Proceedings of the 2nd International Conference on Biological Engineering and Medical Science (ICBioMed 2022), Part II
DOI: 10.54254/2753-8818/4/20220505

DNA Data Storage: The Fusion of Digital and Biological Information

Xiao Li

School of Chemical Engineering and Technology, Tianjin University, Tianjin, 300350, China

3020207051@tju.edu.cn

Abstract. With the development of Internet technology, human information data grows by a huge amount. Traditional data storage media are no longer suitable for large amounts of data storage due to their inherent shortcomings, such as high power consumption, large physical size, and short storage life. DNA information storage, on the other hand, can overcome these shortcomings to a certain extent. This paper introduces the process and mechanism of DNA storage, such as the DNA synthesis method, DNA coding, DNA preservation and sequencing, the history of DNA storage, its defects, and shortcomings. This article introduces the principles and mechanisms of DNA storage, including DNA synthesis method, data storage, and DNA preservation, as well as the history of DNA storage and its shortcomings and prospects for improvement, using the comparison of traditional storage methods and DNA storage methods as an import cut.

Keywords: DNA synthesis, data encoding, DNA preservation, DNA storage history

1. Introduction

The massive growth of information data in recent decades, especially the development of Internet and mobile Internet technologies, has brought an unprecedented huge data storage crisis for human beings. In social and scientific research, traffic monitoring, high-definition scientific research images, etc. have generated massive video image type data. In social scenes, information is disseminated in the medium of photos and videos, and the trend is accelerating. According to international data[1], the total amount of human-generated data reached 33 ZB (1 ZB ≈ 109 TB) in 2018, and this number will grow to a staggering 175 ZB by 2025. while traditional magnetic, electrical, and optical traditional media have defects such as high power consumption, large size, and short life span, in contrast, DNA storage can well overcome the defects in these aspects. DNA information storage possesses high storage density and excellent stability, which makes it a reliable choice for a new storage method. This paper introduces the mechanism of DNA storage, including DNA synthesis methods, DNA data encoding, DNA preservation and sequencing, the history of DNA storage, and its defects and unresolved problems.

2. The principle mechanism of DNA storage

DNA data storage is mainly divided into two parts: synthetic writes and encoded reads. Taking audio and video data storage[2]. As an example, data storage and random reads （Arbitrary positioning to

DOI: 10.54254/2753-8818/4/20220505

target sequence position） are achieved using high-throughput synthetic oligonucleotide pools and standardized logical blocks, using multiple synthetic pools（A high diversity collection of oligonucleotides, also known as a primer pool） and amplification primers to distinguish different kinds of data and allowing multi-level reads of files and logical blocks; using group error correction coding technology（A code that can detect or correct itself at the receiving end after an error occurs during transmission. Codes that are used only to detect errors are often referred to as error detection codes） to achieve logical block（Generally used to refer to the physical device data addressing from the host's perspective. In storage devices, it is often necessary to convert the logical block address provided to the host into the corresponding physical media address where the data is actually stored） recovery; using file multiplexing technology to achieve logical block and data size matching to improve logical data block Utilization rate.

In the synthesis and writing stage, the video files are firstly encoded and converted to their common formats, with H.264/AVC (advanced video codec) encoding for video and AAC (advanced audio codec) encoding for audio. The second step is to convert the data file into a binary bit stream and group it into 35568 bytes each, then use Reed-Solomon code and low-density parity code cascade coding for each group of data to get 73440 bytes size coding block, add 24 bytes to each coding block as a unit to number the block into 30 bytes (corresponding to 120nt sequence), and finally Add a 9nt sequence to the 120nt sequence to distinguish different data blocks. The third step is to add primers to both ends of the 129nt sequence, with a total length of 169nt after addition (using 4 different sets of primers). Step 4, synthesize the sequence. Step 5, synthesis pool storage.

In the data read phase, the first step is read index establishment in parameter selection, including data block, amplification primers, and oligonucleotide synthesis pool selection. The second step is PCR selective amplification and sequencing. In the third step, second-generation high-throughput sequencing of amplification products is performed. The fourth step, is data processing, decoding, and file recovery.

Next, some important steps involved in the storage process such as the DNA synthesis method, data storage, DNA preservation, the development history of DNA storage, and the shortage of DNA storage will be described in detail.

2.1. DNA synthesis method

DNA synthesis from scratch (de novo DNA synthesis) is a technique for synthesizing DNA fragments starting from oligonucleotide strands, which does not require a DNA template. It has undergone three generations of development. The first generation of DNA synthesis uses phosphoramidite synthesis, which means that DNA is immobilized on a solid-phase carrier to complete the synthesis of DNA strands. The second generation DNA synthesis technology is chip-based, including inkjet, photochemical and electrochemical deprotection methods. The third generation DNA synthesis technology is ultra-high throughput synthesis technology, which is a semiconductor combined with the electrochemical method. Both the first and second generation use phosphoramidite chemical synthesis, while the third generation is based on the principle of enzymatic synthesis[3]. The first generation synthesis method is a column synthesis method, i.e., solid phase phosphite tris, which can be performed in a simple reactor vessel and simplifies the multi-step synthesis process compared to its previous liquid phase synthesis. It is currently the main method used for the automated production of oligonucleotides, and the synthesis process consists of four steps: deprotection, coupling reaction, cap addition reaction and oxidation reaction. However, it has disadvantages such as low synthetic throughput, high error rate[4], and high cost[5]. To improve these drawbacks, researchers developed the microarray chip for DNA synthesis, which is the prototype of the second generation synthesizer. hundreds of thousands of nucleic acid probes are fixed on the DNA chip, and these probes are not randomly distributed, but are divided into many partitions with the same sequence within each partition and different sequences between the partitions. When the sample is complemented with the nucleic acid probes on the chip, different signal intensities are presented according to the binding force,

DOI: 10.54254/2753-8818/4/20220505

and the signal can be analyzed and processed to obtain a large amount of information about the gene sequence. Microarray synthesis can not only significantly increase the throughput of nucleotide synthesis, but also significantly reduce the amount of reagents consumed for synthesis. However, it still has shortcomings, such as low yield of single nucleotides, low quality of synthesized oligonucleotides, and oligonucleotide components in the chip are too complex and cannot be separated separately. In parallel with the improvement of the phosphoramidite chemical synthesis method, the development of enzymatic synthesis methods such as terminal deoxyribonucleotidyl transferase (TdT) polymerase[6] is flourishing, which can improve the synthesis efficiency by an order of magnitude compared with the phosphoramidite synthesis method, does not require the use of toxic compounds, and has mild reaction conditions. It is promising to solve many problems of chemical synthesis methods.

2.2. DNA data storage

The principle of DNA data storage is the transformation and flow of digital information between binary code streams, quaternary base sequences, and actual DNA fragments. The process mainly includes (i) information writing, encoding the binary code stream of the information to be stored, obtaining the sequence composed of four bases, and subsequently writing the information into DNA fragments and preserving it by using DNA synthesis technology[7, 8]; (ii) information reading, sequencing the DNA fragments, identifying, assembling, error correction and decoding the sequenced information, and finally reducing it into the original digital information to obtain the original file. At present, DNA information storage modes can be divided into the following three types: "DNA hard disk", "DNA CD" and "DNA tape".

The "DNA hard disk" mode is based on high-throughput DNA chip synthesis technology and high-throughput second-generation sequencing technology. Similar to traditional hard disks, it has high-density storage potential for massive data. However, its data end-to-end reliability is far less than that of a traditional hard disk, and it needs to solve the data reliability problem of DNA as a carrier. For this reason, several information encoding methods in the field of informatics have been introduced to the method. Grass's team[9] at ETH Zurich introduced Reed-Solomon (RS) correction codes to address partial fragment loss and intra-fragment base substitution errors in oligonucleotide strand pools, and Erlich et al[10] introduced fountain codes to better adapt to massive fragmentation storage models and achieve further logical density improvements in data portions.

The main feature of "DNA CDs" is the use of longer DNA fragments, which are written by in vivo assembly of cells, and the fast and low-cost DNA replication capability of the cells themselves to make fast and homogeneous copies of data. Since DNA is preserved by model organisms with low mutation rates, "DNA CDs" can be copied with high fidelity and support the long-term reproduction of data from generation to generation. In order to increase the data storage capacity, increasing the storage capacity of single-cell data and increasing the parallel throughput are the keys to increasing the data storage capacity.

"DNA tapes" are a new model for recording information by "writing" DNA in organisms using dynamic genome engineering[11]. The "writing" consists of targeted inversions, deletions, insertions, and single base mutations of specific DNA, very similar to the process of recording information on a magnetic tape[12]. If genetic circuit design is applied to this field, biological "logic gates" can be combined with "DNA tapes" to provide records for biological cellular computation. However, this model still has many drawbacks, such as low accuracy, delayed data response, and low logical density. In addition, it is very difficult to differentiate between different colonies by adding tags, and random access is very difficult.

2.3. DNA preservation

The advantage of DNA data storage over other storage media is that the data can be preserved for a longer period of time. Currently, there are two main ways of artificially preserving DNA: the extracellular preservation method and the intracellular preservation method.

DOI: 10.54254/2753-8818/4/20220505

DNA molecules are preserved in solution, which has the advantage of supporting random indexing and multiple information reads[13-15]. In order to solve the problem of degradation of DNA samples during transportation, storage, and handling, DNA molecules are usually preserved after lyophilization treatment. This approach is suitable for the long-term stable preservation of samples and allows for rapid and complete recovery of samples. In recent years, researchers at the Swiss Federal Institute of Technology have found that alkaline earth metals can stabilize dry powdered DNA functionality, significantly enhancing DNA stability even at high DNA loadings and 50% relative humidity, facilitating random access to data and information readout.

Intracellular preservation methods relative to in vitro DNA preservation methods can provide random access pathways and real-time recording of biological events using efficient DNA replication, proofreading, and DNA repair mechanisms within cells[13, 16]. The most commonly used writing tools for DNA data storage are DNA targeting and modifying enzymes. And depending on the writing mechanism, the currently reported research directions are broadly divided into two types. The first one is intracellular data storage using recombinases, which can be stored at a specific genomic location[17, 18]. The other one is DNA data storage using the CRISPR-Cas system[19, 20] . It is not only capable of recording cellular genealogy information in large quantities but can also record analog signals by coupling Cas9 expression to cellular signals, for example, non-binary gene mutations can be recorded into DNA[21].

3. History of the development of DNA storage

DNA information storage has a history of more than 50 years[9, 10, 13, 14, 22-33]. in the 1960s, computer scientists Wiene[22] and Neiman[22] first introduced the concept of DNA data storage "genetic memory". in 1995, Professor Baum of Princeton University proposed to build a DNA molecule-based mass database storage system[23]. In 1996, Davis[24] conducted the "Microvenus" experiment to validate the concept of DNA data storage by writing 35bit black and white icons into an 18bp DNA sequence. Clelland's[25] team developed a DNA steganography technique that again proved the DNA data storage concept. In the decade 2001-2010, in vivo DNA storage was dramatically improved in terms of encoding methods and storage capacity[14, 26-29]. 2012, Church[30] et al. stored a book (650KB) in DNA, and in 2013 Goldman[34] and colleagues achieved 720KB in DNA data for high-capacity storage. In 2015 and 2016 Grass[9] et al. and Blawet[31] et al. achieved high-capacity storage in synthetic DNA with error-free retrieval, which became an important node in the field of DNA information storage. 2017, Erlich's[10] team developed the "fountain code " in DNA storage to maximize the data storage capacity of DNA; in the same year, Shipman[32] used the CRISPER-Cas system to write an image and a video file to the E. coli genome. in 2018, the team of Organick[13] wrote over 200 MB of massive data to DNA. In 2020, Qi[33] et al. of Tianjin University used a hybrid culture system to deposit 445KB of data into bacteria.

4. The shortcomings and defects of DNA storage

DNA synthesis is the first step, and the main reverse of its development has further cost reduction, increasing read/write speed, and achieving perfect integration with existing information systems. In terms of cost, the commercial synthesis price of oligonucleotide pools is about $0.002/base, which translates to $0.001/bit (about 8.6 × 106 $/GB)[35] , and the writing cost is high, about 108 times higher than that of hard disks. Improvements can be made by optimizing the synthesis reaction, improving the chip structure, replacing cheap consumables, and optimizing the number of reagents dispensed. In terms of reading and writing speed, DNA information reading relies on sequencing technology, which is slower compared to optical, electrical, and magnetic storage. The read speed based on second-generation sequencing is limited by chemical reactions, and it is difficult to significantly reduce the reaction time, which can meet the demand for large-scale cold data reading by further increasing the throughput; the data reading based on third-generation nanopore sequencing has a greater potential to improve the single-well reading speed. In terms of applications, DNA storage is

DOI: 10.54254/2753-8818/4/20220505

expected to take the lead in cold data storage. the process of integrating DNA as a new medium into modern storage systems is also the process of evolution of information storage systems.

In DNA data coding, the combination of achieving high-density storage and effective data error correction mechanism, random access, data encryption, and erasure are the focus of further research. High code rate RS codes, alternate embedded ARS sequences, and scale-variable combinatorial methods can be used to achieve fast localization of overlapping groups, RS code error correction, and deletion decoding[3]. To avoid reading the entire storage system, indexing methods need to be designed to extract specific DNA sequences and achieve random access. Information encryption relies on fixed biomolecular reaction patterns, and its security will be seriously threatened once adversaries discover relevant decryption methods; meanwhile, the good chemical stability of DNA molecules poses new challenges for data erasure of DNA storage.

In DNA preservation, there is a need to consider the energy consumption problem of large-scale DNA information storage, the stable preservation of DNA on long-time scales, and the realization of physical isolation of data.

5. Conclusion

In late 2018, the National Institute of Standards and Technology (NIST), the International Semiconductor Research Alliance (ISRA), and the U.S. Advanced Research Projects Agency (ARPA) jointly released In May 2021, the Ministry of Science and Technology of China released the "14th Five-Year Plan" National Key Research and Development Program "Bioinformatics Fusion" (BT-IT Fusion) key project application guidelines[36]. This represents the top-level recognition of the future disruptive convergence technology represented by DNA storage in two major economies in the world. Compared with traditional storage methods, DNA storage has lower storage consumption, higher storage density, longer retention time, and smaller storage volume. However, at the same time, DNA storage also has its corresponding defects, such as the inconvenience of sequencing with special sequences, the relatively slow speed of DNA synthesis, the imperfect error correction mechanism of data encoding, and the urgent need to strengthen the ability of data restoration after sequencing. Through the development of the third-generation DNA synthesis technology and the improvement of the second-generation DNA synthesis technology, we can significantly reduce the time and material consumption of DNA synthesis; combine more data encoding mechanisms with DNA sequences to realize the simultaneous improvement of DNA encoding density and error correction ability; explore the preservation of DNA in multiple ways in multiple extracellular and intracellular media to extend the DNA preservation time scale and enhance its preservation The challenges will be solved gradually, and DNA data storage will gradually move from the laboratory to industrialization, bringing a more convenient and faster information storage experience for human beings.

References

[1] Y. Yang and C.H. Fan, Synthetic biology. 2(3): p. 305-308 (2021).
[2] CHEN, W., et al., SCIENTIA SINICA Vitae. 50(1674-7232): p. 81 (2020).
[3] D.M. Chen, et al., Synthetic biology. 2(3): p. 399-411 (2021).
[4] Kosuri, S. and G.M. Church, Nature Methods. 11(5): p. 499-507 (2014).
[5] Saini, N., et al., Nature. 502(7471): p. 389-392 (2013).
[6] Song, X.-P., et al., Chemistry & biodiversity. 9(12): p. 2685-2700 (2012).
[7] T.Y. Zhou, Y. Luo, and X.Y. Jiang, Synthetic biology. 2(3): p. 371-383 (2021).
[8] Lim, C.K., et al., Trends Biotechnol. 39(10): p. 990-1003 (2021).
[9] Grass, R.N., et al., Angewandte Chemie International Edition. 54(8): p. 2552-2555 (2015).
[10] Erlich, Y. and D. Zielinski, Science. 355(6328): p. 950-954 (2017).
[11] Ausländer, S. and M. Fussenegger, Science. 346(6211): p. 813-814 (2014).
[12] Farzadfard, F. and T.K. Lu, Science. 361(6405): p. 870-875 (2018).
[13] Organick, L., et al., Nature Biotechnology. 36(3): p. 242-248 (2018).
[14] Yachie, N., et al., Biotechnol Prog. 23(2): p. 501-5 (2007).

DOI: 10.54254/2753-8818/4/20220505

[15] Tabatabaei Yazdi, S.M.H., et al., Scientific Reports. 5(1): p. 14138 (2015).
[16] Sheth, R.U. and H.H. Wang, Nat Rev Genet. 19(11): p. 718-732 (2018).
[17] Bibikova, M., et al., Science. 300(5620): p. 764 (2003).
[18] Wirth, D., et al., Curr Opin Biotechnol. 18(5): p. 411-9 (2007).
[19] Marraffini, L.A., Nature. 526(7571): p. 55-61 (2015).
[20] Xie, Z.X., et al., G3 (Bethesda). 8(1): p. 173-183 (2018).
[21] Kalhor, R., P. Mali, and G.M. Church, Nature Methods. 14(2): p. 195-200 (2017).
[22] M.Y. Gao, et al., Synthetic biology. 2(3): p. 384-398 (2021).
[23] Baum, E.B., Science. 268(5210): p. 583-585 (1995).
[24] Davis, J., Art Journal. 55(1): p. 70-74 (1996).
[25] Clelland, C.T., V. Risca, and C. Bancroft, Nature. 399(6736): p. 533-534 (1999).
[26] Ailenberg, M. and O. Rotstein, Biotechniques. 47(3): p. 747-54 (2009).
[27] Bancroft, C., et al., Science. 293(5536): p. 1763-1765 (2001).
[28] Gibson, D.G., et al., Science. 329(5987): p. 52-6 (2010).
[29] Wong, P.C., K.K. Wong, and H. Foote, Commun. ACM. 46: p. 95-98 (2003).
[30] Church, G.M., Y. Gao, and S. Kosuri, Science. 337(6102): p. 1628-1628 (2012).
[31] Blawat, M., et al., Procedia Computer Science. 80: p. 1011-1022 (2016).
[32] Shipman, S.L., et al., Nature. 547(7663): p. 345-349 (2017).
[33] Hao, M., et al., Commun Biol. 3(1): p. 416 (2020).
[34] Goldman, N., et al., Nature. 494(7435): p. 77-80 (2013).
[35] Zhirnov, V.V. and D. Rasic, 2018 Semiconductor Synthetic Biology Roadmap. 2018.
[36] L. Qian, et al., Synthetic biology. 2(3): p. 303-304 (2021).