Digital Innovators
Séminaires d'innovation numérique

L'ADN comme technologie d'archivage des données numériques : état de l'art
Pierre-Yves Burgi

7 décembre 2022
12h30 – 13h30

Webinaire Zoom gratuit
http://pin.unige.ch
Agenda:

- Introduction to long-term data preservation
- Current storage technologies
- Why DNA?
- Archiving data in DNA molecules
- Viable solutions for legacy archiving?
Introduction to long-term data preservation
Long-term preservation of electronic documents consists in preserving the document and the information it contains:

• In its physical and intellectual aspects
• On the very long term (hundreds of years)
• In a way so that it is permanently accessible and understandable

To reach this goal we apply the **Reference Model for an Open Archival Information System (OAIS - ISO 14721)**

yareta.unige.ch and olos.swiss are examples of OAIS-compliant archiving systems
Open Archival Information System (ISO 14721)
Backup vs. Archiving

**ARCHIVE**

Preserve information as required by regulations and institutional policies

- Auditable
- Follows a life cycle
- Self-described
- Data integrity

**BACKUP**

Insurance policy against unforeseen system failures

- Incremental
- Multiple snapshots
- Retained on short periods of time
- Not searchable

**FAIR**

- DOI
- OAIS (ISO)
- OAI-PMH
- Data integrity
- Auditable
- Follows a life cycle
- Self-described

**OPEN SCIENCE**

- FAIR
- OPEN SCIENCE

**Redundancy**

- DNA
- S3
- NAS
- LTX
Current storage technologies
Hard Disk and Flash (and other) Devices

HDD/flash areal density perspective

- Tapes
- Optical disks

4 Gb/mm²

Production year

(Adapted from Li, 2016)
Lifetime of current devices

(Meiser et al., 2022)
Technology reaches its limits

TSMC Roadmap

Under 10-15 nm the quantum tunnelling effect affects storage reliability

Samsung dreams of 2 nm chips in 2025, 1.4 nm by 2027
Environmental Impacts

Taiwan Semiconductor Manufacturing Company uses almost 5% of all Taiwan’s electricity, predicted to rise to 10% by 2030, and it used about 63 million tons of water in 2019.

Advanced technologies, such as the Extreme Ultraviolet Lithography (EUV) tools that TSMC bought in 2019, consume 20 times as much power as previous generations of production tech.

Media and hardware used to store and manage the data will be changed every 5-10 years, with the old media/hardware either recycled, incinerated, or dumped in a landfill.
Environmental Impacts

(Tannu & Nair., 2022)
Why DNA?
Density

- 14 atoms / bit
- No tunneling effect because made of molecules
- Created on demand
- About 200 Tb/mm² (vs. 4 Gb/mm² HDD)
- All human-made digital data stored in less than 100 g of DNA
- Human genome: 3.59 pg
Lifetime

(L. Meiser et al., 2022)
No obsolescence (fabric of life)

*Thermus aquaticus*, up to 85 °C and in very acidic environments

Space traveler Deinococcus radiodurans recovered after 1 year of exposure to low Earth orbit (LEO) outside the International Space Station

Resistance to ionization and extreme environments
Low-cost copies based on PCR (Polymerase Chain Reaction*)

- Fixed time process (few minutes) : does not depend on the volume of data
- Average power of approximately 1.0 W per PCR cycle!

* Use the DNA polymerase enzyme e.g., isolated from the *Thermus aquaticus*


Mullis, 1990
Archiving data in DNA molecules
(Doricchi et al. 2022)
I. Sequenced-based archiving
Une seule synthèse

00 00 10 01 11

Synthèses en parallèle

01 10 00 01 11 01
00 10 11 01 11 10
10 01 00 10 10 ...

Adapté par François Képès depuis la présentation de Nick Gold (Catalog DNA)

About 1 error(s) per 500-2000 nucleotides (nts) (0.2 - 0.05 %)
1 Writing

(Bornholt et al 2016)
The logical structure

Primer | Payload (including index and code redundancy)
---|---
TACAATGAATTTGATCGCCC CGAAAGCAAGAAGATCGGACAAGCCCGGCGGTCCCGGCCAGCTCTGAGTCGCGGGTCGTCTCGGCCTAA
TATGTTAGGTCCCTGATTACCTGATAGCGCTTAGCTTTCCCGAGATTCGACAAAGGTCTTTGTACATGGGCATCAAAATTCTATTGTA

A ↔ T
C ↔ G

TACAATGAATTTGATCGCCC
ATGTTACTTAAACTAGCGGG
GGGCGATCAAAATTCTATTGTA
Preserving DNA in silica at 50% rel. humidity (this work)
DNA in solution (Lindahl & Nyberg 1972)
DNA over CaCl₂ (Bonnet et al. 2010)
DNA in moa bone (Allentoft et al. 2012)

Experimental window $t_{1/2}$ ~weeks

Grass et al. 2015
Organick et al. 2020

2⃣
Reading process

DNA Storage System

1. Key:
   - foo.txt

2. Map to Primer Sequence:
   - ATCCGATACT

3. Synthesize Primers

4. PCR Amplification

5. Sequencing

6. Decode

(value: 1101000100...)

(Bornholt et al. 2016)
3 Reading through nanopore technologies

Error rate: 5-15%
Transition A-G more error prone than A-T
II. Structure-based archiving
DNA Origami

(Wang & Ding, 2022)
DNA Origami Method

- Short DNA staple strand (< 20 nt) can be configured (e.g. orientation) and regarded as pixels (voxels in 3D).
- “Pixel” size ≈ 5 nm, about 330 Gbit/cm²
- Reading through AFM or optical methods

(Rafta et al, 2020)
Nanopore-Based DNA Hard Drives (structured-based)

- A DNA strand (20 nt) is defined as "0" without a binding protein and "1" otherwise.
- Nanopores are sensitive to the DNA configuration shape.
- The protein can be removed (erase operation) and added (write operation).
- DNA molecules can move through microfluidics channels.
- About 100 nt per bit.
- Possibility to chemically encrypt the information.
Viable solutions for legacy archiving?
Costs

Today about 1’000 CHF/MB

But 2 main reasons not to worry...
DNA Technology is advancing 7.5 times faster than Moore's Law.

Forecast of DNA Synthesis Cost

- Photolithographic DNA could lead to $10^{-8} \$/ MB (0.01 \$/ TB)

The future of DNA storage (sept. 2018)
Technology obsolescence requires recurrent investments

Adapted from Goldman et al. 2013
Estimated Cost of Writing and Storing 1PB - Cloud Archival (Glacier Deep Archive), Tape (On-Prem) and DNA

From: The future of DNA storage, DNA Storage Alliance (sept. 2018)
Error correction is still an ongoing research field

Generating a library of orthogonal primers for very large datasets is not obvious

The density of information in DNA is enormous, but “only” a small part is currently feasible, so far about 200 PB/g
Amsterdam’s new circular archives building sustainably generates all of its own energy
DNA Range of temperature for long-term preservation

Annual average ambient temperature < 15°C
Conclusions

The environmental impacts of DNA storage are not commensurate with those of silicon/electronic technologies

Six to eight orders of magnitude more dense than traditional storage media

Is already being made available to institutions aiming at very long-term preservation

Its technology is evolving faster than Moore’s Law, so will soon be available to legacy archiving institutions

Get ready for the next storage revolution!
Thanks for your attention!

Interested in a collaboration?

Contact us:

pierre-yves.burgi@unige.ch
lamia.friha@unige.ch

To know more on our DNA project:
Burgi et al. (2022) OAIS-Compliant digital archiving of research and patrimonial data in DNA. Proc. iPres22, pp. 220-224