PhD Thesis Topic: Error Correction for Data Storage in DNA Molecules

- **Thesis in Channel Coding and Computer Science (no prior knowledge in biology required)**
- Expected to start in Fall 2025, at IMT Atlantique in Brest, France
- Joint supervision between IMT Atlantique/Lab-STICC (Brest) and IRISA (Rennes)

I. Data Storage in Synthetic DNA Molecules

Data centers currently account for about 20% of digital energy consumption in France. An alternative, storing information in synthetic DNA molecules, has been actively explored in recent years. In addition to offering a storage density far superior to current technologies (up to several exabits per mm³), DNA is a robust medium capable of withstanding significant temperature variations and is durable over time. It should therefore allow for long-term information preservation and significantly reduce the energy consumption of storage.

A DNA molecule is composed of a sequence of nucleotides, or bases, of types A,C,G, and T. DNA synthesis refers to the process of constructing the molecule corresponding to a given digital quaternary sequence. Currently, the synthesis operation represents the main bottleneck of this technology, as it is slow and costly, although very reliable since it was originally dedicated to the medical field. The reading of the information is then done through a sequencing operation, a technique that introduces a significant proportion of errors (about 5%) in the sequenced data.

II. Error correction

Channel coding involves introducing structured redundancies into the data, which will be exploited during decoding to correct errors introduced during the transmission or storage of data. Modern channel coding solutions such as Turbo codes, LDPC codes, or Polar codes are an essential component of most telecommunications standards (Wi-Fi, mobile radio, etc.) and information storage components (RAM, hard drives, etc.), as they make data transmission and storage reliable. However, data storage in DNA introduces errors (insertions, deletions) that conventional channel codes cannot correct because these errors disrupt their redundancy structure.

That said, an interesting opportunity from the perspective of error correction lies in the fact that sequencing naturally produces a large number of reads of the same molecule, with different errors in each read. A solution from the field of bioinformatics involves using consensus algorithms to reconstruct the input sequence from multiple reads. In this thesis, the idea will be to develop hybrid approaches combining these two complementary solutions (consensus algorithms and channel coding) to more efficiently reconstruct the input data by exploiting both multiple reads and code redundancies.

III. Thesis Environment

The thesis will be carried out within the framework of the PEPR MolécularXiv (see https://pepr-molecularxiv.fr/le-pepr/). The doctoral student will be affiliated with the MEE department of IMT Atlantique in Brest and will also work with the GenScale team at IRISA/INRIA Rennes. This thesis is intended for students with a Master degree, an Engineering degree, or equivalent, who have followed a curriculum in computer science, telecommunications, or signal processing. Prior knowledge of channel coding would be a plus. However, it is not necessary to have prior knowledge of biology to work on this topic.

IV. How to apply

To apply, contact Elsa Dupraz (elsa.dupraz@imt-atlantique.fr) and Dominique Lavenier (dominique.lavenier@irisa.fr), and include the following: CV, academic transcripts, a few lines in the body of the email explaining your interest in this topic, and contact information for one or two referees (internship supervisors, etc.).