Archives doublement durables

Stockage et archivage sur l'ADN : option durable

1. Juin 2022

Dina Andriamahady, Pierre-Yves Burgi, Jan Krause

UNIVERSITÉ DE GENÈVE

h e g
Haute école de gestion Genève

canton de Vaud
Archives Cantonales Vaudoises

Plan - Stockage et archivage ADN

- Introduction à l’archivage ADN – Dina Andriamahady
- Durabilité – Pierre-Yves Burgi
- En pratique – Jan Krause

L’ADN, un excellent candidat pour le stockage à long terme

Les caractéristiques qui font de l’ADN un atout pour le stockage numérique:

- Une grande densité de stockage informationnelle: 1 gramme d’ADN peut contenir 215 millions de Gigabytes;
- Une longévité remarquable: dans les meilleures conditions, une molécule d’ADN peut être conservée des milliers d’années;
- L’ADN est le support de l’information génétique, à l’origine de la perpétuation du vivant, ce qui l’exempt de tout risque d’obsolescence.

“… and ~ 1 kg of DNA would be sufficient to address the world’ storage requirement in 2040 (3×10^{24} bits).” (Panda et al. 2018)

Modèle de référence: Open Archival Information System (OAIS)

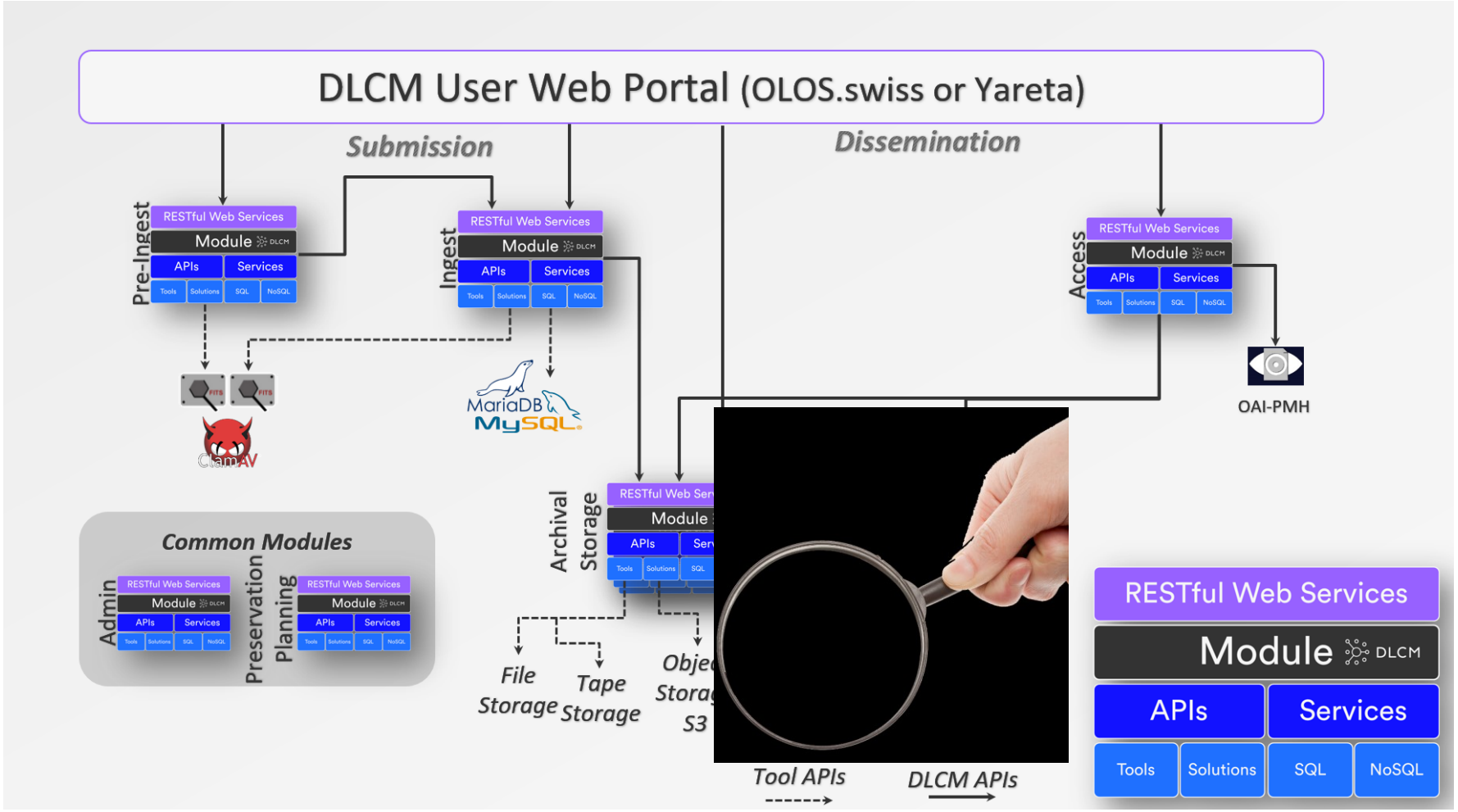
AIP Archival Information Package

ICS › 49 › 49.140

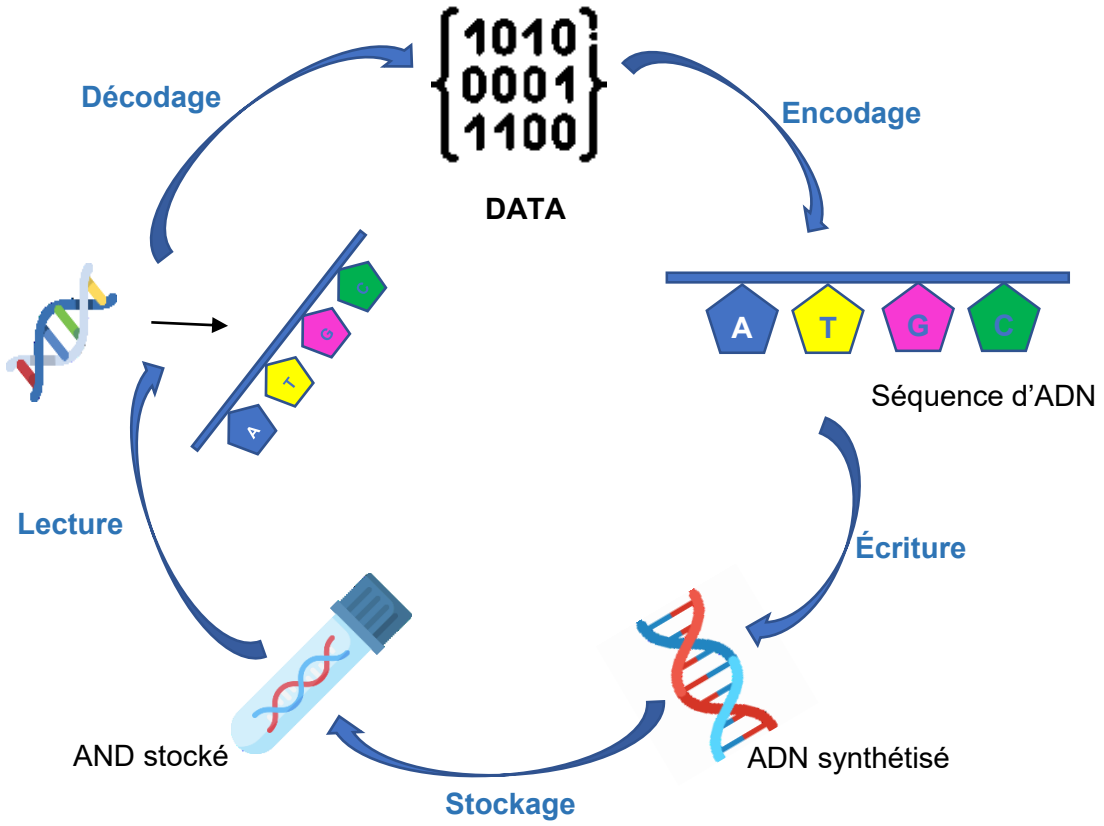
ISO 14721:2012

Space data and information transfer systems — Open archival information system (OAIS) — Reference model

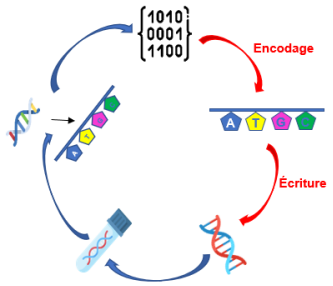
OLOS: la solution suisse pour la gestion des données de la recherche



ADN, support de stockage d’archivage numérique

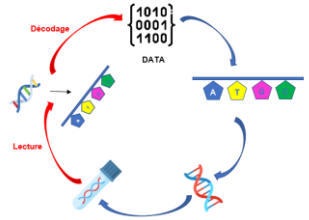


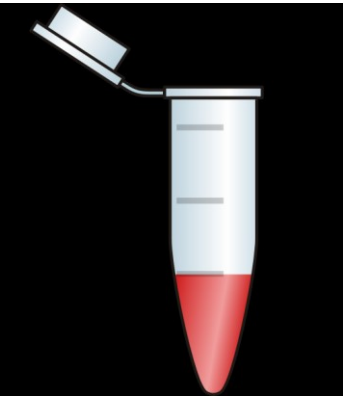
Phases d’encodage et d’écriture



00 ⟶ A
01 ⟶ T
10 ⟶ G
11 ⟶ C

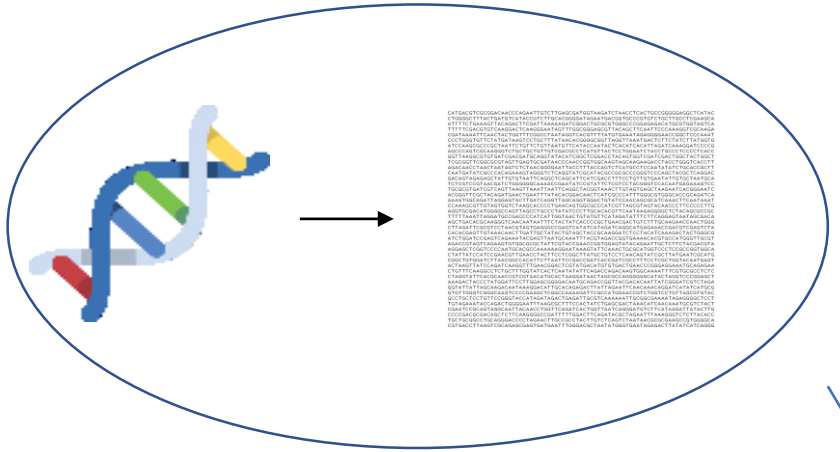
Phases de lecture et de décodage





Molécule d'ADN synthétisée puis stockée

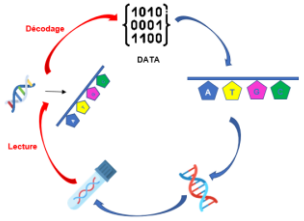
Amplification par PCR



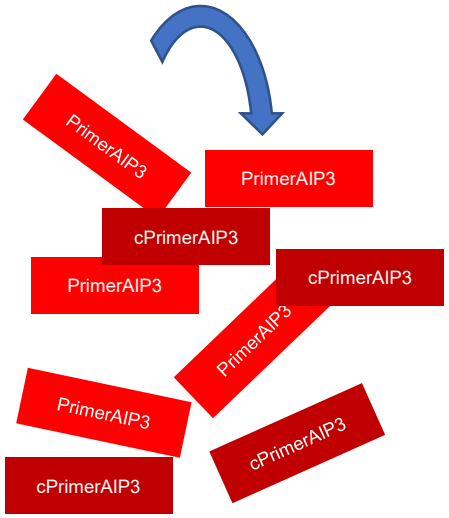
Séquençage de la molécule d'ADN pour en obtenir la séquence

Décodage

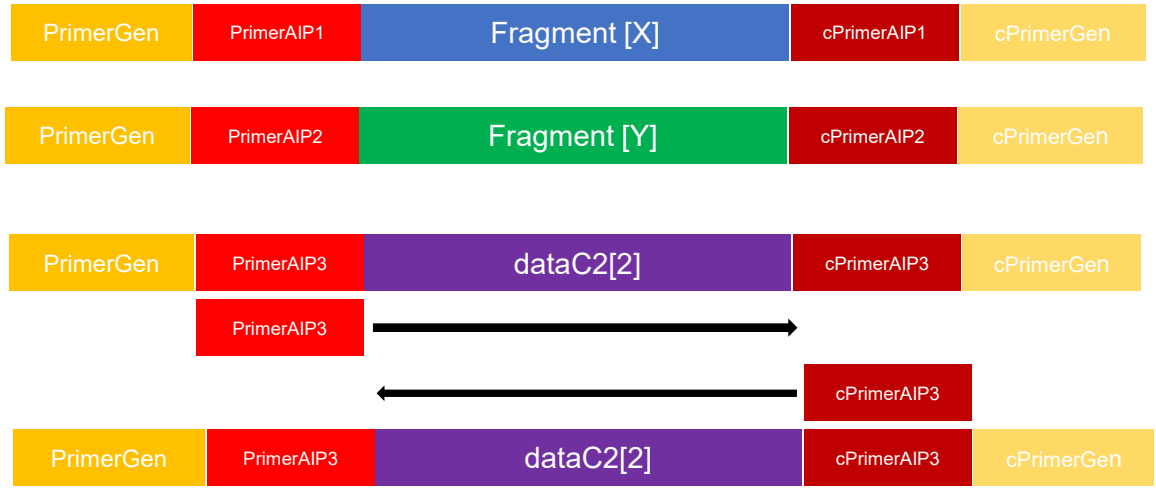
Accès ciblé à une information: la force des primers



Addition de primers spécifiques

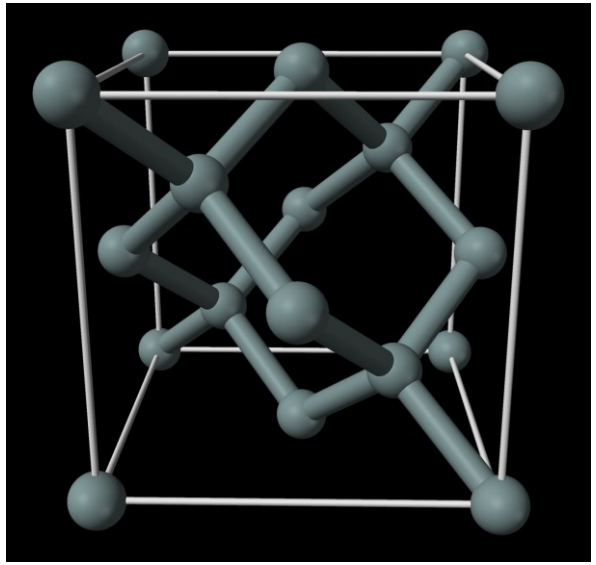


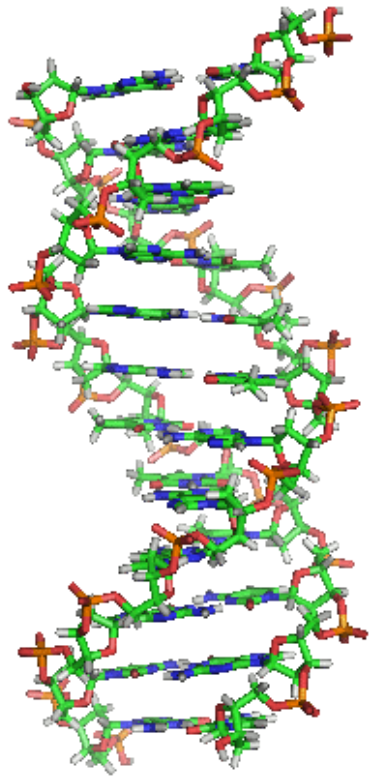
PCR



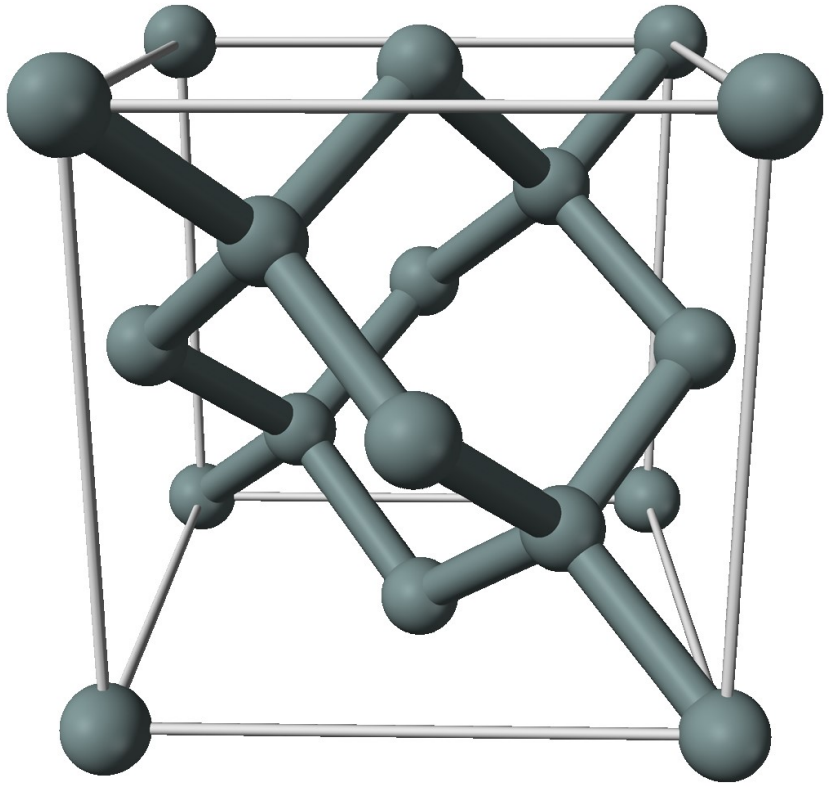
Résultat après un cycle

Silicon versus DNA





1. Silicon



Data growth and wafer demand

Global data storage demands will rise to 175 Zetta Bytes by 2025 with 2.5 Exa Bytes produced everyday

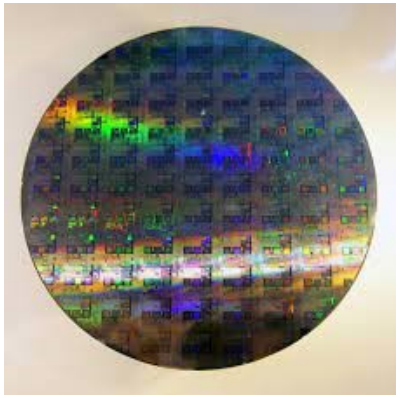
175 ZB will require about 6×10^7 kg of silicon wafers

Giga = 10^9 Tera = 10^{12} Peta = 10^{15} Exa = 10^{18} Zetta = 10^{21} Yotta = 10^{24}

By 2040, global memory usage will surpass the supply of silicon available to store it :

3 Yotta Bytes by 2040, requiring about 10^9 kg flash memory

Estimate supply in Silicon wafers: 10^7 - 10^8 kg



Giga = 10^9 Tera = 10^{12} Peta = 10^{15} Exa = 10^{18} Zetta = 10^{21} Yotta = 10^{24}

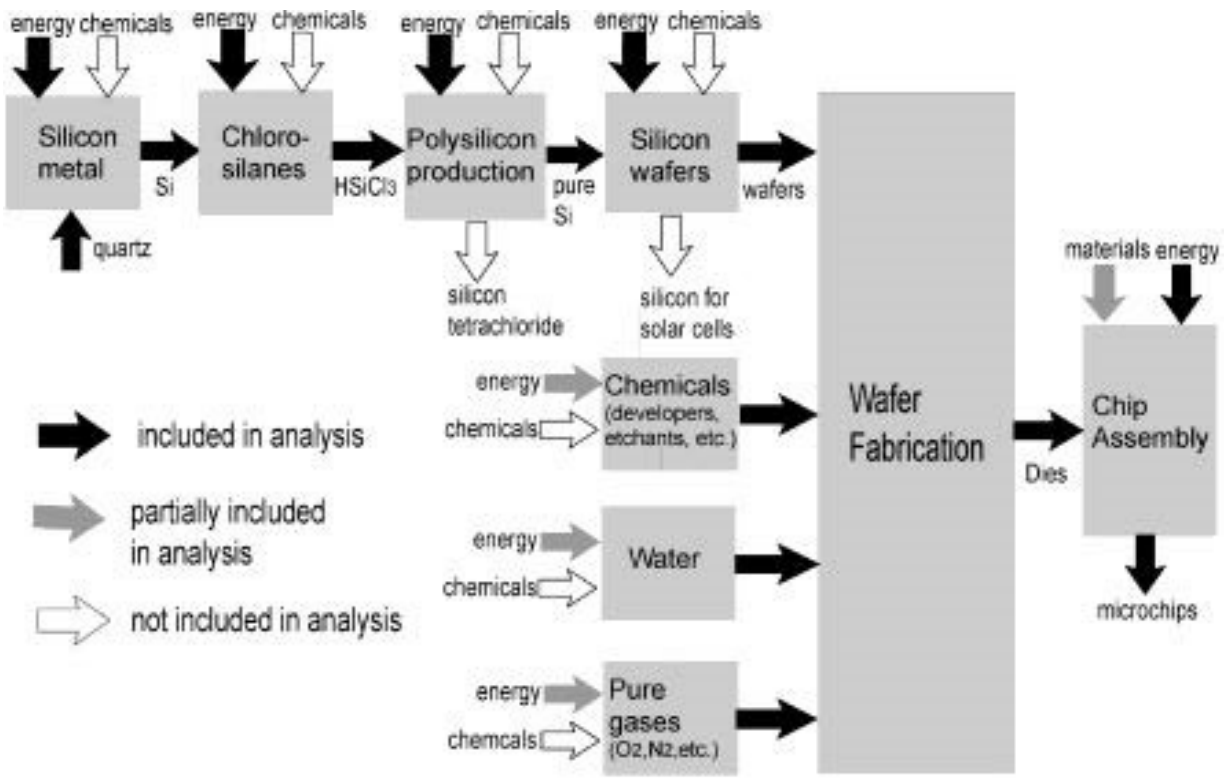
Environmental Impacts

A single 2-g chip requires:

- 1.6 kg secondary fossil fuel
- 72 g chemical inputs

For each cm^2 silicon wafer:

- 1.5 KWh
- 18-27 L of water



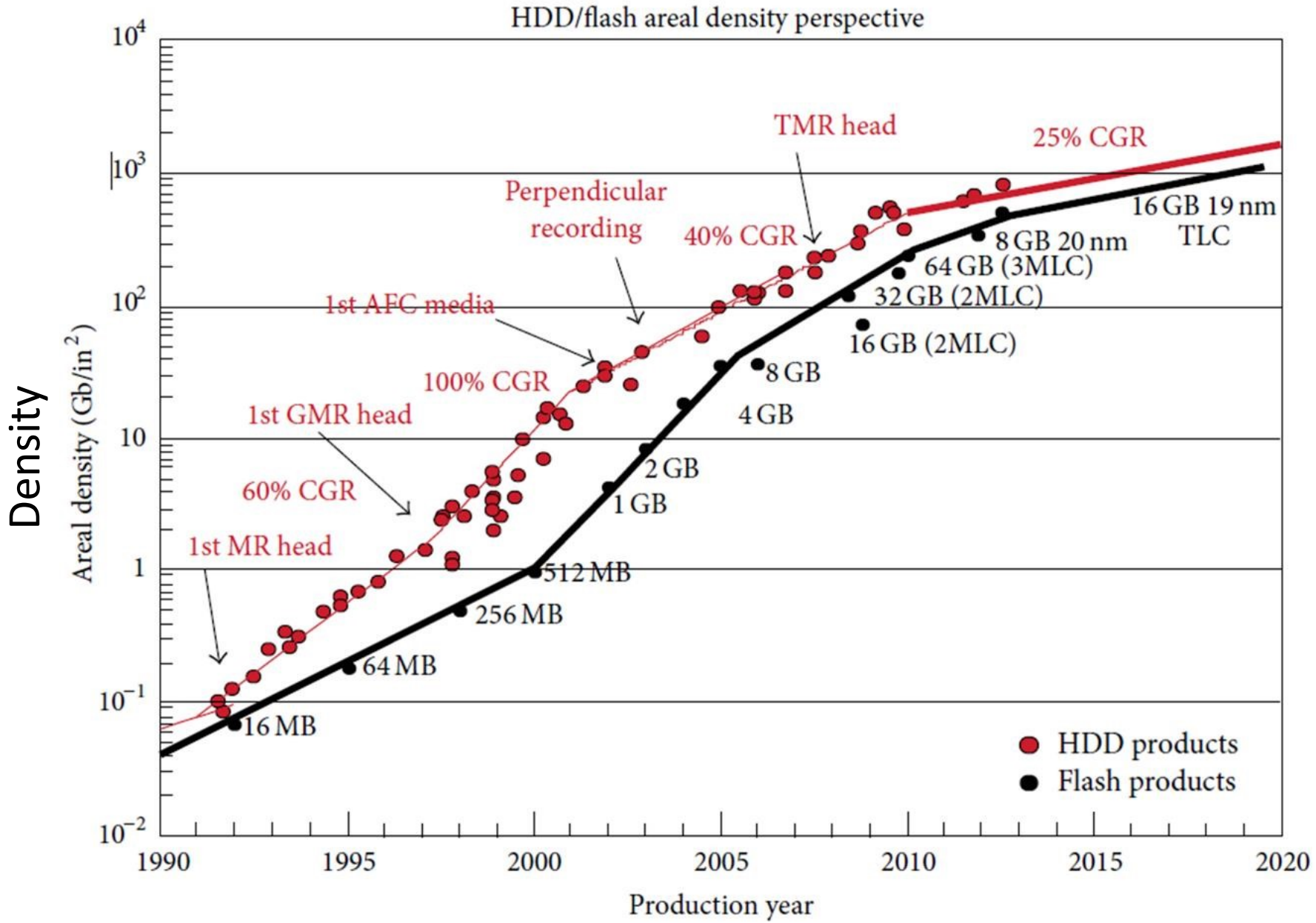
From Williams et al. 2002

Environmental Impacts

Taiwan Semiconductor Manufacturing Company uses almost 5% of all Taiwan's electricity, predicted to rise to 7.2% in 2022, and it used about 63 million tons of water in 2019.

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.

Silicon density reaches its limits

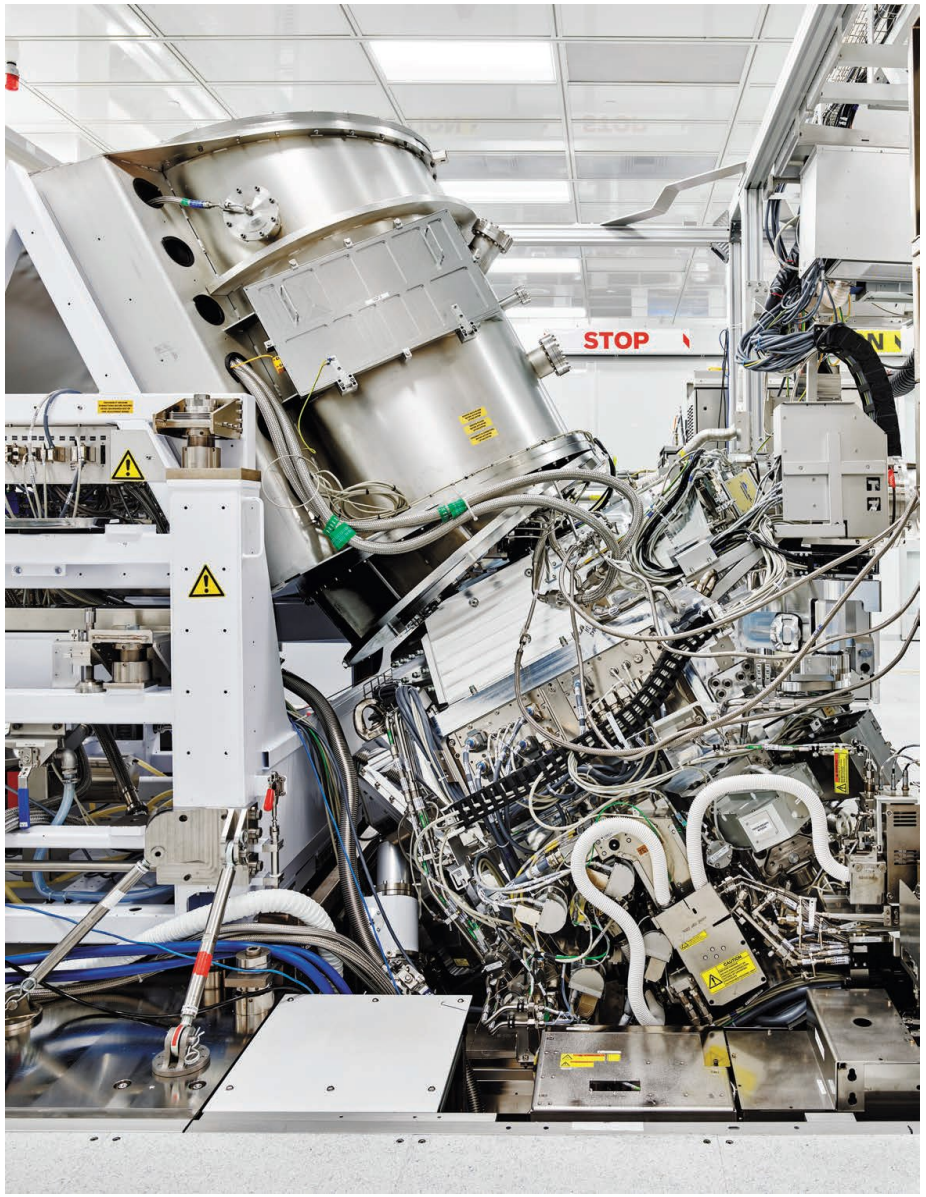


(H. Li, 2016)

Silicon technology reaches its limits

Quantum tunnelling effect:

- Electron-based memory: limited to 10-15 nm technology
- ASML $9 billion of R&D and 17 years of research to get to 13 nm and could reach less than 10 nm by 2025



Storage Obsolescence

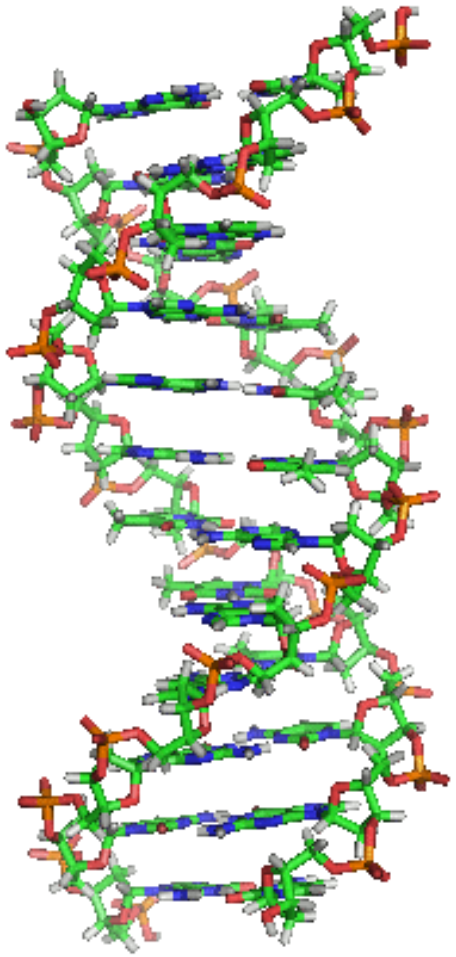
5-50 years

Depends on number of writes, but usually 3-5 years

3-5 years

2 generations (8-10 years)

2. DNA



Data growth and DNA

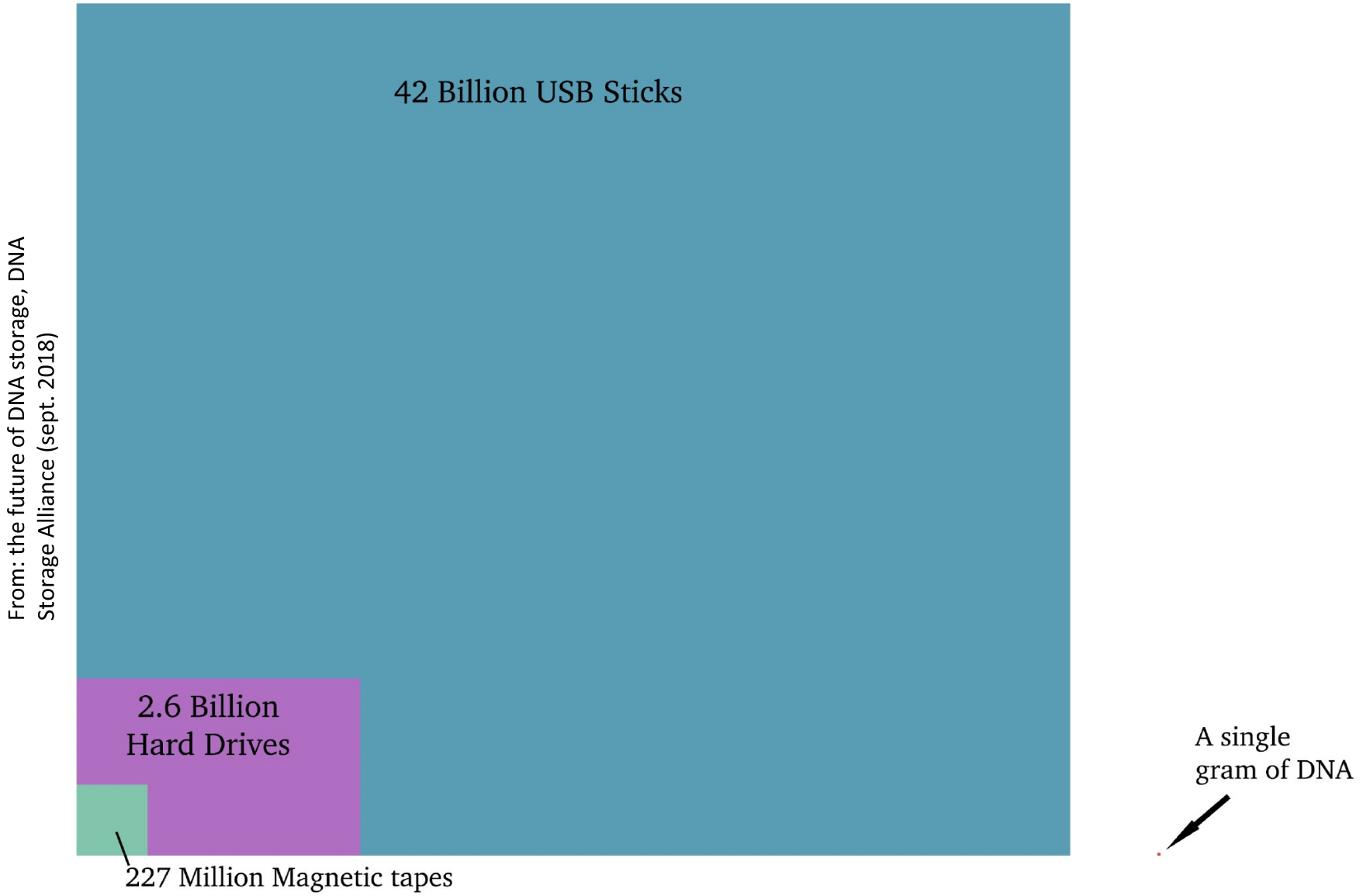
1 g of DNA can contain (in theory) 455 EB

→ 175 ZB can be contained within about 400 g of DNA

DNA is only synthesized on demand

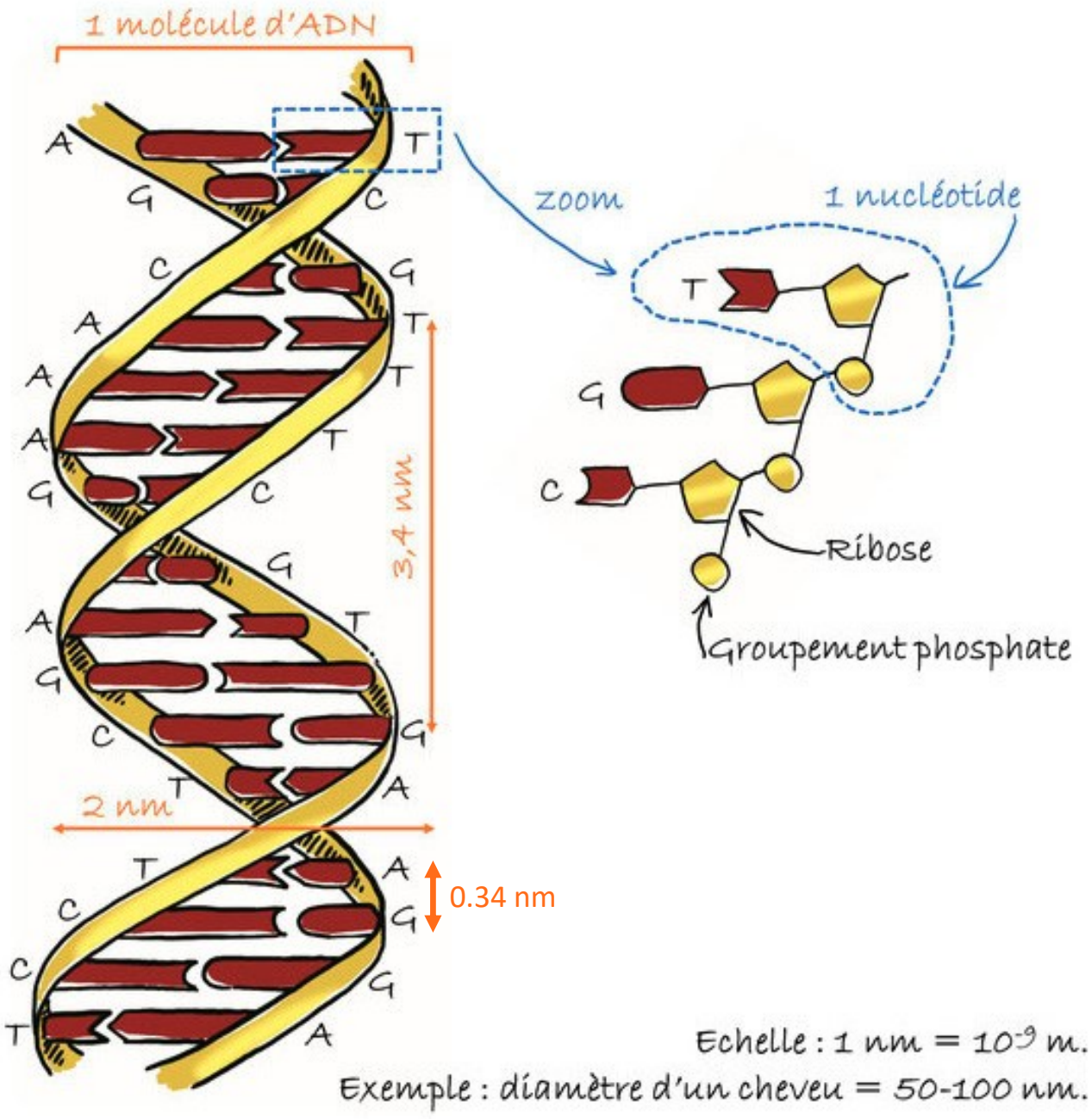
In practice currenly about 30 PB/g

Density comparison: Example for 40 ZB



From: the future of DNA storage, DNA Storage Alliance (sept. 2018)

No quantum tunneling effect **because made of molecules (**14 atoms / bit)



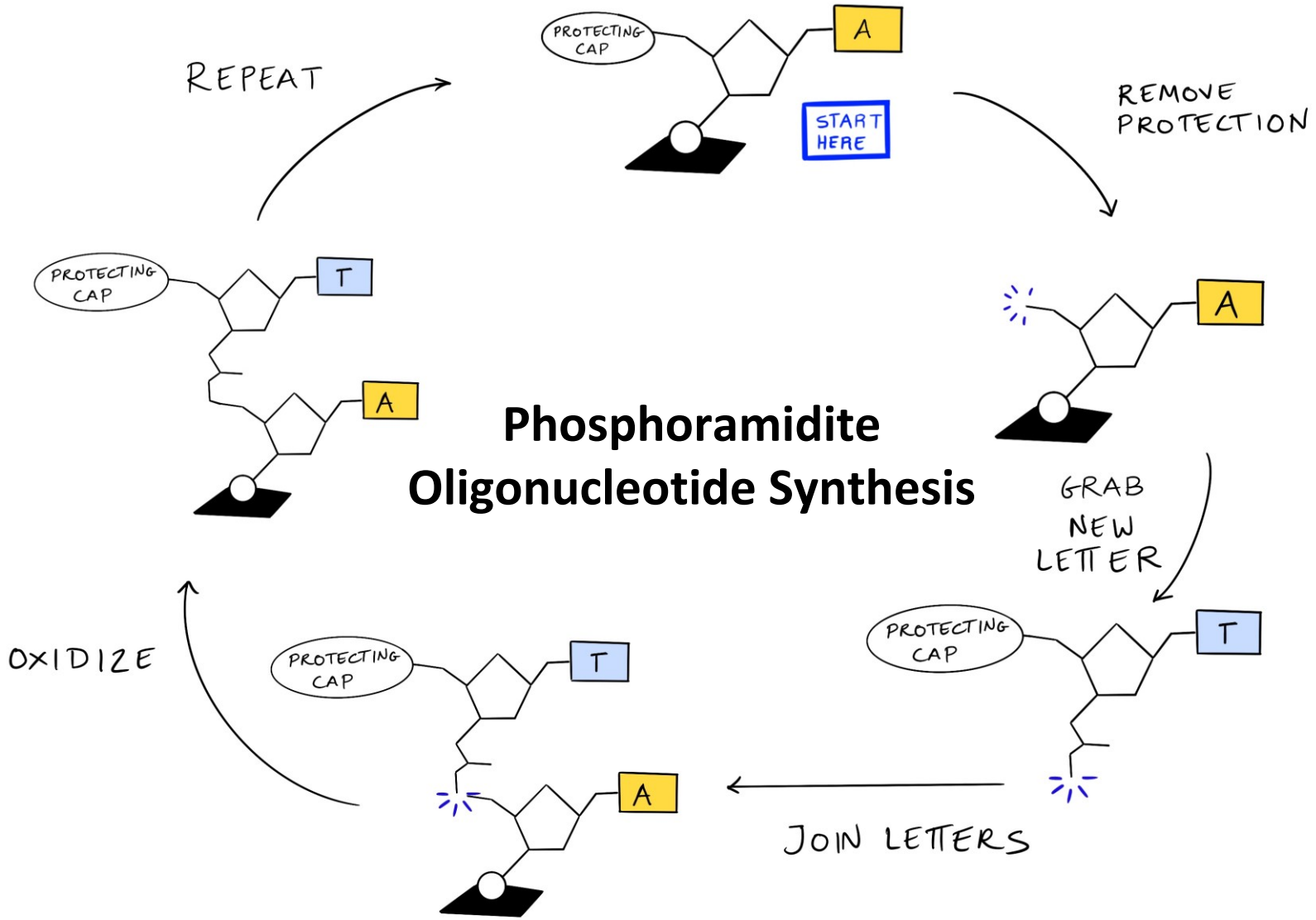
Purine : Adenine (A) and Guanine (G)
Pyrimidine : Thymine (T) and Cytosine (C),
and Uracil (U) instead of T in RNA

Illustration de Jeanne Le Peillet

Environmental Impacts

Some chemistry for the manufacture of oligonucleotides:

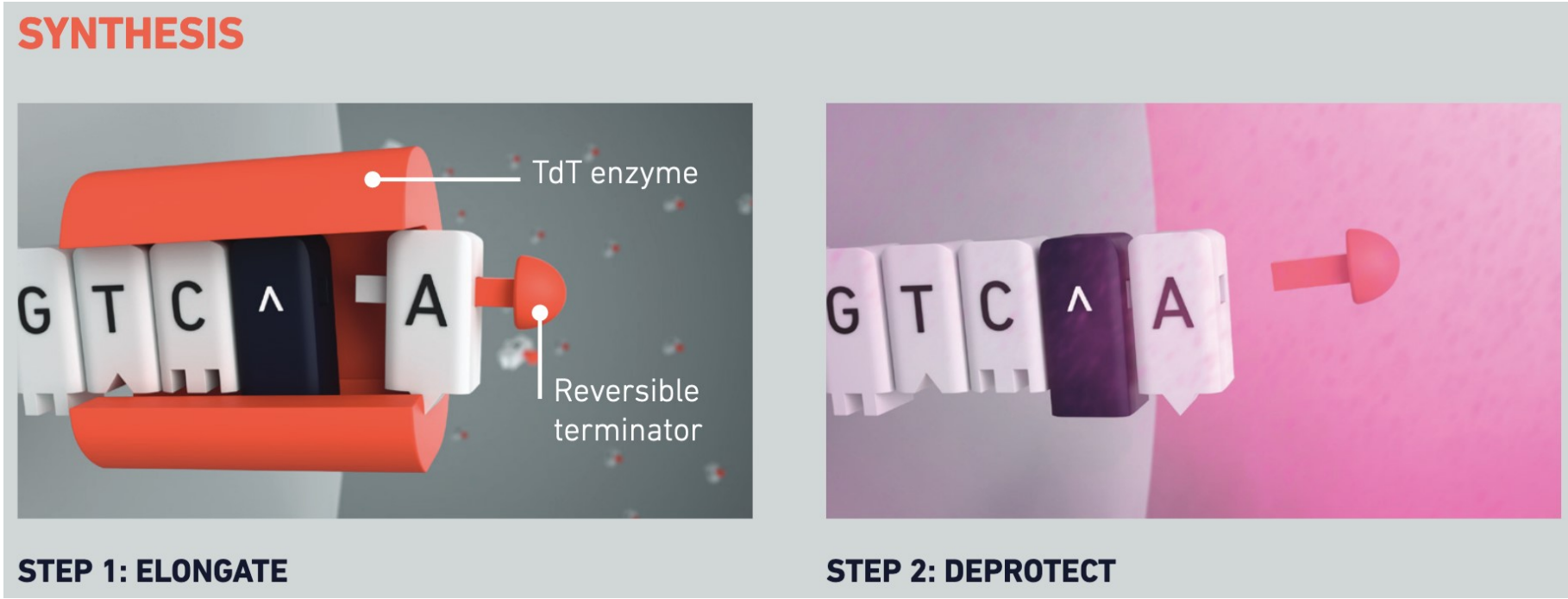
Of the order of 3 L of solvent for 1 g of oligonucleotide



Environmental Impacts

The enzymatic process, inspired by nature will:

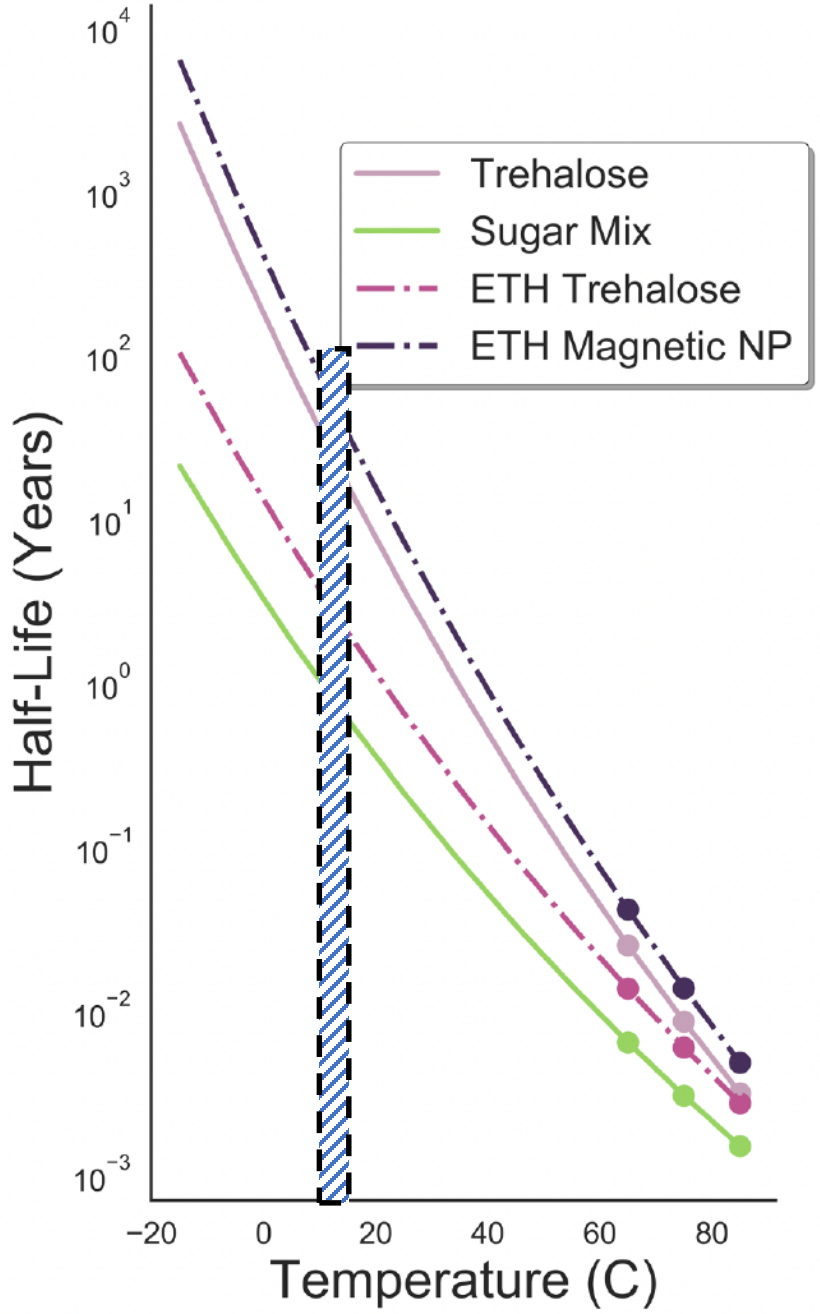
- avoid the use of toxic reagents and thus contribute to reducing the ecological footprint of this industry
- improve speed and efficiency by several orders of magnitude



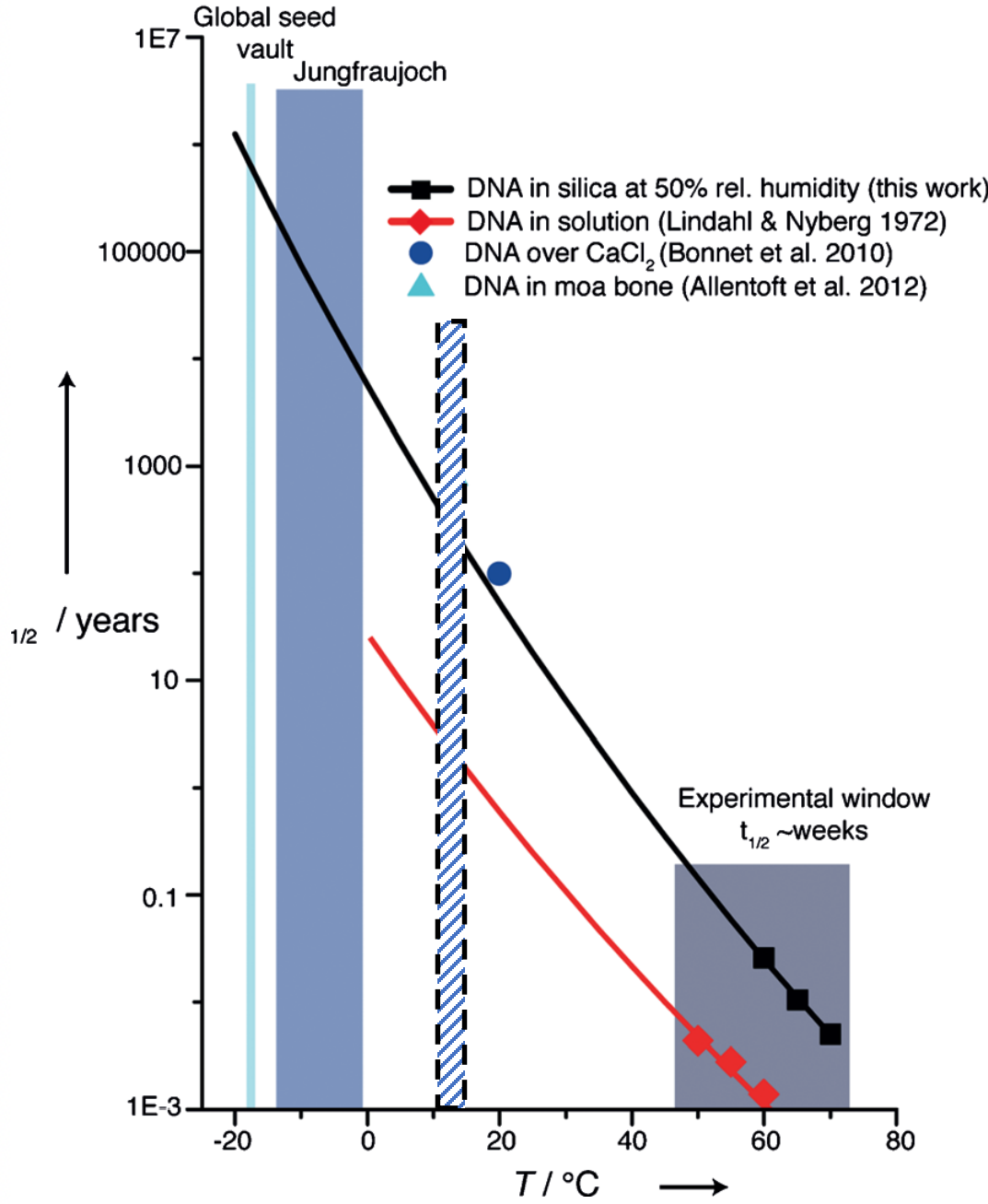
From DNAScript

Storage Obsolescence: Depends on process & temperature

Target temperature range

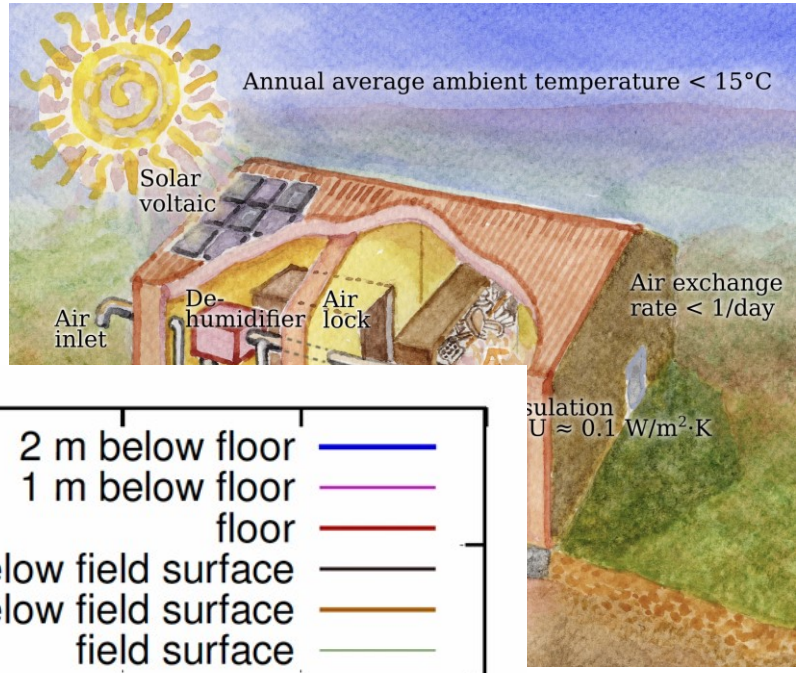


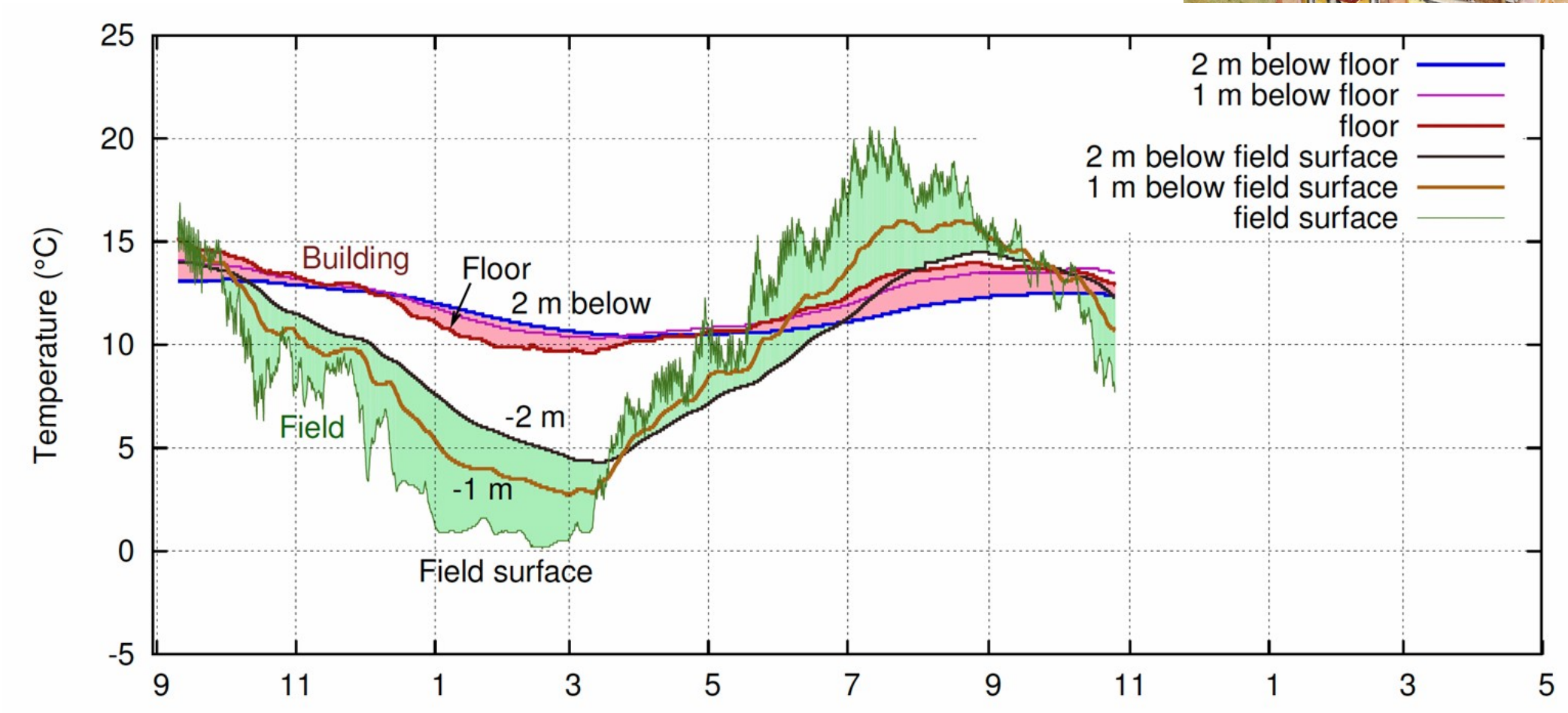
Organick et al. 2020



Grass et al. (2015)

No energy to keep information

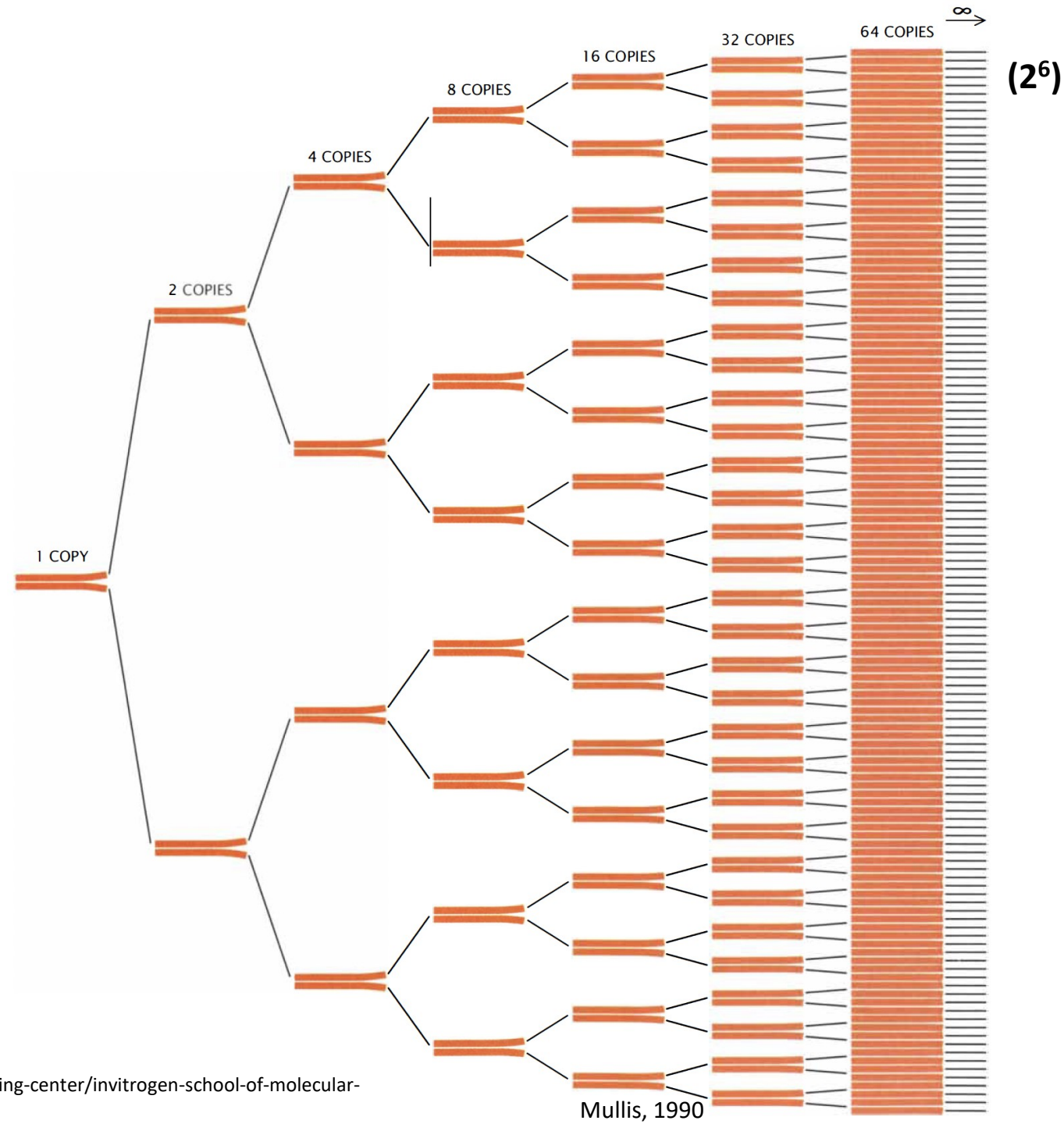




Easy to make copies

Polymerase chain reaction (PCR)

- Fixed time thermal cycling process - does not depend on the volume of data
- Average power of approximately 1.0 W per cycle !



Mullis, 1990

See for instance https://www.thermofisher.com/ch/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/pcr-education/pcr-reagents-enzymes/pcr-cycling-considerations.html

Conclusions for DNA storage

Its environmental impacts are not commensurate with those of silicon/electronic technologies

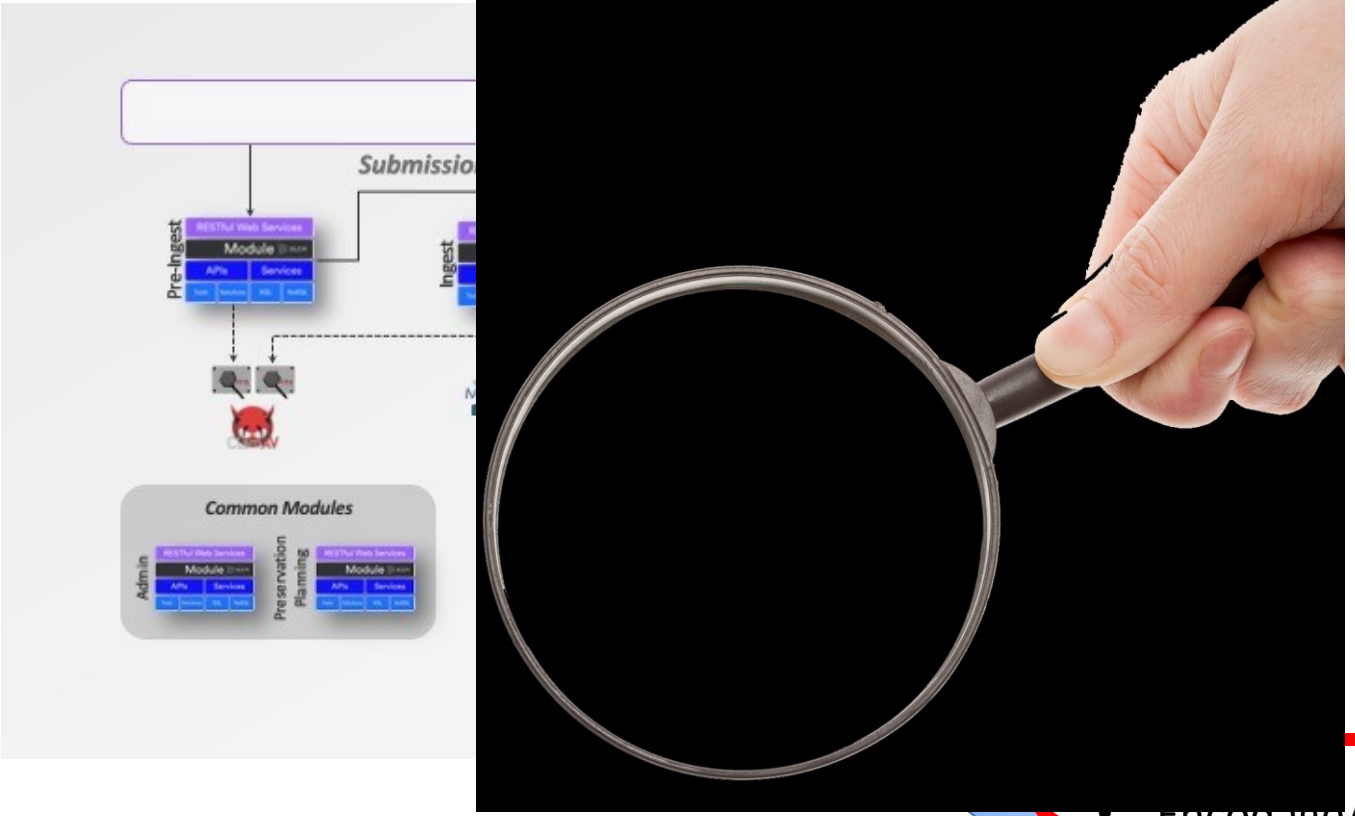
Eight orders of magnitude more dense than traditional media

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

Its technology is evolving faster than Moore’s law

Get ready for the next storage revolution !

Module de stockage : encodage / décodage

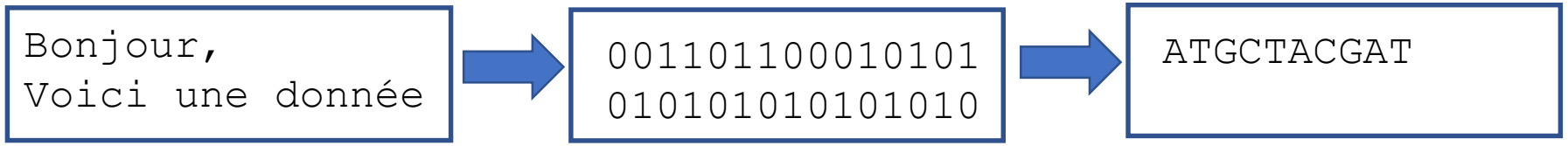


- Encodage/décodage
- Workflow fabrication
- Organisation stockage

Encodage : idée 1

Conversion:

00 -> A
01 -> T
10 -> G
11 -> C



Donnée à encoder

En binaire

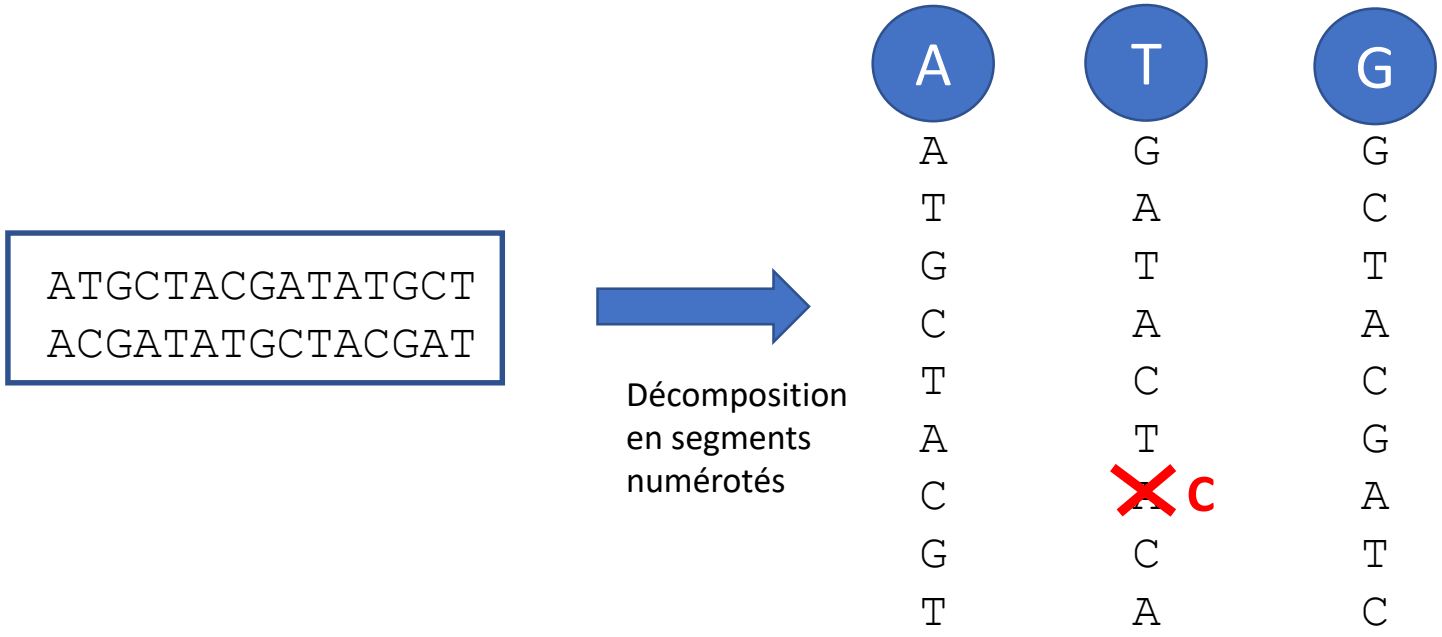
En ADN

Mais:

- Fabrication de long segments difficile et couteuse

Méthode inspirée de: Meiser et al., « Reading and Writing Digital Data in DNA ». Nature Protocols. 01/2020. 10.1038/s41596-019-0244-5

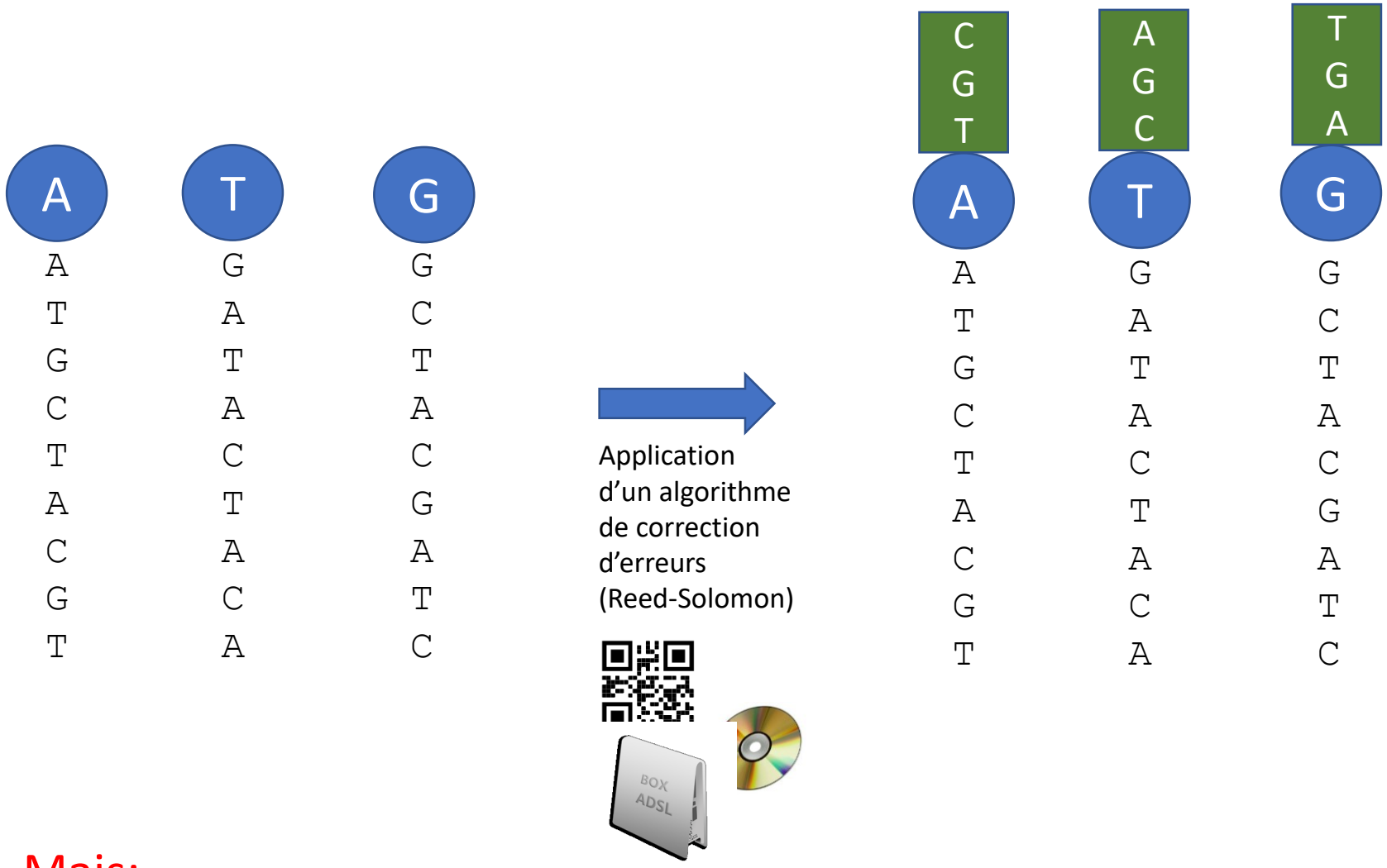
Encodage : idée 2



Mais:

- Si un segment mute, le paquet d'archive est corrompu.

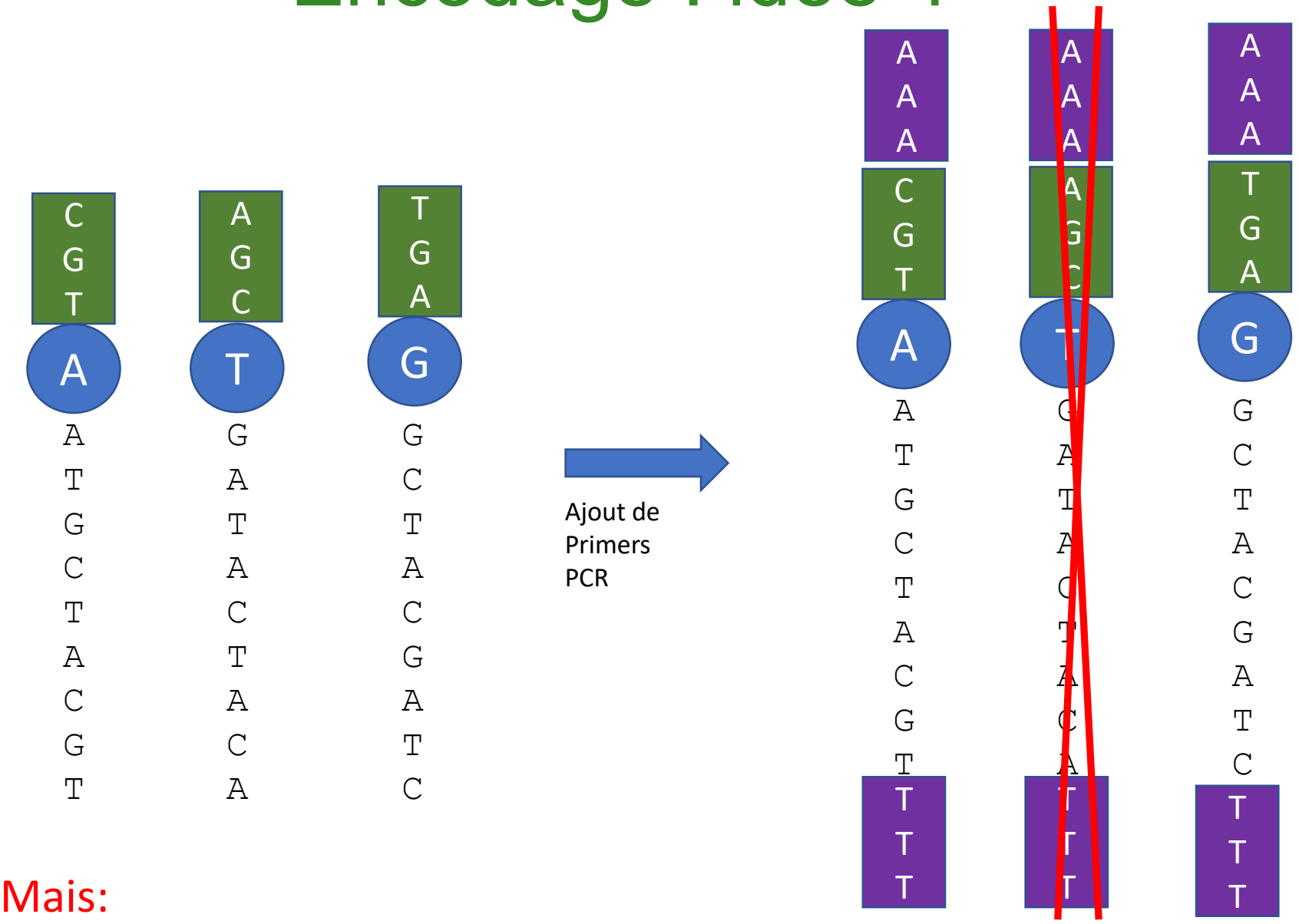
Encodage : idée 3



Mais:

- Comment différentier les paquets d'information?

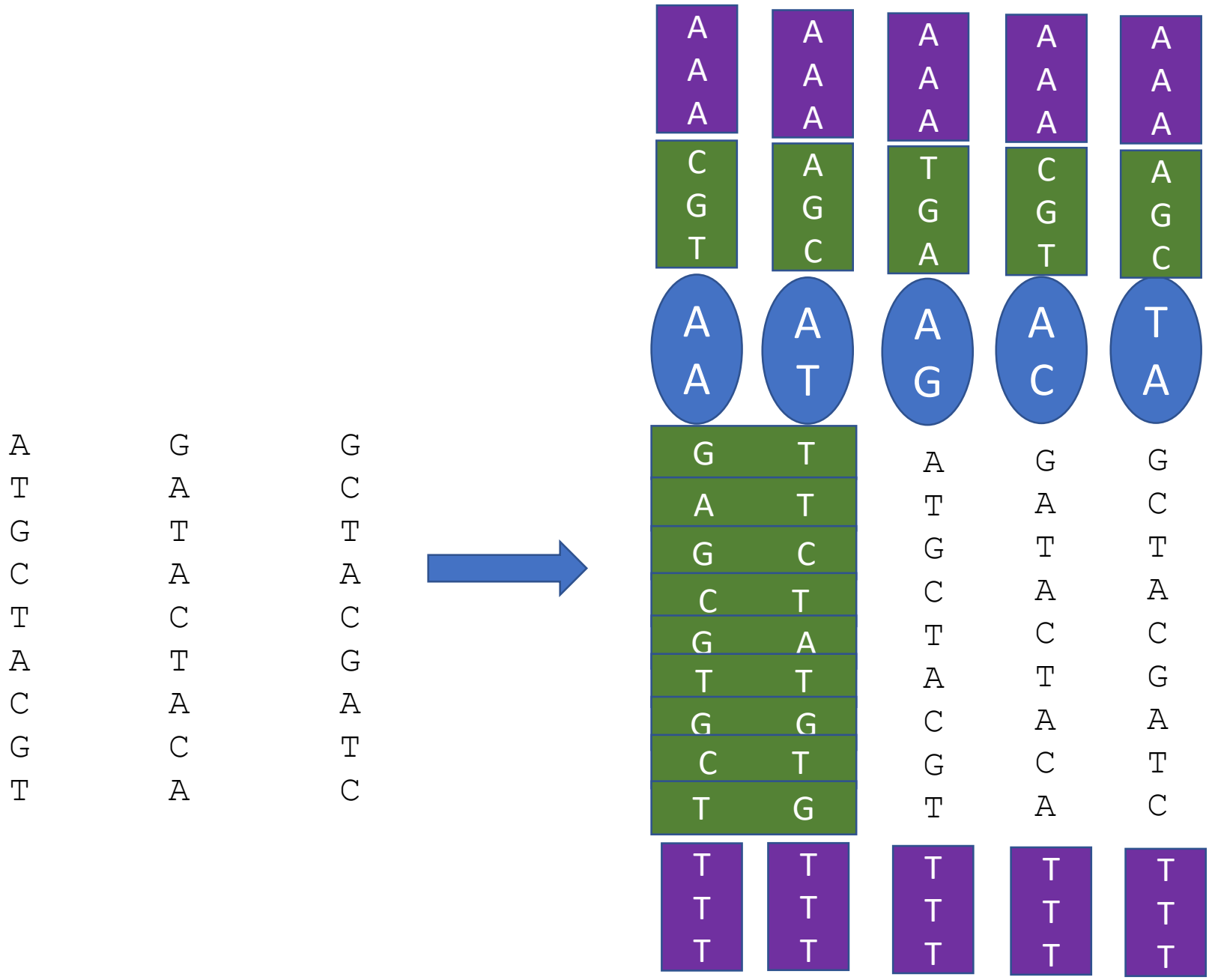
Encodage : idée 4



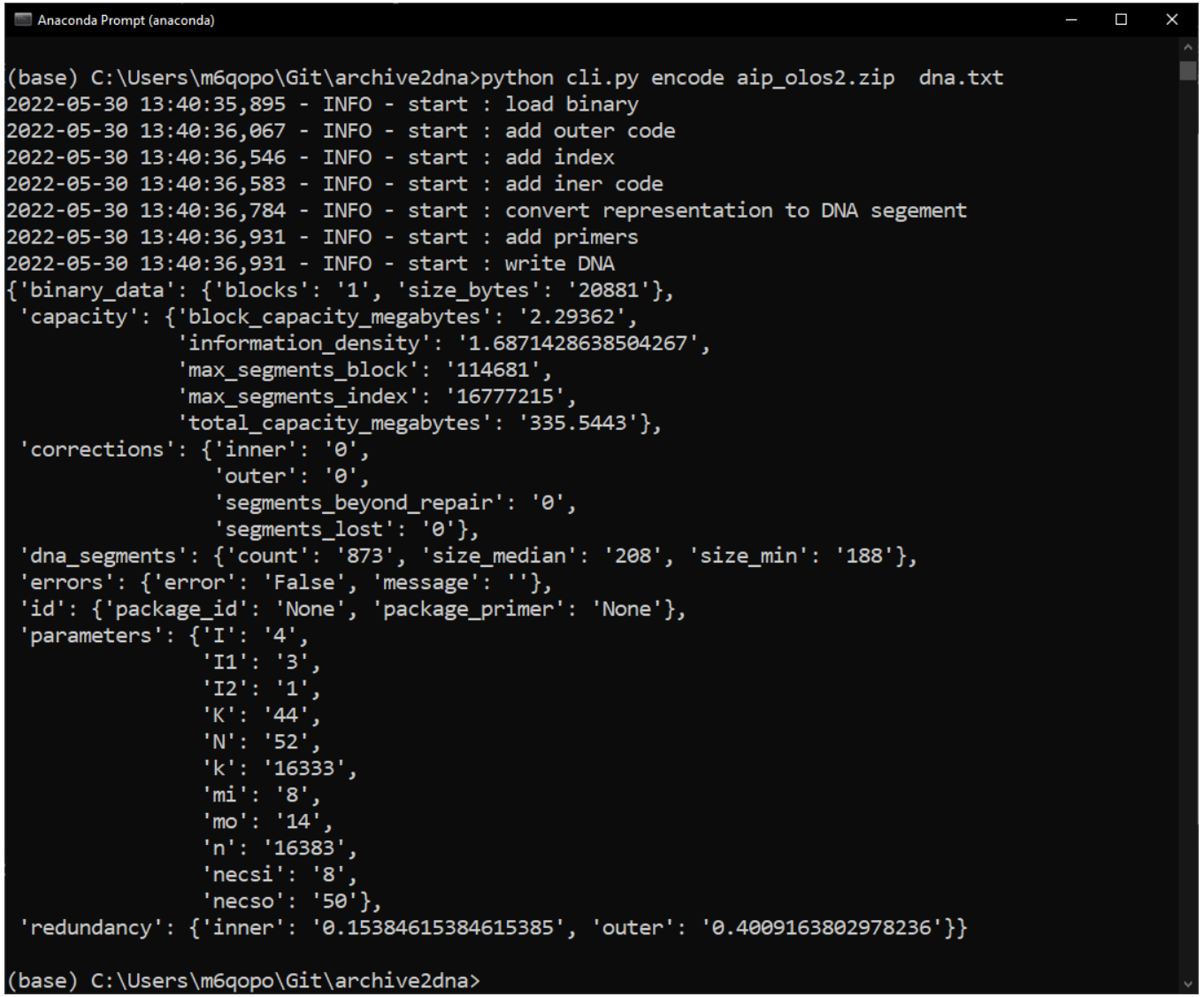
Mais:

- Si on perd un segment le paquet d'information est corrompu.

Encodage : idée 5



Démo



Logiciel libre et open source : https://github.com/jbkrause/archive2dna