Lossless Compression of Nanopore Sequencing Raw Signals

Rafael Castelli, Tomás González, Rodrigo Torrado, Álvaro Martín, and Guillermo Dufort y Álvarez

Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay

Overview

- Problem introduction
- Data compression
- Nanopore Sequencing
- VBZ: The current standard
- Our proposed improvements and innovations
- Results and conclusion

Introduction

- Nanopore sequencing is revolutionizing genomics.
- Oxford Nanopore Technologies (ONT) is a key player.
- Challenge: Massive raw signal data (10x larger than FASTQ).
- Raw signal preservation is important for future analysis due to algorithm improvements
- Our solution: Innovative lossless compression algorithms.

Brief introduction to Data Compression

- There are two types of compression: lossless and lossy.
- A set of values is more compressible if it has more redundancy.
- One approach is to separate the data in different sets which are separated acording to correlation between the values.
- In order to achieve compression of these sets of data, clever encoding algorithms are used to store information in smaller file size.

Nanopore Sequencing

- DNA passes through a nanopore.
- Disrupts electrical current, creating a signal.
- Raw signal is captured and analyzed.
- Basecalling: Decoding the DNA sequence.

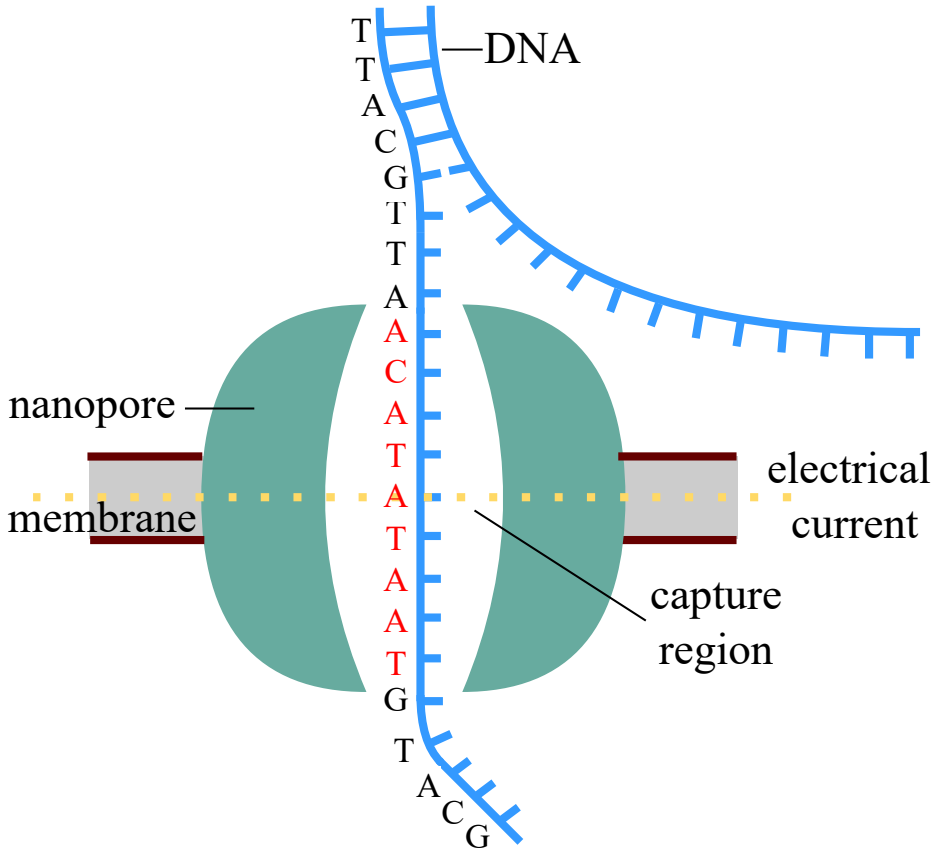
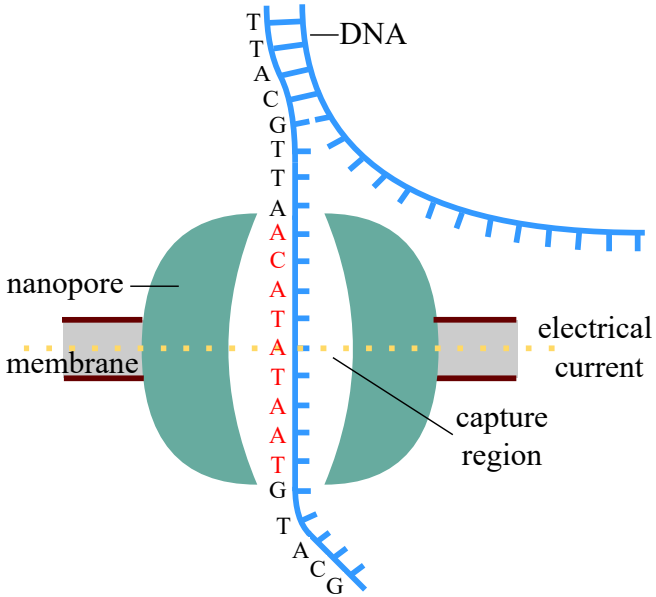
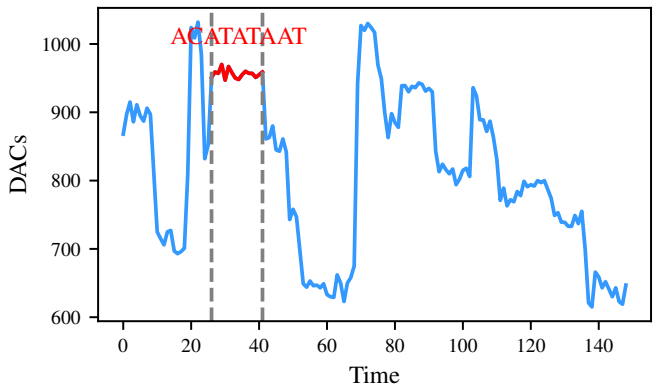


Figure: Nanopore sequencing process

The Raw Signal

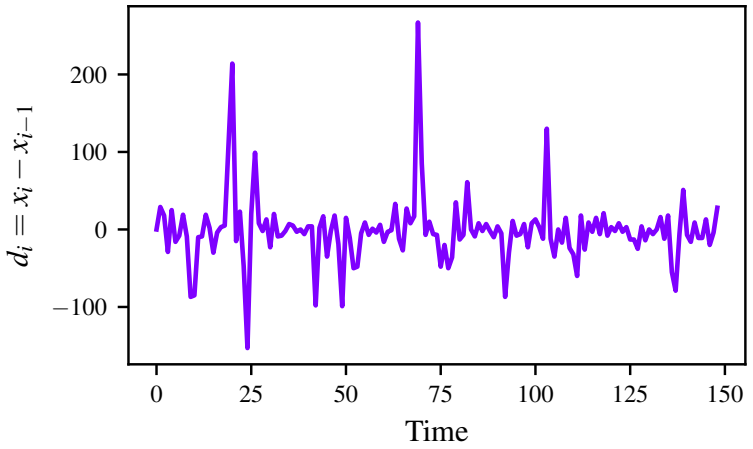
- Each k-mer obeys a specific distribution.
- We can separate in stable and transition values.
- How can we take advantage of this?.

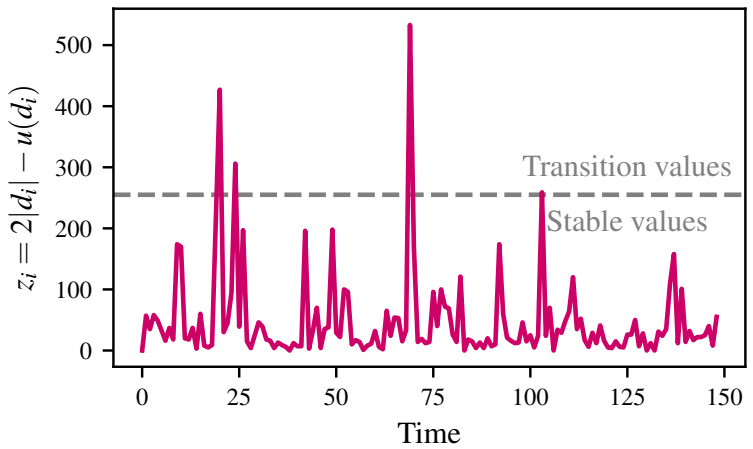




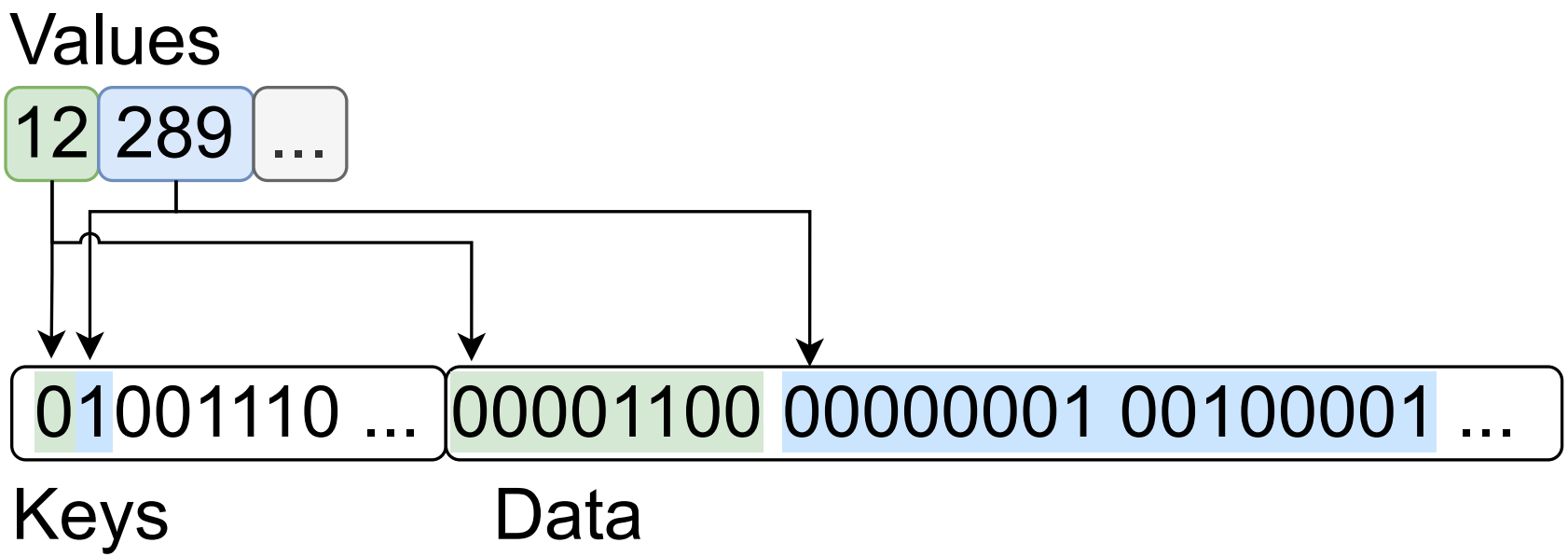
VBZ: The Current Standard

- State-of-the-art compression algorithm.
- Cleverly encodes the difference of the consecutive values.
- Separate values according to fixed threshold, attempting to separate stable and jump values. And stores a buffer of keys to the decode.
- After the values are decoded it runs ZSTD (LZ77 based compression algorithm)





VBZ: The Current Standard



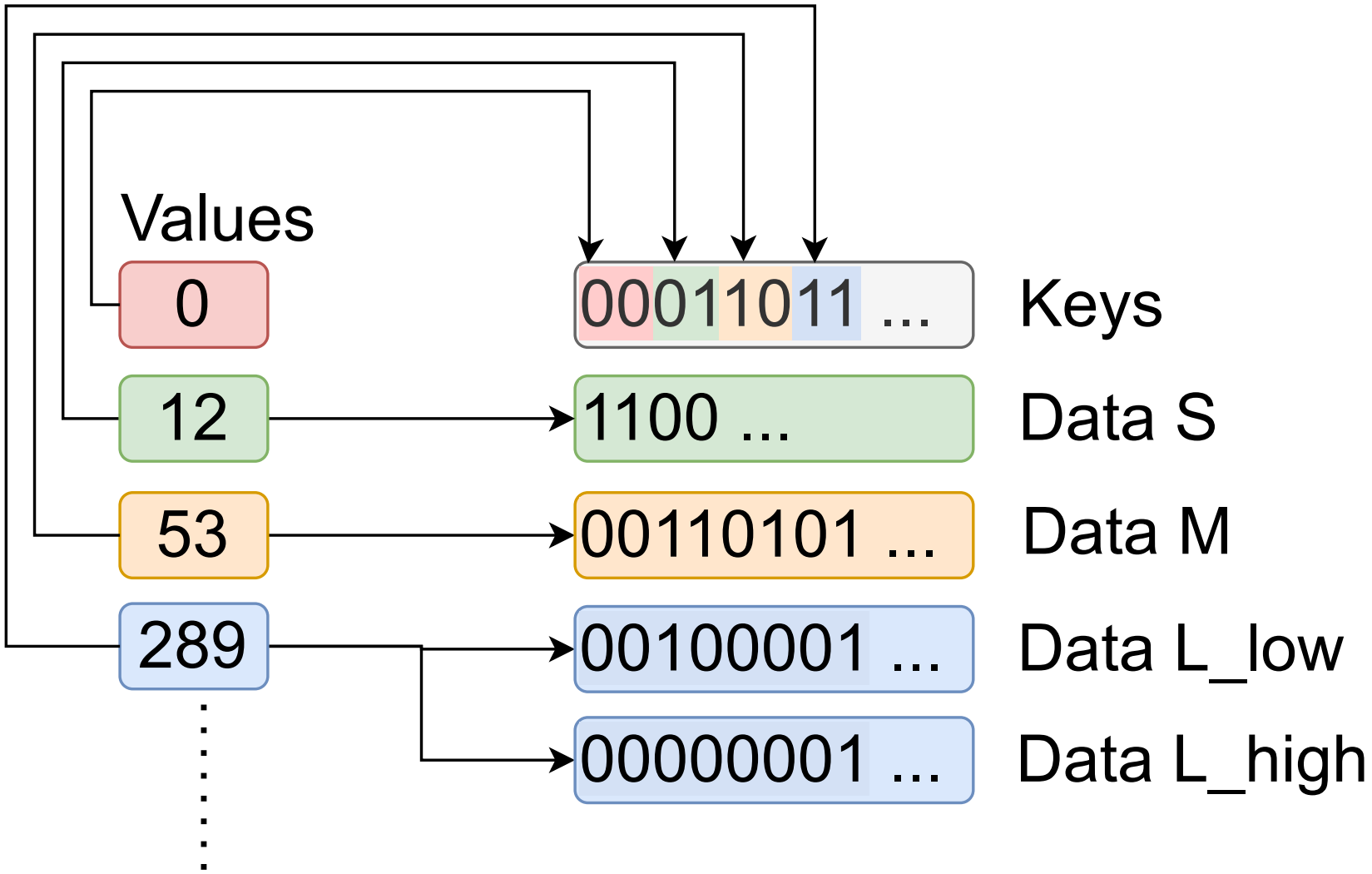
Problems in VBZ

- The ZSTD algorithm is run on keys and data simultaneusly
- The stable and transitional values are compressed together even though they have different statistical properties

Our solutions and improvements

- We run the ZSTD compression algorithm separately on each buffer.
- Separate into more buffers according to signal properties.
- Explore alternative compression techniques.
- We presented a series of new compression algorithms derived from VBZ.

Our solutions and improvements



Experimental Results

- Evaluated on various datasets (different organisms, nanopore models).
- Our methods consistently outperform VBZ in compression efficiency.
- Comparable speed and memory usage.

Table: Compression ratios and percentage relative difference (PRD) with respect to VBZ1.

	Averages	
Compressor	CR	PRD
VBZ1	6.656	
VBZ0	7.450	12.02
C1	6.633	-0.31
C2	6.572	-1.17
C3	6.566	-1.25
C4	6.508	-2.13
C5	**6.487**	**-2.42**
Pgnano	6.457	-2.85

Conclusion

- Lossless compression is vital for nanopore sequencing.
- Our research advances the field.
- Future directions:
  - SIMD optimization.
  - Predictive models for better compression.

Acknowledgements

- This project was funded by ANII
- Special thanks to the rest of the Information theory Group.

Questions?

Thank you!