Analysis of the presence matrix of hashes

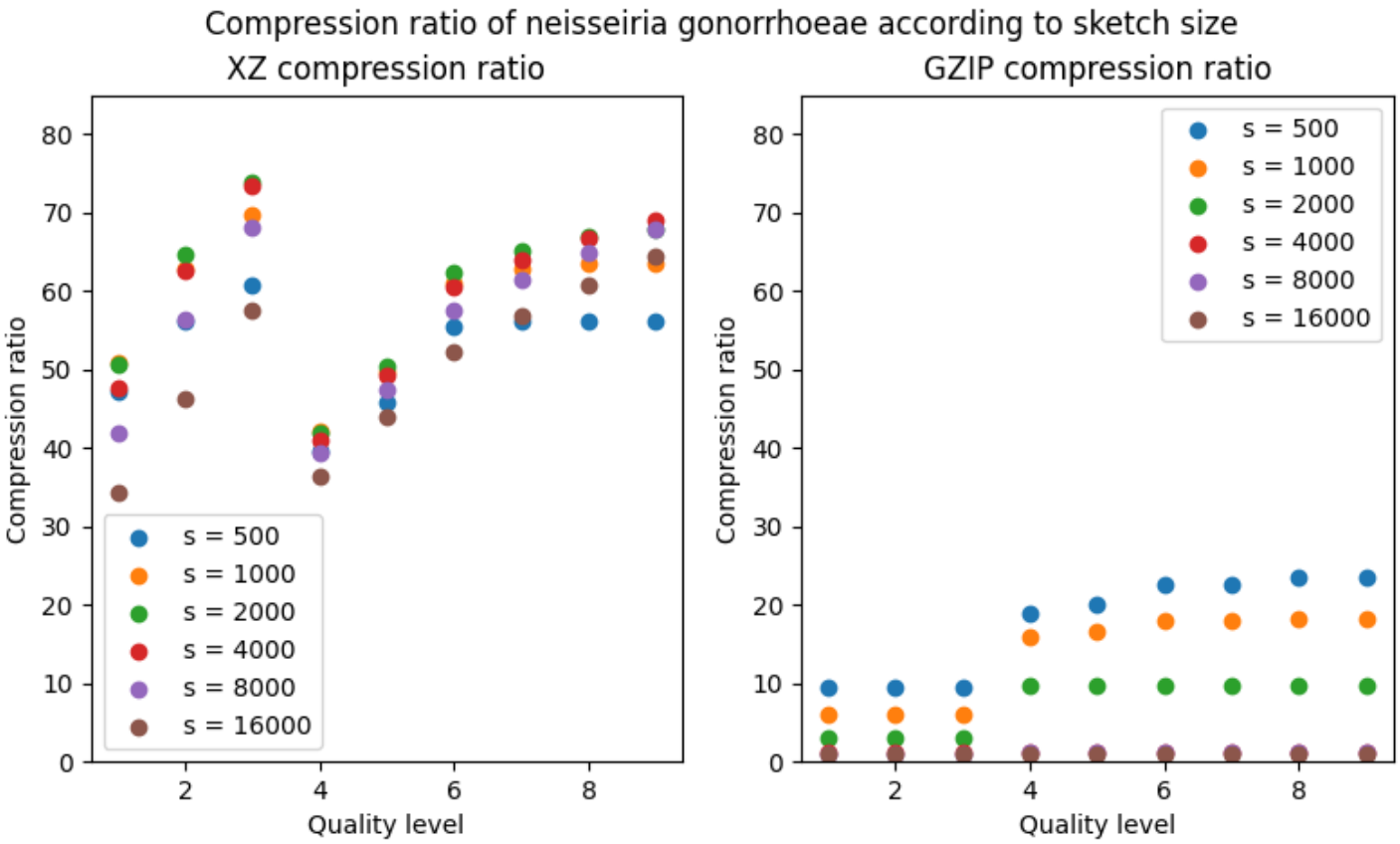
Marie Picard

–

Supervisors : Karel Břinda, Leo Ackermann

November 13, 2025

Compression ratio of XZ and GZIP over an ngono archive (4000 genomes)



Hypothesis

H (informal)

There is a value of sketch size s, around 2000 to 4000, such that the compressibility of sketches is optimal.

Hypothesis

H (informal)

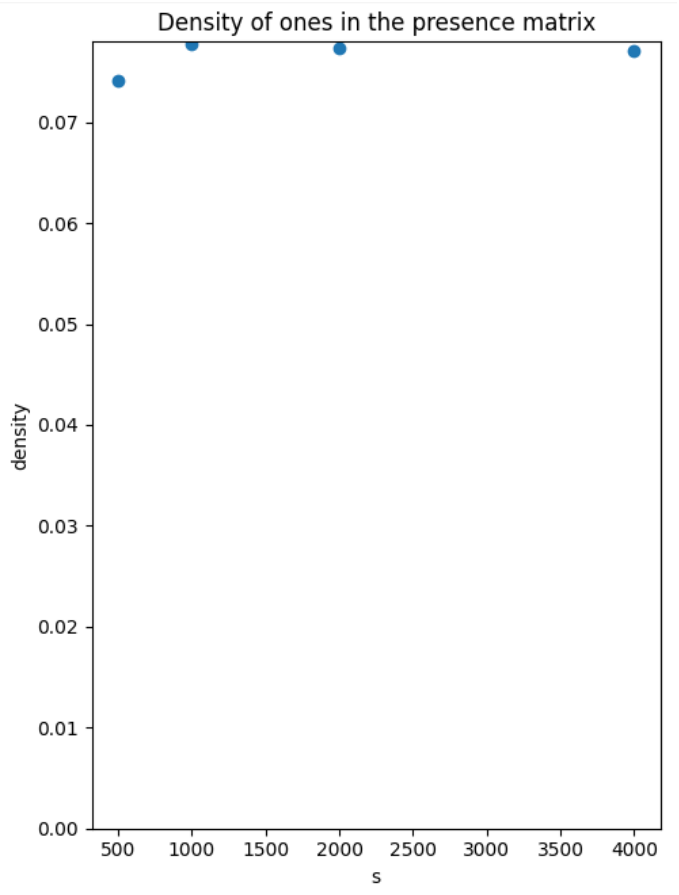
There is a value of sketch size s, around 2000 to 4000, such that the compressibility of sketches is optimal.

H (formal)

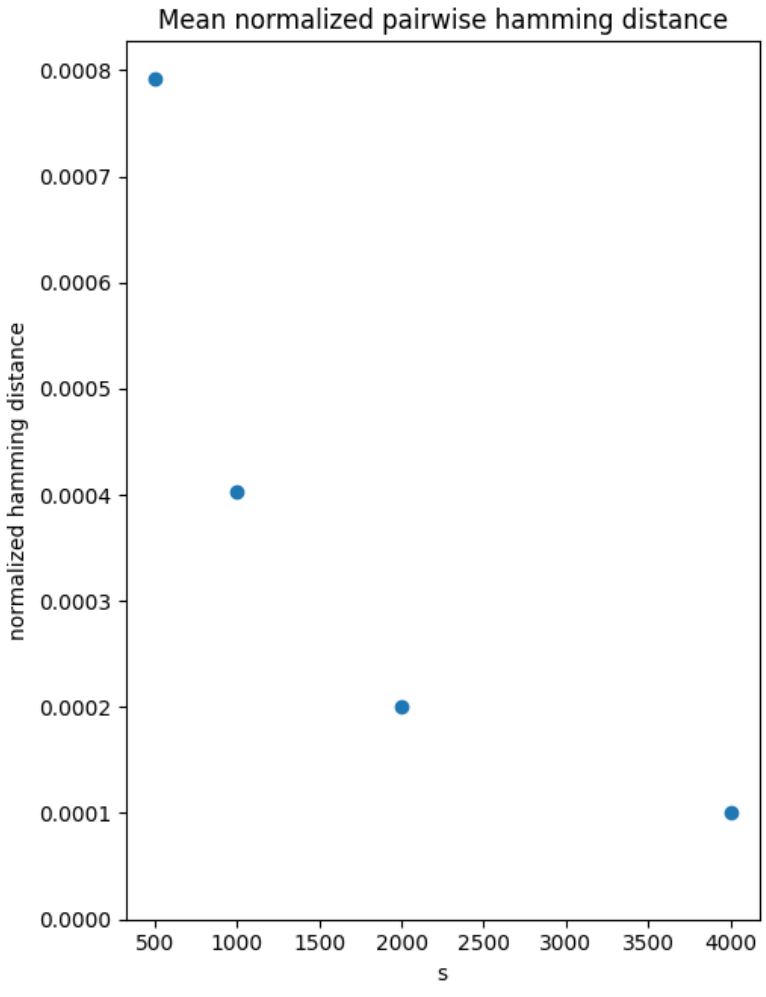
$\forall \mathcal{G}, \exists s \in \mathbb{N}, K(Sketches(\mathcal{G}, s))$ is minimal where K is

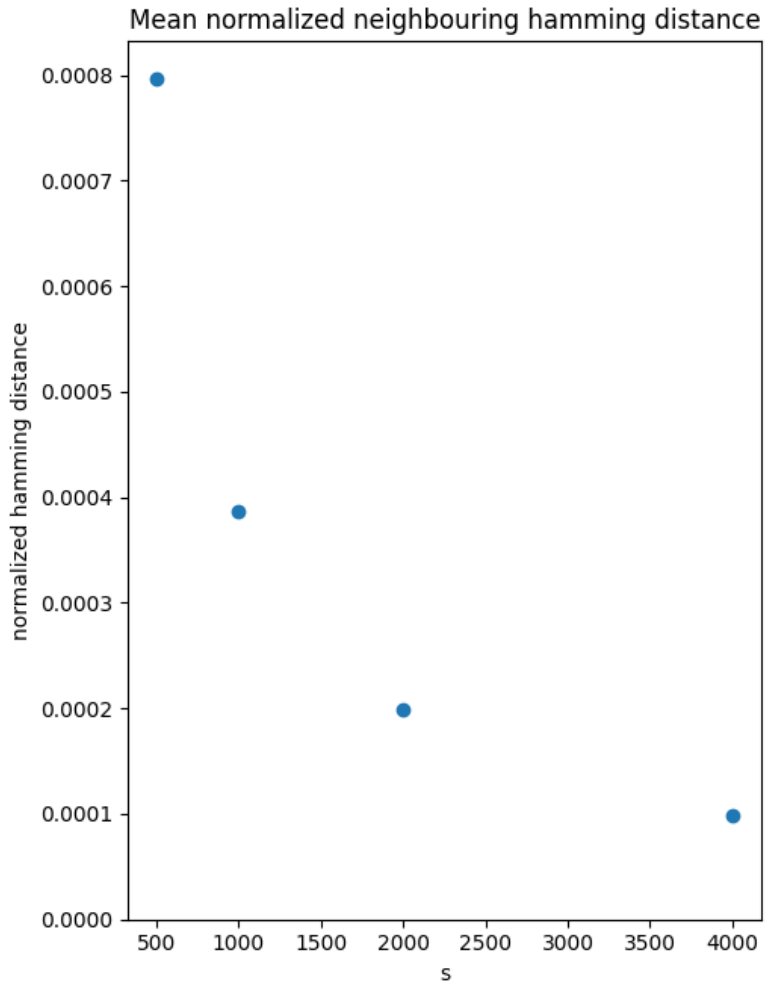
Kolmogorv's complexity, $\mathcal{G}$ is the set of genomes to sketch, and $Sketches(\mathcal{G}, s)$ the archive of all sketches of $g \in \mathcal{G}$ of size s.

Density of the presence/absence matrix



Evolution of Hamming distance





Conclusion

Analysis :

- very low distance and hyperbolic decrease

To do :

- compute for larger values of s
- look at the mathematical aspect of it