Deep learning-aided error-correction for storage on synthetic DNA molecules

The SEED[1] program (academic track)

www.imt-atlantique.fr/seed

PhD topic open for applications until March 20, 2025

1 Definition **1**
- 1.1 Domain and scientific/technical context . . . 1
- 1.2 Scientific/technical challenges . . . 2
- 1.3 Considered methods, targeted results and impacts . . . 2
- 1.4 Environment (partners, places, specific tools and hardware) . . . 2
- 1.5 Interdisciplinarity aspects . . . 2
- 1.6 References . . . 3

2 Partners and study periods **3**
- 2.1 Supervisors and study periods . . . 3
- 2.2 Hosting organizations . . . 3
  - 2.2.1 IMT Atlantique . . . 3
  - 2.2.2 Laboratory for Integrated Micro-Mechatronics Systems (LIMMS) . . . 3

1 Definition

1.1 Domain and scientific/technical context

Data centers today represent around 20% of the energy consumption of digital technology in France. An alternative, the storage of information in synthetic DNA molecules, has been actively explored over the last few years. In addition to offering a storage density far superior to current technologies (up to several exabits per mm^3), DNA is a robust medium capable of withstanding significant temperature variations and it is durable over time. It is therefore expected to significantly reduce the energy consumption of data storage. Research on this area is multidisciplinary by nature, encompassing fields as diverse as biology, chemistry, bioinformatics, signal processing, and error-correction. This PhD falls into the fields of error-correction and Deep Learning. Due to the inherent unreliability of the DNA storage support, the goal will be to develop advanced deep learning models to ensure perfect data recovery after storage.

[1]Co-funded by the European Union under Grant Agreement no. 101126644

1.2 Scientific/technical challenges

A DNA molecule is composed of a sequence of nucleotides, or bases, of types A,C,G, T. DNA synthesis refers to the process of constructing the molecule corresponding to a given digital quaternary sequence. Synthesis is currently the main bottleneck of DNA storage technology. Indeed, while highly reliable due to its origins in the medical field, current synthesis techniques are slow and expensive. To enable DNA storage at scale, new synthesis technologies are under intensive development. These emerging techniques promise faster writing speeds. However, this will come at the cost of a decrease in reliability, as these techniques may introduce a non-negligeable amount of errors, between 5 and 10%, in the stored data. Therefore, addressing this reliability issue is critical to develop synthetic DNA as a large-scale storage medium.

1.3 Considered methods, targeted results and impacts

Channel coding is a technique that consists of introducing structured redundancies into data, which are then used during decoding for error-correction. Modern coding techniques, such as Turbo, LDPC, and Polar codes are essential components to current telecommunications standards as they ensure data reliability. However, DNA storage introduces specific errors, namely insertions and deletions, that break the channel codes structure of redundancy.

An opportunity for error correction arises from the fact that synthesis generates several copies of the DNA molecule, each containing different errors. Given that this problem is very complex to treat from standard error-correction tools , this PhD aims to develop deep learning models capable of jointly exploiting redundancy from both coding and synthesis to accurately reconstruct stored data. Achieving this would provide a scalable and reliable solution for DNA storage, paving the way for its large-scale adoption.

1.4 Environment (partners, places, specific tools and hardware)

The PhD will take place in the team CODES of Lab-STICC, a team specialized in channel coding which has already carried research activities on DNA data storage for several years. The team already has available channel simulators that mimic DNA data storage. This will be helpful to generate data for training Deep Learning models. The team has also developped several error-correction solutions for DNA data storage, which will provide baselines for the proposed Deep-Learning based solutions. The co-supervisor is Anthony Genot from the LIMMS. He comes from the field of biophysics, and he is actively working on DNA data storage, with a special focus on novel synthesis techniques. Therefore, he will bring a deep understanding on the synthesis aspects. Both supervisors are involved in the PEPR MolecularXiv (2022-2029), a French project to make DNA data storage a practical technology. The PhD student will be expected to interact with other teams of the PEPR.

1.5 Interdisciplinarity aspects

The topic is interdisciplinary by nature as it will require to rely on tools from signal processing, deep-learning, and error-correction, with an application into the field of bio-technology. The PhD student will mostly work on developping methods issued from error-correction

and Deep learning, while he/she will also need to understand (at least from a high level) the DNA storage process.

1.6 References

2 Partners and study periods

2.1 Supervisors and study periods

- **IMT Atlantique**: Assoc.-Prof. Elsa Dupraz, IMT Atlantique, Brest, France
- **International partner**: Dr. Anthony Genot and Assoc.-Prof. Kim Soo Hyeon, LIMMS, Tokyo, Japan.

  The PhD student will stay 1 year at the LIMMS.
- **Industrial partner(s)**: for short-term visits have not yet been determined. However, cooperations with non-academic partners on similar topics will be harnessed.

2.2 Hosting organizations

2.2.1 IMT Atlantique

IMT Atlantique, internationally recognized for the quality of its research, is a leading French technological university under the supervision of the Ministry of Industry and Digital Technology. IMT Atlantique maintains privileged relationships with major national and international industrial partners, as well as with a dense network of SMEs, start-ups, and innovation networks. With 290 permanent staff, 2,200 students, including 300 doctoral students, IMT Atlantique produces 1,000 publications each year and raises 18€ million in research funds.

2.2.2 Laboratory for Integrated Micro-Mechatronics Systems (LIMMS)

The LIMMS is an International Research Laboratory on MEMS and NEMS (Micro- and Nano-Electro-Mechanical Systems) between the CNRS and the University of Tokyo's Institute of Industrial Science. It is located in Komaba Campus, Tokyo, Japan. It was created in 1995 and became a laboratory in 2004. Its research activities are focused on three main microsystems related fields: Biology, Energy, and Quantum & Molecular Tech.