Data Storage on DNA

Dominique Lavenier

GenScale

Univ. Rennes, IRISA/CNRS, Inria

CominLabs

UMR IRISA

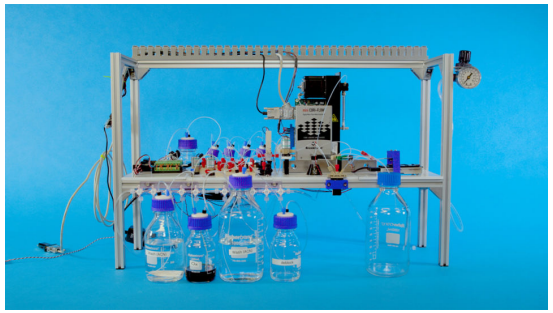
Inria

Agenda

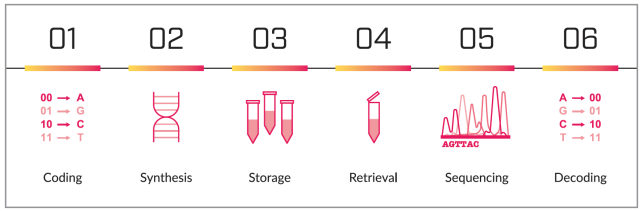
1. Introduction
   - Why new storage medium ?
   - Why DNA ?

2. DNA storage principle
   - How does it work ?
   - What are the main challenges ?

3. dnarXiv project
   - What's going on in Rennes ?



Source: https://news.microsoft.com/innovation-stories/hello-data-dna-storage/



Source: An Introduction to DNA Data Storage, DNA storage alliance, June 21

dnarXiv

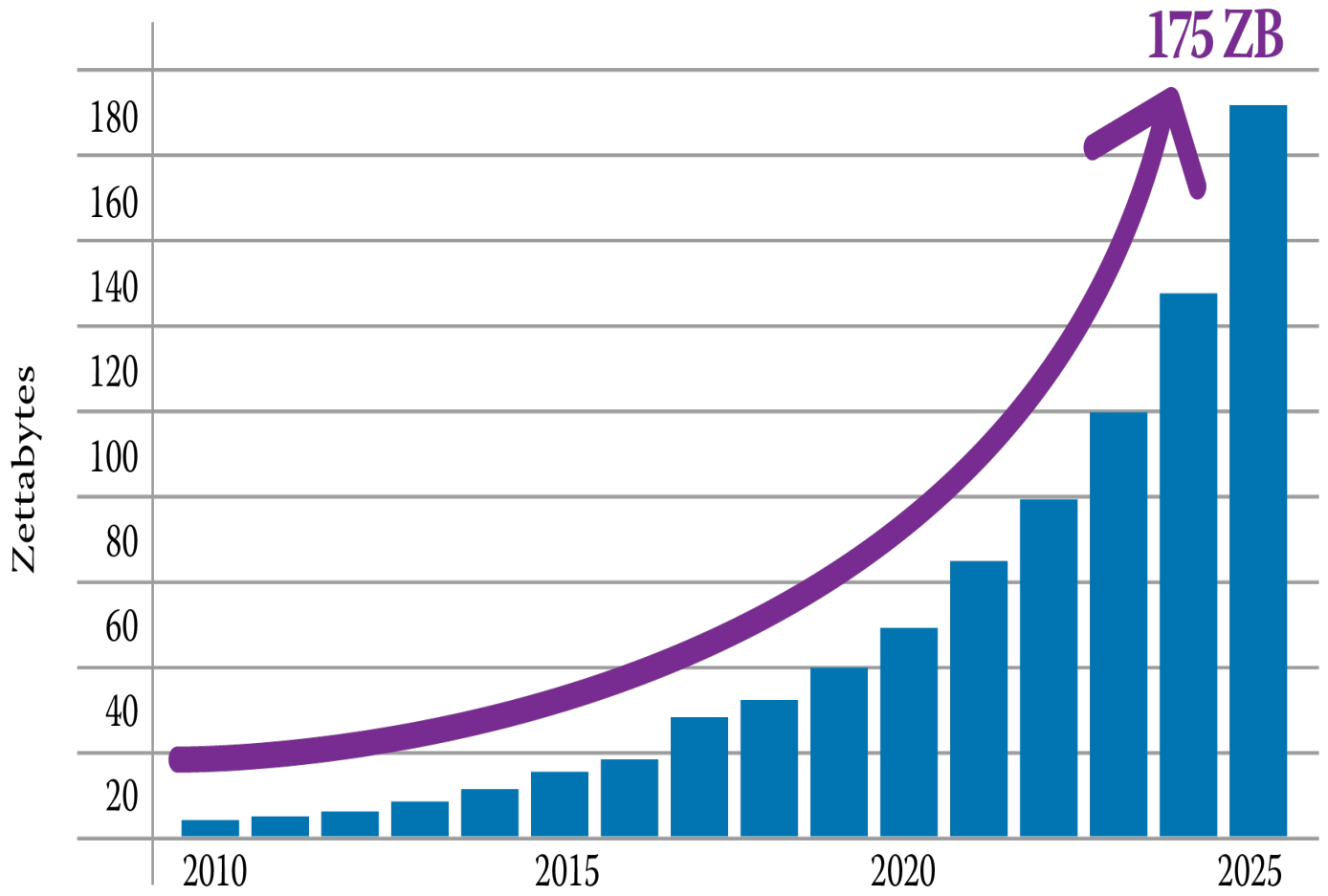
Exponential Data Growth

According to IDC*
- annual data growth rate:
  → 23% [2020-2025]
- should reach 175 ZB in 2025

1 Zetta Bytes (ZB) = 10^{21} Bytes
- *1 billion of TB*

IoT devices are expected to create over 90 ZB of data in 2025

IDC: International Data Corporation



Source: IDC White paper, Data Age 2025, nov. 2018, The Digitization of the world, From Edge to Core

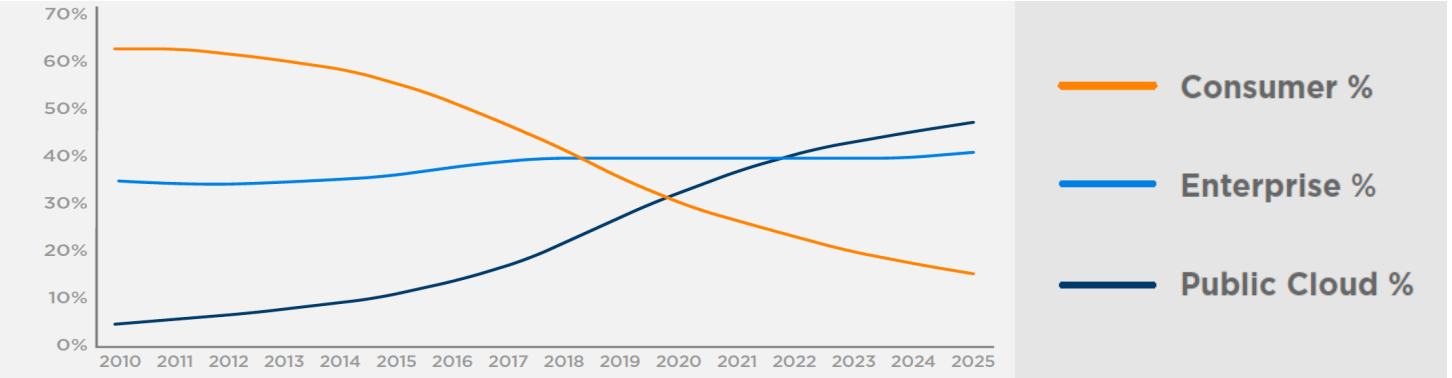
Where the data will be stored ?

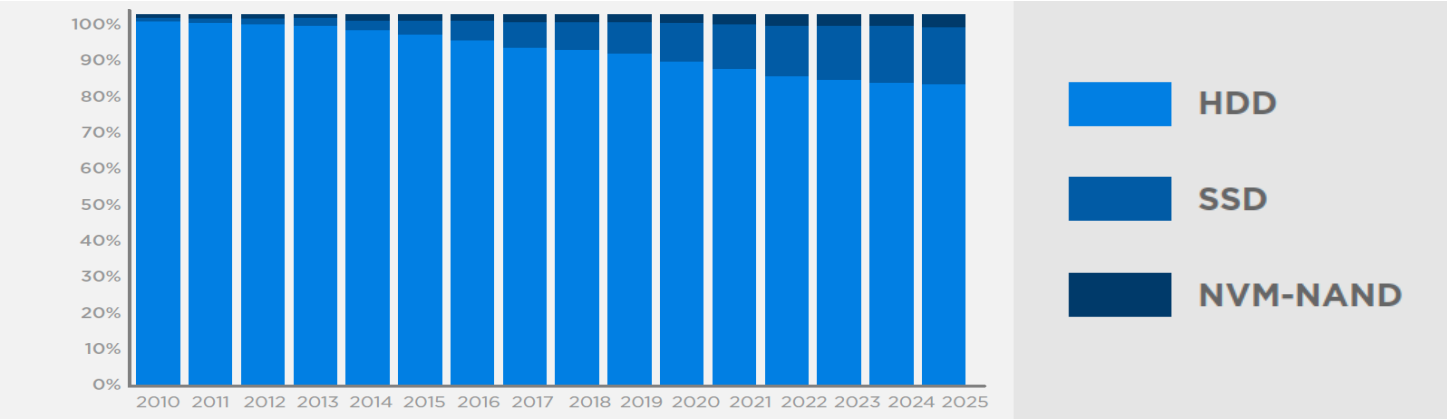
By 2025:

- need to store 22 ZB of data

- 85% of all data worldwide will reside in public or private cloud environment

- 80% of cloud data will be stored on HDD





Source: IDC White paper, Data Age 2025, nov. 2018, The Digitization of the world, From Edge to Core

Challenges for archival storage technologies

Storage maintenance and replacement costs

- HDD / SSD devices life span : 10 years (max)

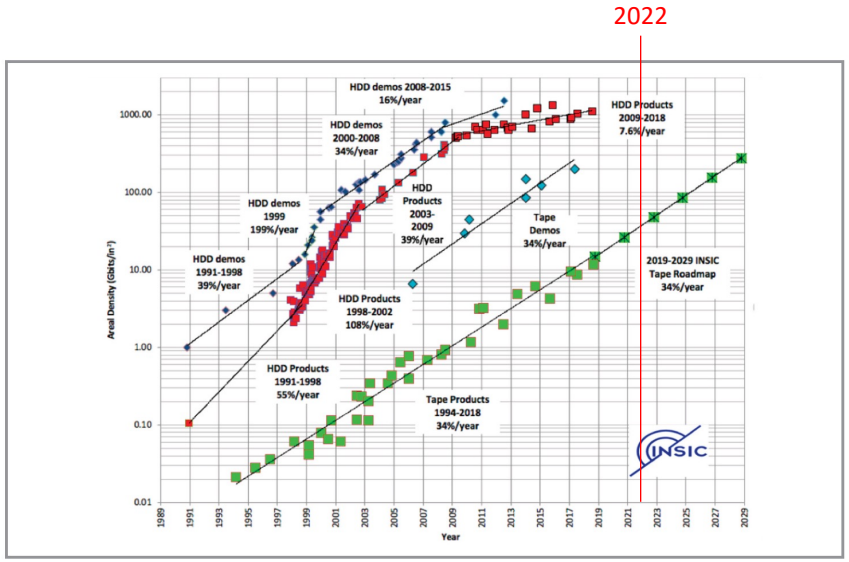
Device storage have to be periodically changed

Density limitations

- Areal density for magnetic media is slowing

Energy and sustainability concerns

- Increasing electric consumption of data centers
- HDD relies on rare earth metals



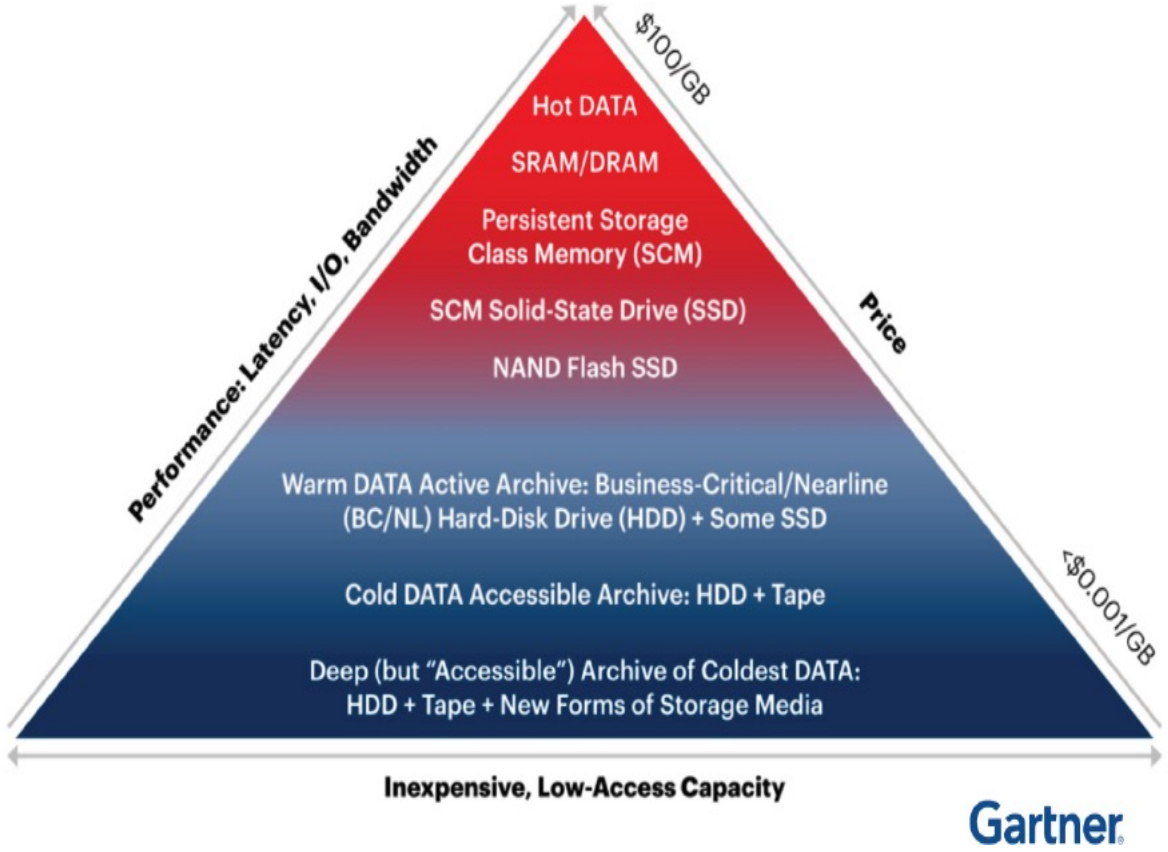
Storage hierarchy

HOT DATA
- data accessed frequently
- storage: high performance devices: SSD

WARM DATA
- data accessed less frequently
- storage: HDD

COLD DATA
- data accessed unfrequently
- storage: tapes

COLDEST DATA
- data never accessed (or with low probability)
- storage: tapes or new storage media



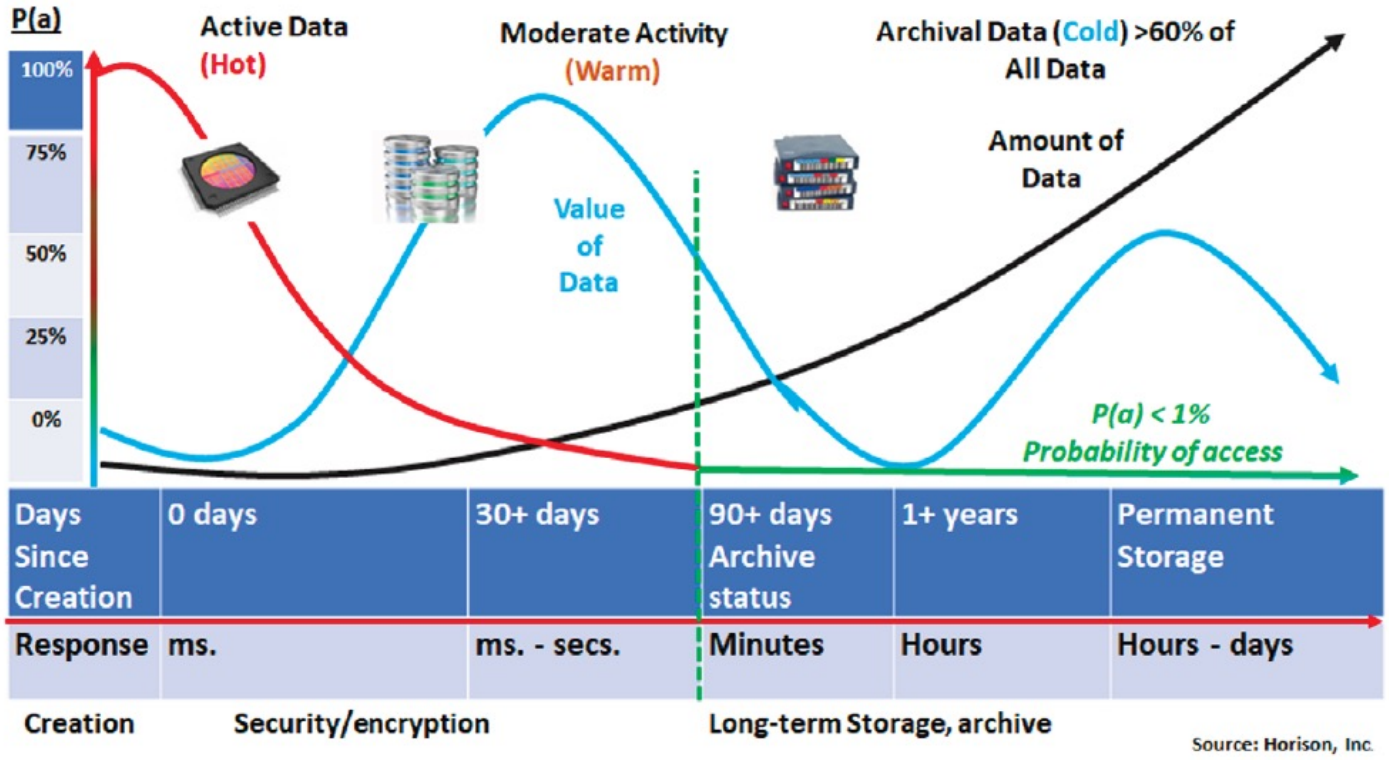
Source: Source: Gartner Product Manager Insight: Tape to the Future? Oct. 2020

Data life span

Most data become archival after 90 days

Over 60% of all data is archival

Retention of 50-100+ years is common



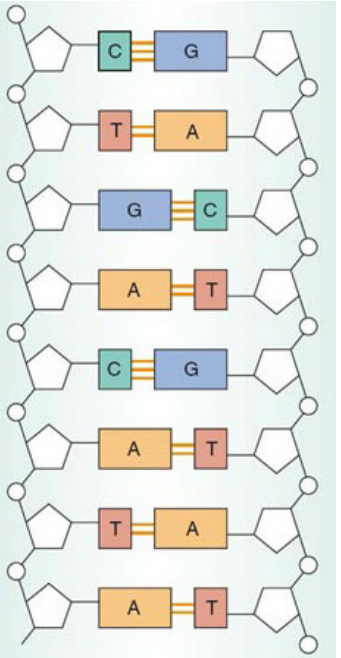
Source: Horison, Inc

DNA as storage media

Advantages

- density
- durability
- medium immutability
- low TCO (Total Cost of Ownership)
- Energy efficiency and sustainability

DNA molecule



Genome size

Virus	3 x 10^4 bp
Bacteria	3 x 10^6 bp
Human	3 x 10^9 bp
Salamander	30 x 10^9 bp
Paris Japonica	150 x 10^9 bp

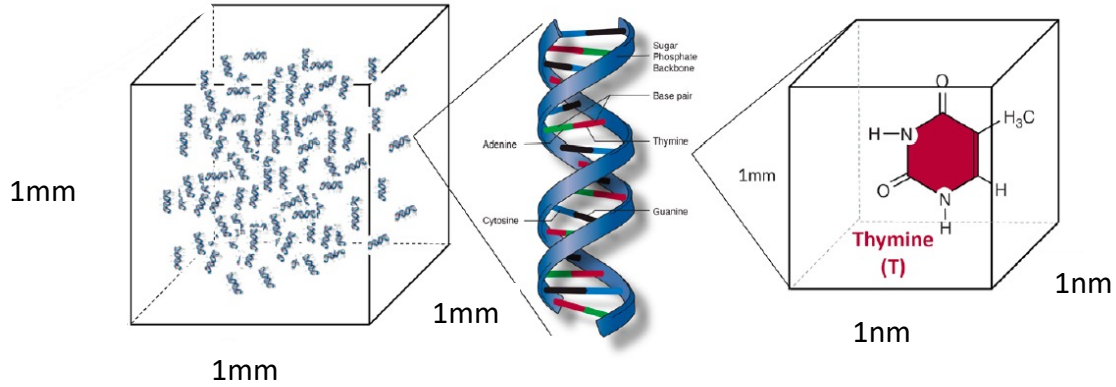
```
4 bases: A C G T
         | | | | base pairs (bp)
         T G C A
```

DNA as storage medium

Advantages

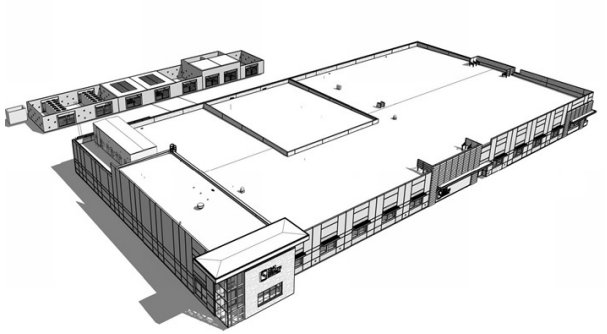
- density
- durability
- medium immutability
- low TCO (Total Cost of Ownership)
- Energy efficiency and sustainability

3D Storage



1 mm^3 → 9TBytes

1 DNA base ~ 10 atoms



Data center: 1 ExaBytes, 100 000 m^2, 10M Watts

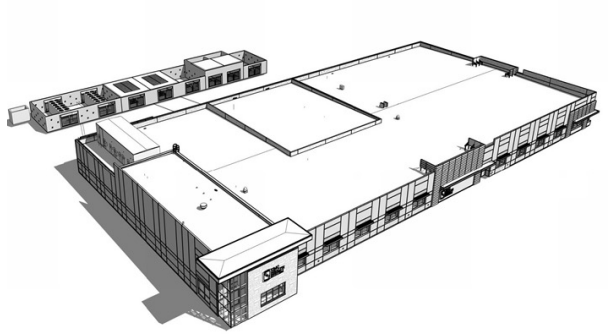
DNA storage

5cm X 5cm X 5cm

DNA as storage medium

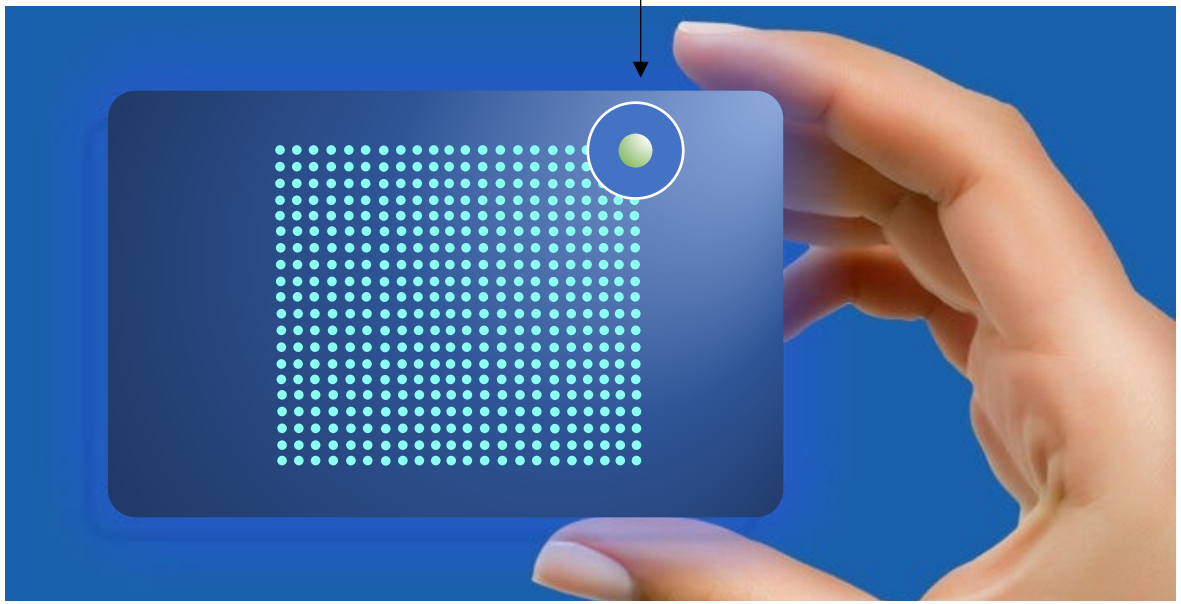
Advantages

- density
- durability
- medium immutability
- low TCO (Total Cost of Ownership)
- Energy efficiency and sustainability



1 ExaBytes of cold data → 2,500 credit cards

1 mm dot → 1 TB of data



400 hard drives on a credit card

DNA as storage medium

Advantages

- density
- durability
- medium immutability
- low TCO (Total Cost of Ownership)
- Energy efficiency and sustainability

Stable at room temperature for a very long period

No need to replicate over time

Stockholm University

CPG Centre for Palaeogenetics


nature

Explore content | About the journal | Publish with us | Subscribe

nature > articles > article

Article | Published: 17 February 2021

Million-year-old DNA sheds light on the genomic history of mammoths

DNA as storage medium

Advantages

- density
- durability
- medium immutability
- low TCO (Total Cost of Ownership)
- Energy efficiency and sustainability

There will always be DNA readers

In the future, DNA readers will be smaller, cheaper, faster

How to read a floppy disk today ?

DNA as storage medium

Advantages

- density
- durability
- medium immutability
- low TCO (Total Cost of Ownership)
- Energy efficiency and sustainability

TCO: the money put on the resources to preserve data for a certain amount of years

No need to periodically replace SSD, HDD, tapes, …

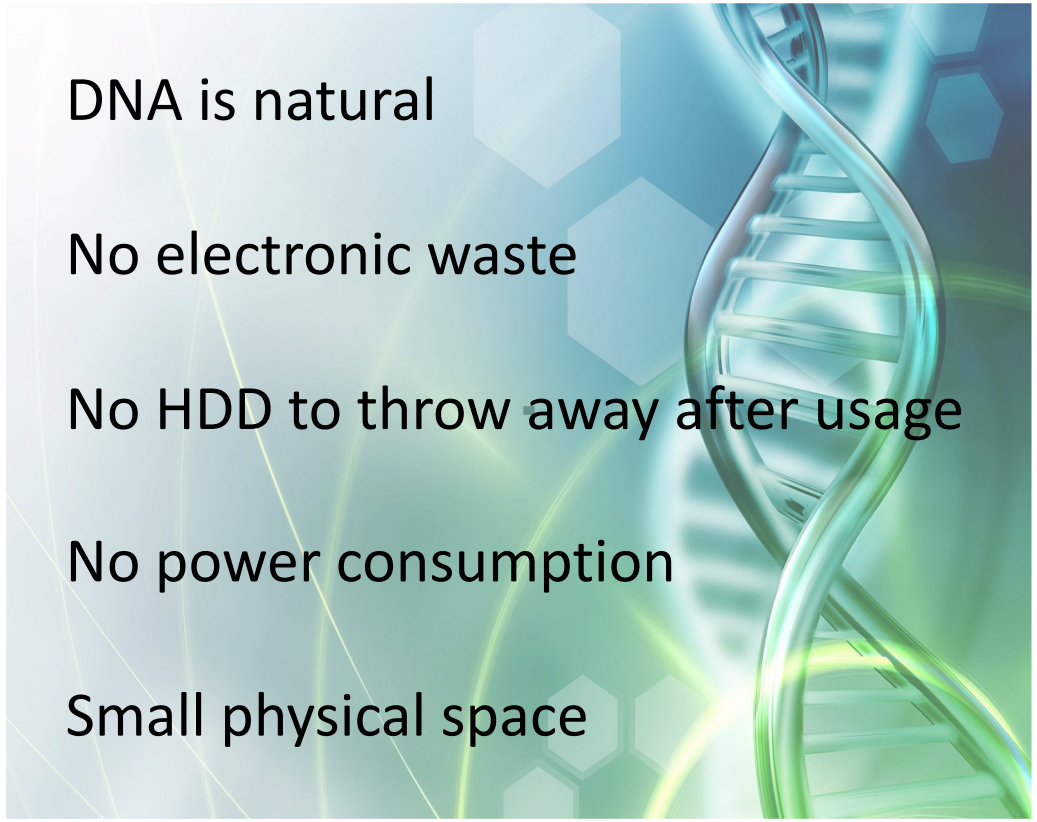
Low electric consumption
- no electronic devices
- no air conditioning

Cheap data replication

DNA as storage medium

Advantages

- density
- durability
- medium immutability
- low TCO (Total Cost of Ownership)
- Energy efficiency and sustainability

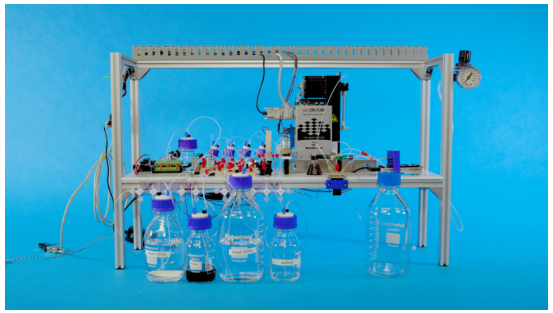


Agenda

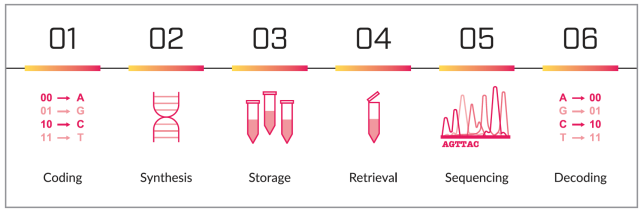
1. Introduction
   - Why new storage medium ?
   - Why DNA ?

2. DNA storage principle
   - How does it work ?
   - What are the main challenges ?

3. dnarXiv project
   - What's going on in Rennes ?



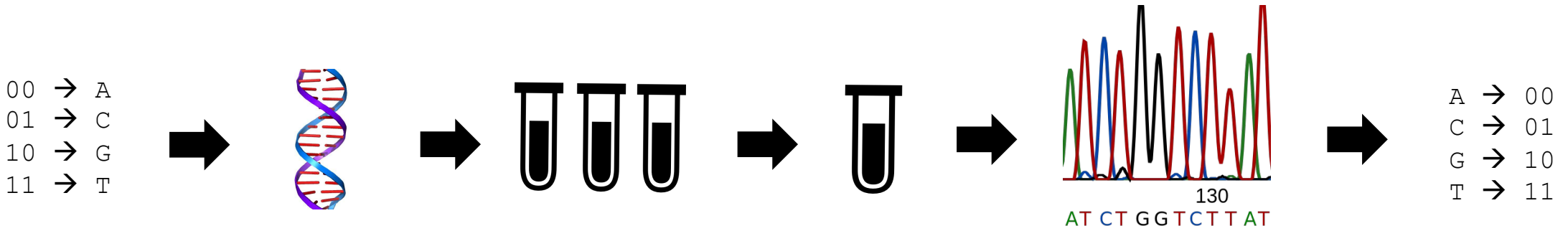
Source: https://news.microsoft.com/innovation-stories/hello-data-dna-storage/



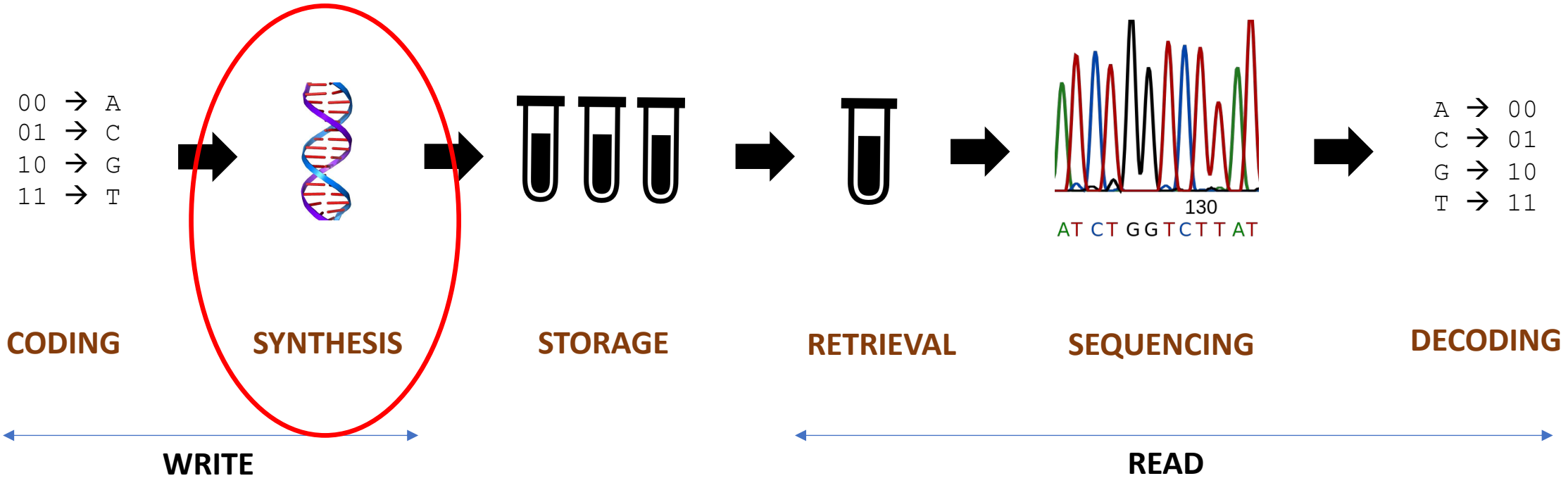
Source: An Introduction to DNA Data Storage, DNA storage alliance, June 21

dnarXiv

Principle of DNA data storage



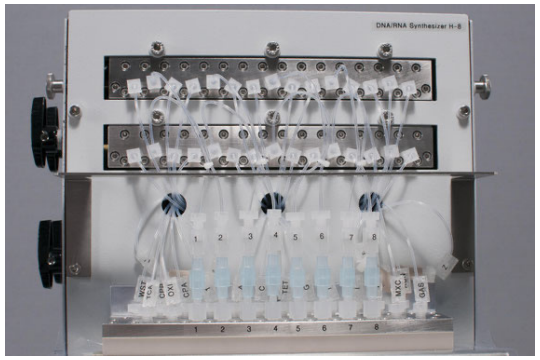
Principle of DNA data storage



Synthesis

ATGGACCGGTGACACCCGGTTGGACCGTGA

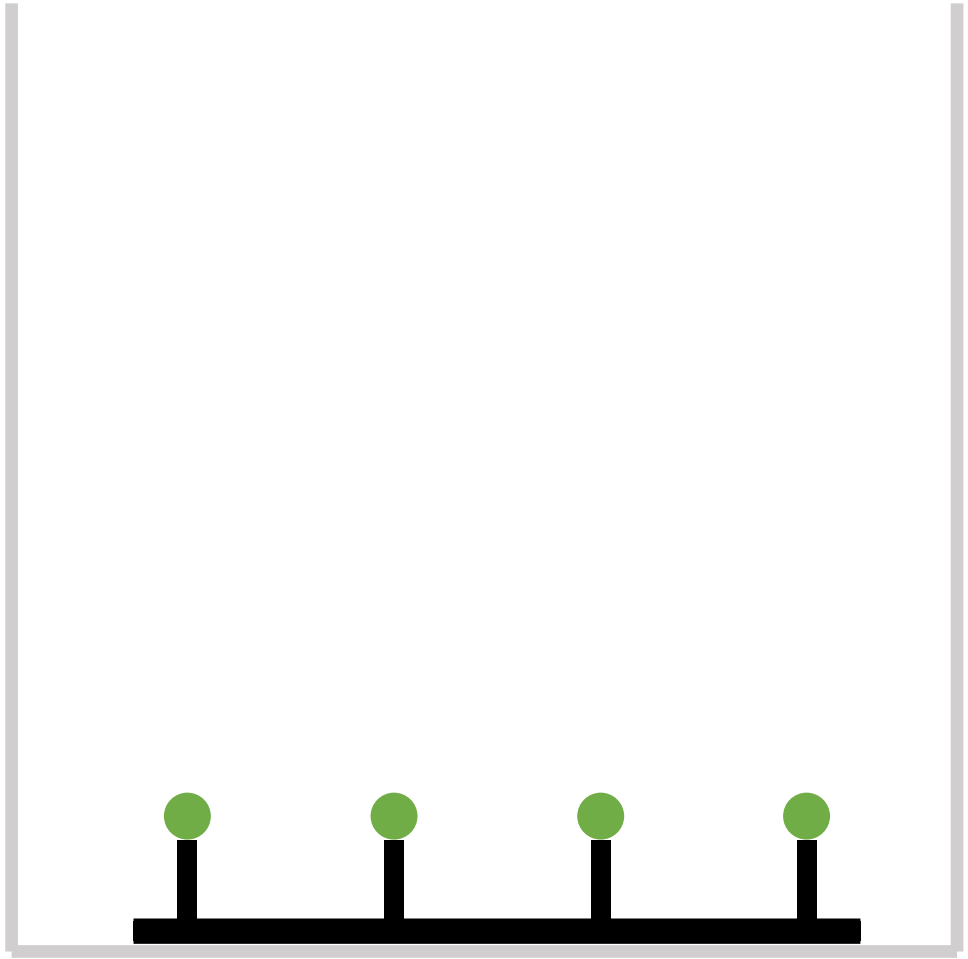
1 text sequence



DNA
Synthesizer

Million/Billions of molecules

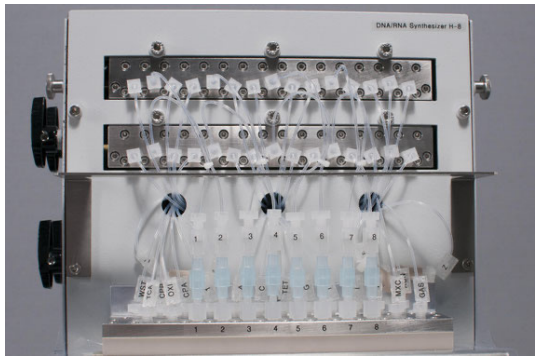
A G T C G T A C . . .



Synthesis

ATGGACCGGTGACACCCGGTTGGACCGTGA

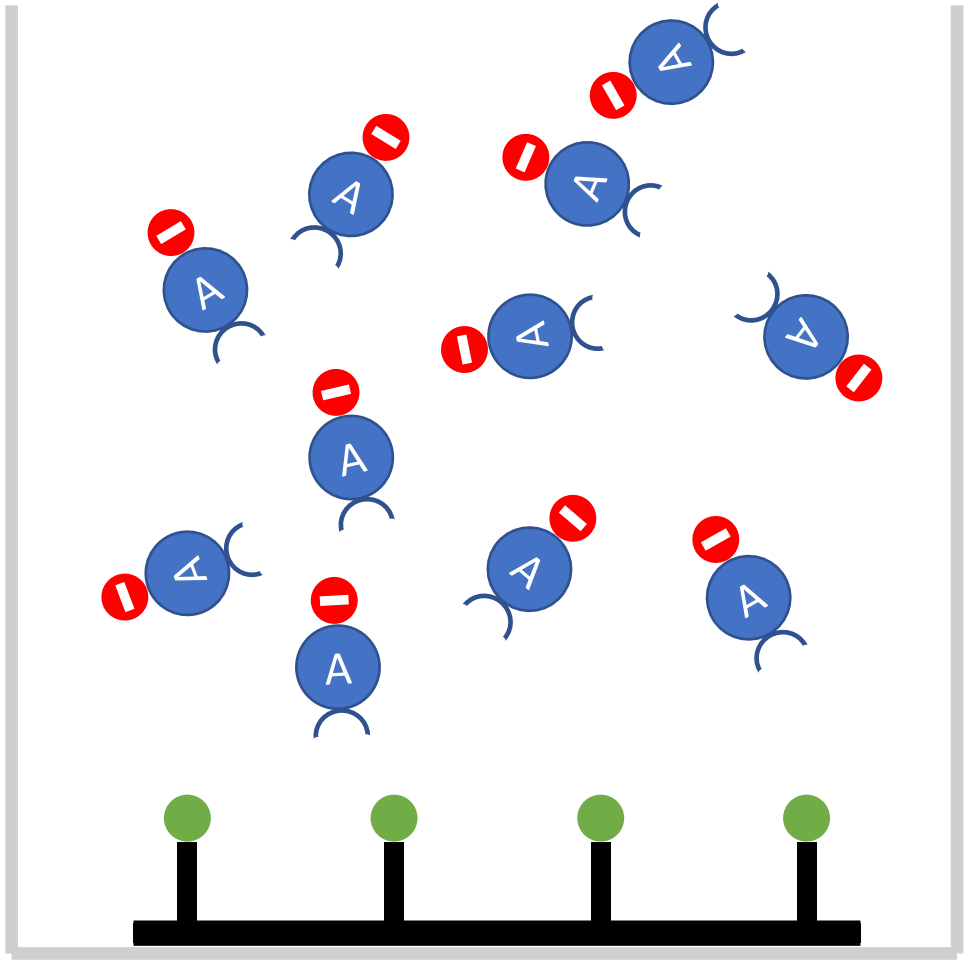
1 text sequence



DNA Synthesizer

Million/Billions of molecules

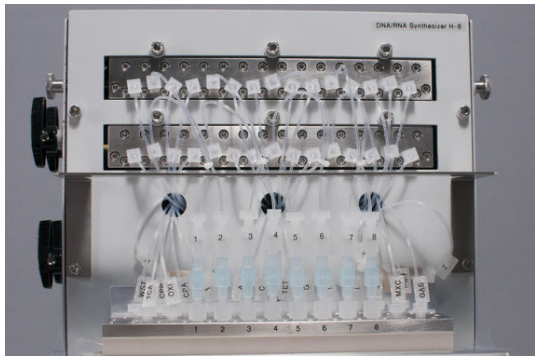
A G T C G T A C . . .



Synthesis

ATGGACCGGTGACACCCGGTTGGACCGTGA

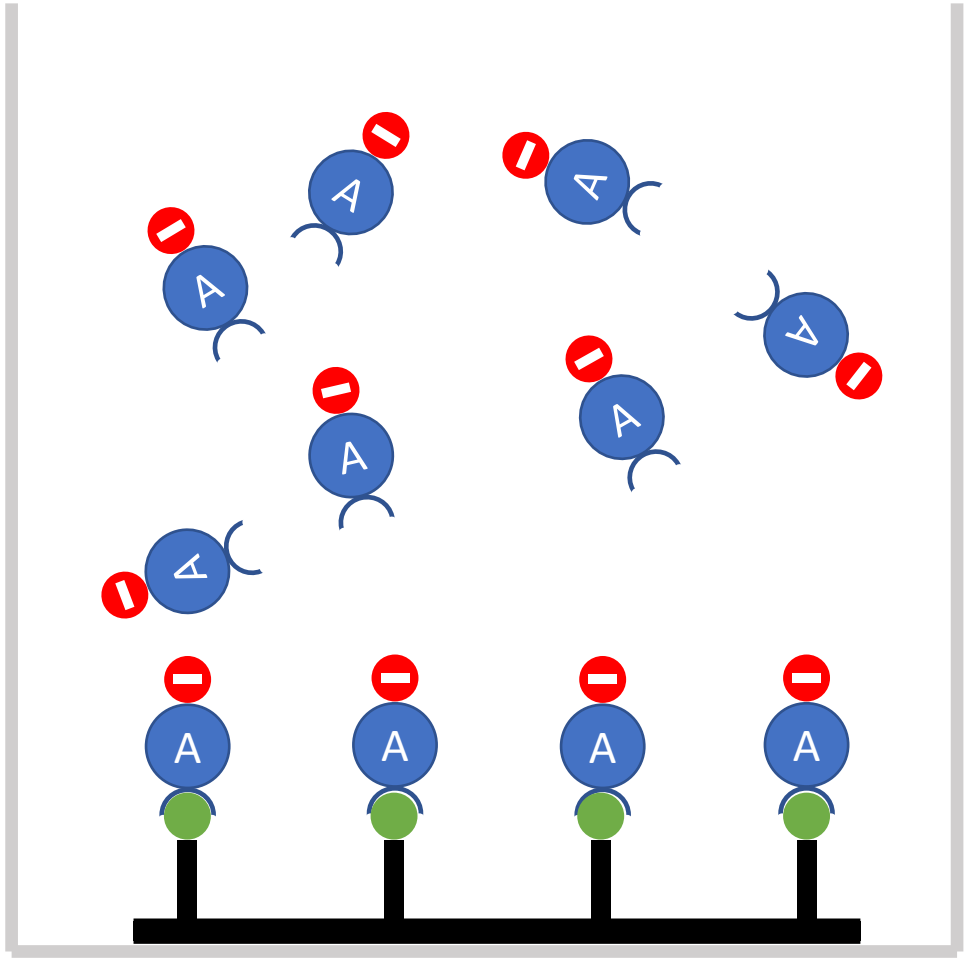
1 text sequence



DNA
Synthesizer

Million/Billions of molecules

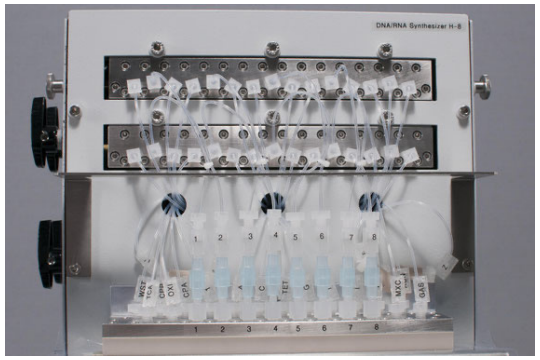
A G T C G T A C . . .



Synthesis

ATGGACCGGTGACACCCGGTTGGACCGTGA

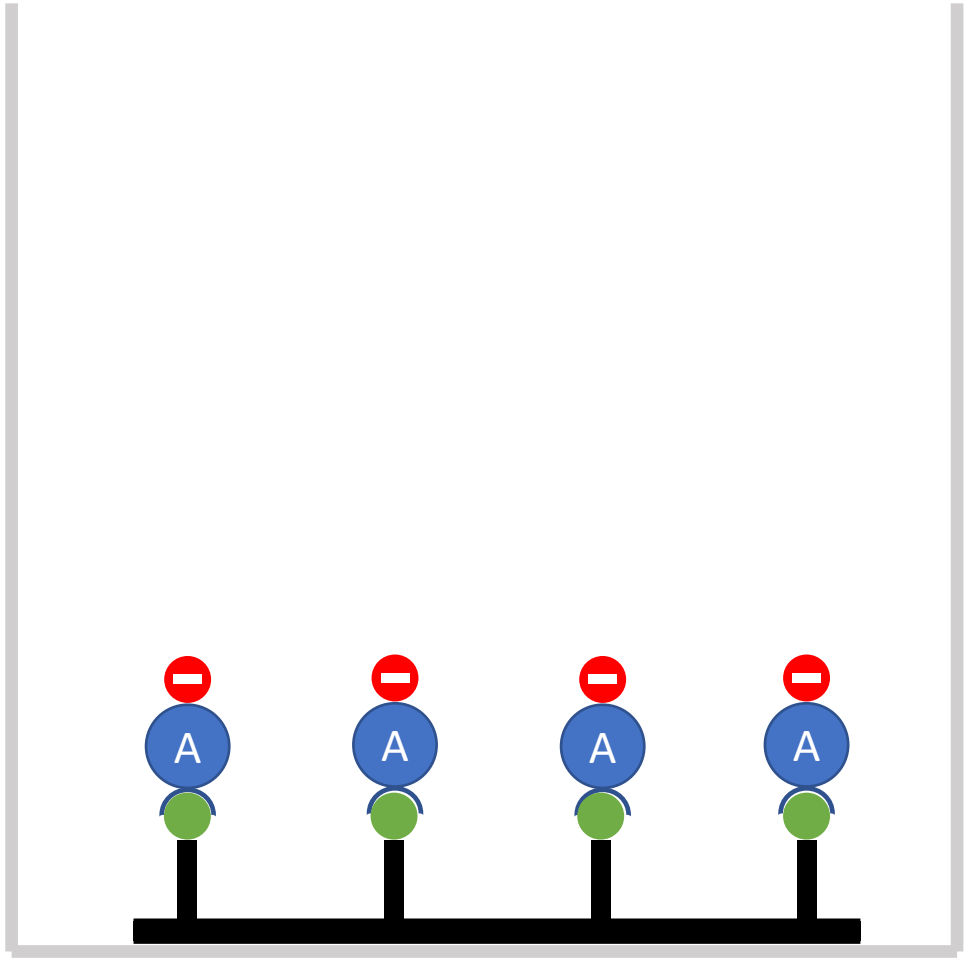
1 text sequence



DNA
Synthesizer

Million/Billions of molecules

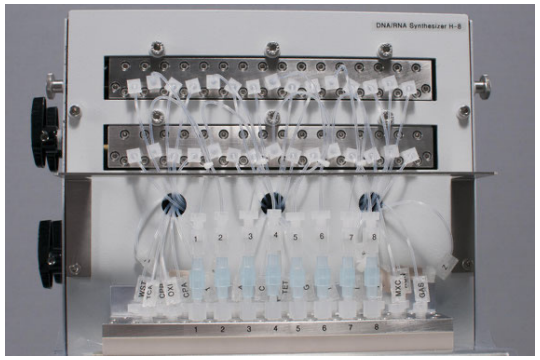
A G T C G T A C . . .



Synthesis

ATGGACCGGTGACACCCGGTTGGACCGTGA

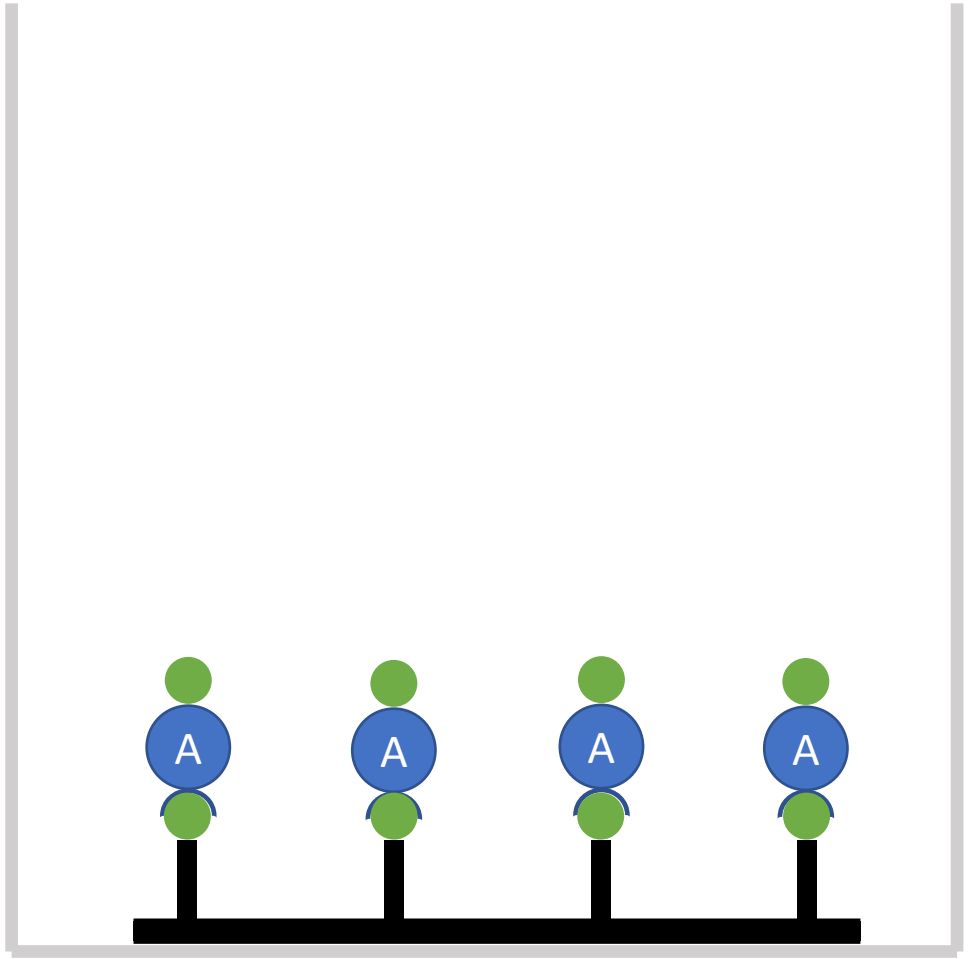
1 text sequence



DNA
Synthesizer

Million/Billions of molecules

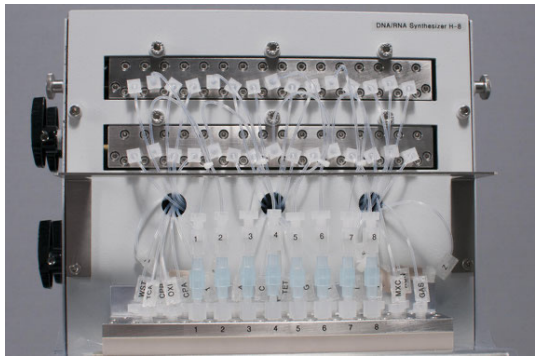
A G T C G T A C . . .



Synthesis

ATGGACCGGTGACACCCGGTTGGACCGTGA

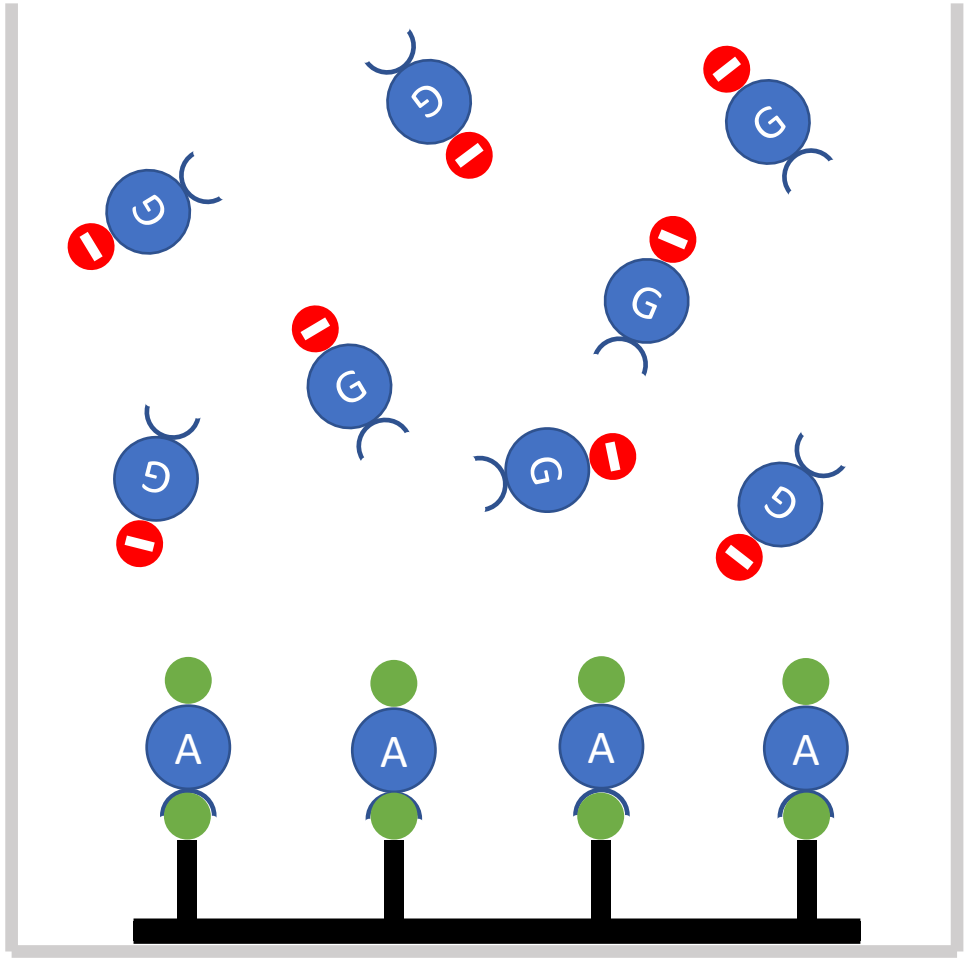
1 text sequence



DNA
Synthesizer

Million/Billions of molecules

A G T C G T A C . . .



Synthesis limitation

Synthesizers can only generate short molecules (oligonucleotides)

- oligonucleotide length < 200 nucleotides

Very slow process

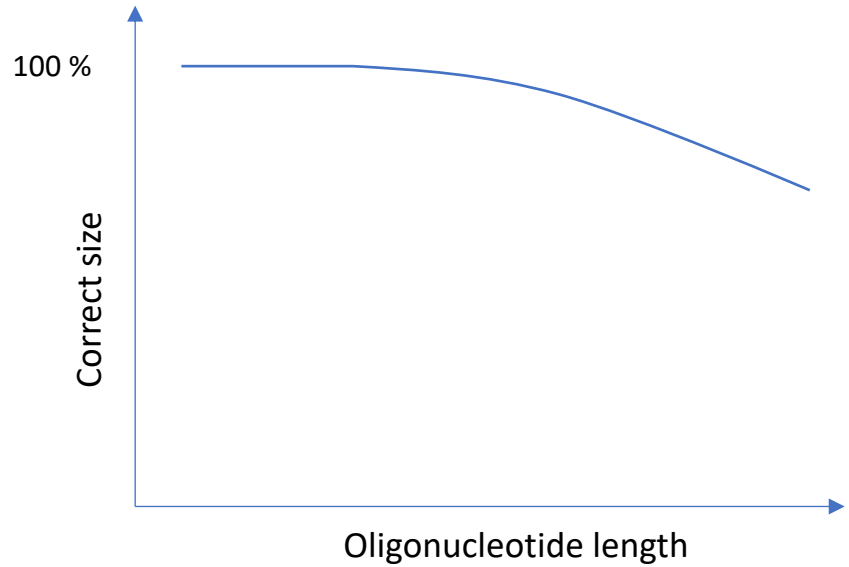
- 1 nucleotide / minute

High cost

- 100 € / Kbyte

Errors

- variation on oligonucleotide length



Synthesis perspectives

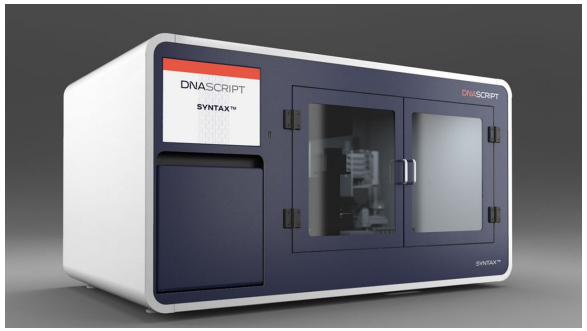
New technologies

- Enzymatic DNA synthesis
  - exploit the capabilities of the TdT enzyme
  - longer oligonucleotides
  - faster

2,5 pt

DNA SCRIPT

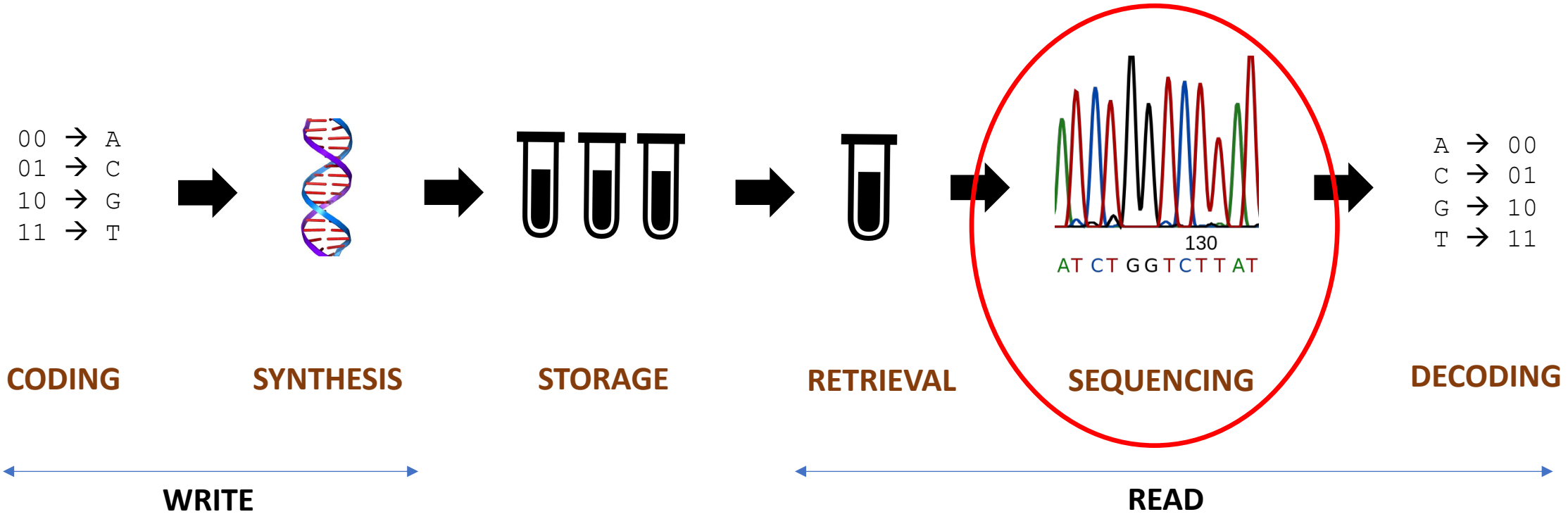
NextGen DNA Synthesis



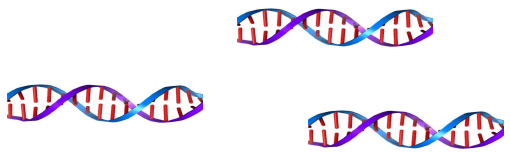
Miniaturization & Parallelization

- Roadmap 2025
  - 1 Tbytes
  - 1 day
  - 1000 $

Principle of DNA data storage



Sequencing



DNA molecules

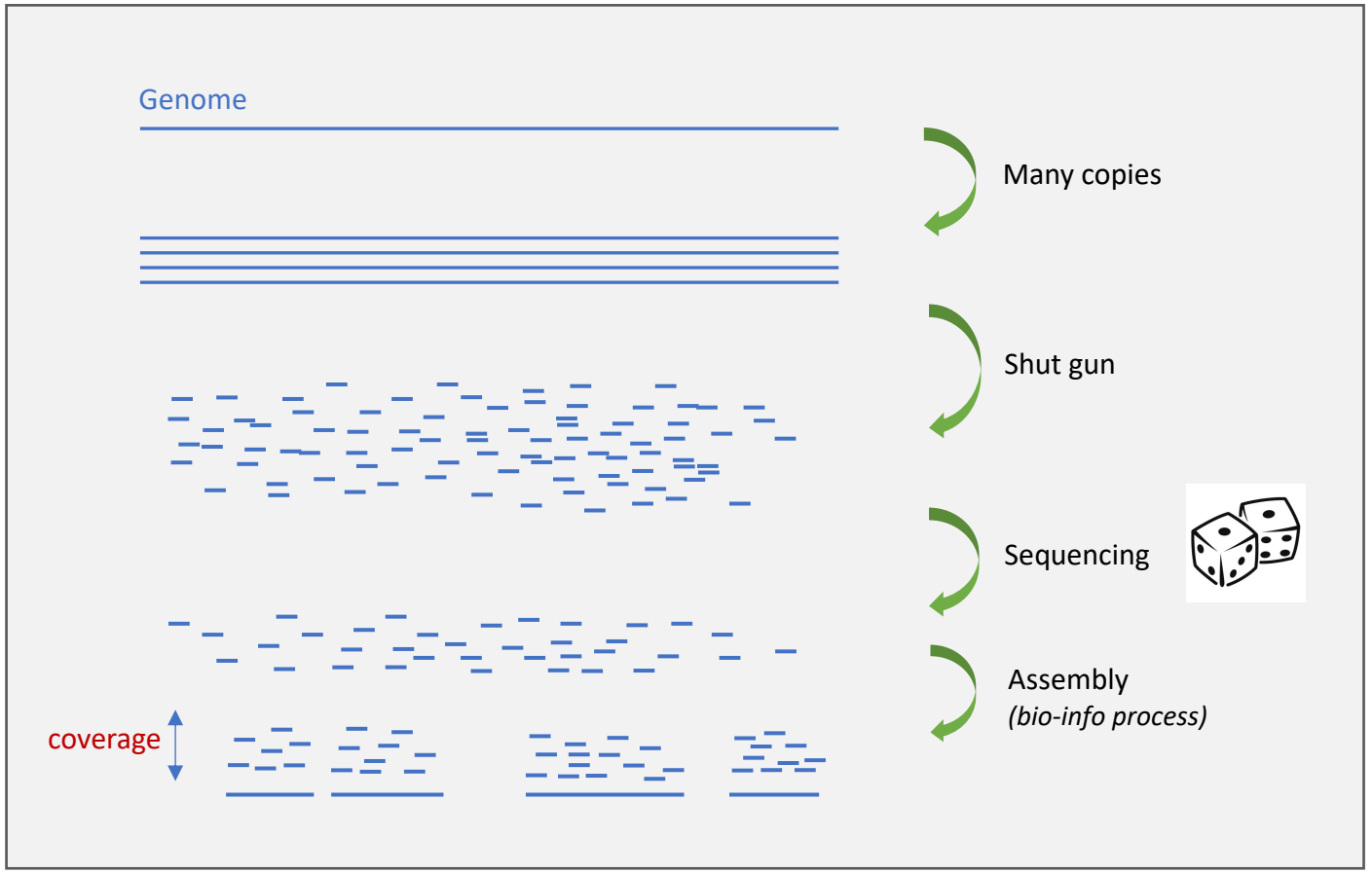
DNA
SEQUENCER

ATGGACCGGTGACACCCGGTTGGACCGTGA

TGGATAGGCCAGGTTAGGCCGGATATGAGC

Reads

Reading the text of a genome



Sequencing Technologies

Short reads (< 300 nt)

- High quality (error rate < 0.1 %)

Long reads (10^5 nt)

- low quality (error rate ~ 5 %)

Errors

Error types
1: substitution
2: deletion
3: insertion

Sequencing cost & throughput

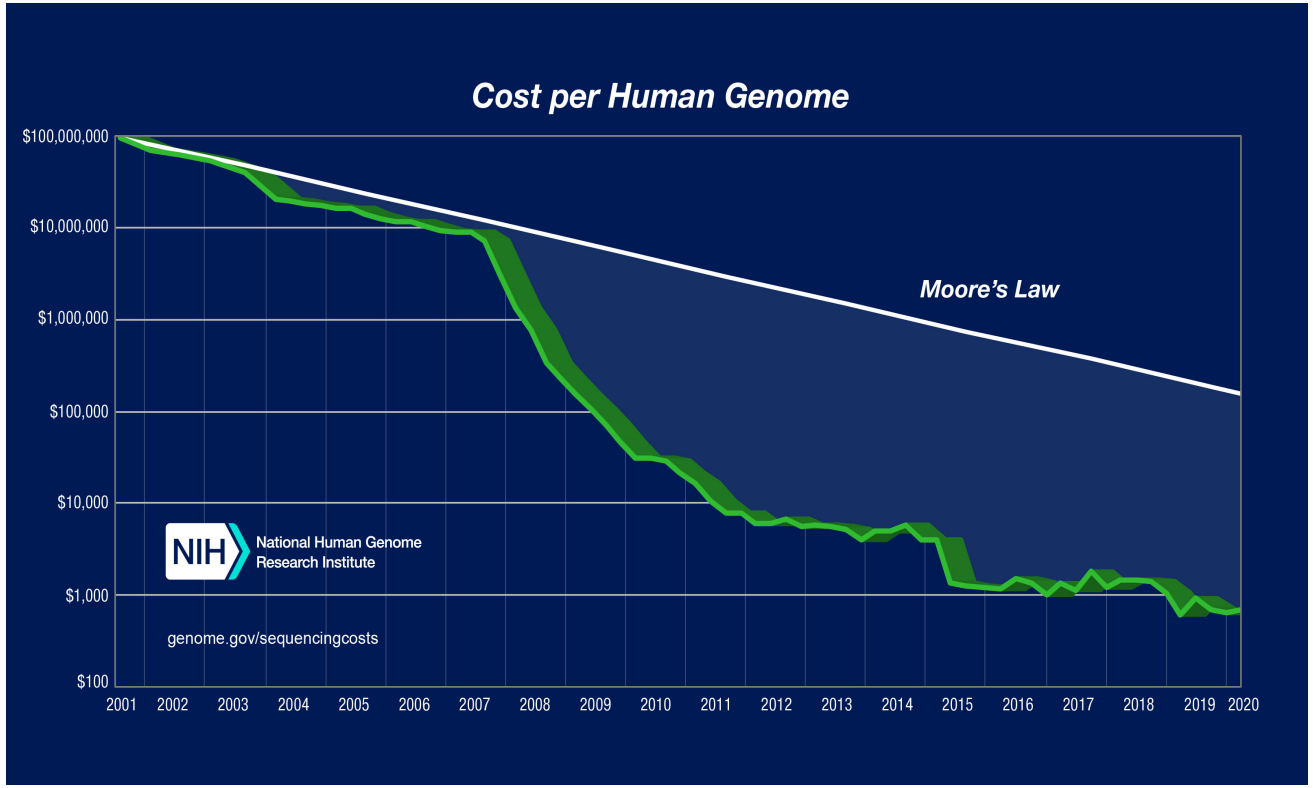
1 Human genome = 3.2 Gbp (~ 800 Mbytes)

Cost

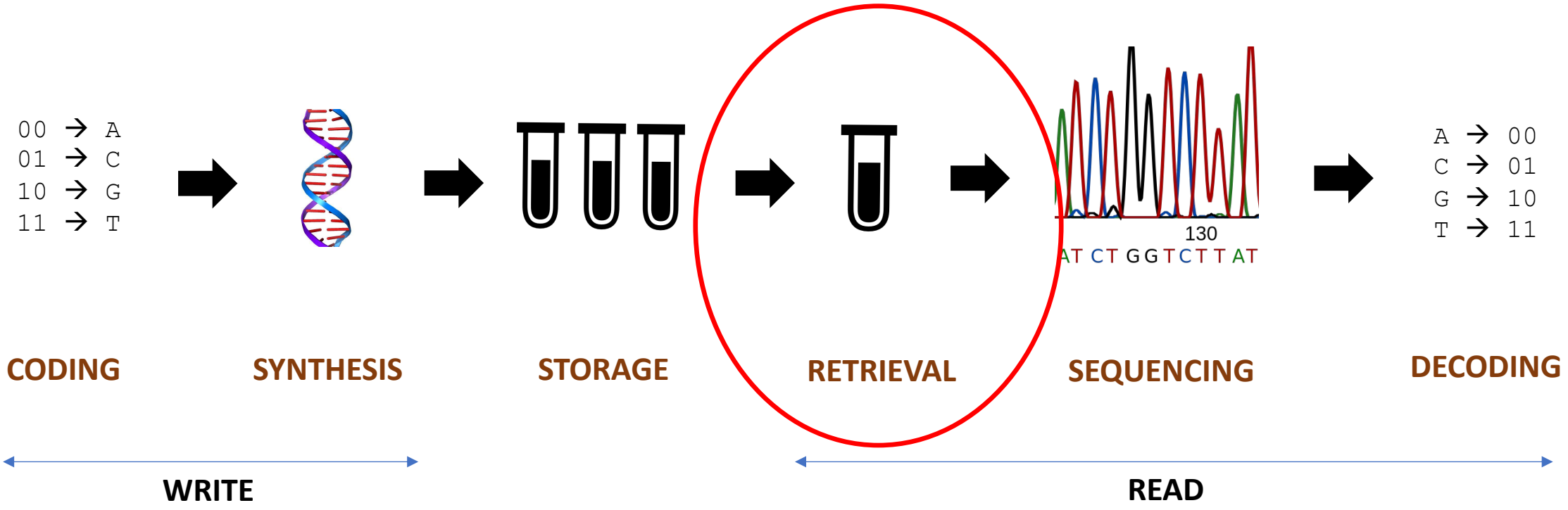
1 Human Genome < 1000 $
→ ~ 1 MBytes / $

Throughput

Illumina (NovaSeq 6000)
70 x 10^3 nucleotides / sec
ONT (Promethion, 48 flow cells)
20 x 10^3 nucleotides / sec



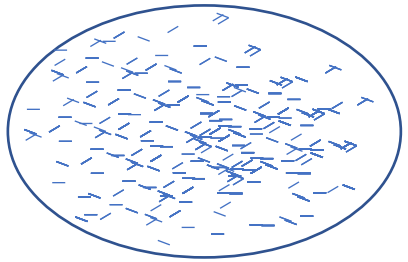
Principle of DNA data storage



Retrieval

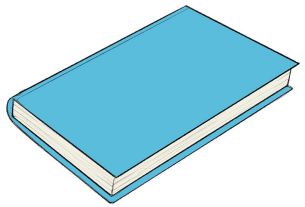
Encoding
+
Synthesis

container



Trillions of DNA molecules

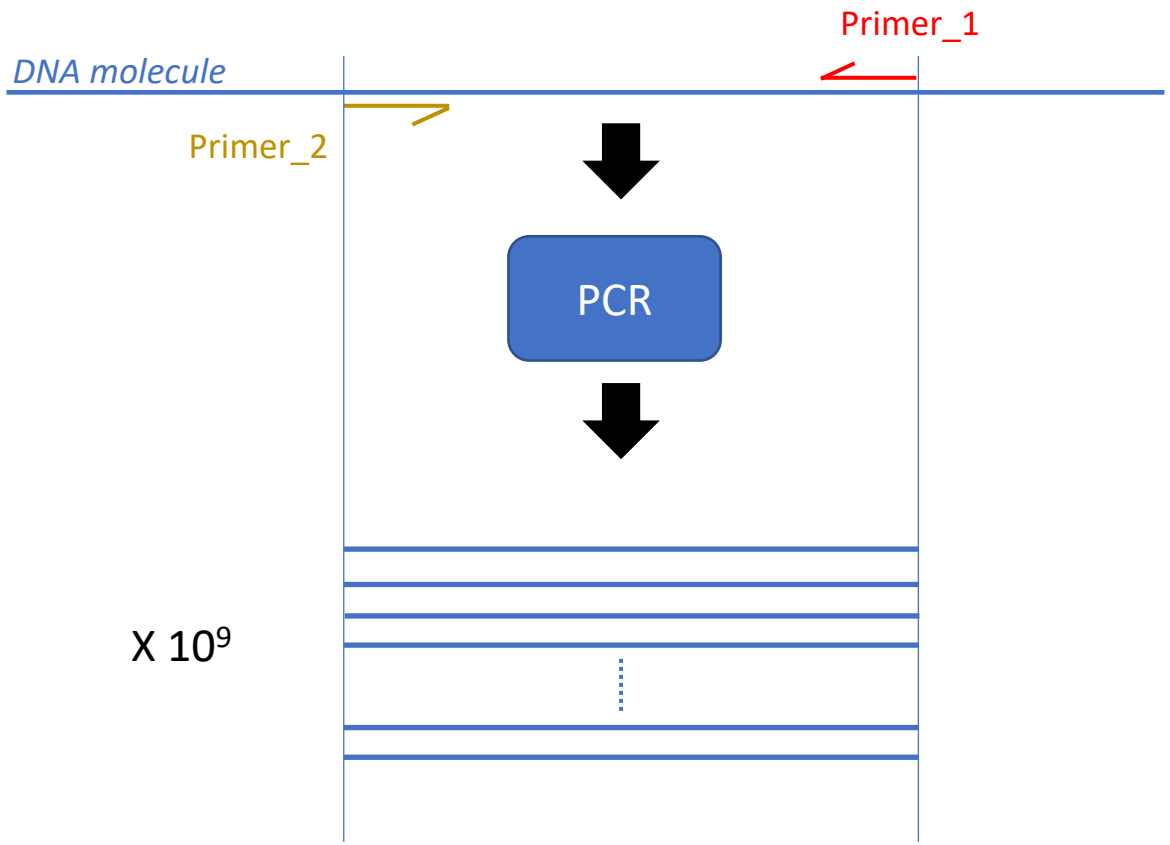
?



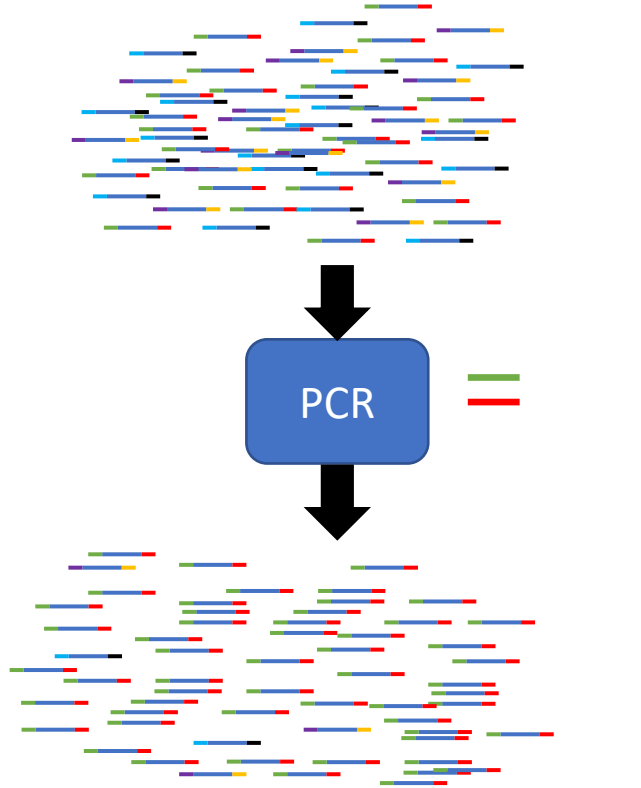
How avoiding to sequence all molecules to retrieve a specific document ?

Selection by PCR

PCR: amplification of DNA molecules

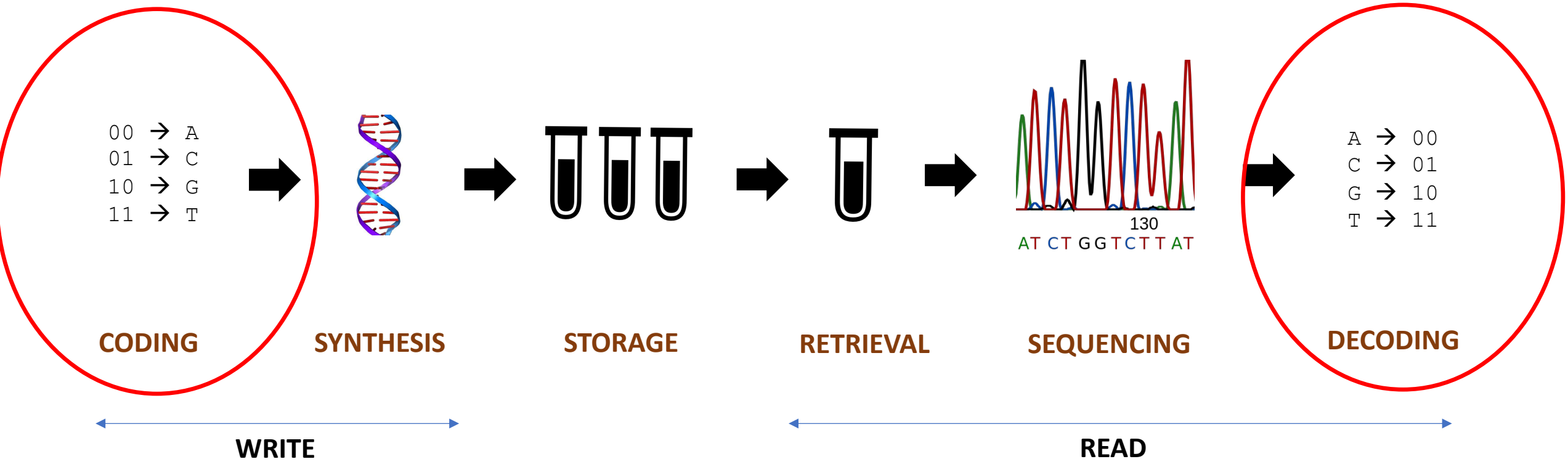


All DNA molecules from the same document will start and end by the same couple of primers

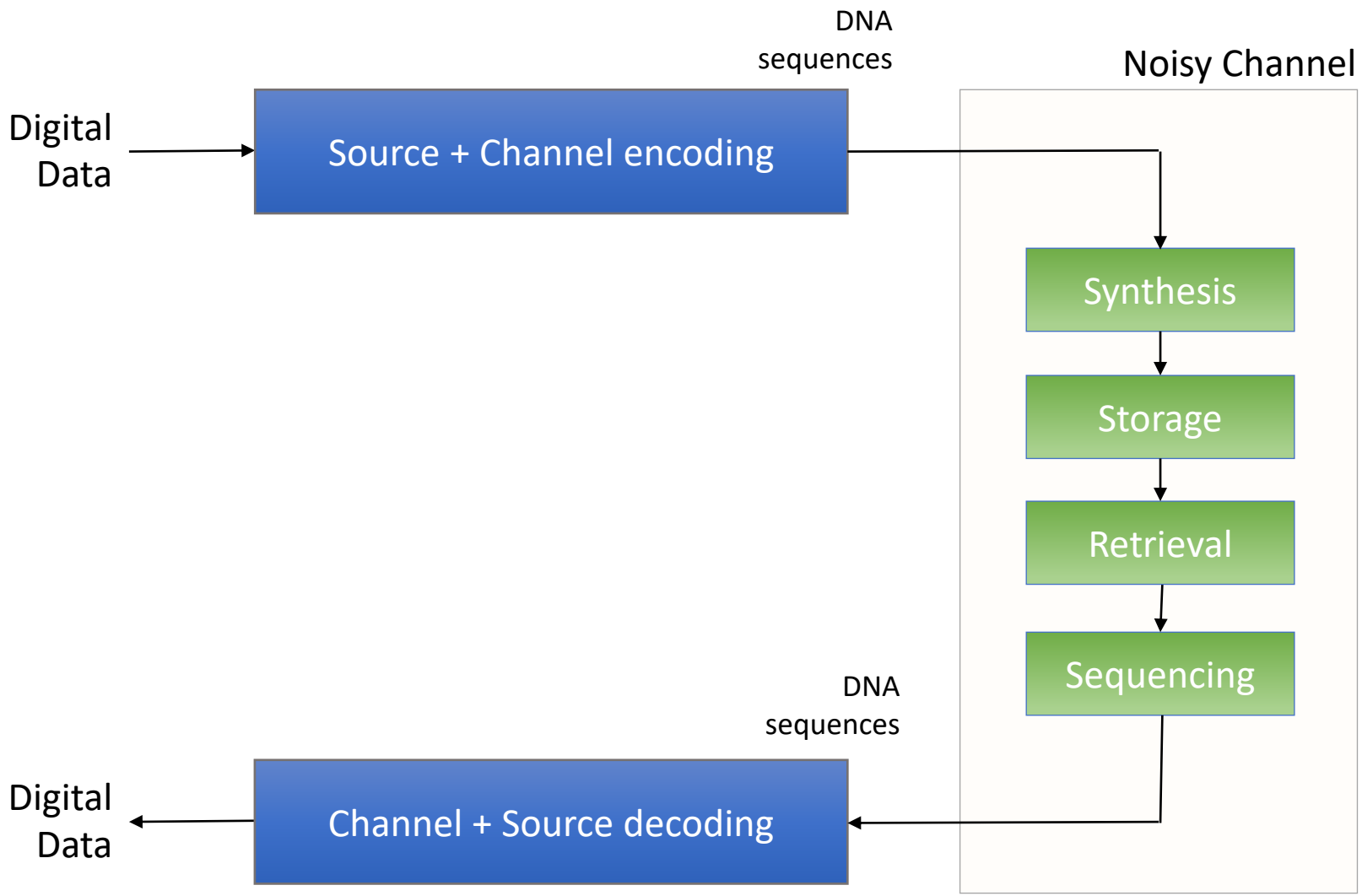


A large majority of

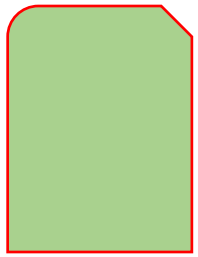
Principle of DNA data storage



Coding / Decoding



Coding scheme



document

Encoding

ATGGACCAGTTGACCGACCATTTGACCAG
TTAGGAACCAGGGATTTAGGAGCCAGGGA
TGGACCACAGGGATTAGGGACCAGGAGGC
⋮
TGCCAATACACCATTACCACACCATTACA

Error Corrector Code

TGGACGATGGACCAGTTGACCGACCATTTGACCAGTGGACCAGTTGAGACATG
TGGACGTTAGGAACCAGGGATTTAGGAGCCAGGGAGCCATTACCATAGACATG
TGGACGTGGACCACAGGGATTAGGGACCAGGAGGCTTGGACCACGGAGACATG
⋮
TGGACGTGCCAATACACCATTACCACACCATTACATAAGACCATTTAGACATG

Length depends of the synthesizer capabilities

ATGGACCAGTTGACCGACCATTTGACCAGTGGACCAGTTGA
TTAGGAACCAGGGATTTAGGAGCCAGGGAGCCATTACCATA
TGGACCACAGGGATTAGGGACCAGGAGGCTTGGACCACGGA
⋮
TGCCAATACACCATTACCACACCATTACATAAGACCATTTA

Coding Challenges

Maximize the ratio: $\frac{\text{Useful information}}{\text{Indexing + CCR}}$

TACGAGACCAGTTGACATTTGACCAGTCAGTTGAGATG

Constraints on sequences

- no (long) homopolymers: ATTAGACTTTTTTTTTCGAGTA
- no specific motifs
- AT / GC ratio between 0.45 & 0.55

Error Corrector Codes

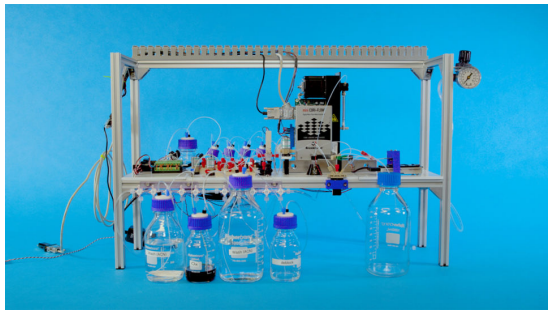
- have to consider the following errors: insertion, deletion, substitution

Agenda

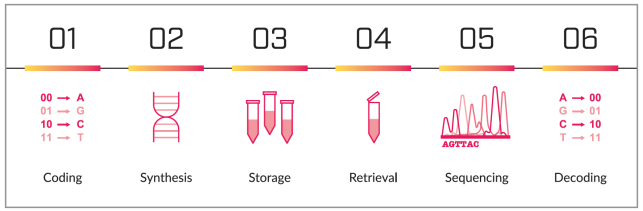
1. Introduction
   - Why new storage medium ?
   - Why DNA ?
   - State-of-the-art

2. DNA storage principle
   - How does it work ?
   - What are the main challenges ?

3. dnarXiv project
   - What's going on in Rennes ?



Source: https://news.microsoft.com/innovation-stories/hello-data-dna-storage/



Source: An Introduction to DNA Data Storage, DNA storage alliance, June 21

dnarXiv

dnarXiv project

Starting date: October 2020

Objective: Investigate information archiving on DNA

Multidisciplinary project

- Signal processing / computer science
- Bio-informatics
- Bio-technology: synthesis & sequencing

dnarXiv research axes

Coding/decoding

Design codes dedicated to DNA storage medium

Bio-technology

Synthesis: store long DNA molecules

Sequencing: 3rd generation sequencing technology

Security

Investigate the possibilities to mixed digital technologies and bio-tech properties

Platform

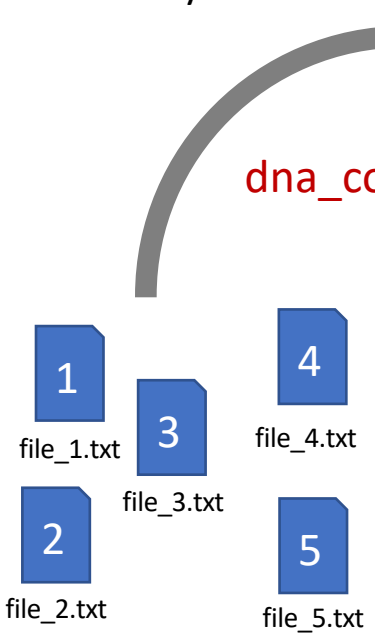
- To integrate codes & Software developed in dnarXiv

- To connect IT and bio-tech parts

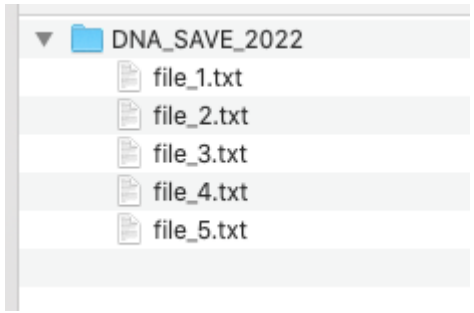
- To conduct experiments
  - wet-lab
  - in-silico

dnarXiv approach

Files that will be archived are copied into a directory

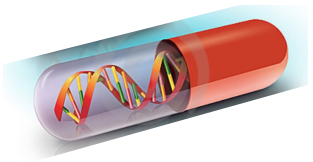


dna_copy



Binary files are transformed into long DNA molecules

dna_store



DNA container

dna_read

Molecules corresponding to one file are extracted and converted to a binary file

dnarXiv bio-technology

Objective: optimize the ratio: $\frac{\text{Useful information}}{\text{Indexing}}$

AGTTGACACCAGTTGACATTTGACCAG

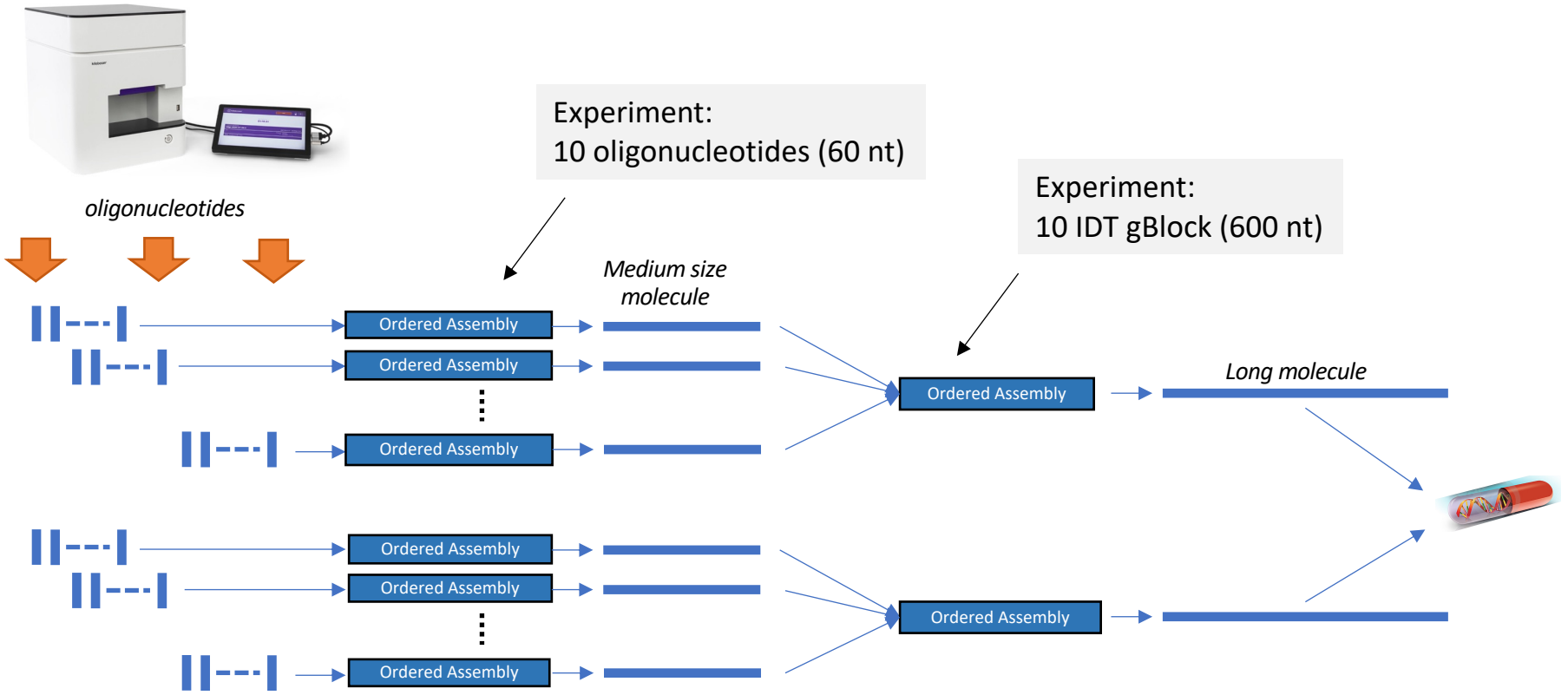
Design long molecules from short oligonucleotides

AGATGCCAGTTGACTTGGGACCAGGATTTAGGGGACACCAGGGGATTGTAGGACCCACGGATTGAATTTGACCAGTTGGACCAGGTTCCATTA

Use 3rd generation sequencing technology

- long reads
- Oxford nanopore technology

Oligos → long molecules



Ordered assembly

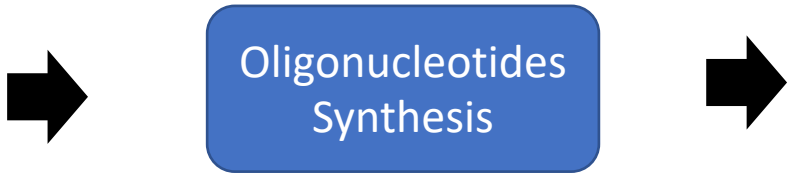
```
AGGACCAGGGATTTAGGCCAGATATGAGGACCGGATTAGGCCCGGGTATATATGATCCATGGGAC...

AGGACCAGGGATTTAG GCCAGATATGAGGACCGGAT TAGGCCCGGGTATATATGAT CCATGGGAC...

AGGACCAGGGATTTAG      GCCAGATATGAGGACCGGAT      TAGGCCCGGGTATATATGAT
|||||||||||||||||     ||||||||||||||||||     ||||||||||||||||||
TCCTGGTCCCTAAATCCGGT  CTATACTCCTGGCCTAATCC     GGGCCCATATATACTAGGTA
```

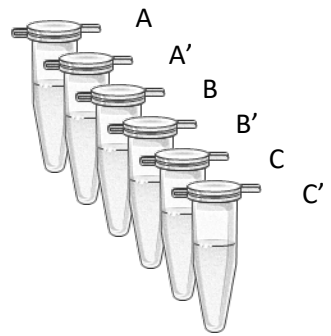
Split sequence

Design Oligos

```
A  : AGGACCAGGGATTTAG
A' : GCCAGATATGAGGACCGGAT
B  : TAGGCCCGGGTATATATGAT
B' : TCCTGGTCCCTAAATCCGGT
C  : CTATACTCCTGGCCTAATCC
C' : GGGCCCATATATACTAGGTA
...
```

List of oligos





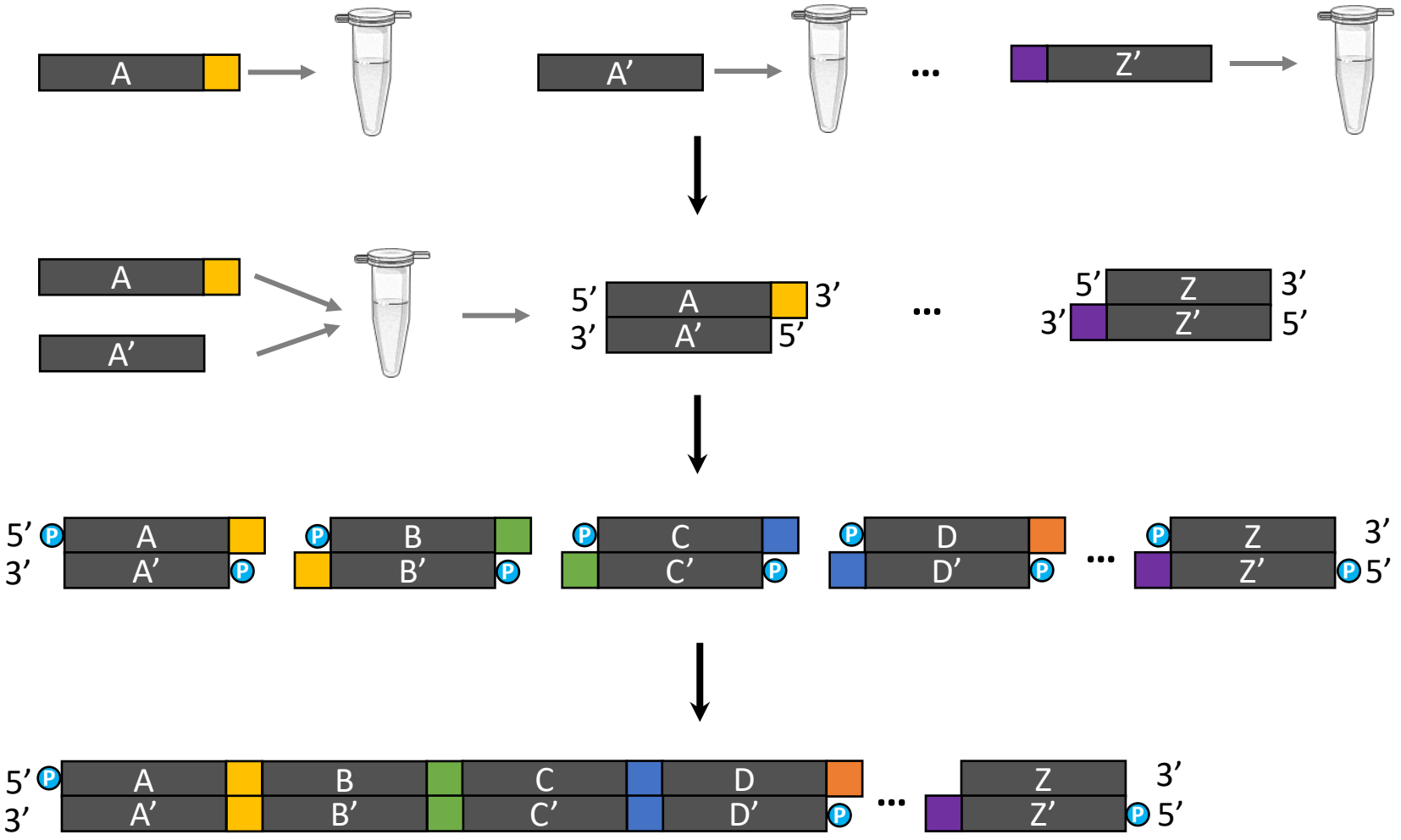
Synthetic DNA assemblies workflow

1 **Design and purchase ss oligo's**
ssDNA

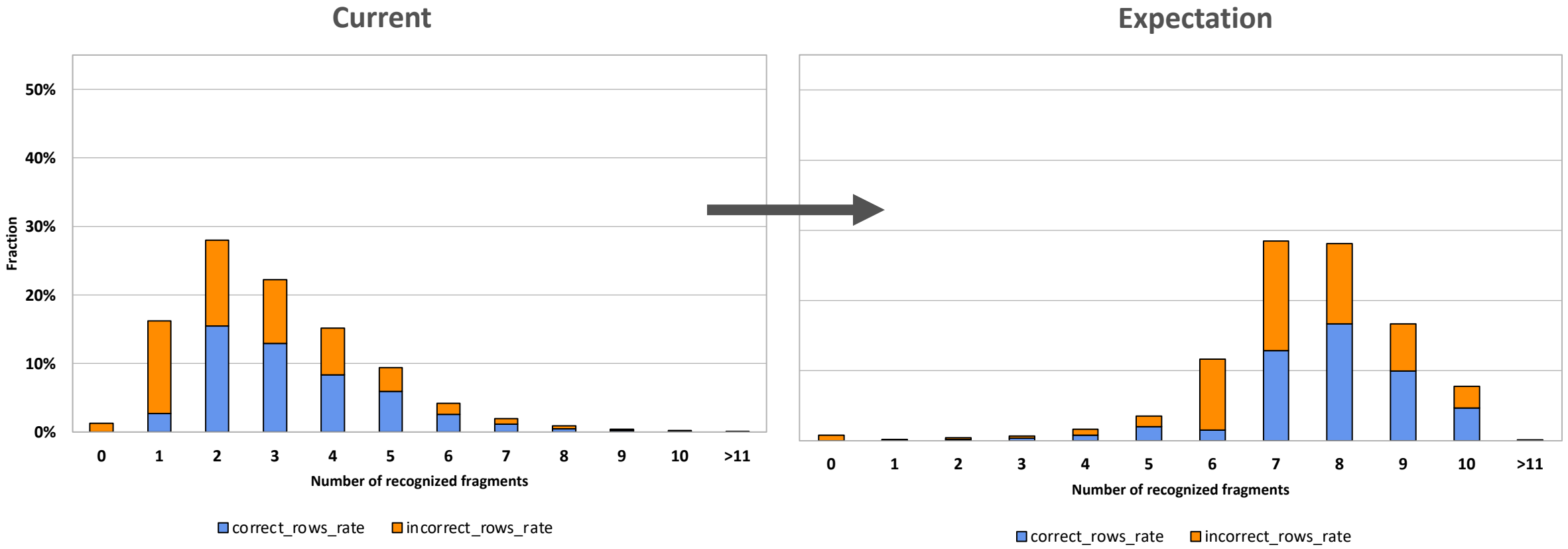
2 **Oligonucleotides annealing**
ssDNA → dsDNA

3 **5' phosphorylation**
T4 PNK

4 **Ligation**
DNA ligase T4 or T7

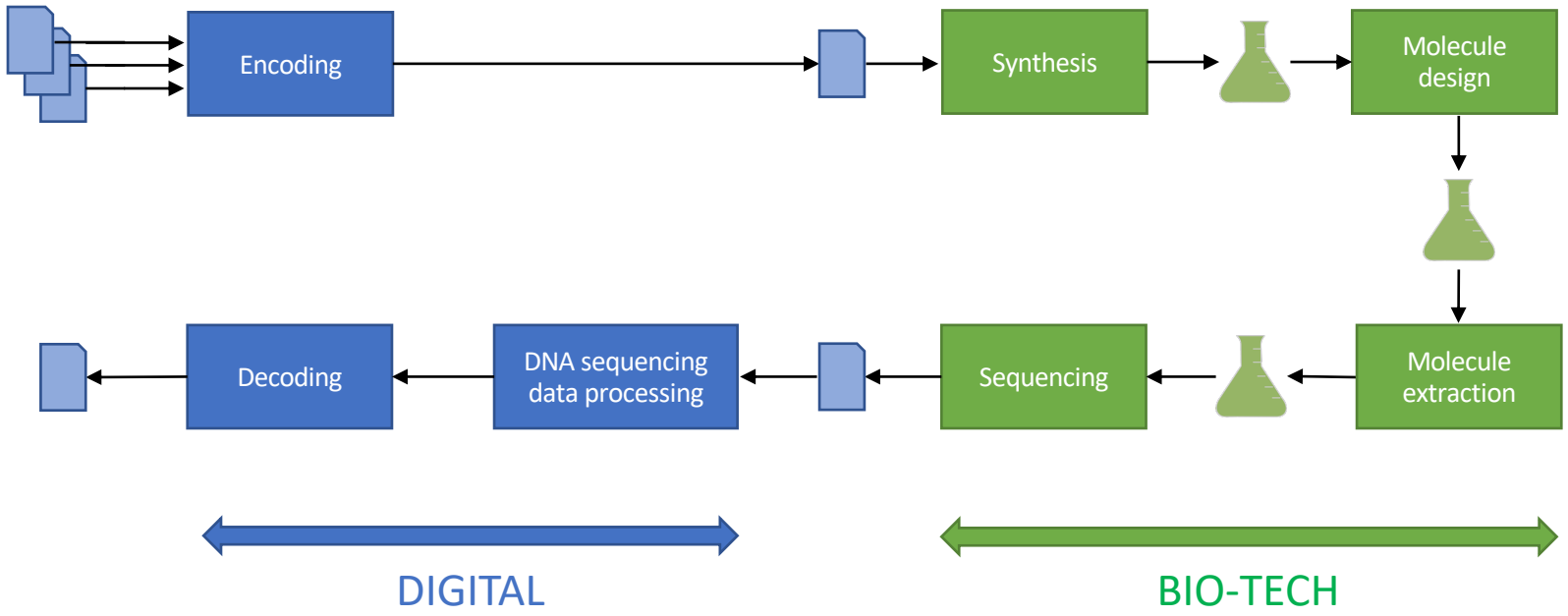


Ordered assembly experiments

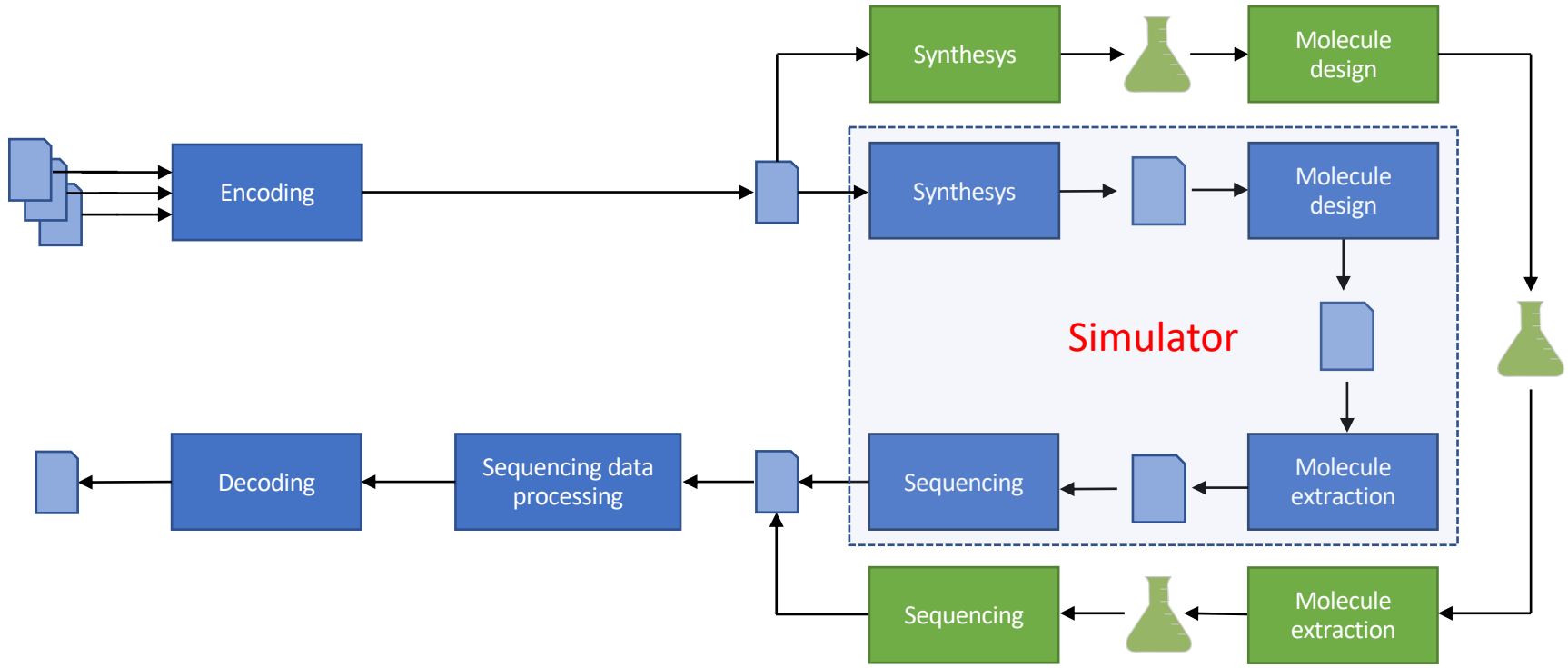


Work on the limiting steps of the process: inhibitor reaction products, ATP degradation, enzymes stability …
Next: try to assemblies more fragments and increase correct row rates.

dnarXiv platform



in-silico alternative



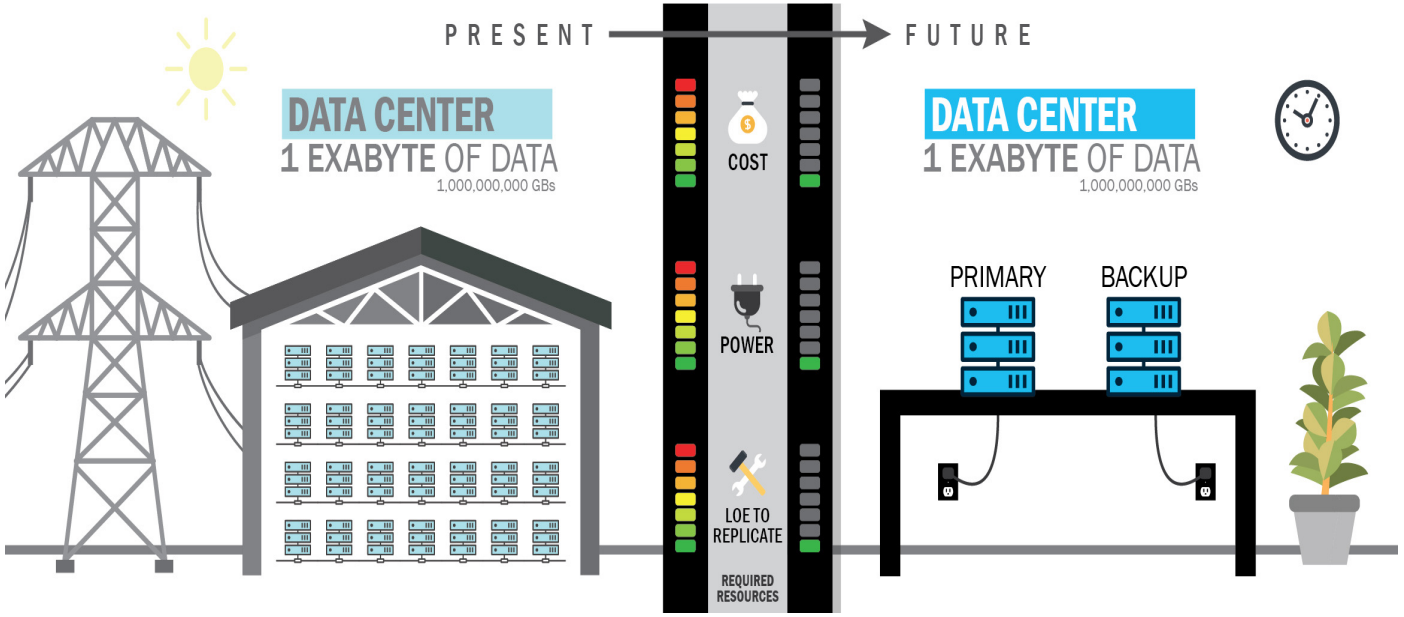
Take home message

Data explosion
- Need for new storage medium

DNA molecules
- density / durability
- cold data (archiving)

Today bottleneck = Synthesis
- slow & costly
- keys: miniaturization, parallelization, automation

New active research field
- international initiatives



https://www.iarpa.gov/research-programs/mist

International initiatives

MIST Molecular Information Storage (IARPA)

- 2020 – 2023 (4 years)
- Roadmap 2024: write 1TB/read 10TB, 1 day, 1000 $

https://www.iarpa.gov/research-programs/mist

DNA Data Storage Alliance

- formed in October 2020
- founders: Illumina, Microsoft, Twist Bioscience and Western Digital
- Today : 50 members

https://dnastoragealliance.org/

EIC Pathfinder Challenge: DNA-based digital data storage

- Deadline: oct 2022

PEPR MoleculArXiv, France

- start year: 2022

Thank you for your attention

dominique.lavenier@irisa.fr

CominLabs

UMR IRISA

Inria

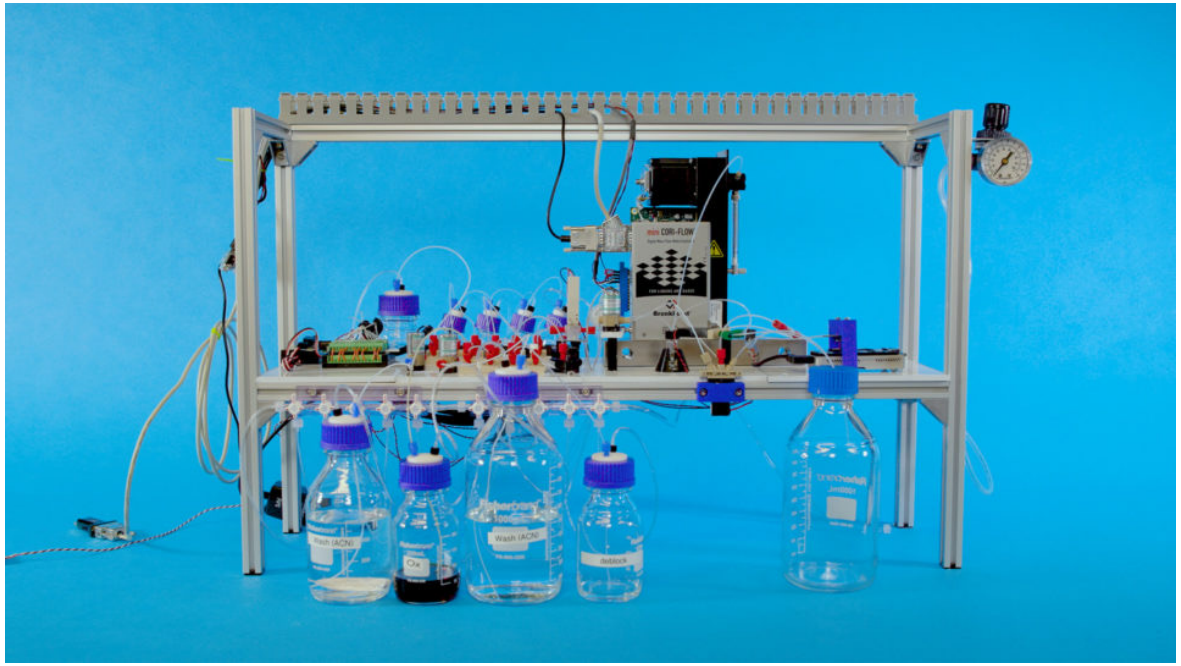
State of the art

MICROSOFT & University of Washington
March 2019

First fully automated system to store and retrieve data in manufactured DNA

Write & read automatically the word “**Hello**”

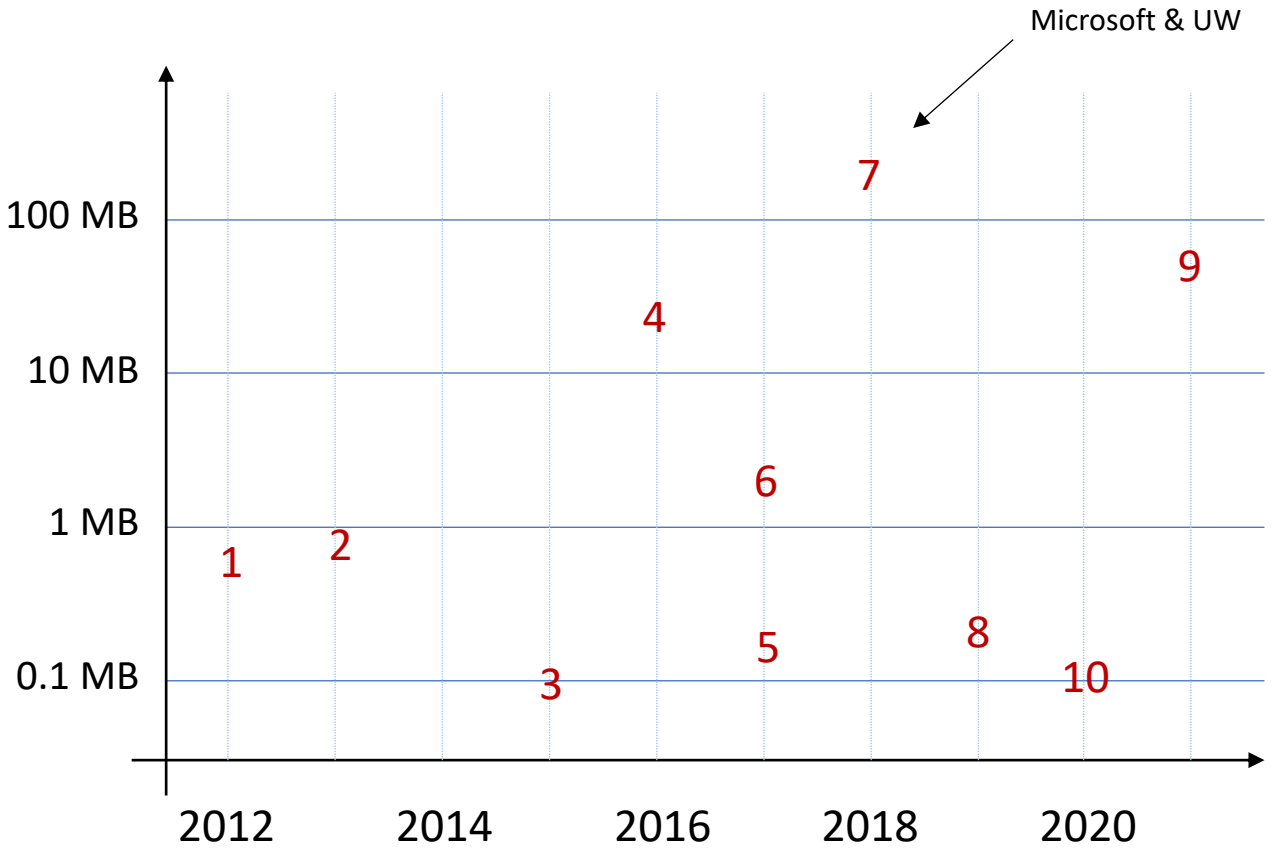
Process time : 21 hours



Takahashi, C.N., Nguyen, B.H., Strauss, K. *et al.* Demonstration of End-to-End Automation of DNA Data Storage. *Sci Rep* **9,** 4998 (2019)

Source: https://news.microsoft.com/innovation-stories/hello-data-dna-storage/

DNA Storage Experiments



		length of seqs.	number of seqs	data stored (in MB)	error correction
1	Church, Gao, and Kosuri [27]	115	54,898	0.65	None
2	Goldman, Bertone, Chen, *et al.* [38]	117	153,335	0.75	Repetition
3	Grass, Heckel, Puddu, *et al.* [39]	117	4,991	0.08	RS
4	Blawat, Gaedke, Hütter, *et al.* [12]	190	900,000	22	RS
5	Bornholt, Lopez, Carmean, *et al.* [14]	120	45,652	0.15	RS
6	Erlich and Zielinski [33]	152	72,000	2.14	Fountain
7	Organick, Ang, Chen, *et al.* [80]	150	$13.4 \cdot 10^9$	200.2	RS
8	Chandak, Tatwawadi, Lau, *et al.* [17]	150	13,716	0.192	LDPC
9	Heckel and Grass [46]	105	$3.88 \cdot 10^9$	63.1	RS
10	Antkowiak, Lietard, Darestani, *et al.* [7]	60	16,383	0.1	RS

I. Shomorony, R. Heckel (**2022**), "Information-Theoretic Foundations of DNA Data Storage", Foundations and Trends® in Communications and Information Theory: Vol. 19: No. 1, pp 1-106
http://dx.doi.org/10.1561/0100000117