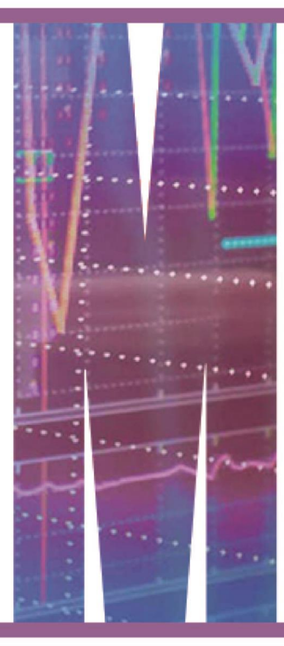
CODING PROBLEMS IN DNA STORAGE WITH NANOPORE SEQUENCING

Author:
Preet Patel

Supervised by:
Professor E. Viterbo



ECSE

INTRODUCTION

- Writing to a DNA strand is an up-and-coming method for data storage.
- Currently a very expensive process.
- Cheaper technologies like nanopore sequencing are becoming prominent.
- DNA strand passes through a microscopic pore, creating a current signal which can be analysed.
- This method is prone to errors and therefore any message would need to be encoded before being transmitted through this channel.

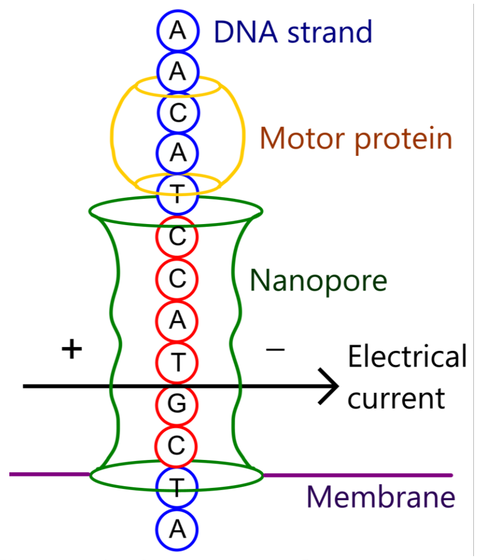


Figure 1: Anatomy of a nanopore.

AIM

The purpose of this project is to investigate the viability of encoding using codons, which are 3 consecutive bases in a natural DNA strand that code for an amino acid. We investigate whether the popularity, GC-content and homopolymerity of these codons impact the quality of a codebook. With these findings, we create an optimal codebook of codons.

METHODS

Experiment 1

- The quality of a codebook is measured via its quality score, Q, which is related to the error rate e by

$$Q = -10 \log_{10} e.$$

- Errors in a nanopore channel include substitution, insertion and deletion errors, as there is no mechanism for synchronisation.

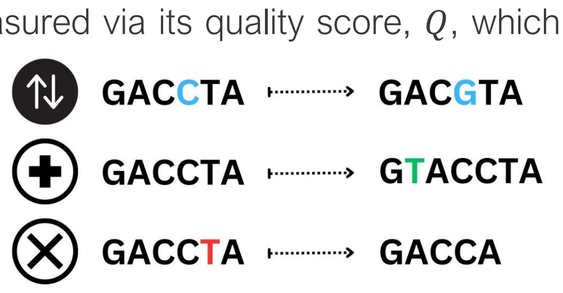


Figure 2: Substitution, insertion and deletion errors.

- Experiment 1 investigates how this quality score varies across different pores and basecalling models, such as Dorado and Scrappie. Four preliminary codebooks are tested.
- This requires building a pipeline to simulate the synthesiser, nanopore and basecaller.

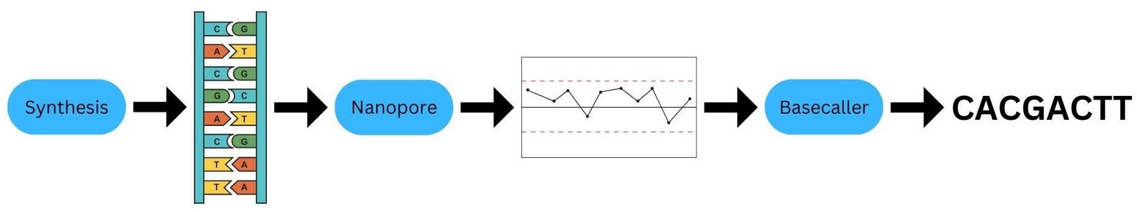


Figure 3: Three main stages of DNA storage with nanopore sequencing.

Experiment 2

- We define three codebook metrics and observe how these impact its quality.
- The first metric is the popularity $\mathcal{p}$, which measures how frequently the codons in the codebook appear in human DNA.
- The second metric is the GC-content $\mathcal{g}$, which measures the proportion of bases which are G or C. Previous research with Guppy found that lower GC-content improves quality score [1].
- The third metric is the homopolymerity $\mathcal{h}$, which measures how often repeated bases (i.e. AA) occur.

Experiment 3

- The final experiment is disjoint from the previous two. It investigates the quality score and errors present in real reads from an R10 nanopore.
- 51614 reads were provided, and the task was to classify them to one of 16 possible reference sequences, by finding the minimum edit distance.
- We investigate methods to reduce classification error.

RESULTS AND DISCUSSION

Experiment 1 produced several performance plots and gave a preliminary indication that certain aspects of codebooks do impact the quality score.

In experiment 2, we found that basecalling strongly favours codebooks with very high or low GC-content, and performs worst when $\mathcal{g}$ is between 40-50%, as is the case for the human genome [2].

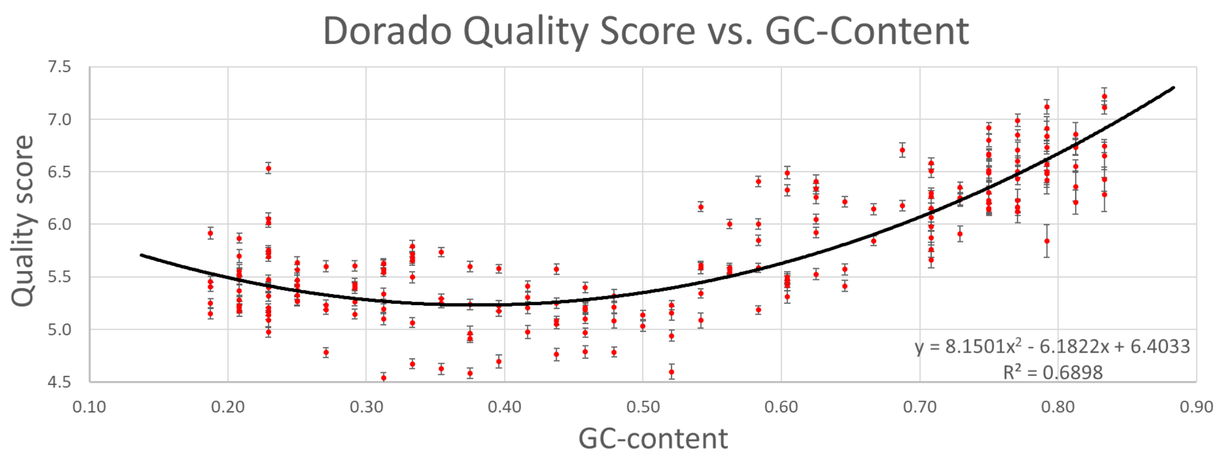


Figure 4: Dorado quality score vs. GC-content for 200 codebooks, with 400 transmissions simulated for each codebook.

Similar analysis was conducted for popularity and homopolymerity, each time finding mild or moderate linear correlation. The fits suggested by the data are given by

$$\hat{Q}_{\mathcal{p}} = 0.56\mathcal{p} + 4.62$$

$$\hat{Q}_{\mathcal{g}} = 8.2\mathcal{g}^2 - 6.2\mathcal{g} + 6.4$$

$$\hat{Q}_{\mathcal{h}} = -2.6\mathcal{h} + 6.1$$

The metrics $\mathcal{g}$ and $\mathcal{h}$ were found to have the strongest correlation coefficients, and 400 codebooks were made favouring these metrics. The highest performing codebook had a quality score of 8.1, equating to half as many errors as an average codebook which is a significant improvement.

In experiment 3, the 51614 reads were basecalled and classified, with a classification accuracy of 98.97%. This can be improved by discounting reads whose edit distances to the two closest reference sequences are comparable. We lose more reads as the discounting threshold is raised but obtain a higher classification accuracy.

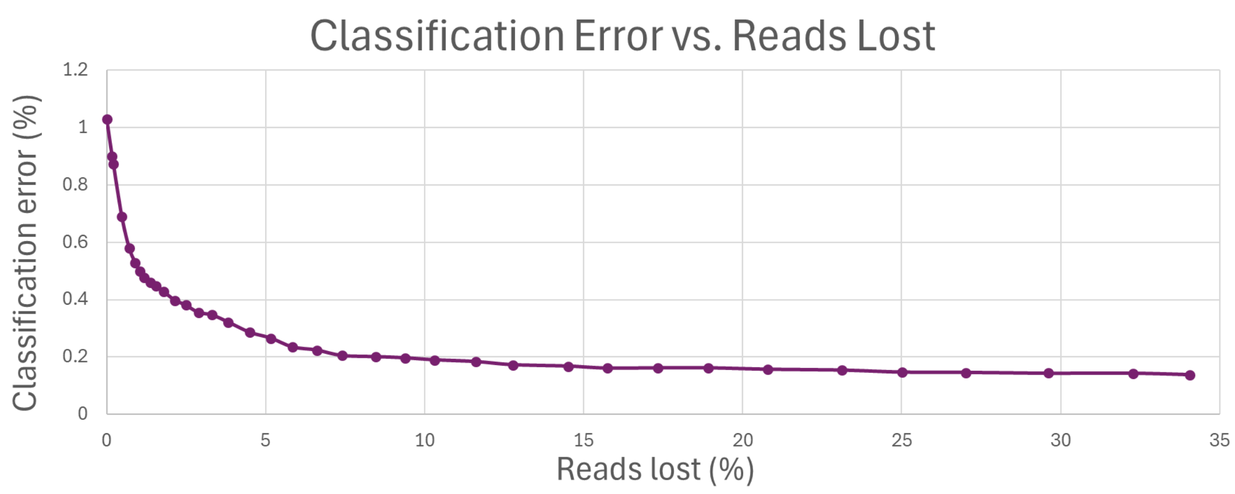


Figure 5: Classification error vs. reads lost.

CONCLUSIONS

- Most codebooks of codons yield moderate quality scores.
- GC-content and homopolymerity have significant impact on a codebook's quality score. These may benefit designers of more general codebooks.
- Classification error of real reads is very low and can be improved by omitting poor reads.

REFERENCES

[1] Borisova et al., "Relative stability of AT and GC pairs in parallel DNA duplex formed by a natural sequence," *FEBS Letters,* vol. 322, no. 3, pp. 304-306, 1993.

[2] Piovesan et al., "On the length, weight and GC content of the human genome," BMC Research Notes, vol. 12, no. 106, 2019.