DNA: The Future Storage Device

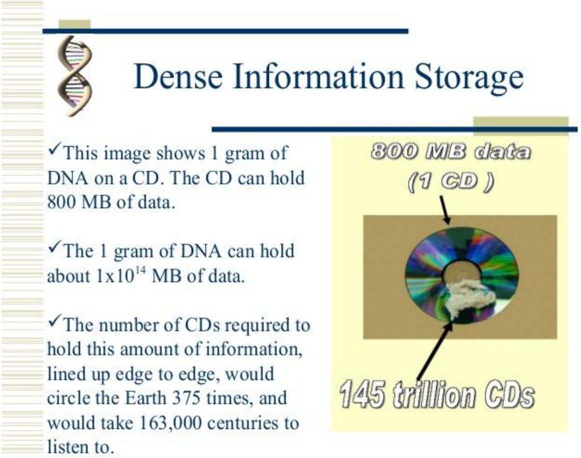
Sakshi Jha[1],
[1]Ganga Institute of Technology and Management,
MDU, Rohtak

Mr. Mahesh Kumar Malkani[2]
[2]Department of Computer Science Engineering and CFIS,
GITAM, MDU, Rohtak

Data Storage has been a great concern since the early ages. With the advent of time, data has been expanding and new storage devices have been satisfying the demand. Demand for data storage is growing on a faster pace .In order to meet this growing need for storage device, a shift from electronic media to molecular DNA has been made. DNA molecule is universal and fundamental data storage unit in biology. 1gram of DNA molecule can store 1 Zettabyte of data. A new encoding scheme that offers controllable redundancy, trading off reliability for density has also been developed. Finally, we highlight trends in biotechnology that indicate the impending practicality of DNA storage for much larger datasets.

I. INTRODUCTION

DNA digital data storage is storing digital data in the base sequence of DNA. Artificial DNA made using commercially available oligonucleotide synthesis machine is used for storage and DNA sequencing machines are used for retrieval. This type of storage system is more compact than current magnetic tape or hard drive due to its huge density .Digital data universe has been growing exponentially. Trends in data storage are setting new milestones. From floppy disk to tape drives and tape drives to cloud storage have been the new developments in the field of storage devices. Most of the world’s data is stored on magnetic tapes and optical disk drives. Despite this improvement, storing Zetta bytes of data would still take millions of years and space. With the increasing technology and tech-literacy, there is a need for revolution in the arena of archiving. DNA is a good storage medium as it is more compact and dense than the available conventional storage devices. Present long term archival storage system require refreshes to remove the corrupted data. In order to preserve the world’s data ,we need to look for those storage devices that are dense and durable.



Dense Information Storage

✓This image shows 1 gram of DNA on a CD. The CD can hold 800 MB of data.

✓The 1 gram of DNA can hold about 1x10^{14} MB of data.

✓The number of CDs required to hold this amount of information, lined up edge to edge, would circle the Earth 375 times, and would take 163,000 centuries to listen to.

Figure.1[4]Density of DNA Storage

Synthetic DNA sequences are a solution to store digital data i.e. increasing at an alarming rate.DNA is a extremely dense with a theoretical limit above 1 EB/mm^3.DNA based storage is of eternal relevance as long as there is a DNA based life , there will be strong resons to read and write data. Digital data in DNA Storage is mapped into DNA nucleotide sequence[1] , synthesizes the corresponding DNA molecule and stores them . Sequencing the DNA molecules and then decoding the digital data back to original data ,these two processes are included in reading the data. Progress in DNA storage has been rapid.The main innovation in this is the use of error correcting encoding scheme to ensure extremely low data loss.This was achieved in 2015 via Reed Solomon Error correction coding and by encapsulating the DNA within silica glass spheres via Solo-gel[2] chemistry.

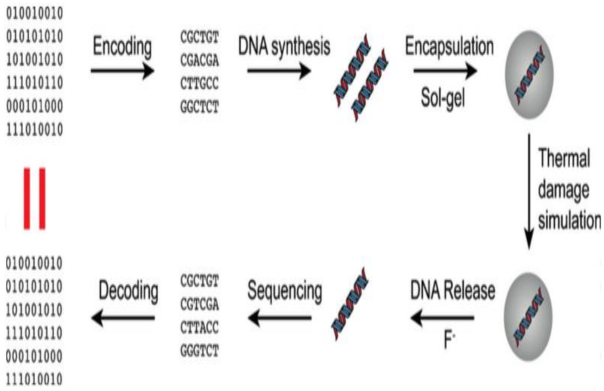


Figure.2[5] Synthesis and Sequencing

With so many developments in the storage and retrieval process from DNA storage device ,there remained a shortcoming. The whole strand of synthetic DNA has to sequenced in order to retrieve only one data set out of the several data sets that are previously encoded.

To overcome this short come, three classical coding schemes: Huffman coding, Differential coding and Single parity-check coding were applied. Data compression is done using Huffman coding. Differential coding eliminates homo-polymers in DNA strings. Single Parity check were used to add controlled redundancy, which in conjunction with four-fold coverage allows for mitigating assembly errors. Due to dynamic changes in biotechnological systems, none of the above coding schemes proved to be a perfect solution. Furthermore, selection of blocks to be rewritten is made possible by the Prefix Encoding format, while rewriting is performed via two DNA editing techniques, the gBlock and OE-PCR(Overlap Extension PCR) methods. Randomly accessing data in DNA based storage is highly problematic. Read latency is much longer than write latency. Polymerase Chain Reaction, PCR amplifies only the desired data. This design both accelerates, reads and ensures that an entire DNA pool need not be sequenced.

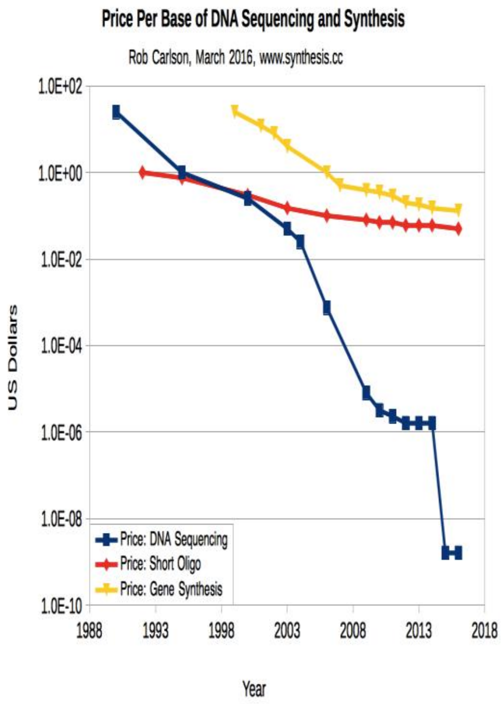
II. DNA SYNTHESIS:

Deoxyribonucleic Acid consists of four types of nucleotides: Adenine(A),cytosine(C),Guanine(G),and thymine(T).A DNA strand ,or oligonucleotide is a linear sequence of these nucleotides. The coupling efficiency of a synthesis process is the probability that a nucleiotide binds to an existing partial strand at each step. DNA already has a digital flavor. The molecules stick to each other in a very programmable manner. The strings of 1s and 0s are mapped into strings of As, Cs, Gs and Ts. Polymerase Chain Reaction technique eases this process of mapping. The physical storage medium is synthesized DNA .Decoding of these strings in 1s and 0s gives the original data .

III. DNA Sequencing: It is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of four bases-Adenine, guanine, cytosine and thymine-in a strand of DNA. Several high throughput sequencing techniques exist but DNA polymerase enzyme technique is widely accepted. Sequencing of DNA that involves DNA polymerase enzymes are commonly referred to as “sequencing by synthesis” .The strand of interest serves as a template for the polymerase, which creates a complement of the strand. Fluorescent nucleotides used during this process emit different color. The complement sequence can be read out optically. Sequencing is error prone, but as with synthesis ,in aggregate ,sequencing produces precise reads of each strand.

Sequencing and Synthesis Improvement Projections: Carlson curve[3] describes the biotechnological equivalent of Moore’s Law and is name after the economist, Rob Carlson. It was predicted that the doubling time of DNA sequencing technologies (measured by cost and performance) would be at least as fast as Moore’s law. Carlson Curve illustrates the rapid decrease in cost and increase in performance of variety of technologies, including DNA sequencing, DNA synthesis and a range of other computations.

Moore’s law started being profoundly outpaced since January 2008 . It was followed by the shift from Sanger’s DNA Sequencing to Next generation Sequencing. These next generation sequencing techniques have helped in plunging down the cost of DNA synthesis. The prices are falling faster than the cost of fabricating transistors did over the past 50 years ,a trend that has been the engine of innovation in computing



Graph1[6]: Trends in DNA synthesis and DNA sequencing

As per the Carlson's latest curve ,it can be figured out that the cost of DNA Synthesis i.e writing the digital data and DNA Sequencing i.e reading the digital data on DNA storage is decreasing at an exponential rate. Important biotechnology applications such as genomics and the development of smart drugs ensure that these improvements are eventually making DNA data storage a viable application.

IV. DNA DATA STORAGE: DNA

synthesis and DNA sequence are the two important tasks in the DNA storage. DNA synthesizer encodes the data to be stored in a DNA, there is a storage container with compartments that store pools of DNA that map to a volume. DNA sequencer reads the DNA sequences and converts them back into digital data.

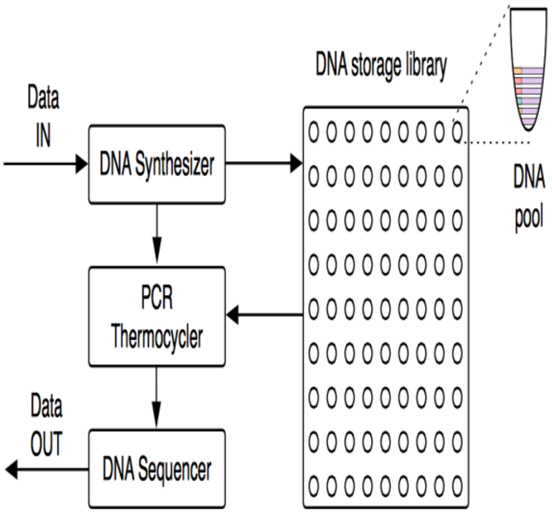


Figure.3 [7]Overview of a DNA Storage System

The basic unit of DNA storage is DNA strand that is roughly 100-200 nucleotides long, capable of storing 50-100 bits total. Therefore, a typical data object maps to a very large number of DNA strands. The DNA strands will be stored in "pools" that have stochastic spatial organization and do not permit structured addressing, unlike electronic storage media. Therefore, it is necessary to embed the address itself into the data stored in a strand.This is how after sequencing, one can reassemble the original data value.

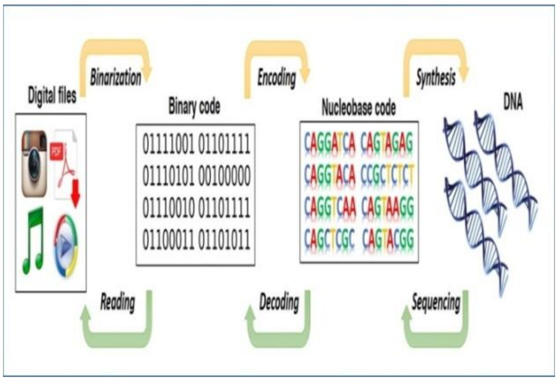
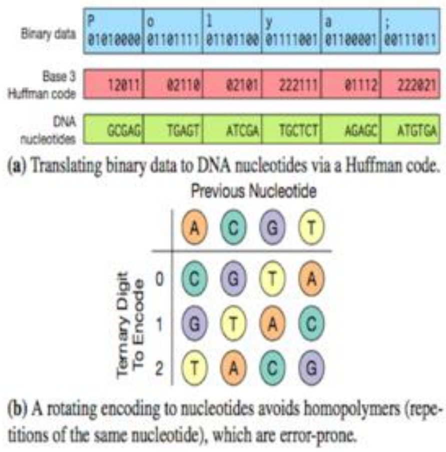


Figure .4[8] Read and Write operation process in DNA data storage

V. DATA REPRESENTATION IN DNA[1]

DNA has many properties that differentiates it from the traditional storage devices. At lowest levels , storage media stores raw bits. The abstraction of DNA is the nucleotide, though a nucleotide is an organic molecule consisting of one base (A,C,G or T) and a sugar phosphate backbone, the abstraction of DNA storage is as a contiguous string of quaternary (base 4) numerals. This section describes the challenges of representing data in DNA. Base 3 is not a multiple of base 2, mapping directly between the bases is inefficient. Instead, Huffman code is used, as it maps each binary byte to either 5 or 6 ternary digits.The rotating nucleotide encoding maps strings to the DNA sequence.The code maps more common ASCII characters to 5 digit strings,offering minor compression benefits for textual data,though the effect an overall storage density is insignificant. The obvious approach to store binary data in DNA is to encode the binary data in base 4 ,producing a string of n/2 quaternary digits from a string of n binary bits. The quaternary digits can then be mapped to DNA nucleotides.However ,DNA synthesis and sequencing processes are prone to a wide variety of errors,requiring a more careful encoding. The likelihood of some forms of error can be reduced by encoding binary data in base 3 instead of base 4 as figure 5(a).Each ternary digit maps to a DNA nucleotide based on a rotating code in Fig.5(b) that avoids repeating the same nucleotide twice.This encoding avoids homopolymers-repitions of the same nucleotide that significantly increase the chance of sequencing errors.



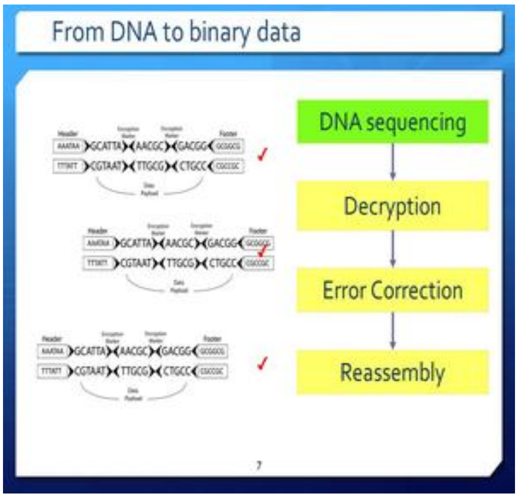
Binary data	P 01010000	o 01101111	l 01101100	y 01111001	a 01100001	; 00111011
Base 3 Huffman code	12011	02110	02101	222111	01112	222021
DNA nucleotides	GCGAG	TGAGT	ATCGA	TGCTCT	AGAGC	ATGTGA

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

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

(b) A rotating encoding to nucleotides avoids homopolymers (repetitions of the same nucleotide), which are error-prone.

Figure 5[9] .Encoding binary data as nucleotides



From DNA to binary data

DNA sequencing

Decryption

Error Correction

Reassembly

Figure 6[10] DNA to binary data

VI. ADDRESSING METHODS IN DNA :[2]

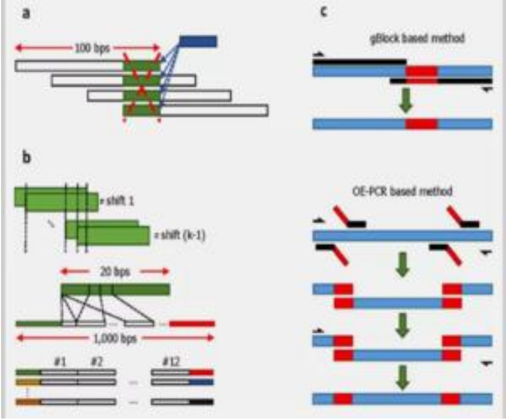
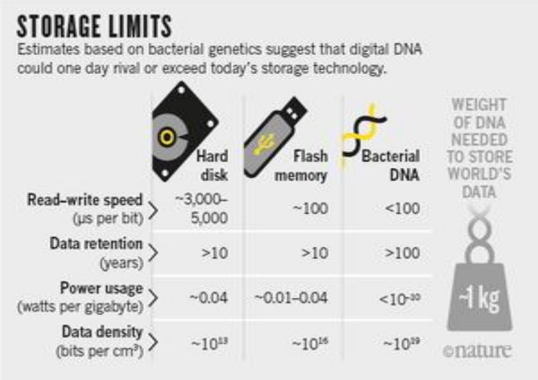


Figure 7[11] Addressing methods in DNA

(**a**) The scheme of uses a storage format consisting of DNA strings that cover the encoded compressed text in fragments of length of 100 bps. The fragments overlap in 75 bps, thereby providing 4-fold coverage for all except the flanking end bases. This particular fragmenting procedure prevents efficient file editing: If one were to rewrite the "shaded" block, all four fragments containing this block would need to be selected and rewritten at different positions to record the new "shaded" block. (**b**) The address sequence construction process we propose which uses the notions of autocorrelation and cross-correlation of sequences. A sequence is uncorrelated with itself if no proper prefix of the sequence is also a suffix of the same sequence. Alternatively, no shift of the sequence overlaps with the sequence itself. Similarly, two different sequences are uncorrelated if no prefix of one sequence matches a suffix of the other. Addresses are chosen to be mutually uncorrelated, and each 1000 bps block is flanked by an address of length 20 on the left and by another address of length 20 on the right (colored ends). (**c**) Content rewriting via DNA editing: the gBlock method for short rewrites, and the cost efficient OE-PCR (Overlap Extension PCR) method for sequential rewriting of longer blocks.

VII. DEVELOPMENTS AND FUTURE :



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 (µs 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	$<10^{-10}$
Data density (bits per cm³)	$\sim10^{13}$	$\sim10^{16}$	$\sim10^{19}$

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

Figure.8[12] Comparison of different Storage devices

Data density and durability of DNA storage seems to be more efficient than other .Microsoft has roughly written 200 megabytes of data in a physical space which is smaller than the tip of a pencil.

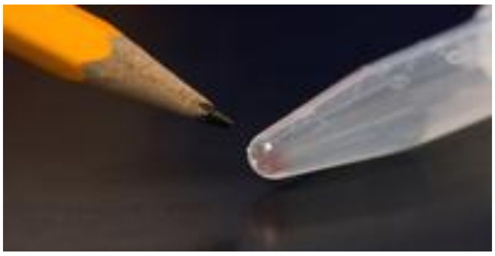


Figure.9[13] a new record for DNA data storage, cramming 200 MB of text and video

VIII. CONCLUSION

DNA –based storage has the potential to be the ultimate archival storage solution. DNA synthesis and DNA sequencing ,both technologies are growing at an alarming rate. Biotechnology has benefitted tremendously from progress in silicon technology developed by the computer industry. Perhaps,now is the time for the computer industry to borrow back from the biotechnology industry to advance the state of the art in computer systems.

IX. ACKNOWLEDGEMENT

I am very grateful to Mr.Mahesh Malkani Asstt. Professor, for his support to write this paper. I am very thankful to Mrs. Neetu Sharma, the Head of Department of Computer science in Ganga institute of technology and management for his motivation and support during the paper.

REFERENCES

[1] Data Representation in DNA https://homes.cs.washington.edu/~bornholt/papers/dnastorage-asplos16.pdf

[2] Addressing Methods in DNA http://www.nature.com/articles/srep14138

[3] Carlson Curve https://en.wikipedia.org/wiki/Carlson_Curve

[4] Figure.1Density of DNA Storage http://www.slideshare.net/DeevenaDayaal/dna-computing-41079997

[5] Figure.2 Synthesis and Sequencing https://www.extremetech.com/extreme/199414-scientists-create-million-year-data-storage-with-dna

[6] Graph 1 Trends in DNA synthesis and DNA sequencing http://blog.dshr.org/2016/09/natures-dna-storage-clickbait.html

[7] Figure.3Overview of a DNA Storage System http://homes.cs.washington.edu/~bornholt/dnastorage-asplos16/

[8] Figure .4 Read and Write operation process in DNA data storage https://www.assignmentexpert.com/blog/how-can-dna-be-used-in-data-storage-and-computing/

[9] Figure 5 Encoding binary data as nucleotides. http://homes.cs.washington.edu/~bornholt/dnastorage-asplos16/

[10] Figure.6 DNA to binary data http://slideplayer.com/slide/10930178/

[11] Figure.7 Addressing Methods in DNA http://www.nature.com/articles/srep14138

[12] Figure.8 Comparison of different Storage devices http://www.nature.com/news/how-dna-could-store-all-the-world-s-data-1.20496

[13] Figure.9 a new record for DNA data storage, cramming 200 MB of text and video https://www.technologyreview.com/s/601851/microsoft-reports-a-big-leap-forward-for-dna-data-storage/