MOL2NET
International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-03

Data compression with Python: application of different algorithms with the use of threads in genome files

Vinicius Seus[1] (viniciusseus@gmail.com)
Alex Camargo[1] (alexcamargoweb@gmail.com)
Diego Mengarda[2] (diegormengarda@gmail.com)
[1]FURG
[2]UNIPAMPA

Brazil

Introduction

This work proposes and evaluates an implementation of different algorithms of **data compression using the Python** programming language allied to the use of threads.

- As a case study it was used genomic data available from the NCBI (National Center for Biotechnology Information) public database.

Introduction

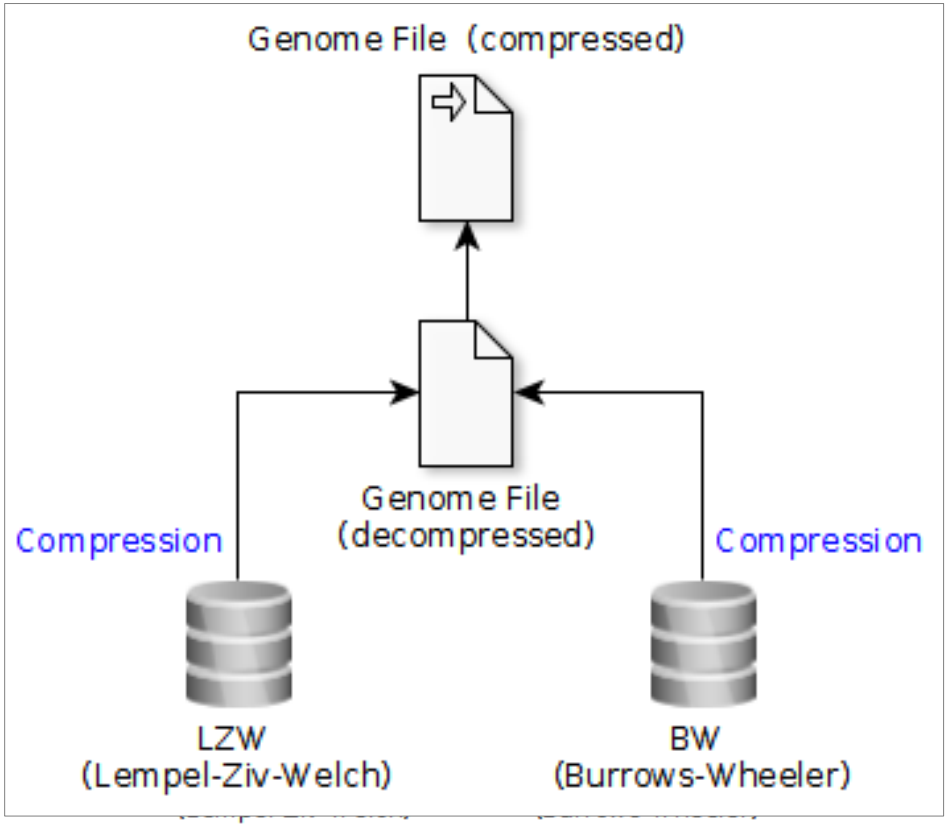

Figure 1. Graphical Abstract

Materials and Methods

For the experimental environment it was used a simplified structure composed of a computer with: 3.07GHz processor (12 cores), 12 GB RAM and 500 GB hard disk.

- Table 1 shows the results of the experiments for the file "ref_ASM45574v1_gnomon_scaffolds.txt" with a total size of 122 MB, referring to an excerpt from the genome identified by "Aligator sinensis", belonging to the family Alligatoridae.

Table 1. Results for the different compression methods

Method	Compression ratio (%)	Compression time (s)
BW	0	-
LZW	39.98	29.72

Conclusions

The main contribution of this work was **to present an algorithm option for data compression based on the Python** programming language.

- By default, the algorithms were not designed to work in parallel, however, with the use of the Python Threading library this was achieved.
- With the experimental environment implemented, it was possible to analyze the performance of both the compression rate and the compression time for each algorithm.
- As future works, we intend to extend the range of algorithms to be studied as well as the application and analysis of decompression metrics with emphasis on public genomic data.

References

BURROWS, Michael; WHEELER, David J. A block-sorting lossless data compression algorithm. 1994.

VERLI, Hugo et al. Bioinformática da Biologia à flexibilidade molecular. Porto Alegre, Brasil, v. 1, 2014.

WELCH, Terry A. A technique for high-performance data compression. Computer, v. 6, n. 17, p. 8-19, 1984.