



HAL
open science

Advanced Coding Schemes for DNA-Based Data Storage Using Nanopore Sequencing Technologies

Dominique Lavenier, Emeline Roux, Laura Conde-Canencia, Belaid Hamoum

► **To cite this version:**

Dominique Lavenier, Emeline Roux, Laura Conde-Canencia, Belaid Hamoum. Advanced Coding Schemes for DNA-Based Data Storage Using Nanopore Sequencing Technologies. Journées CominLabs 2019, Nov 2019, Rennes, France. 2019. hal-02400656

HAL Id: hal-02400656

<https://hal.science/hal-02400656>

Submitted on 9 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



DNA-store: ADVANCED CODING SCHEMES FOR DNA-BASED DATA STORAGE USING NANOPORE SEQUENCING TECHNOLOGIES

IAS, Lab-STICC, UBS

- L. Conde-Canencia
- B. Hamoum



GenScale, INRIA

- D. Lavenier
- E. Roux



Introduction and objectives

Design of coding schemes that allow information to be efficiently stored on DNA molecules, and read back using very low cost sequencing devices based on nanopore technologies.

1. The first part of the project focused on developing codes adapted to the nanopore sequencing constraints.
2. Demonstrate the feasibility of the approach by
 - (1) synthesizing DNA molecules encoded with the proposed coding schemes;
 - (2) sequencing those DNA molecules on a MinION device;
 - (3) correcting the output signal to retrieve the initial information..

Motivation

The current data explosion era is bringing new challenges in data storage technologies and leading research to emerging fields.

The need to explore innovative solutions is now undeniable because the available data storage systems have grossly been outpaced by the ever-increasing data generation.



Microsoft Data Center, Chicago

Principle and advantages



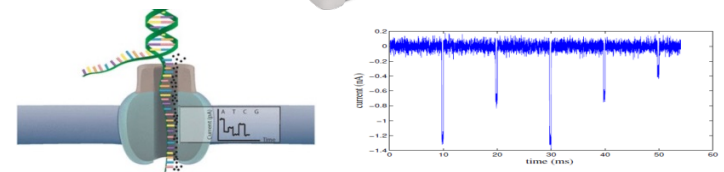
Manufacture DNA
Dehydrate & store
Read DNA

100101010

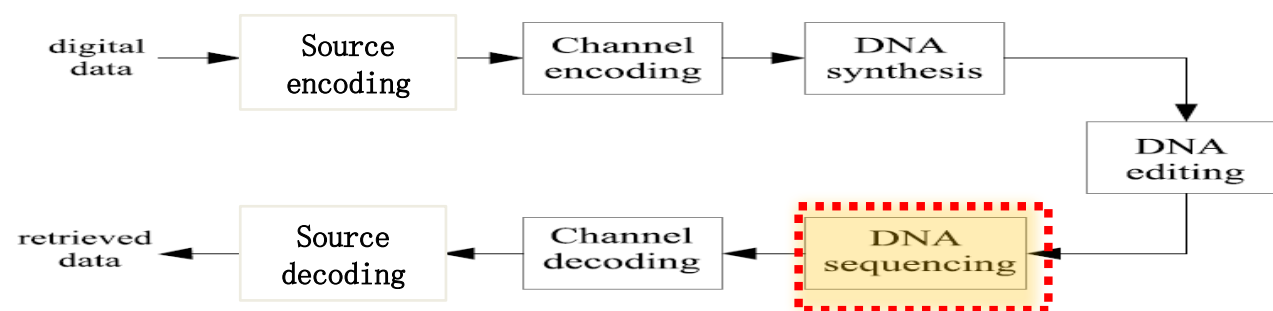
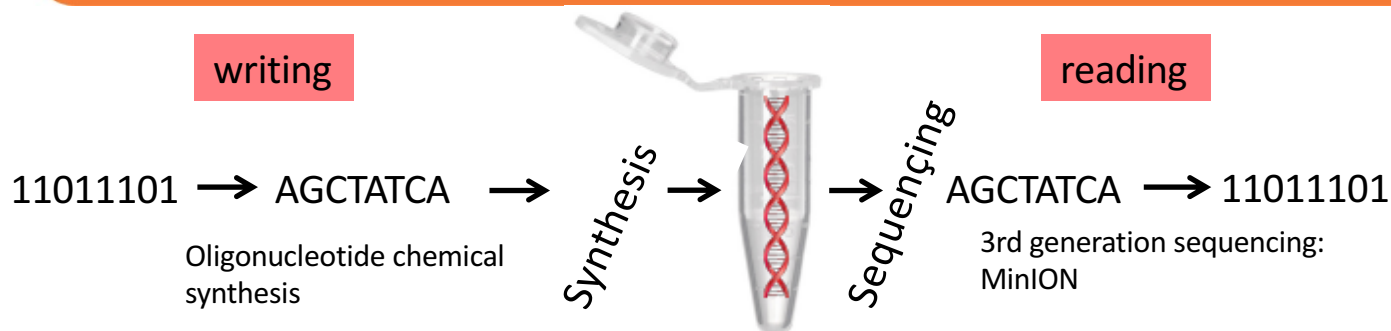
- 10^6 more compact than hard disks
- Capability for longevity > 100 years ...
- Resistance to obsolescence
- Potential revolution in data centers

The MinION: a portable real-time device for DNA sequencing

- ❑ 100 g
- ❑ USB plug
- ❑ ~ 1000 \$



DNA-based data storage chain

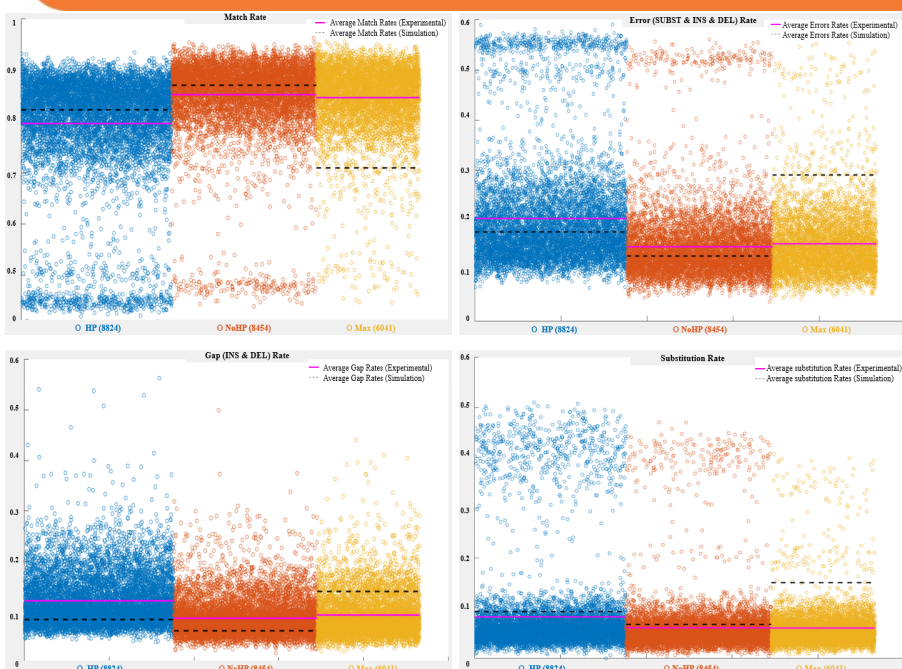


Source coding to avoid homopolymers and improve the MinION output signal

Synthesis and sequencing of real data sequences



Results: match rates and errors



- Simulation results obtained with deep learning tool: DeepSimulator
- Simulation vs. real synthesis and sequencing
- 3 types of errors: substitutions (or mismatches), insertions/deletions (*aka* gaps).
- 3 types of sequences:
 - “HP” sequence with homopolymers,
 - “NoHP” stands for the sequence without homopolymers
 - “Max” stands for the sequence without homopolymers and with maximized inter-k-mer distances (our proposed source coding)
- “NoHP” decrease error rates
- “Max” does not introduce gain

Conclusion and future work

We implemented a complete DNA-based data storage chain: we coded + synthesized + stored + retrieved the data. We analyzed error rates for 3 types of sequences to evaluate the effect of homopolymers and source coding.

This project was extremely enriching for the two partners. We have set the foundations for an interdisciplinary and fruitful collaboration. DNA-based data storage is a revolutionary technology but its success will only come by putting together the knowledge of bioinformatics, biology, information theory and coding.