A DNA Data Storage Channel Model Trained on Genomics Data with Nanopore Sequencing

Belaid Hamoum[1], Elsa Dupraz[2], Laura Conde-Canencia[1]

[1]Lab-STICC, CNRS UMR 6285, Universite Bretagne-Sud, Lorient, France
[2]IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, France.

Abstract

Data storage on DNA molecules recently appeared has a promising emerging technique, as it could allow for higher density, increased durability, and reduced energy consumption compared to conventional data storage solutions. Recent improvements on DNA synthesis and sequencing techniques will make DNA data storage more affordable, especially with the emergence of the MinION nanopore sequencer. The MinION sequencer sequentially reads k bases, called a k-mer, at a time. But as a main drawback, the MinION sequencer introduces about 10% of insertion, deletion, and substitution errors, in the read sequences. These errors can be corrected by combining powerful bio-informatics algorithms with efficient channel codes [1].

Ideally, the design of errors-correction solutions should first be investigated from numerical simulations (in-silico), and then validated through experiments (in-vitro), the latter still being a costly and time-consuming process. Existing DNA storage channel simulators are based on independent and identically distributed (i.i.d.) models [2], or rely on black-box Deep Learning techniques [3], or depend on empirical parameters to be set up by the user [4]. Comparison of these models with experimental datasets show that they do not represent accurately bursts of errors introduced by the nanopore sequencer [5]. Alternatively, in [6], we introduced a novel DNA data storage channel model that relies on Markov models with memory so as to accurately capture the statistical dependency between read k-mers and bursts of sequencing errors. However, in [6] our model was trained on a small amount of experimental data at our disposal. Although this small dataset was shown to lead to a more accurate model than existing solutions, there is still a risk that it introduces undesired bias due to insufficient k-mers coverage.

In this talk, we introduce a novel methodology so as to train our channel model on genomics dataset, which provide a significantly larger amount of data compared to our prior experimental dataset. In particular, we discuss the choice of efficient alignment techniques for genomics data, and address the key issue of selecting only relevant genomics reads so as to accurately train the model. We also present the results of training our model onto the genome of the streptococcus thermophilus bacteria. In future works, we aim to rely on our channel model so as to develop efficient and accurate source and channel coding solutions for DNA data storage.

References

[1] 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," 2020.

[2] W. Song, K. Cai, M. Zhang, and C. Yuen, "Codes With Run-Length and GC-Content Constraints for DNA-Based Data Storage," *IEEE Communications Letters*, vol. 22, no. 10, pp. 2004–2007, 2018.

[3] S. Chandak, J. Neu, K. Tatwawadi, J. Mardia, B. Lau, M. Kubit, R. Hulett, P. Griffin, M. Wootters, T. Weissman, and H. Ji, "Overcoming High Nanopore Basecaller Error Rates for DNA Storage via Basecaller-Decoder Integration and Convolutional Codes," in *ICASSP*, pp. 8822–8826, 2020.

[4] R. R. Wick, "Badread: simulation of error-prone long reads," *Journal of Open Source Software*, vol. 4, no. 36, p. 1316, 2019.

[5] R. Heckel, G. Mikutis, and R. N. Grass, "A Characterization of the DNA Data Storage Channel," *Scientific Reports*, vol. 9, no. 1, p. 9663, 2019.

[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)*, pp. 1–5, 2021.