A Review: DNA Data Storage

Divesh Panwar
Specialist Programmer, Infosys Ltd. Pune, India

Kamaljeet
Biotechnologist, N.I.B, Delhi, India

***Abstract*—Technologies are emerging day by day, and hence the demand for efficient and large capacity storage devices. According to a survey done by IDC in Framingham, Massachusetts humanity is generating data at an exponential rate, which will be beyond the storage capacity of the mediums we are currently using. Thus, an efficient and durable medium for data archival and retrieval is required to manage this exponentially growing data. DNA can be the possible solution. DNA possesses the attributes like extreme density, with a raw limit of 1 extra-byte/mm3, half life span of over 500 years and successive code of 0's and 1's similar to that used in computers. This paper reviews the DNA data storage process and the encoding techniques used in the field of DNA data storage. Current developments and research work done in this field have also been reviewed in this paper.**

Keywords— Deoxyribonucleic Acid, Gigabyte, International Data Corporation.

I. INTRODUCTION

With the digitization of the social orders and worldwide uprisings in the advances, new issues have developed. Putting away the information produced at the worldwide level is the trouble that is begging to be addressed. IDC predicts this enormous information development to outperform 44 trillion gigabytes by the year 2020 [1]. This exponential information development rate effortlessly surpasses our capacity to store it. As far as possible and additionally the lifetime of the capacity medium is another issue that is asking to be tended to. Prior attractive tapes, optical drives and hard disks have been utilized to store information. As of late, the vast majority of the cell phone, PC makers have changed to blaze drives [2]. Hypothetically hard disks have a life expectancy of around 3-5 years while streak drives have a life expectancy of 5 -10 years [3]. There are different issues with the present mediums like they dirty condition and they are restricted assets and, in this way, will end one day. The information stockpiling firms over the globe are cooperating to locate a tough stockpiling medium with high thickness. Apparently, DNA is the best fit, as it is the densest stockpiling medium known till date that can likewise store data. Hypothetically DNA has a thickness of 1 extra Byte/mm3 and has a watched half life expectancy of more than 500 years even under brutal ecological conditions [4].

The natural capacity of DNA and the present methods for information stockpiling are analogs. DNA speaks to qualities and proteins that make a living thing as set of four nucleotides (the fundamental building piece of DNA), like a PC that procedures information as a string of two parallel numbers (0 and 1). The compose procedure of DNA stockpiling incorporates mapping of computerized information into nucleotide arrangements, trailed by integrating a DNA particle and afterward putting away it. Consequently, we can encode 455 extra Bytes of information for each gram of DNA. Also, DNA can be reproduced using polymerase chain reaction methodology. While working with the DNA the power utilization is more effective than cutting edge PCs. DNA is sturdier as it can't be hurt effectively. The whole grouping never gets harmed amid denaturation. In this way, the first succession can be recovered by opening up the rest of the grouping. These points of interest have filled in as a base for examines in this field in the current years.

II. BACKGROUND OF DNA MANIPULATION

A. *DNA Fundamentals*

The natural capacity of DNA and the present methods for information stockpiling are analogs. DNA speaks to qualities and proteins that make a living thing as set of four nucleotides (the fundamental building piece of DNA), like a PC that procedures information as a string of two parallel numbers (0 and 1). The compose procedure of DNA stockpiling incorporates mapping of computerized information into nucleotide arrangements, trailed by integrating a DNA particle and afterward putting away it. Consequently, we can encode 455 extra Bytes of information for each gram of DNA. Besides, DNA can be reproduced utilizing polymerase chain response strategy. While working with the DNA the power utilization is more effective than cutting edge PCs. DNA is sturdier as it can't be hurt effectively. The whole grouping never gets harmed amid denaturation. In this way, the first succession can be recovered by opening up the rest of the grouping. These points of interest have filled in as a base for examines in this field in the current years.

B. *Polymerase Chain Reaction (PCR)*

Created in 1983 PCR is a system used to make duplicates of a specific area of a DNA in vitro [5]. PCR plans to make enough of the objective DNA area so it can be investigated or used in some other way. PCR has numerous applications in the regions of sub-atomic science look into, therapeutic diagnostics, and even some branches of nature. PCR requires a DNA polymerase protein that makes new strands of DNA, utilizing existing strands as layouts [6][7]. The DNA polymerase regularly utilized as a part of PCR is called Taq polymerase. The capacity of warmth soundness makes Taq polymerase perfect for PCR.

PCR requires a groundwork to create a DNA. An introduction (short arrangement of nucleotides) gives an underlying point to DNA blend [8]. PCR response utilizes two groundworks such that the preliminaries flank the objective district. The PCR utilizes Taq polymerase, preliminaries, format DNA, and nucleotides as key fixings, which are then gathered in a tube, alongside cofactors required by the chemical. They are then subjected to rehashed cycles of warming and cooling, permitting DNA to be combined.

C. DNA Synthesis

It is a procedure by which strands of nucleic acids are made. In a cell, DNA union happens in a procedure called replication. The capacity that a nucleotide ties to a current fractional strand at each progression of the procedure chooses the coupling proficiency of the blend. For all intents and purposes, combination is done utilizing countless begin locales which brings about many truncated side effects, alongside many duplicates of the full-length target grouping. In this way, a given combination bunch will create many impeccable strands, and blunders in blending a particular strand can be overlooked [25]. Current cluster blend methods are fit for integrating complex pools of about 105 distinctive oligonucleotides parallelly.

D. DNA Sequencing

It is the method of defining the arrangement of nucleotide bases in a part of DNA. The DNA sample is combined in a tube with primer, DNA polymerase, and DNA nucleotides. The four dye-labeled, chain-terminating dideoxy nucleotides are added in very less quantities. The mixture is denatured first then cooled so that the primer can bind to the single-stranded template. Once the primer bounds, the temperature is raised so that DNA polymerase synthesize new DNA starting from the primer [9]. DNA polymerase will stop adding nucleotides to the chain when it adds a dideoxy nucleotide instead of a normal one. Thus, every strand ends with a dideoxy nucleotide. This process is repeated many times. The next generation sequencing techniques are now available which uses the advance concepts of parallelism and scaling. This not only improved the speed of sequencing but also reduced the cost involved. The cost of sequencing has reduced from $100 million in 2001 to $1245 in 2015 [10].

III. Earlier Approaches in DNA storage

The first indirect implementation of Digital Data storage was in 1999 by Clell and, Risca, Bancroft [11] and [12]. They used the DNA strands (Microdots) to store encoded data. Secret data was communicated during World War II using this technique. The scientists tried to skin valued information in a DNA by means of the 4 bases, PCR Primers and an Encryption key (base triplets to denote English letters and Arabic Numbers). The structurally complex denatured human genomic DNA was used to hide the desired DNA-encoded message. It was then scaled down to a microdot. The PCR sequences served as a base DNA storing the data, each of which was surrounded concealing human DNA, 109 times the size of PCR sequences. This in turn provided privacy and security to the data by providing a complex background. The key concept was that if the PCR arrangements and the specific encryption key are recognized only then the data can be retrieved by the process of electrophoresis irrespective of the concealing DNA mask used [24]. This discovery also made it clear that noises or unwanted sequences cannot affect the encoding process. This research also resolved that DNA data storage is much more isolated and safe than Digital Storage devices.

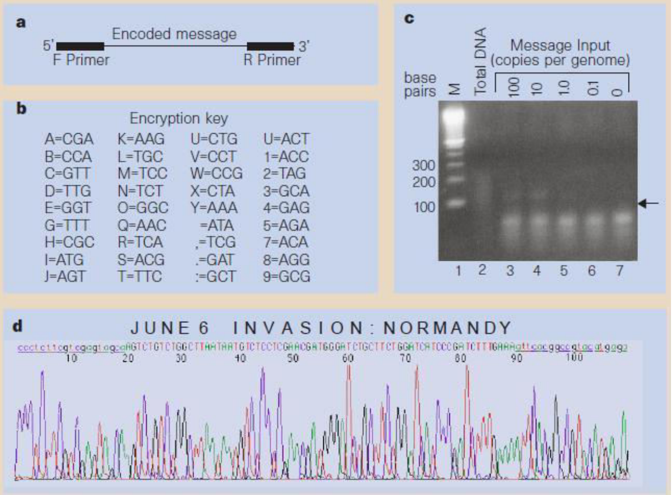


Fig. 1. The research work and coding scheme summarization

The following Fig. 1 illustrates the research code stored and coding scheme used. Bancroft, Bowler, Bloom, and Clelland [13] worked on the same idea in 2001. They used Poly Primer Key (The primer base sequence to access the information on the iDNA) and iDNA (information DNA – the encoded data) to develop an experimentation which includes amplification by PCR. It generated sufficient 'universal' forward and reverse primers to analyze them consuming the obvious encryption scheme. Requested parts of iDNA were produced from each straightly broke down sequencing groundwork. Thus, the data can be decoded from these linearly arranged primer sequences, using the designed Encryption Scheme. Thus, this research proved that development of such storage devices is possible using the simple concepts of biology. The Fig. 2 illustrates the assembly of DNA molecules used for information storage and reading. The next research [13] marked its appearance in the year 2003, done by Wong et al on-Data Storage in DNA. This research pointed out that strands of DNA can break at both ends and can lead to loss of information. The need of an orchestrated quality successions was subsequently required to shield and save the encoded DNA strands from unforgiving conditions. A vector was in this way required which contains the DNA alongside the information and can develop and duplicate to guarantee perpetual quality of data. The researchers used Escherichia coli and Deinococcus radiodurans as vectors due to their ability of quick renewal and tolerance to vacuum, and radiations. They used the oligonucleotide sequences to prepare a digital arrangement similar to the 1's and 0's that forms the ASCII scheme for representing text in silicon-based devices.

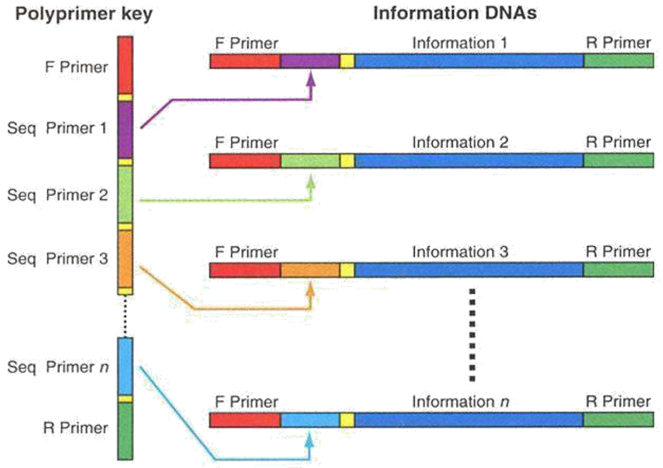
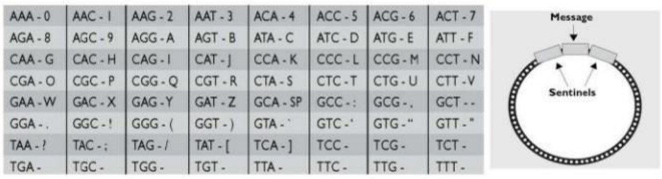


Fig. 2. DNA molecule structure used for data storage

They also managed to find out 25 safe sequences out of 10 billion sequences that were immune to bacteria and could not harm the data. At that point, they cloned the twofold strands into a recombinant plasmid by making a complimentary strand and utilizing limitation proteins (used to embed the coveted encoded sections). Data recovery is done utilizing the PCR procedure. This way they were able to acquire 7 chemically synthesized DNA fragments with 57–99 base pairs (bp) of foreign encoded information in the bacteria . This research represented the procedures for defense of the desired data from extremities in environment, radiation, vacuum, nucleases otherwise harmful for the fragile DNA fragments. The analysis finished up by giving a DNA can be utilized for authentic science. Fig. 3 illustrates the encoding scheme and idea of protecting the sequence.



AAA - 0	AAC - I	AAG - 2	AAT - 3	ACA - 4	ACC - 5	ACG - 6	ACT - 7
AGA - 8	AGC - 9	AGG - A	AGT - B	ATA - C	ATC - D	ATG - E	ATT - F
CAA - G	CAC - H	CAG - I	CAT - J	CCA - K	CCC - L	CCG - M	CCT - N
CGA - O	CGC - P	CGG - Q	CGT - R	CTA - S	CTC - T	CTG - U	CTT -V
GAA - W	GAC - X	GAG - Y	GAT - Z	GCA - SP	GCC - :	GCG - ,	GCT - -
GGA - .	GGC - !	GGG - (	GGT -)	GTA - `	GTC - '	GTG - "	GTT - "
TAA - ?	TAC - ;	TAG - /	TAT - [	TCA -]	TCC -	TCG -	TCT -
TGA -	TGC -	TGG -	TGT -	TTA -	TTC -	TTG -	TTT -

Message
Sentinels

Fig. 3. Encoding Scheme

IV. DNA Storage Encoding

Till now we examined how information can be decomposed into strands of DNA. An encoding system is utilized to store information in DNA. Notwithstanding, the procedure of DNA combination and sequencing are inclined to an extensive variety of mistakes (substitutions, additions, and erasures of nucleotides), and in this way, require a more cautious encoding. In this area, some current encodings utilized for DNA information stockpiling have been talked about.

E. *Base 4 Encoding*

The undeniable way to deal with storing parallel information in DNA is to encode the double information in base 4, creating a string of n/2 quaternary digits from a string of n paired bits. Parallels may be drawn between the way request is diagrammatic by deoxyribonucleic acid and quaternary numerals. The four DNA nucleotides abridged A, C, G and T, might be wont to speak to the quaternary digits in numerical request 0, 1, 2, and 3. This encoding permits the corresponding digit sets 0↔3, and 1↔2 (twofold 00↔11 and 01↔10) coordinate the complementation of the base sets: A↔T and C↔G and might be kept as information in DNA sequences.

F. *Goldman Encoding*

Goldman et al. proposed this encoding scheme [14], appeared in Figure 4. This scheme parts the information DNA nucleotides into covering fragments to give fourfold repetition to each portion. Every window of four sections compares to a strand in the yield encoding. The creators utilized this scheme to effectively recoup a 739kB message. This encoding is utilized as a benchmark since it seems to be, as far as anyone is concerned, the most effective distributed DNA stockpiling system. What's more, it offers a tunable level of excess, by decreasing the width of the sections and along these lines rehashing them all the more frequently in strands of a similar length (for instance, if the covering fragments were half the length of in Figure 4, they would be rehashed in eight strands rather than four).

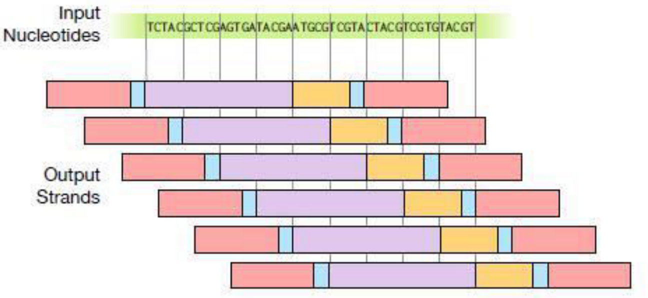


Fig. 4. An encoding proposed by Goldman et al. [10]. The payloads of each strand are overlapping segments of the input stream.

G. *XOR Encoding*

While the Goldman encoding gives high unwavering quality, it moreover brings about critical overhead: each piece in the information string is rehashed four circumstances. This encoding provides comparative stages of excess to earlier work, yet with lessened overhead. This encoding, appeared in Figure 5, gives repetition by a basic selective or operation at the strand levels. It takes the exclusive or A⊕B of the payloads A and B of two strands, which delivers another payload thus

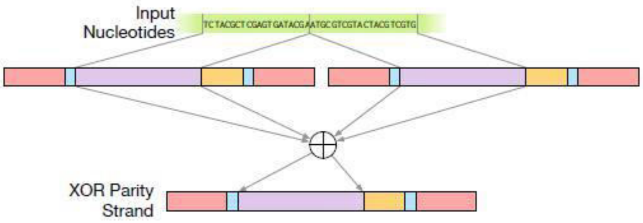


Fig. 5. This encoding incorporates redundancy by taking the exclusive-or of two payloads to form a third. Recovering any two of the three strands is sufficient to recover the third

another DNA strand. The address square of the new strand encodes the locations of the info strands that were the contributions to the selective or; the high piece of the deliver is utilized to show whether a strand is a unique payload or a select or strand. This encoding gives its repetition in a comparable manner to RAID 5: any two of the three strands A, B, and A⊕B are adequate to recuperate the third. The dependability of this encoding is like that of Goldman. The hypothetical thickness of this encoding is considerably higher than Goldman —where in their encoding every nucleotide rehash (up to) four circumstances, in this every nucleotide rehashes an normal of 1.5 circumstances.

V. Related Works

Microsoft and University of Washington researchers came together and researched on storing data on DNA molecules, they managed to store 200MB of data [15]. This was a remarkable milestone for team of researchers led by Karin Strauss as this was the highest amount of data stored and recovered without any error. After encoding it occupied as much as a spot in the test tube [16]. In 1999, the state-of-the-art in DNA-based storage could encode and recuperate a 23 - character message [17]. Later in the year 2000 Leier et al managed to recover three 9-bit numbers [18]. The first significant improvement was observed in the year 2010 when Gibson et al used bacterial genome to successfully recover 1280 encoded characters. But this technique rendered useless when it comes to large scale data as it was employed inside vivo. Further improvements in data storage have been noticed in the year 2012 when Church et al recovered 643 kB message, and another researchers Goldman et al recovered a 739-kB message [18]. However, Chapel et al. needed on manually right ten odds from claiming error, also Goldman et al. lost two successions of 25 nucleotides. Grass et al. recouped an 83kB message without lapse [19]. Their plan employs a Reed-Solomon code [20], striping those whole datasets over 5000 DNA strands. At the same time this plan prompts fantastic redundancy, it defeats those longing for irregular get. In 2013, a team at the European Bioinformatics Institute (EBI) in Hinxton, Britain, claimed to create the largest DNA archive, storing computer files of worth 739 into DNA strands [21]. After the fact to July 2016, a group from claiming Microsoft Furthermore college of Washington analysts guaranteed that they needed reset the previous record, storing 200 megabytes from claiming information clinched alongside DNA [22]. Specialists are presently asserting that they can achieve a density of 215 Petabytes for every gram of DNA [21]. As of late, PC researcher Yaniv Erlich at Columbia University and bioinformatics analyst Dina Zielinski at the Genome Center exploited DNA's normal property to quickly improve itself and recombine in complex however unsurprising examples [23]. They encoded a duplicate of the Kolibri PC working system, an 1895 French novel into a film called "Arrival of a prepare during la Ciotat," a gift card, a 1948 investigation by majority of the data scholar Claude Shannon, and a picture of the plaque conveyed of the edge of the earth's planetary group Eventually Tom's perusing the Pioneer 10 and 11 space probes.

VI. Conclusion

DNA-based capacity can possibly be a definitive documented stockpiling arrangement: it is amazingly dense and durable, stable and vitality proficient. Taken a toll effectiveness likewise indicates guarantee as the exponential drop in DNA amalgamation and sequencing cost has been five-overlay and twelve-overlap separately while electronic media demonstrate just a 1.6-crease drop every year. Considering the approaching furthest reaches of silicon innovation, we trust that cross breed silicon and biochemical frameworks are worth genuine thought: time is ready for PC modelers to consider joining biomolecules as an indispensable piece of PC plan. A few achievements will be required before it turns out to be industrially standard for information recovery. This field has had a million-overlap change in the current years. Computerized Data Storage in DNA innovation demonstrates tremendous advance, since perusing and composing it is propelling ten times each year not at all like the Electronic Technology which is enhancing around 1.5 times each year (Moore's Law). This review paper examines the need and future extent of DNA data storage. Information is initially changed over into hereditary code and after that take this humble content record and physically build the particle it speaks to. Continuing further by dissecting different strategies for encoding information, and evaluating the work done in this field. Consequently, DNA stockpiling is a potential possibility to be considered for putting away the im mense measure of information produced by mankind later on. We are as yet not certain when DNA stockpiling innovation will be accessible as essential stockpiling to the normal open yet considering the mechanical advances in the data science and Biotechnology it is not adequately far.

Acknowledgment

I would like to express my sincere gratitude to my mentor, Er. Akshay Bhardwaj, and my colleague Sunidhi Raju for the guidance and support extended to drive me on the correct path to write this review paper.

References

[1] Pushing the Theoretical Limits of DNA Data Storage http://blogs.discovermagazine.com/d-brief/2017/03/02/dna - data-storage-limits/

[2] IDC. Where in the world is storage? http://www.idc.com/downloads/where_is_storage_infographic _243338.pdf, 2013.

[3] Data storage lifespans: How long will media really last http://www.storagecraft.com/blog/data-storage-lifespan/

[4] 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. Proceedings of the Royal Society of London B: Biological Sciences, 279(1748) pp. 4724–4733, 2012.

[5] What is a DNA? https://ghr.nlm.nih.gov/primer/basics/dna

[6] What is PCR (polymerase chain reaction)? http://www.yourgenome.org/facts/what-is-pcr-polymerase-chain-reaction

[7] Polymerase chain reaction (PCR) https://www.khanacademy.org/science/biology/biotech - dna - technology/dna-sequencing-pcr-electrophoresis/a/polymerase-chain-reaction-pcr

[8] James Bornholt, Randolph Lopez, Douglas M. Carmean, Luis Ceze, Georg Seelig, Karin Strauss, “A DNA-Based Archival Storage System”, 21th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Atlanta, GA, April 2–6, 2016.

[9] "DNA structure and sequencing," by OpenStax College, Biology (CC BY 4.0) http://cnx.org/contents/185cbf87-c72e-48f5-b51e-f14f21b5eabd@10.4.

[10] S. Kosuri and G. M. Church. Large-scale de novo DNA synthesis: technologies and applications. Nature Methods, 11 pp. 499–507, 2014.

[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, “Long-term 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] Q. Guo, K. Strauss, L. Ceze, and H. Malvar. High-density ima ge storage using approximate memory cells. In ASPLOS, 2016.

[15] MIT Technology Review https://www.technologyreview.com/s/601851/microsoft-reports-a-big-leap-forward-for-dna-data-storage/

[16] Microsoft and University of Washington researchers set record for DNA storage https://blogs.microsoft.com/next/2016/07/07/microsoft-university-washington-researchers-set-record-dna-storage/

[17] A DNA Based archival storage system http://homes.cs.washington.edu/~bornholt/dnastorage-asplos16/

[18] Goldman, N. et al. 2013. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature. 494, (2013), pp. 77–80.

[19] Grass, R.N. et al. 2015. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew. Chem. Int. Ed. 54, (2015), pp. 2552–2555.

[20] Reed, I.S. and Solomon, G. 1960. Polynomial codes over certain finite fields. Journal of the Society for Industrial and Applied Mathematics. 8, 2 (1960), pp. 300–304.

[21] To make better computers, researchers turn to molecular biology http://www.csmonitor.com/Science/2017/0302/To-make-better-computers-researchers-turn-to-molecular-biology

[22] New Study Confirms That the Future of Data Storage Is in DNA https://futurism.com/4-the-future-of-data-storage-is-in-the-dna-new-study-confirms/

[23] What’s Stored in DNA? An Old French Movie and a $50 Gift Card https://www.wsj.com/articles/whats-stored-in-dna -an-old-french-movie-and-a-50-gift-card-1488481266

[24] Siddhant Shrivastava, Rohan Badlani, “Data Storage in DNA”, International Journal of Electrical Energy, Vol. 2, No. 2, June 2014.

[25] Sourav Chatterjee, C S Kshyatisekhar Panda, “D.N.A as a Data Storage Medium”, IJSRD - International Journal for Scientific Research & Development, Vol. 4, Issue 08, 2016, ISSN (online): 2321-0613.