# PhD Position: Design of Efficient Error-Correction Codes for DNA Data Storage

MEE department of IMT Atlantique, Team CODES of Lab-STICC, Brest, France
Expected starting date: September 2022 (for a 36 month duration)

## I. Topic

Data storage on DNA molecules is perceived as an emerging and promising technology which should allow for highly increased density and durability compared to conventional storage techniques (HDD, SDD, etc.). The disruptive idea of this technology, which appeared for a long time to be a distant hope, is to build synthetic DNA sequences encoding some relevant information, see https://www.youtube.com/watch?v=r8qWc9X4f6k for a brief introduction.

Recent major improvements in both DNA synthesis and sequencing techniques made DNA data storage affordable, although these techniques still introduce a large amount of errors in the read data sequences. While conventional storage systems are mostly concerned with substitution errors, DNA storage also introduces deletions and insertions which standard Error-Correction (EC) solutions (LDPC, Turbo, Polar, etc.) cannot handle [1].

DNA sequencers output a large number of copies of the same input sequence, with different error realisations. Recently developed EC solutions for DNA storage efficiently exploit those multiple reads so as to correct both substitution, insertion, and deletion errors [2], [3]. However, these solutions assume unrealistic independent and identically distributed (i.i.d.) error models, and may therefore show poor performance in practice. Alternatively, some other works consider advanced bio-informatics techniques such as consensus algorithms so as to correct most insertion and deletion errors, before applying a standard EC solution to correct remaining substitution errors [4], [5]. Although working well in practice, these pragmatic solutions do not exploit any knowledge of the error model in the consensus part, thus leaving a doubt on their optimality.

## II. Objectives of the PhD

In [6], we developed a channel model for DNA storage, which accurately captures error dependency to the successive bases of the read sequences, as well as memory within errors introduced by the sequencer. This model could be used in a first design step before testing the developed EC solutions under costly in-vivo experiments. Therefore, the three main objectives of the PhD will be as follows:

1) Benchmark current existing EC solutions for DNA storage under the channel model of [6], so as to evaluate their efficiency under more realistic error models.
2) Develop efficient EC solutions for DNA storage, which exploit as much as possible the knowledge of the statistical channel model of [6], while keeping a reasonable complexity. At this step, two key questions will be: (i) do we need a two-step coding approach like most current existing solutions? and (ii) how to address the tradeoff between number of read sequences and coding rate?
3) Experimentally validate the developed EC solutions in a full end-to-end DNA storage system, by relying on the expertise and means of the MolecularXiv project.

## III. Context of the work

The PhD will be realized in the context of the PEPR (Projet Exploratoire de Recherche) MolecularXiv. This French project will involve many researchers working in various area: biology, bioinformatics, signal processing, etc. Although the focus of the PhD will be on coding and information theory, the candidate should expect some

interactions with researchers working on the other fields. Feel free to check https://project.inria.fr/dnarxiv/ which presents a prior project on this topic.

## IV. How to apply

The candidate should have earned an MSc degree, or equivalent, in one of the following fields: telecommunications, information theory, applied mathematics, signal processing. To apply, please contact Elsa Dupraz (elsa.dupraz@imt-atlantique.fr), and attach the following:

- Full CV with a list of projects and courses related to the subject of the PhD
- Complete academic record (from bachelor to MSc)
- 1 or 2 reference contacts (former or current internship advisor, teacher, etc.)

## References


[1] H. Mercier, V. K. Bhargava, and V. Tarokh, "A survey of error-correcting codes for channels with symbol synchronization errors," *IEEE Communications Surveys & Tutorials*, vol. 12, no. 1, pp. 87–96, 2010.

[2] Y. M. Chee, H. M. Kiah, and T. T. Nguyen, "Linear-time encoders for codes correcting a single edit for DNA-based data storage," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 772–776.

[3] A. Lenz, I. Maarouf, L. Welter, A. Wachter-Zeh, E. Rosnes, and A. G. i Amat, "Concatenated codes for recovery from multiple reads of DNA sequences," in *2020 IEEE Information Theory Workshop (ITW)*, 2021, pp. 1–5.

[4] S. Chandak, K. Tatwawadi, B. Lau, J. Mardia, M. Kubit, J. Neu, P. Griffin, M. Wootters, T. Weissman, and H. Ji, "Improved read/write cost tradeoff in DNA-based data storage using LDPC codes," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2019, pp. 147–156.

[5] S. H. T. Yazdi, R. Gabrys, and O. Milenkovic, "Portable and error-free DNA-based data storage," *Scientific reports*, vol. 7, no. 1, pp. 1–6, 2017.

[6] B. Hamoum, E. Dupraz, L. Conde-Canencia, and D. Lavenier, "Channel model with memory for DNA data storage with nanopore sequencing," in *2021 11th International Symposium on Topics in Coding (ISTC)*, 2021, pp. 1–5.