Efficient Encoding and Compression Techniques for the SLOW5 File Format

Sasha Jenner

November 9, 2020

Contents

List of Figures

1 Abstract

1.1 Background

Contemporary data storage of raw nanopore signals in the FAST5 file format doesn't benefit from parallel file access. A more computationally resourceful and space efficient file format could result in significant improvements in the runtime and storage size of nanopore sequencing pipelines.

1.2 Methods

Binary and compressed binary equivalents to the existing SLOW5 format were created. The binary SLOW5 format was created by transforming the human-readable data of the

SLOW5 format to its binary equivalent. In-file gzip compression of the binary SLOW5 format was performed per nanopore read entry. The resulting file is gzip-compatible.

Benchmarking of the access time and file size was performed for each SLOW5 format and their corresponding FAST5 files.

1.3 Results

1.4 Conclusion

2 Introduction

DNA sequencing devices from Oxford Nanopore Technologies (ONT) record disturbances in ionic current as DNA molecules are passed through a biological nanopore. These measurements can be translated to determine the sequence of each DNA molecule analysed.

The time series current signal data is written in a format called FAST5, which stores the raw data for each nanopore sequencing read. FAST5 is a Hierarchical Data Format 5 (HDF5) file with a specific schema defined by ONT. HDF5 is a complex file format for storing and managing high volume data. It works well with the time series current signal of a nanopore read, since it uses B-trees to index table objects.

However, the official library used to access the HDF5 file format is not scalable with threads. This means that tasks performed with the FAST5 file format suffer from an inefficient utilisation of parallel resources. Therefore, the process of 'basecalling' FAST5 data into DNA sequence reads (in FASTQ format) and other common analyses that utilise signal-level data (such as DNA methylation calling) are slow and costly.

A new file format is needed to improve the efficiency of these bioinformatics tasks. SLOW5 is a simple file format designed to address this. It is a tab-separated values (TSV) file with a header marked by lines beginning with a '#', followed by the time series current signal data for one nanopore read per line. In order to perform multi-threaded access to the SLOW5 file, the SLOW5 index is used. It is also a TSV file and begins with a single header line prepended by a '#' to define the column names and their order. Each following line represents the location of a particular read in the corresponding SLOW5 file.

In the current readable format, a SLOW5 file containing multiple reads takes up much more memory than its corresponding FAST5 files, each containing one read. Thus, the goal is to efficiently compress the SLOW5 file format to a size smaller than or equal to the total size of its corresponding FAST5 files. Whilst still maintaining multi-threaded access to the SLOW5 file format and achieving the best possible level of performance with these constraints.

When integrated with popular third part tools for nanopore sequencing analysis (e.g. Bonito, Nanopolish and F5C), we anticipate that SLOW5 format will deliver considerable performance advantages that will scale with the number of allocated threads. Given these advantages, the SLOW5 format should be readily adopted by the Nanopore sequencing community.

3 Methods

Conversion of SLOW5 entry to binary form Compress each line of SlOW5 in binary Index file for each step

4 Results

5 Discussion

6 Conclusion

7 References