Digital data storage on synthetic DNA
Technological advances and challenges

M. ANTONINI
MEDIACODING RESEARCH GROUP

am@i3s.unice.fr

Séminaire Aristote
Ecole Polytechnique
April 13th 2023

UNIVERSITÉ CÔTE D'AZUR

cnrs

i3s sophia antipolis

Priority Research Program and Equipment
MoleculArXiv

LIGO ARCHIVE

The first hard disk drive in the history of computing

An IBM RAMAC 350 hard disk is loaded on board a DC7 - Fall 1957

5 MB of storage

1000 kilograms

$10 000 / MB !
(in 1956)

Could only store
ONE SINGLE IMAGE!

STORAGE LIMITS

Estimates based on bacterial genetics suggest that digital DNA could one day rival or exceed today's storage technology.

	Hard disk	Flash memory	Bacterial DNA
Read–write speed (μs per bit)	~3,000–5,000	~100	<100
Data retention (years)	>10	>10	>100
Power usage (watts per gigabyte)	~0.04	~0.01–0.04	$<10^{-10}$
Data density (bits per cm^3)	$\sim10^{13}$	$\sim10^{16}$	$\sim10^{19}$

WEIGHT OF DNA NEEDED TO STORE WORLD'S DATA

~1 kg

©nature 2016

It is estimated that **215 petabytes** (215 million gigabytes) can be stored in a **single gram of DNA!**

☺

- **Ultracompact** (1billion times more than HDD)
- Can last **hundreds of thousands of years** if kept in a cool, dry place

⇒ Sequencing of the DNA of a mammoth (1.2 million years old)

⇒ Sequencing DNA of a horse bone (700,000 years old)

- **Eco-friendly** solution

☹

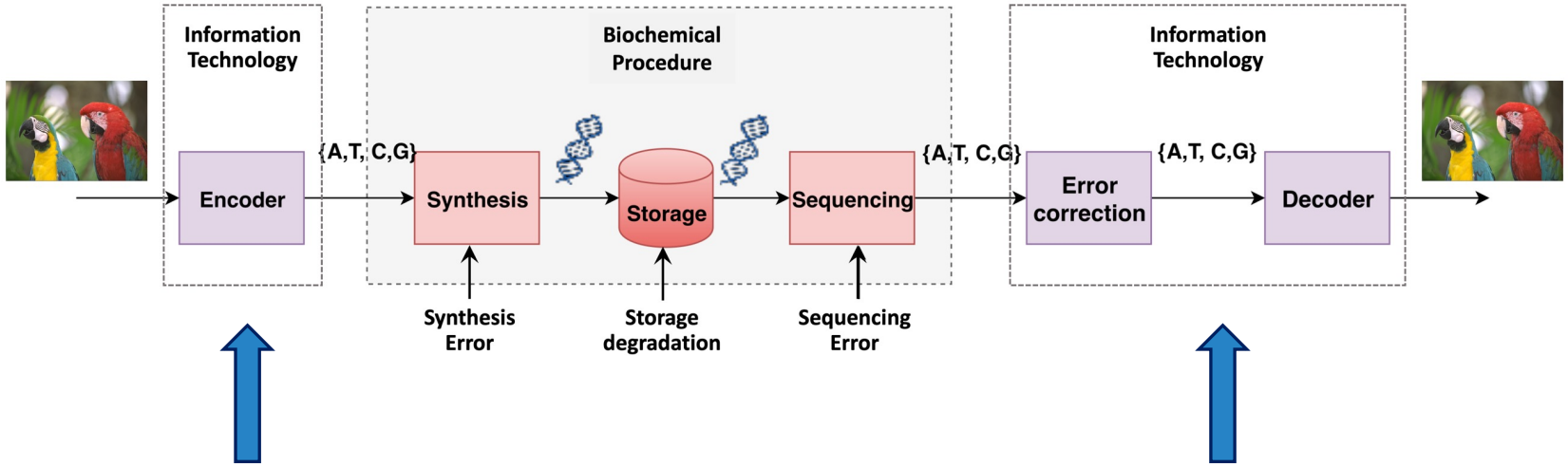
- **Low** synthesis **speed**
- **High price** for synthesizing and sequencing DNA (around $1000 / MB today)

Cold data

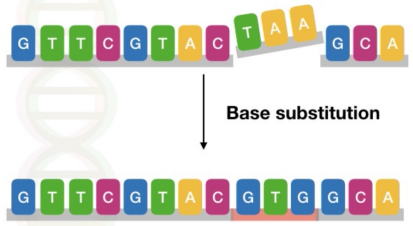
How to store digital data in synthetic DNA?

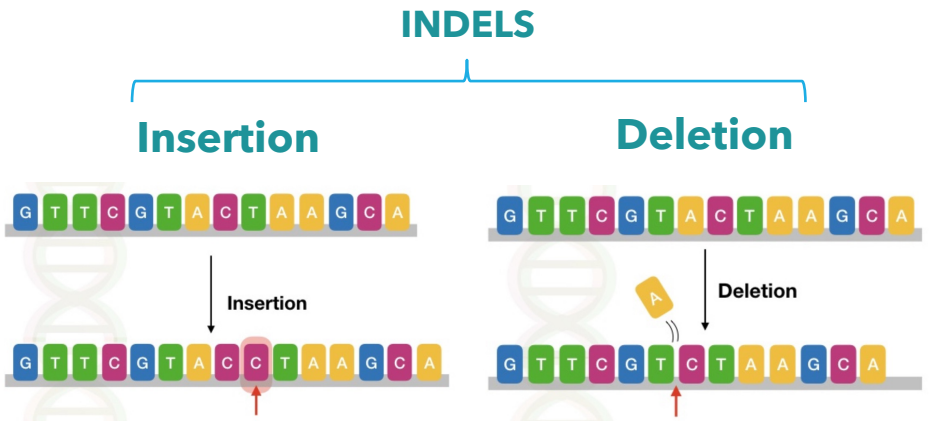
The biochemical process will keep evolving to achieve

- Reduction of errors
- Faster synthesis and sequencing
- Reduction of the costs



Encoding/Decoding **need to keep adapting to changes**





Coding restrictions

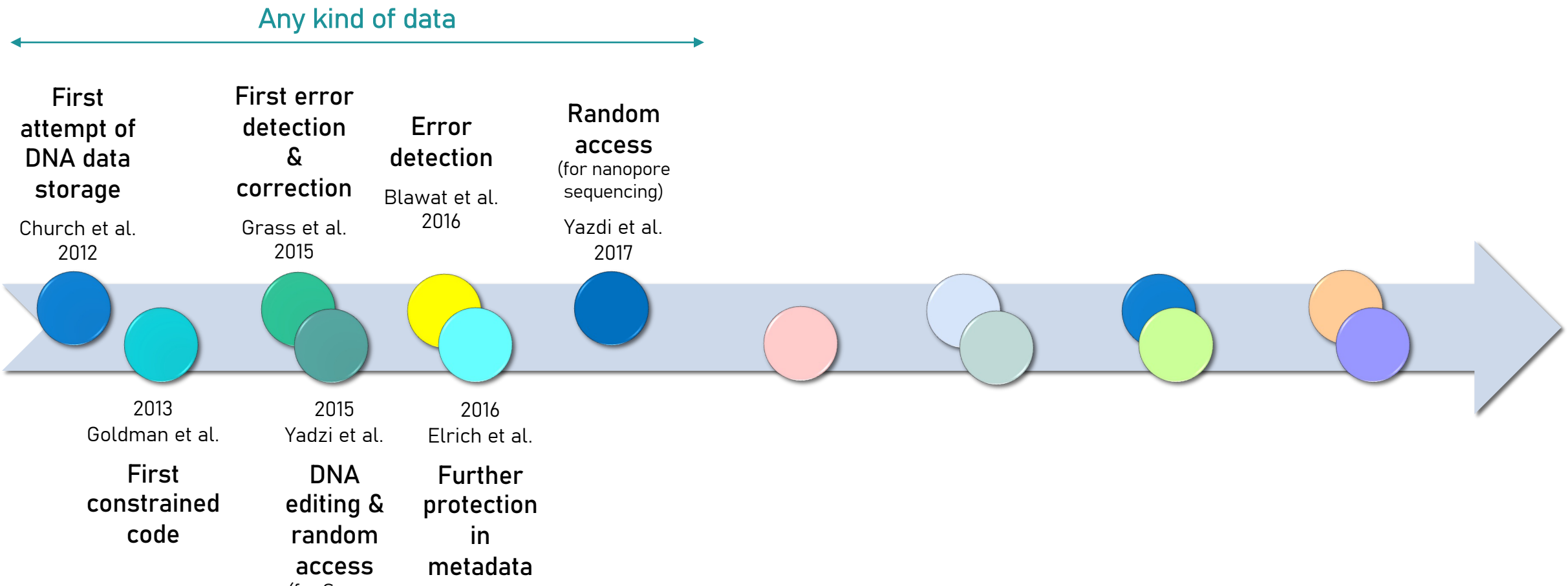
Reducing the **synthesis error**

- **Oligos should be short (length < 300 nucleotides)**
  - Formatting of encoded sequence cutting it into smaller pieces and introducing headers.

Reducing the **sequencing error**

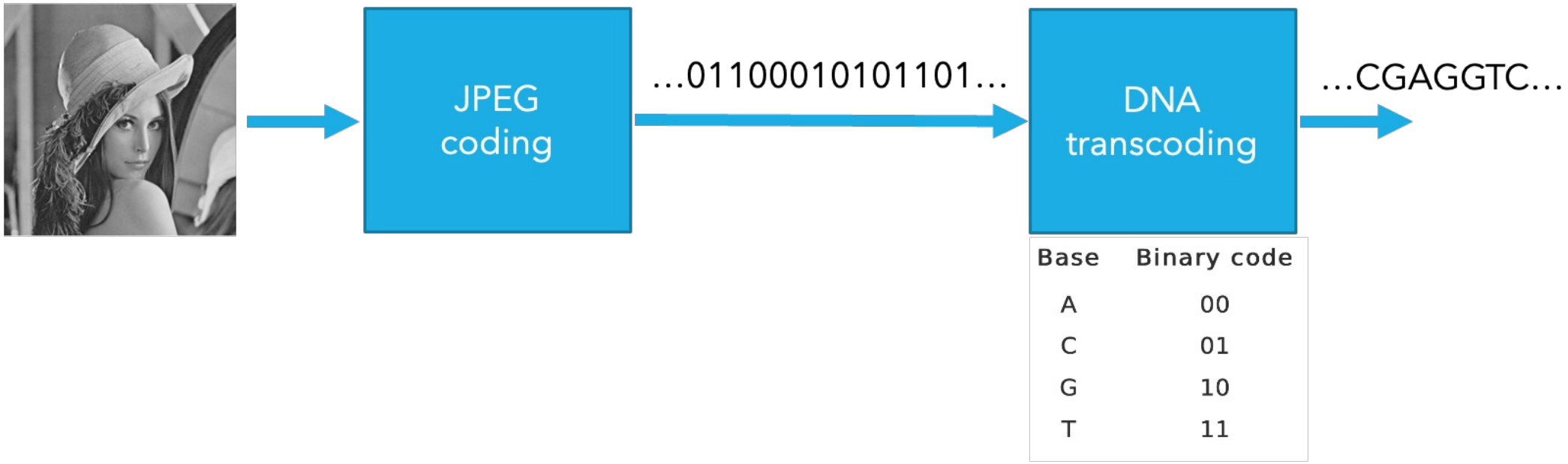
- **Homopolymer runs**
  - Consecutive occurrences longer than 3 or 5 nucleotides (nts) should be avoided. (ex. AAAAA or TTTTTT …)
- **GC content**
  - %{G, C} ≤ %{A, T}
- **Pattern repetitions**
  - Codewords should not be repeated forming patterns (ex. ATCATCATC…)

State of the art on DNA coding



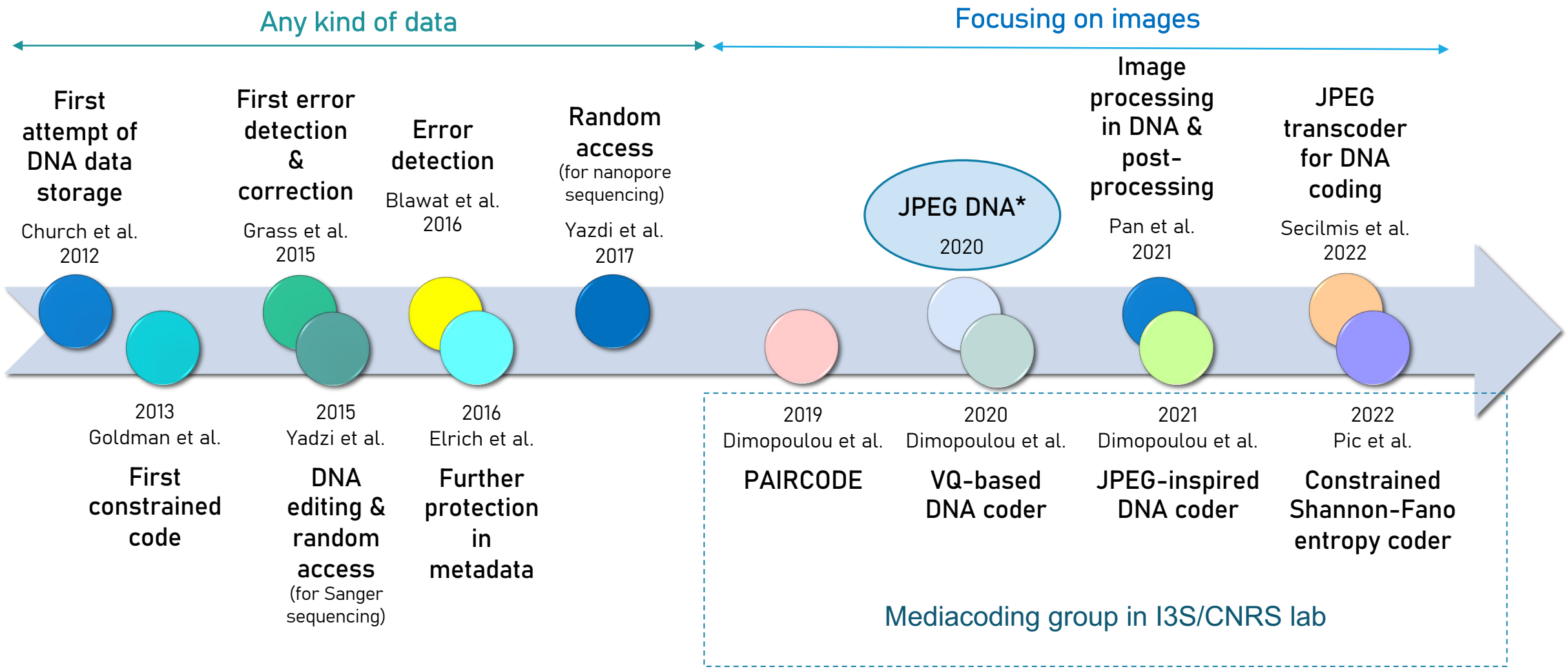
State-of-the-art: transcoding

Most of previous works **transcode** binary information into a quaternary code without taking into consideration the nature of the input data



PROBLEM: **No Quality/Cost control during compression**

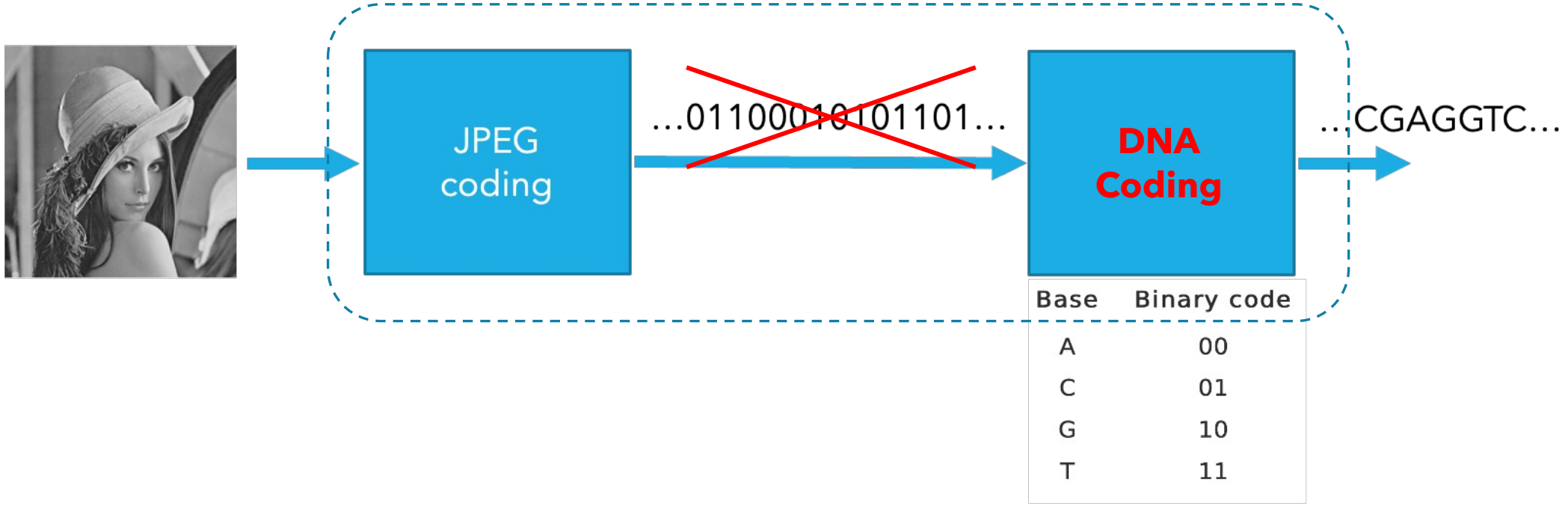
State of the art on DNA coding



* https://jpeg.org/jpegdna/

In loop DNA coding

Direct DNA encoding allows to optimize the **Quality/Cost trade-off during compression** taking into consideration the nature of the input data



Challenges

Consideration of noise and constraints introduced by new chemical processes and sequencing

- Noise models
- Design of error correction quaternary codes
- Robust decoding
- Coding on synthetic polymers: beyond 2 bits / base

Big Data Management

- Structuration of the stored data
- Solutions for random access to data
- Joint sequencing/decoding

I3S laboratory activities on DNA storage

Compression

- Design of image compression solutions
- JPEG DNA standardization

Coding

- Robust and constrained 4-ary and N-ary code construction

Neural network

- Cleaning of the sequenced data
- Noise removal and error correction
- Design of coding/decoding solutions

OLIGO ARCHIVE

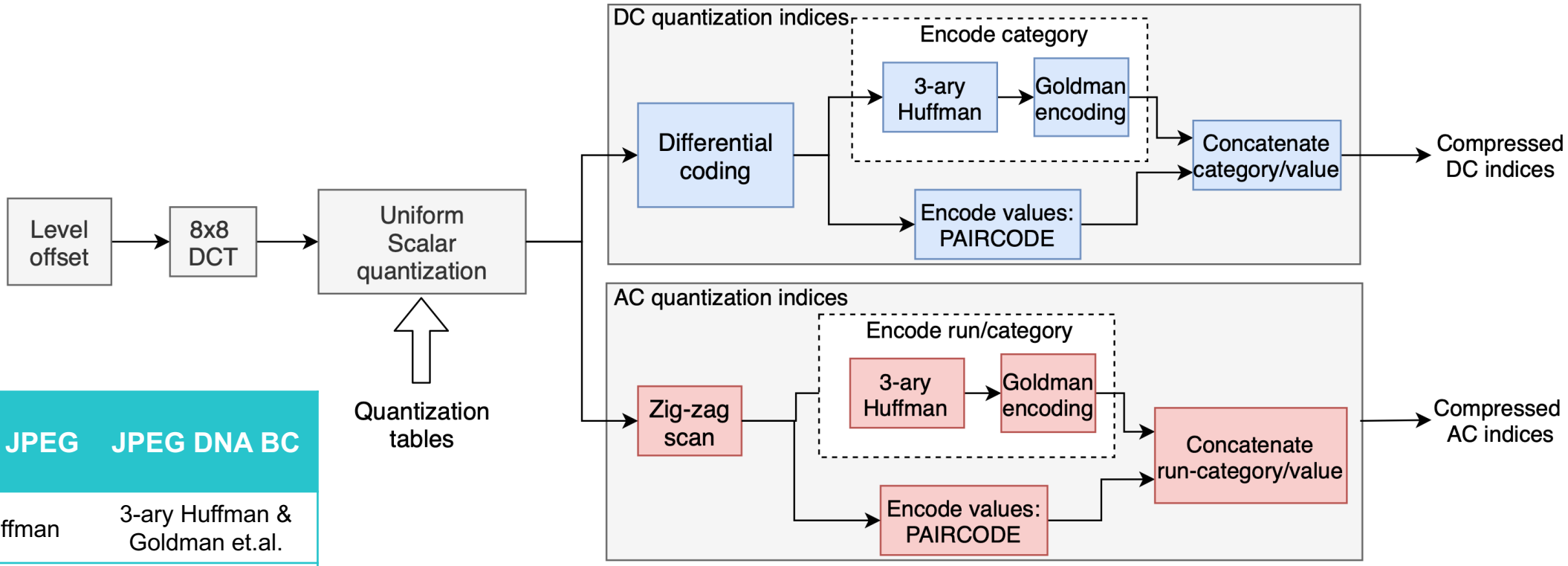
JPEG DNA

PEPR MoleculArXiv

20M€ - 7 years

PEARCODE

*** PATENT** - *Methods for storing digital data as, and transforming digital data into, synthetic DNA,* M. Dimopoulou, M. Antonini, US n°16/811,985, 2020

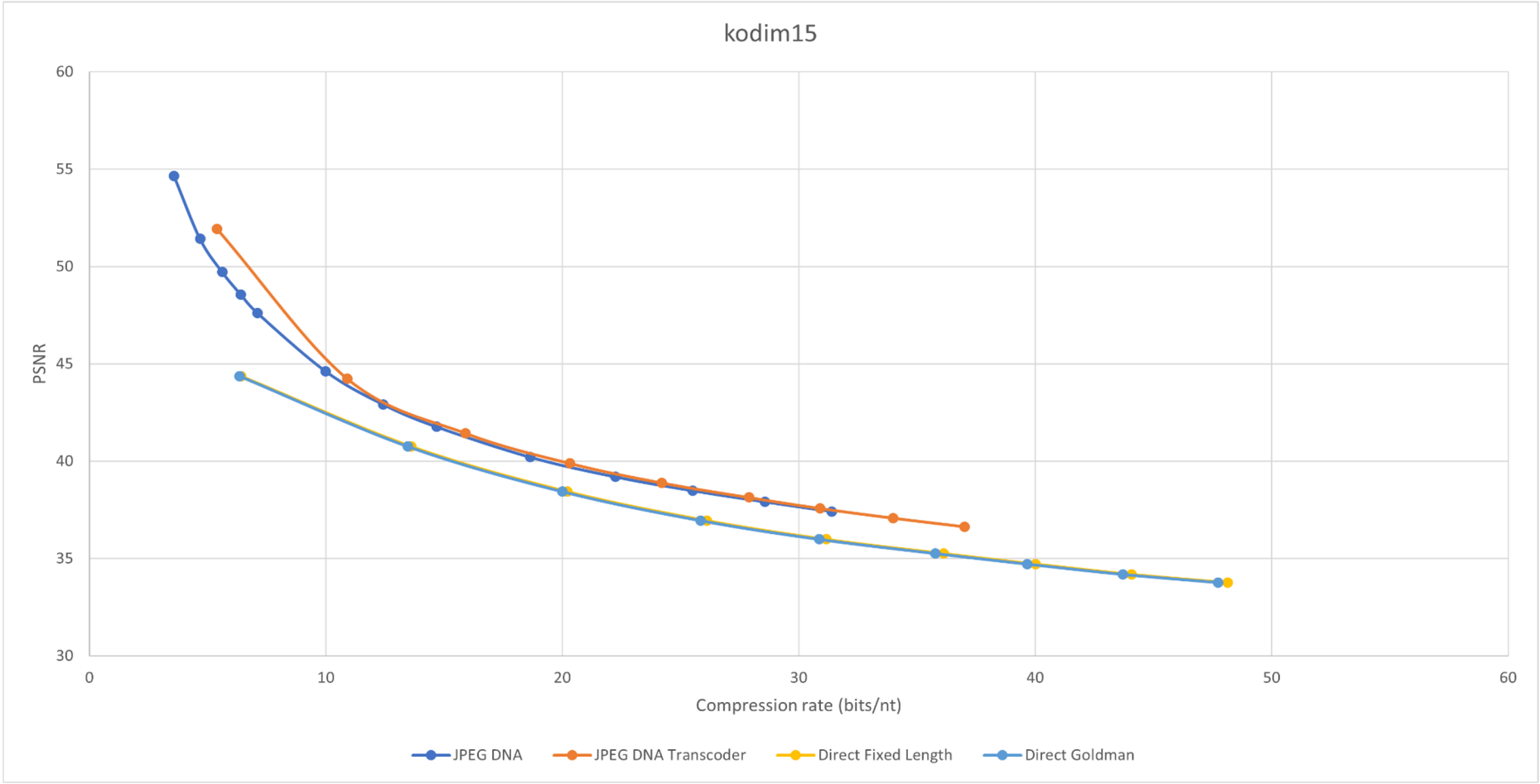


	Classical JPEG	JPEG DNA BC
AC/DC Categories	Binary Huffman	3-ary Huffman & Goldman et.al.
AC/DC Values	Classical binary encoding	PAIRCODE encoding

The JPEG DNA Benchmark Codec*

* www.jpeg.org
and Melpomeni Dimopoulou, Eva Gil San Antonio, Marc Antonini, EUSIPCO 2021 (https://tel.archives-ouvertes.fr/tel-03152789)

Performance

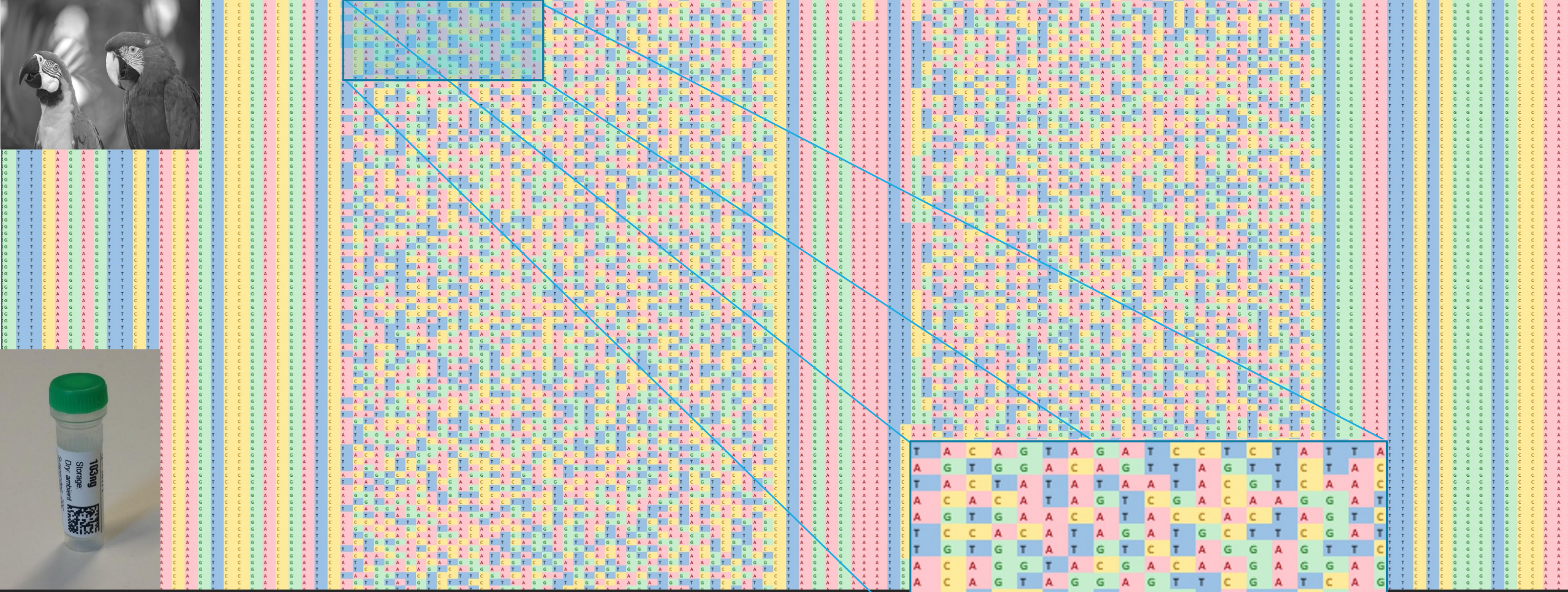


DNA was encapsulated in DNAshell (Imagene)

Wet lab experiment

- To verify the feasibility of storage and reconstruction using our encoding algorithm we performed a wet lab experiment with several images (molecular synthesis)
- For the Kodak image 23 “Parrot”
  - Coding ratio : 10.26 bits/nt
  - PSNR = 41.5dB
  - 2571 oligos (size 200nt) were synthesized*

* Twist Bioscience

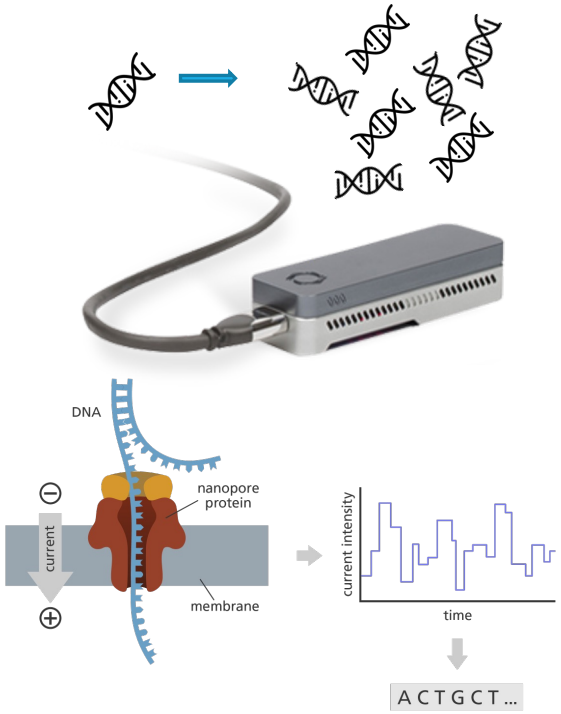


Format of the oligos

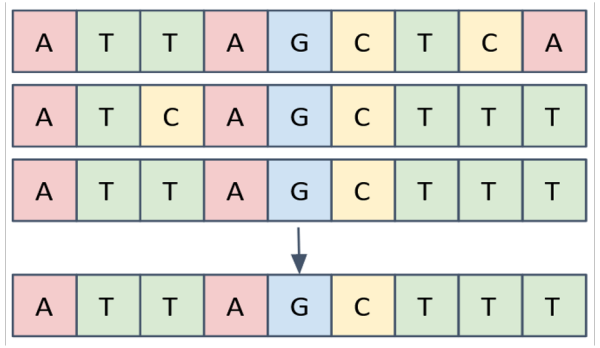
26nt	1nt	3nt	6nt	75nt	4nt	1nt	1nt	21nt
Primer 1	S	H	offset	Payload	ID	P	S	Primer 2

Image decoding

MinION SEQUENCING TECHNOLOGY



CLUSTERING and CONSENSUS FINDING



Assign to each position inside the sequence the most frequent symbol along the cluster

DECODED IMAGE

JPEG DNA BC PSNR = 37.55dB

Collaboration with:

- IPMC (UMR 7275 CNRS and UCA)
- EURECOM

Conclusion

New compression/coding algorithm for the robust encoding of digital images into DNA

Allows to control the trade-off QUALITY/COST

The proposed solution can be applied on any kind of input data format (binary, symbols, quantized samples...)

Tackles the problems of biochemical constraints

Future works

SmidgION Oxford Nanopore Technologies

Improve coding/decoding performance

- Improve robustness of coding solutions
- Robustify to **sequencing** AND **synthesis** AND **storage noise**
- New error correction quaternary codes

Deal with greater length oligos (>300nt)

Nanopore sequencing

- Lower sequencing cost
- **Prone to errors** -> new solutions for robust decoding based on Machine Learning
- Noise models

The PEPR MoleculArXiv

Massive data storage on DNA and artificial polymers

Program manager Marc Antonini

- Specialist in **data compression and coding**, strong experience in **DNA data coding**, chair of the **JPEG DNA** AHG, co-founder of the start-up Cintoo and PearCode

Duration 84 months

Budget 20 M€

Lead institution CNRS

- CNRS gathers skills in computer science, (bio-)chemistry, microfluidic and sequencing
- **16 French laboratories** directly involved including **6 flagship labs** that cover the fields and also involved in the steering committee
- **A potential ecosystem of 50 laboratories**

Leader French laboratories directly involved in MoleculArXiv

The PEPR MoleculArXiv
Create and federate a community

Prime a community
- **Direct funding** to the laboratories involved in the WP
- **Platforms** to sustain the sharing of resources and instruments inside the community

Develop a community
- **Chairs** to promote the recruitment of the future investigators of the field
- **ANR calls** to foster the creation of strong interdisciplinary projects

Enlarge the community
- Organize recurrent **interdisciplinary international Workshops and Summer Schools**
- **Encourage discussions and exchanges between researchers** from different communities
- Create the right environment for the creation of a European flagship project

Foster technological transfer
- Push new technologies to industry (pre-maturation or start-up creation)

Melpo Dimopoulou
Postdoc
UCA-CNRS I3S

Eva Gil San Antonio
PhD student
UCA-CNRS I3S

Xavier Pic
PhD student
UCA-CNRS I3S

THANK YOU