

Archives doublement durables

Stockage et archivage sur l'ADN : option durable

1. Juin 2022

Dina Andriamahady, Pierre-Yves Burgi, Jan Krause



**UNIVERSITÉ
DE GENÈVE**



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

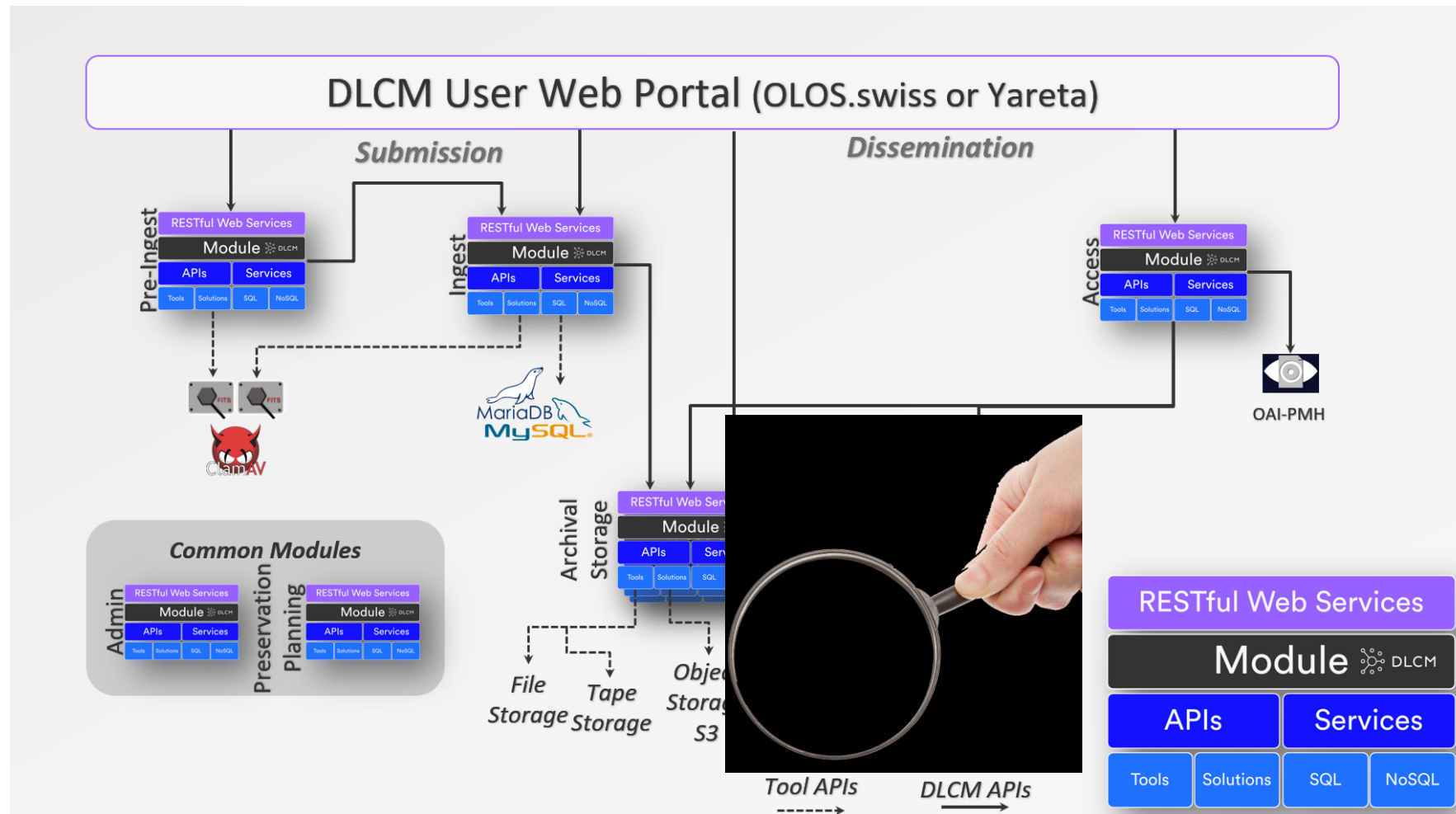


ICS 49 49.140

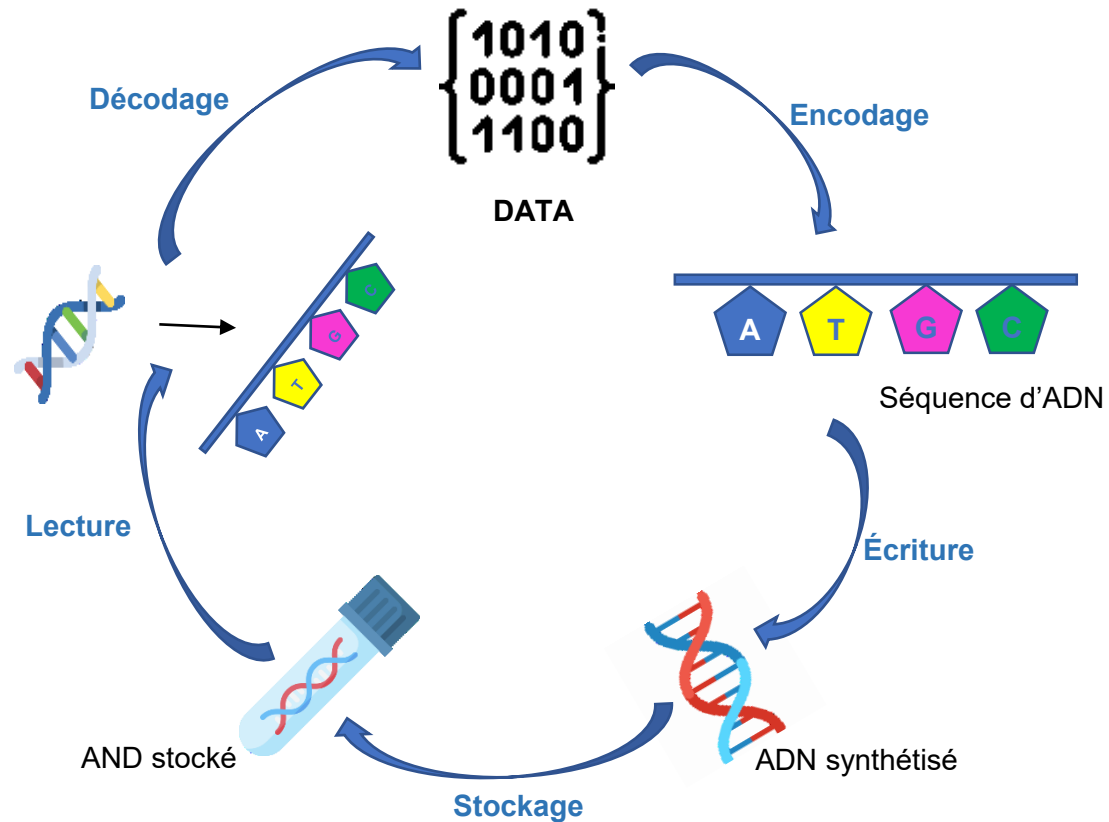
ISO 14721:2012

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

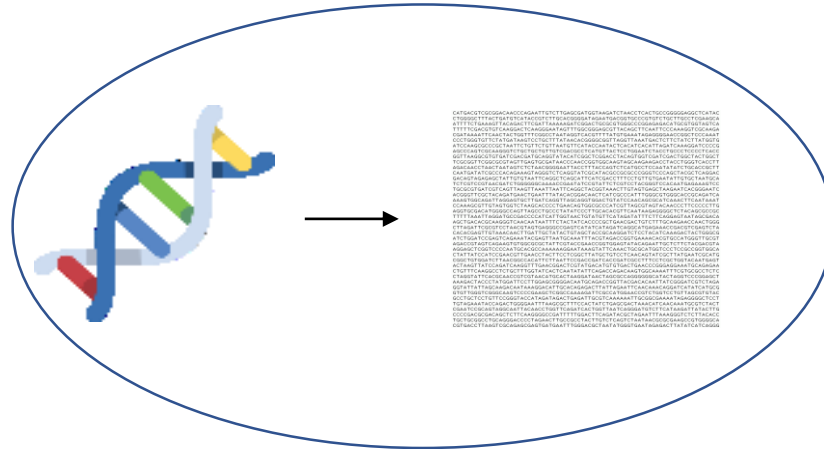
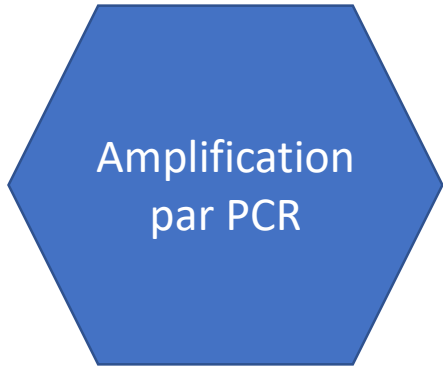
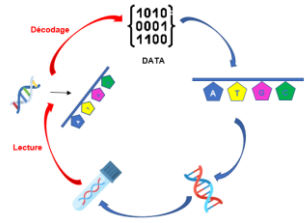
OLOS: la solution suisse pour la gestion des données de la recherche



ADN, support de stockage d'archivage numérique

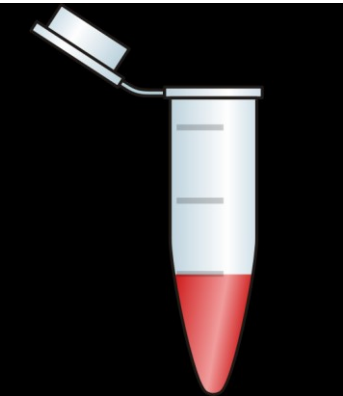


Phases de lecture et de décodage



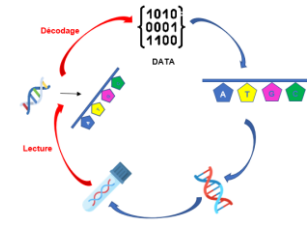
Séquençage de la molécule d'ADN pour en obtenir la séquence

Décodage



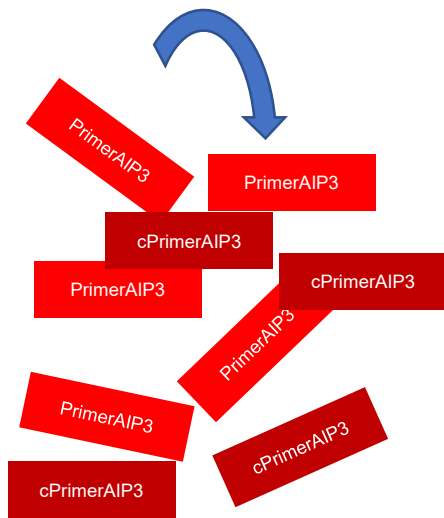
Molécule d'ADN synthétisée puis stockée

```
01110100111001011000011101100110011110100010001011101100100101
111010001010111110101011110101000101011011001100010001010101
011111000010001010001111110010111000011101011101011001011100111
001011010000000011110111101101110100100001110111001011100011100001
111101110110100111010111010001100010001000100010101101000000100
011011010101100100000111010110010010010000001101010100001100101
1110100000110110101010101100101011001010100110001001000000110110
01000101010011101010110110010000110000100010101010100001100101
1110101011010101101101101101101101101101101101101101101101101101
1000010000010110110101010101000100101110101100100000101000011
0110110010010000100001110001010101010001100001001011001001100100100
11100100001000010110111000001101111010110101010000101000100000
1100010110110110111101010101010101010101010101010101010101010101
000001101011001011000110100000100001000010000100001000010000100001
01101100100010110000111000010000100001000010000100001000010000100001
10010001001001001000001101101010101101011010101000010000010110010
011001000000001011101101101101101101101101101101101101101101101101
001010001111010101101101101101101101101101101101101101101101101101
0010100001001000101010101010101010101010101010101010101010101010101
00101100001001000101010101010101010101010101010101010101010101010101
00101100001001000101010101010101010101010101010101010101010101010101
01101001101101010101010101010101010101010101010101010101010101010101
000000101000110100001010000010100011110001110000110001110001110111011
11101011100101101000010001101101101101101101101101101101101101101101
01110101101101010101010101010101010101010101010101010101010101010101
11001000101101010101010101010101010101010101010101010101010101010101
001101000110000010001001111101011100001100011110001010010110001111
001110100000000111101110101010101010101010101010101010101010101010101
```

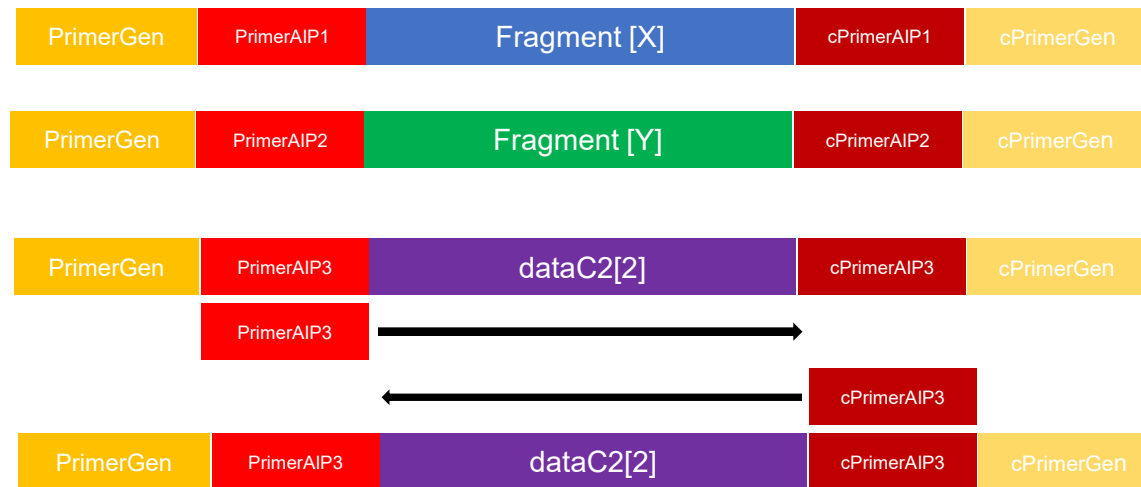



Accès ciblé à une information: la force des primers

Addition de primers spécifiques



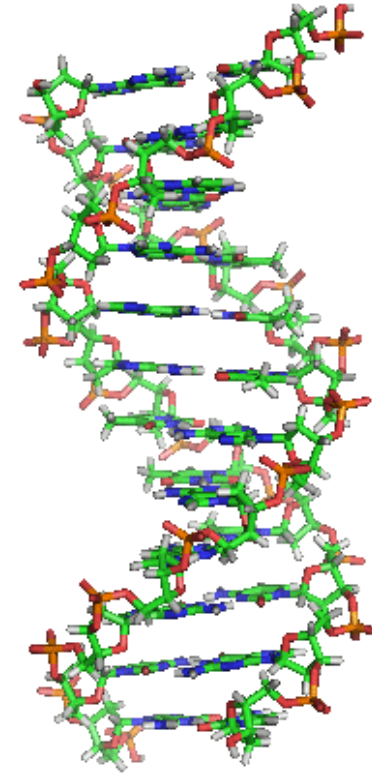
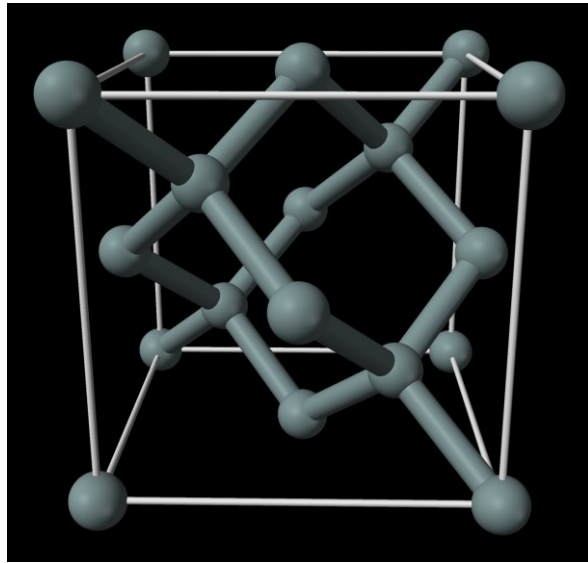
PCR



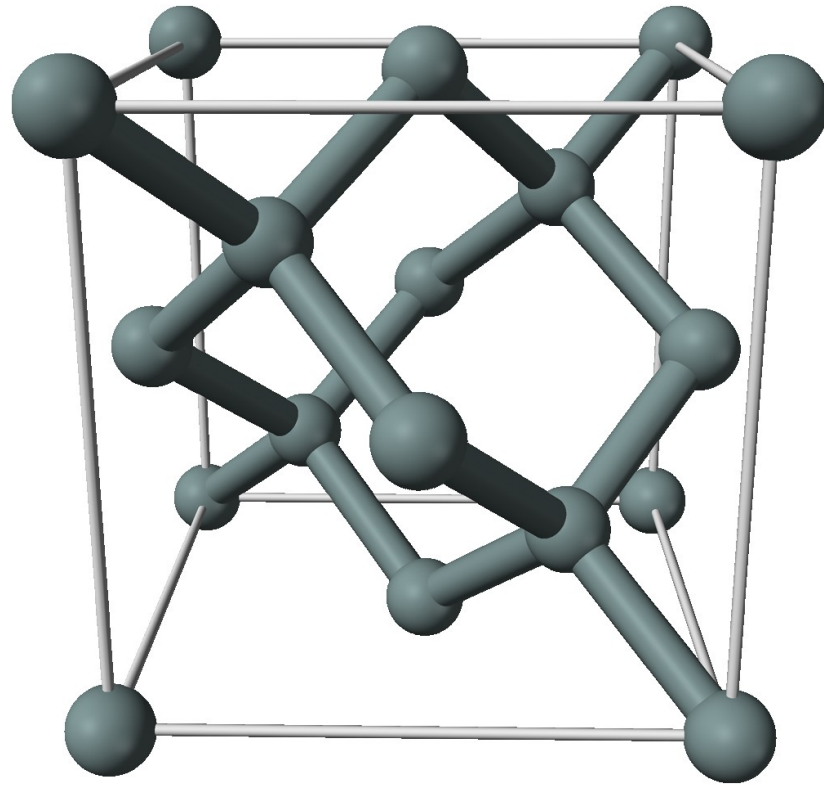
Résultat après un cycle



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}

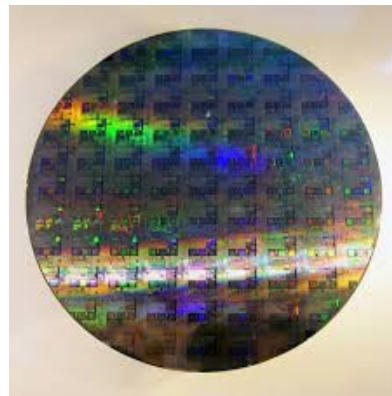
Zetta = 10^{21}

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}

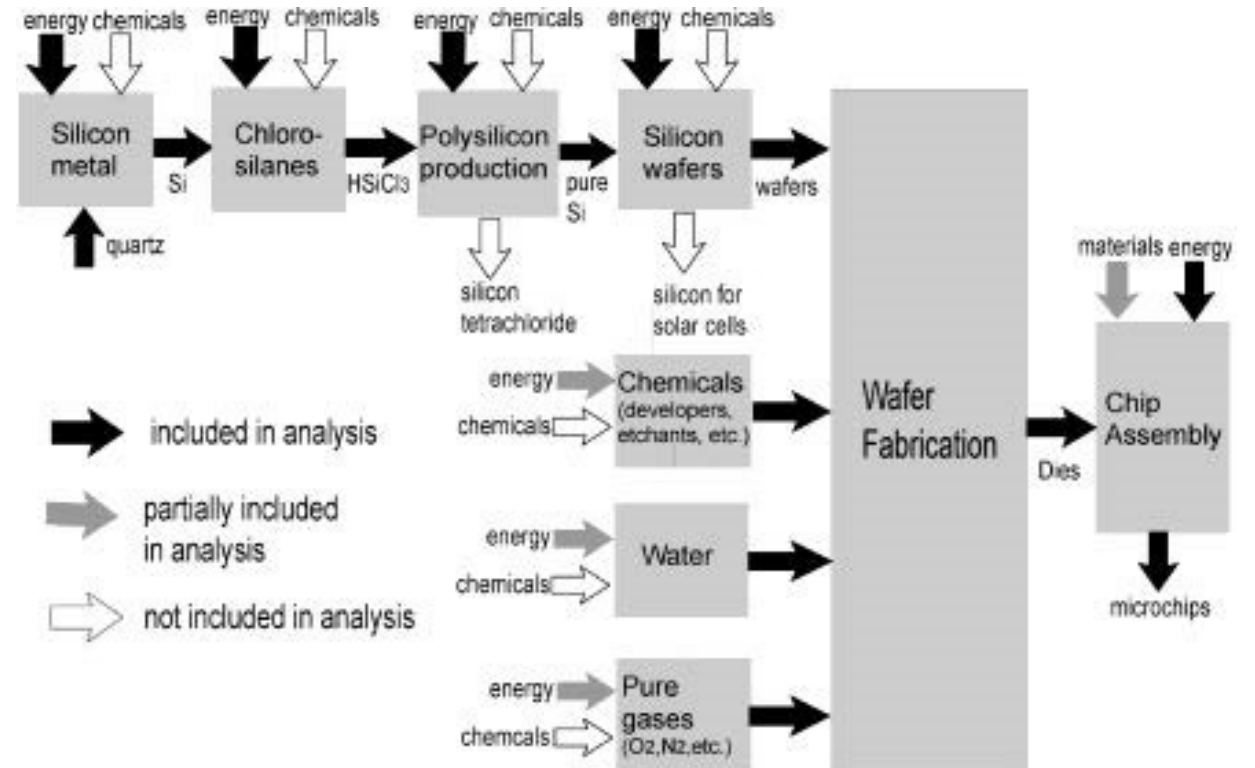
Environmental Impacts

A single 2-g chip requires:

- 1.6 kg secondary fossil fuel
- 72 g chemical inputs

For each cm² silicon wafer:

- 1.5 KWh
- 18-27 L of water

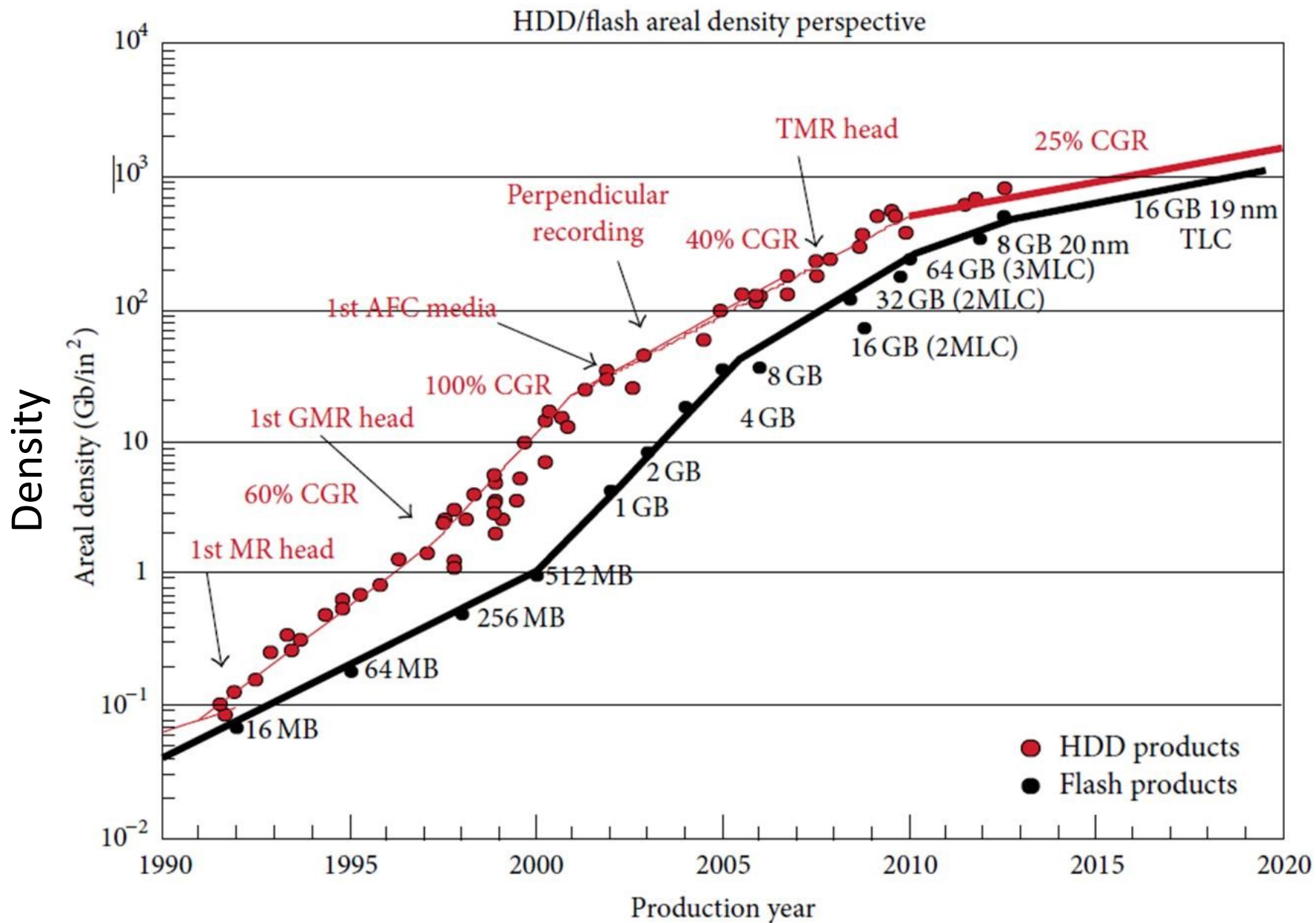


Environmental Impacts

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.

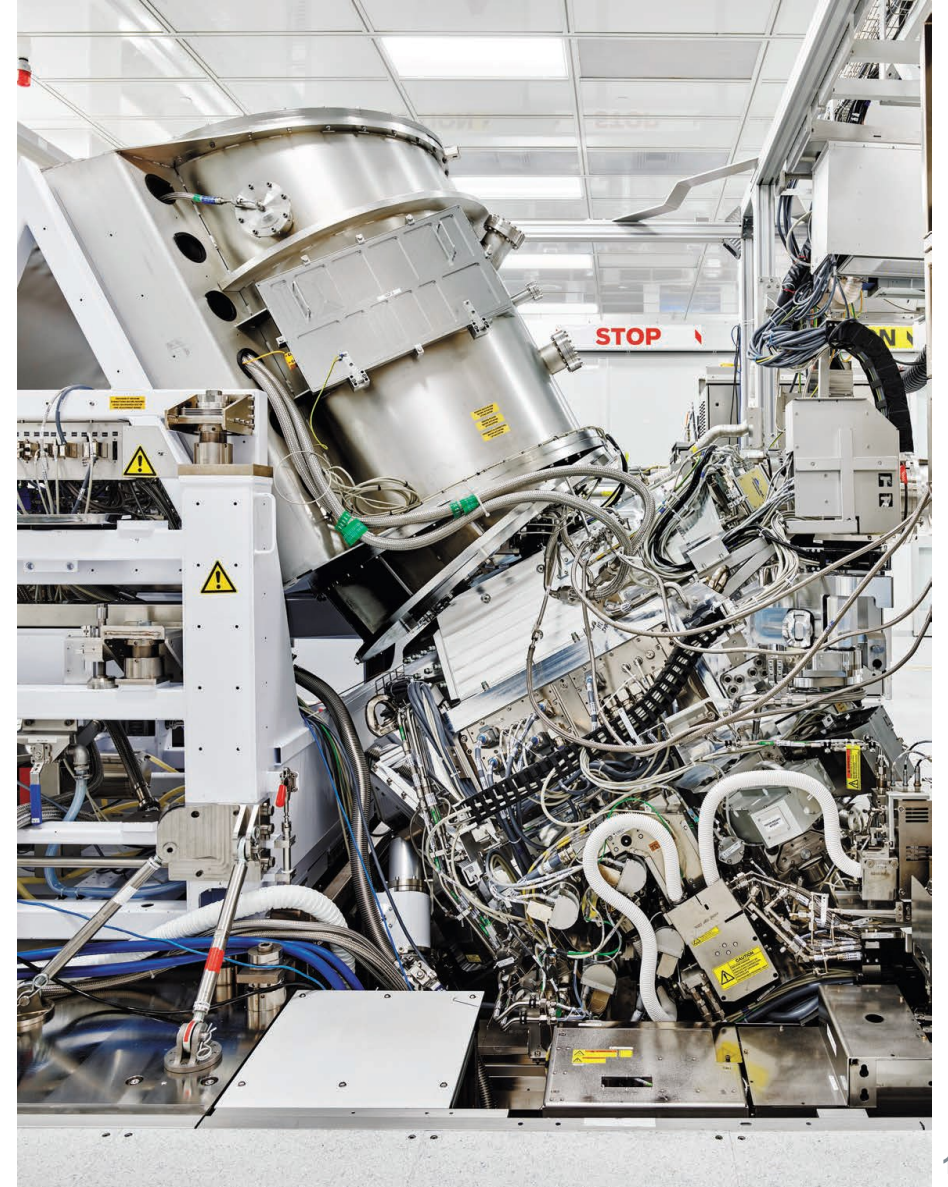
Silicon density reaches its limits



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

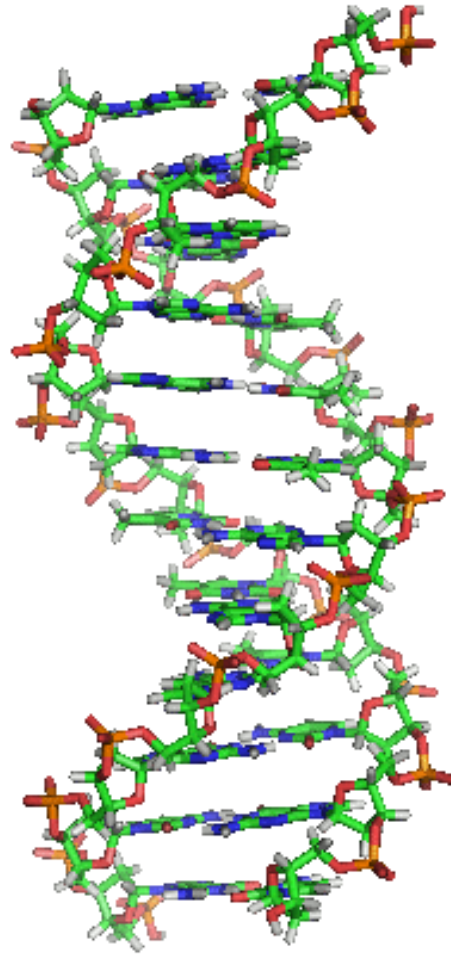


2 generations (8-10 years)



3-5 years

2. DNA



Data growth and DNA

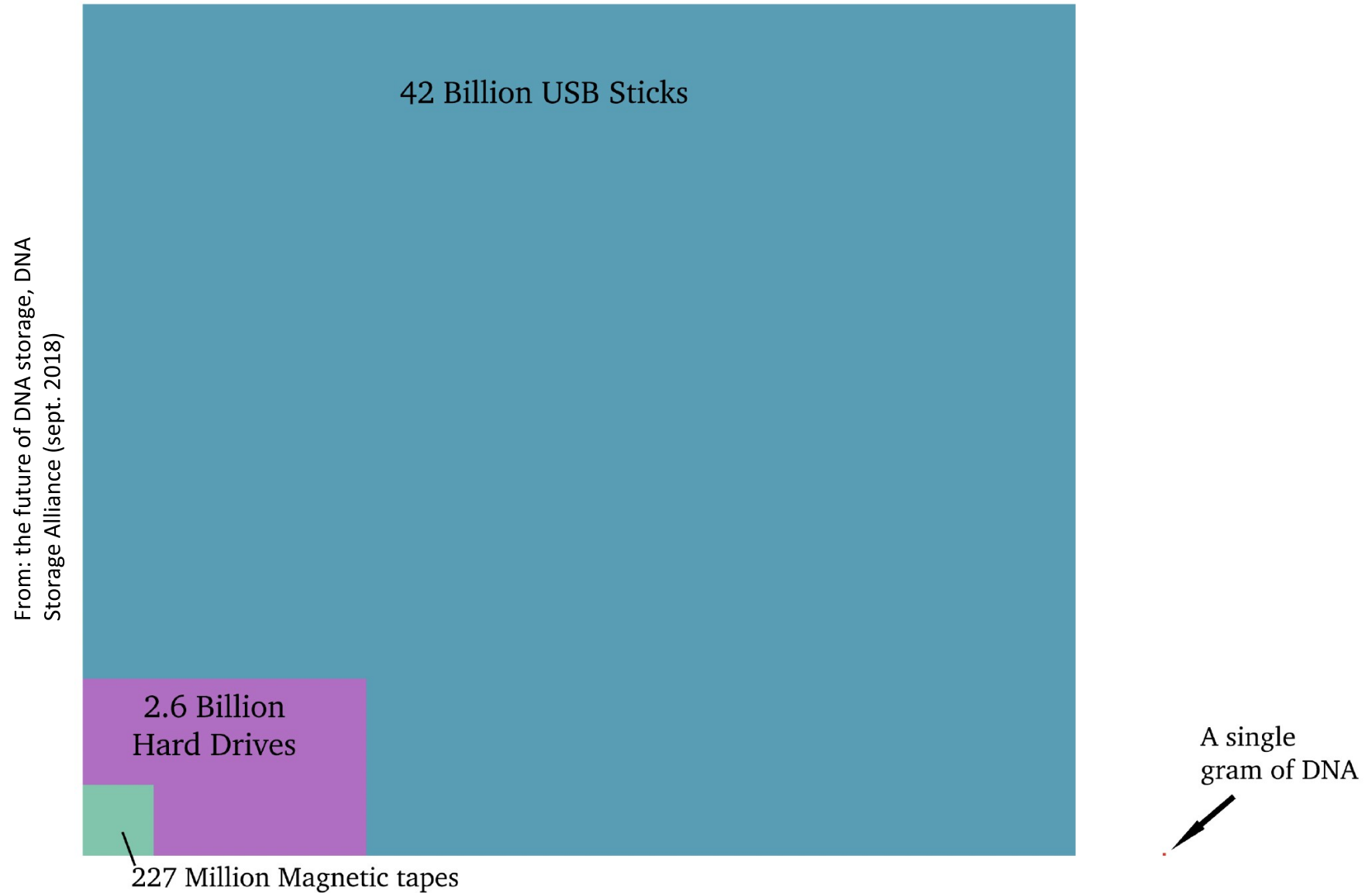
1 g of DNA can contain (in theory) 455 EB

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

DNA is only synthesized on demand

In practice currently about 30 PB/g

Density comparison: Example for 40 ZB



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

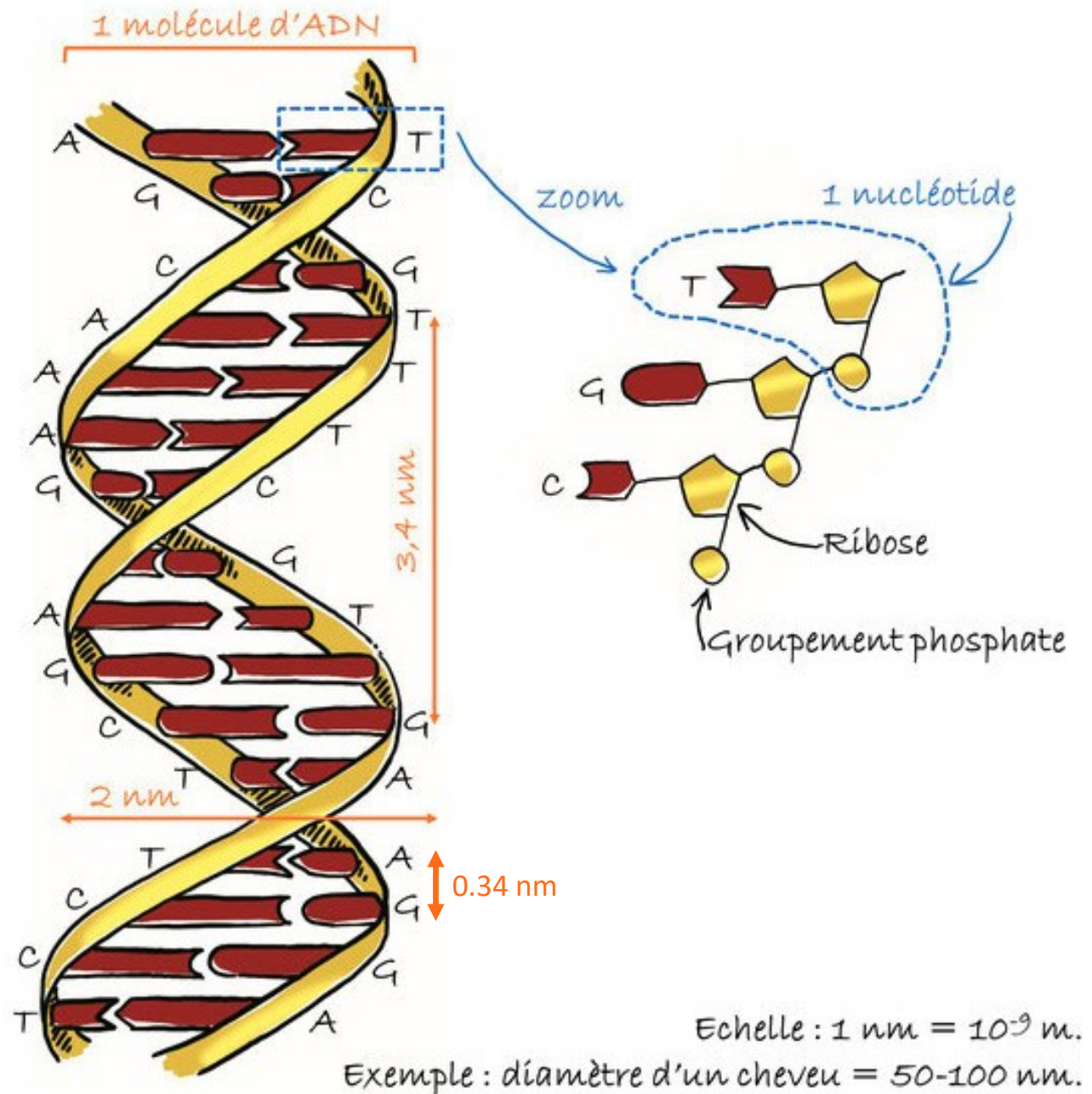
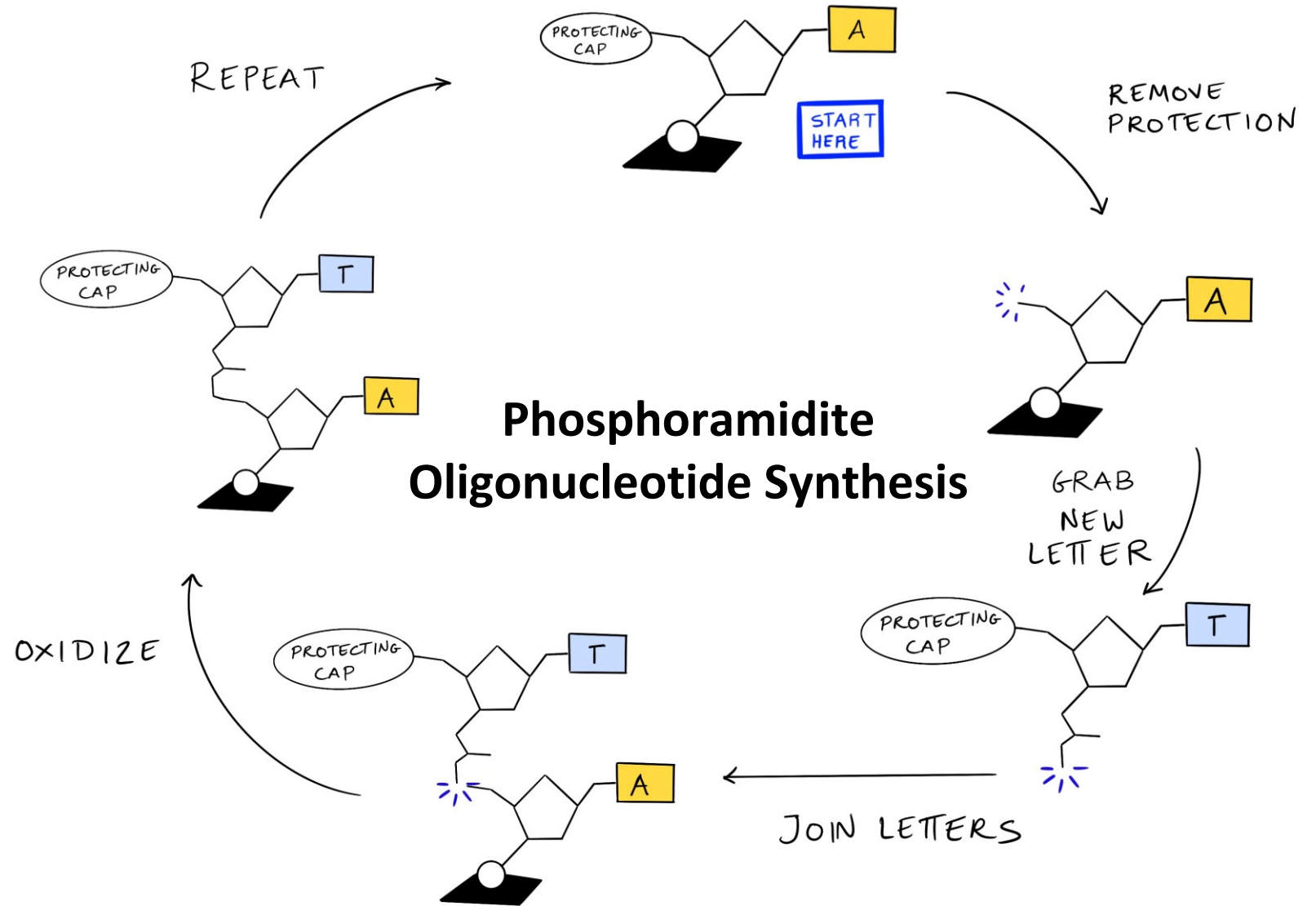


Illustration de Jeanne Le Peillet

Environmental Impacts

Some chemistry for the manufacture of oligonucleotides:

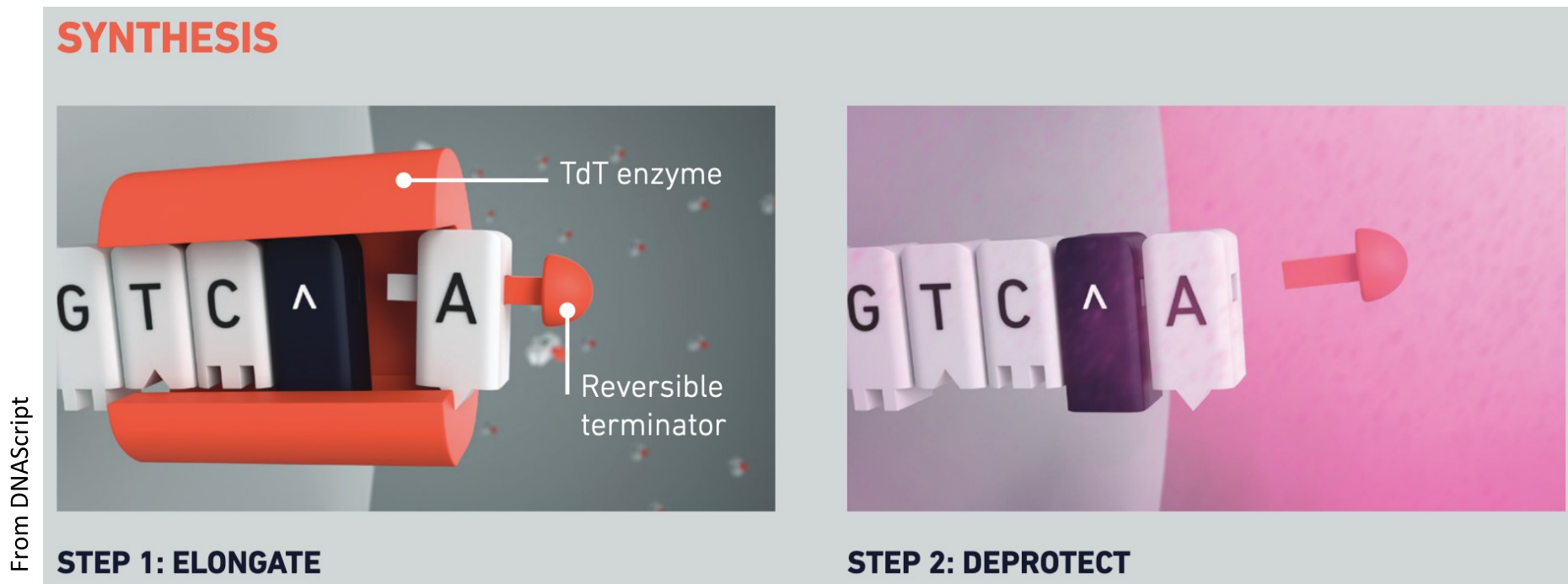
Of the order of 3 L of solvent for 1 g of oligonucleotide




Environmental Impacts

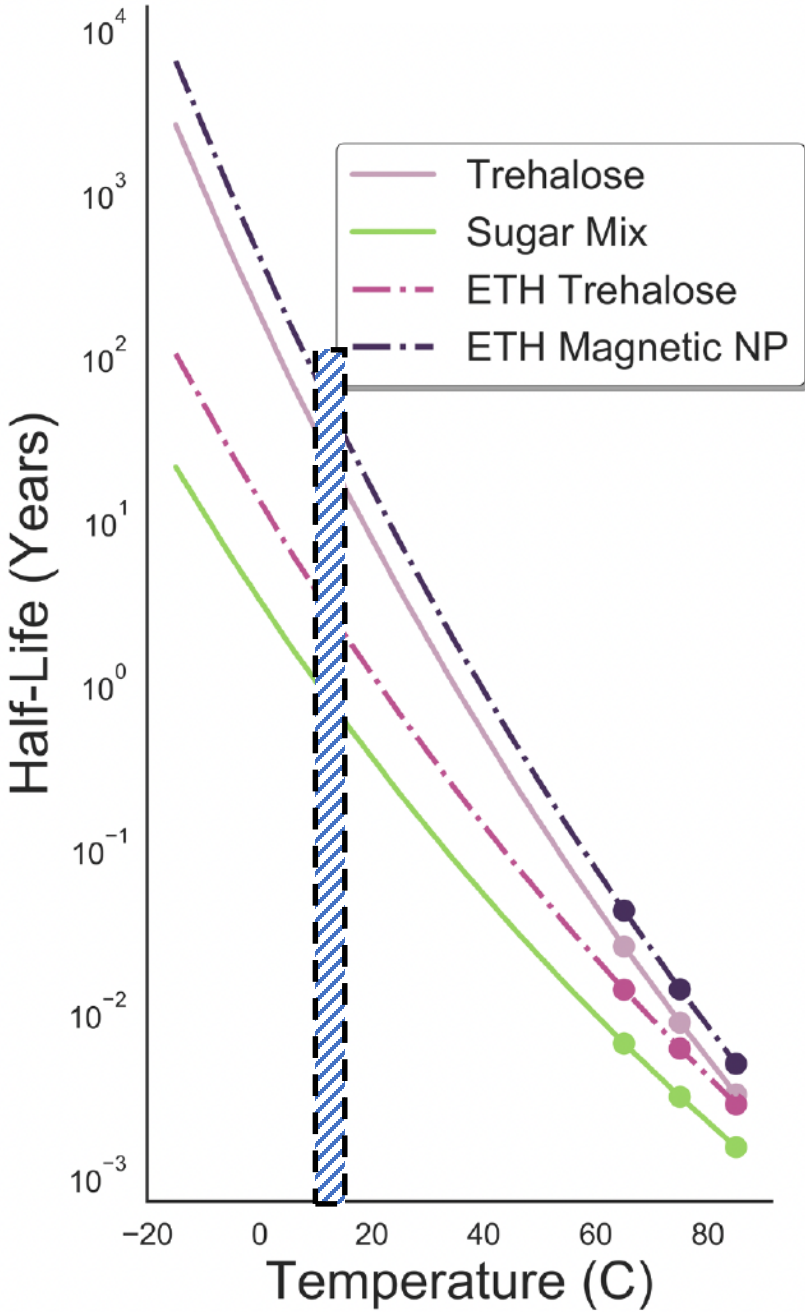
The enzymatic process, inspired by nature will:

- 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

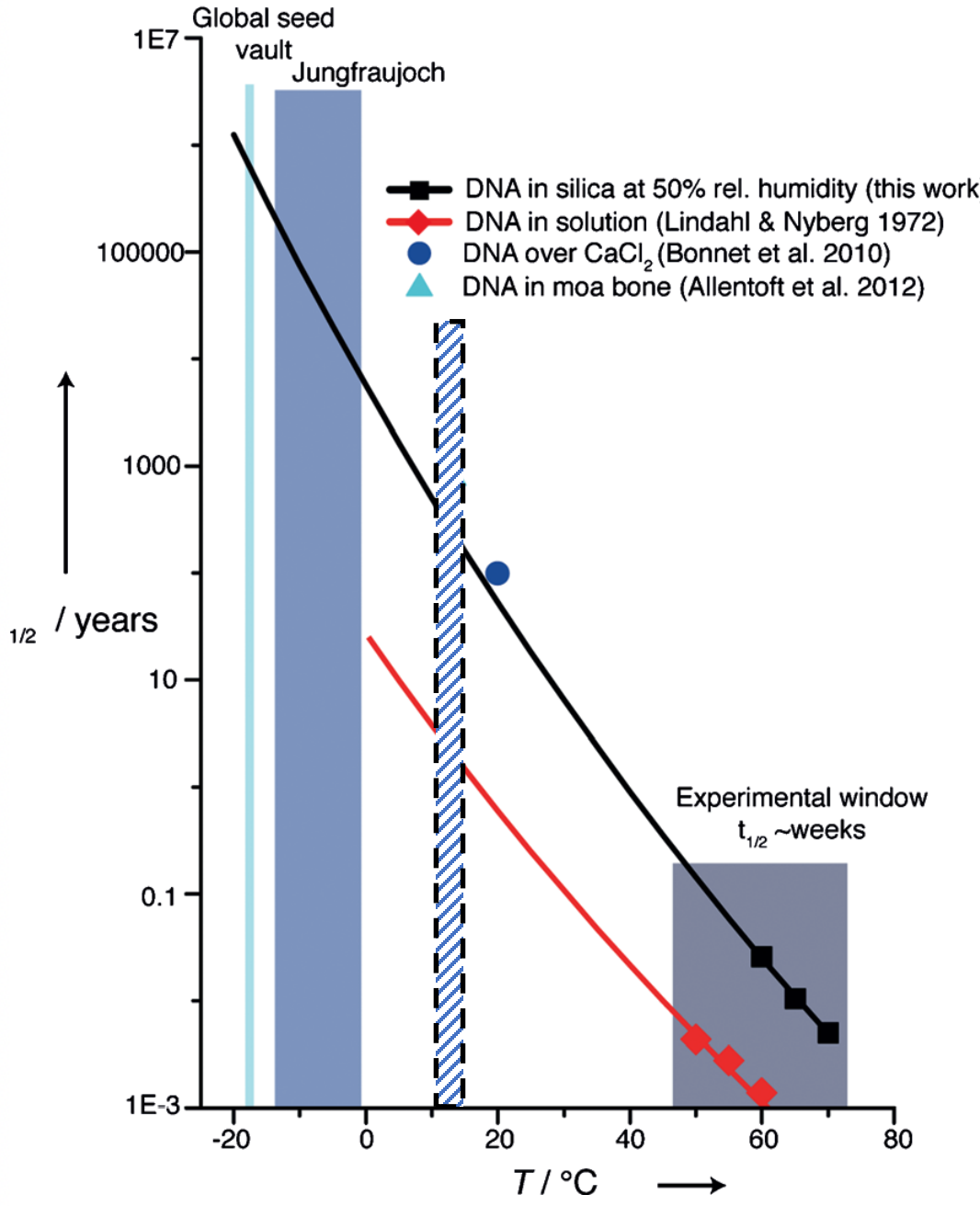


Storage Obsolescence: Depends on process & temperature

 Target temperature range

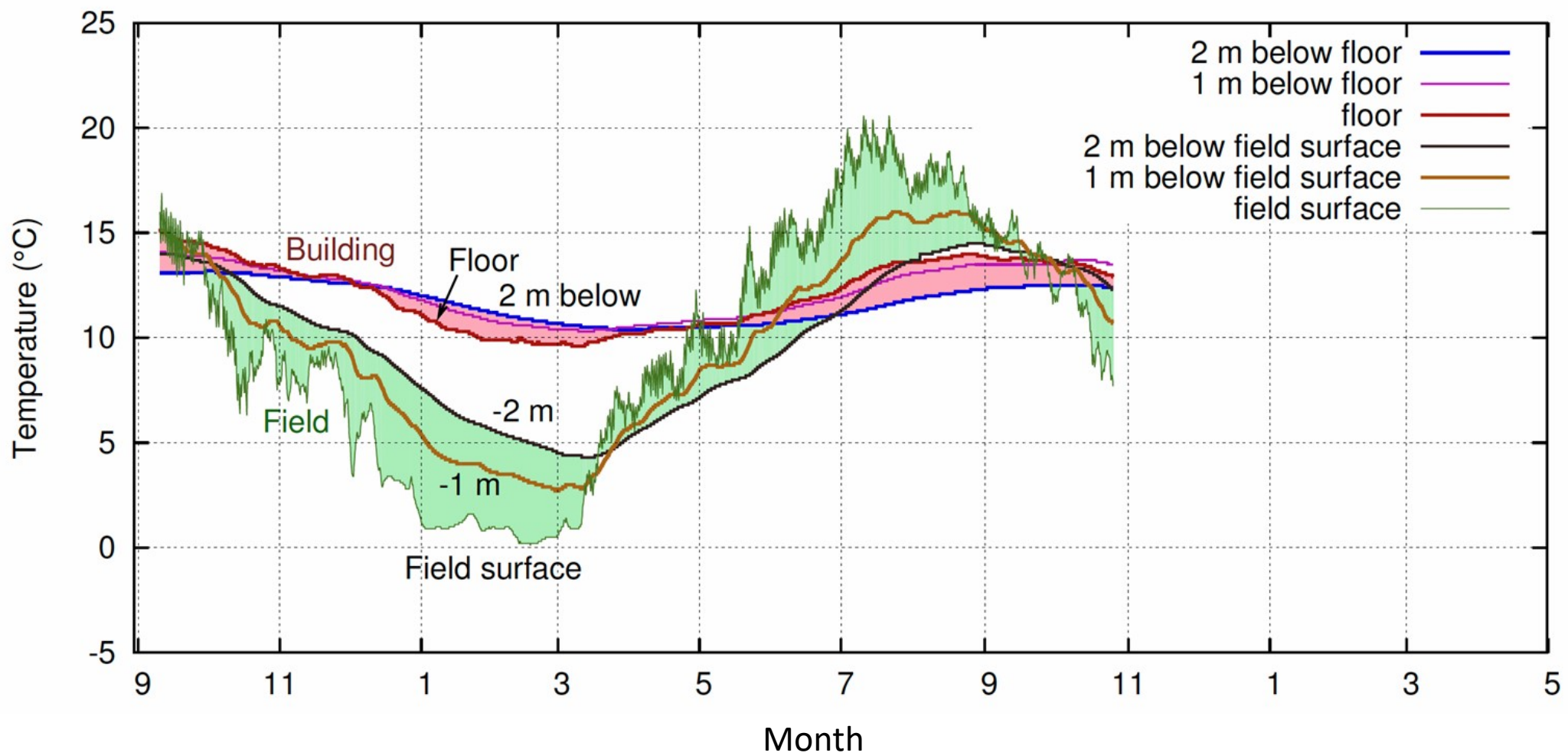
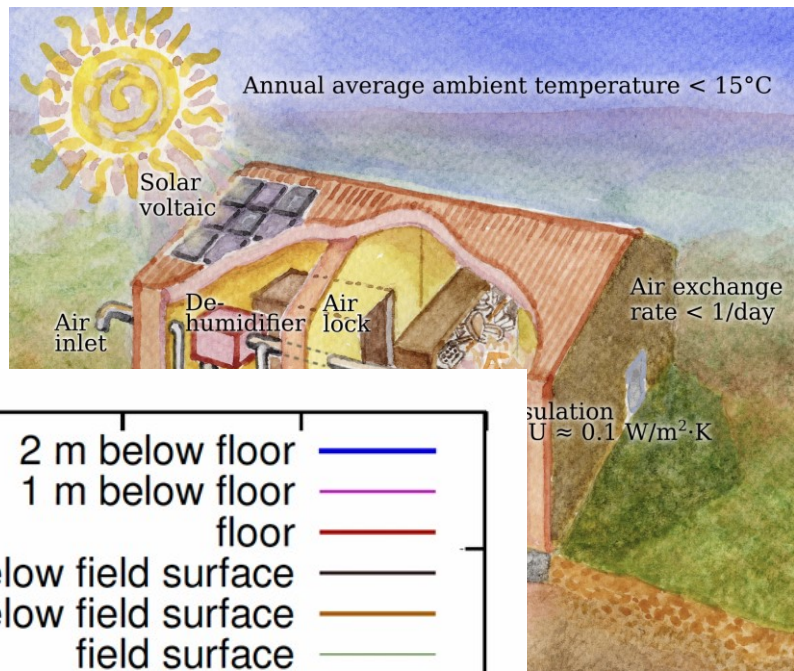


Organick et al. 2020



Grass et al. (2015)

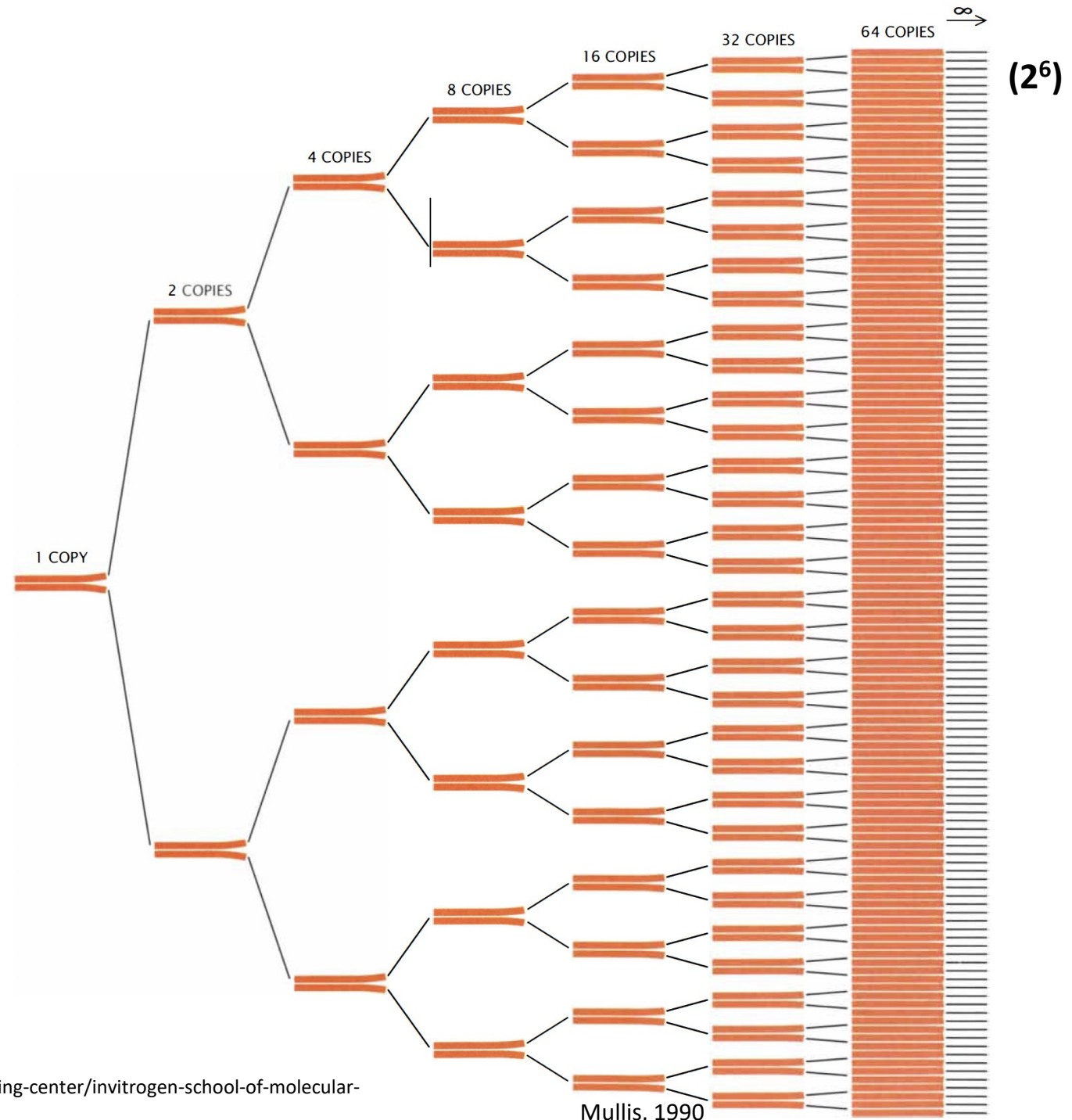
No energy to keep information



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 !



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

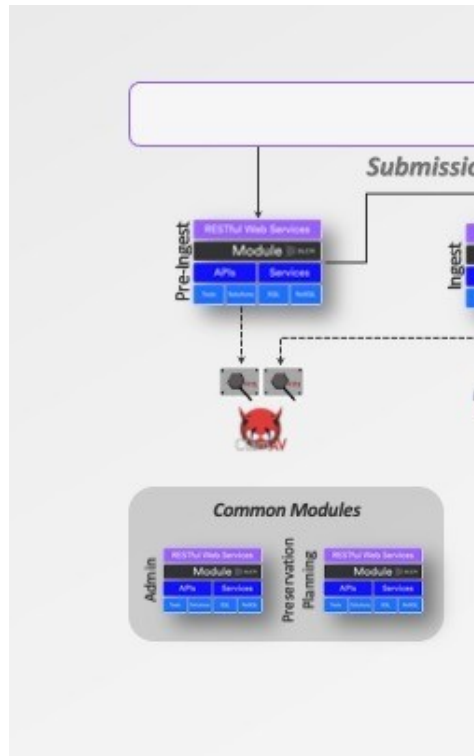
Eight orders of magnitude more dense than traditional media

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

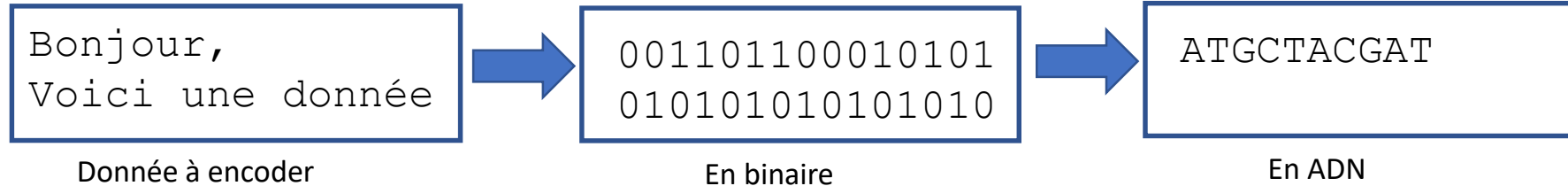
Conversion:

00 -> A

01 -> T

10 -> G

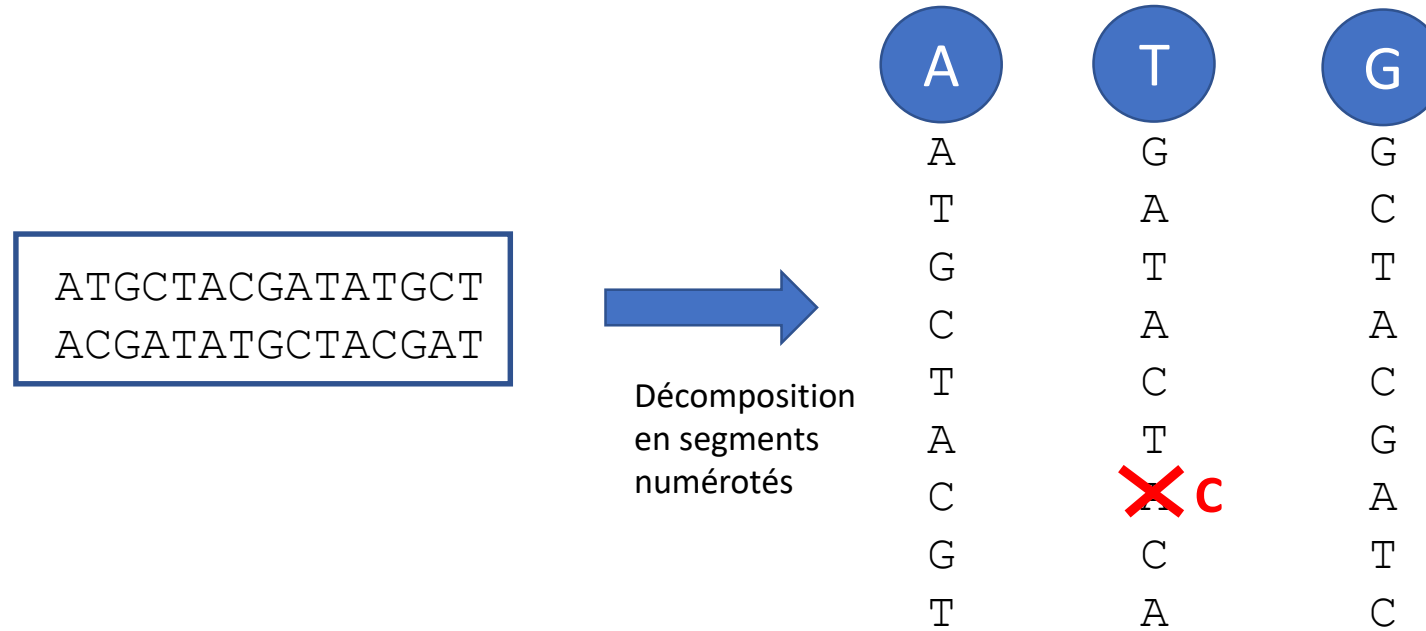
11 -> C



Mais:

- Fabrication de long segments difficile et couteuse

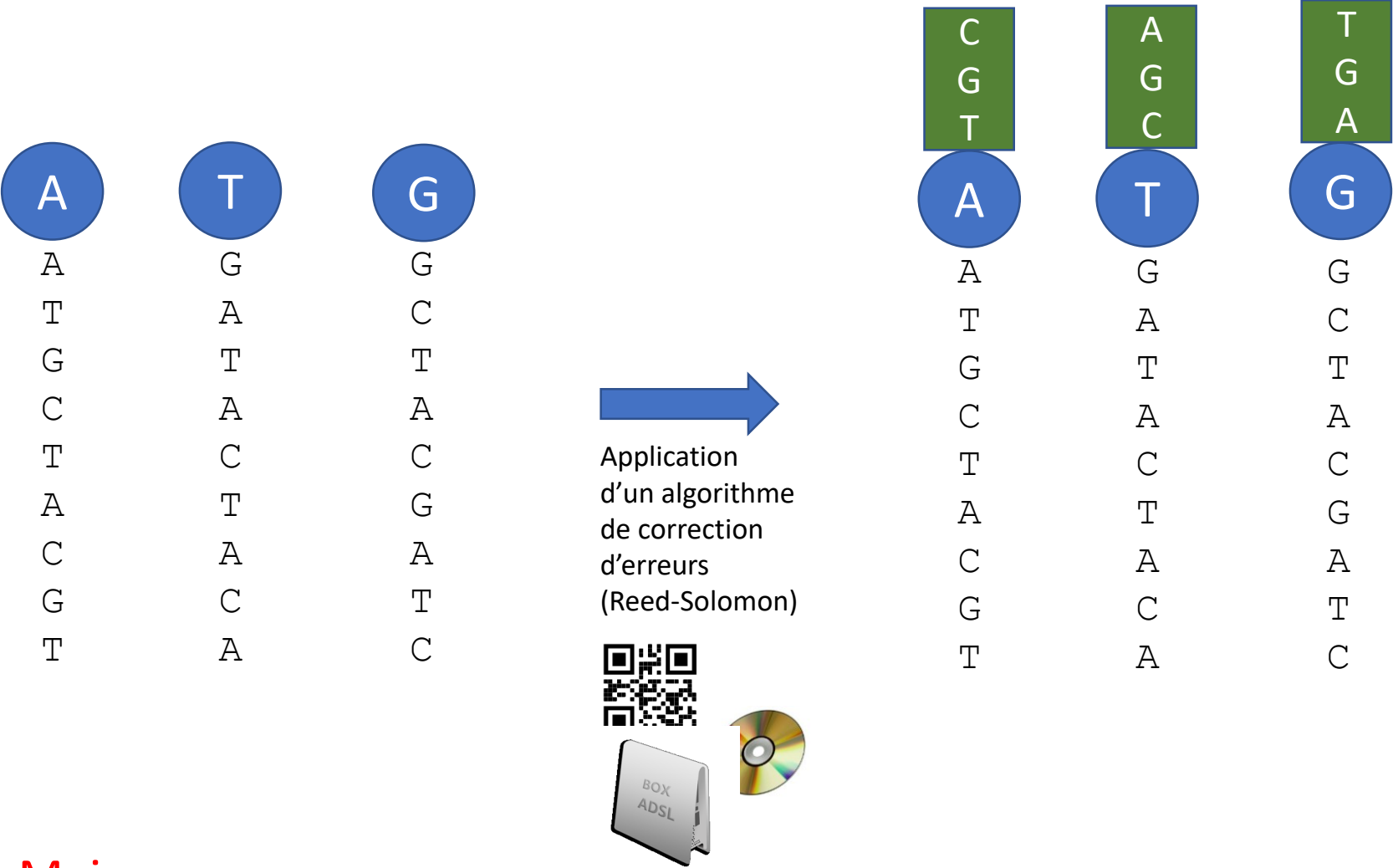
Encodage : idée 2



Mais:

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

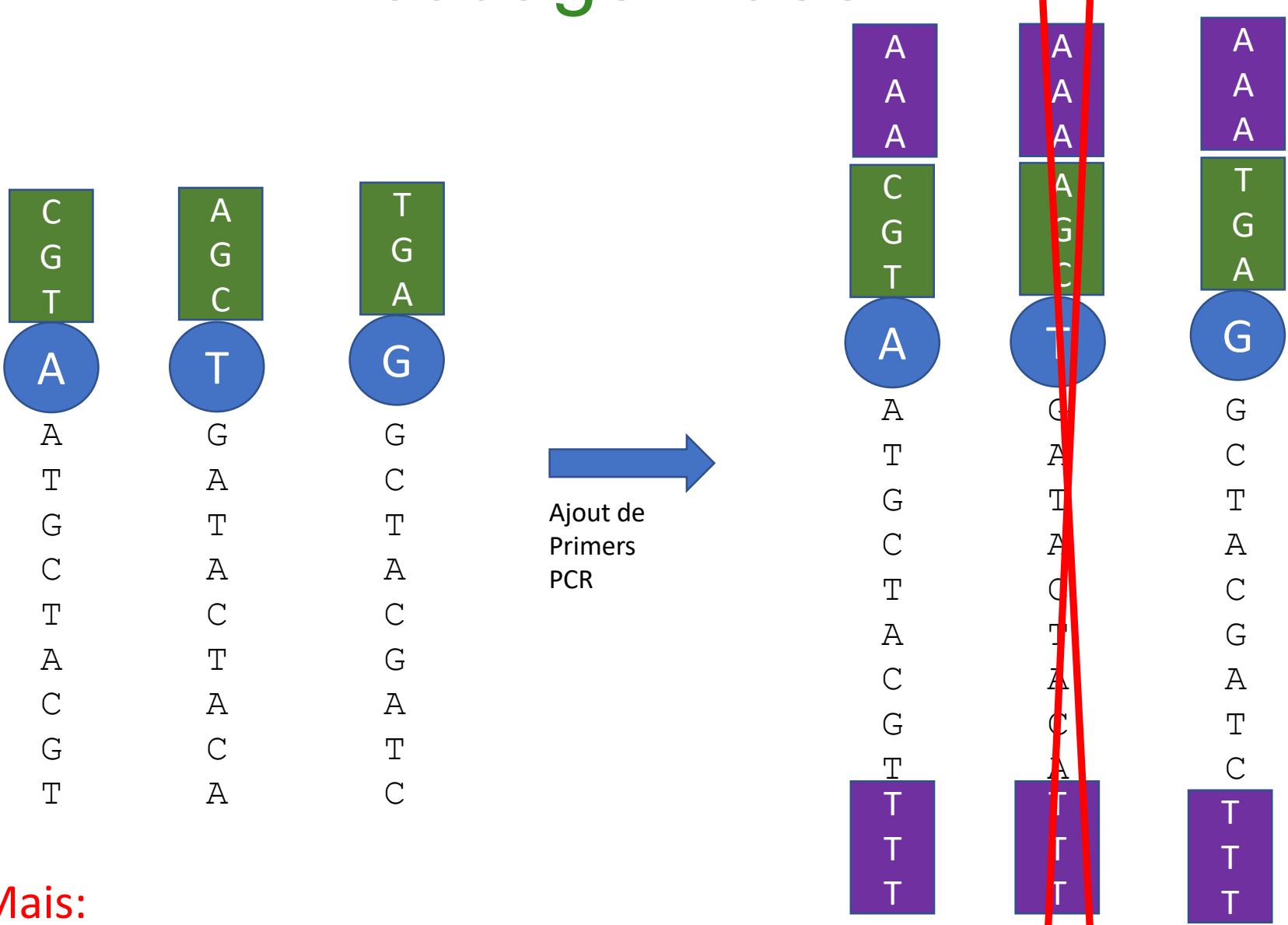
Encodage : idée 3



Mais:

- Comment différencier les paquets d'information?

Encodage : idée 4



Mais:

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

Démo

```
Anaconda Prompt (anaconda)
(base) C:\Users\m6qopo\Git\archive2dna>python cli.py encode aip_olos2.zip dna.txt
2022-05-30 13:40:35,895 - INFO - start : load binary
2022-05-30 13:40:36,067 - INFO - start : add outer code
2022-05-30 13:40:36,546 - INFO - start : add index
2022-05-30 13:40:36,583 - INFO - start : add inner code
2022-05-30 13:40:36,784 - INFO - start : convert representation to DNA segment
2022-05-30 13:40:36,931 - INFO - start : add primers
2022-05-30 13:40:36,931 - INFO - start : write DNA
{'binary_data': {'blocks': '1', 'size_bytes': '20881'},
 'capacity': {'block_capacity_megabytes': '2.29362',
              'information_density': '1.6871428638504267',
              'max_segments_block': '114681',
              'max_segments_index': '16777215',
              'total_capacity_megabytes': '335.5443'},
 'corrections': {'inner': '0',
                 'outer': '0',
                 'segments_beyond_repair': '0',
                 'segments_lost': '0'},
 'dna_segments': {'count': '873', 'size_median': '208', 'size_min': '188'},
 'errors': {'error': 'False', 'message': ''},
 'id': {'package_id': 'None', 'package_primer': 'None'},
 'parameters': {'I': '4',
                'I1': '3',
                'I2': '1',
                'K': '44',
                'N': '52',
                'k': '16333',
                'mi': '8',
                'mo': '14',
                'n': '16383',
                'necsi': '8',
                'necso': '50'},
 'redundancy': {'inner': '0.15384615384615385', 'outer': '0.4009163802978236'}}
(base) C:\Users\m6qopo\Git\archive2dna>
```

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