DIGITAL DATA STORAGE IN DNA

KBIRU GAMBO

Bsc. IT.

PP Savani University, Surat– 394125 India.

MR. MITUL PATEL

Lecturer Department of Information Technology, School of Engineering.

PP Savani University, Surat – 394125. Gujrat India.

Abstract

A review on nowadays digital data storage in DNA, most of current digital data are mainly stored on magnetic and optical media. At the explosive era of digital data, the digital data are generated every day and increased at an exponential rate. These traditional media cannot meet the urgent requirement of big digital data storage. With such advantages as high density, high replication efficiency, long-term durability and long-term stability, deoxyribonucleic acid (DNA) is expected as a novel and potential data storage medium. For the new DNA data storage, the files or any data readable will be converted to binary and then encoded to DNA sequences consisting of Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The data-carrying DNA sequences will be synthesized and stored until data retrieval one day. Once data retrieval, the unique data-carrying DNA fragments will be amplified, sequenced and analyzed. The DNA-based data information will then be decoded into binary and eventually converted to the information readable. Currently, the application of DNA data storage is limited due to such disadvantages as high cost, time-consuming, lack of random-access ability. We still need to face serial tough challenges. However, the seen advances in DNA sequencing technology positively shine the future of DNA digital data storage. DNA is considered as a better storage system is that 215 petabytes (215 million gigabytes) can be stored in just 1 gram of DNA.

Keywords: Digital data storage; Deoxyribonucleic acid (DNA); Binary; Encoding; Decoding; Sequencing.

INTRODUCTION

The excursion of data storage initiated from bones, rocks, and paper. Then this journey deviated to punched cards, magnetic tapes, gramophone records, floppies, and so forth. Afterwards with the development of the technology optical discs including CDs, DVDs, Blu-ray discs, and flash drives came into operation. All of these are subjected to decay. Being nonbiodegradable materials these pollute the environment and also release high amounts of heat energy while using energy for operation [1]. With the employment of digital systems for the purpose of generation, transmission, and storage of information, there rises a need for active and ongoing maintenance of digital media.

With the massive amounts of digital data that has to be stored for future use, a problem arises in the storage of irresistible amounts of data. The demand for data storage is rapidly increasing day by day. The total information storage of the entire world was around **2.7 ZB** in 2012 [2].

Every year the storage necessity is increasing by 50%. Currently almost all of the digital data is stored with a technology that will last only for a limited time period. Memory cards and chips are sustainable for 5 years from their preliminary use [3].

Standard hard drives are prone to damage from high temperatures, moisture, and exposure to magnetic fields and through mechanical failures [4]. Though solid-state drives operate better than hard drives, if not power-driven for more than few months they tend to lose their information. Therefore, researchers' devotion has been driven towards development of a storage mechanism which overcomes the aforementioned drawbacks successfully.

Taking into account the manner fossil bones preserve genetic material for ages, researchers paid their attention towards using **deoxyribonucleic acid (DNA)** as a storage medium [5]. DNA and ribonucleic acid (RNA) are nucleic acids. Alongside proteins, lipids and complex carbohydrates (polysaccharides), nucleic acids are one of the four major types of macromolecules that are essential for all known forms of life [5].

The two DNA strands are known as polynucleotides as they are composed of simpler monomeric units called nucleotides. Each nucleotide is composed of one of four nitrogen-containing nucleobases **(cytosine [C], guanine [G], adenine [A] or thymine [T]),** a sugar called deoxyribose, and a phosphate group [6]. The nucleotides are joined to one another in a chain by covalent bonds [6].

DNA has an unbelievable storage capacity. Castillo states that all the information in the entire Internet could be located in a device which is lesser than unit cubic inch [7].

DNA is witnessed as the optimal medium in this regard fundamentally because instead of using 1 s and 0 s by the computer to store data, DNA consisting of **adenine, guanine, cytosine, and thymine (A, G, C, and T)** already paired into nucleotide base pairs A-T and G-C can be utilized for storing information in a form of binary code [8]. As the urgent need for high-capacity storage medium rises, DNA is considered ideal in this regard as single nucleotide can represent 2 bits of information [8]. Accordingly, **455 EB** of data can be encoded in 1 gram of single stranded DNA (ssDNA). Entire information that is produced by the world over a year can be stored in just 4 grams of DNA [8]. High memory space is offered by DNA as it is 3-dimensional (3D) by structure. DNA offers readable and reliable information for millennia, which

can be extended to almost infinity by drying and protecting from oxygen and water. One of the reasons why DNA is considered as a better storage system is that **215 petabytes (215 million gigabytes)** can be stored in just 1 gram of DNA [9].

Fig. 1: The DNA Data Storage 1

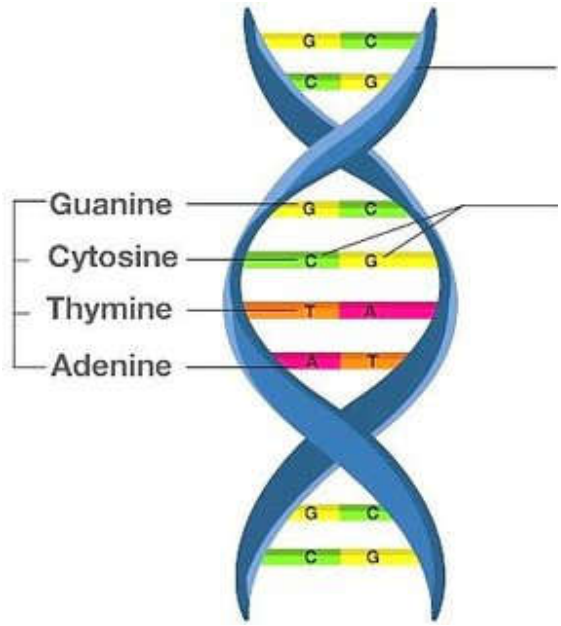


Fig. 2: The DNA Data nucleotides structure

II. BEGINNING OF THE IDEA

Mikhail Samoilovich Neiman, a Russian physicist proposed the idea of the possibility of storing and retrieving information from DNA molecules. This technology was known as MNeimON (Mikhail Neiman Oligonucleotides).

III. OVERVIEW OF DNA DATA STORAGE SYSTEM

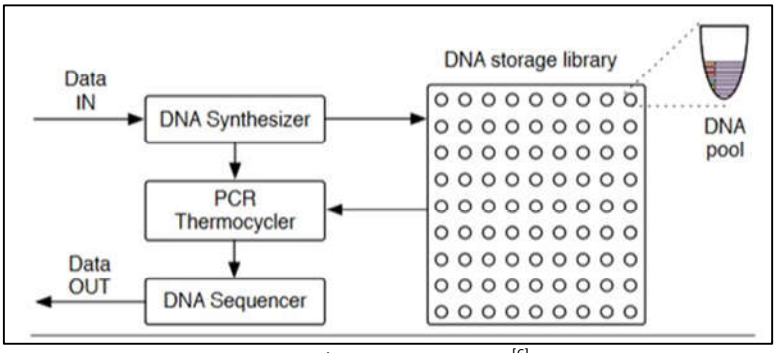


Fig.3 : The Basic Overview [6]

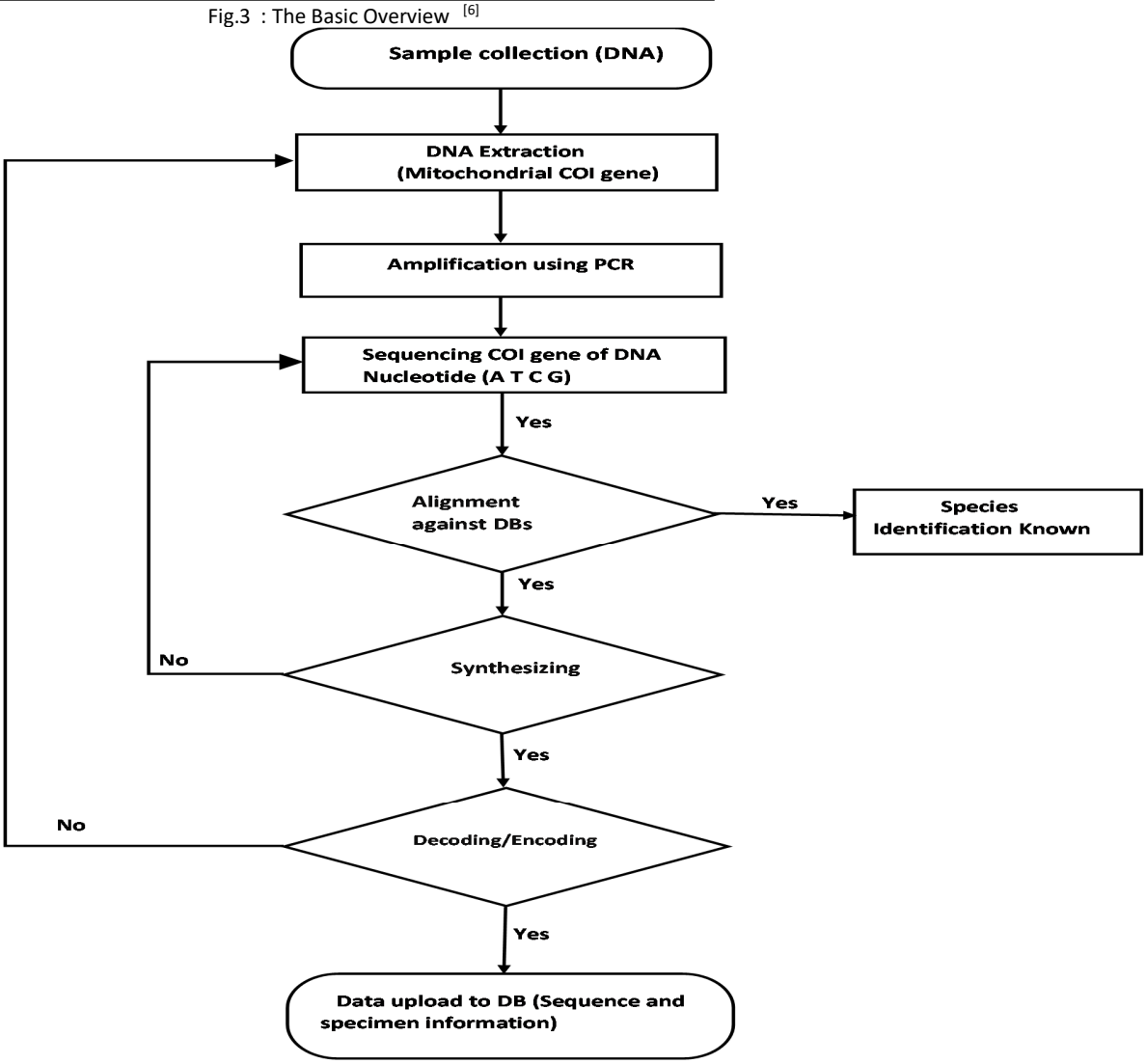


Fig. 4: Flowchart of DNA storage process

HOW DOES IT WORK

At first place, DNA storage is ultra-compact as it can be stored safely for hundreds and thousands of years in a cool, dry place. It will not degrade easily like HDDs, SDDs and other memory devices. Oligonucleotide Synthesis machines are made to upload/store information in DNA and there are highly complex machines, which can retrieve the stored data called as DNA Sequencing Machines. Oligonucleotide synthesis is the chemical synthesis of relatively short fragments of nucleic acids with defined chemical structure.

DNA molecules are long strands or sequences, which are made-up of nucleotides called **Adenine (A), Cytosine (C), Thymine (T) and Guanine (G)**. Sequences of these nucleotides are made rather than creating sequences of 0s and 1s. The way it works is assigning digital data patterns (Binary form of data) to DNA nucleotides.

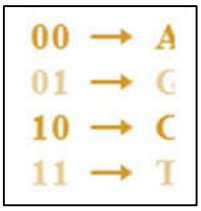
The whole process starts with the simple concept of Preparing Bits to become Atoms. ***A.***

Encoding

Binary codes as 00, 01, 10, 11 are represented using the 4 nucleotides A, T, G and C.

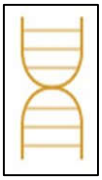
For example, 00 could be equal to A, 01 to C, 10 to T and 11 to G. Therefore, the binary for of digital data (01 11 10 00 11 11 10

11 01 00 01……) is represented biologically as C-G-T-A-G-G-T-G-C-A-C-……. Therefore, this order of nucleotides forms a DNA strand. This is how digital data is encoded.



B. Synthesis

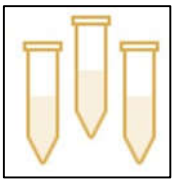
The artificial DNA should be shorter because longer DNA is chemically harder to build. Digital data can be of large sizes, but a single DNA strand can only hold around 20 bytes [5]. So, data is broken into smaller pieces and an indicator is set to the sequence so that it will ensure all the pieces of data can stay in proper order. Hence, the data is synthesized.



C. Storage

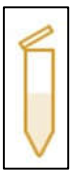
The chemical reactions used in synthesis are driven by a device, which takes the ATGC nucleotides, mixes them in a solution with some other chemicals to control reactions and order of the strands. This process also benefits us by creating backup by creating copies of each strand for another series at once.

Now the created DNA is protected from damages that is caused by light and humidity [6]. Therefore, it is dried and stored in cool place also blocking water and light.



D. Retrieval

The indicator installed during the synthesis of DNA is now used to retrieve the multiple strands of the DNA in a determined databased order [9].



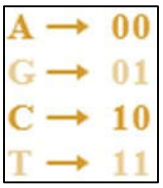
E. Sequencing

To read back the data, a machine called Sequencing Machine is used. It looks like the machines that are used for the present-day analysis of Genomic DNA in different cells [8]. As a result of this process, molecules are identified and a letter sequence is generated. This sequencing is done to obtain the final format of digital data.

Fig.5: DNA Sequencing machine

F. Decoding

The letter sequence generated by the sequencing machine is now decoded back into an ordered sequence of 0s and 1s. As for today, DNA can be destroyed during this process, but as mentioned above, many copies of each sequence are made and they now come into play. If these backup copies are, also depleted, more duplicate copies can be made easily as DNA replication is also a natural process. In this system, the whole DNA is required to be analyzed even if we need to read or access only some part of the information in it. Therefore, some special Biochemistry [10] methods are being developed and studied for accessing only required information at a much faster rate.



- Digital Data to DNA process can be figured out from this diagrammatic procedure representation.
- Binary text file → Base Encoding → DNA Encoding → DNA Fragmenting → Indexing and Storing

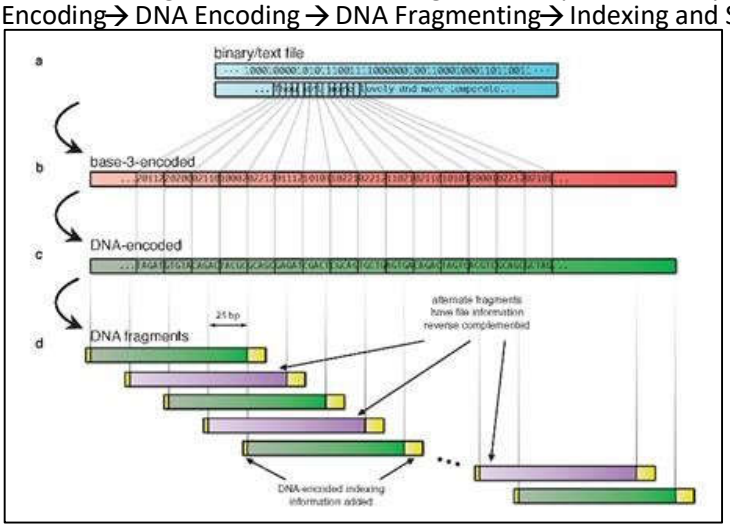


Fig. 6: The Basic Overview [6]

V. CHALLENGES

DNA Data Storage technology is still being developed and experimented even today. The researchers of different universities, companies and organizations are trying to make the whole process completely automated. The process of building DNA and accessing it by reading it at a faster rate is also being improved on every step [17]. But, as per today, the process is still relatively slow compared to Flash Drives.

Many significant changes are to be made yet trying to improve and develop the systems rapidly. The main target of this technology now is to make it faster and cheaper.

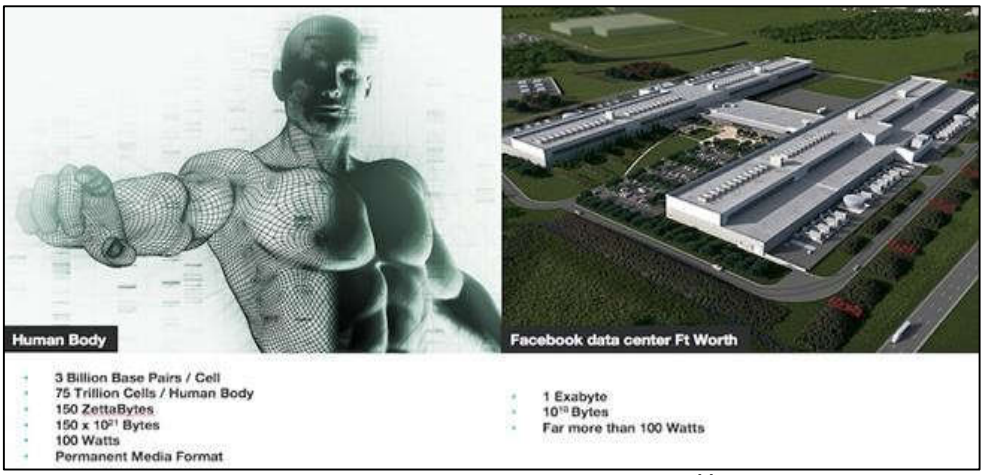


Fig. 7 : Human DNA Storage vs Data Centers [1]

Twist Bioscience is a private company, which develops and manufactures synthetic DNA. Some of the most important products that the company researches and provides are

1) Oligo Pools
2) Genes and Gene Fragments
3) Therapeutic Antibody Design and Optimization Services
4) Higher Density DNA Digital Data Storage

VI. PREPARATION OF DNA FOR DATA STORAGE

Twist Bioscience has its own innovative methods of making synthetic DNA and to get it ready for the processes of the storage mechanisms [9].

Fig. 8: Traditional Method Silicon -based synthesis [7]

As shown in the diagram, Oligo Synthesis (making synthetic DNA) is done and Perfect Gene is prepared which is ready for storing and accessing DNA. Each of this perfect gene is capable of storing 1.8 kb of data. Billions and trillions of such genes together make up a DNA strand, which can store petabytes of data.

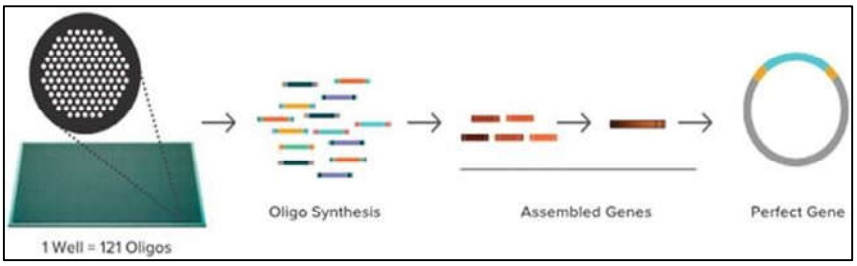


Fig. 9: Manufacture of Artificial Genes and DNA [7]

VII. MICROSOFT DNA RESEARCH

The Microsoft Corporation has started its research and experiments on DNA Data Storage too. In fact, it had the most successful start in this field of technology. Microsoft was able to store 200 megabytes of data related to literary and other articles into DNA. Microsoft also purchased 10 million strands of D

NA (Oligonucleotides) from Twist Bioscience for research and implementation of this technology and to encode digital data. Microsoft is also the company to recover or retrieve 100% of the encoded data successfully. Based on its own study, the company also re-estimated that 1 cubic millimeter DNA can store 1 Exabyte (1 billion gigabytes) of data [1].

Microsoft's digital data continuous to expand exponentially and therefore it is planning to use DNA Data Storage technology to replace its 1000's of acres of land occupied with data centers starting form one of them. It also has a plan to add DNA Data Storage to its cloud services for now. In the beginning, the speed of encoding and retrieval of data from DNA was only about 400 bytes/second. Now they are reaching speeds of 100 megabytes/second, achieving high and successful results. Microsoft had 36 Azure data centers and 8 being ready again, from which 1 data center is going to be a DNA-based data center. University of Washington is also working as a part of Microsoft [1].

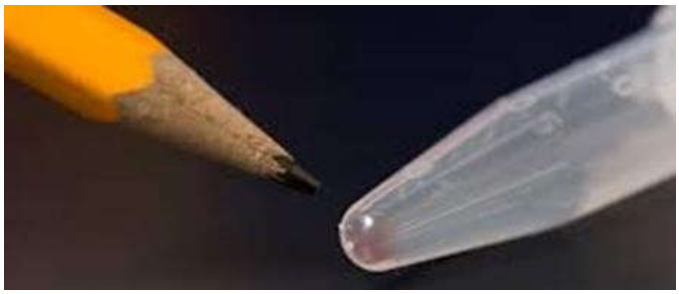


Fig. 10: Microsoft's 1 Gram of DNA [2]

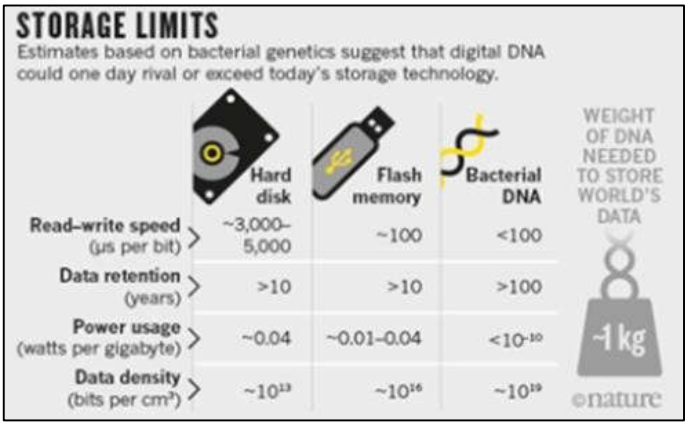


Fig. 11: DNA Data Storage Limits [5]

ADVANTAGES	*DISADVANTAGES*
Highly durable, stable and easily synthesized	***The process of copying and retrieving is a slow process***
Needs easy maintenance and information is stored for thousands of years	***Not yet fully developed and the DNA replicators can be of high costs.***
Highly reliable and 2.2 petabytes of data can be stored in 1 gram of DNA.	***Difficult to identify where the process went wrong at any point of the mechanism.***

Fig. 12: DNA Data Storage facts [3]

CONCLUSION

In this article we discussed various methods of storing digital data onto DNA. The recent achievements of Microsoft have also been discussed. Even though DNA storage has greater advantages of storing data, the cost remains a barrier. Scientists are working on several projects for minimizing the cost of artificial DNA synthesis. DNA can be used as an organic memory to store massive amounts of data. This paper also analyzes the mechanism where living organism could be used as storage devices

while identifying limitations and appropriate applicability. Challenges faced through trying to apply organic memory concepts are also discussed through this paper. Big data storage

and analytics and the way it has led to DNA computing to solve hard computational problems are also discussed here. The outcome of this study is a review article which identifies the limitations of existing encoding algorithms and proposes methods to overcome the identified limitations.

AKNOWLEDGEMENT

All Thanks and Gratitude Be to Almighty.

I would like to express my sincere gratitude to my faculty teacher **Mr. Mitul Patel** who gave me a golden opportunity to work on this paper. I'd also like to express my gratitude to my school principal wholeheartedly.

I must also thank my parent **Hajjiya Hadiza** and **Alhaji Gambo** and friends for the immense support and help during this work. Without their help, completing this project would have been very difficult.

REFERENCES

[1] A. Rosenblum, "Microsoft reports a big leap forward for DNA data storage," 2016, https://www.technologyreview.com/s/601851/

[2]Akram F, Haq I, Ali H, Laghari AT (2018) Trends to store digital data in DNA: an overview. Mol Biol Rep 45: 1479-1490.

[3] Lee Organick1, Siena Dumas Ang2,et.alRandom access in large-scale DNA data storage"2018 Nature America, Inc., part of Springer Nature. www.rte.ie/news/ireland/2018/0219/941956-dna-data/

[5]Journal for Research | Volume 03| Issue 11 | January 2018

ISSN: 2395-7549

[6] J.Davis, "Microvenus," Art Journal, vol. 55,no. 1, pp.70–74, 1996.

[7] E.Kac, "Genesis-art ofDNA," 1999, http://www.ekac.org/geninfo.html.

[8] N. Yachie, Y. Ohashi, andM. Tomita, "Stabilizing synthetic data

in the DNA of living organisms," Systems and Synthetic Biology,vol. 2, no. 1-2, pp. 19–25, 2008.

[9] C. Bancroft, T. Bowler,B.Bloom, andC. T. Clelland, "Long-term storage of information in DNA," Science, vol. 293, no. 5536, pp.1763–1765, 2001.

[10] N. Yatchie, Y.Ohashi, andM. Tomita, "Stabilizing synthetic data

in the DNA of living organisms," Systems and Synthetic Biology, vol. 2, no. 1-2, pp. 19–25, 2008.