Impact of lossy compression of nanopore raw signal data on basecalling and consensus accuracy

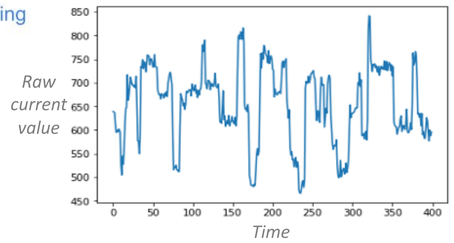
Shubham Chandak, Kedar Tatwawadi, Srivatsan Sridhar, Tsachy Weissman
Department of Electrical Engineering, Stanford University

Motivation

Nanopore Sequencing

- Portable
- Real time
- Long reads (~10kb-100kb)
- Native DNA & direct RNA sequencing
- High error rates (5-10%)

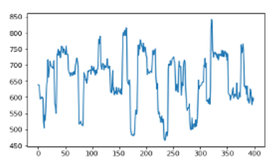
Raw signals in HDF5 files need **1 TB** storage for 30x human WGS nanopore sequencing *after* using state-of-the-art lossless compressors.



Need to retain raw signal to allow reanalysis with future basecallers and other tools

Lossy compression

$x_1, x_2, \ldots, x_n$ → Compress → *compressed bitstream* → Decompress → $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$



Error constraint: $\max_{i=1,\ldots,n} |x_i - \hat{x}_i| \leq \epsilon$

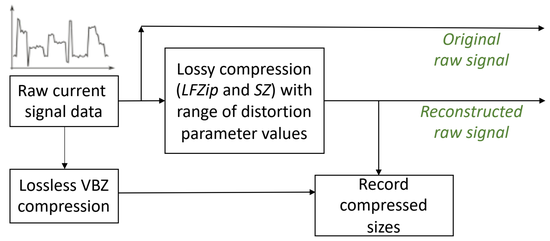
Maximum absolute error

But the actual loss metric is the downstream accuracy

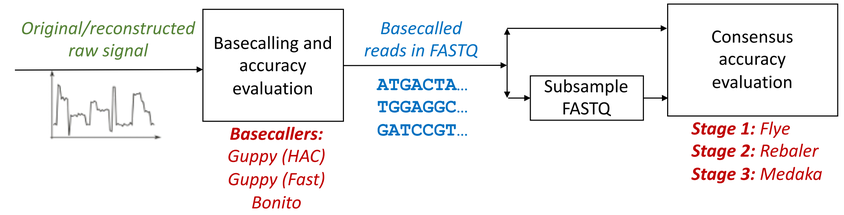
Reference

Shubham Chandak, Kedar Tatwawadi, Srivatsan Sridhar, Tsachy Weissman, Impact of lossy compression of nanopore raw signal data on basecalling and consensus accuracy, ***Bioinformatics***, Volume 36, Issue 22-23, 1 December 2020, Pages 5313–5321.

Evaluation pipeline



Lossless and lossy compression of raw signal data

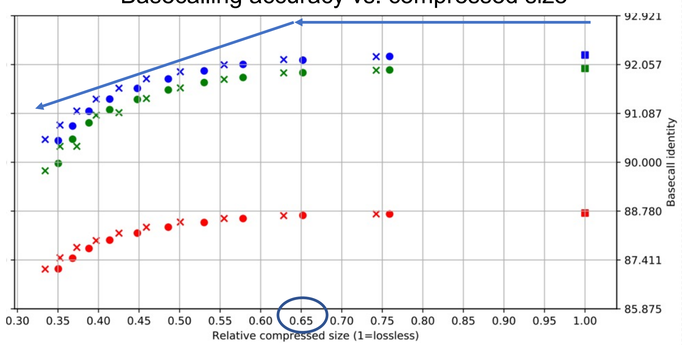


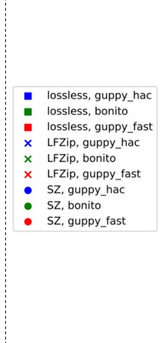
Basecalling and consensus accuracy analysis

4 datasets with high quality ground-truth genome (3 bacterial + 1 human)
Evaluate basecalling and consensus accuracy for lossless and lossily compressed raw data
Test multiple datasets, basecallers, subsampled coverage, assembly tools

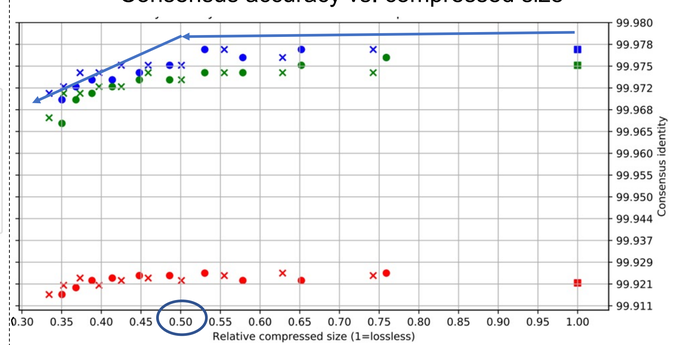
Results

Basecalling accuracy vs. compressed size





Consensus accuracy vs. compressed size



Achieve 35-50% reduction over best lossless compression with negligible loss in accuracy
Consistent observations across datasets, coverage, downstream tools
Evaluation scripts, data, plots: https://github.com/shubhamchandak94/lossy_compression_evaluation