ISO/IEC JTC 1/SC 29/WG 1
(ITU-T SG16)

Coding of Still Pictures

JBIG
Joint Bi-level Image Experts Group

JPEG
Joint Photographic Experts Group

TITLE: DNA-based Media Storage: State-of-the-Art, Challenges, Use Cases and Requirements version 2.0

SOURCE: Contributors: Marc Antonini, Luis Cruz, Melpomeni Dimopoulou, Touradj Ebrahimi (editor), Fernando Pereira (editor), António Pinheiro, Mohamad Raad

PROJECT: JPEG DNA Exploration

STATUS: Public

REQUESTED ACTION: For information and feedback

DISTRIBUTION: WG1

Contact:
ISO/IEC JTC 1/SC 29/WG 1 Convener – Prof. Touradj Ebrahimi
EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland
Tel: +41 21 693 2606, Fax: +41 21 693 7600, E-mail: Touradj.Ebrahimi@epfl.ch
DNA-based Media Storage

State-of-the-Art, Challenges, Use Cases and Requirements

Version 2.0

Executive Summary

DNA is a macromolecule which is essential for any form of life and is made of simple units that line up in a particular order within this large molecule. The order of these units usually carries genetic information for a specific life organism, similar to how the order of letters in a text carries information. In practice, this means that DNA molecules may be artificially created with specific unit orders, notably to store some relevant sequence of information.

In digital media information, notably images, the relevant representation symbols, e.g. quantized DCT coefficients, are expressed in bits (binary units) but they could be expressed in any other units, for example the DNA units which follow a quaternary (4-ary) representation basis. This would mean that artificial DNA molecules may be created with a specific DNA units’ configuration which store some media representation symbols or bits, e.g. the symbols or bits of a JPEG compressed image, thus leading to DNA-based media storage as a form of molecular data storage.

To make it more interesting, the DNA data storage density seems to be extremely high, notably beyond any available storage technology, but also energy friendly and very durable. In this context, DNA storage implies DNA synthesis/storage and DNA sequencing/access, which are currently rather complex and expensive processes, but should become increasingly affordable in the coming years.

This exciting story directly leads to the purpose of this document, which is basically to review and discuss:

1. DNA-based media storage basics, architectures and technology state-of-the-art
2. DNA-based media storage challenges
3. DNA-based media storage use cases and requirements
4. Main players in DNA-based media DNA media storage
5. JPEG role and next steps in JPEG DNA-based media storage activity

As a minimum, JPEG committee can launch an activity to convert its existing image coding formats from compressed binary representation to compressed DNA 4-ary representation, according to appropriate requirements. Standardized image coding solutions along with complementary appropriate tools, such as error resiliency and associated metadata, which particularly suit the requirements of DNA information storage, are good directions for JPEG to explore.
1. Background and Motivation

DNA is a macromolecule which is essential for any form of life and is made of simple units that line up in a particular order within this large molecule. The order of these units usually carries genetic information for a specific life organism, similar to how the order of letters in a text carries information. However, it is also possible to create artificial DNA molecules with specific DNA unit orders, notably to store some relevant sequence of information.

In digital media information, notably images, the relevant representation symbols, e.g. quantized DCT coefficients, are expressed in bits (binary units) but they could be expressed in any other units, for example the DNA units which follow a 4-ary representation basis. This would mean that artificial DNA molecules may be created with a specific DNA unit configuration which store some media representation symbols or bits, e.g. the symbols or bits of a JPEG compressed image, thus leading to DNA-based media storage as a form of molecular data storage. In this context, DNA storage implies DNA synthesis/storage and DNA sequencing/access, which are currently rather complex and expensive processes, but should become increasingly affordable in the coming years.

To make this storage mechanism more interesting, the DNA data storage density seems to be extremely high, notably beyond any available storage technology. Moreover, DNA-based storage is also extremely stable, as demonstrated by the complete genome sequencing of a fossil horse that lived 700,000 years ago [1]. And, even more interesting, storing DNA does not require much energy. On the contrary, current magnetic and optical data-storage systems cannot last for more than a century and they spend large amounts of energy. In summary, DNA-based storage may be a very powerful alternative to the current data-storage solutions, which seem to have rather serious limitations, notably in terms of storage capacity, duration and energy consumption.

This exciting story directly leads to the motivation of this document, which is basically to review and discuss:

1. DNA-based media storage basics, architectures and technology state-of-the-art
2. DNA-based media storage challenges
3. DNA-based media storage use cases and requirements
4. Main players in DNA-based media DNA media storage
5. JPEG role and next steps in JPEG DNA-based media storage activity

2. JPEG Standards for Storage and Archival

JPEG standards have been used in storage and archival of digital pictures as well as moving images. The most popular format for storage and archival of digital pictures is the popular legacy JPEG format as described in ISO/IEC 10918 and, in particular, in parts 1, 3 and 5 of the latter standard.

While the legacy JPEG format is widely used for photo storage in SD cards, as well as archival of pictures by consumers, JPEG 2000 as described in ISO/IEC 15444 is used in many archival applications, notably for preservation of cultural heritage in the form of visual data as pictures and video in digital format. Notable examples are the Library of Congress, Library and Archives Canada, Chronicling America website and the
Google Library Project. Because of its use in digital cinema, JPEG 2000 is also used for archival of movies in digital form.

In terms of technology, both legacy JPEG and JPEG 2000 formats are based on a transform-quantization-entropy coding pipeline with JPEG using the Discrete Cosine Transform (DCT) and JPEG 2000 a Discrete Wavelet Transform (DWT), followed by quantization, coefficient reordering and entropy coding. The legacy JPEG format has been extended to define JPEG XT, as described in ISO/IEC 18477, to include features attractive for archival applications such as lossless coding, while being backward compatible with the popular legacy JPEG format.

The latest JPEG image coding format called JPEG XL, as described in ISO/IEC 18181, also offers a number of attractive features important to archival applications, such as lossless compression and lossless transcoding from legacy JPEG to JPEG XL, resulting in smaller file sizes without numerical loss in the pixel values.

3. DNA-based Media Storage Technologies

Deoxyribonucleic acid (DNA) is a molecule composed by two polynucleotide chains that coil around each other to form a double helix, carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms and many viruses. DNA and ribonucleic acid (RNA) are nucleic acids. Alongside proteins, lipids and complex carbohydrates (polysaccharides), nucleic acids are one of the four major types of macromolecules that are essential for all known forms of life [2].

A so-called base or nucleotide is a unit of the DNA molecule. There are four different DNA bases: Adenine (A) and Guanine (G) are the larger purines; Cytosine (C) and Thymine (T) are the smaller pyrimidines, see Figure 1. The sequence of bases (for example, CAG) is the genetic code corresponding to a specific DNA.

![Figure 1 - DNA molecule and its units/nucleotides](image_url)

The DNA fragments (i.e. sequences of the A, G, C and T units) with stored data may be synthesized/written/printed onto a DNA microplate (i.e. a flat plate with multiple "wells" used as small test
tubes) or kept in a test tube and stored somewhere cool, dark, and dry, such as a refrigerator.

Recovering/reading the DNA stored information involves rehydrating the sample, amplifying the fragments using PCR (Polymerase Chain Reaction), and then sequencing and reassembling the full nucleotide code. Provided the user knows the strategy employed to generate the DNA code, they can then decode the original message.

In practice, the mechanism that Nature has been using to store the information of life may be now used to store any other type of information. The objective is to store information in synthetic DNA molecules created in a lab, naturally not DNA from humans or other living things. Just as with other storage systems, the data can be encrypted before it is written to the storage medium [4].

Moreover, this biological mechanism has the very appealing feature of reaching spectacular data storage densities, much beyond the current electronics mechanisms. According to [1], "... all the world’s current storage needs for a year could be well met by a cube of DNA measuring about one meter on a side."

However, there are several challenges to be overcome to successfully store media information using DNA [5]. The DNA coded information stream must respect some biological constraints on the combinations of the A, G, C, T bases that form a DNA fragment to reduce the synthesis and sequencing errors. There is also a need to overcome “biological errors”, mainly substitutions or indels (insertions or deletions of (a) nucleotide(s)), when storing and recovering information in DNA [6], that is, DNA needs to be viewed as a naturally noisy channel for which appropriately resilient codes need to be defined.

3.1 DNA-based Media Storage: End-to-end Architecture

The overall workflow of an end-to-end DNA-based media storage architecture includes the following phases:

1. **Encoding** - This phase corresponds to the conversion of the visual information into a DNA representation composed by molecules made of sequences of A, G, T and C; the visual information may be available in multiple representation forms and encoding may happen following the architectures described below.

2. **DNA synthesis** - This phase corresponds to the artificial creation of DNA molecules.

3. **Encapsulation** - This phase corresponds to the storage of the synthesized DNA molecules in a medium to preserve them.

4. **Thermal damage simulation** - This phase targets simulating the degradation that may happen in a real DNA storage system.

5. **DNA release** - This phase corresponds to the extraction of the stored DNA molecules from the storage medium.

6. **Sequencing** - This phase targets determining the nucleic acid sequence, this means the order of nucleotides in the released DNA.

7. **Decoding** - This phase corresponds to the conversion of the DNA units’ sequence back to a convenient form of visual information representation. Eventually, because of robustness issues, error resiliency tools need to be to limit the errors.

DNA media storage may happen according to three basic architectures, notably:
Transcoding-based Architecture - The data to be stored is already available as bits and the DNA coding process corresponds, in practice, to a numerical base transcoding process, notably from base 2 to base 4. This is the most common type of solution reported in the literature, see Figure 2. This type of solution has the following features:

- May have a better integration in the external data ecosystem, e.g. image ecosystem, as the data to be DNA stored (and also read from the DNA store) may be available in an off-the-shelf coding standard, e.g. JPEG, JPEG 2000;
- Less compression efficient since the important biological constraints are applied to pre-coded data in another numerical base and not directly to the symbols to code;
- Blind to the characteristics of the input data to code, e.g. its statistics, as only the binary representation is available at DNA coding time;
- Does not allow to control the quality/rate as the image has been previously coded and lossless only transcoding is happening here.

Constrained Coding-based Architecture - The data to be stored is available as symbols, e.g. quantized coefficients, and the DNA coding process corresponds to a constrained 4-ary coding process (e.g. entropy based) where the quaternary code is directly created based on statistical information or other constraints, e.g. a fixed length may be used, but also the biological constraints. In this architecture, the top/bottom left bits in Figure 2 should be substituted by some coding symbols, e.g. the quantized transform coefficients from an existing coding standard such as JPEG or JPEG 2000. This type of solution has the following features:

- May have a poorer integration in the external data ecosystem, e.g. image ecosystem, as the data pipeline does not explicitly offer a binary representation (it may be created by applying binary entropy coding on the decoded symbols); as such there is no full compatibility with existing image coding standards;
- May consider DNA storage related constraints on the 4-ary constrained-coding process creating the DNA code;
- May be more compression efficient since the critical biological constraints are directly applied to the 4-ary coding process without the binary coding intermediate stage; this may be important due to the high cost of the synthesis and sequencing processes;
- May allow quality/rate control, e.g. in the transform coefficients quantization process.

Sample-based Architecture - The data to be stored is available as component samples, e.g. YUV or RGB samples, and the DNA coding process involves all the modules of the coding pipeline. In this architecture, the top/bottom left bits in Figure 2 should be substituted by component samples. This type of architecture has the following features:

- Further reduces the interoperability/compatibility with the image coding ecosystem since all the coding modules and not only the constrained/entropy coding module as in the constrained coding-based architecture may be different from available coding solutions; an example could be a deep learning-based codec;
- This architecture only makes sense assuming that it is possible to develop more efficient image coding solutions for the modules before the entropy coding module; however, if this would be true, these advances could be also used for regular image coding standards, reverting this architecture to the constrained coding-based architecture.
3.2 DNA-based Media Storage: Technology Overview

3.2.1 Introduction

Multimedia information storage needs have been increasing rapidly over the last few years, calling for further research on novel storage approaches that should allow low-cost, long-term, high-reliability data storage. Among the competing technologies, storage in synthetic DNA strands is very well positioned due to the very high storage density (bits/gram) and long lifetime of the support medium. There are however significant hurdles that need to be overcome to make the technology usable [8], [9] in commercial applications, as described in Section 4 dedicated to the coding-related DNA-based media storage challenges. Figure 3 shows the lifecycle of DNA digital data storage according to [10].

DNA synthesis is an error-free procedure as long as the DNA strands to be synthesised do not exceed the length of 150-300 nts. For longer sequences, the synthesis error increases exponentially. Consequently, to eliminate this error to zero, 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 during decoding.

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

- **Homopolymers**: Consecutive occurrences of the same nucleotide should be avoided.
- **G, C content**: The percentage of G and C in the oligos should be lower or equal to the one of A and T.
- **Pattern repetitions**: The codewords used to encode the oligos should not be repeated, forming the same pattern throughout the oligo length.
Taking into account all the rules above, the sequencing errors can be reduced. Consequently, to be efficient, any DNA coding algorithm should respect the above rules to reduce as much as possible the probability of sequencing errors. This means that the source information has to be adapted to the DNA medium, e.g. by segmenting the encoded source information, applying error control codes and mapping the bits or symbols into the sequence of bases after some coding similar to line-codes used in telecommunications.

Note: Add text to talk about the dependency of some of the above constraints on the approach used for sequencing.

3.2.2 Digital to DNA Mapping

Encoding

It is clear that the encoding of digital data into 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 synthesised should not be larger than 150-300 nts. Consequently, the structure of a reliable encoder for DNA coding contains the following sub-parts (see Figure 4).

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 (see Figure 10).

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. The addition of redundancy is then necessary for the detection of errors and can be easily achieved using the method of Polymerase Chain Reaction (PCR) amplification, which is applied during both DNA synthesis and sequencing. Consequently, at the output of the sequencer, there will be multiple copies of each synthesised 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 can be based on computing a consensus sequence using all of the erroneous copies of each oligo or on finding the most frequent among all copies. This process can be 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 techniques 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 into DNA. An overview of the decoding process is presented in Figure 5.
3.2.3 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.

The first application of DNA data storage

In 2012, George Church et al. encode for the first time a 659-Kbyte book that was co-authored by Church into 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 [8]. The encoded sequence was then written onto a microchip as a series of DNA fragments using an ink-jet DNA 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.

This work make a pioneering step to prove the feasibility of using DNA as an alternative means of storage while also demonstrating the extraordinary capacity compared to conventional storage devices and revealing that sequencing may be an error prone process. By analysing the different errors which occurred during sequencing, this work provided a first study of the main constraints to be respected during the encoding.

First biologically constrained DNA encoding

In 2013, Goldman et al. [6] proposed a novel algorithm for encoding digital data into binary while respecting the main sequencing constraints. The encoding proposed using a 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 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 Figure 6.

![Figure 6 - Digital encoding in DNA [10].](image)

**Introduction of Reed-Solomon codes for DNA encoding**

To deal with the remaining sequencing errors, in 2015, Grass and his team [11] 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 mapping the data to blocks containing 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, thus guaranteeing that homopolymers are avoided. Each encoded column represents a DNA fragment to be synthesized and stored in silica to ensure long-term storage without DNA corruption. 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. The storage workflow is shown in Figure 7.
In 2016, Blawat et al. [12] proposed another interesting method for constructing a robust quaternary code by encoding each byte of some digital data to 5 nucleotides using the following algorithm. To begin with, the first three pairs of bits are encoded to the corresponding nucleotides from Table 1 in Figure 8 and represent the first, second and fourth nucleotide, respectively, in the encoded DNA word. Then the last pair of bits can be encoded to a pair of nucleotides among four different options as presented in Table 2 of Figure 8 and will be placed in the third and fifth position of the resulting DNA word.

As a result, for each byte, four different DNA words are provided. To ensure that the limitation concerning the maximum run-length is respected, the four options are filtered so to not create homopolymers.

To do so, the authors propose keeping only the options that do not violate the following rules: i) the first three nucleotides shall not be the same; and ii) the two last nucleotides shall not be the same.

With the above described constraints, at least two 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 byte positions and from cluster B to represent 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 propose robustifying the addressing headers using Reed Solomon codes to allow a more reliable decoding.

**First robust random-access implementation**

At the same year (2015) Yazdi et al. [13] 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 specially 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 losing the full oligo during sequencing.

In a later study published in 2017 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.

**DNA coding using Fountain codes**

Still in 2016, Columbia University researchers Yaniv Erlich and Dina Zielenski proposed a method based on a Fountain code [14], an error-correcting code used in video streaming. As part of their method, they use the code to generate many possible oligos on the computer, and then screen 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. This encoding architecture is presented in Figure 9.

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 symbols 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, reaching a capacity of 1.98 bits/nt.
Efficient end-to-end DNA coding workflow

In 2016, Borhholt et al. from Microsoft Research proposed 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 [15]. Later in 2017, researchers at Microsoft Research 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 [16], [17]. Finally, in 2019, a Microsoft Research 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 described in [18].

Closed loop optimization of DNA image coding

All the above studies provide some way for building a quaternary encoding of digital data by respecting the biological constraints. 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 solutions and improve the quality of the stored data.

As the main DNA data storage drawback is the high synthesis cost, the encoding methods proposed in the literature attempt to improve the storage capacity while also being robust to sequencing errors. To this end, some 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 the recent study of Dimopoulou et al [19], [20], a source allocation algorithm was developed 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, they also proposed a new efficient algorithm for the construction of a robust fixed length DNA code that facilitates the nucleotide allocation method. Then, two different mapping methods were introduced. The first deals with pattern repetitions which
might be the cause of error increase in the Illumina sequencers and has not been tackled by previous studies.; the second aims at decreasing the visual impact of substitution errors which may remain after error correction [18], [21], [22].

Finally, a new formatting structure was presented for cutting the encoded information into oligos and adding the needed headers, which suits the proposed encoding.

3.2.3 Segmentation and Reassembly

Synthesizing long chains of DNA was/is challenging and long chains are prone to single and multiple base errors and erasures. To address these problems, most works on DNA data storage rely on short DNA chains to represent the data, mandating the use of segmentation of the data prior to mapping/synthesis into DNA strands. The original order of the data can be recovered if some kind of addressing or indexing is used to signal the segments order. More complex and higher-level indexing schemes can be used as shown in Figure 10 which depicts a DNA fragment format used in an image DNA storage method [19] that includes primers end markers as well as an ID field used to identify images.

![Figure 10 - Dimopoulou et al DNA packet format [19].](image)

3.2.4 Error Control

To retrieve information stored in DNA, first PCR (Polymerase Chain Reaction) has to be employed to multiply the DNA strands to reach numbers beyond the detectability thresholds of the equipment in charge of the next step, i.e. sequencing. After PCR, multiple copies of each strand, possibly with errors, are aligned and, as illustrated in Figure 11, some sort of voting or parity scheme is used to obtain the error-corrected strand.

![Figure 11 - Strand alignment during sequencing [23].](image)

More sophisticated methods for error control can be used, such as Reed-Solomon codes, applied as suggested by [23] and shown in Figure 12.
3.2.5 Multimedia Storage in DNA

An early effort to store audio information (music) in DNA was made by Twist Bioscience, Microsoft, the University of Washington, EPFL, and the Montreux Jazz Digital Project as reported in [24]. This project recorded both Deep Purple’s “Smoke on the Water” and Miles Davis’ “Tutu” songs in DNA, “making scientific history.”

At least two types of approaches can be used for image data storage in DNA. The simplest one involves storing bitstreams representing images obtained by the use of e.g. JPEG encoders, using the DNA storage procedures summarized before. However, to ensure better adaptation to the characteristics of the storage medium, i.e. DNA, and possibly achieve higher storage efficiencies, it is better to design coding algorithms specific for DNA storage. Risking leaving out other relevant works, the methods proposed by Dimopoulou et al. in 2018, 2019 and 2020 [5], [19] [20] should be cited. The solution described in [19] is particularly interesting as it is based on a DWT image decomposition where the DWT coefficients are scalar quantized and an optimal nucleotide allocation is employed to minimize the distortion values and to constrain the length of the nucleotide strand for each sub-band given by the encoder. This allocation affects the choice of the quantization step size. The nucleotides generated are then transformed to synthetic DNA, after splitting into smaller segments, usually with less than 150 nucleotides, to control the sequencing error rate. The fragment reassembly is made possible by the addition of headers to the oligos as shown in Figure 7. The headers contain the localization of each split segment encoded information, allowing further information recovery and
decoding. Moreover, the stored data is also amplified, thus creating several copies using PCR to deal later on with sequencing errors. An early and simple example of applications of DNA storage to encode movies is briefly described in [25].

4. Coding-related DNA-based Media Storage Challenges

The main coding-related challenges in DNA-based media storage are:

1. **DNA-based writing/synthesis and reading/sequencing costs** - While the cost of DNA-based writing and reading are currently prohibitive for large amounts of data, it has been reducing and it may be affordable in the future, at least for specific applications scenarios.

2. **DNA-based writing/synthesis and reading/sequencing speed** - The DNA-based writing and reading processes are currently slow.

3. **Biological-related errors constrained coding** - The biological properties of the nucleic acid and the molecular machinery used to read and write may create errors specific of this technology; for example sequences containing lots of G nucleotides are difficult to write, for example, because they often produce secondary structures that interfere with synthesis [26]. Thus, the coding processes defining the sequence of DNA bases to be used have to consider the relevant limitations and constraints in terms of DNA bases combinations while maximizing the stored data density. This may look as some kind of constrained entropy coding for a 4-ary representation basis if statistics are considered.

4. **Random access** - The basic storage processes are not amenable to random access, which requires special attention as it is a fundamental coding related functionality; in this case, it is critical to be able to read a part of the data without having to read the full data.

5. **DNA-based Media Storage: Use Cases**

DNA-based representations of media data might provide efficient means for huge storage of data. Synthetic DNA provides a very high storage density compared with the traditional electronic and magnetic based methods. Furthermore, provides a long-term support for data, which is not comparable with the traditional storage devices. According to [19], DNA has the theoretical ability to store more than 450 Exabytes in 1 gram, which is not comparable with current HDD technology that requires 600 grams for a 10TB storage. Moreover, DNA can last for centuries, which is not comparable with the typical duration of the current storage devices. Finally, it is becoming fast, easy and cheaper to perform in-vitro replications of DNA. In fact, DNA-based storage is considered as one of the solutions to the growth of digital data that some believe to reach over 170 zetabytes in 2025 [10]. Most of this data is related to the proliferation of media information over the social networks. However, most of this media information is almost never accessed (the so-called cold data) and its storage does not require very efficient access. Currently, DNA still faces the lack of random access which limits efficient access times.

While storage is the key denominator, there are different relevant use cases depending on specific requirements in terms of storage longevity, target quality, etc.
5.1 Long Term Media Preservation Archives

Considering the complexity of the storing/synthesis and reading/sequencing processes, DNA-based storage seems to firstly target large scale, long-term preservation archives with DNA-based storage confined to one or a few central storage units where information is only intended to be accessed infrequently [26]. In this case, longevity is a key requirement and no quality degradation seems to be acceptable. Lossless coding may also be a relevant requirement. National archives clearly fit in this use case.

5.2 Social Networks Cold Media Storage

With the explosion of social networks, huge amounts of personal media data are created, which should be stored for long periods, e.g. the lifetime of users. However, most of this data, getting old with time, are infrequently accessed, thus justifying the so-called cold storage. In this case, some quality degradation may be acceptable over time. Companies like Facebook, Instagram, etc, may fit in this use case.

6. DNA-based Media Storage: Requirements

Although this is still rather preliminary, the potential list of requirements may include:

1. **Compression efficiency** - The standard **shall** offer significantly increased compression efficiency regarding simple solutions in the literature, e.g. based on binary coding.

2. **Random access** - The standard **shall** allow the access to specific parts of the information without having to decode the full coded information.

3. **Biological constraints** - The standard **shall** consider the relevant biological constraints on the coding process to avoid affecting the stability of the sequence and synthesising and sequencing errors, e.g. avoiding long homopolymers (repeats of the same nucleotides > 3) and extreme G-C content.

4. **Error resilience** - The standard **shall** offer some degree of error resilience regarding reading/sequencing errors.

5. **Scalability** - The standard **shall** allow scalable representations of the information where reading only part of the full information offers a lower quality or resolution of the full represented information.

6. **Ambiguity** - The standard **shall** allow decoding without any ambiguity, i.e. a decoded bit may not be both '0' and '1'.

7. **Artificial recognition** - The standard **shall** allow the encoding output to be unambiguously recognized as artificial DNA; this is relevant as the artificial DNA stream should not be confused with natural DNA streams.

7. Relevant DNA-based Media Storage Companies, Initiatives and Consortia

Researchers at the University of Washington and Microsoft Research have developed a fully automated end-to-end system for writing, storing and reading data encoded in DNA [4]. According to [4], “Microsoft is exploring ways to close a looming gap between the amount of data we are producing that needs to be
preserved and our capacity to store it. That includes developing algorithms and molecular computing
technologies to encode and retrieve data in fabricated DNA, which could fit all the information currently
stored in a warehouse-sized data center into a space roughly the size of a few board game dice.”

A number of companies, including Microsoft and Twist Bioscience, are working to advance DNA-storage
technology [1].

A consortium, named DNA Data Storage Alliance, is being created to define an interoperable end-to-end
architecture for data storage based on DNA and to accelerate the creation of an ecosystem. The DNA Data
Storage Alliance will be a global ecosystem of companies and academic researchers, setting industry-leading
DNA data storage software and hardware standards and specifications that enable and streamline the use of
DNA to store digital data.

Oligoarchive [27], Intelligent DNA Storage for Archival, is a European Commission funded FET project that
aims at defining an architecture with the same name for efficient DNA storage of digital information.

8. What Role for JPEG and Next Steps

Because of its past successful history of offering efficient image and image sequence formats for storage and
archival applications, the JPEG committee is well positioned to address standardization challenges related to
multimedia content efficient representations and, in particular, for image and image sequences in the context
of DNA storage.

As a minimum, JPEG committee can launch an activity to convert its existing image coding formats from
compressed binary representation to compressed DNA 4-ary representation according to appropriate
requirements. Standardized image coding solutions along with complementary appropriate tools, such as
error resiliency and associated metadata, which particularly suit the requirements of DNA digital information
storage, are good directions for JPEG to explore.

As a next step, the applications of DNA digital information storage need to be explored more in detail with
particular emphasis on image and video content as information. They should then be ordered in terms of time
to market and maturity and efforts should be focused on a specific use case that can gather a critical mass of
stakeholders while remaining open to other use cases.

Based on the latter, various workshops and discussion sessions can be organized with experts and end users
in order to better understand the market needs and how a JPEG standard can help create or accelerate an
ecosystem for media storage on DNA. Once the latter is identified, the standardization process can start with
precise milestones to be identified for each stage.

Note: please see [28], [29],[30],[31],[32] for some videos of DNA-based media storage.
References


28. “DNA data storage is the future!”, https://www.youtube.com/watch?v=aPWA-n9oo4k.
29. N. Goldman, “DNA Hard Drives”, https://www.youtube.com/watch?v=tBvd7OSDGgQ.


31. “Microsoft and UW demonstrate first fully automated DNA data storage”, https://www.youtube.com/watch?v=60Gi5lqL-dA.