Monitoring and Controlling Nanopore Sequencing Runs

Matej Fedor[1]
Supervisor: Tomáš Vinař[1]
[1] Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia

Introduction

- Modern nanopore sequencers provide users with the option to **decide if a DNA sequence is rejected or sequenced**. Decisions are made based on the raw nanopore signal in real time.
- This feature of nanopore sequencers is called selective sequencing. One can adaptively sample the desired DNA sequences and reject the undesired ones. **Thus achieving higher sequencing coverage in a shorter time and at a lower cost**. The potential coverage gain achieved by adaptive sampling increases with the decision speed.
- Typically, the similarity between reads and desired reference genomes is evaluated to make a decision. The need for fast decisions about a read's biological origin based on its small portion constitutes a difficult problem in the field of bioinformatics.
- Developing an efficient decision-making algorithm is a very difficult task. **One must conduct numerous experiments using a physical sequencer to test and calibrate the algorithm before its deployment**. We find that the ongoing research in the area might be hampered by its high cost and the need for advanced expertise in both the fields of biology and informatics.

Virtual Sequencing

- **We introduce a virtual sequencer** to facilitate the development and testing of decicision-making algorithms. It emulates the real MinION nanopore sequencer, together with its selective sequencing capabilities.
- We use stored data from previous sequencing runs to emulate them as they originally took place. In addition, **we allow a decision-making algorithm to intervene by rejecting DNA sequences**. The decisions of the algorithm impact the emulated run. The user can observe the effects of the adaptive sampling strategy such as **coverage gain, decision accuracy, mean length of desired reads** and others.
- **We increase time and cost efficiency of decision-making algorithm development**. We make selective sequencing experiments accessible by avoiding the need for a physical sequencer. We significantly reduce the required expertise in the field of biology.
- We connect a **well known adaptive sampling tool** Readfish [Payne et al.,] to the virtual sequencer and attempt to enrich the *Saccharomyces cerevisiae* in the ZymoBIOMICS sample during an emulated sequencing run.

Attributes	Original Run	Adaptive Sampling Run
Desired read count	4.31k	14.35k
Undesired read count	189.11k	635.27k
Desired Avg. Length	3562.47b	3459.40b
Undesired Avg. Length	3725.40b	663.10b
Desired Bases	15.35M	49.67M
Undesired Bases	704.52M	421.26M
Coverage gain	1.00x	3.24x

Table 1: Demonstrating the use of the virtual sequencer with Readfish

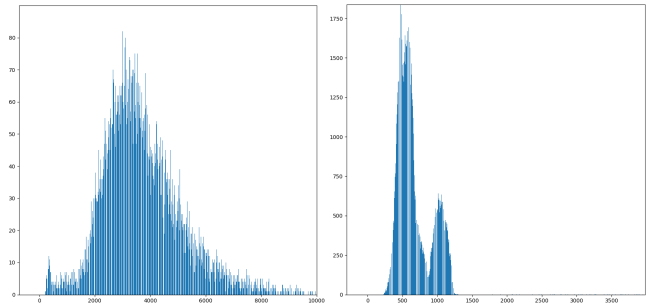

Figure 1: Undesired read length distribution in the original sequencing run (left side) and the adaptive sampling run (right side).

Our Adaptive Sampling Tool

- We utilize the virtual sequencer in the development of our decision-making algorithm.
- We explore the use of machine learning for adaptive sampling task. We study **if we can trade the generality of the decision-making algorithm for its speed**. We develop an algorithm named Selectify specialized for use with the *SARS-CoV-2* reference genome.
- We propose a **compact Convolutional Neural Network design** and a training method. To our knowledge, we are the first to test such a machine learning model in realistic conditions due to our access to the sequencing emulator.
- We manage to boost decision-making performance, **requiring less time and a smaller portion of sequenced read** to make a decision compared to the Readfish tool. Selectify **classifies the desired *SARS-CoV-2* reads with satisfactory accuracy**, but proves to be **unable to generalize the information about diverse biological backgrounds** in the sequenced samples that is not completely covered by the training dataset.
- The lack of generalization capabilities makes our model unsuitable for use with diverse and fast-evolving clinical samples researched in the **area of epidemiology**.

Attributes	Readfish	Selectify
Accuracy	86.58%	75.27%
Sensitivity	98.78%	91.00%
Specificity	37.90%	8.13%
Acceleration	1.00x	1.66x

Table 2: Classification measurements of the Readfish and Selectify adaptive sampling tools

References

[Payne et al.,] Payne, A., Holmes, N., Clarke, T., Munro, R., Debebe, B., and Loose, M. Nanopore adaptive sequencing for mixed samples, whole exome capture and targeted panels. https://www.biorxiv.org/content/10.1101/2020.02.03.926956v2.abstract.