Digital Preservation with Synthetic DNA

Eugenio Marinelli
eugenio.Marinelli@eurecom.fr
EURECOM
France

Eddy Ghabach
eddy.ghabach@eurecom.fr
EURECOM
France

Thomas Bolbroe
tbo@sa.dk
Rigsarkivet
Denmark

Omer Sella
osella@imperial.ac.uk
Imperial College
London, UK

Thomas Heinis
t.heinis@imperial.ac.uk
Imperial College
London, UK

Raja Appuswamy
raja.appuswamy@eurecom.fr
EURECOM
France

ABSTRACT

The growing adoption of AI and data analytics in various sectors has resulted in digital preservation emerging as a cross-sectoral problem that affects everyone from data-driven enterprises to memory institutions alike. As all contemporary storage media suffer from fundamental density and durability limitations, researchers have started investigating new media that can offer high-density, long-term preservation of digital data. In the European Union-funded Future and Emerging Technologies project OligoArchive, we are exploring one such media, namely, synthetic Deoxyribo Nucleic Acid (DNA). In this paper, we provide an overview of the ongoing collaboration between project OligoArchive and the Danish National Archive in preserving culturally important digital data with synthetic DNA.

KEYWORDS

DNA storage, long-term archival, preservation, SIARD-DK

ACM Reference Format:
Eugenio Marinelli, Eddy Ghabach, Thomas Bolbroe, Omer Sella, Thomas Heinis, and Raja Appuswamy. 2021. Digital Preservation with Synthetic DNA. In *Proceedings of 37ème Conférence sur la Gestion de Données – Principles, Technologies et Applications (BDA '21).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Today, we live in an increasingly digital society. Digital data pervades all disciplines and has established itself as the bed rock that drives our society, from enabling data-driven decisions based on machine learning, to encoding our collective knowledge compactly in a collection of bits. Thus, preservation of digital data has emerged as an important problem that must be addressed by not just memory institutions today, but also by institutions in several other sectors.

In order to preserve digital data, it is necessary to first store the data safely over a long time frame. Historically, this task has been complicated due to several issues associated with digital storage media. All current media technologies suffer from density scaling limitations resulting in storage capacity improving at a much slower rate than the rate of data growth. For instance, Hard Disk Drive (HDD) and magnetic tape capacity is improving only 16-33% annually, which is much lower than the 60% growth rate of data [8]. All current media also suffer from media decay that can cause data loss due to silent data corruption, and have very limited lifetime compared to the requirements of digital preservation. For instance, HDD and tape have a lifetime of 5–20 years. A recent survey by the Storage and Networking Industry Association stated that several enterprises regularly archive data for much longer time frames [16]. Thus, the current solution for preserving data involves constantly migrating data every few years to deal with device failures and technology upgrades. A recent article summarized the financial impact of such media obsolescence on the movie industry [14].

In project OligoArchive[3], we are exploring a radically new storage media that has received a lot of attention recently—Deoxyribo Nucleic Acid (DNA)[5, 7, 10, 13]. DNA is a macro-molecule that is composed of smaller molecules called *nucleotides*. There are four types of nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). DNA used for data storage is typically a single-stranded sequence of these nucleotides, also referred to as an *oligonucleotide* (oligo). DNA possesses several key advantages over current storage media. First, it is an extremely dense three-dimensional storage medium with a capacity of storing 1 Exabyte/mm^3 which is eight orders of magnitude higher than magnetic tape, the densest medium available today [6]. Second, DNA is very durable and can last millennia in a cold, dry, dark environment. A recent project that attempted to resurrect the Woolly Mammoth using DNA extracted from permafrost fossils that are 5000 years old is testament to the durability of DNA even under adverse conditions [15]. Thus, data once stored in DNA can be left untouched without repeated migration to deal with technology upgrades. Third, as long as there is life on earth, we will always have the necessity and ability to sequence and read genomes, be it for assembling the genome of a previously-unknown species, or for sequencing the genome to detect diseases causing variations. As a result, unlike contemporary storage technologies, where the media that stores data and the technology to read data are tightly interlinked, DNA decouples media (biological molecules) from read technology (sequencing), thus reducing media obsolescence issues.

In this work, we provide an overview of the ongoing collaboration between the Danish National Archive and project OligoArchive in demonstrating a holistic solution for long-term preservation of

BDA '21, October 25–28, 2021, Paris, France

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

culturally significant data using DNA. We present a motivating use case for long-term digital preservation, outline the challenges involved in using DNA as a digital storage medium, and present the end-to-end pipeline we have put in place to overcome these challenges.

2 CONTEXT AND BACKGROUND

2.1 Danish National Archive Use Case

The Danish National Archives is a knowledge center documenting the historical development of the Danish society. The archive collects, preserves and provide access to original data with the purpose of supporting current and future possible needs of the Danish community - public authorities as well as private citizens. A huge part of this work includes the preservation of digitally created and retro-digitized data securely and cost-effectively. Thus, the archive has received and preserved such data since the 1970s.

The archival material used for this work consists of selected hand-drawings made by the Danish king Christian IV (1577-1648). Although his reign was marked by military defeat and economic decline, Christian IV stands out as one of the most prominent, popular and admired characters in the line of Danish kings. The hand-drawings date to the period 1583-1591 where the king was 6–14 years old. The material is a part of a larger archival unit consisting of numerous documents and records [1]. The specific image [2] used for this experiment presents a naval battle between several warships. Besides emphasizing the young king's admiration for warfare and naval tactics, the material further indicates his high level of cultural education as well as his talent for drawing. At Danish National Archive, the material is thus ranked as having "Enestående National Betydning" (meaning unique national significance).

2.2 DNA storage challenges

Using DNA as a digital storage medium requires mapping digital data from its binary form into a sequence of nucleotides using an encoding algorithm. Once encoded, the nucleotide sequence is used to *synthesize* DNA using a chemical process that assembles the DNA one nucleotide at a time. Data stored in DNA is read back by *sequencing* the DNA molecules and decoding the information back to the original digital data.

A simple way to convert bits into nucleotides is to adopt a direct mapping that converts 2 bits into a nucleotide, for instance 00 to A, 01 to T, 10 to G, and 11 to C. This way a binary sequence is translated to an arbitrary sequence of nucleotides. However, such a simple approach is not feasible due to several biological limitations imposed by DNA synthesis and sequencing steps. First, DNA synthesis limits the size of an oligo between hundred to few thousands of nucleotides. Therefore, data must be divided into several pieces, with each piece being stored in an oligo. However, unlike current storage devices, oligos do not have logical addressing. Hence, index information that helps to identify the order in which the oligos, and hence the corresponding data bits, must be reassembled back during recovery must be stored together with the data bits and integrated in each oligo.

[1] https://www.flickr.com/photos/statensarkiver
[2] https://www.flickr.com/photos/statensarkiver/28273082238

Second, oligos with repeat sequences (like ACACAC), or long consecutive repeats of the same nucleotide (like AAAAA), and oligos with extreme GC content, where the ratio of Gs and Cs in the oligo is less than 30% or more than 70%, are known to be difficult to synthesize, sequence, and process correctly. Thus, when constructing oligos, constraints need to be enforced to minimize homopolymer repeats and balance GC content. Further, care must be taken to minimize similarity across oligos as having too many oligos with similar nucleotide sequences can exacerbate sequencing errors and make it difficult to identify the original oligo.

Third, sequencing and synthesis are not error free even for well-formed oligos, as they introduce substitution errors, where a wrong nucleotide is reported, or indel errors, where spurious nucleotides are inserted or deleted. Both sequencing and synthesis also introduce bias. Some oligos are copied multiple times during synthesis, while others are not. Similarly, some oligos are read thousands of times during sequencing while others are not sequenced at all. Thus, it is important to use error correction codes in order to recover data back despite these errors.

In addition to the aforementioned media-level challenges in using DNA as a digital storage medium, there are also other problems associated with digital preservation that DNA does not solve. Any digital file stored on DNA is an encoded stream of bits whose interpretation makes sense only in the context of the application used to render, manipulate, and interact with that file format. While DNA might be able to store data for millennia, the associated applications and file formats might become obsolete. Thus, in addition to preserving data, it is also necessary to preserve the meaning of data by ensuring that data is stored in a preservation-friendly, non-proprietary format. Digital data can also be altered due to a variety of reasons and additional data-integrity techniques should be put in place to ensure that data retrieved from DNA can be trusted to be the same as the original source. The digital preservation community has long pioneered file formats, information systems, and operational methodologies for solving such format obsolescence issues[1]. Thus, a holistic DNA-based preservation solution should build on such techniques to solve both media and format obsolescence issues.

3 DESIGN

In this section, we will describe the end-to-end pipeline we have put in place to overcome the aforementioned challenges.

3.1 Overcoming format obsolescence with SIARD-DK

In Denmark, all public institutions and organizations that produce data worthy of persevering are legally bound to submit them to a public archive. As the vast majority of data in the Danish public sector are organized as databases with or without files in various formats, the focus has been on archiving these data in a standardized, system-independent and cost efficient manner. As a result, the archive has implemented a Danish version of the SIARD format (Software Independent Archiving of Relational Databases)[12] named SIARD-DK for storing of such data. SIARD is an open format, designed for archiving relational databases in a vendor-neutral form and is used in the CEF building block "eArchiving".

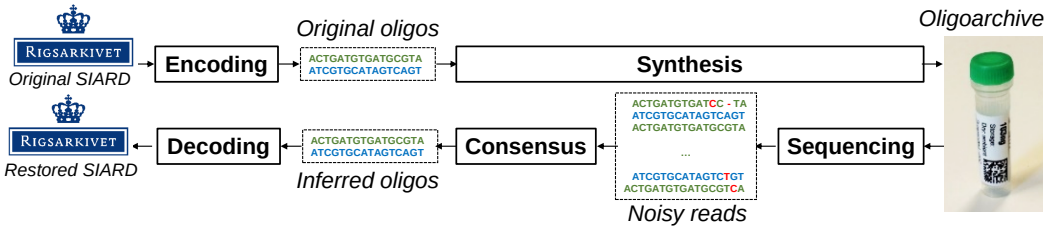

Figure 1: DNA Storage Pipeline

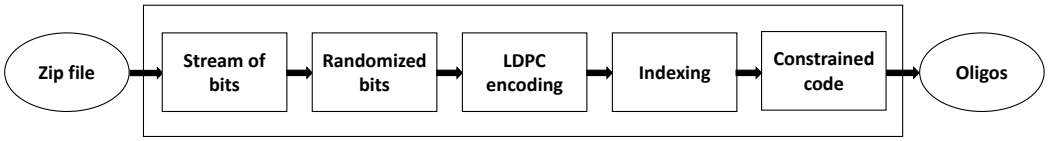

Figure 2: Encoding bits into oligos

The first step in preserving data is extracting it and creating an SIARD-DK Archival information Package (AIP). In the creation of this particular AIP, the digitized material was converted to TIFF format. Information relevant to the images such as the preservation format of the files, their title, creator and original size, descriptive information, etc., was extracted and packaged together with relevant documentation in the AIP-format. This was done using proprietary tools developed at the Danish National Archive. The usage of SIARD for storing the files guarantee that the material is preserved in a rich format with relevant metadata stored in a standardized, system and vendor independent way. The resulting AIP is a single ZIP64 file that internally contains the TIFF images, in addition to XML and XSD files that store the schema of the archive and metadata information. This allows for strict validation of the AIP. Further, an MD5 value of each file is stored inside the archive and serves as the fixity to verify data integrity on retrieval.

3.2 DNA data storage pipeline

The end-to-end DNA media storage pipeline is presented in Figure 1. In the rest of this section, we will provide an overview of both the write path that takes as input the SIARD zip file and stores it in DNA, and the read path that restores back the zip file from DNA.

3.2.1 Write Path. In order to store the archive on synthetic DNA, the zip file is first encoded from binary into a quaternary sequence of oligonucleotides, and then synthesized to generate synthetic DNA. The steps for encoding the SIARD archive file into oligos is presented in Figure 2. During encoding, the file is read as a stream of bits and pseudo randomized. Randomization is applied to reduce the number of homopolymer repeats within oligos, and similarity across oligos when generating the oligos at a later stage. After randomization, error correction encoding is applied to protect the data against errors. We use large-block length Low-Density Parity Check (LDPC) codes [9] with a block size of 256,000 bits as the error correction code, as it has been shown to be able to recover data in the presence of intra-oligo errors, or even if entire oligos are missing[4]. We configure LDPC to add 10% redundancy to convert each sequence of 256,000 bits into 281,600 bits with data and parity. Each 281,600 bit sequence is then used to generate a set of 300-bit sequences, where each 300-bits is composed of 281 data bits and a 19-bit index that is used to order the sequences. Each 300-bit sequence is then passed to a constrained code that converts it into an oligonucleotide sequence.

We provide a practical overview of the constrained code here, deferring rigorous mathematical definition to future work. The constrained code is essentially a finite state machine that views each oligo as a sequence of short symbols that are concatenated together. In our current configuration, the constrained code breaks up each 300-bit sequence into a series of ten 30-bit integers. Each 30-bit integer is fed as input to a deterministic finite state machine that takes as input the previous 16-nucleotide sequence, and a set of constraints (homopolymer repeat, GC ratio, etecetera) that each symbol must meet and produces a valid 16-nucleotide sequence corresponding to the 30-bits as output. Thus, each 300-bit sequence is encoded as concatenation of ten such symbols, each with a length of 16 nucleotides, leading to an oligo that is 160 nucleotides long. We would like to explicitly point out here that the length of an oligo is a configurable parameter. Thus, while we use 160 nucleotides in our current system due to favorable pricing provided by our synthesis provider, our encoder can generate shorter or longer oligos if necessary, and automatically adjust various aspects (like the 19-bit index and 281-bit data size) based on desired oligo length.

3.2.2 Read Path. To retrieve back the SIARD archive, the DNA is sequenced in order to retrieve back the nucleotide sequence of oligos. As mentioned before, sequencing produces noisy copies of the original oligos that can contain insertion, deletion, or substitution errors, which are referred to as *reads*. In order to infer the original oligos from the reads, we use a consensus procedure.

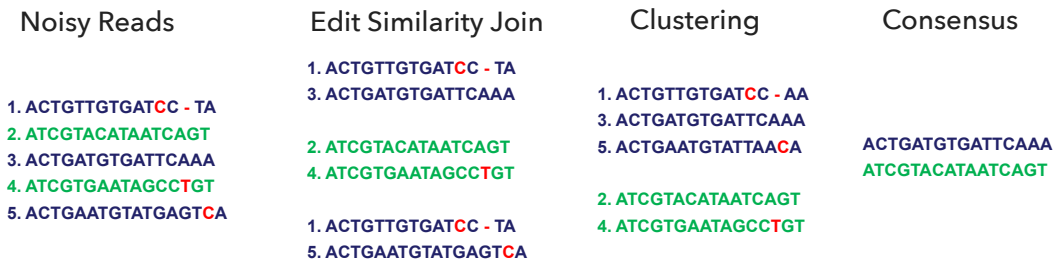

Figure 3: Various steps in the consensus procedure.

Concretely, we structure this process as sequence of three algorithms, as depicted in Figure 3. First, we identify all pairs of strings that are similar to each other. As modern sequencers produce hundreds of millions of reads, this first task is an extremely computationally intensive due to use of edit distance as a metric for comparing strings. Thus, we have developed an efficient similarity join algorithm, called OneJoin[11], that exploits the fact that due to randomization during encoding, reads corresponding to the same original oligo are "close" to each other despite some errors and "far" from the reads related to other oligos. The results obtained from the join algorithm are then used to quickly identify clusters of strings that are similar to each other. Each cluster thus groups all reads belonging the same oligo. Finally, we apply a position-wise consensus procedure that uses multiple reads to infer the original oligo in each cluster using a sequence alignment procedure [17]. We would like to point out here that not all oligos need to be correctly inferred. In fact, as we show later, some original oligos might not appear at all in the inferred set, and other inferred oligos might have errors. We rely on the parity added by LDPC codes at a higher level to recover data despite these errors.

The inferred oligos are then passed to the decoder which reverses the steps shown in Figure 2. The constrained code is first used to convert each 160nt oligo back into 300 bit sequences. The index stored in each 300 bits is used to reassemble bits back in the correct order. The LDPC decoder is then used to recover back data even if some bits were wrongly decoded, or some bits were zeroed out as corresponding oligos were missing. The decoded data is then derandomized to obtain a stream of bits that corresponds to the SIARD zip file.

4 EVALUATION

At this stage of our collaboration, we have assembled the entire pipeline. We are in the process of carrying out a real-life, large-scale synthesis and simulation experiments. As we do not have results from real experiments yet, we will provide a preliminary, simulation-driven evaluation in this section. Note that we simulate only the synthesis and sequencing steps in Figure 1. We encode/decode the real dataset using our pipeline.

The raw SIARD archive that is fed as input to our pipeline is 12.9MB in size. With redundancy added by LDPC, the resulting binary data to be stored on DNA is 14.19MB in size. We encode the SIARD archive generating 404,863 oligos, each with a length of 160 nucleotides. This corresponds to a density of 1.837 bits-per-nucleotide. Using these original oligos, we then generate five million reads by using a short-read simulator[3] tool, that adds random errors such as insertion, substitution, and deletion in each read to mimic the actions of an Illumina DNA sequencer. This corresponds to an average coverage of 11×, meaning that each oligo, on average is covered by 11 noisy copies. Note that, while the average coverage is 11×, the overall coverage typically follows a negative binomial distribution, with some oligos being covered hundreds of times, and some being not covered at all. Using the consensus procedure described earlier, we obtain the inferred oligos using this simulated dataset. We then use the constrained code to convert these inferred oligos into 300-bit sequences, and reassemble them in order based on the 19-bit index. At this stage, we can have a situation where an oligo is missing due to sequencing simulation bias, or an oligo could not be converted back into 300-bits by the deterministic finite state machine due to errors. In both cases, there will be a corresponding index whose data bits cannot be recovered. We insert a sequence of zero bits for such indices and rely on the redundancy added by LDPC to recover back the missing data.

Table 1 shows statistics for the rebuilding process of SIARD archive. The figures are obtained by comparing the inferred oligo for a certain index with the corresponding true oligo. We see that we are able to infer 404,727 oligos that correspond to 99.97% of the original oligos perfectly without errors. In addition 128 oligos were inferred with some errors, and 9 oligos were completely missing. Note that errors in an oligo does not imply that the entire oligo is different from the original one as they typically differ only in few nucleotides. For this reason, we also report the difference between inferred data and original in terms of number of bits. Despite these errors, our decoder was able to recover back the original archive completely due to error correction provided by the LDPC code.

#Original Oligos	404863
#Correctly Inferred Oligos	404726
#Incorrectly Inferred Oligos	128
#Missing Oligos	9
#Incorrect bits	8640

Table 1: Statistics for decoding of SIARD archive

[3] https://sourceforge.net/projects/bbmap/

5 CONCLUSION AND FUTURE WORK

In this work, we provided an overview of the ongoing collaboration between project OligoArchive and the Danish National Archive in using DNA to preserve culturally significant digital data. Building on prior work on molecular information storage and digital preservation, we presented a holistic, end-to-end pipeline for preserving both data and the meaning of data on DNA, and tested the pipeline using simulation studies. There are several avenues of future work we are pursuing. First, as described earlier, we are in the process of carrying out a large-scale experiment to validate our pipeline using real data. Second, we are investigating various optimizations to both encoding and consensus algorithms to support alternate synthesis and sequencing technologies with potentially higher error rates. Finally, while we addressed the question of preserving data in this work, we left open the question of preserving the decoding algorithm itself. Recent work has investigated the design of nested universal emulators that can be used to preserve and emulate such decoders using analog media like film or archival paper[2]. Thus, we are investigating methods to combine DNA-based digital data storage with analog media-based decoding logic storage.

ACKNOWLEDGMENTS

This work was partially funded by the European Union's Horizon 2020 research and innovation programme, project OligoArchive (grant agreement No 863320).

REFERENCES

[1] 2015. *Digital Preservation Handbook*. Digital Preservation Coalition.
[2] Raja Appuswamy and Vincent Joguin. 2021. Universal Layout Emulation for Long-Term Database Archival. In *CIDR*.
[3] R. Appuswamy, Kevin Lebrigand, Pascal Barbry, Marc Antonini, Oliver Madderson, Paul Freemont, James MacDonald, and Thomas Heinis. 2019. OligoArchive: Using DNA in the DBMS storage hierarchy. In *CIDR*.
[4] Shubham Chandak, Kedar Tatwawadi, Billy Lau, Jay Mardia, Matthew Kubit, Joachim Neu, Peter Griffin, Mary Wootters, Tsachy Weissman, and Hanlee Ji. 2019. Improved read/write cost tradeoff in DNA-based data storage using LDPC codes. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing*.
[5] George M. Church, Yuan Gao, and Sriram Kosuri. 2012. Next-Generation Digital Information Storage in DNA. *Science* 337, 6102 (2012).
[6] Semiconductor Research Corporation. 2018. 2018 Semiconductor Synthetic Biology Roadmap. https://www.src.org/program/grc/semisynbio/ssb-roadmap-2018-1st-edition_e1004.pdf.
[7] Yaniv Erlich and Dina Zielinski. 2017. DNA Fountain enables a robust and efficient storage architecture. *Science* 355, 6328 (2017).
[8] Robert E. Fontana and Gary M. Decad. 2018. Moorеâs law realities for recording systems and memory storage components: HDD, tape, NAND, and optical. *AIP Advances* 8, 5 (2018).
[9] Robert Gallager. 1962. Low-density parity-check codes. *IRE Transactions on information theory* 8, 1 (1962), 21–28.
[10] Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M. LeProust, Botond Sipos, and Ewan Birney. 2013. Toward Practical High-capacity Low-maintenance Storage of Digital Information in Synthesised DNA. *Nature* 494 (2013).
[11] Eugenio Marinelli and Raja Appuswamy. 2021. OneJoin: Cross-architecture, scalable edit similarity join for DNA data storage using oneAPI. In *ADMS*.
[12] Library of Congress. 2015. SIARD (Software Independent Archiving of Relational Databases) Version 1.0. https://www.loc.gov/preservation/digital/formats/fdd/fdd000426.shtml. [Online; accessed 28-May-2021].
[13] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, Christopher N Takahashi, Sharon Newman, Hsing-Yeh Parker, Cyrus Rashtchian, Kendall Stewart, Gagan Gupta, Robert Carlson, John Mulligan, Douglas Carmean, Georg Seelig, Luis Ceze, and Karin Strauss. 2014. Random access in large-scale DNA data storage. *Nature Methods* 11, 5 (2014).
[14] Marty Perlmutter. 2017. The Lost Picture Show. https://tinyurl.com/y9woh4e3.
[15] Beth Shapiro. 2015. Mammoth 2.0: will genome engineering resurrect extinct species? *Genome Biology* (2015).
[16] SNIA. 2017. 100 Year Archive Requirements Survey 10 Years Later. https://tinyurl.com/yytsbvmb.
[17] Yiqing Yan, Nimisha Chaturvedi, and Raja Appuswamy. 2021. Accel-align: A fast sequence mapper and aligner based on the seed-embed-extend method. *BMC Bioinformatics* (March 2021).