RESEARCH ARTICLE OPEN ACCESS

DNA: The New Age Data Storage System

Ms. Meghana Bhandari[1], Ms. Leena Bhandari[2]
1Department of Biotechnology, Sinhgad College of Engineering, Vadgaon, Pune-411041
2 Department of Computer Engineering, Sinhgad Institute of Technology and Science, Pune-411041

Abstract:

The quantum of digital data being produced has long been surmounting the immensity of available storage devices. Demand for data storage has radically overhauled the capacity of existing storage media. Despite periodical developments that lead to improvement in the efficiency of optical discs, storing a zettabyte of data would still take many millions of its units, thus occupying significant physical space and a considerable amount of man power to ensure the safety of the space as well as the credibility of stored data. On the other hand, deoxyribonucleic acid (DNA), a thread-like chain of nucleotides is a molecule that stores magnanimous amount genetic information used in the growth, development, functioning and reproduction of all known living organisms and many viruses; in the form of its own language or codes. It can therefore be proclaimed as the “Hard disk of Human Life”. It is very small in size, does not require a huge repository and considering the life span of a human, data extracted from age old fossils and other sorts of archaeological studies, the average shelf life of a DNA molecule can span up to several hundred years, without hampering the plausibility of the stored data. In this review we further explore the idea of using DNA as a storage medium for an enormous volume of data. This paper also discusses the feasibility and the challenges associated with this approach.

Keywords **— DNA computing, Big data, storage, encoding, DNA based archive**

I. INTRODUCTION

Magnetic discs, floppies, CDs, hard discs and pen drives are proof that the constant progress in the field of data storage has always led to compactness of the storage device and successive increase in its individual capacity to store data. However, now we have reached a point where the amount of data being produced is in no way even close to the rate at which new storage devices are concocted. Hence, there have been conscious efforts going on since the past few years to explore new methodologies to store large chunks of data and one such proposed method that has garnered much attention of late is the use of DNA for data storage.

Deoxyribonucleic acid (DNA) is a double stranded helical structure of nucleotides that contain four different nitrogen bases namely Adenine (A), Thymine (T), Guanine (G) and Cytosine(C) which carry the genetic blueprint of life. A normal human DNA is approximately 3 metres in length and consists of around 30,000 genes. The genetic information is stored in these genes in the form of triple letter genetic codes or codons that code for various metabolic proteins [6]. Another intriguing feature of DNA is that despite of its enormous length it is neatly packaged into condensed structures called chromosomes in a cell with a diameter ranging from 10-100μm. The two essential features of DNA that contribute towards showing a great potential in transforming this double stranded helix into a reliable storage system are : 1) the extensive quantity of data that a DNA strand can store 2) the minimal amount of physical space that a DNA takes up.

A DNA storage system takes data as input, synthesizes DNA molecules to represent that data, and stores them in a library or pool of DNA. To retrieve the data, the system screens molecules from the pool, amplifies them with PCR (Polymerase Chain Reaction), and sequences them back to digital data. As there are four nucleotides and simply two binary digits, the conversion of one to later appear arduous; we instead convert binary data to base 3 and employ a rotating encoding from ternary digits to nucleotides. This encoding avoids homopolymer repetitions that might have occurred during the conversion of nucleotides to binary digits thereby significantly reducing the chances of errors while retrieving data. A reasonably efficient strand length after complete DNA synthesis is 120 to 150 nucleotides, which gives a maximum storage capacity of 237 bits of data in a single molecule using this ternary encoding. The write process therefore fragments input data into small blocks that correspond to separate DNA molecules. This blocking approach also ensures added redundancy.

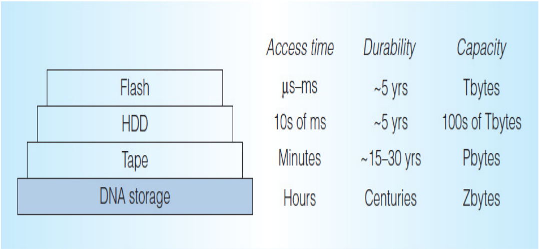


Fig 1 Progress in Storage technology[1]

II. Historical Background

Most of the papers relating to this topic have been published by the Microsoft research centres since they have successfully carried out Data storage experiments in their wet labs and have set a new benchmark with more than 200MB data recovered error free in collaboration with the University of Washington.

Sr. no.	Paper Title	Year	Paper theme/ idea	Advantages and limitations
1	Review of Big Data Storage based on DNA Computing	2015	They proposed an approach as simple method to store data into DNA. The experiment work is done to validate the proposed approach.	Results clearly show advantages and merits of the proposed model.
2	A DNA-Based Archival Storage System	2016	This paper presents architecture for a DNA-based archival storage system.	They also propose a new encoding scheme that offers manageable redundancy, compensating reliability for density.
3	Encoding Movies and Data in DNA Storage	2016	A systems level design is presented for encoding movies and digital information in DNA storage.	They have only discussed from an architectural point of view without getting into the details.
4	Toward a dna-based archival storage system	2017	This paper presents a dna-based archival storage system, performs wet lab experiments to showcase its viability, and identifies technology trends that might lead to increase in its practical usage.	This paper does not get into the details of the wet lab experiments but presents the whole idea and proves its feasibility.
5	Random access in large-scale DNA data storage	2018	They encoded and stored 35 distinct files (over 200 MB of data), in more than 13 million DNA oligonucleotides.	They designed and validated a large library of primers that enable individual recovery of all files stored

Table 1 Previous Records

III. Methodologies Used

The proposed DNA archival storage system makes use of the phenomenon of data encoding and decoding in alliance with the DNA nucleotides. The following block diagram represents the main principle behind this ideology.

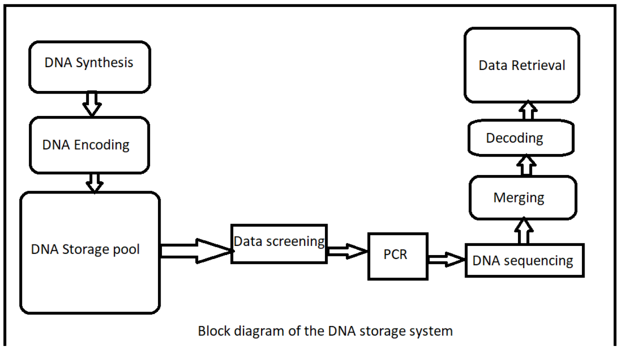


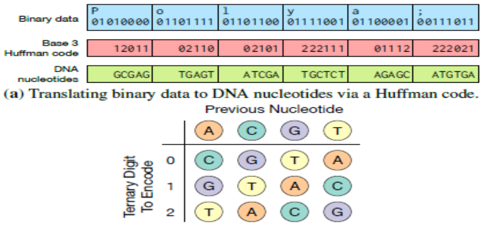
Fig 2 Block diagram of DNA storage system

3.1. *DNA Synthesis*

Chemical synthesis of arbitrary strands of DNA is possible yet not literally feasible as it is slow process and does not provide a good yield. DNA replication, being a semi-conservative technique requires a template strand to initiate the process of replication. The coupling efficiency of a synthesis process is based on the probability that a nucleotide binds to an existing template strand at each step of the process. Although the coupling efficiency for each step can be higher than 99%, it may reduce at times as a result of inefficiencies in maintaining proper temperature and pH conditions for replication, inaccurate binding of primer to the initiator region, unavailability of successive nucleotides or DNA polymerase degradation. These small errors may lead to an exponential decrease of product yield with increasing length thus limiting the size of oligonucleotides that can be efficiently synthesized. In order to avoid this problem, synthesis of a given sequence is done by utilizing a large number of parallel primers which results in many pruned byproducts in addition to multiple copies of the full length target sequence thereby ensuring a fairly good yield.

3.2 Data Encoding

Like every other storage media, DNA also needs encoding and decoding. The only difference in this scenario is that, instead of converting the data to binary bits, data is converted to the alphabets of nucleotides. For encoding, varieties of algorithms have been proposed such as Huffman encoding, Goldman encoding, XOR encoding, etc. The researchers of Microsoft found out that the XOR encoding is the best method which provides the accuracy and also reduces the overheads caused by the Goldman encoding.



Binary data	P 01010000	o 01101111	l 01101100	y 01111001	a 01100001	; 00111011
Base 3 Huffman code	12011	02110	02101	222111	01112	222021
DNA nucleotides	GCGAG	TGAGT	ATCGA	TGCTCT	AGAGC	ATGTGA

(a) Translating binary data to DNA nucleotides via a Huffman code.

Previous Nucleotide

Ternary Digit To Encode	A	C	G	T
0	C	G	T	A
1	G	T	A	C
2	T	A	C	G

Fig 3 i) Translating binary data to DNA nucleotides via a Huffman code. (ii)A rotating encoding to nucleotides avoids homopolymers (repetitions of the same nucleotide), which are error-prone[1]

The encoding algorithm for the same can be given as:

(1) Read data stream: A

(2) Check size of the data [r,c,n] = size (A) where r = rows, c = columns, n = number of matrix

(3) Calculate DNA sequences size for image

(4) Create a zero matrix of DNA sequences size

(5) While DNA sequences size = max

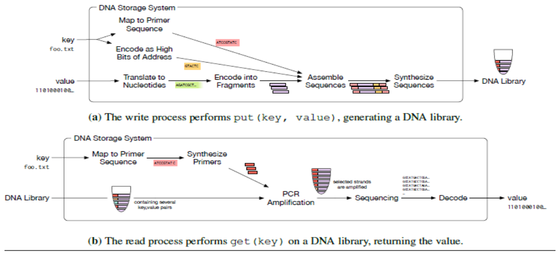
(6) Convert even smallest piece of data to binary form

(7) Insert binary DNA code of an individual data cell to DNA sequence

(8) Continue till all of max size of DNA sequence is reached

3.3. Data Screening from DNA Pool

After encoding data into DNA nucleotides and storing these molecules into pools of DNA, the desired data can be retrieved by selectively screening these pools with complementary primers that bind to the necessary data strand. However, DNA is a very lengthy molecule which makes it complicated to remember which data is stored in which part of the strand. Therefore the DNA archival storage system proposes a key or indexing feature for every data that is encoded within a strand. It implies that a predetermined promoter region i.e. a short length strand of nucleotides whose sequence is already known and which can induce the binding of strand with complementary primer, which maintains an index of the encoded data is attached to every strand. The primer bound strand is then amplified using polymerase chain reaction. These amplified strands are sequenced to retrieve data.



(a) The write process performs put(key, value), generating a DNA library.

(b) The read process performs get(key) on a DNA library, returning the value.

Fig 4 Data retrieval technique[1]

3.4. DNA Sequencing

The strand of interest serves as a template for the DNA polymerase, which creates a complement of that strand; fluorescent nucleotides are majorly used during this synthesis process. Since each type of fluorescent nucleotide emits a different colour, it is possible to read out the complement sequence optically. Sequencing is error-prone, but when compared with synthesis, in totality, sequencing typically produces sufficient precise reads of each strand. The sequenced data then needs to decoded in order to retrieve the required data in binary format. The decoding algorithm that helps retrieve the desired data from DNA storage system is given as:

(1) Read DNA sequence

(2) Calculate size : length, size of individual cells

(3) Convert DNA to real data decode

(4) Convert one cell to data

(5) Insert the converted value to template data matrix

(6) Stop when entire DNA is converted

IV. CONCLUSIONS

DNA-based storage has the potential to be the ultimate archival storage solution as it is extremely dense and durable. However, this technique is not practically feasible. This is because the methods currently employed for DNA synthesis, PCR and sequencing are really expensive. Besides the DNA also has a tendency of getting mutated with respect to time. This may cause major errors in data retrieval processes or may even cause significant loss of data. DNA can thrive viably in very specific conditions of temperature and pH, hence it is very necessary that these conditions be maintained in the DNA storage pools otherwise there are chances of denaturation of these strands. Apart from the data encoding and decoding stages in this system rest entire work requires the participation of highly skilled and experienced professionals which is one of the many reasons that makes this project expensive and difficult to implement. Although its implementation in normal computers seems far from reality, this may be employed at datacenters. Hopefully in the future, such datacenters would take up a lot lesser space as compared to the ones we see today. If only the facility of data retrieval is improved, the system can be implemented in cloud storage. DNA storage continues to be an emerging, innovative, and viable technology as the cost of high-throughput DNA synthesis and DNA sequencing may decrease in the near future with the rapid progress in the field of biotechnology and the conception of innumerable varied sequencers and synthesizing techniques.

REFERENCES

1. *James Bornholt, Randolph Lopez, Douglas M. Carmean, Luis Ceze, Georg Seelig,Karin Strauss. Toward a dna based archival Storage system: IEEE 2017*

2. *Naveen Goela, Jean Bolot. Encoding Movies and Data in DNA Storage: 2016*

3. *Hanadi Ahmed Hakami, Zenon Chaczko, Anup Kale Review of Big Data Storage based on DNAComputing: Asia-Paci_c Conference 2015*

4. *James Bornholt, Randolph Lopez, Douglas M. Carmean, Luis Ceze, Georg Seelig,Karin Strauss. .A DNA-Based Archival Storage System: Microsoft Research Publication, 2016*

5. *Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin,Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bich-lien Nguyen, Christopher N Takahashi1, Sharon Newman, Hsing-Yeh Parker, CyrusRashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, DouglasCarmean, Georg Seelig, Luis Cez1 & Karin Strauss. Random access in large-scale DNAdata storage: Microsoft Research Publication, 2018*

6. *Griffiths AJF, Miller JH, Suzuki DT, et al. An Introduction to Genetic Analysis. 7th edition. New York: W. H. Freeman; 2000. Genetic code. Available from: https://www.ncbi.nlm.nih.gov/books/NBK21950/*