EURASIP Journal on Image and Video Processing

REVIEW Open Access

Data and image storage on synthetic DNA: existing solutions and challenges

Melpomeni Dimopoulou* and Marc Antonini

*Correspondence:
dimopoulou@i3s.unice.fr

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis (I3S), UMR 7271, Université Côte d'Azur, CNRS, Euclide B, 2000 Route des Lucioles, 06900 Sophia Antipolis, France

Abstract

Storage of digital data is becoming challenging for humanity due to the relatively short life-span of storage devices. Furthermore, the exponential increase in the generation of digital data is creating the need for constantly constructing new resources to handle the storage of this data volume. Recent studies suggest the use of the DNA molecule as a promising novel candidate which can hold 500 Gbyte/mm^3 (1000 times more than HDD drives). Any digital information can be synthesized into DNA in vitro and stored in special tiny storage capsules that can promise reliability for hundreds of years. The stored DNA sequence can be retrieved whenever needed using special machines that are called sequencers. This whole process is very challenging, as the process of DNA synthesis is expensive in terms of money and sequencing is prone to errors. However, studies have shown that when respecting several rules in the encoding, the probability of sequencing error is reduced. Consequently, the encoding of digital information is not trivial, and the input data need to be efficiently compressed before encoding so that the high synthesis cost is reduced. In this paper, we present a survey on the storage of digital data in synthetic DNA, explaining the problem which is tackled by this novel field of study, present the main processes included in the storage workflow as well as the history of different studies and the most well-known algorithms that have been proposed in the bibliography on DNA data storage.

Keywords: DNA data storage, Robust encoding, Quaternary code

1 Introduction

The problem of data explosion constitutes one of the greatest challenges of digital evolution. The continuous greedy use of the internet, including digital platforms and social networks is leading to an immersive increase in the generation of digital data which needs to be handled and stored efficiently. This information overload is massively stored in the servers of big data centers where it is being organized and archived using different technologies according to the data's nature and demand. According to an article published in [1], the rapid rise in smartphone usage, IoT adoption, and big data analytics have led to a massive growth in data centers, and they come with a cost. This article presents the following statistics as provided by IDC:

- In 2012 there existed 500.000 data centers to handle global traffic while today there exist more than 8 million.
- The yearly CO_2 emissions of data centers reaches the amounts of CO_2 produced by the global airline industry.
- Every year, millions of data centers worldwide are draining country-sized amounts of electricity. Several models even predict that data center energy-usage could engulf over 10% of the global electricity supply by 2030 if left unchecked.
- 90% of the existing data have been only generated in the last 2 years.
- The amount of energy used by data centers continues to double every 4 years.

Along with the above numbers, it is also known that storage media have a limited lifespan which varies from 3 to 5 years for HDD drives [2] and 20–30 years for back-up tape drives [3]. To reassure reliability of the stored data, it is therefore necessary that data centers frequently replace the different storage units, a fact that leads to a huge hardware waste. Furthermore, the replacement of older storage units yields the need for migrating the data into the new units, a process which is expensive both in terms of money and energy. All these facts, reveal that the enormous increase in the generation of data is causing significant pollution to the environment. Due to the resulting environmental impact, increased pressure has been placed on companies to follow a green policy by building green or sustainable data centers which utilize energy-efficient technologies.

1.1 What is cold data? Problem definition

For managing, storing and re-purposing digital content, industries and data centers differentiate the data into three levels, hot, warm and cold, based on interest or access priorities. The frequency of data demand (metaphorically called data temperature) denotes the most appropriate unit to which each type of data should be stored. More precisely, hot data refer to assets that require the fastest storage as they are accessed most frequently. It is thus stored in the nearest or closest spots from the accessing points such as solid-state or flash drives and CPU. Warm data represent information that is less accessible and is stored on a bigger storage capacity or file servers for relatively cost-efficient concern. Finally, the data which are very rarely or even never accessed and does not require on-line workflow is placed on the slowest low-cost options of storage mediums such as tape and optical discs and is termed as cold data.

The largest part of digital information consists of cold data and in spite of its infrequent use, this information must be nevertheless stored in back-up tape drives due to security and regulatory compliance reasons. Old photographs stored by users on Facebook is one such example of cold data; Facebook recently built an entire data center dedicated to storing such cold photographs [4]. Furthermore, as the percentage of cold data has reached the 80% over the last years, it is clear that the total cost for preserving this type of information increases significantly along time! However, all current storage media used for cold data storage (Hard Disk Drives or tape) suffer from two fundamental problems. First, the rate of improvement in storage density is at best 20% per year, which substantially lags behind the 60% rate of cold data growth. Second, current storage media have a limited lifetime of 5 (HDD) to 20 years (tape). As data are often stored for much longer duration (50 or more years), due to legal and regulatory compliance reasons, it must be migrated to new storage devices every few years, thus, increasing the price of ownership. It is therefore necessary to find new resources for the storage of digital data which exhibit higher capacity and longer life-span. It is necessary to note that while electricity and energy consumption in data centers are not directly an issue produced by back-up drives which are kept off-line, it is important to consider that the frequent migration of the data to new storage units requires the consumption of energy and electricity. Furthermore, most of the world's "cold" data are not well organized so as to be kept in off-line tape drives but gathers much space in on-line servers instead occupying significant disk space. Thus, the organization of this high percentage of data and the use of the extremely compact DNA solution as a security back-up unit to store it could significantly ease the workload of data centers and liberate enough space to host newly generated data decreasing this way the rate of constructing new ones. Some interesting solutions to the problems addressed above will be presented in the following paragraph.

1.2 Existing solutions

Longevity of data storage is not only important for financial or environmental reasons, but it is also crucial for preserving fundamental and invaluable cultural heritage for next generations. To deal with this problem, scientists have been studying the use of alternative means of higher durability.

Several projects, for instance at the University of Southampton [5] or at Hitachi [6], are currently considering new forms of very long-term digital storage, using molding silica glass, which estimated storage length time in the range of 100 million years. However, these projects are currently stymied by an important problem related to space: both developed at most a storage capacity that does not exceed 40 MBytes per inch, i.e., a very low value compared to the one Terabyte per square inch capacity reached by any standard hard disk.

Another very interesting solution proposes the use of the DNA molecule which is life's information-storage material as an alternative approach for digital data storage. Interestingly enough, recent works have proven that storing digital data in DNA is not only feasible, but also very promising as the DNA's biological properties allow the storage of a great amount of information in an extraordinary small volume, for centuries or even longer, with no loss of information. This paper aims to present some novel algorithms and techniques for the storage of digital information in DNA and thus the next sections are dedicated to explaining the term of DNA data storage as also in analyzing the most important assets and challenges.

2 DNA coding

DNA (deoxyribonucleic acid) is the support of heredity in living organisms. It is a complex molecule corresponding to a succession of four types of nucleotides (nts), adenine (A), thymine (T), guanine (G), cytosine (C). DNA can be double strand if one single strand binds on a complementary one according to the complementary base pairing rule (Chargaff's rule) [7] which denotes that DNA base pairs are always adenine with thymine (A-T) and cytosine with guanine (C-G). It is this quaternary genetic code that inspired the idea of DNA data storage which suggests that any binary information can be encoded into a DNA sequence of A, T, C, G.

More specifically, some important advances in the field of synthetic biology have allowed artificial synthesis of DNA strands in a laboratory (in vitro). The produced DNA is synthetic, but shares the same extraordinary properties as the real one. The only difference would be the fact that artificial synthesis does not require any particular DNA templates, allowing virtually any quaternary sequence of A, T, C, G to be synthesized in the laboratory. This means that the produced DNA will not necessarily contain any genes, which are DNA sequences responsible for producing life. On the contrary, any sequence of nucleotides can be assembled into a DNA strand. Consequently, using this technique any digital information can be synthesized into DNA on the condition that it has been previously encoded into a quaternary representation, a process called DNA coding. Once synthesized into the form of DNA, the encoded sequence can be stored in storage units such as materials or special small containers which can protect the DNA and provide a long-life storage. An example of such storage is the insertion of the DNA into special tiny capsules named as "DNA-shell" provided by the company Imagene which protect the molecule from contacts with water and oxygen and can ensure reliability for hundreds of years. The stored DNA can be retrieved whenever needed using some special machines, the sequencers. DNA sequencing is the biological process which allows reading any DNA strand and decoding it to provide their quaternary content. The two fundamental biological processes of DNA synthesis (writing) and sequencing (reading) work similarly to a noisy channel and thus construct an encoding workflow for digital storage.

2.1 Advantages

DNA possesses four key properties that make it a very promising candidate for archival storage of digital data:

- First, it is an extremely dense three-dimensional storage medium that has the theoretical ability to store 455 Exabytes in 1 g. In contrast, a 3.5" HDD can store 10 TB and weighs 600 g today.
- Second, DNA can last several centuries even in harsh storage environments. The decoding of the DNA of a woolly mammoth that had been trapped in permafrost for 40,000 years [8] is only one example which proves DNA's longevity in contrast to HDD and tape drives which have a life-span of 5 and 20 years, respectively.
- Third, it is very easy, quick, and cheap to perform in vitro replication of DNA; tape and HDD have bandwidth limitations that result in hours or days for copying large Exabyte-sized archives.
- Finally, DNA is life's information-storage material the main composition of which will never change. This comes in contrast to other means of storage which tend to change over the years according to the technological progress and so do the corresponding decoding devices. This means that in the long term, due to the incompatibility of the stored data with the new decoders, the stored content might not be decodable. For example almost 20 years ago, computers used to have a special input for floppy discs which is no longer the case. Consequently, any information that was stored in floppy disks is no longer accessible. On the contrary, DNA will exist forever in living organisms and even if the methods used for sequencing will further

Table 1 Comparison of DNA to other means of digital data storage table and data taken from [9]

	Hard disk	Flash memory	Bacterial DNA
Read/write speed (μs/bit)	$\sim$3,000-5,000	$\sim$100	< 100
Data retention (years)	> 10	> 10	> 100
Power usage (watts per gigabyte)	$\sim$0.04	$\sim$0.01–0.04	$< 10^{-10}$
Data density (bits per cm^3)	$\sim 10^{13}$	$\sim 10^{16}$	$\sim 10^{19}$

improve, the new sequencer's will always be adapted for decoding the exact same molecule.

The above properties reveal that storing digital data in DNA is an extremely promising solution. According to an article published in the journal of *Nature* [9], in a very rough theoretical estimation, scientists claim that 1 kg of DNA would be enough for storing all the world's digital information. Table 1 shows the results of the study published in [9] which compares the molecule of DNA to some widely used storage devices, the hard disks and flash memories.

2.2 The challenges

As described in section 2, DNA synthesis and sequencing are the key procedures which allow the archiving of digital data in DNA. While fundamental to the field of biology, those two processes introduce some important challenges.

To begin with, DNA synthesis requires the construction of DNA strands (oligos) of no more than 200–300 nts. This restriction stems from the fact that the error of the synthesis increases exponentially with the increase in the length of the oligos. To achieve a construction with low error probability, it is thus necessary to cut the encoded quaternary strand into smaller chunks and also yields the need for introducing some special headers to allow correct reconstruction at the decoding.

Secondly, both DNA synthesis and sequencing include some fine and delicate biological manipulations and thus those two processes are expensive and require several dollars per synthesized/sequenced oligo. It is therefore necessary to efficiently compress the data to be archived before it is stored in DNA.

Another important drawback rises from the process of DNA sequencing which is prone to errors creating insertions, deletions or substitutions of nucleotides in the decoded sequence. Luckily there are some special rules for the encoded strands which allow reducing the probability of error, but unfortunately without eliminating it. Those rules will be described in a later section.

Finally, a last but not negligible challenge lies in the longevity of DNA. While being an important asset which allows storage of digital data for centuries and maybe over, it also requires that the know-how of the decoding process should be passed on to the next generations to allow long-term decoding of data that had been stored many years ago. It is therefore very important, to safely preserve this information in durable materials while also ensuring that it is encoded in a way that will be easy for any new user to retrieve and understand. An interesting study on this particularly difficult challenge has been presented in [10]. Another interesting idea could be storing the decoding information in silica glass. Some interesting works for storing information in silica glass have been proposed in [11].

3 A constrained problem

DNA synthesis is a procedure that produces very low error rates as long as the DNA strands to be synthesized do not overpass the length of 150–300 nts. For longer sequences the synthesis error increases exponentially. Consequently, to reduce the probability of error, the DNA sequences to be synthesized need to be cut into short pieces and formatted in such a way that the initial sequence can be correctly reconstructed in the decoding part.

On the contrary, the biological procedure of DNA sequencing introduces much error which can not be neglected and therefore there is a need for dealing with the erroneous oligos produced by the sequencer. Studies have shown that the three main factors causing errors in the sequenced oligos are the following:

- *Homopolymers* Consecutive occurrences of the same nucleotide should be avoided [12].
- *G, C content* The percentage of G and C in the oligos should be lower or equal to the one of A and T [12].
- *Pattern repetitions* The codewords used to encode the oligos should not be repeated forming the same pattern throughout the oligo length [13].

Taking into account all the above rules, the sequencing error can be reduced. Consequently, to be efficient, any DNA coding algorithm should respect the above rules to reduce as much as possible the probabilities of sequencing error. In addition to this, it is important to mention that to achieve a reliable decoding it is also necessary to introduce some redundancy to the encoded data so as to allow the application of error detection and correction. To this end, the encoding workflow which is adopted by the state-of-the-art works when addressing DNA data storage has been structured according to those constraints. The main structure of this workflow is detailed in the following sections.

3.1 General workflow

In the previous sections, we presented the reasons for which DNA is an eco-friendly solution offering the possibility of storing a great amount of information in a very small volume while also promising longevity of the stored data. We also explained that DNA coding is a multi-disciplinary subject which is inspired by the quaternary code of DNA and highly depends on the biological processes of DNA synthesis and sequencing. Those two methods are reminiscent of a digital noisy source channel which adds any type of noise to the transmitted data. Therefore, the process of DNA data storage can be thought of as a classical encoding workflow for the transmission of 4-ary data through a noisy channel. The general coding scheme for DNA data storage is depicted in Fig. 1.

Although this process might seem simple, it is very important to denote that this is only a very rough and simple presentation of the general DNA coding workflow. However, as briefly explained in Sect. 2.2, DNA synthesis and sequencing are very delicate and complex processes which introduce some important constraints when it comes to

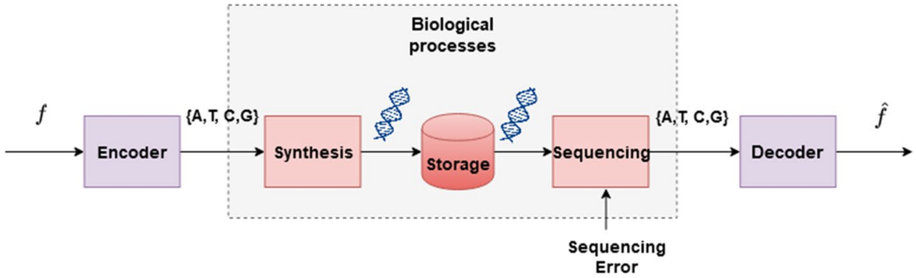


Fig. 1 Main component parts of a typical DNA storage process

the encoding of digital data. In the next sections, we will describe how the encoding and decoding of the general workflow is adapted to the needs of DNA data storage forming a more complete and extended workflow.

3.2 Encoding

Until this point, it is clear that the encoding of digital data in DNA is strongly constrained by the biological part of the process. More precisely, to sum up the main obstacles which have been discussed in the previous sections, the encoding should provide a quaternary code which will respect the sequencing restrictions to ensure robustness and the length of the DNA oligos to be synthesized should not be higher than 150–300 nts. Consequently, the structure of a reliable encoder for DNA coding contains the following sub-parts.

The first step in the encoding workflow is the construction of a dictionary of codewords composed by the symbols A, T, C and G similarly to the nucleotides of the DNA molecule. Those codewords should provide a robust encoding when assembled at a long sequence. This means that the quaternary strands should not contain homopolymers, high G,C content compared to the content of A and T and finally it should not contain repeated patterns.

The next sub-process of a DNA workflow is a mapping function which assigns input symbols to codewords of the quaternary code. This function can be a simple one-to-one function or a more sophisticated one.

Finally, as the oligo length is restricted due to the synthesis limitations to avoid errors, it is necessary to adopt some formatting function for cutting the produced long encoding into shorter oligos and adding special headers for the reconstruction of the input at decoding. Those headers can contain information for the address of the data chunk in the original long sequence, information for any necessary encoding parameters as well as information about the input characteristics as for example the size. A general overview of the encoding of data in DNA is described by Fig. 2.

3.3 Decoding

Since DNA data storage is a process which is prone to both writing and reading errors, the decoding should include some techniques to predict, detect or even to correct the sequenced data. As explained in Sect. 3, the addition of redundancy is necessary for the detection of errors and can be easily achieved using the method of polymerase

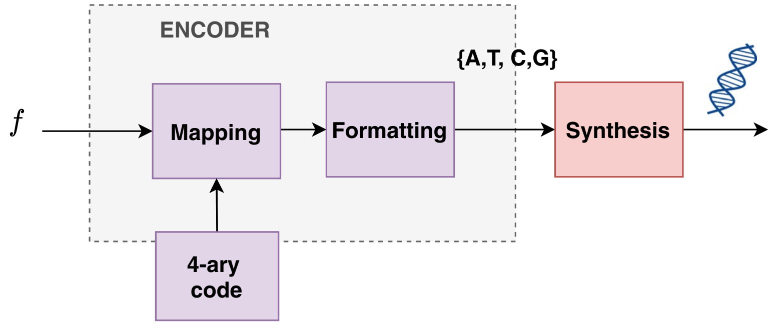


Fig. 2 Encoding of digital data in DNA

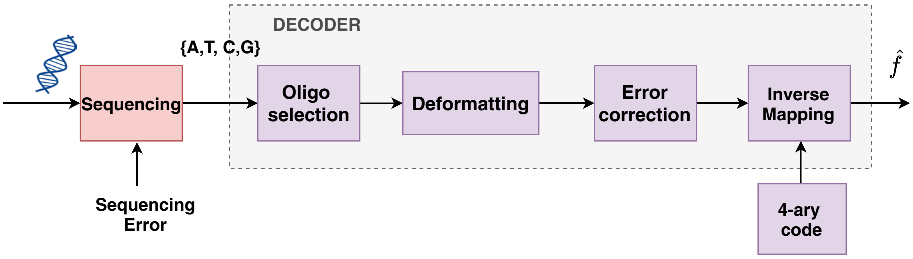


Fig. 3 Decoding of digital data stored in DNA

chain reaction (PCR) amplification which is using an enzyme which, called Polymerase, for the replication (amplification) of DNA strands. This particular enzyme can initialize the process of replication in the presence of some specific short DNA sequences which are called primers. PCR is applied during both DNA synthesis and sequencing. Consequently, in the output of the sequencer there will be multiple copies of each synthesized oligo. Each copy might contain different types of errors in various positions and this yields the need for selecting the most representative copy for each oligo. This selection occurs after aligning the different erroneous copies which is a process that allows computing a consensus sequence for each oligo. Another simpler method for the oligo selection is based on finding the most frequent among all copies. Those processes can be then followed by some error-correction algorithm to treat any remaining errors for obtaining an error free decoding. It is important to mention that the efficiency of the error correction highly depends on the methods and machines that have been used during sequencing as some particular sequencers can cause higher error rates than others and can therefore create stronger distortion. Finally, using the inverse mapping function one can retrieve the digital information which had been stored in DNA. An overview of the decoding process is described by Fig. 3.

4 Existing works

DNA data storage is a relatively new field of research and thus the state of the art is limited to a few pioneering works which have, however, contributed widely to this emerging topic.

4.1 First references to the idea of DNA data storage

The idea for storing digital data using the DNA molecule ages back in the late 1950s when Soviet physicist Mikhail Samoilovich Neiman and cybernetician Norbert Wiener expressed ideas regarding the possibility of recording, storage, and retrieval of information on synthesized DNA and RNA molecules [14, 15]. However, the first attempt of DNA data storage came in 1988 when the artist Joe Davis and researchers from Harvard collaborated for storing a 5 × 7 matrix in a DNA sequence in *E. coli*, which once decoded, formed a picture of an ancient Germanic rune representing life and the female Earth [16]. In the matrix, ones corresponded to dark pixels while zeros corresponded to light pixels. In 2007, at the University of Arizona scientists created a device which used addressing molecules to encode mismatch sites within a DNA strand. These mismatches were then able to be read out by performing a restriction digest, thereby recovering the data. This was the starting point for various interesting works that followed, introducing multiple novel encoding algorithms that brought DNA data storage to practice and contributed widely to this emerging topic. In the following sections, we will present the most widely used studies in the bibliography and briefly analyze the proposed solutions.

4.2 The first application of DNA data storage by Church et al.

In 2012, George Church et al. encode for the first time a 659-Kbyte book that was co-authored by Church in DNA. In their experiment, the authors used a very simple encoding, by randomly translating zeros to A or C and ones to T or G [12]. The encoded sequence was then written onto a microchip as a series of DNA fragments using an ink-jet printer. The encoding resulted in 54,898 oligonucleotides, containing 96 bases of data along with a special 22-base sequence at each end to allow the fragments to be copied in parallel using the PCR amplification, and a unique, 19-base "address" sequence to denote the segment's position in the original document.

The resulting PCR amplified oligos were then read back using an Illumina sequencer to retrieve the original text. The storage density of the DNA fragments produced by this method was estimated to be more than 700 terabytes per cubic millimeter. This result represented the largest volume of data ever artificially encoded in DNA, and proved that data density for DNA is several orders of magnitude greater than that of state-of-the-art storage media as shown in their plot in Fig. 4.

Not only did this work make a pioneering step to prove the feasibility of using DNA as an alternative means of storage while demonstrating the extraordinary capacity compared to conventional storage devices, but it also revealed that sequencing can be an error-prone process. By analyzing the different errors which occurred during sequencing, this work provided a first study of the main constraints to be respected during the encoding.

After this important first step, several works followed to propose new encoding techniques, attempting to provide a robust encoding which would allow reducing the sequencing errors obtained in this study.

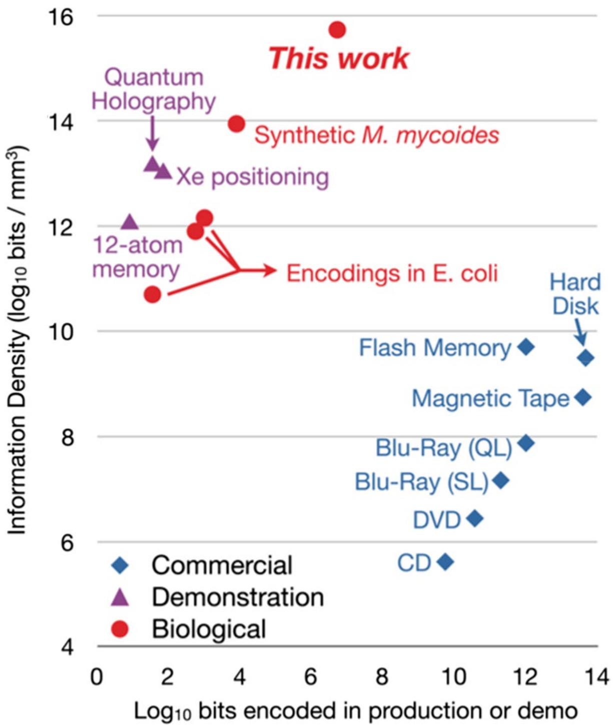


Fig. 4 Results of the study carried out by Church et al. Image taken from the authors' publication in [12]

4.3 First biologically constrained encoding by Goldman et al.

In 2013, Goldman et al. [17] proposed a novel algorithm for encoding digital data in binary so as to respect the main sequencing constraints. The encoding proposed using ternary Huffman algorithm to encode each byte of a binary sequence into the digits 0, 1 and 2. Those digits are then associated to three of the symbols A, T, C and G omitting the symbol that has been used for the encoding of the previous digit, so as to ensure that no base is used twice in a row. This strategy avoided the creation of homopolymers while still making use of DNA's four-base potential. To enhance the reliability of the oligos and determine the data's position in the original file, Goldman's team synthesized oligonucleotides carrying 100 bases of data, with an overlap of 75 bases between adjacent fragments, so that each base was represented in four oligonucleotides creating a fourfold redundancy. Even so, the researchers lost two 25-base stretches during sequencing, which had to be manually corrected before decoding. The encoding followed in this study is explained in Fig. 5.

The code construction proposed by this work has been thereby used by Microsoft researchers in their later works.

4.4 Introduction of Reed–Solomon codes by Grass et al.

To deal with the remaining sequencing errors, in 2015, Grass and his team [18] have proposed for the first time the use of Reed–Solomon codes to introduce error correction in the encoding. More precisely, in this work the authors proposed the mapping

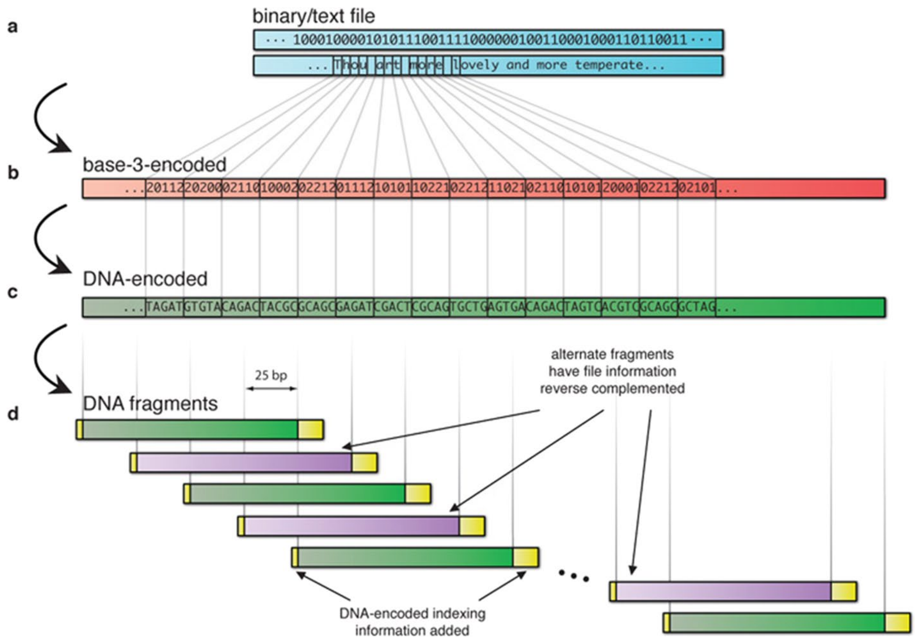

a
binary/text file
··· 1000100001010111001111000000100110001000110110011 ···
... Thou art more lovely and more temperate...
b
base-3-encoded
...20112202000211010002022120111210101102210221211021021101010120001022120210 1...
c
DNA-encoded
...TAGATGTGTACAGACTACGCGCAGCGAGATCGACTCGCAGTGCTGAGTGACAGACTAGTCACGTCGCAGCGCTAG...
d
DNA fragments
25 bp
alternate fragments have file information reverse complemented
DNA-encoded indexing information added

Fig. 5 Goldman et al. encoding. Image taken from the authors' publication in [17]

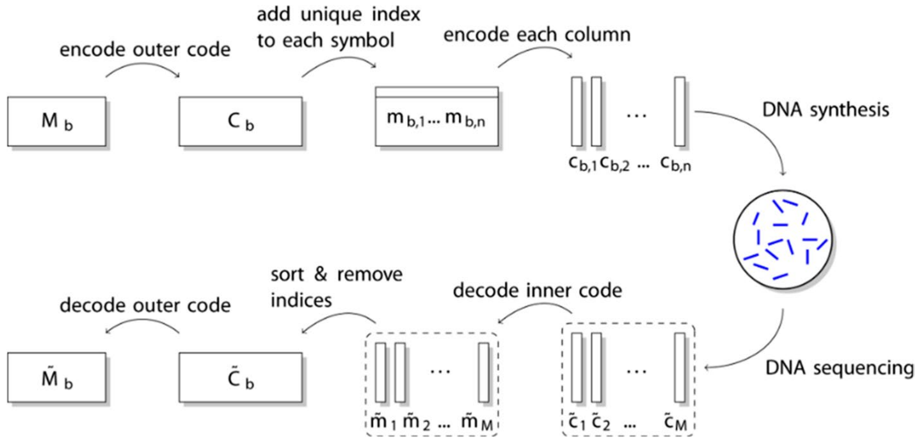
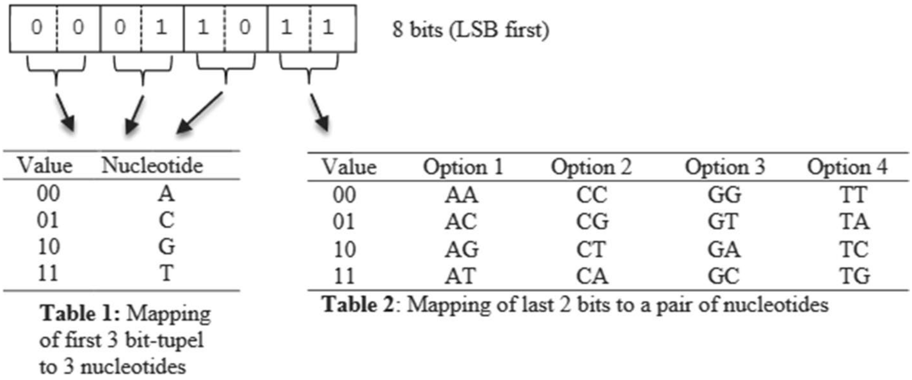


Fig. 6 Grass et al. encoding. Image taken from the authors' publication in [18]

of the data to blocks which contain elements from Galois Field 47 [GF(47)]. The column of each block is extended using a unique index consisting of elements in GF(47). The extended columns are then encoded to DNA by mapping each of the GF(47) elements to a triplet of nucleotides while ensuring that there is no repetition of the same base in the two last positions ensuring that homopolymers are avoided. Each encoded column represented a DNA fragment to be synthesized and stored in silica to ensure long-term storage without corruption of the DNA. In their study the authors reported perfect retrieval of 83 kB of data encoded using a Reed–Solomon code, an error-correcting code used in CDs, DVDs, and some television broadcasting technologies such



Value	Nucleotide
00	A
01	C
10	G
11	T

Table 1: Mapping of first 3 bit-tupel to 3 nucleotides

Value	Option 1	Option 2	Option 3	Option 4
00	AA	CC	GG	TT
01	AC	CG	GT	TA
10	AG	CT	GA	TC
11	AT	CA	GC	TG

Table 2: Mapping of last 2 bits to a pair of nucleotides

Fig. 7 Blawat et al. encoding. Image taken from the authors' publication in [22]

as the Advanced Television Systems Committee (ATSC) broadcasting. The storage workflow is shown in Fig. 6.

4.5 First random-access implementation by Yazdi et al.

At the same year (2015) Yazdi et al. [19] have introduced an important way for allowing random access using specific and robust addressing in the encoding! In their study, the authors proposed the addition of some especially designed primers in both ends of the encoded data to allow selective PCR amplification of particular oligos instead of amplifying the full oligo pool. The primers were specially designed to be robust to sequencing errors and the encoding DNA words for each oligo depend on the corresponding primer. More precisely, for each oligo the DNA code is constructed by ensuring there is no correlation of the payload to the oligo's addressing primer as this would create secondary structures which can be catastrophic and can lead to full oligo loss during sequencing.

In a later study published in 2017 [20], the authors provided an experiment testing the efficiency of their proposed encoding using the MinION-Oxford Nanopore's handheld sequencer for the reading of the DNA while also using JPEG compression to reduce the synthesis cost. This study has devised error-correcting algorithms specifically for the kinds of mistakes the MinION makes. The result is an error-free read-out, demonstrated earlier this year when the team stored and sequenced around 3.6 kB of binary data coding for two compressed images. Finally, in a co-authored study of Chao Pan in [21], the research group proposed the use of inpainting techniques to correct discolorations of the decoded image which occurred by corruptions introduced during sequencing.

4.6 Reed–Solomon codes on headers by Blawat et al.

In 2016 Blawat et al. [22] published another interesting method for constructing a robust quaternary code. In their work, the authors presented a new method for creating a quaternary code by encoding each byte of some digital data to 5 nucleotides using the following algorithm. To begin with each of the first three pairs of bits are encoded to 1 nucleotide and placed in the first, second and fourth position, respectively, of the resulting DNA word. Then the last pair of bits can be encoded to a pair of nucleotides and will be placed in the third and fifth position of the resulting DNA word. The above encodings are performed as depicted in Fig. 7. As a result, for each byte, there are provided 4

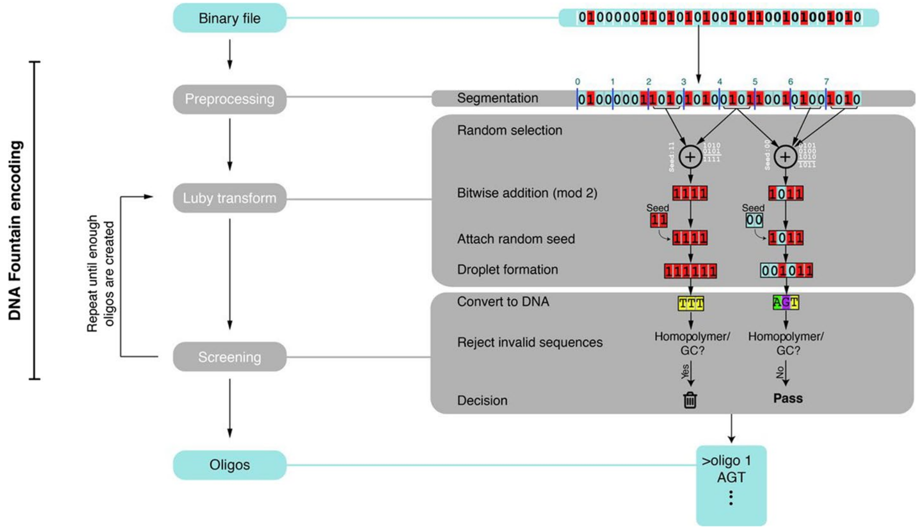


Fig. 8 Erlich et al. encoding [23]

different DNA words. To ensure that the limitation concerning the maximum run-length is respected, the 4 options are filtered so as to not create homopolymers.

To do so the authors propose keeping only the options that do not violate the following rules:

- The first three nucleotides shall not be the same.
- The two last nucleotides shall not be the same.

With the above described constraints, at least 2 valid DNA symbols can be found for every data byte, thus introducing some redundancy which can be used for error detection. More precisely, the authors proposed separating the different codeword options into different predefined clusters and encode each input byte using the encoding of a specific cluster according to the byte's position. For example, one option would be to use codewords from cluster A to represent even positions and cluster B for odd byte positions. Thus, in the case where an error alternates a codeword expected to be found in one cluster to another one that belongs to some other cluster, error detection is possible. Furthermore, in this work the authors proposed robustifying the addressing headers using Reed–Solomon codes to allow a more reliable decoding.

4.7 DNA coding using fountain codes by Erlich et al

At the same year (2016), Columbia University researchers Yaniv Erlich and Dina Zielenski published a method based on a fountain code [23], an error-correcting code used in video streaming. As part of their method, they used the code to generate many possible oligos on the computer, and then screened them in vitro for desired properties. Focusing only on sequences free of homopolymers and high G content, the researchers encoded and read out, error-free, more than 2 MB of compressed data—stored in 72,000 oligonucleotides—including a computer operating system, a movie, and an Amazon gift card. Their encoding is depicted in Figure 8 and follows the following steps.

First, the input binary file is segmented in partitions. Then using a Luby transform, droplets of bits are created by selecting randomly segments from the input sequence and bit-wise adding them attaching also the random seed used for the selection. The resulting bit droplets are then encoded into quaternary and scanned for satisfying the biological constraints of GC content and homopolymers. Encoded droplets which do not respect the above restrictions are discarded while the rest are used for creating the oligos. This process is repeated until enough oligos are produced resulting in a densely compressed encoding that reaches a capacity of 1.98 bits/nt.

4.8 Efficient end-to-end workflow by Microsoft researchers

In 2016, Borhholt et al. in a Microsoft research presented a DNA-based archiving system using the quaternary code introduced by Goldman et al. In this study, they improved the encoding by avoiding the fourfold redundancy using themselves addressing primers for allowing random access [24]. Researchers in Microsoft have then in 2017 presented some extra studies to improve their results using a clustering algorithm to cluster and correct the multiple reads provided by the sequencer allowing a better reconstruction quality [25, 26]. Finally, in 2019, a Microsoft team successfully encoded the word "hello" in snippets of fabricated DNA and converted it back to digital data using a fully automated end-to-end system, which is described in [27].

4.9 Forming a novel DNA database by Appuswamy et al.

In [28], the authors present a novel way for using DNA coding to encode structured database information and implement database operations. In this work, the structured database information (i.e., relational tables) and the database operations are encoded in two different ways. The first encoding exploits the inherent structure in databases. In other words, each attribute of a record in a table can be linked to the corresponding record using the primary key. Thus, attributes of the same record can be distributed across different DNA sequences without the need for addressing, using only the primary key, reducing this way the space needed for the address.

The information is compressed using dictionary encoding and the dictionary is encoded in DNA as well. Subsequently, as many attributes as possible are stored in a DNA sequence along with the primary key (to link together attributes of the same record). A parity nucleotide is also added to each DNA sequence for error detection. After sequencing, the parity nucleotide and length of the DNA sequence are used to discard invalid sequences. The remaining sequences are aligned to compute a consensus. In the experiments, based on a subset of the database benchmark TPC-H, multiple tables are encoded, synthesized, sequenced and fully recovered.

This work takes the first step in addressing DNA data storage from the perspective of data management systems by presenting an architecture for using DNA as the archival tier of a relational DBMS. The experiments show that it is not only feasible to archive and restore data using synthetic DNA, but also exploit database knowledge for optimizing the encoding and decoding process, and even execute SQL operations directly over DNA

4.10 Closed-loop optimization encoding solutions for image storage in DNA

All the studies of the state of the art which have been described above, are providing some way for building a quaternary encoding of digital data by respecting the biological restrictions discussed in Sect. 3. Each one of those encodings exhibits different advantages and weaknesses and since the subject is still very new, it is necessary to provide new encoding ideas which can help enriching the existing studies and improve the quality of the stored data.

As the main drawback of DNA data storage is the high synthesis cost, the encoding methods proposed in the bibliography attempt to improve the storage capacity while also being robust to sequencing errors. To this end, most of the above studies have proposed compressing images with JPEG before encoding. However, no study has proposed a method for controlling this compression such that it provides a closed-loop solution which can allow selecting the best compression parameters for a given coding potential. In our studies in [29], we included a source allocation algorithm which offers the possibility of not only reducing the synthesis cost, but also promising an optimal quality of the stored image for a predefined encoding rate and thus a given synthesis cost. As a low complexity source allocation requires a fixed-length code, we also propose a new efficient algorithm for the construction of a robust fixed-length DNA code that facilitates the nucleotide allocation method. We also introduce two different mapping methods. The first one which is presented in [29] deals with pattern repetitions which might be the cause of error increase in the Illumina sequencers and has not been tackled by previous studies, and the second one which is presented in [30] aims in decreasing the visual impact of substitution errors which may remain after error correction. The reason for implementing a fixed-length encoder stems from the fact that variable-length coding is less robust to sequencing errors. In other words, in case of an error, variable-length coding is prone to losing important information about the structure of the encoded data which can result in wrong reconstruction of the input image. To prove this claim, in [31] we also implemented a variable-length encoder which is inspired by the classical binary JPEG encoder. This idea has been created thanks to the JPEG Ad Hoc group which has recently shown interest in building a new JPEG standard for the purpose of image coding in DNA. Our proposed solution uses a modified workflow of the classical JPEG standard for binary coding which optimizes the compression of the input image according to a constrained quaternary code, producing a compressed nucleotide stream which is robust to sequencing error.

5 Comparing the different DNA coding solutions

In this section, we will compare and comment on the assets and flaws of some of the solutions presented in the previous sections. As explained the first attempt of encoding digital data in DNA is described by Church et al. in [12]. The main importance of this work is that it made the very first step in initializing research in this emerging field of storage. In this work, each binary bit is encoded to one nucleotide giving a total coding potential of 1 bit/nucleotide. To improve the coding potential, as well as the robustness of the encoding to errors, following works have adopted some more complicated encoding algorithms. More precisely Goldman et al. in [17], have proposed an algorithm that provides a quaternary encoding which avoids homopolymer runs to improve the quality of sequencing. This work reached a coding potential of 1.58 bits/nt. However, this encoding algorithm does not allow control of the C,G percentage and can create pattern repetitions which is an ill-case leading to a higher error probability at the phase of sequencing [13]. However, since the works of Goldman et al. use Huffman codes which rely on the frequency distribution of the input in such an encoding can allow an efficient variable-length encoding which relies on the characteristics of the source. Another asset of this algorithm is the fact that it can be applied to any type of input data without being restricted to binary representations. Nevertheless, the use of Huffman codes yields the need for transmitting the distribution of the source to the decoder.

The works of Yazdi in [19] introduces the use of addressing fields to allow random access in the reading and writing of the DNA oligos. As the addressing primers contain fundamental information which should be correctly retrieved, the authors propose a novel encoding for DNA data storage which is built such that secondary structure is avoided in the encoded DNA strands. More precisely, the DNA code differs for each oligo and is constructed according to the oligo's address field. More precisely, the code is constructed ensuring that there is no strong correlation between the encoding codewords and the addressing header which could lead to the oligo binding on itself and therefore leading to important loss in sequencing. According to a later publication [20], this encoding can reach a coding potential of 1.57 bits/nt. While this encoding avoids undesirable cross-hybridization problems during the process of oligo selection and amplification and can allow some limited error correction, one possible drawback is the fact that the code is varying according to the addressing primer it is not fixed throughout the encoding process.

Blawat et al. [22] have proposed using 5 nucleotides to encode 8 bits of information using a method for avoiding homopolymers. Furthermore, the encoding inserts some randomization in the selection of the codewords which could potentially be exploited for avoiding pattern repetitions as well as for correcting some types of errors that may occur. The coding potential of this method is 5 nucleotides per 8 bits of binary sequence which is equivalent to 1.6 bits/nt. Nevertheless, a drawback of this algorithm is the fact that it can only be applied for transcoding. In other words, it can only be applied to encode binary information to a quaternary DNA representation.

In the works of Grass et al. [18], the encoding is performed using Reed–Solomon codes. This encoding achieves a coding potential of 1.187 bits/nt introducing some extra redundancy in order to allow error correction. Nevertheless, similarly to [22], it is only applicable to binary stream.

Bornholt et al. in [24] have applied the same encoding as in [17], improving the encoding scheme and avoiding the fourfold redundancy which is suggested by the latter and synthesizes each DNA chunk in 4 shifted copies of the initial sequence. For further information about the fourfold redundancy, the reader can refer to [17].

Erlich et al. [32] have implemented an encoding using fountain codes to reach an extremely high coding potential of 1.98 bits/nt. Similarly to most of the previously mentioned works, despite the efficiency in terms of information density, this type of encoding is only applicable to binary information while also being very expensive in computational cost.

Table 2 Comparison to previous works—coding potential: maximal information content of each nucleotide before indexing or error correcting

Parameter	Church et al. [12]	Goldman et al. [17]	Yadzi et al. [19]	Grass et al. [18]	Bornholt et al.[24]	Blawat et al. [22]	Erlich et al. [32]	Our work (raw data)
Input data (Mbytes)	0.65	0.75	0.017	0.08	0.15	22	2.15	0.26
Coding potential (bits/nt)	1	1.58	1.57	1.78	1.58	1.6	1.98	1.6
Redundancy	1	4	1	1	1.5	1.13	1.07	1
Error correction	No	Yes	Yes	Yes	No	Yes	Yes	No

Redundancy: excess of synthesized oligos to provide robustness to dropouts. Error correction/detection: the presence of error-correction code to handle synthesis and sequencing errors. Full recovery: DNA code was recovered without any error. Net information density: input information in bits divided by the number of synthesized DNA nucleotides (excluding primers)

In our work in [29], we introduced an algorithm robust to sequencing noise which respects all the necessary constraints imposed by the sequencing process. It has the important asset of being applicable to any input representation without being restricted to binary inputs. It is fixed-length and simple in computational cost and therefore it can be embedded to any encoding workflow. More precisely, as proposed in our works in [29] and [31] when used in a "closed-loop" optimization of the encoding, it can allow controlling the high DNA synthesis cost while maximizing the encoding quality. Our encoder allows an efficient coding potential of 1.6 bits/nt. However, as in every fixed-length algorithm, it cannot reach an extremely high value as the one proposed by [23].

A comparison of the coding potential that has been reached using the different encoding approaches proposed by the state of the art is presented in Table 2.

6 Conclusions and discussion

As DNA data storage is a very challenging multi-disciplinary field of study that highly depends on biological manipulations, it is expected to evolve along with the changes in the methods and the machines used for DNA synthesis and sequencing. Thus the encoding methods might change in the following years so as to respect different encoding constraints. In addition to this, since this is a relatively new topic of research with great potential in future applications, it is sure to attract much interest in the next few years, hoping that the more it will be studied, the more the cost of biological processes such as DNA synthesis and sequencing will be reduced.

It is very encouraging to notice that there is already a great interest in the topic by many different research groups around the world. Namely, during our studies, our research group had the chance to collaborate with some of those teams through the OligoArchive project Horizon 2020[1] which is founded by the European Union. This collaboration includes the I3S/CNRS laboratory, the IPMC/CNRS and EURECOM which are located in Sophia Antipolis in France, as well as the Imperial College of

[1] https://oligoarchive.eu

London in the UK and the startup of Helixworks which is synthesizing DNA and is located in Ireland. This project aims in the creation of a prototype system that will allow the research of the whole cycle from encoding to the sequencing of data to DNA. Furthermore, recently, the JPEG community has launched an Ad Hoc Group on Digital Media Storage using DNA [33], in which we have the honor to participate as invited experts on the topic of DNA data storage. At the same time, Microsoft, Western Digital, Twist Bioscience and Illumina have formed an alliance for building an efficient prototype for DNA data storage. We therefore hope that these collaborations will create some fruitful ideas that will help the field advance rapidly enough to be soon used in practice.

Another important point to be discussed is the fact that DNA data storage is intended for the archiving of digital data to be decoded in the very long term. The reading of DNA will always be guaranteed, since the molecule of DNA exists in every living organism and thus there will always be some machine for its reading. However, it is fundamental to find a way for ensuring that the decoder as well as the information for the decoding will be available when needed to allow correct reconstruction after reading. Some interesting works on digital preservation propose some solutions for creating durable ways for storing the information for decoding while also expressing it in a way that will be understandable by anyone in the future that might have no previous knowledge on the encoding. Such solutions include the storage of the decoding information in microfilms, as proposed by the company of EUPALIA in France [10] and in their latest collaboration with EURECOM in [34] or in silica glass. Some interesting works for storing data in silica glass have been proposed in [11].

In all, DNA data storage is a very promising new field of research which is expected to play a significant role in the solution of fundamental challenges of digital data storage. However, since it constitutes a multi-disciplinary subject which is highly constrained by some limitations of the biological manipulations, there are multiple challenges to be addressed in the encoding of digital data in DNA. All the works described in this article set the groundwork for further improvement for proving that DNA data storage is no longer considered as a science-fiction scenario, but instead will make a huge breakthrough in next years to come by giving new breath to the existing storage solutions.

Abbreviations

DNA	Deoxyribo nucleic acid
IDC	International Data Corporation
PCR	Polymerase chain reaction

Author contributions

Both authors have participated in the structure of the manuscript. Both authors have read and approved the final manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant agreement No. 863320

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 24 February 2021 Accepted: 4 October 2022
Published online: 29 October 2022

References

1. M. McNerney, The data center dilemma: is our data destroying the environment? The DataCenter Knowledge (2019)
2. S. Jeremy, How long do hard drives last? Lifespan and Signs Of Failure. Prosoft Eng. Inc (2017)
3. I. Group, Lifespan of LTO tapes. ISC Group (2012)
4. R. Miller, Facebook builds exabyte data centers for cold storage. Retrieved June **8**, 2014 (2013)
5. Y. Lei, M. Sakakura, L. Wang, Y. Yu, R. Drevinskas, P.G. Kazansky, Low-loss geometrical phase elements by ultrafast laser writing in silica glass. In: CLEO: Applications and Technology, pp. 4–4 (2019). Optical Society of America
6. Y. Shimotsuma, K. Miura, H. Kazuyuki, Nanomodification of glass using fs laser. Int. J. Appl. Glas. Sci. **4**(3), 182–191 (2013)
7. E. Chargaff, R. Lipshitz, C. Green, Composition of the desoxypentose nucleic acids of four genera of sea-urchin. J. Biol. Chem. **195**(1), 155–160 (1952)
8. E.M. Prager, A.C. Wilson, J.M. Lowenstein, V.M. Sarich, Mammoth albumin. Science **209**(4453), 287–289 (1980)
9. A. Extance, How DNA could store all the world's data. Nature **537**(7618) (2016)
10. V. Joguin, Passive digital preservation now & later (2019)
11. A. Chatzieleftheriou, I. Stefanovici, D. Narayanan, B. Thomsen, A. Rowstron, Could cloud storage be disrupted in the next decade? In: 12th {USENIX} Workshop on Hot Topics in Storage and File Systems (HotStorage 20) (2020)
12. G.M. Church, Y. Gao, S. Kosuri, Next-generation digital information storage in DNA. Science, 1226355 (2012)
13. T.J. Treangen, S.L. Salzberg, Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. **13**(1), 36 (2012)
14. M. Neiman, On the molecular memory systems and the directed mutations. Radiotekhnika **6**, 1–8 (1965)
15. N. Wiener, Machines smarter than men? interview with dr. norbert wiener. noted scientist. US News & World Report, 84–86 (1964)
16. G.M. Skinner, K. Visscher, M. Mansuripur, Biocompatible writing of data into DNA. J. Bionanosci. **1**(1), 17–21 (2007)
17. 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**(7435), 77 (2013)
18. 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**(8), 2552–2555 (2015)
19. S.H.T. Yazdi, Y. Yuan, J. Ma, H. Zhao, O. Milenkovic, A rewritable, random-access DNA-based storage system. Sci. Rep. **5**, 14138 (2015)
20. S.H.T. Yazdi, R. Gabrys, O. Milenkovic, Portable and error-free DNA-based data storage. Sci. Rep. **7**(1), 1–6 (2017)
21. C. Pan, S. Yazdi, S.K. Tabatabaei, A.G. Hernandez, C. Schroeder, O. Milenkovic, Image processing in dna. arXiv preprint arXiv:1910.10095 (2019)
22. 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. Proc. Comput. Sci. **80**, 1011–1022 (2016)
23. Y. Erlich, D. Zielinski, Capacity-approaching DNA storage. bioRxiv, 074237 (2016)
24. J. Bornholt, R. Lopez, D.M. Carmean, L. Ceze, G. Seelig, K. Strauss, A DNA-based archival storage system. ACM SIGOPS Oper. Syst. Rev. **50**(2), 637–649 (2016)
25. 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.: Scaling up DNA data storage and random access retrieval. bioRxiv, 114553 (2017)
26. C. Rashtchian, K. Makarychev, M. Racz, S. Ang, D. Jevdjic, S. Yekhanin, L. Ceze, K. Strauss, Clustering billions of reads for DNA data storage. In: Advances in Neural Information Processing Systems, pp. 3362–3373 (2017)
27. C.N. Takahashi, B.H. Nguyen, K. Strauss, L. Ceze, Demonstration of end-to-end automation of DNA data storage. Sci. Rep. **9**(1), 1–5 (2019)
28. R. Appuswamy, K. Le Brigand, P. Barbry, M. Antonini, O. Madderson, P. Freemont, J. McDonald, T. Heinis, Oligoarchive: Using DNA in the DBMS storage hierarchy. In: CIDR (2019)
29. M. Dimopoulou, M. Antonini, P. Barbry, R. Appuswamy, A biologically constrained encoding solution for long-term storage of images onto synthetic DNA. In: EUSIPCO 2019 (2019)
30. M. Dimopoulou, E.G. San Antonio, M. Antonini, A quaternary code mapping resistant to the sequencing noise for dna image coding. In: 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6 (2020). IEEE
31. M. Dimopoulou, E. Gil San Antonio, M. Antonini, A jpeg-based image coding solution for data storage on DNA. In: EUSIPCO (2021)
32. Y. Erlich, D. Zielinski, DNA fountain enables a robust and efficient storage architecture. Science **355**(6328), 950–954 (2017)
33. Dna-based media storage: State-of-the-art, challenges, use cases and requirements version 2.0. JPEG Ad Hoc group, ISO/IEC JTC 1/SC29/WG1M89031 (2020)
34. R. Appuswamy, V. Joguin, Universal layout emulation for long-term database archival. In: Submitted on ArXiV, 8 September 2020 (2020). http://www.eurecom.fr/publication/6335

Publisher's Note