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

# Coding of Still Pictures

| **JBIG** | **JPEG** |
|---|---|
| Joint Bi-level Image Experts Group | Joint Photographic Experts Group |

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

**SOURCE:** Contributors: Luis Cruz, Touradj Ebrahimi (editor), Fernando Pereira (editor), Antonio Pinheiro, Mohammad Raad

**PROJECT:** Requirements

**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

### 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 4-ary representation basis. This would mean that DNA molecules may be created with a specific DNA units' configuration which stores some media representation symbols, e.g. the symbols of a JPEG 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. In this context, DNA storage implies DNA synthesis/storage and DNA reading/access, which are rather complex processes, yet becoming increasingly affordable in recent 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 and technology state-of-the-art
2. DNA-based media storage challenges
3. Main players in DNA-based media DNA media storage
4. DNA-based media storage use cases and requirements
5. JPEG role and next steps

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. Standardized image coding approaches along with appropriate tools, such as error resiliency and associated metadata, that particularly suit the requirements of DNA digital information storage are also among 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. 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 4-ary representation basis. This would mean that DNA molecules may be created with a specific DNA units' configuration which stores some media representation symbols, e.g. the symbols of a JPEG 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 reading/access, which are rather complex processes, yet becoming increasingly affordable in recent 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 more than 500,000 years ago [DNA data storage]. 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 huge 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 and technology state-of-the-art
2. DNA-based media storage challenges
3. Main players in DNA-based media DNA media storage
4. DNA-based media storage use cases and requirements
5. JPEG role and next steps

## 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 form of visual data as pictures and video in digital format. Notable examples are Library of Congress, Library and Archives Canada, Chronicling America website and the Google Library Project. Because of its use in digital cinema, it 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 of 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 [DNA].

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 for a specific DNA.

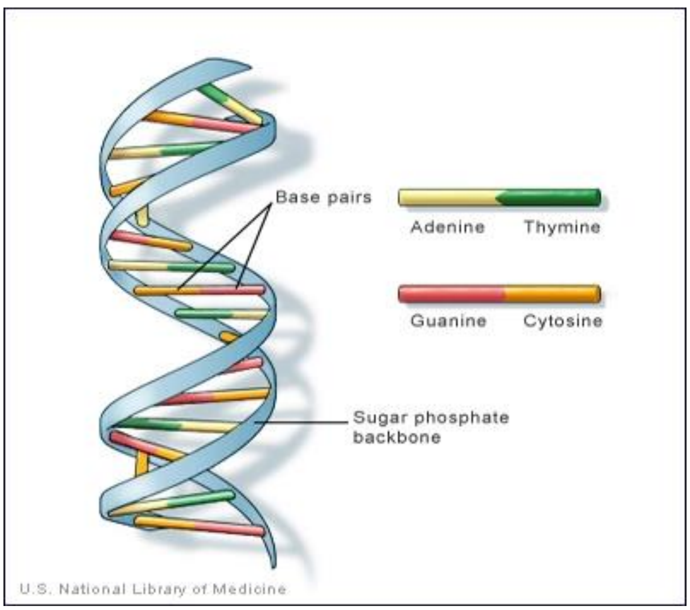


Figure 1 - DNA molecule and its units/nucleotides. [http://www.core.org.cn/NR/rdonlyres/Biology/7-A12Fall-2005/D4134A30-F348-4615-8B0A-D0CB5ED86081/0/chp_dna.jpg]

The DNA fragments (i.e. sequences of the A, G, C, T letters,) with stored data may be written/printed onto a DNA microchip or kept in a test tube and stored somewhere cool, dark, and dry, such as a refrigerator.

Recovering/reading the 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, 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, 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 [DNA Hello].

Moreover, this biological mechanism has the very appealing feature of reaching spectacular data storage densities, much beyond the current electronics mechanisms. According to [DNA data storage], "..., 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."

There are physical challenges to be overcome to successfully store digital information using DNA [DNA coding]. The coded information stream must respect the constraints on the combinations of A, G, C, T bases that form a DNA fragment. There is also a need to overcome "biological errors" when storing information in DNA [Low maintenance] - 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

1. The overall workflow of an end-to-end DNA-based media storage architecture is described in Figure 2 and includes the following phases:. **Encoding** - This phase corresponds to the conversion of a digital representation of visual information into a DNA representation composed of molecules made of sequences of A, G, T and C.
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 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 the digital representation of the information. Eventually, because of robustness issues, error resiliency tools are also used.

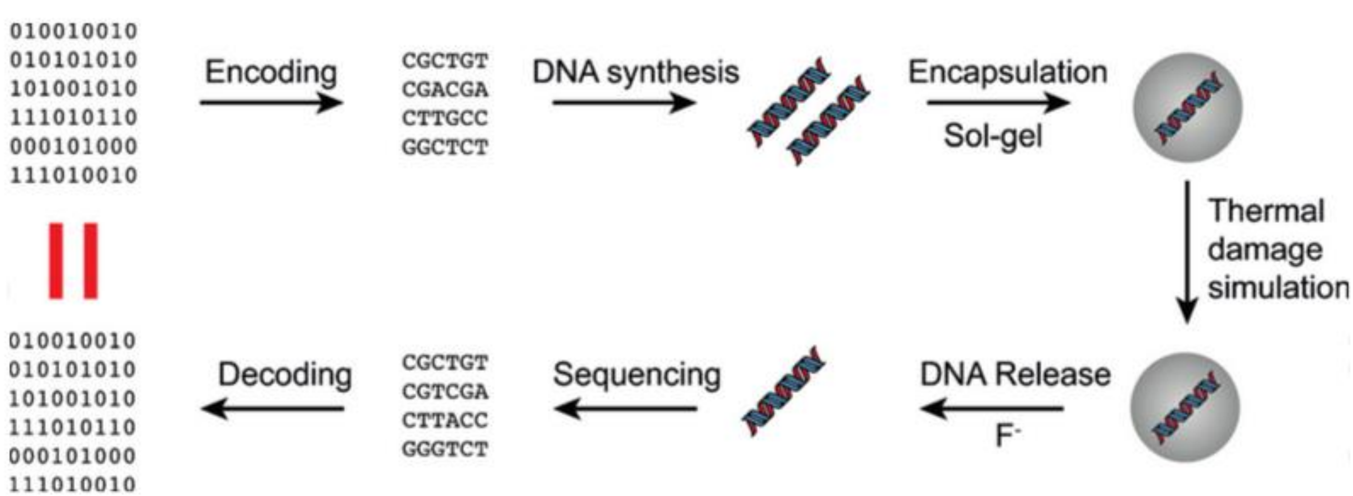

Figure 2 - End-to-end DNA-based media storage architecture [https://www.nonteek.com/en/dna-storage-solution-to-big-data-in-a-strand/]

### 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 research on novel storage approaches that will 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 [Church 2012], [Bancroft 2001] in commercial applications, as described in section "Coding-related DNA-based Media Storage Challenges". Figure 3 shows the lifecycle of DNA digital data storage according to [Campbell 2020].

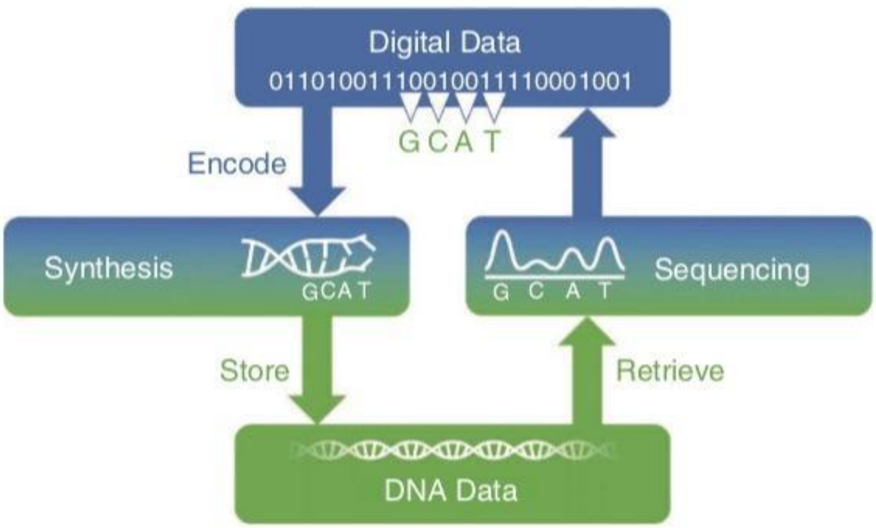


Figure 3 The DNA data storage lifecycle (from [Campbell 2020])

Due to the chemical characteristics of DNA molecules, long repetitions of sets of DNA bases (homopolymers) should be avoided in order to reduce sequencing errors at read-time. For the same reason the frequency of some combinations of bases like GC pairs should not dominate those of other bases or combinations of bases. 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 into the sequence of bases after some coding similar to line-codes used in telecommunications is used.

#### 3.2.2 Digital to DNA mapping

Since most digital data is expressed in binary format and the DNA alphabet counts four bases/symbols A,T,G and C, there has to be a mapping from the binary alphabet to alphabets related to the primitive DNA alphabet. A trivial mapping/code assigns to each of the {A,T,G,C} symbols a different combination of two bits. There are alternative solutions with [Church 2012] mapping 0 bits to A or C and 1 bits to T or G, which according to the authors increases the stability of the synthesized DNA. In [Goldman 2013] the authors propose converting all binary information to base-3 digits which are then mapped to three nucleotides as illustrated in Figure 4.

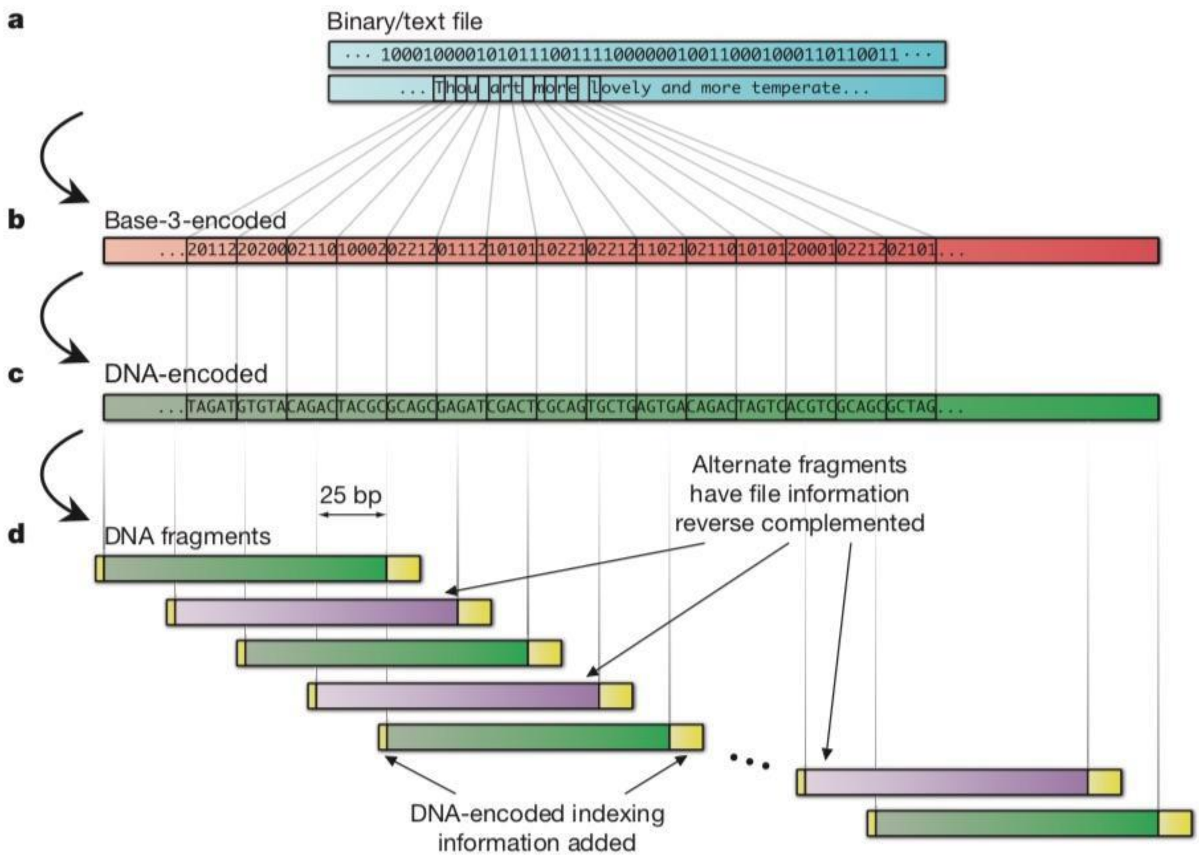


*Figure 4 Digital Encoding in DNA (from [Goldman 2013])*

More recently [Dimopoulou 2019] used two alphabets, the basic DNA set of bases {A,T,G,C} complemented by a second set of two-base combinations {AT,AC,AG,TA,TC,TG,CA,CT,GA,GT} construct to avoid long repetitions of the same base and the GC combination. A similar approach is reported in [Anavy 2019] which used a composite alphabet presenting evidence demonstrating an increase in DNA stability with resulting smaller error rate.

### 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. Both the segmentation procedure and the auxiliary segment indices are illustrated in Figure 5. More complex and higher-level indexing schemes can be used as shown in

Figure 3 depicting a DNA fragment format used in an image DNA storage method [Dimopoulou 2019] that includes primers end markers as well as an ID field used to identify images.


*Figure 5 Dimopoulou DNA packet format (from Dimopoulou 2019)*

### 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, sequencing. After PCR multiple copies of each strand, possibly with errors, are aligned and as illustrated in Figure 6 and some sort of voting or parity scheme is used to obtain the error-corrected strand.

*Figure 6 Strand alignment during sequencing (from Anavy 2020)*

More sophisticated methods for error control can be used like Reed-Solomon codes applied as suggested by [Anavy 2020] and shown graphically in Figure 7.

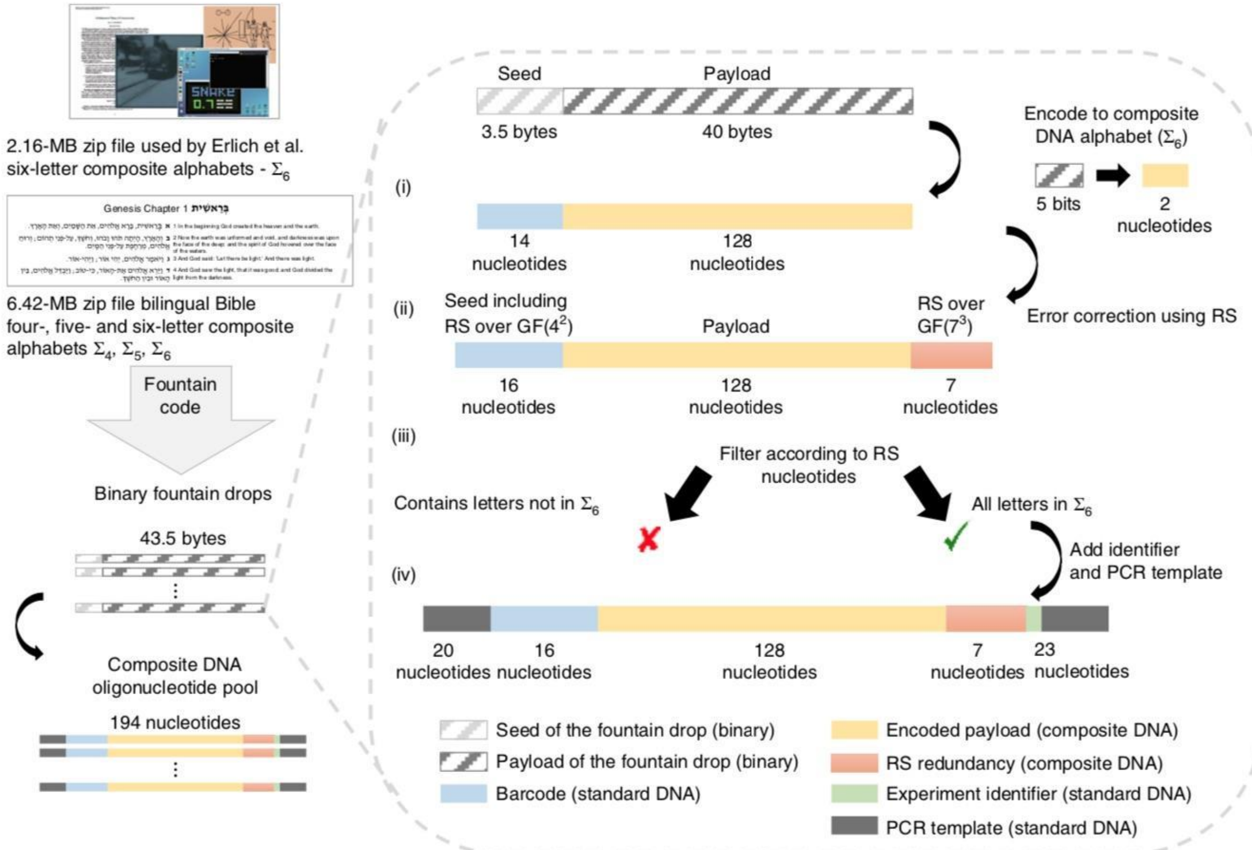


*Figure 7 Encoding pipeline using error-correcting coding according to Anavy (from Anavy 2020)*

### 3.2.5 Multimedia Storage in DNA

An early effort to store audio information (music) in DNA was brought about by Twist Bioscience, Microsoft, the University of Washington, EPFL, and the Montreux Jazz Digital Project as reported in [Twist 2017]. 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 and obtained by the application of e.g. JPEG encoders, using the DNA storage procedures summarized before. However, to ensure better adaptation to the characteristics of the storage medium, DNA, and possibly achieve higher storage efficiencies it is better to design compression algorithms specific for DNA storage. Risking leaving out other relevant works, we should cite the methods proposed by Dimopoulou et al. in 2018, 2019 and 2020 [Dimopoulou 2018] [Dimopoulou 2019] [Dimopoulou 2020]. The solution described in [Dimopoulou 2019] 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 arte 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 5. The headers contain the localization of each split segment encoded information, allowing further information recovery and decoding. Moreover, the stored data is also amplified 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 described briefly in [Goela 2016].

## 4. Coding-related DNA-based Media Storage Challenges

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

A. **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 reduced and it may be affordable in the future, at least for specific applications scenarios.

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

C. **Biochemical-related errors constrained coding** - The biochemical 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 [DNA reality]. 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.

D. **Random access** - The basic storage processes are not amenable to random access which requires special attention as 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. 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 [DNA Hello]. According to [DNA Hello], "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 [DNA data storage].

A consortium named *DNA Data Storage Alliance* is being put in place 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[1] - Intelligent DNA Storage for Archival is an EC funded FET project that aims at defining an architecture with the same name for efficient DNA storage of digital information.

## 6. DNA-based Media Storage: Use Cases

Considering the complexity of the storing and reading processes, DNA-based storage seems to firstly target large scale, long-term archival with DNA-based storage confined to one or a few central storage units where information is only intended to be accessed infrequently [DNA reality]; however, with new technologies emerging, other types of applications may become relevant.

**U1) Media storage.**

This is the obvious use case. 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, that is not comparable with the traditional storage devices. According to [Dimopoulou, 2019] 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 zettabytes in 2025 [Campbell, 2020]. 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. Nevertheless research to improve random access is still immature and it is likely to be improved.

## 7. DNA-based Media Storage: Requirements

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

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

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

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

D. **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**.**

## 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

[1] https://oligoarchive.github.io

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. Standardized image compression approaches along with appropriate tools such as error resiliency and associated metadata that particularly suit the requirements of DNA digital information storage are also among 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.

## References

*[DNA data storage]"DNA data storage is closer than you think", https://www.scientificamerican.com/article/dna-data-storage-is-closer-than-you-think/*

*[DNA] "DNA", https://en.wikipedia.org/wiki/DNA#Nucleobase_classification*

*[DNA reality] "Making DNA data storage a reality", https://www.the-scientist.com/cover-story/making-dna-data-storage-a-reality-30218*

*[DNA Hello] "With a "hello", Microsoft and UW demonstrate first fully automated DNA data storage", https://news.microsoft.com/innovation-stories/hello-data-dna-storage/*

*[DNA coding] M. Dimopoulou, M. Antonini, P. Barbry, R. Appuswamy, "DNA Coding for image storage using image compression techniques", CORESA, Poitiers, France, Nov. 2018.*

*[Low maintenance] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E.M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA", Nature, 494(7435), 77–80, 2013.*

*[DNA future] "DNA data storage is the future!",* https://www.youtube.com/watch?v=aPWA-n9oo4k

*[DNA HD] N. Goldman, "DNA Hard Drives",* https://www.youtube.com/watch?v=tBvd7OSDGgQ

*[Digital DNA] D. Zielinski, "How we can store digital data in DNA",* https://www.youtube.com/watch?v=wxStlzunxCw

*[Automated DNA] "Microsoft and UW demonstrate first fully automated DNA data storage",* https://www.youtube.com/watch?v=60Gi5lqL-dA

*[String DNA] "Storing data in DNA",* https://www.youtube.com/watch?v=vjc4LIcoux4

*[Anavy 2019] Anavy, L., Vaknin, I., Atar, O. et al. Data storage in DNA with fewer synthesis cycles using composite DNA letters. Nat Biotechnol 37, 1229–1236 (2019). https://doi.org/10.1038/s41587-019-0240-x*

*[Campbell 2020] M. Campbell, "DNA Data Storage: Automated DNA Synthesis and Sequencing Are Key to Unlocking Virtually Unlimited Data Storage" in Computer, vol. 53, no. 04, pp. 63-67, 2020.doi: 10.1109/MC.2020.2967908*

*[Church 2012] G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," Science, vol. 337, no. 6102, pp. 1628–1628, aug 2012.*

*[Goldman 2013] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low- maintenance information storage in synthesized DNA," Nature, vol. 494, no. 7435, pp. 77–80, jan 2013.*

*[Dimopoulou 2018] Melpomeni Dimopoulou, Marc Antonini, Pascal Barbry, Raja Appuswamy. DNA CODING FOR IMAGE STORAGE USING IMAGE COMPRESSION TECHNIQUES. CORESA 2018, Nov 2018, Poitiers, France.*

*[Dimopoulou 2019] M. Dimopoulou, M. Antonini, P. Barbry and R. Appuswamy, "A biologically constrained encoding solution for long-term storage of images onto synthetic DNA," in 2019 27th European Signal Processing Conference (EUSIPCO). IEEE, sep 2019.*

*[Dimopoulou 2020] M. Dimopoulou, M. Antonini, P. Barbry and R. Appuswamy, "Storing Digital Data Into DNA: A Comparative Study Of Quaternary Code Construction," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020*

*[Goela 2016] N. Goela and J. Bolot, "Encoding movies and data in DNA storage," 2016 Information Theory and Applications Workshop (ITA), La Jolla, CA, 2016, pp. 1-1, doi: 10.1109/ITA.2016.7888163.*

*[Twist 2017] "Deep Purple's "Smoke on the Water" Becomes a Piece of Scientific History https://www.twistbioscience.com/blog/company-news-updates/deep-purples-smoke-water-becomes-piece-scientific-history*