LUPINE PUBLISHERS

Research and Reviews on Healthcare: Open Access Journal

DOI: 10.32474/RRHOAJ.2019.03.000163

ISSN: 2637-6679

Review Article

# Deoxyribonucleic Acid (DNA): The Magician in the Field of the Digital Data Storage

**Lichun Sun[1,2]*, Jun He[2], Jing Luo[3] and David H Coy[1]**

[1]*Department of Medicine, Tulane University Health Sciences Center, New Orleans, USA*

[2]*Sino US Innovative Bio-Medical Center and Hunan Beautide Pharmaceuticals, China*

[3]*Department of Health Information and Technology, EXCELth Primary Health Care, USA*

***Corresponding author:** Lichun Sun, Department of Medicine, School of Medicine, Tulane University Health Sciences Center, Sino US Innovative Bio Medical Center and Hunan Beautide Pharmaceuticals, Hunan, China

**Received:** May 20, 2019 **Published:** May 29, 2019

## Introduction

It is an explosive era of the big digital data with an exponential growth rate. The data produced in the couple years between 2015 and 2017 are more than all the preceded data produced in the entire human history. Nowadays, over 40 exabytes (EBs) of new data per day are produced all over the world. There were totally 33 zettabytes (ZBs, or 33000 EBs) of digital data worldwide in 2018. It is estimated that there will be almost 175 zettabytes (ZBs) of data in 2025. As the continuous increase of data, scientists realized that, to store these huge data is a mission impossible with current technologies and is becoming a server headache for people from academy, industry and almost all other fields in this digital universe. For instances, the data storage center built by IBM in 2011 just only has 120 petabytes (PBs) (about 0.1 EB) of the data-storing capacity. In 2013, Facebook built a large data storage center as well with only the capacity to store 1 EB of data. The current data storage spaces, technologies and approaches cannot meet the urgent requirement of storing these huge data.

### The current digital data storage

Currently, most digital data are stored on traditional magnetic and optical media such as tapes, DVDs, HDD (hard disk drive), and USB flash drives [1-4]. These media have quick retrieval times, but with very limited data storage capacity. A CD may store several hundred megabytes (MBs) of data. A flash drive can maximally store less than one hundred gigabytes (GBs) of data, with an HDD holding couple terabytes (TBs) of data. Meanwhile, these media also lack long-term durability. They are also easily damaged, resulting in data loss [3]. As the broad use of computers, smart phones, Internets and other electronic devices, we are surrounded in this digital data world. These digital data are stored on the silicon-based chips with the binary numeric system that uses only two digital numbers, or 0 and 1[5]. As reported, the microchip-grade silicon is rare in nature and will be run out in 2040 [6]. A new media or a new technology is needed to fix the coming data storage crisis. Due to its unique advantages [7], deoxyribonucleic acid (DNA) is expected to be the magic digital data storage medium to eventually solve data storage problem that have long plagued scientists worldwide.

### Deoxyribonucleic Acid (DNA) and data storage

To most of natural organisms, deoxyribonucleic acids (DNAs) are their genetic materials to pass their genetic characteristics from one generation to next generation and has double-stranded helical structures, with each strand having a backbone and a side chain. The former consists of the phosphate groups and the deoxyribose groups. The latter consists of four types of nitrogenous bases including adenine (A), cytosine (C), guanine (G), and thymine (T). Each double-stranded DNA replicate in a semi-conservative manner [2]. Each of two parental strands serves as a template for the synthesis of new daughter DNA molecules. The complementary nucleotides are added to the daughter strand via pairing with the opposite of the bases on the parental strand under the base-pairing rules of A with T and G with C. The new double-stranded DNA molecule consisting of one parental strand and one daughter strand is the same as their parent DNA. Given that DNA is highly conserved with high density, high stability, high replication efficiency, and long-term durability [2,7,8], DNA was considered as the ideal data storage medium with high data fidelity and long-term storage ability. DNA can magically store one EB of data per cubic millimeter (mm3) or 215PBs per gram of DNA. As mentioned above, an entire huge data center built by Facebook in 2013 can only archive one EB of data. And indeed, for the first time, DNA had been demonstrated to be capable of storing digital data in 1988 [9]. Data-storing DNA currently is a single-stranded type of synthetic nucleotide sequences [3]. DNA data storage is the process that encodes and decodes binary data to and from the synthesized DNA sequences. For data encoding, the data information are converted to binary codes with 0 and 1, and then encoded to DNA nucleotide sequences, with the four bases (A, C, G, T) instead of 0 and 1, further synthesize and store the DNA sequences (Figure 1) [4-10]. For data decoding, the DNA sequences stored are amplified by PCR, sequenced, decoded, and eventually converted the sequencing information into the original digital data as required (Figure 1) [2,3,8].

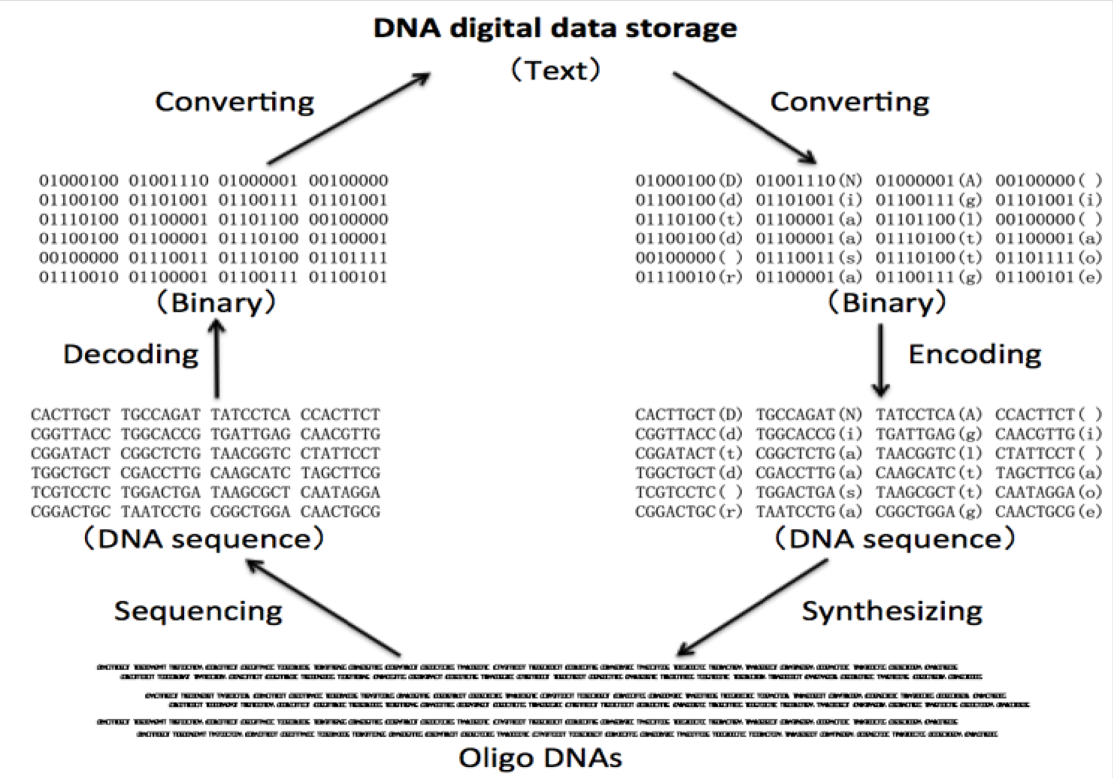
DNA digital data storage
(Text)
Converting
Converting
01000100 01001110 01000001 00100000
01100100 01101001 01100111 01101001
01110100 01100001 01101100 00100000
01100100 01100001 01110100 01100001
00100000 01110011 01110100 01101111
01110010 01100001 01100111 01100101
01000100(D) 01001110(N) 01000001(A) 00100000( )
01100100(d) 01101001(i) 01100111(g) 01101001(i)
01110100(t) 01100001(a) 01101100(l) 00100000( )
01100100(d) 01100001(a) 01110100(t) 01100001(a)
00100000( ) 01110011(s) 01110100(t) 01101111(o)
01110010(r) 01100001(a) 01100111(g) 01100101(e)
(Binary)
(Binary)
Decoding
Encoding
CACTTGCT TGCCAGAT TATCCTCA CCACTTCT
CGGTTACC TGGCACCG TGATTGAG CAACGTTG
CGGATACT CGGCTCTG TAACGGTC CTATTCCT
TGGCTGCT CGACCTTG CAAGCATC TAGCTTCG
TCGTCCTC TGGACTGA TAAGCGCT CAATAGGA
CGGACTGC TAATCCTG CGGCTGGA CAACTGCG
CACTTGCT(D) TGCCAGAT(N) TATCCTCA(A) CCACTTCT( )
CGGTTACC(d) TGGCACCG(i) TGATTGAG(g) CAACGTTG(i)
CGGATACT(t) CGGCTCTG(a) TAACGGTC(l) CTATTCCT( )
TGGCTGCT(d) CGACCTTG(a) CAAGCATC(t) TAGCTTCG(a)
TCGTCCTC( ) TGGACTGA(s) TAAGCGCT(t) CAATAGGA(o)
CGGACTGC(r) TAATCCTG(a) CGGCTGGA(g) CAACTGCG(e)
(DNA sequence)
(DNA sequence)
Sequencing
Synthesizing
Oligo DNAs

**Figure 1:** The schematic process of DNA data storage. It is to encode and decode binary data to and from the synthesized DNA sequences. For data-encoding, A text string with 24 bytes ("DNA digital data storage") is converted to binary bits that are subsequently encoded to an oligo DNA using a 1 bit per base, with purine (A, G) being assigned as 1s and pyrimidine (C, T) as 0s. The oligoDNA is chemically synthesized and the text contents are then stored as an oligo DNA fragment for the long-term storage. For data-decoding, the DNA fragment required will be amplified by PCR, sequenced and decoded to binary bits once, one day, we need to retrieve data in order to output binary data to be readable. Eventually, reading the data from DNA sequence library is to sequence the unique DNA molecules, convert the sequencing information into the original digital data as required [2,3,8]. This figure is cited from reference [21].

## The progress of DNA data storage

Compared with the traditional media, DNA could be the potential and promising medium for digital data storage [6-10]. Actually, the Soviet physicist Mikhail Neiman described his idea on the use of DNA as data storage medium in 1960s. To encode data information into DNA has been reported and demonstrated in 1986. The bio-artist Joe Davis collaborated with the scientist Dan Boyd and his coworkers to convert the artwork Microvenus into an easily recognized DNA sequences [11]. Since then, scientists made a great progress in this field. In 1999, Dr. Clelland et al. [12] use a DNA-based, doubly steganographic technique to send a DNA-encoded secret message. They encoded the message into DNA, eventually decoded and successfully got the message retrieval [12]. In 2010, Dr. Gibson et al. [13] encoded and decoded 7920 bits of data into and from oligo DNA fragments [13]. In 2012, Dr. Church and his team used a novel encoding scheme to successfully convert an entire book including about 53,400 words, 11 JPEG images, and 1 JavaScript program into a 5.27 MB bitstream, and encode them into 54,898 oligo DNA fragments [9]. Each oligo DNA has an addressing code (19 nucleotides) and flanking common sequence (22 nucleotides) for amplification and sequencing. All data were then decoded and recovered with only 10-bit errors out of 5.27 million bits [9]. In 2015, Dr. Grass et al. [14] encoded 83 kB of data into 4991 DNA fragments with each 158 nucleotides long, retrieved the information with error free, even keeping DNA fragments in silica at 70 °C high for one week [14]. In 2016, Dr. Church and coworkers stored 22MB of a MPEG compressed movie in oligoDNA fragments. They synthesized 900 000 230nt oligo DNA fragments with each 230 nucleotides long. They further retrieved them with data being error-free [15]. In 2016 again, Scientists from Microsoft and University of Washington stored the largest data (200MB) in DNA. They encoded and decoded a song "This Too Shall Pass" from a music video of the American rock band OK Go [16,17]. In 2017, Dr. Yazdi et al. [18] made a portable, random access and error-free DNA data storage system via allowing random access by storing data in gBlock codewords (long DNA strings), Reducing the cost of synthesizing DNA by compression and subsequent constrained coding, and allowing portability of the system by using error-prone nanopore sequencers [18]. In 2018, it is a key turning point for DNA to be commercially used as data storage medium. The first automated DNA data encoding/decoding device has been invented. Dr. Takahashi and the team scientists from Microsoft and University of Washington developed an end-to-end automation of DNA data storage device consisting of three core modules (a DNA preparation and sequencing module, an encoding/decoding software module, and a DNA preparation and sequencing module) (Figure 2) [19]. They stored and retrieved a 5-byte word "HELLO" (01001000(H) 01000101(E) 01001100(L) 01001100(L) 01001111(O) in binary bits) into and from a DNA sequence (GCA GAC GCC CGT ACG TAC GTT CAC CGT GCG TCT TCA CCG TGC GTC). Most significantly, this is first fully automated technology to encode, store, read and recover the message [19]. In 2019, Dr. Appuswamy and his coworkers demonstrated the ability to store structured data in synthetic DNA. They developed a tool Oligo Archive for using DNA in the database management system as the archival tier of a relational database. They used this tool to archive 12KB of TPC-H database and 2 images into DNA, perform in vitro computation, and retrieve data back, demonstrating that this system can archive and restore data using synthetic DNA, and can also exploit database knowledge for optimizing the DNA data encoding/decoding process, and can directly execute SQL operations over DNA. We are seeing the critical advances in the use of DNA as a promising data storage medium [19].

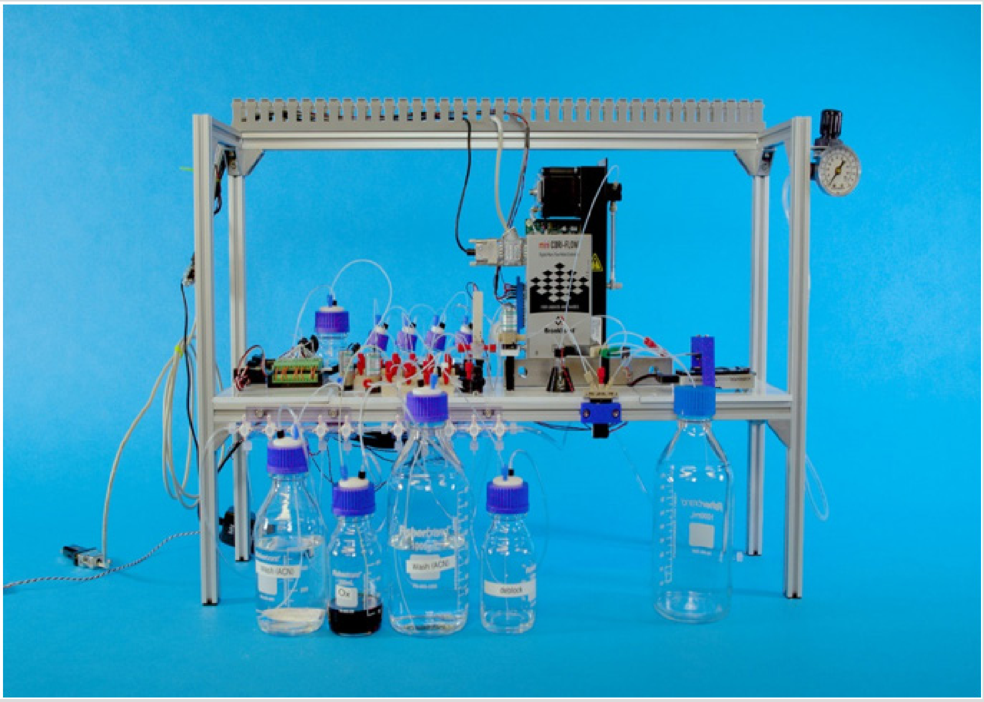

**Figure 2:** The completed end-to-end automated DNA data storage system that was made by scientists from Microsoft and University of Washington, including synthesis, storage and sequencing. It is the first fully automated DNA data storage system. This device can completely write, store and read DNA data information. It is cited from Dr. Takahashi [19].

## The challenges and the prospective of DNA data storage

However, the use of DNA as a data storage medium is still on the early stage before being commercially applied. We need to solve several challenges including high cost, low throughput, the limited access to data storage, short synthetic oligoDNA fragments, error rate in synthesis and sequencing [8-18]. For example, there is an error rate of about 1% per nucleotide in DNA synthesis and sequencing, with the occurrence of insertion, deletion, and substitution of nucleotides. Also, it is time-consuming for data to write into and retrieve from oligoDNA library. Access time for traditional media is much quicker. Access cost one millisecond for flash drive and several minutes for tape, but tens of hours for DNA data storage. Data random access is also a tough challenge in DNA data storage. In particular, the use of DNA in data storage is much more expensive than the traditional magnetic and optical media such as flash drive, hard disk drive [3-6]. This seriously hampered its commercial use. For example, to synthesize 2 MBs of data cost $7000, with another $2000 to retrieve it in 2017 [19]. It costs approximately $10,000 to write and read a 5-byte word with the fully automated DNA data storage device developed by Microsoft and University of Washington [20]. However, following with the advances in the development of DNA synthesis and sequencing technology, scientists can eventually go through these challenges and let the magic DNA become the perfect data storage medium. Currently, the handy and portable DNA sequencers are available [21]. The automated DNA data storage device has been invented. Once matured and commercially applied, these technologies can further reduce the cost of DNA sequencing and simplify retrieval of DNA information. Thus, in the near future, DNA serving as a data storage medium will be a golden opportunity in this era of big data.

## References

1. Mayer C (2016) An Epigenetics-Inspired DNA-Based Data Storage System. Angew Chem Int Ed Engl 55(37): 11144-11148.
2. Swati a, Mathuria F, Bhavani S, Malathy E, Mahadevan R, et al. (2017) A review on various encoding schemes used in digital DNA data storage. International Journal of civil engineering and Technology 8(12): 108-114.
3. Appuswamy RL, Barbry K, P Antonini M, Madderso O, Freemont P, et al. (2019) OligoArchive: Using DNA in the DBMS storage hierarchy. CIDR 2019, Biennal Conference on Innovative Data Systems Research, California, USA.
4. De Silva PY, GU Fanegada (2016) New Trends of Digital Data Storage in DNA. Biomed Res Int p. 8072463.
5. Kuang SY, G Zhu, ZL Wang (2018) Triboelectrification-Enabled Self-Powered Data Storage. Adv Sci (Weinh) 5(2): 1700658.
6. Panda DM, KA Baig, MJ Swain, A Behera D, Dash, M, et al. (2018) DNA as a digital information storage device: hope or hype? Biotech 8: 9.
7. Zakeri B, TK Lu (2015) DNA nanotechnology: new adventures for an old warhorse. Curr Opin Chem Biol 28: 9-14.
8. Organick L (2018) Random access in large-scale DNA data storage. Nat Biotechnol 36(3): 242-248.
9. Church GM, Gao Y, Kasur S (2012) Next-generation digital information storage in DNA. Science 337(6102): 1628.
10. Akram FH, I Ali H, Laghari AT (2018) Trends to store digital data in DNA: an overview. Molecular Biology Reports 45: 12.
11. Davis J (1996) Microvenus. Art Journal 55(1): 5.
12. Clelland CT, V Risca, C Bancroft (1999) Hiding messages in DNA microdots. Nature 399(6736): 533-534.
13. Gibson DG (2010) Creation of a bacterial cell controlled by a chemically synthesized genome. Science 329(5987): 52-56.
14. Grass RN (2015) Robust chemical preservation of digital information on DNA in silica with error-correcting codes. Angew Chem Int Ed Engl 54(8): 2552-2555.
15. Blawat MG, K Chen XM, Turczyk B, Invers, S, Pruitt BW, et al. (2016) Forward Error Correction for DNA Data Storage. Procedia Computer Science 80: 1011-1022.
16. Callaham J (2016) Microsoft has found a way to store 200MB of data in synthetic DNA.
17. Langston J UW (2016) Microsoft researchers break record for DNA data storage.
18. Yazdi S, Gabrys R, Milenkovic O (2017) Portable and Error-Free DNA-Based Data Storage. Sci Rep 7(1): 5011.
19. Takahashi CN (2019) Demonstration of End-to-End Automation of DNA Data Storage. Sci Rep 9(1): 4998.
20. Erlich YD Zielinski (2017) DNA Fountain enables a robust and efficient storage architecture. Science 355(6328): 950-954.
21. Sun LH, Coy DH, DNA (2019) The Digital Data Storage. Health Science J.

DOI: 10.32474/RRHOAJ.2019.03.000163

**Citation:** Lichun Sun, Jun He, Jing Luo, David H Coy. Deoxyribonucleic Acid (DNA): The Magician in the Field of the Digital Data Storage. Res & Rev Health Care Open Acc J 3(3)- 2019. RRHOAJ.MS.ID.000163. DOI: 10.32474.RRHOAJ.2019.03.000163.