Stockage de données sur ADN Synthétique

GT du GDR IASIS (2025 & 2026)

- Responsables : M. Antonini, E. Dupraz, D. Lavenier

Initiative du PEPR MoleculArXiv

- https://pepr-molecularxiv.fr/

Objectifs

- Susciter l'intérêt de la communauté sur le sujet
- Explorer des sujets non couverts par le Projet Ciblé 2 du PEPR
  - De la donnée numérique à l'ADN synthétique
    https://pepr-molecularxiv.fr/projet/de-la-donnee-numerique-a-ladn-synthetique/

Agenda

date	horaires	sujet	Intervenants
11 03 2025	14h00 - 15h30	Introduction au domaine	M. Antonini, E. Dupraz, D. Lavenier
10 04 2025	10h00 - 11h30	Production des données: synthèse, séquençage	P. Barbry, A. Genot
16 05 2025	10h00 - 11h30	Reconstruction des données	D. Lavenier
24 06 2025	10h00 - 11h30	Codage canal	E. Dupraz
23 09 2025	14h00 - 15h30	Codage source	M. Antonini
23 10 2025	14h00 - 15h30	Théorie de l'information	S. Kas Hanna

Liste de diffusion pour la formation : gt-gdr-iasis-dnastorage@inria.fr

The first HDD in the history of computing

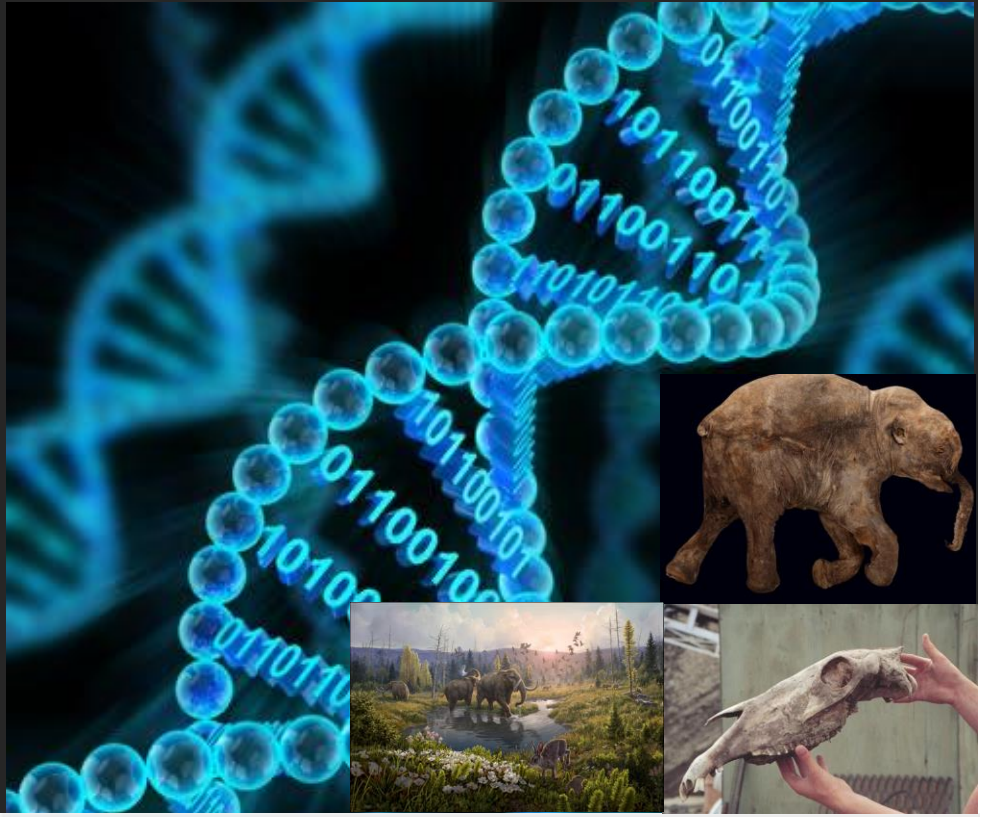
An IBM RAMAC 350 hard disk is loaded on board a DC7 - Fall 1957

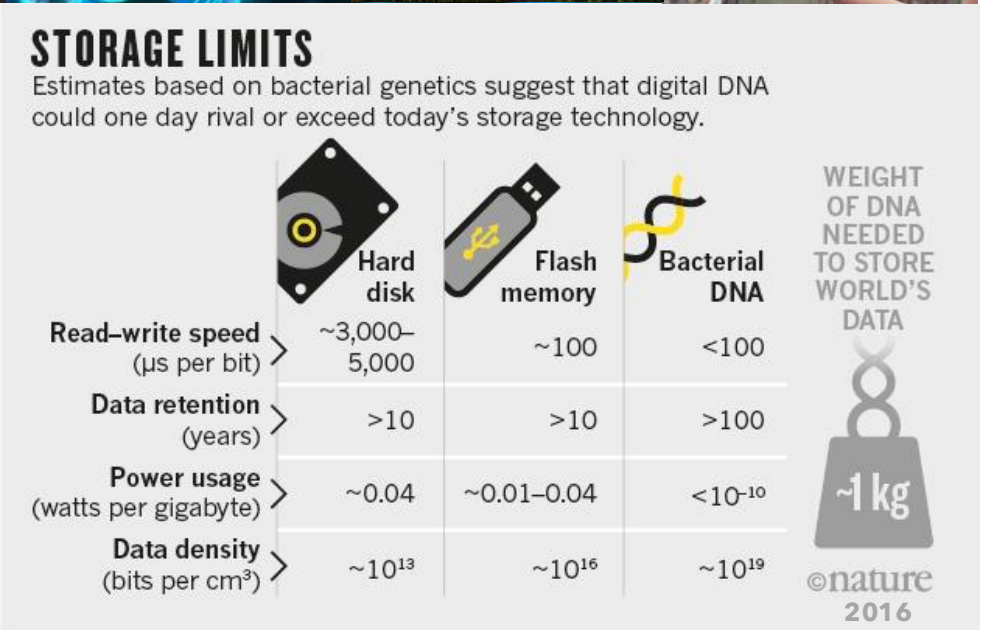
5 MB of storage

1000 kilograms

$10 000 / MB !
(in 1956)

Could only store
ONE SINGLE IMAGE!





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)

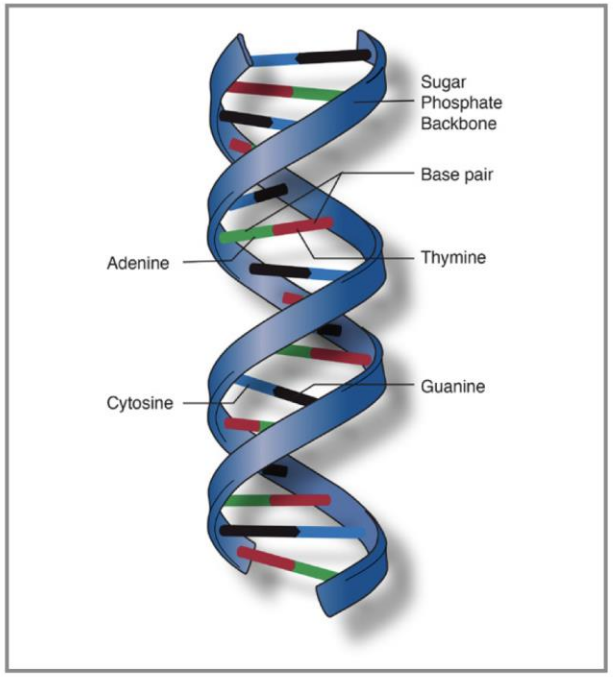
⇒ A vanished Arctic world in Greenland revealed by 2-million-year-old DNA

- **Eco-friendly** solution

☹

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

What's DNA?

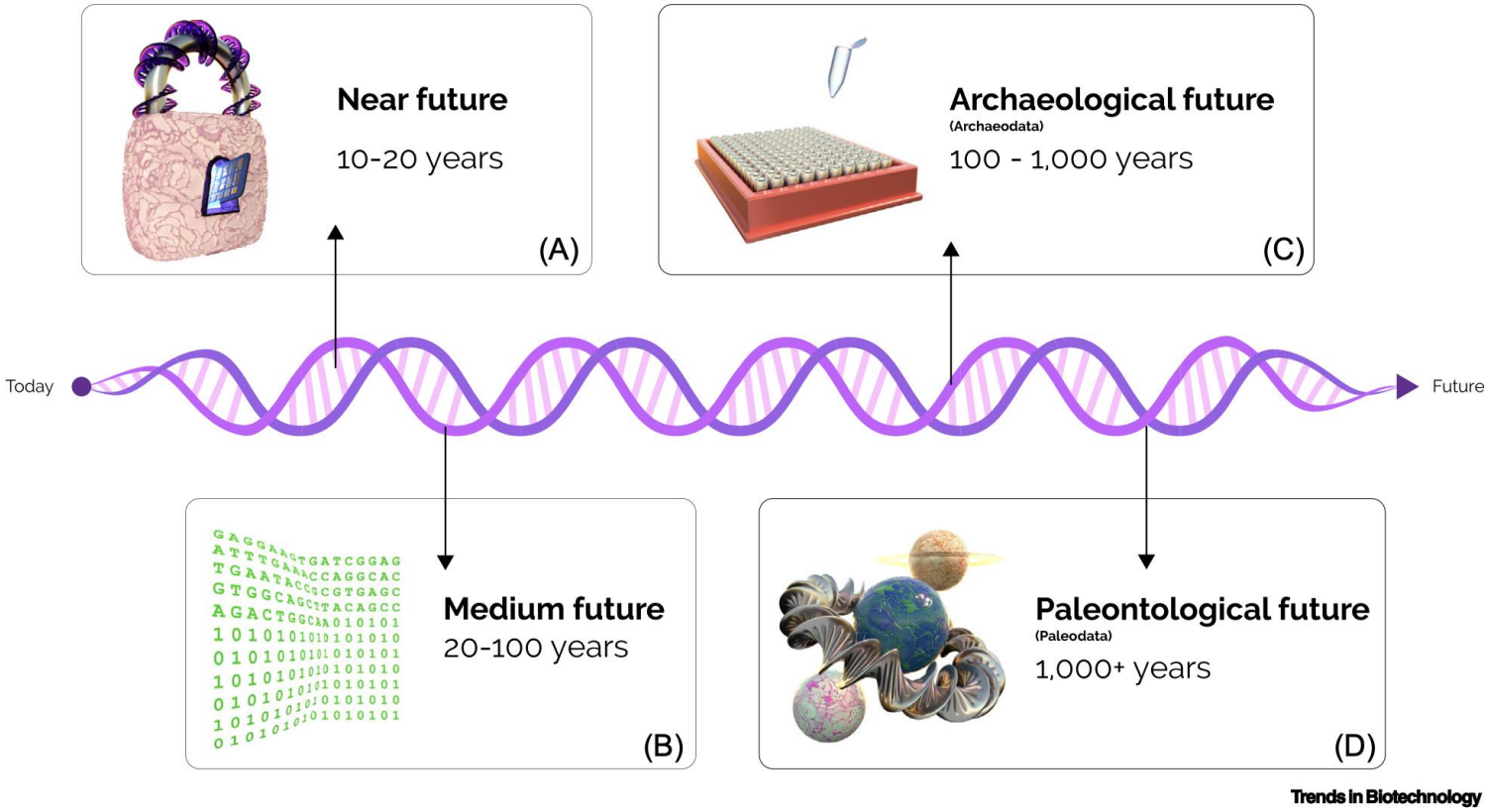


Courtesy: National Human Genome Research Institute

- DeoxyriboNucleic Acid (DNA)
  - Biological molecule holding information about life
  - Strand of 4 different monomers, the **nucleotides** (nts)
    - Adenine (A)
    - Thymine (T)
    - Cytosine (C)
    - Guanine (G)
- Advances in the field of chemistry allow synthesis (writing) and sequencing (reading) of any DNA sequence in vitro
- In the context of digital data storage, DNA is manufactured:
  The creation of the DNA data storage medium does not require or use - nor does the resulting stored data result in the creation or modification of - any cells, organisms, or life

Timeline

The different timings of the use of DNA data storage require distinct factors to be standardized



Format obsolescence needs data migration

VHS/Betamax
~1976-2016

wikipedia

CD-ROM
1982-??

DVD
1995-??

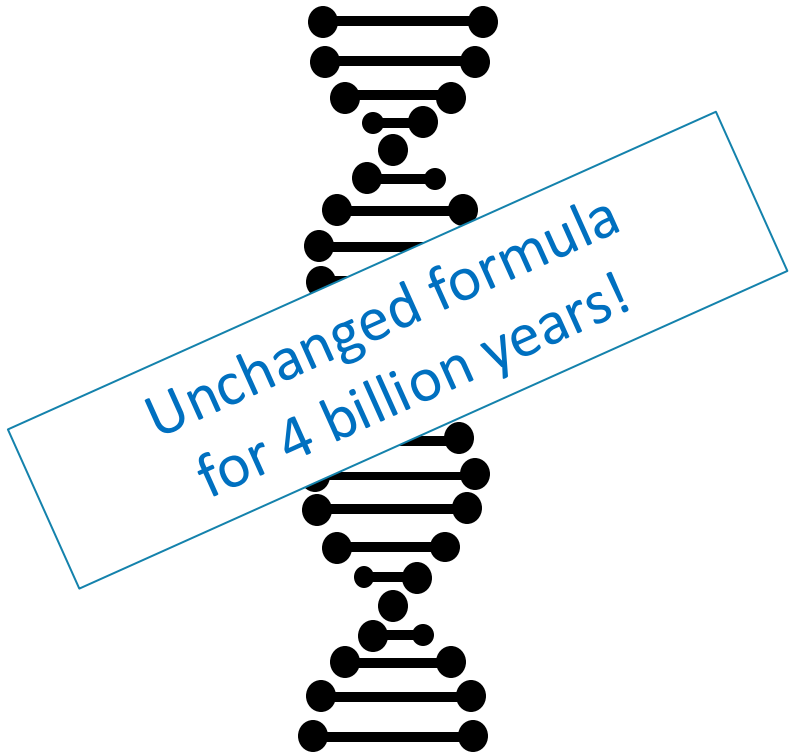
Blu-Ray
2006-??

DTF
1994-2004

SDLT
1984-2007

LTO
2000-??

DNA: Nature's choice for its data



Cold data

A single gram of DNA

VS

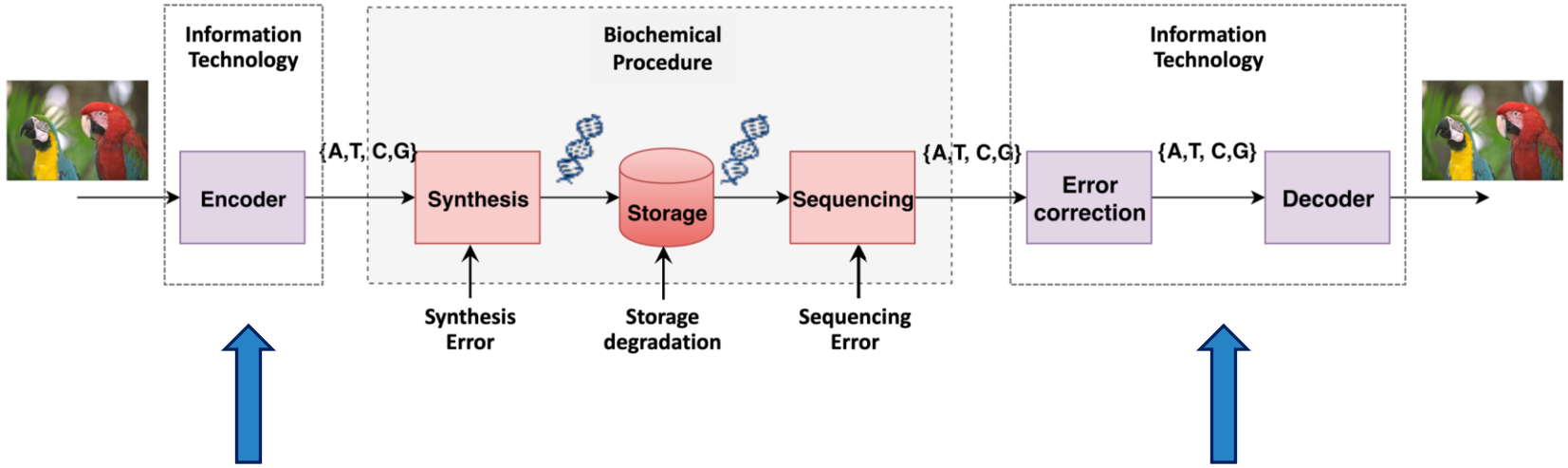
Main challenges

- Improve the writing (synthesis) speed – 100s for a unique bit
- Decrease the cost – around 1000€ per MB
- 3rd generation of sequencing able to manage Tera bases
- Random access to the stored data
- Security of the stored data
- Adapted compression/coding solutions

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**

- **Synthetic oligonucleotides must be shorts (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…)

Synthetic DNA Synthesis

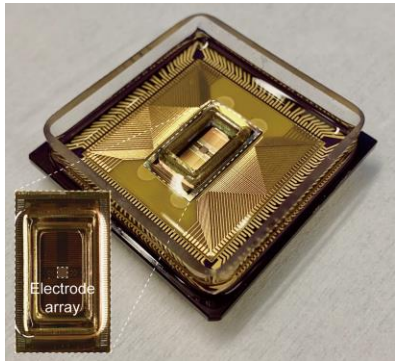
Text file

DNA molecules

Synthetic DNA Synthesis

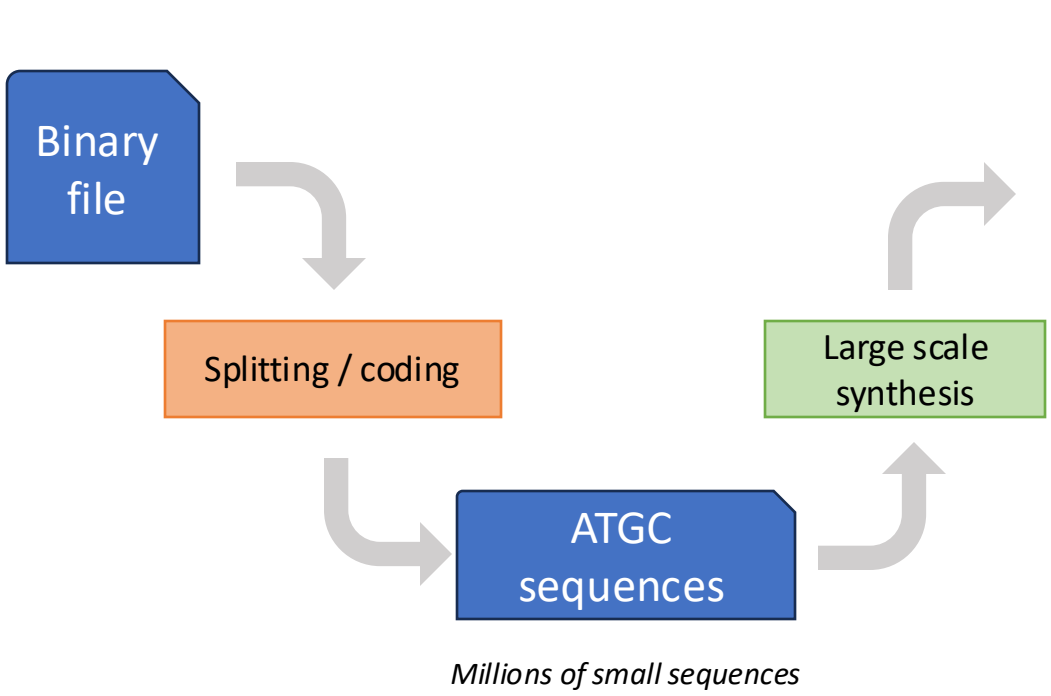
- Only short oligonucleotides can be produced
  - Single strand molecules of 20 to 300 nt
  - Quality decrease with the length
- Slow process
  - Phosphoramidite chemistry, enzymatic synthesis
  - A few tens of minutes to generate a single short oligo
- Costly
- Polluting chemistry

Miniaturization and Parallelism



Dedicated CMOS chips

- Millions of oligonucleotides can be built simultaneously



nucleotides are present in thousands of copies

Synthesis Errors

Truncated nucleotides
Some oligonucleotides will be missing

DNA Sequencing

DNA molecules


```
TGACCAGGAT
ATGAGAGCCA
GGATTAGGAC
CGGGGGAGAG
AGAGGAGACC
 TAGGAGATC
```

Text file
(billions of reads)

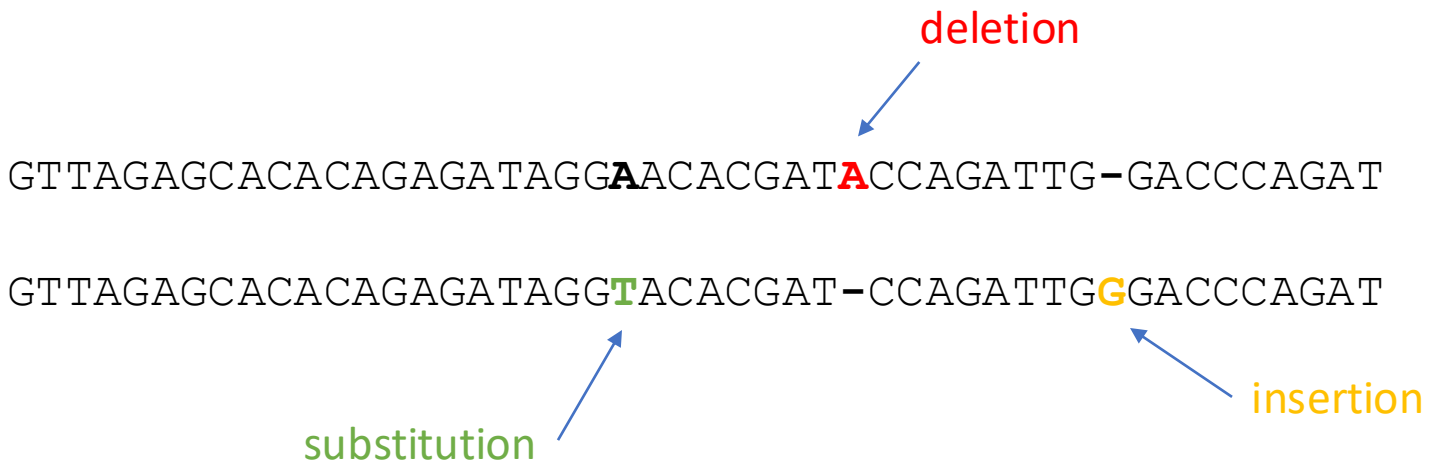
DNA sequencing

Random process

- Input = N sequences → output P sequences (P<<N)

Sequencing Errors

- Substitution
- Insertion / Deletion



Biochemical constraints

- No homopolymers
  - Both synthesis and sequencing manage poorly suite of identical nt

ATGGACACGATTTTTTTTTTGCCATATTACCAT

- Balance between the number of AT and CG in an oligonucleotide
  - GC% in the range [40%-60%]

Recap

Noisy channel

coding → Synthesis + storage + sequencing + random access → cleaning → decoding

Millions of sequences

```
GGTAGGACACAGGATAGGACA
CCATTAGGACACAGAGTTAGA
CCGATATAGGCCATAHACACC
. . .
GGATAGACACAGGAGGATTAG
```

Constraints:
Small DNA sequences
No homopolymers
40% < GC < 60%

Billions of sequences

```
GGTAGGACTCAGGATAGGA
CCAGTAGGACACAGAGTTAGA
CCGATAT-AGGCCATTACACC

. . .

GGATAGACACAGGAGGATTAG
```

Many Errors
substitution
insertion / deletion

Millions of sequences

```
GGTAGGACACAGGATAGGACA
CCATTAGGACACAGAGTTAGA
CCGATATAGGCCATAHACACC
. . .
GGATAGACACAGGAGGATTAG
```

Missing sequences
Few errors

Data cleaning

- Objective:
  - Select useful DNA fragments
    - Right length
    - No contaminated sequences
  - Correct as many errors as possible
- How:
  - By exploiting the redundancy of sequencing data

Processing of sequencing data

TGGCCACAGAGATGGATTAG
TTAGAGAGACAAGAGGACACG
GGATTATAGACCAATGGGACAC
GGATGAGACACAGATTAGGAG
TAAAAGATACCAATTAGGATAA
TTAGAGAGACAGAGGACCACG
GGATAAGACCAATCGGACAC
TGGCACACCGAGATGGATTAG
TTAGAGAGACAGAGGACACG
GGATATAGACCATTGGGACAC
TAAAGATACCAATTAGGTTAA
TGGCAGCCAGAGATGGATTAG
TTAGAGTGACCAGAGGCACACG
GGATATAGTCCAATGGGACAC
TAAAATACCAATTAGGATTAA
GGATACAGACCAATGCGACAC
TAATAGATACCAATTAGGATAA
TTAGAGAGACAGAGGTCCACG

CLUSTERING

TGGCCACAGAGATGGATTAG
TGGCACACCGAGATGGATTAG
TGGCAGCCAGAGATGGATTAG

TTAGAGAGACAAGAGGACACG
TTAGAGAGACAGAGGACCACG
TTAGAGAGACAGAGGACACG
TTAGAGTGACCAGAGGCACACG
TTAGAGAGACAGAGGTCCACG

GGATTATAGACCAATGGGACAC
GGATAAGACCAATCGGACAC
GGATATAGACCATTGGGACAC
GGATATAGTCCAATGGGACAC
GGATACAGACCAATGCGACAC

GGATGAGACACAGATTAGGAG

TAAAGATACCAATTAGGATAA
TAAGATACCAATTAGGTTAA
TAAATACCAATTAGGATTAA
TATAGATACCAATTAGGATAA

CONSENSUS

TGGC CA CAGAGATGGATTAG
TGGCACACC GAGATGGATTAG
TGGCAG CCAGAGATGGATTAG
TGGCACACCAGAGATGGATTAG

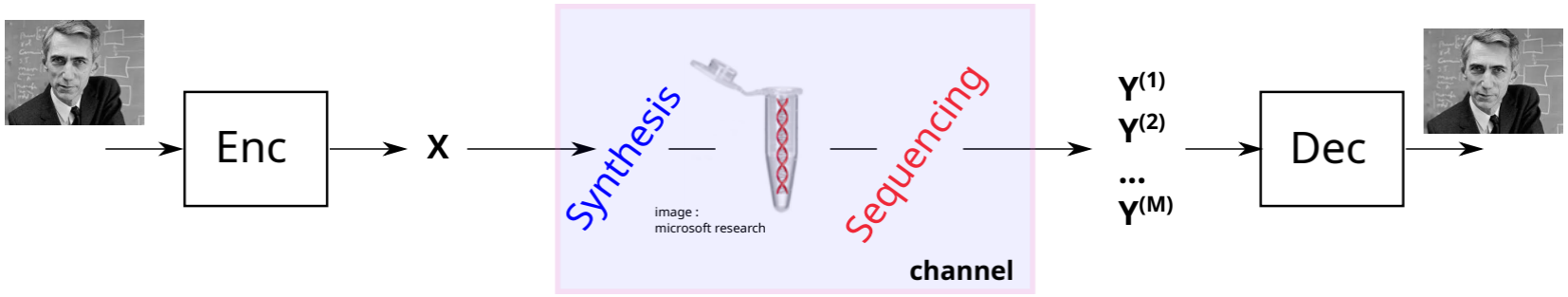
TTAGAGAGACAAGAGGAC ACG
TTAGAGAGACA GAGGACCACG
TTAGAGAGACA GAGGAC ACG
TTAGAGTGACCAGAGGCACACG
TTAGAGAGACA GAGGTCCACG
TTAGAGAGACAAGAGGTCCACG

GGATTATAGACCAATGGGACAC
GGATAA GA CCAATCGGACAC
GGA ATAGAACCATTGGGACAC
GGATATAGTACCAATGGGACAC
GGATACAGA CCAATGCGACAC
GGATACAGAGCCAATGCGACAC

GGATGAGACACAGATTAGGAG

TAAAGATACCAATTAGGA TAA
TAA GATACCAATTAGG TTAA
TAAA TACCAATTAGGATTAA
TATAGATACCAATTAGGAT AA
TAAAGATACCAATTAGGATTAA

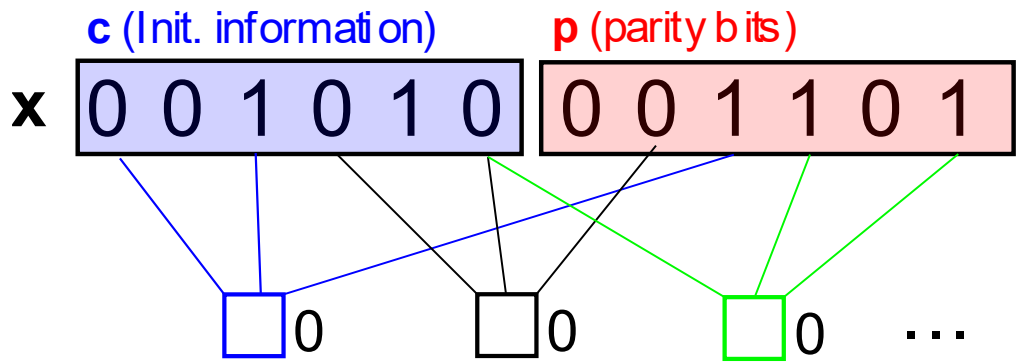
Channel model



- **Channel model?**
  - Useful for source/channel code design
  - Enable to run in-silico simulations before in-vivo experiments

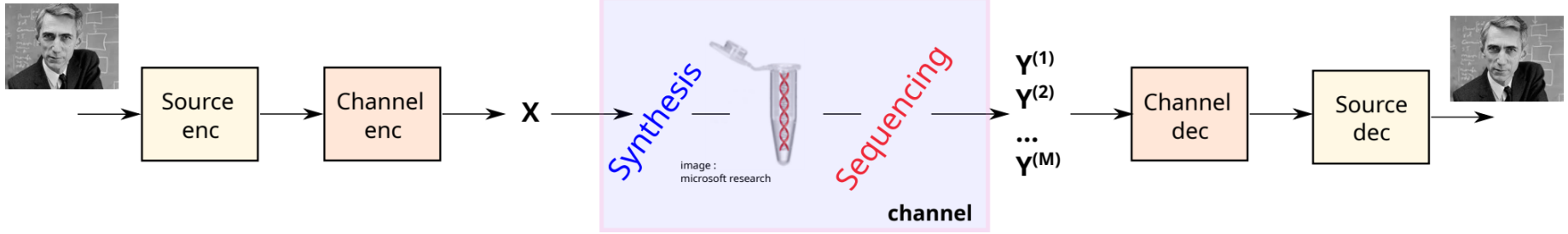
- **Families of solutions**
  - Model-based approaches P(Y|X)
  - Deep-Learning based approaches

Channel coding



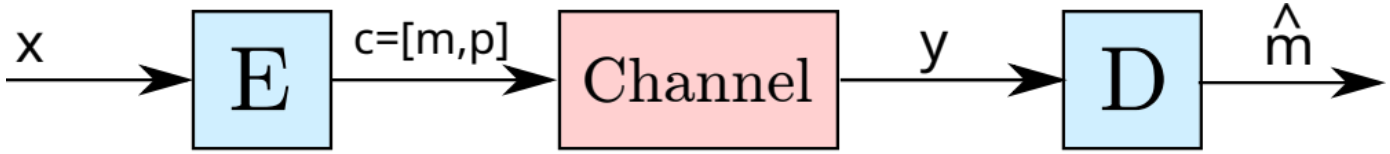
- **Challenges**
  - How can we design channel codes that can handle insertions and deletions?
  - How can we exploit multiple reads?
  - How can we combine bio-informatics methods with channel coding?

Source coding



- **Challenges**
  - How can we handle biocheminal constraints (no homopolymers, GC-AT balance)?
  - Can we design codes that allow for random access to the data?
  - Should be consider joint source-channel coding?

Information theory



Channel capacity: $C = \sup_{p_{\mathbf{X}}} I(\mathbf{X};\mathbf{Y})$

- **Challenges**
  - Can we evaluate the capacity of the DNA storage channel?
  - Can we use the information-theoretic analysis to improve code design?