Benchmarking sketch compression using Gzip and XZ

Marie Picard

–

Supervisors : Karel Břinda, Leo Ackermann

October 9, 2025

Data

Input

A tarball of 8000 phylogenetically-ordered mash sketches ($s = 1000$) of *neisseria gonorrhoeae* assembled genomes (HQ)

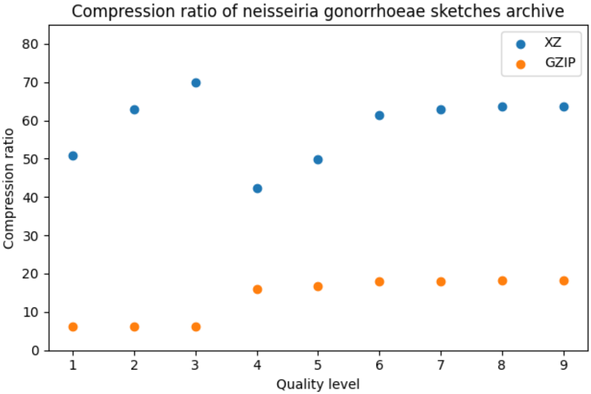
Output

The compression ratio of the archive $\mathcal{A}$: $\frac{size(\mathcal{A}_{uncompressed})}{size(\mathcal{A}_{compresssed})}$

Parameters

- compression scheme : XZ (5.6.4) or GZIP (1.10)
- quality of compression (in range 1 to 9)

1st results : phylogenetically ordered archive

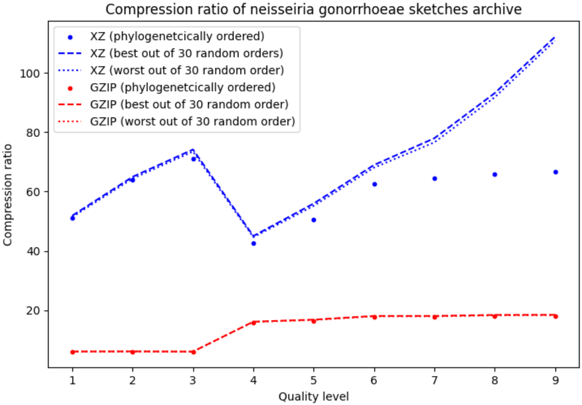


Data

Input

A tarball of 8000 $\prec$-ordered mash sketches ($s = 1000$) of *neisseria gonorrhoeae* assembled genomes (HQ)

Randomly ordered archive



Questions and future work

Questions :

- Why a drop with XZ, for $quality = 3$?
- Why is phylogenetic order so much less efficient (XZ) ?

Questions and future work

Questions :

- Why a drop with XZ, for *quality* $= 3$?
- Why is phylogenetic order so much less efficient (XZ) ?

To do :

- Test with a wider range of data and more random experiments
- Separate headers from sketches
- Look more thouroughly into XZ algorithm