DATA STORAGE IN THE DNA

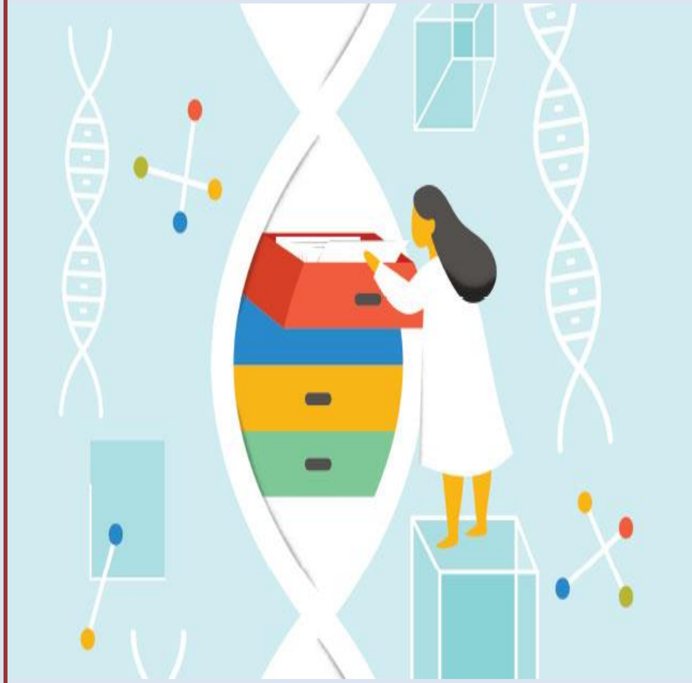
AND

PREVENTING DATA LOSS

Research under: Siva Rama Krishnan S

Slot: F1+TF1

Course code:



By

18BIT0094-SUCHETAN

18BIT0126-ROHITH

18BIT0137-RITHVIK

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

ABSTRACT

One of the major problems faced today in data storage is the data loss due to crashing of storage devices like magnetic disc or optical disc. Everyday lot of data is produced and this requires high density storage devices which can retain values for a long time. In recent years scientists have turned their attention towards the biomaterials for data storage. The growing demand of digital information storage worldwide has led to the development of technology of using DNA as novel storage media. A new type of storing digital information within the DNA of living organism is attracting more attentions. DNA is a suitable storage method due to its high data density, environment compatibility and long-term storage potential. However, despite the low cost and high-throughput advantages, this type of DNA storage also has shortcomings such as limited DNA quantity, difficulty replicating, etc. Pilot studies of DNA storage in representative living organisms such as E. coli, yeast and Arabidopsis, are conducted. This is achieved by the process called induced mutation which is controlled by electrical signals.

Data stored in DNA is more reliable than any other devices. This invention provides a hope for future data storage which will be safe for millions of years, as they are transmitted over various generations. We can cultivate the bacteria in which the data are stored in various places like fish tank, garden, stomach and plants and seeds also. Moreover, a naked DNA molecule can be greatly affected by environmental influences, thus resulting in DNA mutations and changes in the stored information. Research demonstrates the great potential of plants and seeds in circumventing these drawbacks. It shows that artificially encoded data can be stored and multiplied within plants.

There by in our project we are going to discuss about various research in this field and the scope of the research. And to provide the

information about the future storage tech that is the DNA storage technologies.

Objective

The main objective of this project is to find out how DNA acts as a storage medium and to know what are the advantages and disadvantages of doing so. This research paper also deals with the evolution of DNA storage throughout the time and what is the current technique being used for it.

INTRODUCTION

Human civilization went through paradigm shifts with new ways of storing and disseminating information. To survive in the complex and ever-changing environment, our ancestors created utensils out of wood, bone and stone, and used them as media for recording information. This was the beginning of human history. With the development of computer technology, the information age has revolutionized the global scene. Digital information stored in magnetic (floppy disks), optical (CDs) and electronic media (USB sticks) and transmitted through the internet promoted the explosion of next-generation science, technology and arts.

With the total amount of worldwide data skyrocketing, traditional storage methods face daunting challenges. International Data Corporation forecasts that the global data storage demand will grow to 175 ZB or 1.75 × 1014 GB by 2025 (in this review, ‘B’ refers to Byte while ‘b’ refers to base pair). With the current storage media having a maximal density of 103 GB/mm3, this will far exceed the storage capacity of any currently available storage method. Meanwhile, the costs of maintaining and transferring data, as well as limited lifespans and significant data losses, also call for novel solutions for information storage.

On the other hand, since the very beginning of life on Earth, nature has solved this problem in its own way: it stores the information that defines the organism in unique orders of four bases (A, T, C, G) located in tiny molecules called deoxyribonucleic acid (DNA), and this way of storing information has continued for 3 billion years. DNA molecules as information carriers have many advantages over traditional storage media. Its high storage density, potentially low maintenance cost and other excellent characteristics make it an ideal alternative for information storage, and it is expected to provide wide practicality in the future.

The Steps in Using Synthetic DNA for Storage:

- Convert your binary code into nucleotide code.
- Separate your codes into bits and insert an address code for each fragment (including spacers for amplification).
- Synthesize your DNA as short fragments of oligonucleotides.
- Store in the freezer.
- When you need the archive, PCR amplify it and sequence it.
- Analyze the sequence data and re-assemble it.
- Finally, turn it back to binary code!

LITERATURE SURVEY

Research paper	Summary
[1] Cox, J. P. (2001). Long-term data storage in DNA. TRENDS in Biotechnology, 19(7), 247-250. https://www.researchgate.net/publication/11927065_Long-term_data_storage_in_DNA	Discussion about how DNA could be an excellent method for data storage and also addresses two major problems associated with it. The first problem is retrieval of data from the genome is a very tedious process although that can be expected to improve by some methodologies. Another problem being the cost factor since such technology could be very attracting and thus cost

	goes up and the cost factor becomes very high. Challenges: here the challenges are 1. Retrieval of data from the genome as it is a tedious process 2. And this DNA storage is very costly. 3. Encoding techniques takes a long time
[2] Bornholt, J., Lopez, R., Carmean, D. M., Ceze, L., Seelig, G., & Strauss, K. (2016). A DNA-based archival storage system. ACM SIGARCH Computer Architecture News, 44(2), 637-649. https://www.microsoft.com/en-us/research/publication/dna-based-archival-storage-system/	Examination about the exponentially increasing demand for data storage and the unsatisfying capacity of existing storage media. To keep up with the demand for the storage, using DNA to archive data is an attractive possibility. In this paper, they presented an architecture for a DNA-based archival storage system. The system is designed as a key-value store leveraging common biochemical techniques to provide random access. Challenges: here the challenges are 1. Random access is not east to achieve 2. Laboratories are limited for research on DNA storage 3. Access to data as and then in small intervals is not widely appreciated as it takes lot of time and cost
[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	the analogy between the digital data and genetic data. Digital information is stored as binary digits (0's and 1's) and the genetic data is stored in the form of molecular polymers consisting four bases adenine (A), cytosine (C), guanine (G), and

characters using degenerate bases. Scientific reports, 9(1), 6582. https://www.researchgate.net/publication/326329872_Addition_of_Degenerate_Bases_to_DNA-based_Data_Storage_for_Increased_Information_Capacity	thymine (T).Each pair corresponding to a bit of information. Challenges: here the challenges are 1. Not everyone can understand how DNA support stores the data and therefore more skilled workers are needed in this particular field. 2. People with sufficient knowledge to propose encoding techniques are limited and more researches are needed in this domain to develop rapidly and instantly. 3. Not every organization can afford the research on this because of lack of infrastructure and this is not cost effective.
[4] Yazdi, S. H. T., Yuan, Y., Ma, J., Zhao, H., & Milenkovic, O. (2015). A rewritable, random-access DNA-based storage system. Scientific reports, 5, 14138. https://ieeexplore.ieee.org/document/9151948?denied=	described the first DNA-based storage architecture that enables random access to data blocks and rewriting of information stored at arbitrary locations within the blocks. Their system is based on new constrained coding techniques with DNA editing methods that ensure data reliability, specificity, and sensitivity of access, and at the same time provide exceptionally high data storage capacity. Challenges: here the challenges are 1. Here in the architecture we don’t discuss about the encoding techniques supported by the system 2. As we see our computer the hashing techniques followed are of various

	type like that the encoding techniques are also of various types based on the encoding technique the system can be made in low cost by cost cutting factors in the encoding and decoding fields 3. And speed of data writing and retrieval is not discussed 4. No discussion on the data security and integrity
[5] Richard P. Feynman- There's Plenty of Room at the Bottom https://www.researchgate.net/publication/291938281_There's_Plenty_of_Room_at_the_Bottom_An_Invitation_to_Enter_a_New_Field_of_Physics	Richard P. Feynman was the first one to coin the term DNA storage in 1959. Feynman considered the possibility of direct manipulation of individual atoms as a more powerful form of synthetic chemistry than those used at the time. Challenges: here the challenges are 1. During that time the resources are limited not much information can be gathered form this paper but there is a theoretical proof that DNA can be a next generation storage technology. 2. Information provided only about that there is possibility of data storage in the DNA and no sufficient proofs are derived from this paper. But based on the discussion we can get to know that the DNA can also be used as a storage medium.
[6] Pavani Yashodha De Silva and Gamage Upeksha	shows how DNA emerges as the prospective medium for data storage with

Ganegoda Faculty of Information Technology, University of Moratuwa, Katubedda, Moratuwa, Sri Lanka. https://www.researchgate.net/publication/307871084_New_Trends_of_Digital_Data_Storage_in_DNA	its striking features. Diverse encoding models for reading and writing data onto DNA, codes for encrypting data which addresses issues of error generation, and approaches for developing codons and storage styles have been developed over the recent past. DNA has been identified as a potential medium for secret writing, which achieves the way towards DNA cryptography and stenography. DNA utilized as an organic memory device along with big data storage and analytics in DNA has paved the way towards DNA computing for solving computational problems. Challenges: here the challenges are 1. Data writing and retrieval speeds are not mentioned for the given models 2. Type of data that was encoded is not mentioned 3. No information about what happens to data after a long time of storage like the corruption rate, error rates are not mentioned
[7] Vishal Bhatnagar Shri Venkateshwara University. https://www.researchgate.net/publication/283024493_A_novel_DNA_sequence_dictionary_method_for_securing_data_in_DNA_using_spiral_approach_and_framework_of_DNA_cryptogr	Data storage is becoming a crucial issue in modern day age of computers. The storage technologies and medium are trying to cope with the demand generated in the market due to growing storage requirement with more advanced technological development. The storage is also governed with the need for secure and efficient storage for which the

aphy	technology people are striving. The use of the DNA technology in the various fields has created interest for the technology savvy to find and use it in storage medium. The use of the DNA for the storage medium is considered to be of utmost relevance in modern day era. Considering the above facts, the authors were motivated to carry out an extensive survey, classifying the literature dealing with the use of DNA as storage medium according to various topics by proposing a classification framework which identified the contribution of information security in securing the DNA as storage medium. Challenges: here the challenges are 1. Discussed about what is importance of the DNA storage but not mentioned about anything about data like • Data integrity • Data privacy • Data security • Reliability • Efficiency • Availability 2. The main part of DNA Storage is the DNA synthesis that had to be done for Storage purposes and in this we still follow old methods which involve lot of costly chemicals and time.so it is better to get the cost-

	effective DNA synthesizer.
[8] Molecular Digital Data Storage using DNA by Luis Ceze, Jeff Nivala and Karin Strauss. https://www.researchgate.net/publication/332997415_Molecular_digital_data_storage_using_DNA	DNA storage as Molecular data storage is an attractive alternative for dense and durable information storage, which is sorely needed to deal with the growing gap between information production and the ability to store data. DNA is a clear example of effective archival data storage in molecular form. In this Review, we provide an overview of the process, the state of the art in this area and challenges for mainstream adoption. We also survey the field of in vivo molecular memory systems that record and store information within the DNA of living cells, which, together with in vitro DNA data storage, lie at the growing intersection of computer systems and biotechnology. Challenges: here the challenges are 1. As the paper focus on archival system but it not provide the information about the encoding and decoding techniques that has to be used for greater efficiency. 2. Archival systems do not support the easy access of data as the data cannot be retrieved and written to storage. As this less interval access leads to the cost growth and effort to be done is high.
[9] A Potomac Institute for Policy Studies Report [9] Shimanovsky, B., Feng, J.,	Potomac Institute for Policy Studies described how the demand for digital data storage is currently outpacing the world's

& Potkonjak, M. (2002, October). Hiding data in DNA. In International Workshop on Information Hiding (pp. 373-386). Springer, Berlin, Heidelberg. https://potomacinstitute.org/reports/43-pips-reports/189-the-future-of-dna-data-storage	storage capabilities, and the gap is widening as the amount of digital data produced grows exponentially. Current technologies will be insufficient to address these challenges. DNA offers an abundant, sustainable, and stable data storage solution, with storage density orders of magnitude better than today's best methods. As a reference, all of the world's data today could theoretically be stored in 1 kg of DNA. Although the science behind DNA data storage has been proven, its commercial viability is currently limited. This is because DNA technology Executive Summary has been developed to support applications in the life sciences industry and not for data storage purposes. Challenges: here the challenges are 1. Only theoretical proofs have been provided rather than some mathematical proofs. 2. Discussion mainly on the how vast the DNA can be used for data Storage but no information about how the data is written to DNA. 3. No information on how to get the DNA for the Storage.

METHODOLOGIES

Microvenus Project

This project was initiated by Joe Davis to store an image in DNA, which had the core intention of storing abiotic data in DNA. Encoding was based on the molecular size of bases, C→1, T→2, A→3, and G→4. Each and every nucleotide was assigned a phase structure C → X, T → XX, A → XXX, G → XXXX. Encoding was achieved by placing a nucleotide at each repeated position of 1 and 0 bits, for instance, 100101 = CTCCT and 10101 = CCCCC. In the process of decoding, “C” could be decoded as “1” or a “0”, because only the number of repeated bits was taken into consideration at the time of encoding. For instance, CTCCT could be decoded as 011010 or 100101. Therefore, this scheme of encoding was inaccurate as it was not distinctively decodable.

Genesis Project

This was introduced by Eduardo Kac. He created an “artist’s gene,” a synthetic gene by converting a sentence from the bibliographical book “Genesis” into Morse code and then converting the Morse code into DNA base pairs. Hyphen and full stop were represented by bases T and C while replacing words pace and letters pace with A and G, respectively. Synthetic genes were fused into bacteria and in the presence of an ultraviolet light, mutations were caused in bacteria. These mutations have in turn caused changes in the sentence. When the genes were decoded back into the Morse code and then back into English, the original sentence has been changed . Genesis project was inaccurate as the original sentence was altered during mutation at the presence of ultraviolet light.

Current data storage technique:

The DNA in our cells contains the instructions for building all the proteins that keep us running. DNA is made up of repeating sequences of the nucleic acid’s adenine, guanine, cytosine, and thymine (A, G, C, and T) which are sometimes called base pairs. Each sequence of three bases translates to a different amino acid, which are the building blocks of proteins. It’s data storage just like what we do with hard drives but with much higher potential density. The four-lettered nucleobase alphabet of DNA (A, C, G and T) can be transformed into binary code for example, as 00 for A, 01 for C, 10 for G and 11 for T. Scientists looked at the algorithms that were being used to encode and decode the data and first converted the files into binary strings of 1s and 0s compressing them into one master file and then split the data into short strings of binary code. They devised an algorithm called a DNA Fountain which randomly packaged the strings into droplets, to which they added extra tags to put the file back together. They started with six files including a full computer operating system and a computer virus.

In all, the researchers generated a digital list of 72,000 DNA strands, each 200 bases long. They sent these as text files and later, the sequences were fed into a computer which translated the genetic code back into binary and used the tags to reassemble the six original files. The approach worked so well that the new files contained no errors and were also able to make a virtually unlimited number of error free copies of their files.

Some other methods or means listed below:

Encoding Model	Advantages	Disadvantages
Microvenus project	Laid the foundation for storing the abiotic information in DNA	Being inaccurate and not distinctively decodable
Genesis project	Laid the research work to explore the intricate relationship between biology, belief systems, information technology, dialogical interaction, ethics, and the internet	Inaccurate as the original sentence was altered during mutation at the presence of ultraviolet rays
PCR based encoding model	High security because of the size of the microdots and even if an adversary identifies the microdot it	• Insertion of errors in template region making it unmanageable to

	would be extremely difficult without the knowledge of the primer sequence	recover the encoded data • Need of the knowledge of primers • Widespread experimental obstacles and practical problems • Need of PCR • Data breakage that could occur in encoding and decoding due to errors of humans
Allignment based encoding model	Independence of polymerase chain reaction greater speed and lower cost of reading DNA data and lower cost of synthetic DNA positions of the data breakages that could be identified easily by the alignment results although they were not recoverable	• Multiplication of cassettes leads to redundant volumes • Parity effects cost a certain volume of data sequence • Data recovery rate is fragile and is proportional to data breakage which occurs through DNA deletion of long ranges • Sequencing of the entire genome is required to retrieve

		data • There is size limit of the cassette oligonucleotides being used to encode the message. If it increases a certain limit there is a possibility of it to appear by chance in host genome
Rewritable and random access based DNA storage system	• Random access to data blocks of DNA which promotes nonlinear access • Rewriting capability of information into random locations • Cross hybridization problems that are eliminated in this method by prohibiting redundancy of information • Being used to store frequently updated data which needs to memorize the editing history	High cost
Next generation digital	• Employment of one-bit representation for base	• Cost is unfeasible • Time for reading and writing onto DNA is

information storage model	• High scalability • High data storage density • Highly reliable • Each copy having the capability to correct the errors in the other copy as the errors are almost never coextensive	high
Encoding scheme for small text files	• High volume data storage density • Not needing ample context information for encoding purpose • Maximum efficiency of compression • Reducing cost factor	Have to proceed in implementing the biological protocols to insert the sequence in genome of bacteria.

Now when we go for what are the companies or organizations that are in this field finding their way to storage the data in the DNA.
Not many because this was an emerging field and as a result it's still in experimental stage and many R&D labs are needed and more cost to establish the labs and conduct research. So not many organizations are there but form that few the top organizations are as below:

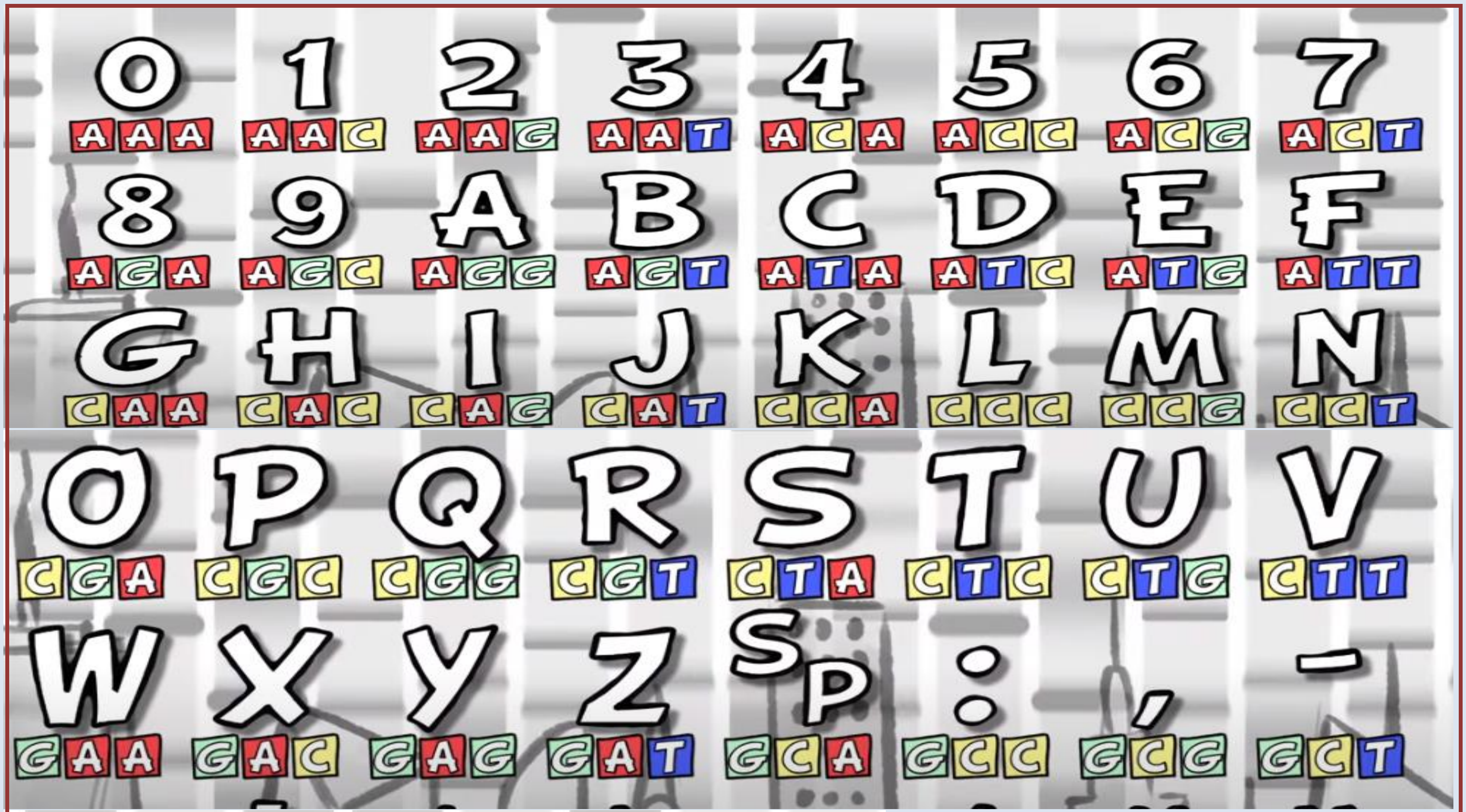
Sector	Organization	Primary effort
Private	Microsoft	R&D with the eventual goal of a proto-commercial DNA data storage system
Private	Semiconductor	R&D in advanced data storage

	Research Corporation	solutions
Private	Catalog	Commercialization of DNA data storage technology
Private	Iridia	Commercialization of DNA data storage technology
Public	NSF, IARPA, DARPA, NIH	Funding support to key players in the DNA data storage field
Academic	University of Washington	Research that is pushing towards increasing the volume of information stored in DNA
Academic	Harvard University	R&D of DNA synthesis technology and novel mechanisms of encoding and retrieving information from DNA
Academic	ETH ZURICH	Research on storing varying types of files in DNA

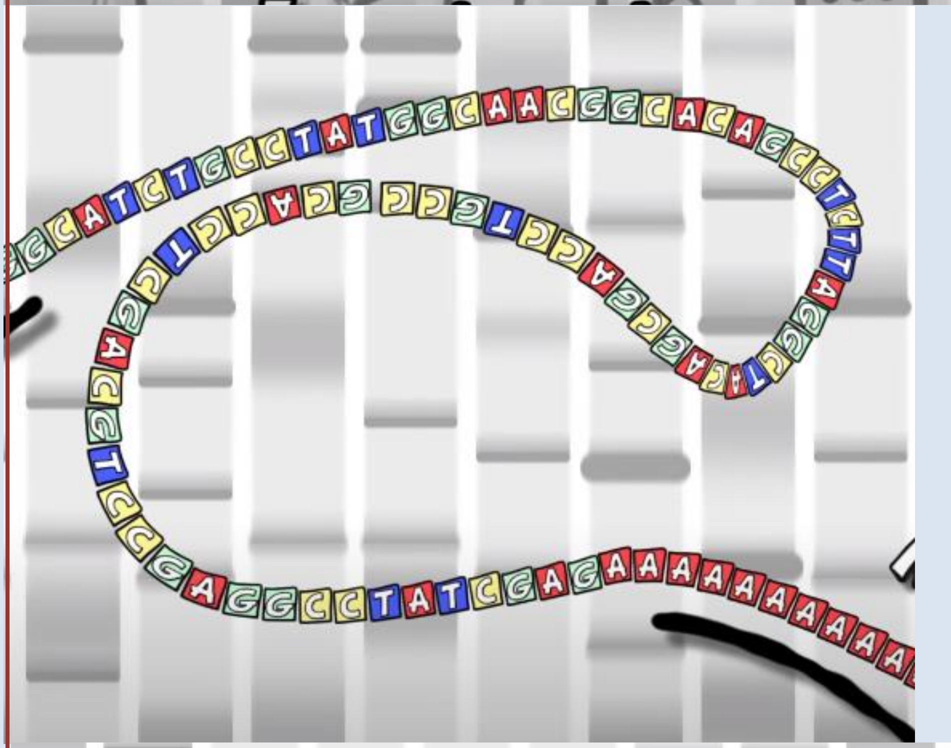
KEY PLAYERS IN DNA STORAGE:

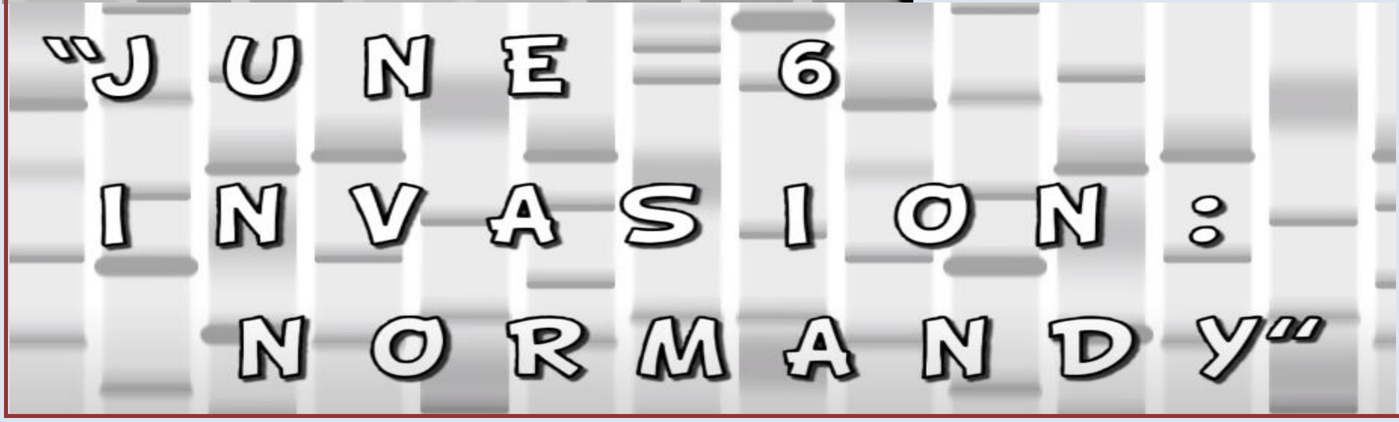
- ACADEMIA-> University of Washington, Harvard university, Columbia University, University of Illonois(Urbana-Champaign), ETH Zurich
- Research Consortium-> Semiconductor Research Corporation
- Industry-> Microsoft, Micron technologies, Apple, Facebook, Google, Intel, and IBM
- Start-ups-> Catalog, Iridia, HelixWorks
- US-Government-> DARPA, IARPA, NIH, NSF
- Foreign-> European Bioinformatics Institute

BELOW CAN BE A PROTOTYPE MODEL OF HOW DATA IS STORED IN THE DNA



0	1	2	3	4	5	6	7
AAA	AAC	AAG	AAT	ACA	ACC	ACG	ACT
8	9	A	B	C	D	E	F
AGA	AGC	AGG	AGT	ATA	ATC	ATG	ATT
G	H	I	J	K	L	M	N
CAA	CAC	CAG	CAT	CCA	CCC	CCG	CCT
O	P	Q	R	S	T	U	V
CGA	CGC	CGG	CGT	CTA	CTC	CTG	CTT
W	X	Y	Z	Sp	:	,	–
GAA	GAC	GAG	GAT	GCA	GCC	GCG	GCT


GGCATCTGCCTATGGCAACGGCACAGCCTCTTAGGCTACAGCGACCTGCCGCACCTCGACGTCCGAGGCCTATCGAGAAAAAAAAAAAA



"JUNE 6

INVASION:

NORMANDY"

FINDINGS:

- 20-Variable 3-SAT Problem: This is the largest problem solved yet with a DNA computer which is 20 variables for three-satisfiability problem. This problem is an NP (Nondeterministic Polynomial) time-complete computational problem. As the problem complexity is very high, even with fastest sequential algorithms, exponential time to solve this problem is required. This was the reason it leads to examining the performance of DNA computers. Subsequences based separation used in Striker Model is used to solve this problem. Two basic operations are used by Striker Model for computations: application of strikers and separation based on subsequence. 20-variable SAT problem uses separations. Oligonucleotide probes restrained in polyacrylamide gel-filled glass modules are employed for carrying out separations. Electrophoresis moves the DNA strands through the module. Strands which are complementary to the immobilized modules hybridized while the strands which do not contain complementary strands pass by. Electrophoresis at a higher temperature which is higher than the melting temperature of the probes is used to free netted strands. Remaining strands are transported to other modules. This problem maximizes the use of parallelism offered by DNA computers.

- Data storage crises are going to occur: The rise of the internet age, and its associated technologies and platforms, has led to an explosion in the amount of digital data being produced. By 2025, humans are expected to produce 160 zettabytes of data each year. The demand for digital data storage is currently outpacing our storage capabilities, and the gap is widening as the amount of digital data produced grows exponentially. Modern digital information storage technologies (e.g., flash memory) depend on microelectronics made from silicon. Analysts estimate that storing the world’s data in flash in 2040 would require more than 1,000 kilograms (kg) of wafer-grade silicon. The projected supply of single crystal wafer grade silicon in 2040 is 108 kg.3 New, sustainable materials will be required to support the world’s information technology base and storage of digital data.

- History of data storage in DNA:

Date	Size (MB)	Group	Data description
1988	.0000004	Harvard University	Encoded image
1999	.00009	Ars Electronica	Encoded text from genesis
2003	.0001	Pacific Northwest National Laboratory	Part of “it’s a small world”
2005	.0001	DNA2.0(Now ATUM)	Poem “tomten”
2009	.0002	University of Toronto	Text,music,image
2010	.0009	The j. carig venter Institute	Watermarking by synthetic genome
Aug-2012	.66	Harvard University	Book (53,426 words,11 jpg images) and javascript program
Feb-2013	.74	European Bioinformatics Institute	Shakespeare’s sonnets, 26-sec audio clip of a MLK sppech, Watson and Cricks Paper on the Structure of DNA

Feb-2015	.08	ETH Zurich	Swiss federal character of 1291, Archimedes palimpsest
Apr-2016	.15	Microsoft, University of Washington	Image files
Jun-2016	22	Harvard University, Technicolor	Mpeg compressed movie sequence
Mar-2017	2.14	New York Genome Center, Columbia University	Graphical os, movie, pdf, text and malware
Mar-2017	200	Microsoft, University of Washington	Universal declaration of human rights(in 100 languages), hd music vedio, db of seeds stored in Svalbard global seed vault
Feb-2018	400	Microsoft, University of Washington	Unspecified
Apr-2018	15	ETH Zurich, Rice University	Music album

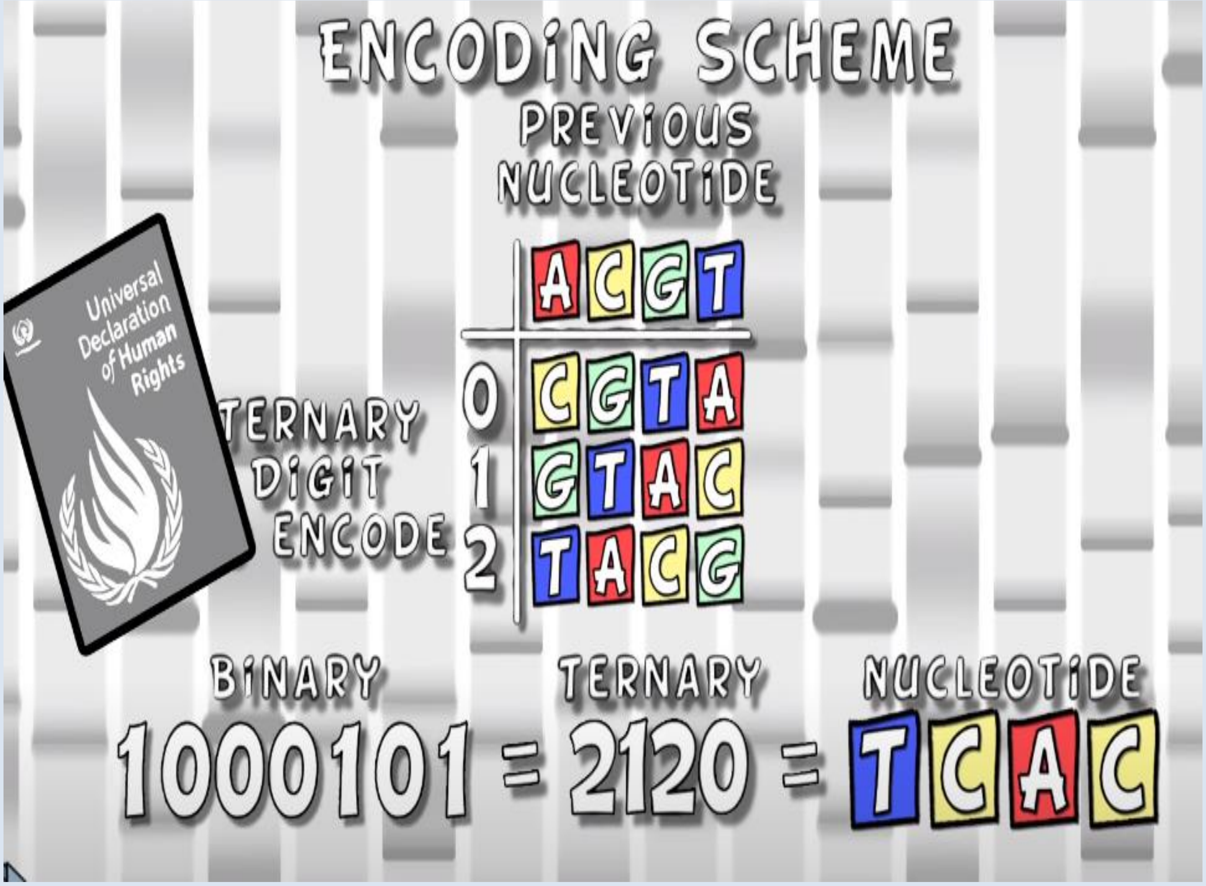
- The science behind storing data in DNA has been proven: Researchers have demonstrated that DNA is a scalable, random-access and error-free data storage system. DNA is also stable for thousands of years and offers utility in long-term data storage. Advancements in next generation sequencing have enabled rapid and error-free readout of data stored in DNA. As the data storage crisis worsens in the coming years, DNA will be utilized to store vast amounts of data in a highly dense medium.

- The first commercial DNA storage company called Catalog is poised to take orders in 2019: Catalog is building a proprietary DNA data storage machine in partnership with Cambridge Consultants that will synthesize 1TB of data per day at a cost of a few thousand dollars. This will revolutionize our approach to archival data storage and pave the way forward towards further advancements in the field.
- Currently, the cost of DNA synthesis is limiting advancement within this field: DNA synthesis methods are still reliant on methods based on organic chemistry, that are three decades old. Our ability to synthesize novel DNA sequences that store data is significantly limited by the inefficiencies of non-biological DNA synthesis methods. These inefficiencies have limited DNA data storage to the domains of research laboratories and significantly limit the data file size that can be stored in DNA.
- Technology that utilizes engineered biological enzymes to synthesize DNA fragments will radically decrease costs and propel the field forward: Biology-inspired engineering approaches to synthesizing DNA will be the catalyst that drives down cost of DNA synthesis. Several independent groups are developing DNA synthesis technologies that utilize enzymes to construct novel DNA sequences. This 2nd generation synthesis technology already has commercial developers and is also actively being developed in academia. This technology is the first major breakthrough in DNA synthesis in decades and should lead to significant reductions in costs and facilitate development of new technologies that support storing data in DNA.

- New technologies for faster reading of data stored in DNA are also needed to advance the field forward: In conjunction with the development of cheap DNA synthesis technologies, technologies that are compact, fast, and efficient must be developed to allow for easy read-out of data stored in DNA. Development of advanced coding schemes and operating systems tailored to DNA storage devices are also necessary. Random access retrieval of data stored in DNA facilitated by biologically-inspired mechanisms must be developed. Furthermore, the density of data stored in DNA can be exponentially increased by exploring the utility of DNA modification techniques.

- As storing data in DNA becomes cost-effective, this field's technological advantages will revolutionize our approach to data access and computing: DNA offers compelling advantages over today's best methods for data storage such as orders of magnitude greater storage density, and long-term stability. These are significant competitive advantages that should facilitate the commercial usage of DNA as data storage medium, especially as DNA synthesizing technology becomes cost-effective. New supporting technologies will be developed in conjunction, such as computers with operating systems tailored for random access retrieval of data stored in DNA. These advancements will usher in a new paradigm for computing with little to no limitations on the volume of data that we can produce, store, and access.
- DNA is made up of chains of four base nucleotides: adenine, guanine, cytosine, and thymine (labeled A, G, C, and T, respectively). For data storage purposes, special algorithms convert the binary digital files of 1s and 0s into the four bases; say, 00 for A, 01 for G, 10 for C and 11 for T. For encoding data, information is transmitted by synthesizing DNA strings with specific base patterns. The files can then be decoded using modern DNA sequencing technology. The theoretical maximum capacity of information storage per nucleotide is two bits.
- In 2016, Microsoft announced a record of storing 200 megabytes (MB) of data using about 1.5 billion unique pieces of DNA. This year, researchers from Columbia University and the New York

Genome Center have reported the development of the DNA Fountain algorithm, which approaches 85% of the theoretical storage limit per nucleotide—60% better than previous studies. In addition, the information storage and the retrieval was 100% reliable and error-free.



TERNARY DIGIT ENCODE	PREVIOUS NUCLEOTIDE A	C	G	T
0	C	G	T	A
1	G	T	A	C
2	T	A	C	G

BINARY 1000101 = TERNARY 2120 = NUCLEOTIDE TCAC

- Above is an example how the normal data is encoded into Genitical data.

BENEFITS

DNA has many advantages for storing digital data.

- It is ultra-compact,It is stable
- Storage density: DNA’s information storage density is several orders of magnitude higher than any other known storage technology.
- It can last hundreds of thousands of years if kept in a cool, dry place.
- As long as human societies are reading and writing DNA, they will be able to decode it.
- DNA won’t degrade over time like cassette tapes and CDs, and it won’t become obsolete.
- DNA computing consumes significantly less energy than the electronic computers. Energy consumed by DNA computers is billion times comparatively less than other electronic computers. The storage space needed to store information is less than trillion times over electronic computers. Furthermore, DNA computers offer parallelism at a high level. Millions and trillions of molecules perform chemical reactions parallel.
- DNA memory can be effectively utilized in commercial applications and in national security for information hiding purposes and for data stenography. Deinococcus bacteria can live and multiply without human interference. This property can be used to preserve data at nuclear catastrophe.
- Researchers have recently shown that one gram of DNA is capable of storing 215 petabytes of digital data. In other words, all of the information humans have ever recorded could be contained in a single room if stored as DNA. Apart from being the densest known storage medium, the information encoded in DNA can last practically forever when kept in a cool, dry, and dark place, as shown by the ability to reconstruct a human genome from a bone of more than 400,000 years of age.

CHALLENGES

- The biggest hurdle faced while storing data in DNA are cost and efficiency.
- Encoding the given data into DNA sequence requires lots of time and hence is incredibly slow. Studies shows that the rate of data encoding in DNA is 400 bytes which is millions of times slower than the microsecond timescales for reading and writing bits in a silicon memory chip.
- The other challenges include making it exact, even a slight error while encoding the data in DNA could have sever impacts on the original data.
- The other challenges include the identification part, how one will identify the correct DNA strand used to store them and how to remove the other. To overcome these challenges, researchers introduced two techniques named as DNA Enrichment and Nested Separation or together called as Dense.
- During DNA synthesis and sequencing, errors tend to crop up.
- Reading and writing DNA is a fairly slow process compared to other forms of data storage and it might not be suited where information is needed quickly. DNA storage can work best for archival purpose.
- Portability - hard to share the data since only a very limited number of places can write / read digital data in DNA.
- Expensive - about 10 cents for encoding a million base pairs, which is equivalent to about a dollar per megabyte ($1K per GB, and $1M per TB).
- DNA synthesis for the storage purpose consumes a lot of budget and still now no cost-effective methods have been discovered.

APPLICATIONS

NEW TECHNIQUE OF WATERFORD RESEARCHERS 2018:

On 20th February 2018 Irish scientists at Waterford institute announced a new technique of storing and recovering data in DNA strand using bacterial molecules. Their survey shows that human in 2025 the proliferation of data will become 160 zettabytes. The researchers encrypt an easy message - during this case "Hello World" - into the plasmids and store them during a strain of the E. coli bacterium referred to as Novablue that's cornered during a specific location that becomes the archive storage location.

Another variety of E. coli bacterium, HB101, that is mobile, is then discharged and travels to the Novablue. Once it meets it, the plasmids containing the information are transferred from the Novablue to the HB101 through an association method called conjugation.

The HB101 then swims to a tool capable of extracting the plasmids and reading the info they store. The movement of the bacterium and therefore the conjugation is controlled and created potential by the utilization and placement of 2 completely different antibiotics, antibacterial and antibiotic, among the archive storage and retrieval space.

Novablue is immune to Achromycin, whereas HB101 is in a position to resist antibiotic. In order to complete its swim across the archive enclosure, the HB101 should thus initial conjugate with the Novablue so as to select up its resistance to Achromycin.

Illinois improve viability of DNA data storage

However, using DNA as storage is currently costly and there is a lack of processing systems suited for this technology.

The University of Illinois at Urbana-Champaign is undertaking a $1.5 million effort to produce new DNA-based storage nanoscale devices using chimeric DNA, a hybrid molecule made from two different sources. As part of the three-year project, “SemiSynBio: An on-chip nanoscale storage system using chimeric DNA," the team will design a method to read, write, and store data in a more cost-effective way than current DNA storage techniques.

MICROSOFT'S REPORT 2016

Microsoft with its researchers and scientists has reported a big leap forward for DNA data storage, DNA could be a better way to store data for the long term than the magnetic tape companies rely on today. Microsoft has stored 200MB of data and demonstrated that on July 7 2016.

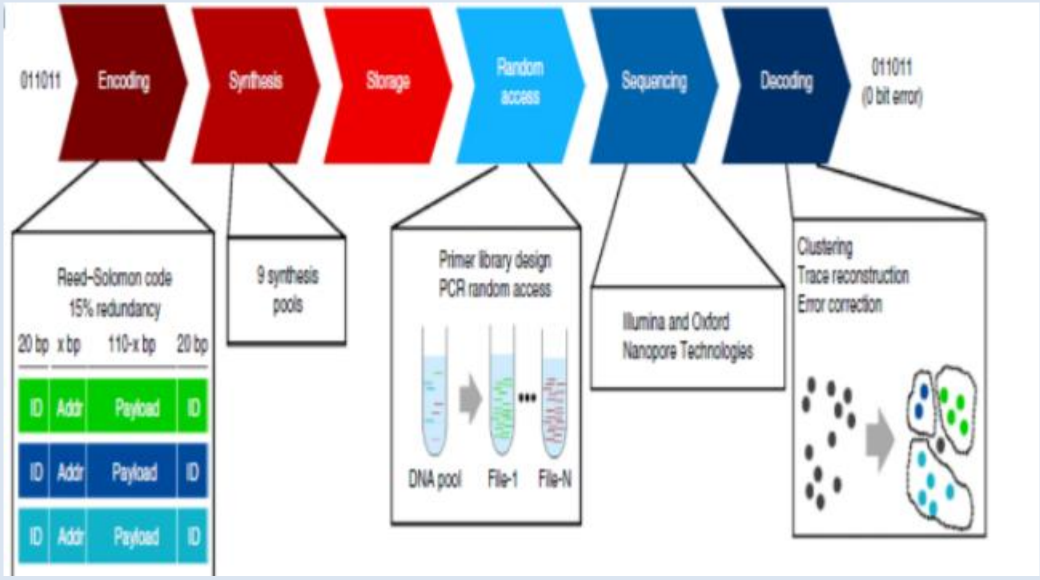
MICROSOFT'S REPORT 2017

March 2, 2017,Columbia University School of Engineering and Applied Science has proposed an algorithm designed for streaming video on a cell phone can unlock DNA's nearly full storage potential by squeezing more information into its four base nucleotides. They demonstrate that this technology is also extremely reliable.

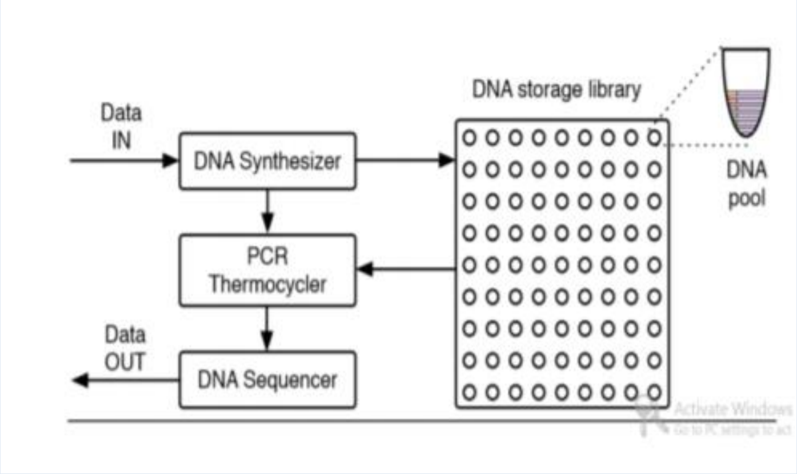
MICROSOFT AND TWIST BIO- SCIENCE REPORT 2018

Researchers at Microsoft in collaboration with the U.S Biotech Company named Twist biotech have succeeded in storing 400 megabytes of data. This project named "Random Access in Large scale DNA Storage" has overcome the disadvantage of the previous method proposed in 2016. The researchers and the scientists used durable synthetic DNA strand for data storage. The conventional technique in the previous research, retrieving the stored data in the DNA molecule requires sequencing the entire pool of DNA. This approach encoded about 35 different files of 400 MB into 13 million strands DNA molecules and retrieves the files individually with zero error.

Phases of random access in DNA:



Prototype for DNA STORAGE can be like this:



DNA SYNTHESIZER

The DNA synthesizer will encode the input data.

DNA POOL

The pool contains a collection of DNA molecule that can be mapped on to the storage library.

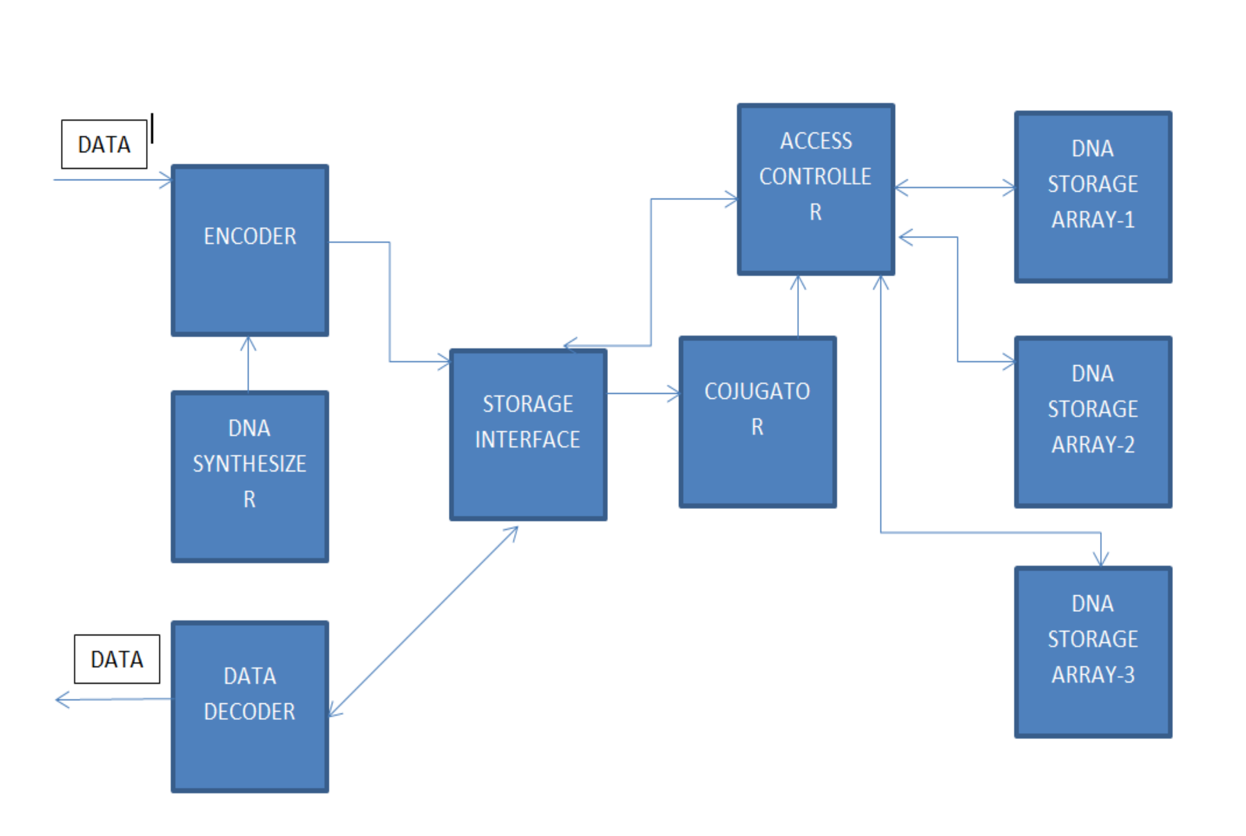
DNA SEQUENCER

The DNA sequencer will sequences the data and converts them back to digital data.

ENCODING MECHANISM

The binary data is converted into a ternary data with the help of Huffman encoding. This ternary data is further converted into nucleotide code.

PROPOSED ARCHITECTURE BY OUR UNDERSTANDING



ENCODER

Encodes the data that reaches it

Decoder

Decodes the data

DNA Synthesizer

Synthesis of DNA for storing data

Storage interface

This helps to interact with the DNA storage array and also this contains a certain amount of Cache memory so that if the data that has been written to the DNA storage array needed as and then without disturbing the DNA storage array this Cache will provide data.

Access controller

This controls the access as we know here we use three storage arrays two of which are redundant so based on the interval between the access time the particular array is selected so as to ensure that no array is disturbed frequently.

Conjugator

The conjugator helps us to produce the redundant data in DNA array by using method of conjugation. By this we need to produce a DNA sample with data and we copy the data into two more DNA samples by conjugation and each sample is saved into each storage array

Storage array

The array that contains all the DNA samples that are storing the data

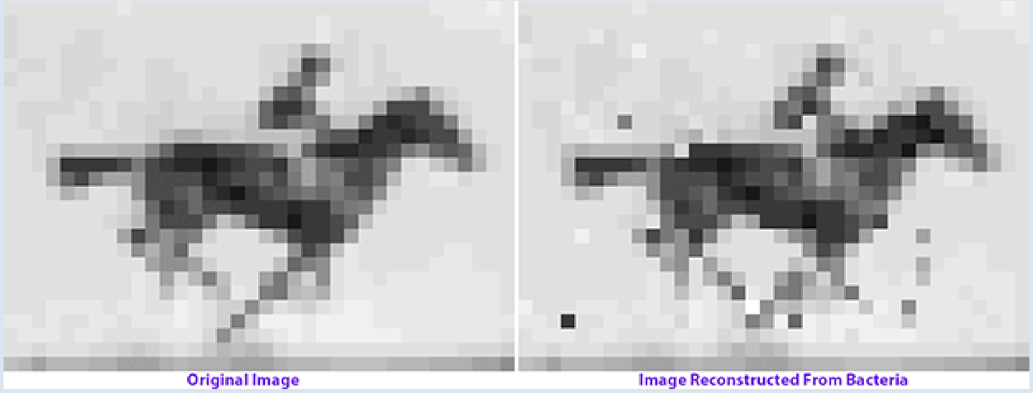
What are the Challenges that are being solved by the proposed architecture

- Making a secondary storage for DNA by general architecture need encoding to be done again but in our architecture the conjugator does the job with low cost
- When DNA is exposed to environment in frequent intervals it under go mutation so in our architecture there is a access controller which does not allow the data to be accessed from particular data array for many times.
- Encoding and then going for retrieval costs high so we maintain a Cache memory in storage interface that store the frequently encoded data so that if data is needed it is retrieved from the Cache.
- Here even if the primary data source fails or corrupted we will have our backup data in other arrays
- Even if any array fails or corrupted we can make a new array by using the conjugator to copy data into new array instead of going for retrieving the data from present storage and then encoding them into new array
- Access controller ensure that no particular array is accessed in short intervals and avoid exposure and mutation

CASE STUDIES

UNIVERSITY OF HARVARD:

1. Researchers at Harvard University uploaded a GIF into a living E.Coli bacteria. Not only did they upload the files, they were also able to download them from copies of same bacteria.



2. They did it using a revolutionary technology called CRISPR-Clustered Regularly Interspaced Short Palindrome Repeats.
3. But currently DNA storage is slow and expensive. And researchers at Harvard are not intending to store movies in bacteria, but using it as a medical record of all the genetic activity inside a cell.

MICROSOFT & UNIVERSITY OF WASHINGTON:

1. In 2012, UK scientists encoded 739 kb of data into DNA strands including all 154 Shakespeare sonnets and an excerpt from Martin Luther King's "I Have A Dream" speech.
2. And 4 years later, in 2016 researchers at Microsoft and University of Washington broke that record. They used binary coding to capture 200mb of data including Universal "Declaration Of Human Rights" and a High Definition "Ok Go" music video, all in strings of DNA.

3. The current theoretical limit of DNA's storage capacity is so high that we could fit 100 million HD movies on a pencil eraser. It is even conceivable that one day we can fit all of the information available on the internet into the space of a shoe box.

UNIVERSITY OF LJUBLJANA:

1. A group of researchers in Slovenia have demonstrated how genes can be inserted into bacteria, transferred into plants, and passed on to subsequent generations.
2. They encoded a short snippet of computer code into the DNA of tobacco plants. When the plants were cloned from cuttings and the DNA re-sequenced, converted back to binary, and executed as a Python script, the program ran flawlessly — printing "Hello World" on a computer screen.
3. Because DNA is so small compared to silicon-based storage media, they claimed, a handful of seeds could hold all of humanity's current digital archives. But even more exciting is the prospect of growing living libraries, where a garden or forest is made up of plants that hold vast stores of information. A visitor might pick a leaf or a flower, pop it in a hand-held DNA reader, and be able to read a classic book, browse Wikipedia, or listen to music.
4. On the other hand, the prospect of inserting digital content into living creatures is fraught with moral and practical implications, as is any genetic modification technology. One primary concern is how individual organisms and the environment might be affected by these extra genes. Researchers wrote that the experiment did not affect the plants' "vigour and fertility".
5. A more practical concern is that natural mutations would gradually corrupt the data as it was passed down through the generations. That's a tough nut to crack — but seeds, at least, could be frozen to preserve their information infused DNA for centuries.

STANFORD UNIVERSITY:

1. Storage project funded by NASA: Scientists have synthesized a molecular system that, like DNA, can store and transmit information. This unprecedented feat suggests there could be an alternative to DNA-based life, as we know it on Earth- a genetic system for life that may be possible on other worlds.
2. DNA is a complex molecule that stores and transmits genetic information, is passed from parent to offspring in all living organisms on Earth, and its components include four key ingredients called nucleotides- all standards for life as we know it.

PREVENTING DATA LOSS

- The DNA STORAGE can be a good archival storage as per the present research
- And DNA mutation leads to corruption of data and data loss so as to prevent that we store the DNA that contains the DNA in the cold temperature to stop mutation. So when we say we store in the cold storage it means we cannot access data in frequent intervals because if we access the DNA storage in frequent intervals it leaves the DNA to expose to the outer environment and my lead to data loss by mutation
- Conjugation is the method through which we can transfer the data from the one DNA to another by this method we need not to use encoding techniques for all DNA molecules that we use to store a redundant data.
- So when we are going to copy redundant data we can go for conjugation but conjugation should be done in controlled environment and there are many limitations for conjugation you must select two DNA samples so that after conjugation we must able to separate them.

- Finally we can say that until and unless we expose the DNA strand to the environment for a long time our data is safe, and uncorrupted.

CONCLUSIONS

From the above study we can get to know that the DNA storage is a viable and reliable data storage for our future. And we can also understand how far the research was done and we can identify the flaws in the process like the data storing in DNA is not cost effective, DNA synthesis takes lot of time, we are still working on Cost-effective encoding methods, DNA storage is vast that around some handful of DNA strands could store the data that humans had seen forever since now. In future we will use DNA readers to go through the DNA and to get data from the storage. Many companies are also working the manufacturing of portable DNA reader machines. And here form the research we come to how DNA storage become much more important than other storages. Basically we are going to produce around 160 zettabytes of data by 2025 which cannot be handled by our present infrastructures. So we can finally conclude that DNA is the Future of our data storage techniques which presently developing rapidly by many companies around the world as then many improvements are done and progress is rapid now a days because of developing technologies and evolving ideas in the researches. Not only companies many institutions are also working in this field to achieve the reliable DNA data storage.

REFERENCES

[0] A Survey Paper on DNA-Based Data Storage Siva Rama Krishnan S-1 Shubham-2, Jagrit-3,[1]Assistant Professor (Senior) School of Information Technology and Engineering VIT, Vellore-By our faculty.

[1] Cox, J. P. (2001). Long-term data storage in DNA. TRENDS in Biotechnology, 19(7), 247-250.

[2] Bornholt, J., Lopez, R., Carmean, D. M., Ceze, L., Seelig, G., & Strauss, K. (2016). A DNA-based archival storage system. ACM SIGARCH Computer Architecture News, 44(2), 637-649.
[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] Yazdi, S. H. T., Yuan, Y., Ma, J., Zhao, H., & Milenkovic, O. (2015). A rewritable, random-access DNA-based storage system. Scientific reports, 5, 14138.

[5] Richard P. Feynman- There's Plenty of Room at the Bottom
[6] Pavani Yashodha De Silva and Gamage Upeksha Ganegoda Faculty of Information Technology, University of Moratuwa, Katubedda, Moratuwa, Sri Lanka.
[7] Vishal Bhatnagar Shri Venkateshwara University.

[8] Molecular Digital Data Storage using DNA by Luis Ceze, Jeff Nivala and Karin Strauss.
[9] A Potomac Institute for Policy Studies Report [9] Shimanovsky, B., Feng, J., & Potkonjak, M. (2002, October). Hiding data in DNA. In International Workshop on Information Hiding (pp. 373-386). Springer, Berlin, Heidelberg.
[10] 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.

[11] Shipman, S. L., Nivala, J., Macklis, J. D., & Church, G. M. (2017). CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. Nature, 547(7663), 345-349.

[12] Sun, J., Wang, Q., Diao, W., Zhou, C., Wang, B., Rao, L., & Yang, P. (2019). Digital information storage on DNA in living organisms. Medical Research Archives, 7(6).

[13] Waltz, E. (2017). Biocomputer and memory built inside living bacteria [News]. IEEE Spectrum, 54(9), 11-12.

[14] Fister, K., Fister, I., & Murovec, J. (2017). The Potential of Plants and Seeds in DNA-Based Information Storage. In Understanding Information (pp. 69-81). Springer, Cham.

[15] O'Neill, S. (2016). I plant memories in seeds. New Scientist, 229(3056), 27.

[16] Taluja, S., Bhupal, J., & Krishnan, S. R. (2020, February). A Survey Paper on DNA-Based Data Storage. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-4). IEEE.

[17] Ceze, L., Nivala, J., & Strauss, K. (2019). Molecular digital data storage using DNA. Nature Reviews Genetics, 20(8), 456-466.

[18] Meiser, L. C., Antkowiak, P. L., Koch, J., Chen, W. D., Kohll, A. X., Stark, W. J., ... & Grass, R. N. (2020). Reading and writing digital data in DNA. Nature Protocols, 15(1), 86-101.

[19] Stanley, P., Strittmatter, L. M., Vickers, A. M., & Lee, K. C. (2020). Decoding DNA data storage for investment. Biotechnology Advances, 107639.

[20] Chen, K., Zhu, J., Bošković, F., & Keyser, U. F. (2020). Nanopore-Based DNA Hard Drives for Rewritable and Secure Data Storage. Nano Letters, 20(5), 3754-3760.

[21] Li, B., Song, N. Y., Ou, L., & Du, D. H. (2020). Can We Store the Whole World's Data in {DNA} Storage?. In 12th {USENIX} Workshop on Hot Topics in Storage and File Systems (HotStorage 20).
[22] Sharma, D., Kumar, R., Gupta, M., & Saxena, T. (2020). Encoding scheme for data storage and retrieval on DNA computers. IET nanobiotechnology, 14(7), 635-641.
[23] https://www.youtube.com/watch?v=r8qWc9X4f6k
[24] https://www.youtube.com/watch?v=wxStlzunxCw

[25] https://www.youtube.com/watch?v=gK3dcjBaJyo
[26] https://www.youtube.com/watch?v=DMYgjOHgHxc
[27] https://www.youtube.com/watch?v=N7zJLSEZKYQ

[28] Kim, J., Bae, J. H., Baym, M., & Zhang, D. Y. (2020). Metastable hybridization-based DNA information storage to allow rapid and permanent erasure. Nature Communications, 11(1), 1-8.

[29] Tabatabaei, S. K., Wang, B., Athreya, N. B. M., Enghiad, B., Hernandez, A. G., Fields, C. J., ... & Milenkovic, O. (2020). DNA punch cards for storing data on native DNA sequences via enzymatic nicking. Nature communications, 11(1), 1-10.

[30] Dimopoulou, M., Antonini, M., Barbry, P., & Appuswamy, R. (2020, May). Storing Digital Data Into DNA: A Comparative Study Of Quaternary Code Construction. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4332-4336). IEEE.

****THANKS FOR OUR INSTRUCTOR'S SUPPORT****

******Siva Rama Krishnan S******

****

ADDITIONAL INFORMATION UPON PRESENT SITUATIONS IN THE FIELD

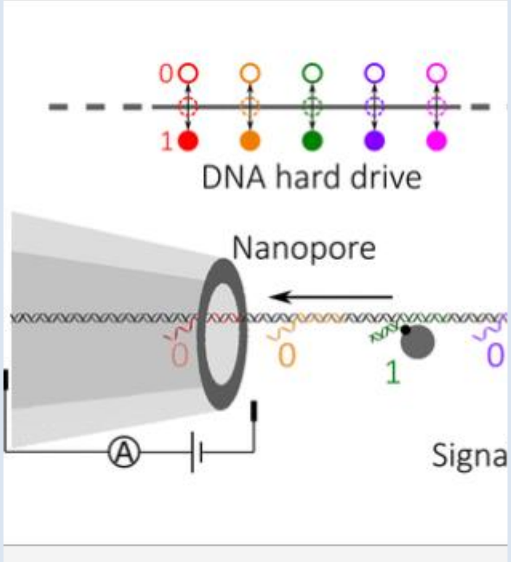
Research paper	summary
[16] Taluja, S., Bhupal, J., & Krishnan, S. R. (2020, February). A Survey Paper on DNA-Based Data Storage. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-4). IEEE. https://ieeexplore.ieee.org/abstract/document/9077678	2.5 quintillion bytes of data are generated on the internet every day. So, the demand for data storage is growing exponentially, but the capacity of existing storage media is not keeping up. A revolution in the field of data storage is the need of an hour. This paper surveys a very unique technique of storing digital data in DNA sequences. The idea is to replicate nature's way of storing information. It has been around with us for ages. This paper elaborates on the procedure, applications, and challenges that are associated with this fictitious idea of data storage. Using DNA to archive data is an attractive possibility because it is extremely dense, with a raw limit of 1 Exabyte/mm3 (109 GB/mm3), and long-lasting, with an observed half-life of over 500 years
[17] Ceze, L., Nivala, J., & Strauss, K. (2019). Molecular digital data storage using DNA. Nature Reviews Genetics,	Molecular data storage is an attractive alternative for dense and durable information storage, which is sorely

20(8), 456-466. https://www.nature.com/articles/s41576-019-0125-3	needed to deal with the growing gap between information production and the ability to store data. DNA is a clear example of effective archival data storage in molecular form. In this Review, we provide an overview of the process, the state of the art in this area and challenges for mainstream adoption. We also survey the field of in vivo molecular memory systems that record and store information within the DNA of living cells, which, together with in vitro DNA data storage, lie at the growing intersection of computer systems and biotechnology.
[18] Meiser, L. C., Antkowiak, P. L., Koch, J., Chen, W. D., Kohll, A. X., Stark, W. J., ... & Grass, R. N. (2020). Reading and writing digital data in DNA. Nature Protocols, 15(1), 86-101. https://www.nature.com/articles/s41596-019-0244-5	archiving digital information in the form of DNA and for subsequently retrieving it from the DNA. In principle, information can be represented in DNA by simply mapping the digital information to DNA and synthesizing it. However, imperfections in synthesis, sequencing, storage and handling of the DNA induce errors within the molecules, making error-free information storage challenging. The procedure discussed here enables error-free storage by

	protecting the information using error-correcting codes. Specifically, in this protocol, we provide the technical details and precise instructions for translating digital information to DNA sequences, physically handling the biomolecules, storing them and subsequently re-obtaining the information by sequencing the DNA. Along with the protocol, we provide computer code that automatically encodes digital information to DNA sequences and decodes the information back from DNA to a digital file. The required software is provided on a Github repository. The protocol relies on commercial DNA synthesis and DNA sequencing via Illumina dye sequencing, and requires 1–2 h of preparation time, 1/2 d for sequencing preparation and 2–4 h for data analysis. This protocol focuses on storage scales of ~100 kB to 15 MB, offering an ideal starting point for small experiments. It can be augmented to enable higher data volumes and random access to the data and also allows for future sequencing and synthesis technologies, by changing the

	parameters of the encoder/decoder to account for the corresponding error rates.
[19] Stanley, P., Strittmatter, L. M., Vickers, A. M., & Lee, K. C. (2020). Decoding DNA data storage for investment. Biotechnology Advances, 107639. https://www.sciencedirect.com/science/article/pii/S0734975020301415	DNA's perpetual role in biology and life science is well documented, its burgeoning digital applications are beginning to garner significant interest. As the development of novel technologies requires continuous research, product development, startup creation, and financing, this work provides an overview of each respective area and highlights current trends, challenges, and opportunities. These are supported by numerous interviews with key opinion leaders from across academia, government agencies and the commercial sector, as well as investment data analysis. Our findings illustrate the societal and economic need for technological innovation and disruption in data storage, paving the way for nature's own time-tested, advantageous, and unrivaled solution. We anticipate a significant increase in available investment capital and continuous scientific progress, creating a ripe environment on which DNA data storage-

	enabling startups can capitalize to bring DNA data storage into daily life.
[20] Chen, K., Zhu, J., Bošković, F., & Keyser, U. F. (2020). Nanopore-Based DNA Hard Drives for Rewritable and Secure Data Storage. Nano Letters, 20(5), 3754-3760. https://pubs.acs.org/doi/abs/10.1021/acs.nanolett.0c00755	Nanopores are powerful single-molecule tools for label-free sensing of nanoscale molecules including DNA that can be used for building designed nanostructures and performing computations. Here, DNA hard drives (DNA-HDs) are introduced based on DNA nanotechnology and nanopore sensing as a rewritable molecular memory system, allowing for storing, operating, and reading data in the changeable three-dimensional structure of DNA. Writing and erasing data are significantly improved compared to previous molecular storage systems by employing controllable attachment and removal of molecules on a long double-stranded DNA. Data reading is achieved by detecting the single molecules at the millisecond time scale using nanopores. The DNA-HD also ensures secure data storage where the data can only be read after providing the correct physical molecular keys. Our approach allows for easy-writing and easy-reading,

	rewritable, and secure data storage toward a promising miniature scale integration for molecular data storage and computation.  0 1 DNA hard drive Nanopore 0 0 1 0 A Signa
[21] Li, B., Song, N. Y., Ou, L., & Du, D. H. (2020). Can We Store the Whole World's Data in {DNA} Storage?. In 12th {USENIX} Workshop on Hot Topics in Storage and File Systems (HotStorage 20). https://www.usenix.org/conference/hotstorage20/presentation/li	The total amount of data in the world has been increasing rapidly. However, the increase of data storage capacity is much slower than that of data generation. How to store and archive such a huge amount of data becomes critical and challenging. Synthetic Deoxyribonucleic Acid (DNA) storage is one of the promising candidates with high density and long-term preservation for archival storage systems. The existing works have focused on the achievable feasibility of a

	small amount of data when using DNA as storage. In this paper, we investigate the scalability and potentials of DNA storage when a huge amount of data, like all available data from the world, is to be stored. First, we investigate the feasible storage capability that can be achieved in a single DNA pool/tube based on current and future technologies. Then, the indexing of DNA storage is explored. Finally, the metadata overhead based on future technology trends is also investigated.
[22] Sharma, D., Kumar, R., Gupta, M., & Saxena, T. (2020). Encoding scheme for data storage and retrieval on DNA computers. IET nanobiotechnology, 14(7), 635-641. https://digital-library.theiet.org/content/journals/10.1049/iet-nbt.2020.0157	There has been exponential growth in the amount of data being generated on a daily basis. Such a huge amount of data creates a need for efficient data storage techniques. Due to the limitations of existing storage media, new storage solutions have always been of interest. There have been recent developments in order to efficiently use synthetic deoxyribonucleic acid (DNA) for information storage. DNA storage has attracted researchers because of its extremely high data storage density, about 1 exabyte/mm3

	and long life under easily achievable conditions. This work presents an encoding scheme for DNA-based data storage system with controllable redundancy and reliability, the authors have also talked about the feasibility of the proposed method. The authors have also analysed the proposed algorithm for time and space complexity. The proposed encoding scheme tries to minimise the bases per letter ratio while controlling the redundancy. They have experimented with three different types of data with a value of redundancy as 0.75. In the randomised simulation setup, it was observed that the proposed algorithm was able to correctly retrieve the stored data in our experiments about 94% of the time. In the situation, where redundancy was increased to 1, the authors were able to retrieve all the information correctly in the proposed experiments.