

Application of gene editing technology to DNA digital data storage

Jiao Jiang *

College of Life Sciences, South China Agricultural University, Guangzhou, China

* Corresponding author: alieu72555@student.napavalley.edu

Abstract. While the archival digital storage industry is approaching its physical limits, demand is increasing significantly, so alternatives are emerging. The modern world is in dire need of durable, scalable and economical alternative storage media. Deoxyribonucleic acid (DNA), a promising storage medium, offers superior information durability, capacity and energy consumption, making it a promising candidate for long-term data storage. However, the design and realization of DNA digital data storage face many problems, but gene editing technology, as a technology that makes modifications to genes directly from the molecular level, provides a breakthrough in solving these problems. In this paper, I show some methods for designing DNA digital data storage based on gene editing technology. The method utilizes gene editing technology to modify DNA molecules to improve their storage capacity and stability. At the same time, this paper also introduces the application cases of gene editing technology in DNA bio storage devices and looks forward to its future development.

Keywords: Data storage, DNA digital data storage, gene editing.

1. Introduction

Data storage is an indispensable part of modern society. With the advent of the digital age, the amount of data people needs to store is increasing, and traditional data storage devices are struggling to meet this demand. Consequently, scientists have begun to seek new data storage technologies. DNA-based biological storage devices, due to their high density, high capacity, and long-term storage capabilities, have received widespread attention [1].

However, the design and implementation of DNA-based biological storage devices face numerous challenges, such as the stability and storage capacity of DNA molecules. To resolve these issues, scientists have begun to explore the application of gene-editing technology in DNA-based biological storage devices.

Gene-editing technology is a method that can modify DNA sequences, which can be reformed by cutting, inserting, and replacing DNA molecules. The utilization of gene-editing technology can enhance the storage capacity and stability of DNA molecules, providing a novel approach to the design and implementation of DNA-based biological storage devices [1].

This paper will introduce the design methods of DNA digital data storage based on gene-editing technology and discuss application cases of gene-editing technology in DNA digital data storage. Finally, we will forecast the future development of gene-editing technology in DNA digital data storage.

2. DNA digital data storage

2.1. Definition

Digital data storage refers to the process of storing and preserving digital information in a format that can be retrieved and utilized in the future. This is crucial in the modern world, where vast amounts of data are generated daily, ranging from text documents and images to videos and sensor data. Various types of digital data storage solutions have been developed to cater to different needs and requirements. These storage types can be broadly categorized into four main categories: magnetic storage, optical storage, solid-state storage, and emerging storage technologies. Each type has distinct characteristics that make it suitable for specific applications (Table 1).

Table 1. the characteristics of each digital data storage type

Storage Type	Characteristics
Magnetic Storage	Uses magnetic fields to encode data on a metal surface.
	Offers relatively high capacity and low cost per GB.
	Examples include hard disk drives (HDDs) and tape.
Optical Storage	Utilizes laser technology to read and write data on optical discs.
	Provides good durability and long-term data retention.
	Examples include CDs, DVDs, and Blu-ray discs.
Solid-State Storage	Employs NAND flash memory to store data electronically.
	Offers fast access times and high reliability.
	Examples include USB drives, SSDs, and memory cards.
Emerging Storage Technologies	Encompasses new technologies like DNA digital data storage.
	Aim to address the limitations of traditional storage.
	Examples include holographic storage and 3D XPoint.

DNA digital data storage represents a groundbreaking approach to storing digital information by leveraging the inherent properties of DNA molecules. DNA, famously known as the genetic blueprint of life, exhibits several remarkable characteristics that make it an attractive candidate for data storage.

2.1.1. High density

DNA possesses an exceptional storage density, far surpassing conventional storage media. It is estimated that a single gram of DNA could potentially store several terabytes of data, making it an exceptionally compact storage solution.

2.1.2. Long-term stability

DNA is renowned for its long-term stability. Under optimal conditions, it can preserve data for thousands of years, making it an ideal choice for archival purposes. This longevity far exceeds that of many traditional storage mediums.

2.1.3. Low energy consumption

Unlike electronic storage, which requires continuous power to maintain data integrity, DNA storage does not rely on constant energy consumption, contributing to its sustainability and cost-effectiveness over time.

2.1.4. Resistance to obsolescence

DNA storage is inherently resistant to technological obsolescence. As long as the genetic code remains understood, data stored in DNA can be retrieved and decoded regardless of future technological advancements.

2.2. Design

The design of a DNA-based biological storage device primarily includes the following steps: choose the DNA molecule, encode the data into a DNA sequence, synthesize the DNA sequence into a DNA molecule, and store the DNA molecule in a biological body. The selection of DNA molecules and the encoding of DNA sequences are crucial to the performance of the DNA-based biological storage devices [2].

2.2.1. Choose the DNA molecule

Selecting the right DNA molecules for designing a DNA-based biological storage device is a critical step in ensuring the success of the storage system. Several factors need to be considered when making this choice, including stability, storage capacity, and compatibility with biological systems.

DNA stability is crucial because it ensures the long-term integrity of the stored information. DNA can degrade over time due to various factors, such as exposure to UV radiation, heat, or enzymatic degradation [3]. Circular DNA and linear DNA can have different stability profiles: Circular DNA

molecules, such as plasmids or circular genomes of some viruses, can be relatively stable. Their closed-loop structure can protect them from certain types of enzymatic degradation. However, they may still be susceptible to other factors, like UV radiation or chemical damage. Linear DNA, which resembles the genomic DNA found in most organisms, may be less stable in certain conditions compared to circular DNA. It can be more susceptible to enzymatic degradation due to exposed ends. However, advancements in DNA preservation techniques, such as using synthetic caps to protect linear DNA ends, can enhance its stability.

The choice of DNA molecule should also consider the storage capacity required for the specific application. DNA has an exceptionally high information density, making it a promising medium for data storage. The storage capacity depends on the length of the DNA molecule and the encoding scheme used. Modern DNA synthesis techniques allow for the creation of custom DNA sequences of varying lengths [3].

If the DNA-based storage device will be used within a biological context, compatibility with biological systems is essential. This includes ensuring that the chosen DNA molecules do not interfere with the normal functioning of living organisms or cause adverse reactions. Additionally, the DNA sequences should not contain elements that could trigger immune responses or other biological reactions.

The encoding and retrieval methods for the stored data should also be considered. Different DNA sequences may be more or less suitable for specific encoding and decoding techniques, such as DNA barcoding or sequencing.

To enhance the reliability of DNA-based storage, error correction mechanisms should be implemented. This involves using redundant information and error-checking codes to detect and correct any errors that may occur during storage or retrieval.

The cost of synthesizing and manipulating DNA molecules should be factored in, as it can significantly impact the feasibility of a DNA-based storage system. Advances in DNA synthesis technologies may influence the choice of DNA molecules.

In summary, selecting the right DNA molecules for a DNA-based biological storage device involves a careful evaluation of stability, storage capacity, compatibility with biological systems, encoding and retrieval methods, error correction, and cost considerations.

2.2.2. Encode the data into a DNA sequence

Encoding the data into a DNA sequence is a pivotal and intricate phase in the development of a DNA-based biological storage device. This process essentially serves as the bridge between digital information and the biological realm. Two predominant encoding methodologies, direct encoding and indirect encoding, are frequently employed in this context [4].

Direct encoding represents the straightforward and immediate transformation of raw data into a DNA sequence. This entails the conversion of digital information, typically stored as binary code, into a DNA sequence composed of adenine (A), thymine (T), cytosine (C), and guanine (G). In essence, direct encoding operates at the core level, translating the fundamental binary units of data into the genetic alphabet of DNA [4].

On the other hand, indirect encoding follows a more convoluted path. Initially, data is converted into an intermediary form, such as text or image, before being further translated into a DNA sequence. This method introduces an additional layer of abstraction, often facilitating the representation of complex or structured data, which may not be as straightforward to directly encode. Indirect encoding can be likened to a two-step translation process, wherein the original data is first transcribed into a more human-readable or interpretable format, and subsequently, this intermediary representation is converted into a DNA sequence.

In the realm of DNA-based data storage, the choice between direct and indirect encoding depends on a multitude of factors, including the nature of the data, the efficiency of encoding and decoding processes, and the requirements of the specific application. Both methods possess their unique advantages and challenges, making them valuable tools in the endeavor to harness DNA as a robust and compact storage medium for digital information [5].

2.2.3. Encode the data into a DNA sequence

Synthesizing the DNA sequence into a DNA molecule is another key step in the design of a DNA-based biological storage device. Scientists have developed many DNA sequence synthesis techniques, such as chemical synthesis and PCR amplification. During the synthesis of the DNA sequence into a DNA molecule, quality control and purification are also necessary to ensure its quality and purity [5].

2.2.4. Store the DNA molecule in a biological body

Storing the DNA molecule in a biological body is the final step in the design of a DNA-based biological storage device. The synthetic DNA molecules are introduced into the cells of the chosen host organism. This is often done through techniques such as microinjection or viral vectors, depending on the organism. The goal is to integrate the encoded DNA sequences into the host's genome so that they become a permanent part of its genetic material. Once integrated, the DNA sequences are stored within the host organism's cells, specifically within its DNA. These sequences are replicated and passed on as the organism grows and reproduces. This natural replication process ensures long-term storage of the data.

2.3. Application and the meaning of DNA digital data storage

DNA digital data storage is instrumental in securely preserving data for future generations. It is an ideal choice for archiving valuable information, such as scientific research data, historical documents, and cultural heritage, ensuring their integrity over extended periods. Moreover, DNA storage serves as a robust backup solution for critical digital data. It offers redundancy, safeguarding against digital storage failures, data loss, or corruption [6]. In addition, in scenarios where space and weight limitations are critical, such as space exploration missions or compact data centers, DNA storage's high data density becomes advantageous, allowing more data to be stored in less physical space. Additionally, DNA storage can facilitate secure data distribution, particularly when traditional digital transmission methods are insecure or impractical. This can be valuable for sharing sensitive information or large datasets. Lastly, DNA data storage is a valuable component of cold storage solutions where data is stored offline and infrequently accessed [6]. It ensures data remains intact and available when needed, even after prolonged periods.

The significance of DNA digital data storage lies in its potential to address several critical challenges associated with traditional data storage methods.

2.3.1. Unprecedented data density

DNA can store vast amounts of data in an incredibly compact space. This high data density is crucial in a world where the volume of digital information is growing exponentially, and efficient storage solutions are needed to accommodate this growth.

2.3.2. Long-term data preservation

DNA has the potential to preserve data for thousands of years under the right conditions. This makes it an ideal choice for archiving critical information that needs to be stored and accessible for future generations, such as scientific research data, historical records, and cultural heritage.

2.3.3. Durable and stable storage

DNA is highly stable and resistant to environmental factors that can degrade traditional storage media. It can withstand extreme conditions, including temperature variations and radiation, making it suitable for scenarios where data integrity is paramount.

2.3.4. Reduced environmental impact

DNA storage can be more environmentally friendly than traditional data centers that require vast amounts of energy for cooling and maintenance. The compact nature of DNA storage reduces the physical infrastructure needed, potentially reducing the carbon footprint associated with data storage.

2.3.5. Data security

DNA storage can enhance data security, as it is inherently difficult for unauthorized access. This can be particularly valuable for safeguarding sensitive information and protecting against cyber threats.

2.3.6. Space-efficient solutions

In situations where physical space is limited or costly, such as in space exploration missions or compact data centers, DNA storage's efficiency in terms of space utilization becomes a significant advantage.

2.3.7. Innovative research opportunities

The development of DNA digital data storage opens up new frontiers in scientific research and technology development. It fosters interdisciplinary collaboration between biology, chemistry, and computer science, driving innovation in various fields.

2.4. Current shortage

However, it is essential to recognize that DNA digital data storage presents its own set of challenges (Table 2).

Table 2. Current shortage

Challenge/Issue	Description
Cost	High costs associated with DNA synthesis and sequencing technologies need significant reduction for DNA storage to become economically viable for large-scale data storage.
Speed	Reading and writing data to DNA is a time-consuming process compared to traditional digital storage technologies like hard drives and SSDs. Improving speed is crucial.
Error Rates	Addressing error rates during the writing and reading processes is important to ensure data integrity over long-term storage.
Scalability	Demonstrating scalability to efficiently store large datasets is essential to compete with existing data storage solutions.
Infrastructure	Building the necessary infrastructure and expertise for DNA data storage requires significant investment, which can be challenging for many organizations.
Environmental Concerns	DNA synthesis requires resources and can have environmental implications, necessitating environmentally sustainable approaches for DNA storage.

3. Background of gene editing

Gene editing technology can be used to correct errors that may occur during the writing or reading of DNA data, influence the design of DNA sequences to optimize data storage, add security features for encoding data, and enhance the efficiency and accuracy of the DNA synthesis process.

3.1. Current methods of gene editing

Gene editing methods are molecular tools that allow precise manipulation of DNA sequence and function. The most widely used gene editing method is CRISPR-Cas systems, which are RNA-guided endonucleases that can be programmed to target specific DNA sequences [7]. CRISPR-Cas systems have been adapted for use in DNA digital data storage by encoding binary data as a series of DNA sequences that are recognized and cleaved by specific CRISPR-Cas systems [8]. The major advantage of CRISPR-Cas systems is their high specificity and efficiency in targeting DNA sequences, allowing for precise control of DNA sequence and function.

Another gene editing method that has gained attention in DNA digital data storage is base editing, which enables the direct conversion of one DNA base to another without cleaving the DNA backbone [9]. Base editing offers the potential for more efficient and accurate encoding of digital data into DNA sequences by eliminating the need for error-prone enzymatic steps involved in traditional DNA

synthesis methods. However, base editing is currently limited by its low editing efficiency and specificity, as well as its inability to introduce novel DNA sequences [10].

Recently, prime editing has emerged as a new gene editing method that can be used for precise sequence modifications without cleaving the DNA backbone. Prime editing uses a fusion protein of Cas9 and reverse transcriptase to introduce new DNA sequences at specific locations in the genome. Prime editing has the potential to overcome the limitations of traditional gene editing methods by enabling precise and efficient introduction of novel DNA sequences. However, prime editing is still in the early stages of development and requires further optimization for use in DNA digital data storage.

3.2. The application of gene editing in biostorage

In 2015, scientists successfully applied CRISPR-cas9 genome editing technology and overlap extension PCR technology to modify the content of DNA storage files by using the prefix coding addressing technology of unrelated primer sequence sets. The mainstream DNA editing technologies currently used for DNA storage include CRISPR-cas9 genome editing technology and overlap extension PCR technology [11]. The storage capacity and stability of DNA molecules can be improved by using gene editing technology, which provides new ideas for the design and realization of DNA biological storage devices.

3.2.1. Using gene editing technology to increase the storage capacity and the stability of DNA molecules

Using gene editing technology can change the structure and sequence of DNA molecules, thereby increasing its storage capacity and stability. For example, scientists have used gene editing technology to modify the sequence of DNA molecules into a structure similar to DNA hexagons, thereby increasing its storage density and stability [11]. In addition, scientists have also used gene editing technology to modify the sequence of DNA molecules into structures similar to DNA nanorods, thereby increasing their storage capacity and stability.

3.2.2. Application cases of gene editing technology in DNA digital data storage

At present, there are already many application cases of DNA bio storage devices based on gene editing technology.

For example, scientists have used gene editing technology to encode DNA sequences into silk protein sequences, synthesize them into DNA molecules, and then store the DNA molecules in silkworms. This DNA biological storage device has the advantages of high density, high capacity, and long-term storage.

To design robust cell storage, scientists have gained insights from the immune mechanism of living cells, especially CRISPR. The storage capacity has been significantly improved by using the directed spacer acquisition in the CRISPR array to record temporal information in the bacterial genome. This scheme allows any information to be encoded into synthetic oligonucleotides of a pre-set length and introduced into cells overexpressing the Cas1-Cas2 protein complex through electroporation. These synthetic oligonucleotides are then integrated as prototype spacers into expanded CRISPR arrays. As new spacers are always added to the front end of the array, the order of integrated spacers can reflect the historical timeline of the acquisition events. Since spacer acquisition occurs randomly at the single-cell level, reading the stored content depends on the analysis of the CRISPR array in the cell population [12].

4. Summary

In this article, by summarizing DNA digital data storage and gene editing technology, focuses on the gene editing technology in DNA digital data storage applications possible, for its storage design provides new ideas. However, there are still many challenges facing DNA digital data storage, including the high cost of synthesis and sequencing; the long-term stability of DNA as a storage

medium is affected by environmental factors, such as fluctuations in temperature and humidity, which require appropriate environments. With the continuous development of gene editing technology, its application in DNA biological storage devices will continue to expand. In the future, gene editing technology may be used to develop more efficient and stable DNA biostorage devices. At the same time, gene editing technology may also be used to develop other types of biological storage devices, such as using RNA molecules to store data.

References

- [1] TENG Yue, YANG Shan, LI Jin-Yu, et al. Principle and Progress of DNA Data Storage [J]. *Progress in Biochemistry and Biophysics*, 48 (05): 494 - 504, 2021.
- [2] Liu Y, Ren Y, Li J, et al. In vivo processing of digital information molecularly with targeted specificity and robust reliability. *Sci Adv.*; 8 (31): eabo7415, 2022.
- [3] Meiser, L.C., Nguyen, B.H., Chen, Y.-J., Nivala, J., Strauss, K., Ceze, L., Grass, R. N. Synthetic DNA applications in information technology. *Nature Communications*, 2022.
- [4] Shipman SL, Nivala J, Macklis JD, Church GM. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature*. 547 (7663): 345 - 349, 2017.
- [5] Yoo E, Choe D, Shin J, Cho S, Cho BK. Mini review: Enzyme-based DNA synthesis and selective retrieval for data storage. *Comput Struct Biotechnol J.*; 19: 2468 - 2476, 2021.
- [6] De Silva PY, Ganegoda GU. New Trends of Digital Data Storage in DNA. *Biomed Res Int.*; 2016: 8072463, 2016.
- [7] Doudna JA, Charpentier E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science*. 346 (6213): 1258096, 2014.
- [8] Goldman, N., Bertone, P., Chen, S. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* 494, 77 – 80, 2013.
- [9] Komor, A., Kim, Y., Packer, M. et al. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533, 420 – 424, 2016.
- [10] Anzalone AV, Randolph PB, Davis JR, et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*; 576 (7785): 149 - 157, 2019.
- [11] Sadremomtaz, A., Glass, R.F., Guerrero, J.E., Lajeunesse, D.R., Josephs, E.A., Zadegan, R. Digital data storage on DNA tape using CRISPR base editors, 2023.
- [12] Sheth, R.U., Wang, H.H. DNA-based memory devices for recording cellular events. *Nature Reviews Genetics* 19, 718 – 732, 2018.