become viable.

Another challenge is that of the total time required in the process of storing data in DNA based systems, which is many magnitudes more, as compared to its conventional counterparts. Even if it is placed in conjunction with the current memory hierarchy and placed at the bottom of the pyramid, still the write and read time needs to improve many folds for this technology to be both scalable and practical.

The machines used today in the process of storing data in DNA are throwing very less error rates but even these error rates are significant when we consider the complications related to the process of storing data and therefore the machines need to be made much more efficient and accurate in order to achieve better rates of successful conversions.

V. CONCLUSION AND FUTURE WORK

We are past the days where DNA based storage was considered a part of people's imagination. DNA based storage devices are here to stay and like any other revolution in the past, it also has to face some big challenges in order to be adopted by the world. The benefits this technology provides such as stability, storage density, energy efficiency and robustness are too significant to ignore and it is for sure that this technology will be used for archival purposes. In conjunction with present tape technologies. Expecting major leaps in development of this technology won't be asking too much looking at its figures in the past where it has given over a million fold improvement in recent years. Against the 1.5 times a year progress in electronic technology, this technology has grown 10 times every year since we started reading and writing. A major contribution to this progress goes to the improvement in cost efficiency. We have recorded five and twelve times reduction in the cost of synthesis and sequencing respectively. The migration to DNA based storage would also lead to growth in the DNA computing field. This would further speed up the process of reducing synthesis and sequencing costs.

REFERENCES

1. Data Never Sleeps 5.0. https://www.domo.com/learn/data-never-sleeps-5?aid=ogsm072517_1&sf100871281=1
2. Next Chapter. How DNA Storage will transform data centre design beyond recognition. https://data-economy.com/next-chapter-dna-storage-will-transform-data-centre-design-beyond-recognition/
3. What is DNA? https://ghr.nlm.nih.gov/primer/basics/dna
4. DNA data storage breaks records. https://www.nature.com/news/dna-data-storage-breaks-records-1.11194
5. L. M. Adleman, Molecular computation of solutions to combinatorial problems. Science 266, 1021 (1994). doi:10.1126/science.7973651 Medline
6. J. Bonnet et al., Chain and conformation stability of solid-state DNA: Implications for room temperature storage. Nucleic Acids Res. 38, 1531 (2010). doi:10.1093/nar/gkp1060 Medline
7. S. Pääbo et al., Genetic analyses from ancient DNA. Annu. Rev. Genet. 38, 645 (2004). doi:10.1146/annurev.genet.37.110801.143214 Medline
8. DNA Synthesis.http://www2.csudh.edu/nsturm/CHEMXL153/DNASynthesis.htm
9. DNA Sequencing.https://www.genome.gov/10001177/dna-sequencing-fact-sheet/
10. L. M. Adleman, Molecular computation of solutions to combinatorial problems. Science 266, 1021 (1994). doi:10.1126/science.7973651 Medline
11. C. T. Clelland, et al., “Hiding messages in DNA microdots,” Nature, vol. 399, no. 1033, pp. 533–534, 1999.
12. C. Bancroft, T. Bowler, B. Bloom, and C. T. Clelland, “Longterm storage of information in DNA,” Science, vol. 293, no. 5536, pp. 1763-1765, 2001.
13. P. C. Wong, K. K. Wong, and H. Foote, “Organic data memory, using the DNA approach,” ACM, vol. 46, no. 1, pp. 95-98, 2003.
14. G. M. Church, Y. Gao, and S. Kosuri, “Next-generation digital information storage in DNA,” Science, vol. 337, no. 6102, pp. 1628, 2012.
15. N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E.M. LeProust , B. Sipos, E. Birney, “Towards practical, high-capacity, low-maintenance information storage in synthesized DNA”, Nature 494:77–80
16. Yaniv Erlich, Dina Zielinski, “DNA Fountain enables a robust and efficient storage architecture”, Science, 355:950-954
17. L. Organick, et al. ,”Random access in large-scale DNA data storageNature Biotechnology”, Nature Biotechnology, 36:242-248