smallWig: Parallel Compression of RNA-seq WIG Files

Zhijing Wang*, Tsachy Weissman*, Olgica Milenkovic†

*Department of Electrical Engineering, Stanford University
†Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign

RNA-seq Expression

- Whole Transcriptome Shotgun Sequencing
- RNA presence and quantity
- Biological discovery
- Genomic Medicine
- Expression data
  - Indicator of abundance of RNA molecules expressed from given chromosome locations
- Current format: Wig[2], bigWig[3]
- Other quantitative measurements
  - Proteomic, metabolomic data

System Model

Input: Expression Seq.

Run-Length Encoder

Expression Seq. or Run-Length Seq.

Expression Diff. Seq. or Run-Length Diff. Seq.

Delta Encoder

Random Query Enabled?

Block Separator

Arithmetic Encoder

Empirical PMF Generator

Combiner

Output: Compressed File

Compression: Necessity and Measurement

Necessity
- Storage: in the hg19 browser in ENCODE[4], about 10% of all tracks and 31% of space are for bigWig[5]
- Communication: BGI produces ~2000 human genomes a day, which takes weeks to transmit with Internet
- Measurement
  - Compression rate
  - Compress/decompress time
  - Random query functionality: allows to retrieve information corresponding to a subsequence of chromosome locations

Experimental Results

- Compared to bigWig, the proposed algorithm with arithmetic coding obtains 17x improvement on compression rate
- The running time has up to 3.8x improvement for encoding and 1.5x for decoding
- If allowing more elaborated compressor, the compression rate can be improved by 23x

Acknowledgements

The work is supported by NIH BD2K grant number 1U01CA198943-01.

References

[8] Available at: http://web.stanford.edu/~zhiyingw/smallWig/