Outline of Research

Sasha Jenner

May 2, 2022

1 Research Problem

The primary research problem is to develop an improved lossless encoding strategy for nanopore signal data. This requires some unpacking and can be sub-divided into the following smaller research problems.

1.1 Determine the salient features of nanopore signal data

The first problem is to determine the features and repeated patterns of nanopore signal data. Understanding the data is the most crucial step in developing a data-specific lossless encoder. Once the features have been recognised, they can be directly exploited by an encoding strategy in order to reduce the data's redundancy and bit rate. The intention is to encode each nanopore *read* separately in order to maintain random parallel access. With this in mind, there are potentially patterns between independent reads which should first be determined.

1.2 Apply and evaluate appropriate existing encoding strategies

The next problem involves experimenting with and evaluating existing lossless encoding strategies which are suitable for nanopore signal data. Most likely, there already exists a strategy in the literature which when applied to nanopore data would prove better than the state-of-the-art. This problem involves further investigation into the literature of integer and signal data compression. It also requires developing a iterative evaluation framework which can quickly determine whether an existing strategy implementation is worth further investigation. Furthermore, it may prove that an existing

strategy forms the basis for an improved encoder when modified specifically with the salient features of nanopore signal data in mind.

1.3 Develop a new lossless algorithm (or modify an existing one) which is better than VBZ

This is the final problem and main contribution I intend to make. It is clear from the literature review that few lossless encoders specific to nanopore signal data have been considered. A new lossless algorithm which is better than the state-of-the-art known as VBZ (`https://github.com/nanoporetech/vbz_compression/`) would significantly alleviate the storage issues and/or long analysis times faced by the nanopore sequencing community.

2 Evaluation Process

A rigorous evaluation process which determines what constitutes a 'better' lossless encoding strategy must be outlined. There are many ways of evaluating an encoder. Fundamentally, these include the algorithm's

- data compression ratio;
- space and
- time complexity; and
- passes over the data.

The data compression ratio is a measure of how well the algorithm compressing the input data. It is simply defined as

$$\text{Compression Ratio} = \frac{\text{Uncompressed Size}}{\text{Compressed Size}}.$$

The space complexity of the algorithm measures how much memory it requires. Traditionally, this is expressed using big O notation but a more practical hard limit should be imposed. Considering the size of a nanopore read is usually around 256KB when stored in uncompressed binary, space is not usually a problem. In this case, streaming should not be a requirement on the algorithm and $O(n)$ is fine.

Similarly, the time complexity of an algorithm measures how many computational operations it performs. This is also often described in big O notation as a function of the size of the input data. A similar more practical

measure is the number of passes over the data or simply the time taken since most encoders are $O(n)$ anyway. One pass is optimal for nanopore signal data since it contains some degree of randomness. However, several passes may not be much slower and needs to be properly investigated.

In addition, since the compression ratio and time taken is a trade-off for typical data compression techniques. A composite measure *compression-time ratio* is introduced as

$$\text{Compression-Time Ratio} = \frac{\text{Compression Ratio}}{\text{Time Taken}}.$$

In this case, if the compression ratio and time taken are multiplied by a constant factor, the compression-time ratio remains constant. For example, an improvement in the compression ratio by a factor of two can be compensated by up to a doubling of the time taken before the compression-time ratio decreases. This is not an ideal measure but is useful for the comparison of techniques especially as a function of some tunable parameter such as the compression level.

All of these metrics apart from the compression ratio may be different for compression and decompression and should be evaluated separately. In particular the following criterion is used to determine a 'better' encoding strategy.

1. The strategy must be lossless.
2. The best compression ratio is higher than that of VBZ on average across datasets.
3. The compression-time ratio is higher than that of VBZ on average across both datasets and compression levels for both compression and decompression.
4. $O(n)$ space complexity.

The datasets used should represent typical nanopore signal data from a range of Oxford Nanopore machines; biopolymers such as DNA and RNA; and biological species. It should be large enough, at least several datasets, to conclude confidently the performance of the encoding strategies being tested.