DNA: - The Future Supernatural Storage

Amit Toppo, Elizabeth Jacob
Department of Computer Science, St. Joseph’s College (Autonomous), Bangalore
Department of Computer Science, St. Joseph’s College (Autonomous), Bangalore

Abstract:
Demand for digital data storage is escalating exponentially due to the generation of massive digital data at an unimaginable pace, but the capacity of existing storage medium cannot store them up for long-term. These growing challenges have forced the scientists to look for the other alternatives to store digital data which could not only store data but also archive it for long-term and for eternity. Actually, this challenge has focused growing interests on DNA (Deoxyribonucleic Acid) as attractive alternatives for digital data storage because DNA has incredible capacity for high density data storage, and has efficient density of storage compared to any conventional system and modern storage media. Today, DNA-based data storage is an emerging near-future non-volatile storage medium which is potentially unprecedented highly-denser, durable, and replication efficient. This storage medium uses synthesizing DNA strings which contain user data and subsequently retrieving the stored data using high-throughput sequencing mechanisms.

Therefore, our research paper would closely give the prospects and overviews of DNA-based storage medium as the future supernatural storage for long-term, highly denser and secure digital data storage medium. We would also evaluate some of the system designs to understand the error-correction characteristics of encoding schemes, assess their overheads; and thus, make few projections about future feasibilities based on technology trends carried forward in the past.

Keywords — DNA-Based Storage, Nucleotides, Strands, Molecules, Synthesis, Encode, Decode, Long-Term,
corresponding DNA molecules and then stores them away. The decryption of data from DNA-based storage involves sequencing the DNA molecules and the decoding the data back to the original digital information. One of the most significant advantages of using synthesized DNA as a storage is that the storage density is very high and secure. Therefore, DNA-based storage has eternal relevance to store the data as long as there is DNA-based life; thus, making this as a strong future prospect for storing digital data in synthesized DNA will come true in near future.

Therefore, our research paper would closely give the prospects and overviews of the DNA-based storage medium as the future supernatural storage for long-term, highly denser and secure digital data storage medium. We would evaluate some of the system designs to understand the error-correction characteristics of encoding schemes, assess their overheads; and thus, make few projections about future feasibilities based on technology trends carried forward in the past. Our results would clearly give the prospect for the impending practicality of synthesized DNA-based archival storage medium as the greater solution to exponential growth in demand for digital data storage in the modern techno world. We will also highlight some of the important findings and researches are done in the past to make the DNA-based storage medium a reality.

II. A REVIEW OF EXISTING SYSTEMS AND THE CURRENT STATUS OF DNA-BASED STORAGE MEDIUM

Over the years, several types of research have been carried forward to make DNA-based storage a reality.

A) In the year 1999, the idea of Digital Data Storage was first implemented by Clell, Rsca, and Bancroft. The process demonstrated how any message can be hidden in DNA Microdots [3]. They were able to successfully encode words of information in short DNA strands (Microdots). The researchers implemented the concept of hiding secret data used in the World War II, to hide valuable information in DNA strands. To do this, they had to use 4 bases, PCR Primers and an Encryption key (base triplets used for English alphabets and Arabic numbers). The DNA-encoded message was hidden within the human genomic DNA and then scaled down that sample to a Microdot. This process seemed to be in the very complex background to encrypt information in DNA molecules; thus, ensuring data security and privacy. Their very important finding was PCR primer sequences and the specific encryption key which made it clear that the desired DNA can be encoded easily despite noise or unwanted sequences it might be surrounded with. Furthermore, they also brought forward the concept that DNA storage can be much more private and secure than any Digital Storage in Silicon devices that too without an encryption technique. The figure given below demonstrates how their research, coding scheme and the code they stored (Fig. 1).

B) The similar idea and a DNA based Storage Mechanism was further carried forward for encryption and storage of digital data by Bancroft, Bowler, and Clellan in the year 2001. By the use of iDNA (Information DNA) and the Poly Primer Key (Primer base sequence to access the information on iDNA), they managed to devise an experiment based on the same mechanism such that a forward and reverse primers were obtained using encryption scheme [3]. The encryption schemes are shown in the Fig. 1.

It was for the first time that a DNA based data storage technique was ever used which proved that development of DNA based storage device is possible using the simple principles of biology. But this experiment was not taken into careful account for the safety and tolerance of the DNA to extreme external and internal conditions. However, this very experiment paved the way forward for the future prospects in the new field of DNA computing. The very structure of DNA molecules for information storage and read are shown in Fig. 2.

C) In the year 2009, an Improved Huffman Coding was proposed by Ailenberg and Rotstein to facilitate storage of different types of Digital Media in DNA such as text, pictures, and audio, etc [5]. This was done using Huffman Code where DNA Codons for all the Characters on the keyboard were defined. The Plasmid Library was synthesized and arranged for storing and retrieving data by designing primers incorporated in the message segment which facilitates speedy, efficient retrieval of data. In this process, 7 bases were used for every 2 characters because Huffman Codes are better than their earlier counterparts.
D) A project done in the year 2010 by Craig Venter Institute managed to encode a 7920-bit watermark in the Genome sequence of the Bacterium [4]. With this synthesis artificial DNA, researchers were able to generate synthetic cells which could differentiate watermarked from natural cells and DNA. It was for the first time a complete synthetic cell was generated. Therefore, this project proved to be the largest project till data to encode a large amount of data into DNA nucleotides.

E) Furthermore, in the year 2011, three researchers Church, Gao, and Kosuri converted 11 JPG images with an HTML coded draft of more than 53,000 words book [6]. They were also able to convert a JavaScript program into a 5.27-megabit stream by assigning 0 as A or C (Adenine or Cytosine) and 1 as T or G (Thymine or Guanine). This was because the sequence like AAACCTGG read in one direction is the same as that of CCAGTTT read in reverse direction (3’-5’ direction). The converted streams were encoded into Oligonucleotides, each stream containing 96-bit data block. Then the entire stream was encoded onto 159-nt Oligos. Due to the huge length size, the sequence was chunked into 96-nt pieces where 19-nt bits represented the location of the text in the book. The 22-nt bits sequence were amplified and sequenced so that synthesizing the oligonucleotide library was easy. The book was read by limited-cycle PCR and then sequenced on a single lane using sequencing technology. This resulted in 100% recovery with 10 bits per 5.27 million errors due to single sequence coverage. Because this process was expensive, so the researchers split the DNA sequence into small chunks containing every 96 bases long. Therefore, each chunk included a 19-bit bar code or address. Further, DNA was synthesized, inkjet-printed on a glass DNA microchip and then cleaved off and dried to form a 50-nanogram clump smaller than a speck of pollen.

F) In the year 2013, Nick Goldman created an info system which was more feasible, had lots of capacity and needed less maintenance that present storage media [2]. He was able to encode 740 kilobytes of Hard disk stored digital data into a DNA code by synthesizing then and retrieving them completely. Goldman encoding technique splits the input DNA nucleotides into overlapping segments to provide fourfold redundancy for each segment. Because each window of four segments corresponds to a strand in the output encoding. Using this encoding, British scientists saved all 154 of Shakespeare’s sonnets on to DNA. That means that comparing loved ones to summer’s day will still be possible probably in the year 3016. Therefore, this encoding is considered as the most successful DNA storage technique.

G) In July 2016, a group of Microsoft Researchers and the University of Washington in collaboration with Twist Bioscience, a San Francisco start-up held the world record for successfully storing and retrieving 200 Megabytes of digital data in synthesized DNA cells [8]. The present encoding and decoding process of DNA nucleotides are shown in the Fig. 3.

The size of the synthesized DNA was said to be containing data comparable with the tip of a pencil. This research team has produced 10 million strands of synthetic DNA. Doug Carmean of Microsoft Research seems to have said that the company was ‘still years away from a commercially viable product’. The early tests have proved that recovering data from photos and videos without errors is possible as techniques are ‘improving quite rapidly.

III. PROPOSED SYSTEM/MODEL FOR NEAR FUTURE

If we look at all these major researches and findings, it is inevitable that DNA-based storage system would become a near-future archival medium soon. Along with the major findings, it poses several challenges, some due to its own physical composition, whereas some due to technological ineptness to unleash its full capacity at present. Every process of encoding, amplifying, sequencing, restructuring and finally decoding involves significantly more time and slower processing than the present storage mediums [1]. But DNA is unlikely to take over the optical, magnetic or quantum formats in the near future. There are many errors associated with the current technologies dealing with DNA. For example, Presence of Homo Polymers, sequencing errors, error due to lower access rates, etc. Even though DNA living cells have auto-correction enzymes, but there are no such artificial enzymes existing for artificial DNA. Due to its complex structure, it is more prone to mutations in extreme conditions leading to data alteration in a mutation.

The major challenge for practical DNA-based information storage is the difficulty of synthesizing very long sequences of DNA to a specified design. The other factors are high computational costs which are estimated to be $12,500 per MB for information storage in DNA and $220 per MB for decoding information.
The following challenges need to be tackled and advanced such as Rules for coding of text, music, and images with the improved Huffman coding, the index plasmid, Illustration of the improved Huffman coding, Unique sequencing primers for information retrieval, Information retrieval by sequencing, etc. Many other major challenges need to be addressed so that the DNA-based data storage become more feasible, less complex, less time-consuming in encoding and decoding. The researchers and scientists also need to address these challenges and develop efficient technologies and mechanisms which could make it possible to store digital data into DNA nucleotides efficiently and for long-term.

A) Representing data in DNA as the present mechanism

While DNA has many properties that make it different from existing storage media, there are parallels between traditional storage and DNA storage. At the lowest levels, traditional storage media store raw bits. The storage device abstracts the physical media, which could be a magnetic state, or the charge in a capacitor, or the stable state of a flip-flop, and presents to the storage hierarchy raw digital data. In a similar way, the abstraction of DNA storage 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 [1]. This section describes the challenges of representing data in DNA and presents several encodings that overcome these challenges.

B) Representation:

The University of Washington & Microsoft Research, on “A DNA-Based Archival Storage System”, Atlanta, CA, USA, April 2-6, 2016 states:

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 (e.g., mapping 0, 1, 2, 3 to A, C, G, or T respectively). For example, the binary string '01110001' maps to the base-4 string '1301', and then to the DNA sequence CTCTG. The code maps more common ASCII characters to 5-digit strings, offering minor compression benefits for textual data, though the effect on overall storage density is insignificant. The encoded binary digits are shown in Fig. 4.

C) Let us see how the simple encoding and decoding of digital data is processed at present.

DNA consists of four types of nucleotides: Adenine, Cytosine, Guanine, and Thymine, usually designated as A, C, G and T. DNA strand or molecules are linear sequences of these smaller molecules also called as nucleotides. Rather than creating sequences of 0s and 1s, as done in digital media, DNA storage uses sequences of the nucleotides. The sequencing of the nucleotides is done in several ways, but the concept is to designate digital data patterns to DNA nucleotides [7]. For example, 00 could be equivalent to A, 01 to C, 10 to T and 11 to G. Therefore, to store any picture value, we can encode it as digital data, like a JPEG file. The file is usually, a long string of 0s and 1s. We can assume that the first eight bits of the data are 01110100; so, we can break these binary values into pairs such as 01 11 01 00 which should also correspond to the nucleotide values C-G-T-A. Actually, this is the way the values of nucleotides are joined together to form a DNA Strand or molecule [7]. Furthermore, Digital files could be very large-even terabytes in size for large databases. But each DNA molecules must be much shorter holding about 20 bytes each strand. This would avoid difficulties in building DNA strand sequence chemically. Therefore, we need to make sure that the data files are broken into smaller chunks, and then add to each an indicator of where the corresponding sequence it falls. While reading the DNA-stored information, that indicator would ensure that all chunks of data are joined together in their proper sequence.

(i) Process of Encoding & Storing the Data

Once the order of the letters has been determined, the DNA sequences are manufactured letter by letter with chemical reactions. These reactions are done by equipment that takes in container A’s, C’s, G’s and T’s and then mixes them in a
liquid solution along with other chemicals to control the reactions [7]. This would specify the order of the physical DNA molecules. This process has the benefit of DNA storage. Backup copies because this doesn’t make one strand at a time but the reactions result with many identical strands at once before generating many copies of the next strand in the lines. DNA strands need to be protected from damage such as humidity and light; therefore, the DNA strands are dried and put into a container which can keep them cold and blocks water and light as well.

**Process of Decoding and reading data**

In order to decode the data back out of storage, highly sequencing machine is used for analysis of genomic DNA cells. This process identifies the strands, which generates a letter sequence per molecule. Once the sequence is generated, decoding is done into a binary sequence of 0s and 1s in order [7]. In this process DNA cells can damage but then those backup copies come into play because there are many copies of each sequence. If at all the DNA backup copies are depleted, the duplicate copies can be made to refill the storage just as nature copies DNA all the time. At present, DNA retrieval systems require reading all the data stored in a specified location, even if we want only a small amount of it. We have developed techniques- based on well-studied biochemistry which allows us to identify and read only specific pieces of information a user wants to retrieve from DNA storage.

**IV. FUTURE ENHANCEMENT**

Data storage in DNA has to face challenges like other revolutions of technologies. The future enhancements include a study on some of the challenges faced by this technology such as Slower retrieval process, Sequencing errors, Errors due to low access rate, Difficulty in synthesizing long sequences of DNA, High cost, etc.

It is fundamentally inefficient because it uses base-4 storage device. The best storage occurs for base-3. Theoretically, a few grams of DNA can store the world's information. And with proper coding and decoding schemes, DNA based storage technology can be used in various fields such as medical, agriculture, R&D, etc. This technology can be used to store, retrieve and hide data. Since it is invisible to human eye, data security and privacy can be ensured. Studies have shown that the cost of using this technology is dropping every year when compared to the electronic media that is being used in the present day. This technology calls for an ultimate paradigm shift in computing.

**V. CONCLUSIONS**

It is clear that data storage in DNA is no more confined to science fiction but is being realized and improvised at very promising rates by research teams all over the world. DNA-based storage has the potential to be the ultimate archival storage solution because it is extremely dense and durable. This idea has received positive criticism from the general public which can be inferred from their responses on the different science websites. Most DNA retrieval systems require reading all the data stored in a specified location, even if we want only a small amount of it. We have developed techniques- based on well-studied biochemistry which allows us to identify and read only specific pieces of information a user wants to retrieve from DNA storage.

At the moment, the DNA-based storage system is experimental due to the current state of DNA synthesis and sequencing; both technologies are improving at an exponential rate with much advances in the biotechnology. It is the right time given the impending limits of silicon technology, to consider incorporating biomolecules as a part and partial of computer design. We are hopeful that it becomes a reality one day in near future.

Therefore, in the future experiments, DNA storage system needs to be completely automated and the process of both encoding and decoding must be improved because both are prone to error, relatively slow and time-consuming. These are significant challenges but we need to be optimistic because all the relevant technologies improvements are growing rapidly. DNA data storage doesn’t need the perfect accuracy that biology requires. Therefore, researchers would find even cheaper and faster means to encode and decode the information in nature’s oldest data storage system.

**REFERENCES**


