Materials Today Bio 15 (2022) 100306

Contents lists available at ScienceDirect

Materials Today Bio

journal homepage: www.journals.elsevier.com/materials-today-bio

Design considerations for advancing data storage with synthetic DNA for long-term archiving

Chisom Ezekannagha [a,*], Anke Becker [b], Dominik Heider [a], Georges Hattab [a]

[a] Department of Mathematics and Computer Science, Philipps-Universität Marburg, Hans-Meerwein-Str. 6, D-35043, Marburg, Germany
[b] Center for Synthetic Microbiology (SYNMIKRO), Philipps-Universität Marburg, Karl-von-Frisch-Str. 14, D-35043, Marburg, Germany

ARTICLE INFO

Keywords:
Data
Synthetic DNA
Encoding
Data storage
Design considerations

ABSTRACT

Deoxyribonucleic acid (DNA) is increasingly emerging as a serious medium for long-term archival data storage because of its remarkable high-capacity, high-storage-density characteristics and its lasting ability to store data for thousands of years. Various encoding algorithms are generally required to store digital information in DNA and to maintain data integrity. Indeed, since DNA is the information carrier, its performance under different processing and storage conditions significantly impacts the capabilities of the data storage system. Therefore, the design of a DNA storage system must meet specific design considerations to be less error-prone, robust and reliable. In this work, we summarize the general processes and technologies employed when using synthetic DNA as a storage medium. We also share the design considerations for sustainable engineering to include viability. We expect this work to provide insight into how sustainable design can be used to develop an efficient and robust synthetic DNA-based storage system for long-term archiving.

1. Introduction

The digital data generated in modern society is rapidly increasing [1]. Various conventional storage technologies have been developed, including those based on magnetic, optical, and solid-state devices (e.g., tape, Blu-ray disc, flash memory), respectively [2]. Tape technology, primarily used for long-term archiving, has seen significant improvements in density, with tape memories reaching 330 terabytes [3]. However, the sheer volume of data stored on these devices will soon exceed the amount of daily data in modern society. It is estimated that three quarters of humanity will be connected and that the amount of data generated will reach 5000 zettabytes ($\sim 10^{24}$) by 2040 [4]. Storing zettabytes of data requires the use of significant physical space. However, storage density is only one feature of long-term archiving: long-term costs and durability are critical. The retention time of most storage media is short. For example, magnetic tapes have a retention time of about 10–30 years. In addition, physical data centers for centralized storage are constantly being built, although this storage is a large consumer of electricity and requires significant cooling. Although the technologies used in traditional storage media have advanced rapidly, current storage media are all approaching their density limits. Consequently, there is a growing need for storage media with significantly improved information density, durability and energy costs.

1.1. DNA as an alternative storage medium

New alternative methods of data storage, such as molecular or atomic media, have been considered sustainable because they increase storage capacity, reduce energy cost, and increase durability [2]. As shown in Fig. 1, data storage using synthetic DNA has received a lot of attention and is considered a sustainable alternative and storage medium [5–10]. As the carrier of genetic information in our cells and, more generally, of life as we know it, the DNA molecule has evolved as a natural storage medium for genetic information and serves as a model for biological life. Synthetic DNA has been used to store digital information in a panoply of ways. Thanks to modern advances in biological techniques, synthetic DNA can be amplified, edited, and synthesized *de novo* outside of living cells. Indeed, digital data is encoded as a synthesized oligonucleotide, which is a polynucleotide with relatively small number of nucleotides. This qualifies as *in vitro* DNA data storage. These oligonucleotides can be stored in capsules under a protective atmosphere or embedded in a protective material to further improve its storage potential. While the large majority of research efforts focuses on *in vitro* storage, DNA-related operations within living cells (*in vivo*) are gaining traction. This corresponds to inserting synthetic DNA into the plasmid of living cells by means of genetic engineering [11,12]. *In vivo* storage has been experimentally demonstrated for either watermarking [13], or only small amounts of data [14].

* Corresponding author.
E-mail address: chisom.ezekannagha@uni-marburg.de (C. Ezekannagha).

https://doi.org/10.1016/j.mtbio.2022.100306
Received 21 January 2022; Received in revised form 5 May 2022; Accepted 22 May 2022
Available online 27 May 2022

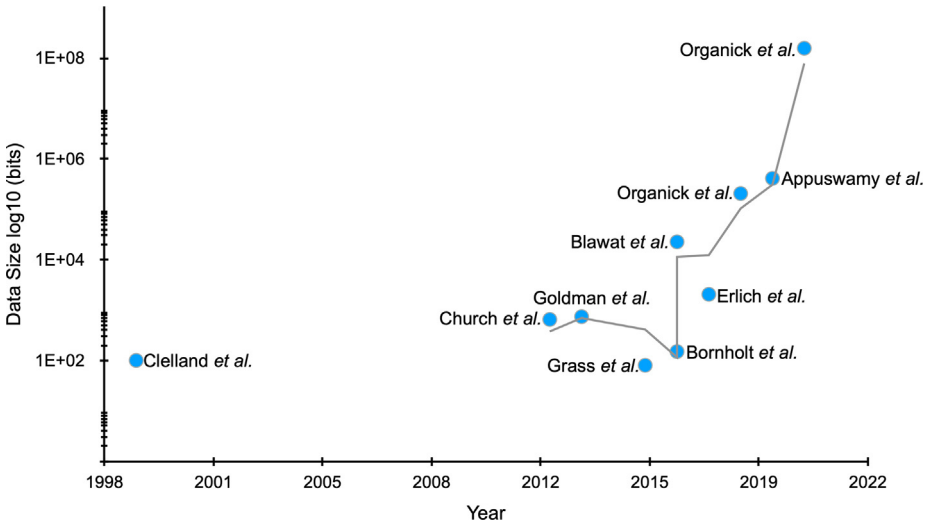

Fig. 1. Evolution of the amount of data stored in DNA. The fitted line shows the trend in data size from related work.

1.2. Advantages of DNA-based data storage

Hyper Dense - Thanks to its exceptional information density reaching up to 455 exabytes per gram (EB g^{-1}), synthetic DNA is an attractive alternative to conventional data storage media [10]. This value is about six orders of magnitude higher than the best conventional storage medium, i.e., magnetic tape [2,15–17].

Ultra-persistent - In terms of longevity, DNA can last about ten thousand times longer than traditional storage media. These aspects characterize DNA as a medium with high technical feasibility and economic viability. For example, DNA molecules of more than 560,000 years were recovered and analyzed from ancient samples [18]. Although storing DNA at room temperature does not involve any resource consumption, it reduces longevity. In realistic scenarios, such as long-term data archiving, synthetic DNA must be stored at moderately low temperatures to balance its protection and the energy costs of maintaining it. Using synthetic DNA as a storage medium is a sustainable solution.

Massive redundancy - Indeed, DNA is naturally replicated in cells before they divide, making the copying or replication of DNA a fast and inexpensive process. In parallel, replication can also be achieved *in vitro* by the Polymerase Chain Reaction (PCR). Thanks to this method, a huge replica of DNA can be prepared with high precision from a single copy [19,20].

Many research efforts as summarized in Table 1 have addressed the encoding of digital information in synthetic DNA [8,9,21,22]. Because the process of storing information is not entirely error-free, several design considerations are necessary for the reliable design of storage systems based on synthetic DNA. Indeed, errors may be introduced at various steps, such as DNA synthesis or sequencing. For example, the coverage of oligonucleotide (oligo) sequencing has been shown to be uneven, and in turn, resulting in the need for modern error-correcting codes (ECCs) capable of handling sequence read errors [22–25]. Current methods that employ synthetic DNA as a storage medium typically require hundreds or thousands of sequencing reads for each sequence to capture under-represented sequences. As a result, this particular inefficiency often stems from the lack of clear design considerations that ought to be adopted for synthetic DNA. Besides, the physicochemical properties of synthetic DNA affect its longevity and usage as a storage medium. As a matter of fact, the synthesized oligos do behave differently under various environmental and storage conditions [26]. This leads to considering these conditions as a design factor that can provide better solutions for encoding methods and physical storage device design. Additionally, as different types of digital information are employed, it is essential to consider their respective domain specifications. In the example of image data, the engineered solutions should consider the image size, the image type, etc [27,28]. To accommodate the growing need for efficient data storage approaches, encoding methods capable of error correction continue to be developed. Indeed, several ECCs have

Table 1
Summary of notable related work for synthetic DNA-based data storage. The table is presented in descending order based on the size of stored data. The encoding alphabet includes binary or alphanumeric encoding. The storage refers to the mechanism used for storing the DNA, either in an organism, buffer solution, or silica beads. Sequencing technologies include Illumina's technology based on the sequencing by synthesis (SBS) principle, and ONT's nanopore technology. The Error Correction refers to related work with codes that are able to detect and correct errors. The information density refers to the average number of binary information (bits) encoded in a nucleotide (nt). This binary information totals the data payload and the additional sequences for index, error correction, and primers. The synthetic DNA is stored in * silica beads and ** an organism (*in vivo*).

Related Work	Writing (encoding alphabet)	Sequencing Technology	Error Correction	Data Size	Information density (bits/nt)
Organick et al.	[0, 1]	SBS	1	150 GB	0.003
Appusawamy et al.	[0, 1]	SBS	1	400 MB	1
Organick et al.	[0, 1]	SBS/Nanopore	1	200.2 MB	0.81
Blawat et al.	[0, 1]	SBS	1	22 MB	0.89
Erlich and Zielinski	[0, 1]	SBS	1	2.15 MB	1.18
Goldman et al.	[0, 1]	SBS	1	739 kB	0.19
Church et al.	[0, 1]	SBS	0	650 kB	0.60
Bornholt et al.	[0, 1]	SBS	0	150 kB	0.57
Grass et al.*	[0, 1]	SBS	1	80 kB	0.86
Yadzi et al.	[0, 1]	Nanopore	0	3 kB	1.72
Clelland et al.	[A-Z, a-z, !]	SBS	0	4.625 B	1.27
Davis**	[0, 1]	–	0	1.125 B	1.25

been proposed and adopted, yet it is not easy to compare the overall performance of these ECCs due to different software architectures and varying evaluation metrics. Therefore, choosing suitable codes sets forth new challenges for the development of adapted and robust DNA data storage systems. Although major breakthroughs continue to be made to optimize the storage of digital information in synthetic DNA, different processing steps significantly impact the capabilities of this molecular storage medium. Additionally, it is also clear that there is currently no single effort capable of addressing all problems associated with synthetic DNA storage as different processing steps will fulfill different applications. In this context, there is a pressing need to achieve sustainable synthetic DNA storage solutions. Indeed, clear design considerations for usage and standardization with respect to the structure, synthesis, assembly, and evaluation performance will offer fresh perspectives and directions.

Furthermore, in the future, sustainable data storage using synthetic DNA should meet methodological requirements at multiple levels. Sustainability using synthetic DNA as a storage medium can be achieved at synthesis and storage. In the latter, and compared to conventional data storage media, synthetic DNA does not require continuous energy, nor does it produce large amounts of carbon dioxide emissions. This makes using synthetic DNA as a storage medium bearable for the environment and economically viable. This puts synthetic DNA in a category of its own, meaning a green data storage medium. Therefore, in this review, we discuss in detail the design considerations and their importance for each processing step. We also present aspects of information theory that we consider as evaluation metrics that can be taken into account for the efficient and robust design of a synthetic DNA-based storage system for long-term archiving.

2. Overview Process of a synthetic DNA-based storage system

The technological process of a synthetic DNA-based storage system consists of several steps: encoding, synthesizing the short DNA molecules or oligos, storing, reading the oligos, and decoding them into the original data. Operations such as modification, amplification, or destruction of the oligos can be associated with these processes. An overview of this process is illustrated in Fig. 2 and is described below. All figures created within this manuscript are print-friendly [29].

2.1. Encoding

The DNA molecule is made up of four different nucleotides: adenine (A), thymine (T), cytosine (C) and guanine (G). Compared to traditional storage media, the concept of storing data in DNA requires a specific arrangement of these nucleotides while following biological and technical constraints. The general idea is to transfer a binary sequence of information, that is to say a sequence of 1s and 0s, into quaternary DNA base sequences. Still, current DNA synthesis technology is limited to short DNA molecules or oligos with a high degree of precision. Typically, the binary source data is broken down into fragments and transferred into oligos of about 300 nucleotides in length. For accurate data recovery, strategies based on adding an addressable index to each fragment or storing overlapping fragments in different oligos are frequently used [9]. Such strategies have proven to be practical for large-scale storage. In addition, error detection and correction algorithms, such as the Fountain [22] or Reed-Solomon (RS) [25] codes, are typically applied to handle errors in subsequent processes, for example DNA synthesis and sequencing. Once the process of encoding digital information in DNA is completed, the DNA fragments are synthesized.

2.2. DNA synthesis

DNA synthesis enables the writing of relatively short nucleic acid fragments with a defined sequence of nucleotides. This process results in oligos and is based on two synthesis methods: chemical or enzymatic. Chemical synthesis relies on phosphoramidite chemistry methods [30, 31] and is usually performed using either traditional column-based synthesis or array-based synthesis. It involves adding successive nucleotides to the end of the synthetic DNA. The added nucleotide has a terminal group (e.g., dimethoxtrityl group) that blocks the addition of a second nucleotide. Then, the excess free nucleotide is removed by

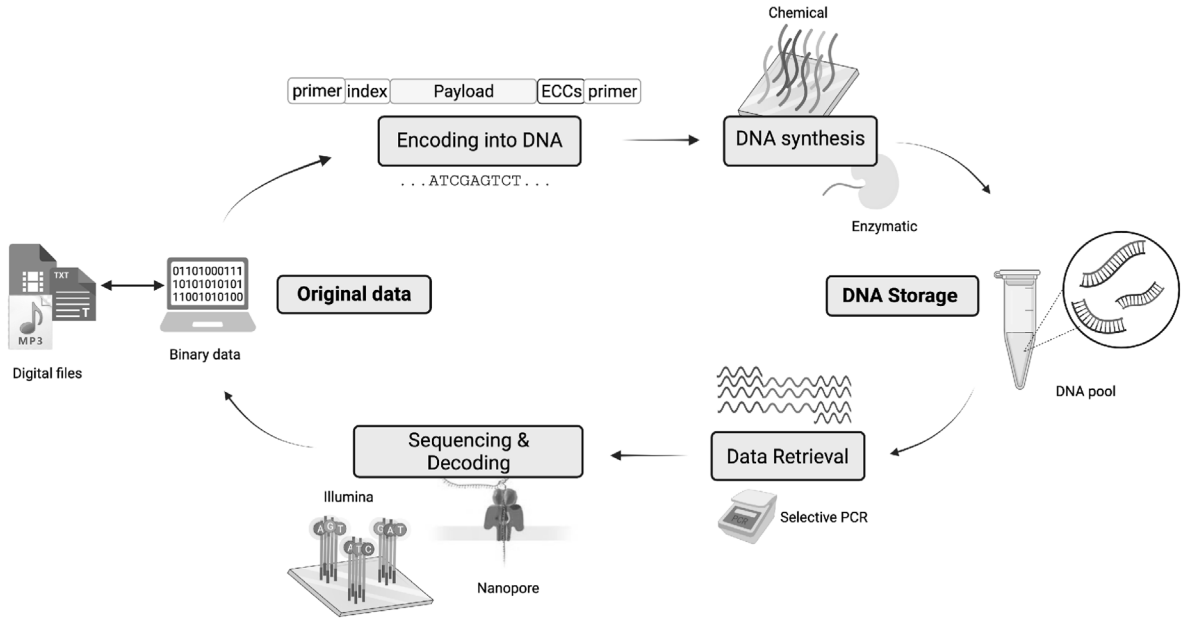

Fig. 2. Overview Process of a synthetic DNA-based storage system. Digital files are converted into binary data. The binary data is encoded into DNA sequences, additional DNA sequences including index, error correction, and primer for DNA amplification are also appended to the oligos. Oligos are synthesized chemically or enzymatically. Synthetic DNA from chemical or enzymatic synthesis processes is usually single-stranded. It is often stored in the single-stranded step, or the complementary strand is enzymatically synthesized, generating double-stranded DNA before storage. The selective extraction or amplification of the suitable oligos from the pool is enabled by PCR. The obtained oligos are sequences using a DNA sequencer. Then the obtained oligo sequences are decoded to obtain the original binary data.

washing, and the protective group is then removed under a chemical reaction. The following nucleotides are added one after the other in repeated cycles. In contrast, enzymatic synthesis was invented in the early 2010s and relied on the unique role of the DNA Polymerase which is present in immune cells, namely the Terminal deoxynucleotidyl Transferase. Compared to the chemical method, the enzymatic method has attracted much interest because of its advantages in speed, efficiency, and costs [32,33]. It was quickly developed and conceived as a convincing candidate for DNA synthesis. Although initial efforts result in high error rates, it is today becoming a commercially attractive solution with error rates of 0.7% (DNA Script Syntax Enzymatic DNA Synthesis technology). In order to maximize the use of the synthesized oligos, additional quality control efforts were explored. These efforts are aimed at reducing the error rate and improving the purification methods. High Pressure Liquid Chromatography (HPLC) or Polyacrylamide Gel Elctrophoresis (PAGE) methods are used to eliminate the vast majority of synthesis errors. Although promising approaches are under development, these methods are currently inadequate for complex oligos libraries. Advanced approaches based on simultaneous purification [34]and purification by synthesis and selection [35] have been investigated to improve DNA purification.

2.3. Storage

Once synthesized, the synthetic DNA is stored by using either a chemical or a physical method. The synthesized DNA is encapsulated in silica nanobeads in the chemical storage system, then referenced and distributed in microplates or microtiter plates. Before being read, the DNA stored in the nanobeads must be extracted by a chemical reagent that can preserve the synthetic DNA while dissolving the silica [36]. The physical method uses stainless steel storage capsules on the outside and glass on the inside, the size of a button cell [11]. Although an inert gas usually surrounds the capsule to protect it, it is not exclusive to the physical method. Moreover, light, water, and oxygen have a deleterious effect that causes chemical reactions responsible for breaks and mutations in synthetic DNA, making the information content of the oligonucleotide difficult to decipher. However, proper storage conditions combined with an appropriate storage method promise an exceptional lifespan of up to thousands of years for synthetic DNA.

2.4. Retrieval

To selectively transcribe the data stored onto the synthetic DNA, specific DNA fragments from the corresponding oligonucleotide pool must be physically extracted and assembled. This process is similar to random access in conventional digital storage media. However, locating specific DNA fragments of the desired data is difficult in molecular storage due to the lack of physical organization in a given pool or sample. To retrieve the data, different approaches based on DNA extraction or selective PCR amplification of the required synthetic DNA sample using a specific primer are often used during the retrieval process [24]. Other random access approaches based on microarray [37], immobilization of DNA molecules [38], digital microfluidic droplets [39], and DNA barcoded silica beads [40] have been explored to improve and enable random access.

2.5. DNA sequencing and decoding

After storing information for a certain time, sequencing technologies are used to read the corresponding sequence used for storage. Indeed DNA sequencing consists of determining the DNA sequence from the selected sample. In the case of synthetic DNA, the selected sample is a set of oligonucleotides (oligos). Sequencing of the oligos is carried out by the sequencer and produces a set of reads [41]. Thanks to high-throughput sequencing technology, the sequencer infers the sequence of base pairs for all or part of a single DNA fragment. This corresponds to reading a large volume of oligos or oligonucleotidic fragments. Then, these reads are decoded to extract the original information using decoding algorithms.

3. Considerations to advance synthetic DNA-based storage systems

The use of synthetic DNA as a storage medium requires a number of design considerations that determine its viability. This section discusses the various considerations for encoding, storage, sequencing, error correction, and evaluation metrics.

3.1. Considerations for encoding

One of the crucial steps of digital information storage in synthetic DNA is to assign binary values to the four nucleotides A, T, G, and C. It involves encoding these nucleotides into DNA strands with a specific sequence, where each sub-sequence of four nucleotides represents a byte. DNA synthesis is time-consuming and the most expensive part of the synthetic DNA-based storage process [42]. As previously mentioned, current DNA synthesis relies on chemical or enzymatic methods. However, while enzymatic synthesis is still gaining ground, the chemical method using traditional column-based or array-based synthesis is well established. Column-based synthesis is performed in columns and allows the addition of a single nucleotide to the column at a time until hundreds of nucleotides are obtained. To produce many DNA strands, array-based synthesis is generally favored over column-based synthesis thanks to its high throughput and low cost. In array-based synthesis, rather than synthesizing a single molecule, a diverse group of nucleotides with different sequences are synthesized in parallel. In addition, it is challenging to synthesize longer oligos [43,44]. Chemical synthesis of DNA is limited to about 300 nucleotides as its upper limit, and this results in a low yield and synthetic errors [45,46], which may impair the encoding of large amounts of data in a single sequence.

To achieve high information density and address errors obtained from DNA synthesis, researchers proposed varying combinations of conventional and composite bases to improve the existing encoding strategies [47,48]. Choi et al. used an alphabet of 11 additional composite bases to encode files of 854 kilobytes (KB), which significantly shortened the length of the DNA and resulted in a dramatic jump in theoretical information density to 3.9 bits per character ($\log_2 15 \approx 3.9$ bits) [47]. Furthermore, Anvay et al. designed a new nucleotide alphabet which contains both the four basic nucleotides and composite nucleotides [48]. In addition to the four base nucleotides, the composite bases M (50% G + 50% T) and K (50% A + 50% C) were used to encode a 2.12 Megabyte (MB) file. In this way, the storage density was improved by 24% compared to Erlich's earlier work [22]. The composite base design could store more data in a given length, but it requires more DNA copies and greater sequencing depth to read the data. It is important to note, in addition to conventional and composite bases, unconventional bases play an important role and offer additional solutions. Unlike composite bases where a nucleotide may be a combination of two natural nucleotides, the unconventional bases are chemically different from the natural nucleotides. These different solutions resulted in an encoding with increased alphabet size for data storage. In one effort, this has been achieved by introducing synthetic orthogonal nucleotide pairs [49,50]. In another effort, Hoshika et al. synthesized eight new nucleotides (S, B, J, V, K, X, Z, and P) by modifying the chemical groups of natural nucleotides. While the aim is to increase the encoding alphabet size, it is essential to avoid specific constraints related to the physicochemical DNA characteristics, which in turn may compromise its ability as a storage medium [49].

To further increase information density, Chaput et al. synthesized new nucleotides by artificially modifying their sugars. The unconventional nucleotides called Xeno-nucleic acids (XNA) were designed to avoid hybridizing with natural DNA while being incorporated by a living organism [50]. Although such nucleotides have been deliberately

modified for different applications, their adoption for synthetic DNA storage could improve the information density of synthetic DNA without interacting with and contaminating the DNA of the host organism. However, an important limiting factor concerns sequencing technologies, which encounter problems reading them.

3.2. Considerations for storage and access

The primary application of synthetic DNA-based storage systems is long-term archiving designed to preserve digital information over many decades and potentially centuries. Current applications with the goal of archiving information are limited by the high costs of synthesizing and sequencing DNA. However, such costs can be justified if they are incurred infrequently and subsidized on a long-term basis [51]. To optimize storage conditions, an important criterion is the long-term stability of synthetic DNA. A contrast can be made with naturally occurring DNA. Indeed, ancient DNA was found in fossilized bones buried in cool, dry soil and dated over 500,000 years old [52]. This finding is used as evidence of the long-term stability of DNA molecules [53,54]. However, estimates of the stability of DNA from biological samples may be considered too optimistic for synthetic DNA-based storage systems. While ancient DNA can be recovered from fossils, extensive degradation of a sample leads to detecting only the most abundant DNA molecules (mitochondrial). This fact is reflected in the estimated half-life of an ancient DNA found in fossils and approximated at $\sim$ 500 years. This estimate corresponds to a fragmentation rate (k) of 5.5×10^{-6} per nucleotide and per year for a 242 nucleotide mitochondrial DNA sequence [55]. Therefore, it is a substantially lower estimate compared to the 500,000 years needed to recover any viable genetic material. In the specific context of long-term archival storage, the percentage of the data that can be recovered depends on the amount of physical redundancy and the encoding method. It is important to note that fossilized DNA data suggests stability of a few hundred years.

To increase the synthetic DNA stability, its design can be fine tuned using highly controlled materials and environmental conditions. While the stability of biological samples is in general limited, several approaches have been inspired by fossilized DNA. Many have been proposed, most involving dehydration to reduce hydrolysis of the phosphate backbone [56]. We detail below three potential solutions that improve the stability of synthetic DNA. First, synthetic DNA was stored in biopolymer storage arrays, such as the commercial product DNA Stable, was adsorbed onto Flinders Technology Associate (FTA) filter cards and then incorporated into silk arrays or stored as a lyophilized powder [53,57]. Second, the synthetic DNA is stabilized by the adsorbing the DNA molecule onto a matrix. For example, a sensitivity analysis at increasing temperatures and over a period of 40 days, from 25, to 37, to 45 °C, resulted in the recovery of 80% of the DNA that was embedded in silk [58]. When unprotected, the recovery dropped to only 20%. Synthetic DNA stored in silk was also protected against UV radiation. Third and last, salts have been shown to offer another viable solution and mitigate special conditions such as high light and humidity. Salts have stabilizing effects for dehydrated DNA and are known to particularly maintain high DNA loading (DNA mass/total mass of the storage system) while keeping it relatively accessible [59]. Of the many different approaches that have been tested, the main one consistently demonstrates that the highest stability is achieved using the encapsulation technique. It consists of encapsulating the synthetic DNA in an inorganic matrix comprising iron oxide, silica, or both. Many research efforts have found that encapsulation can significantly improve the stability of DNA [25,36,60,61]. For example, Grass et al. estimated that encapsulating in silica particles could preserve DNA for 20–90 years at room temperature, 2000 years at 9.4 °C, and over 2 million years at −18 °C [25,36]. Puddu et al. directly compared the stability of encapsulated DNA to that of un-encapsulated DNA for 30 min at 100 °C [62]. In this case, estimates obtained by aging models have projected that 80% of the encapsulated DNA may be recovered when the synthetic DNA is protected. On the contrary, and when unprotected, only 0.05% of the synthetic DNA survived. DNA encapsulation is currently feasible and is a suitable method for long-term storage, giving the synthetic DNA-based storage system relatively high robustness and high information density. A research effort has investigated accelerated aging of various oligonucleotidic populations which encode digital data. It provided evidence of not only the half-lives, but also evidence of a complete file recovery [25]. The intersection of related work shows that the amount of physical redundancy introduced and the density sacrificed to enable degradation-tolerant encodings have both enabled long-term data storage using the DNA molecule. This work discusses considerations of storage based on a long-term application. However, the storage of DNA ranges from long-term archival storage to frequent and dynamic access of storage. Matange et al. provides an extensive review on the different storage formats. For example, storing DNA in aqueous solution for the purpose of frequent access of storage (i.e., access storage multiple times per year) [26]. The main advantage of storing DNA in an aqueous solution is the ease of retrieving information by random access. Polymerase chain reaction is the most widely used method for random access in synthetic DNA storage systems and is scalable with some modifications [63]. Random access is achieved by using a specific PCR primer to amplify and select the DNA sequence that can then be sequenced and decoded. Evolutionary methods for random access have been developed recently, such as the use of biotin-labeled primers for selection [64], or replacing the use of PCR amplification with biotin enrichment for selection [65].

3.3. Considerations for sequencing

Reading the information stored in synthetic DNA begins with DNA sequencing. Errors occur in DNA sequencing due to high GC content and long repeated bases (i.e., homopolymers). Indeed, the insertion and deletion error probability increases considerably for homopolymers with more than six repeated nucleotides. This mostly occurs for polynucleotidic sequences with G (polyG) because they lead to a greater representation of G's forming guanine tetraplex structures of high thermal stability but high instability for sequencing [66]. Moreover, the sequencing coverage of DNA strands with GC content of $<$20% or $>$75% is considered much lower with conventional sequencing methods [67].

DNA sequencing has favorable characteristics and has significantly increased read lengths and read speed. For example, the read length of Oxford Nanopore Technologies (ONT) can reach a thousand base pairs (bps) or more, which is several orders of magnitude larger than what can be achieved using NGS. This advantage has led to the development of new strategies for encoding data in synthetic DNA. For instance, Chen et al. attached short DNA hairpins to double-stranded DNA with different stem lengths. When this structured DNA passes the nanopore, the hairpins block the nanopore and cause the secondary current to decrease with the length of the hairpin. Thus, they represented 8 and 16 bps as '0′ and '1' bits, respectively. After optimization, 56 bits of data were encoded in a DNA strand of 7228 nucleotides [68]. In the example of ONT, current nanopore fluctuations, noises, nanopore defects, and uncontrolled speeds can also affect the outcome of the sequencing. Such occurrences may lead to the subsequent addition of the wrong nucleotide. Substantial error correction is highly necessary to improve the accuracy of the application of DNA as a storage medium. Therefore, the DNA strands in the storage device should contain an appropriate level of data redundancy to compensate for losses and eliminate errors that occur during DNA sequencing.

3.4. Considerations for error correction

Error correction is one of the integral functions of the data encoding and decoding stages. Both DNA synthesis and sequencing are error-prone. These errors arise due to strand breaks or losses and as a result of insertions, deletions and substitutions within strands. It should also be noted that homopolymers or high GC content can lead to low synthesis

yields and sequencing difficulties. Such biological constraints should be carefully considered and avoided when designing ECCs for synthetic DNA-based storage systems. Synthetic DNA as a storage medium exploits ECCs that are able of detect and correct errors. Adding ECCs in the encoded information is essential for error-free recovery of information. In addition, because DNA synthesis is expensive, an encoding scheme is needed to associate as many bits as possible with a nucleotide. An important aspect to consider is information density, which measures the number of bits (information) that can be stored per nucleotide (b/nt). Various factors affect the information density. For example, adding extra information to the oligos (payload) that contain the original data, results in an overall decrease in information density. Therefore, the information density is highly dependent on the properties of the encoding scheme and should be taken into consideration.

Many efforts have been put into developing error-correcting codes (ECC) that use high data redundancy to achieve an error-tolerant storage system [23–25,69]. Data redundancy can be divided into two types: logical redundancy or physical redundancy. Physical redundancy is the existence of many copies of oligos encoding the same information and is not considered adequate to remove all potential errors in DNA storage. To improve robustness and efficiency, logical redundancy is preferred. It comes from adding extra information to the oligos (payload) that contains the original data. This extra information includes: short DNA or oligos used to amplify the files by PCR, an index sequence indicating the relative position of the data, and a piece of sequence for error correction. This combination can ensure the reliability of the stored data. This additional information is linked to the payload to form a complete DNA fragment in the storage.

The encoding from Church et al. is the first to store relatively large amounts of information (5.27 megabit). It is important to note that the goal is for long term archiving of data. This encoding uses one nucleotide to encode one bit (i.e A or C = 0 and G or T = 1), thus it is possible to encode the same information as different sequences. The choice of whether a zero encodes A or C (or one encodes G or T) is randomly made unless a homopolymer of length four needs to be avoided. The randomness also ensures a relatively balanced GC content. To address error correction, each sequence is synthesized many times to enable easy identification of errors by comparing multiple sequences through a consensus. One drawback of the encoding is the lack of ECC. The encoding found a total of 22 errors (mostly caused by homopolymers) in the sequencing results even with abundant physical redundancy, indicating that a more robust technique is needed [8]. This encoding reported a theoretical information density of 1 b/nt. However, the redundancy and use of no ECCs also adversely affected the information density.

To increase fidelity during data retrieval, several error-correction methods that add redundancy have been proposed. Many propose practical improvements that serve as a basis for better encoding schemes. The latter are illustrated in Fig. 3 and feature Hamming codes [70], Huffman codes [71], logical exclusive-or (XOR) operation [24], Reed-Solomon (RS) codes [72,73], and fountain codes [74,75].

Goldman et al. encoded digital information of 739 kilobytes. The encoding used Huffman codes to transform each base into a ternary digit such that it is different from previous base to avoid homopolymers. This results to long sequence been partitioned into short DNA oligo of 100 bp in length, the first 75 bp overlaps with the previous oligo and the last 75 bp overlaps with the next oligo, resulting in a fourfold redundancy where every 25 nucleotide sequence will be present in the four DNA segment (Fig. 3A). The fourfold redundancy allows effective error correction as any error from synthesis or sequencing can be corrected by comparing multiple sequences through a consensus [9]. The encoding scheme cannot prevent abnormal GC content and only uses simple checksum coding to detect an error. The information density of this encoding depends on the encoded data.

Bornholt et al. introduced the concept of random-access to synthetic DNA storage system. While the goal is still long-term archival of information, the major contribution of this work was the introduction of random-access which enables a new level of efficiency. DNA is stored in a single pool and requires the sequencing of the whole pool to retrieve just to retrieve a subset of the data. Storing different data subset in separate pool is too big of a trade-off regarding density. To enable the retrieval of different data subsets, Bornholt et al. reported a solution that attach

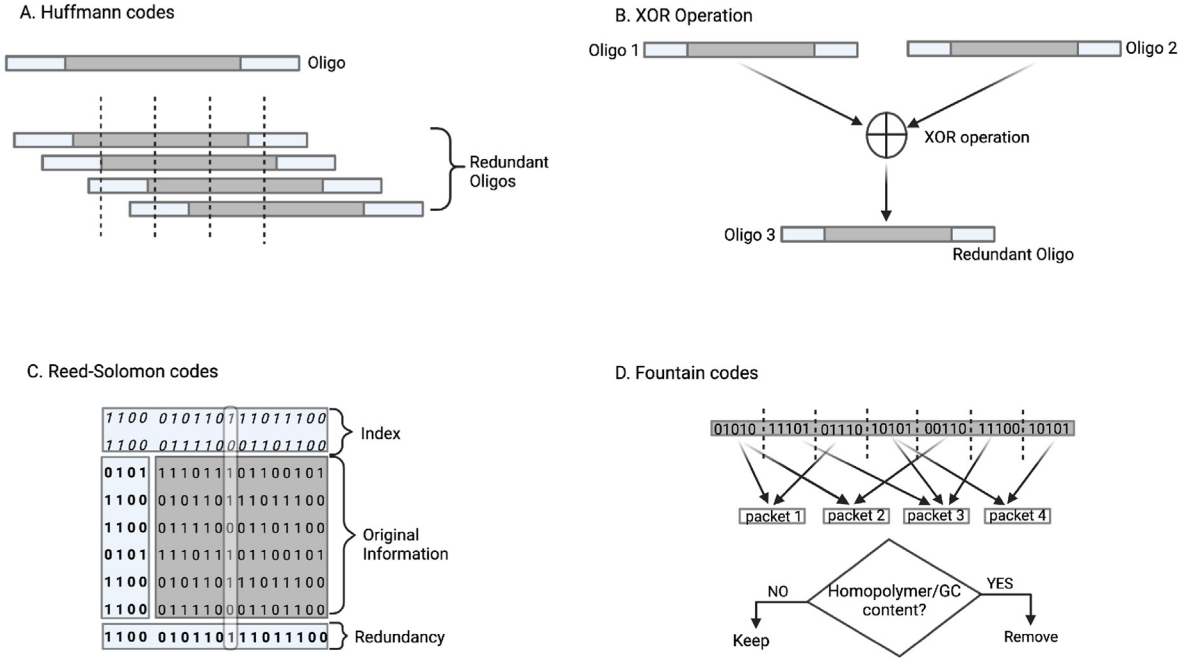

Fig. 3. Error correction methods for synthetic DNA-based storage systems. (A) Huffman codes based on overlapping DNA fragments, resulting in a fourfold redundancy to prevent error and data loss. (B) Code based on exclusive-or operation where any two out of the three oligos can recover the original information. (C) Two Reed-Solomon codes were added to the original information in orthogonal directions alongside the indices. (D) Fountain codes groups data into 'resource packets' and restores the original information after obtaining a sufficient number of packets. An additional screening step was added to exclude unqualified fragments.

different PCR primers to each sequence for individual retrieval. To encode information, the data is split into blocks which includes the following: PCR primers, an oligo payload (containing actual data), 2 sense nucleotides and an address to identify its position in the original data. The encoding of information still relies on Huffman code where each base transformed to ternary digit based on the previous based to avoid homopolymers. Bornholt et al. reported an encoding scheme inspired by Goldman's encoding using an XOR operation to yield redundancy [24]. The encoding uses two original oligos to generate a new oligo through an XOR operation so that any two out of the three DNA fragments can recover the original data (Fig. 3B). This encoding reduced the redundancy of the original data from 4-fold to half, therefore providing an efficient means in terms of information density when compared to Goldman encoding.

With an emphasis on error detection and correction, RS coding, which is widely used in the field of information communication [76] to encode information, has been applied for the first time to DNA-based synthetic storage. Grass et al. presented a new encoding that focuses on embedding DNA in silica and simulating DNA aging for archival purposes. The encoding uses RS codes generated on the Galois field (GF) for error correction [25]. In general, two sets of RS codes, namely the "inner code" and the "outer code", respectively, have been added to map the information symbols to orthogonal directions (Fig. 3C). The outer code is also mapped onto the indices. Homopolymers and in-balanced GC content were avoided by using a GF code wheel. During decoding, the inner code is decoded first, followed by the outer code. As a result of this work, the inner code corrected 0.7 errors per sequence, and the outer code corrects the total loss of 0.7%. Furthermore, even with the aging simulation obtained by keeping the DNA at high temperature, no errors were found during the information retrieval. Therefore, this shows that the method, even under these conditions, can recover the information without errors. This method significantly improved the encoding efficiency and corrected single errors and error bursts such as sequence degradation. In another work, Blawat et al. developed an ECC based on the RS codes for storing data in DNA [69]. Notably, each byte (8 bits) of the original data was converted into a sequence of five nucleotides. Two other criteria were applied to prevent homopolymers: the first three nucleotides should not be the same, and the last two nucleotides should not be the same. This improvement proved the viability of the forward error correction code. In addition to needing random access in synthetic DNA-based storage, Organick et al. reported another ECC based on RS codes that allows random access and recovery of target files in a large-scale system. In this instance, unique PCR primers are assigned to individual files after rigorous screening, which allows random access to the target file(s) [21], thereby setting a new milestone for the possibility of storing large amounts of data in DNA.

Erlich et al. relied on fountain codes and used the Luby Transform code, which resulted in a rateless erasure code [22]. The main concept is to group the signal sources into smaller sets, namely, resource packets. After receiving a sufficient number of packets, the original information can be successfully restored (Fig. 3D). Considering the limitations of synthesis and sequencing, an additional screening step was added to exclude sequences consisting of high GC content and homopolymers. The main advantage of fountain code is its very low redundancy and the fact that it can deal with errors based on deletions and insertions.

In synthetic DNA-based storage and retrieval, errors in the form of substitution, insertion and deletion are indispensable. These errors occur during the synthesis and sequencing processes. In this section, we have discussed different encoding approaches to mitigate these errors. There is always a trade-off between redundancy and information density. Therefore, choosing the right encoding approach depends on the data type and usage scenario. However, considering synthetic DNA-based storage only for long-term archiving as an application, there is a trade-off between accuracy and redundancy. DNA fountain code appears to be the only ECC in the synthetic DNA storage domain that can robustly handle loss on retrieval.

3.5. Considerations for evaluation metrics

A representation of information into DNA is focused on storage and error-free retrieval of data encoded in the four DNA nucleotides. With the rise of reasonably practical solutions that employ DNA as a storage medium, information-theoretic aspects are now considered. That is to say, a synthetic DNA molecule conceptually corresponds to an information channel. Additional research on information encoding into the DNA provides sufficient knowledge and techniques for identifying and comparing multiple encoding schemes based on a single attribute. This fact has been explored for capacity analysis where the Shannon capacity has been considered as a metric. It helps establish the upper limit on the amount of information that can be reliably stored in DNA under a given error rate [77,78].

This concept leads to considering the essential function of coding theory for the transmission of information from one source to another over a given channel. In turn, to model and evaluate the properties of one or more channels and their suitability to a specific ECCs, such concepts are needed [79]. A particular focus is directed to the impact of insertions, deletions, and substitution errors which affect, if not impair, the whole data storage process [80,81]. Furthermore, different distance measures in information theory have been used to design ECCs for the DNA storage channel. These evaluation metrics comprise: the Hamming distance [82, 83], the Levenshtein distance [84], and the Damerau-Levenshtein distance [85].

The major advantage of these codes is the ability to detect and correct a limited number of errors. Moreover, they guarantee a constant minimum distance. For example, Bystrykh et al. proposed to adapt the Hamming binary code to the DNA quaternary metric [86], thus preserving the minimum distance and the ability to correct individual errors at the DNA level. In addition, Song et al. introduced a new metric namely the sequence-subset distance which generalizes the Hamming distance to a distance function defined between any two sets of unordered vectors. It establishes a unified framework for the design ECCs for DNA storage channel [87]. These studies show that the error-correcting ability of such codes is entirely determined by their minimum distance. However, codes designed around the Hamming distance are capable of correcting only substitution errors. As indicated above, insertions and deletions (indels) might be a continuous problem for synthetic DNA-based storage channel. Therefore it is very important to develop a coding scheme resistant to this specific type of error. Buschmann et al. reported a code that follows ideas from the Levenshtein code [84] specifically designed for the DNA context which recovers errors in a DNA sequence [88].

Yet another concern is DNA degradation and its effects on evaluating the encoding scheme. It arises due to DNA aging caused by exposure to increased levels of radiation, humidity, or high temperatures. This degradation results in a changed structures of the DNA sequence. That is to say, if a sequence breaks in three positions, either the adjacent broken sequences swap positions, or one or both sequences decay leading to a bursty deletion. This motivated the study of codes based on the Damerau distance. Besides integrating substitution, insertions, and deletions errors, the Damerau distance accounts for adjacent transposition edits. Gabrys et al. designed codes based on the Damerau distance for joint block deletion and adjacent block transposition [89].

To optimize the information capacity, it is important to consider the shared or mutual information between the inputs and outputs of the channel. Mutual information is an important metric to determine information capacity and should be maximized [76]. It measures the accuracy with which the output of the channel, i.e. the reading of a DNA by sequencing, represents the input of the channel or the preset DNA sequence. However, the information can be distorted during the process of writing and reading the DNA sequences, resulting in shifts between the channel, which reduces the average mutual information during transmission [90]. The impact of indels on the information storage is in much greater magnitude than that of mismatches, as the loss or gain of consecutive sequences may impair the whole DNA molecule. Indels in

DNA storage are equivalent to "erasure channels" in information science. Although error correction codes have different approaches to correct information loss, the result is the most relevant. In this section, we have discussed the concept of evaluation metrics for DNA storage. This concept leads to considering the essential function of information theory for the transmission of information from one source to another over a given channel.

4. Challenges and future direction

Although DNA based storage has vast application prospects and value, the DNA molecule can be considered as a green data storage. Indeed, it has a theoretically higher density, lower electricity consumption, and longer retention time than conventional storage media. Yet, there remain many challenges to address before the broader use and adoption of synthetic DNA for long-term archiving to become possible. Future improvements depend on progress in all steps of the data storage process, including data encoding, DNA synthesis, storage, data recovery, DNA sequencing, and data decoding.

First, an important consideration is the physical storage and preservation of the synthetic DNA molecules away from adverse factors. This will ensure the long-term stability of the data storage. Although there has been mounting evidence that recovery of ancient DNA, hundreds of thousands of years old, is possible [55]. Indeed, DNA can degrade much faster than that, but it depends on the conditions to which it is exposed. For example, high humidity, high temperatures, and exposure to ultraviolet light can contribute to its degradation [17,91].

To address this issue, various research efforts have proposed a variety of methods to ensure the suitable conditions for DNA preservation [25,91,92]. These methods range from chemical solutions including dehydration or lyophilization, to additives or chemical encapsulation with protecting materials such as silicon dioxide [25]. However, a major point to such methods is that the preservation of DNA by encapsulation effectively prevents direct access to the data files. For instance, DNA encapsulated in silica cannot be directly amplified and retrieved by PCR if it is not separated from the beads. Therefore, an optimal preservation method ought to strike a balance between the long-term stability of the device and access to the synthetic DNA carrier in which the data resides.

Second, error correction based on data redundancy is essential in terms of coding schemes to ensure the correct recovery of data stored in DNA oligonucleotides. In general, more data redundancy allows for better correction of errors arising from DNA synthesis or sequencing. However, more data redundancy also results in less amount of information encoded in the DNA sequences. Therefore, there is a compromise between error-correction capability and logical data density. Newly discovered synthetically engineered nucleic acids could expand the choice of bases for encoding and thus increasing the theoretical limit [49,93]. Such nucleic acid candidates could help increase the coding efficiency for DNA digital storage in the near future. Additionally, the performance of ECC differs when applied in different coding schemes. This is especially relevant at the decoding stages as it may be time-consuming and computationally intensive. Another important aspect of future encoding is constrained encoding. This involves designing codewords that consider additional constraints different from the biological constraints and with the aim of improving storage stability or lowering synthesis and sequencing errors. These additional new constraints are based on minimum distance [94], chaos game representation [95], and thermodynamic properties [96–98]. Therefore, the choice of the coding scheme and the ECC should be systematically considered to achieve coding efficiency and optimal error tolerance.

Third, it is most likely that the time used to read the data stored in DNA sequences will continue to be high. However, as long as the throughput of DNA synthesis and sequencing is high, synthetic DNA-based storage can potentially replace traditional media for archiving data. This is because long-term archive storage can tolerate a longer access time and would benefit considerably from the lower energy costs of "at rest" data. The array-based DNA synthesis allows the parallel synthesis of a large group of DNA molecules of different sequences. Thus the consumption of reagents is lower than the column-based synthesis method. Although advancements based on low-cost synthesis technologies have been made [45,99], it remains challenging to enhance the synthesis scale to a competing level versus the scale of conventional media in terms of encoded data. The advancement in this case likely relies on improving the enzymatic synthesis, which may significantly increase the length, speed and quality of DNA synthesis. Indeed, the increase in length, speed and higher accuracy will reduce synthesis cost, which is an important factor. This factor is currently limiting both the widespread application and adoption of synthetic DNA-based storage, and it is capping the limit of experimentally validated data storage using synthetic DNA as a storage medium [2]. In addition, high-throughput sequencing technology has made a notable advancement by significantly reducing the cost and time of sequencing. However, the speed of reading the data is still lower than that of conventional media. As of yet, sequencing based on Oxford Nanopore technologies has been a promising technique to solve this problem. Yet many challenges hinder its adoption for DNA-based storage. Hence, there is an urgent need for further technical developments in DNA synthesis and sequencing to reduce the cost of storing synthetic DNA to an acceptable level.

Fourth, physical libraries for DNA data storage need to provide a pathway to full automation and scalability without substantial impact on density. This research topic is still largely open. Many researchers are committed to the realization of fully automated DNA information storage. Recently, Takahashi et al. released the first demonstration of a fully automated DNA-based storage system [100]. The automation of such systems poses multiple challenges to enable their use in large-scale archival systems. Such environments typically operate with minimal human intervention. On the contrary, most synthetic DNA storage steps rely on human participation (laboratory workers), excluding synthesis and sequencing. Although there is still a long way to go, the automation of information storage and reading is of great significance to the industrialization of synthetic DNA storage.

Fifth and last, other exciting applications of DNA data storage can be seen in the rise of DNA computing. It allows the parallelism of DNA hybridization processes to implement general purpose computations such as logic gates [101] and neural networks [102]. For example, researchers have shown that hybridization reactions can form complex cascades called DNA strand displacement [103], which allow for changes in DNA topologies and thus computations. In other words, encoding information in the form of DNA topological modifications opens up new computational possibilities.

5. Other directions in different knowledge domains

There exists a varied number of alternative storage media which have been covered extensively [2]. We focus on plant science as a use case and a point of discussion. Plants have been used in science, industry, and art to store specific labeled sequences. In science, watermarking systems for labeling DNA has been a common practice in yeast, plant, and even in human genomes [104,105]. Indeed, DNA watermarking systems exist to discriminate natural occurring plants from transgenic plants [106]. In this light, it is noteworthy to cover DNA signatures or watermarking. In industry, a leading name that keeps recurring is Monsanto. Indeed, watermarking opened up ethical questions as to commercializing immaterial goods in the economy [107]. Another notable application of DNA storage is in molecular tagging systems [108]. DNA-based tags are applied to objects and can be used to be read so to determine their identity and provenance. In art, molecular genetics has allowed the artist to engineer the plant genome and create new life forms [109]. There have been efforts to encapsulate the relevant information in different plant parts, such as an apple tree and the pollen grain. In one project, Shiho Fukuhara and Georg Tremmel applied for funding and scientific collaboration for a project entitled "Biopresence." In 2003, they proposed a

method using a genetic coding technique to store a person's DNA after death in an apple tree that can serve as a "living memorial" or "transgenic tombstone" [110]. In another project entitled "Natural History of Enigma" in 2006, Eduardo Kac genetically created a hybrid between a petunia flower and himself, expressing his DNA only in the red veins of the flower [111]. Such a project raised ethical questions that served as a valuable way to explore the ethical issues involved in storing and hybridizing DNA from two biological species.

6. Conclusion

The explosion of digital data generated worldwide has currently overwhelmed conventional storage systems. New means of digital data storage are required to keep up with such a pace and reduce the carbon emissions currently produced by large storage parks. Synthetic DNA data storage discussed in this review is a promising alternative to complement current storage media, which are now approaching their density limit. The long retention and high durability time of DNA make it a natural choice for sustainable long-term archival. Therefore, in this review, we introduced the technological process of synthetic DNA-based storage systems. Then we focused on the design considerations considered for each step and their importance for the efficient and robust design of synthetic DNA-based storage for long-term archiving. Finally, we presented aspects of information theory as well as distance metrics which we referred to as evaluation metrics that help provide sufficient knowledge and techniques to identify and compare several coding systems. As a novel molecular data storage medium, the use of synthetic DNA for data storage benefits from standards and design considerations. Indeed, extensive breakthroughs and standardization at every step of the process are still required to achieve a large-scale adoption. In summary, the presented process of using synthetic DNA as a storage medium offers the opportunity to identify and address issues at every step to meet the task of long-term archiving. In order to achieve excellent information density and stability, new technological breakthroughs will benefit the field, pushing synthetic DNA-based storage media to become the ultimate solution for long-term data storage and archiving needs.

Authors contribution

Conceptualization, C.E. and G.H.; Writing–Original Draft, C.E. and G.H.; Writing–Review & Editing, C.E., A.B., D.H., and G.H.; Visualization, C.E. and G.H.; Supervision, G.H.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

All authors are members of the MOSLA consortium, which has received funding from the Hessian Ministry for Science and the Arts (LOEWE).

References

[1] M. Hilbert, P. López, The world's technological capacity to store, communicate, and compute information, Science 332 (2011) 60–65.
[2] A. Anžel, D. Heider, G. Hattab, The visual story of data storage: from storage properties to user interfaces, Comput. Struct. Biotechnol. J. 19 (2021) 4904–4918, https://doi.org/10.1016/j.csbj.2021.08.031. URL: https://www.sciencedirect.com/science/article/pii/S2001037021003627.
[3] S. Furrer, M.A. Lantz, P. Reininger, A. Pantazi, H.E. Rothuizen, R.D. Cideciyan, G. Cherubini, W. Haeberle, E. Eleftheriou, J. Tachibana, N. Sekiguchi, T. Aizawa, T. Endo, T. Ozaki, T. Sai, R. Hiratsuka, S. Mitamura, A. Yamaguchi, 201 gb/in2 recording areal density on sputtered magnetic tape, IEEE Trans. Magn. 54 (2018) 1–8, https://doi.org/10.1109/TMAG.2017.2727822.
[4] D.R.J.G.J. Rydning, The Digitization of the World from Edge to Core, Framingham: International Data Corporation, 2018, p. 16.
[5] J. Davis, Microvenus, Art J. 55 (1996) 70–74.
[6] C.T. Clelland, V. Risca, C. Bancroft, Hiding messages in DNA microdots, Nature 399 (1999) 533–534.
[7] C. Bancroft, T. Bowler, B. Bloom, C.T. Clelland, Long-term storage of information in dna, Science 293 (2001) 1763–1765.
[8] G.M. Church, Y. Gao, S. Kosuri, Next-generation digital information storage in DNA, Science 337 (2012) 1628, 1628.
[9] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E.M. LeProust, B. Sipos, E. Birney, Towards practical, high-capacity, low-maintenance information storage in synthesized DNA, Nature 494 (2013) 77–80.
[10] L. Organick, Y.J. Chen, S.D. Ang, R. Lopez, X. Liu, K. Strauss, L. Ceze, Probing the physical limits of reliable DNA data retrieval, Nat. Commun. 11 (2020) 1–7.
[11] T.J. Anchordoquy, M.C. Molina, Preservation of DNA, Cell Preserv. Technol. 5 (2007) 180–188.
[12] S.L. Shipman, J. Nivala, J.D. Macklis, G.M. Church, Crispr–cas encoding of a digital movie into the genomes of a population of living bacteria, Nature 547 (2017) 345–349.
[13] D. Heider, A. Barnekow, Dna watermarks: a proof of concept, BMC Mol. Biol. 9 (2008) 1–10.
[14] S.S. Yim, R.M. McBee, A.M. Song, Y. Huang, R.U. Sheth, H.H. Wang, Robust direct digital-to-biological data storage in living cells, Nat. Chem. Biol. 17 (2021) 246–253.
[15] A. Extance, How DNA could store all the world's data, Nat News 537 (2016) 22.
[16] T.R. Gregory, J.A. Nicol, H. Tamm, B. Kullman, K. Kullman, I.J. Leitch, B.G. Murray, D.F. Kapraun, J. Greilhuber, M.D. Bennett, Eukaryotic genome size databases, Nucleic Acids Res. 35 (2007) D332–D338.
[17] V. Zhirnov, R.M. Zadegan, G.S. Sandhu, G.M. Church, W.L. Hughes, Nucleic acid memory, Nat. Mater. 15 (2016) 366–370.
[18] L. Orlando, P. Darlu, M. Toussaint, D. Bonjean, M. Otte, C. Hänni, Revisiting neandertal diversity with a 100,000 year old mtDNA sequence, Curr. Biol. 16 (2006) R400–R402.
[19] L. Garibyan, N. Avashia, Research techniques made simple: polymerase chain reaction (pcr), J. Invest. Dermatol. 133 (2013) e6.
[20] H. Ochman, A.S. Gerber, D.L. Hartl, Genetic applications of an inverse polymerase chain reaction, Genetics 120 (1988) 621–623.
[21] L. Organick, S.D. Ang, Y.J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M.Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, et al., Random access in large-scale DNA data storage, Nat. Biotechnol. 36 (2018) 242–248.
[22] Y. Erlich, D. Zielinski, DNA fountain enables a robust and efficient storage architecture, Science 355 (2017) 950–954.
[23] S.H.T. Yazdi, R. Gabrys, O. Milenkovic, Portable and error-free DNA-based data storage, Sci. Rep. 7 (2017) 1–6.
[24] J. Bornholt, R. Lopez, D.M. Carmean, L. Ceze, G. Seelig, K. Strauss, A DNA-based archival storage system, in: Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems, 2016, pp. 637–649.
[25] R.N. Grass, R. Heckel, M. Puddu, D. Paunescu, W.J. Stark, Robust chemical preservation of digital information on DNA in silica with error-correcting codes, Angew. Chem. Int. Ed. 54 (2015) 2552–2555.
[26] K. Matange, J.M. Tuck, A.J. Keung, Dna stability: a central design consideration for dna data storage systems, Nat. Commun. 12 (2021) 1–9.
[27] M. Dimopoulou, M. Antonini, P. Barbry, R. Appuswamy, DNA coding for image storage using image compression techniques, in: CORESA 2018, 2018.
[28] M. Dimopoulou, M. Antonini, Efficient storage of images onto DNA using vector quantization, in: 2020 Data Compression Conference (DCC), IEEE, 2020, p. 363, 363.
[29] G. Hattab, T.M. Rhyne, D. Heider, Ten Simple Rules to Colorize Biological Data Visualization, 2020.
[30] M.H. Caruthers, The chemical synthesis of DNA/RNA: our gift to science, J. Biol. Chem. 288 (2013) 1420–1427.
[31] S. Beaucage, M. Caruthers, Deoxynucleoside phosphoramidites—a new class of key intermediates for deoxypolynucleotide synthesis, Tetrahedron Lett. 22 (1981) 1859–1862.
[32] A.S. Mathews, H. Yang, C. Montemagno, Photo-cleavable nucleotides for primer free enzyme mediated DNA synthesis, Org. Biomol. Chem. 14 (2016) 8278–8288.
[33] E.A. Motea, A.J. Berdis, Terminal deoxynucleotidyl transferase: the story of a misguided DNA polymerase, Biochim. Biophys. Acta Protein Proteonomics 1804 (2010) 1151–1166.
[34] A. Pinto, S.X. Chen, D.Y. Zhang, Simultaneous and stoichiometric purification of hundreds of oligonucleotides, Nat. Commun. 9 (2018) 1–9.
[35] H. Choi, Y. Choi, J. Choi, A.C. Lee, H. Yeom, J. Hyun, T. Ryu, S. Kwon, Purification of multiplex oligonucleotide libraries by synthesis and selection, Nat. Biotechnol. 40 (2022) 47–53.
[36] W.D. Chen, A.X. Kohll, B.H. Nguyen, J. Koch, R. Heckel, W.J. Stark, L. Ceze, K. Strauss, R.N. Grass, Combining data longevity with high storage capacity—layer-by-layer DNA encapsulated in magnetic nanoparticles, Adv. Funct. Mater. 29 (2019) 1901672.
[37] A. El-Shaikh, M. Welzel, D. Heider, B. Seeger, High-scale random access on dna storage systems, NAR Genom. Bioinf. 4 (2022) lqab126.
[38] Y. Choi, H.J. Bae, A.C. Lee, H. Choi, D. Lee, T. Ryu, J. Hyun, S. Kim, H. Kim, S.H. Song, et al., Dna micro-disks for the management of dna-based data storage with index and write-once–read-many (worm) memory features, Adv. Mater. 32 (2020) 2001249.

[39] S. Newman, A.P. Stephenson, M. Willsey, B.H. Nguyen, C.N. Takahashi, K. Strauss, L. Ceze, High density dna data storage library via dehydration with digital microfluidic retrieval, Nat. Commun. 10 (2019) 1–6.

[40] J.L. Banal, T.R. Shepherd, J. Berleant, H. Huang, M. Reyes, C.M. Ackerman, P.C. Blainey, M. Bathe, Random access dna memory using boolean search in an archival file storage system, Nat. Mater. 20 (2021) 1272–1280.

[41] J. Guo, N. Xu, Z. Li, S. Zhang, J. Wu, D.H. Kim, M.S. Marma, Q. Meng, H. Cao, X. Li, et al., Four-color DNA sequencing with 3'-o-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides, Proc. Natl. Acad. Sci. Unit. States Am. 105 (2008) 9145–9150.

[42] H. Tsunoda, T. Kudo, A. Ohkubo, K. Seio, M. Sekine, Synthesis of Oligodeoxynucleotides using fully protected Deoxynucleoside 3'-Phosphoramidite building blocks and base recognition of Oligodeoxynucleotides incorporating N3-Cyano-Ethylthymine, Molecules 15 (2010) 7509–7531.

[43] A. Shivalingam, T. Brown, Synthesis of chemically modified DNA, Biochem. Soc. Trans. 44 (2016) 709–715.

[44] S. Kosuri, G.M. Church, Large-scale de novo DNA synthesis: technologies and applications, Nat. Methods 11 (2014) 499–507.

[45] P.L. Antkowiak, J. Lietard, M.Z. Darestani, M.M. Somoza, W.J. Stark, R. Heckel, R.N. Grass, Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction, Nat. Commun. 11 (2020) 1–10.

[46] R. Heckel, G. Mikutis, R.N. Grass, A characterization of the DNA data storage channel, Sci. Rep. 9 (2019) 1–12.

[47] Y. Choi, T. Ryu, A.C. Lee, H. Choi, H. Lee, J. Park, S.H. Song, S. Kim, H. Kim, W. Park, et al., High information capacity DNA-based data storage with augmented encoding characters using degenerate bases, Sci. Rep. 9 (2019) 1–7.

[48] L. Anavy, I. Vaknin, O. Atar, R. Amit, Z. Yakhini, Data storage in DNA with fewer synthesis cycles using composite DNA letters, Nat. Biotechnol. 37 (2019) 1229–1236.

[49] S. Hoshika, N.A. Leal, M.J. Kim, M.S. Kim, N.B. Karalkar, H.J. Kim, A.M. Bates, N.E. Watkins, H.A. SantaLucia, A.J. Meyer, et al., Hachimoji DNA and RNA: a genetic system with eight building blocks, Science 363 (2019) 884–887.

[50] J.C. Chaput, P. Herdewijn, M. Hollenstein, Orthogonal Genetic Systems, ChemBioChem (2020) 21.

[51] J. Byron, E.L. Miller, D.D.E. Long, Measuring the cost of reliability in archival systems, in: Proceeding of the Conference on Mass Storage Systems and Technologies (MSST '20), 2020.

[52] M. Hofreiter, D. Serre, H.N. Poinar, M. Kuch, S. Pääbo, Ancient DNA, Nat. Rev. Genet. 2 (2001) 353–359.

[53] E. Willerslev, A.J. Hansen, R. Rønn, T.B. Brand, I. Barnes, C. Wiuf, D. Gilichinsky, D. Mitchell, A. Cooper, Long-term persistence of bacterial DNA, Curr. Biol. 14 (2004) R9–R10.

[54] T. van der Valk, P. Pečnerová, D. Díez-del Molino, A. Bergström, J. Oppenheimer, S. Hartmann, G. Xenikoudakis, J.A. Thomas, M. Dehasque, E. Sağlıcan, et al., Million-year-old DNA sheds light on the genomic history of mammoths, Nature 591 (2021) 265–269.

[55] M.E. Allentoft, M. Collins, D. Harker, J. Haile, C.L. Oskam, M.L. Hale, P.F. Campos, J.A. Samaniego, M.T.P. Gilbert, E. Willerslev, et al., The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils, Proc. Biol. Sci. 279 (2012) 4724–4733.

[56] P.L. Antkowiak, J. Koch, P. Rzepka, B. Nguyen, K. Strauss, W.J. Stark, R.N. Grass, Anhydrous calcium phosphate crystals stabilize dna for dry storage, Chem. Commun. 58 (2022) 3174–3177.

[57] L. Organick, B.H. Nguyen, R. McAmis, W.D. Chen, A.X. Kohll, S.D. Ang, R.N. Grass, L. Ceze, K. Strauss, An empirical comparison of preservation methods for synthetic DNA data storage, Small Methods 5 (2021) 2001094.

[58] Y. Liu, Z. Zheng, H. Gong, M. Liu, S. Guo, G. Li, X. Wang, D.L. Kaplan, DNA preservation in silk, Biomater. Sci. 5 (2017) 1279–1292.

[59] A.X. Kohll, P.L. Antkowiak, W.D. Chen, B.H. Nguyen, W.J. Stark, L. Ceze, K. Strauss, R.N. Grass, Stabilizing synthetic DNA for long-term data storage with earth alkaline salts, Chem. Commun. 56 (2020) 3613–3616.

[60] J. Koch, S. Gantenbein, K. Masania, W.J. Stark, Y. Erlich, R.N. Grass, A DNA-of-things storage architecture to create materials with embedded memory, Nat. Biotechnol. 38 (2020) 39–43.

[61] D. Clermont, S. Santoni, S. Saker, M. Gomard, E. Gardais, C. Bizet, Assessment of DNA encapsulation, a new room-temperature DNA storage method, Biopreserv. Biobanking 12 (2014) 176–183.

[62] M. Puddu, W.J. Stark, R.N. Grass, Silica microcapsules for long-term, robust, and reliable room temperature rna preservation, Adv. Healthcare Mater. 4 (2015) 1332–1338.

[63] K.J. Tomek, K. Volkel, E.W. Indermaur, J.M. Tuck, A.J. Keung, Promiscuous molecules for smarter file operations in dna-based data storage, Nat. Commun. 12 (2021) 1–10.

[64] K.J. Tomek, K. Volkel, A. Simpson, A.G. Hass, E.W. Indermaur, J.M. Tuck, A.J. Keung, Driving the scalability of dna-based information storage systems, ACS Synth. Biol. 8 (2019) 1241–1248.

[65] K.N. Lin, K. Volkel, J.M. Tuck, A.J. Keung, Dynamic and scalable dna-based information storage, Nat. Commun. 11 (2020) 1–12.

[66] K. Poon, R.B. Macgregor Jr., Unusual behavior exhibited by multistranded guanine-rich DNA complexes, Biopolymers: Original Research on Biomolecules 45 (1998) 427–434.

[67] Y. Benjamini, T.P. Speed, Summarizing and correcting the GC content bias in high-throughput sequencing, Nucleic Acids Res. 40 (2012) e72, e72.

[68] K. Chen, J. Kong, J. Zhu, N. Ermann, P. Predki, U.F. Keyser, Digital data storage using DNA nanostructures and solid-state nanopores, Nano Lett. 19 (2018) 1210–1215.

[69] M. Blawat, K. Gaedke, I. Huetter, X.M. Chen, B. Turczyk, S. Inverso, B.W. Pruitt, G.M. Church, Forward error correction for DNA data storage, Procedia Comput. Sci. 80 (2016) 1011–1022.

[70] D. Heider, A. Barnekow, Dna watermarking: challenging perspectives for biotechnological applications, Curr. Bioinf. 6 (2011) 375–382.

[71] M. Ailenberg, O.D. Rotstein, An improved Huffman coding method for archiving text, images, and music characters in DNA, Biotechniques 47 (2009) 747–754.

[72] I.S. Reed, G. Solomon, Polynomial codes over certain finite fields, J. Soc. Ind. Appl. Math. 8 (1960) 300–304.

[73] V. Guruswami, M. Sudan, Improved decoding of Reed-Solomon and algebraic-geometric codes, in: Proceedings 39th Annual Symposium on Foundations of Computer Science (Cat, No. 98CB36280), IEEE, 1998, pp. 28–37.

[74] D.J. MacKay, Fountain codes, IEE Proc. Commun. 152 (2005) 1062–1068.

[75] A.G. Dimakis, V. Prabhakaran, K. Ramchandran, Distributed fountain codes for networked storage, in: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, IEEE, 2006, p. V. V).

[76] I. Csiszár, J. Körner, Information Theory: Coding Theorems for Discrete Memoryless Systems, Cambridge University Press, 2011.

[77] F. Balado, On the shannon capacity of dna data embedding, in: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 1766–1769.

[78] F. Balado, Capacity of dna data embedding under substitution mutations, IEEE Trans. Inf. Theor. 59 (2012) 928–941.

[79] P. Siegel, Codes for mass data storage systems (kh schouhamer immink; 2004) [book review], IEEE Trans. Inf. Theor. 52 (2006) 5614–5616.

[80] A. Lenz, P.H. Siegel, A. Wachter-Zeh, E. Yaakobi, Coding over sets for DNA storage, IEEE Trans. Inf. Theor. 66 (2020) 2331–2351, https://doi.org/10.1109/TIT.2019.2961265.

[81] M. Kovačević, V.Y.F. Tan, Codes in the space of multisets—coding for permutation channels with impairments, IEEE Trans. Inf. Theor. 64 (2018) 5156–5169, https://doi.org/10.1109/TIT.2017.2789292.

[82] R.W. Hamming, Error detecting and error correcting codes, The Bell system technical journal 29 (1950) 147–160.

[83] R.W. Hamming, Coding and Information Theory, Prentice-Hall, 1986.

[84] V.I. Levenshtein, et al., Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet Physics Doklady, Soviet Union, 1966, pp. 707–710.

[85] F.J. Damerau, A technique for computer detection and correction of spelling errors, Commun. ACM 7 (1964) 171–176.

[86] L.V. Bystrykh, Generalized DNA barcode design based on hamming codes, PLoS One 7 (2012), e36852.

[87] W. Song, K. Cai, K.A.S. Immink, Sequence-subset distance and coding for error control in DNA-based data storage, IEEE Trans. Inf. Theor. 66 (2020) 6048–6065.

[88] T. Buschmann, L.V. Bystrykh, Levenshtein error-correcting barcodes for multiplexed DNA sequencing, BMC Bioinf. 14 (2013) 1–10.

[89] R. Gabrys, E. Yaakobi, O. Milenkovic, Codes in the Damerau distance for deletion and adjacent transposition correction, IEEE Trans. Inf. Theor. 64 (2017) 2550–2570.

[90] Y. Dong, F. Sun, Z. Ping, Q. Ouyang, L. Qian, DNA storage: research landscape and future prospects, Natl. Sci. Rev. 7 (2020) 1092–1107.

[91] J. Bonnet, M. Colotte, D. Coudy, V. Couallier, J. Portier, B. Morin, S. Tuffet, Chain and conformation stability of solid-state DNA: implications for room temperature storage, Nucleic Acids Res. 38 (2010) 1531–1546.

[92] N.V. Ivanova, M.L. Kuzmina, Protocols for dry DNA storage and shipment at room temperature, Mol. Ecol. Resour. 13 (2013) 890–898.

[93] D.A. Malyshev, K. Dhami, T. Lavergne, T. Chen, N. Dai, J.M. Foster, I.R. Corrêa, F.E. Romesberg, A semi-synthetic organism with an expanded genetic alphabet, Nature 509 (2014) 385–388.

[94] D. Limbachiya, M.K. Gupta, V. Aggarwal, Family of constrained codes for archival dna data storage, IEEE Commun. Lett. 22 (2018) 1972–1975.

[95] H.F. Löchel, M. Welzel, G. Hattab, A.C. Hauschild, D. Heider, Fractal construction of constrained code words for dna storage systems, Nucleic Acids Res. 50 (2021) e30.

[96] B. Cao, X. Li, X. Zhang, B. Wang, Q. Zhang, X. Wei, Designing uncorrelated address constrain for dna storage by dmvo algorithm, IEEE ACM Trans. Comput. Biol. Bioinf 19 (2020) 866–877.

[97] B. Cao, X. Zhang, J. Wu, B. Wang, Q. Zhang, X. Wei, Minimum free energy coding for dna storage, IEEE Trans. NanoBioscience 20 (2021) 212–222.

[98] Q. Yin, Y. Zheng, B. Wang, Q. Zhang, Design of constraint coding sets for archive dna storage, IEEE ACM Trans. Comput. Biol. Bioinf (2021), 1-1.

[99] B.H. Nguyen, C.N. Takahashi, G. Gupta, J.A. Smith, R. Rouse, P. Berndt, S. Yekhanin, D.P. Ward, S.D. Ang, P. Garvan, et al., Scaling dna data storage with nanoscale electrode wells, Sci. Adv. 7 (2021), eabi6714.

[100] C.N. Takahashi, B.H. Nguyen, K. Strauss, L. Ceze, Demonstration of end-to-end automation of dna data storage, Sci. Rep. 9 (2019) 1–5.

[101] D.Y. Zhang, G. Seelig, Dynamic dna nanotechnology using strand-displacement reactions, Nat. Chem. 3 (2011) 103–113.

[102] K.M. Cherry, L. Qian, Scaling up molecular pattern recognition with dna-based winner-take-all neural networks, Nature 559 (2018) 370–376.

[103] B. Wang, C. Chalk, D. Soloveichik, Simd—- dna: single instruction, multiple data computation with dna strand displacement cascades, in: International Conference on DNA Computing and Molecular Programming, Springer, 2019, pp. 219–235.

[104] L. Hauben, M. Steenackers, J. Swings, Pcr-based detection of the causal agent of watermark disease in willows (salix spp.), Appl. Environ. Microbiol. 64 (1998) 3966–3971.
[105] M. Liss, D. Daubert, K. Brunner, K. Kliche, U. Hammes, A. Leiherer, R. Wagner, Embedding Permanent Watermarks in Synthetic Genes, 2012.
[106] N. Yamamoto, H. Kajiura, S. Takeno, N. Suzuki, Y. Nakazawa, A watermarking system for labeling genomic dna, Plant Biotechnol. (2014) 14–609.
[107] J.M. Schubert, Appropriating and Commercialising Immaterial Goods in Knowledge Economies, 2007.
[108] K. Doroschak, K. Zhang, M. Queen, A. Mandyam, K. Strauss, L. Ceze, J. Nivala, Rapid and robust assembly and decoding of molecular tags with dna-based nanopore signatures, Nat. Commun. 11 (2020) 1–8.
[109] A. Dunne, F. Raby, Speculative Everything: Design, Fiction, and Social Dreaming, MIT press, 2013.
[110] C. Holden, Transgenic tombstone, Science 300 (2003) 1501.
[111] E. Kac, Transgenic art, História, Ciências, Saúde-Manguinhos 13 (2006) 247–256.