On the compressibility of sketches - erratum

RLE and Hamming distance

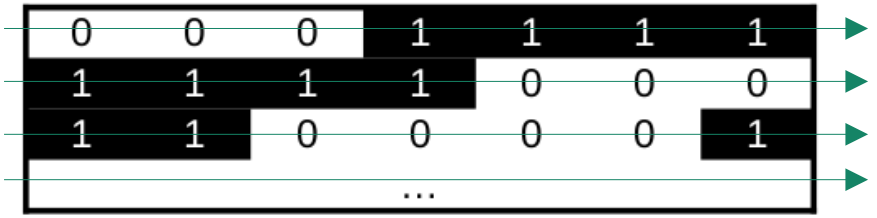
Marie Picard

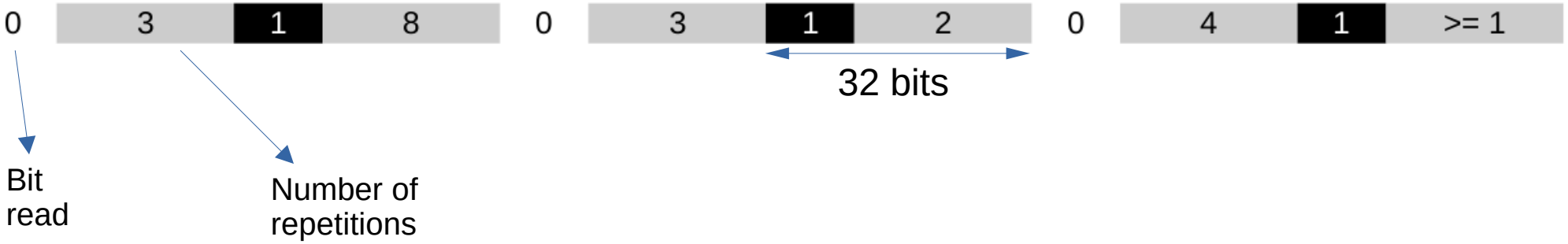
–

Supervisors : Karel Břinda, Leo Ackermann

November 27, 2025

RLE


0 0 0 1 1 1 1
1 1 1 1 0 0 0
1 1 0 0 0 0 1
...


0
3
1
8
0
3
1
2
0
4
1
>= 1
32 bits
Bit read
Number of repetitions

distance δ

$$\delta = \frac{1}{2} \sum_{i=1}^{|\mathcal{S}|-1} d_{Hamming}(S_i, S_{i+1})$$

Line changes

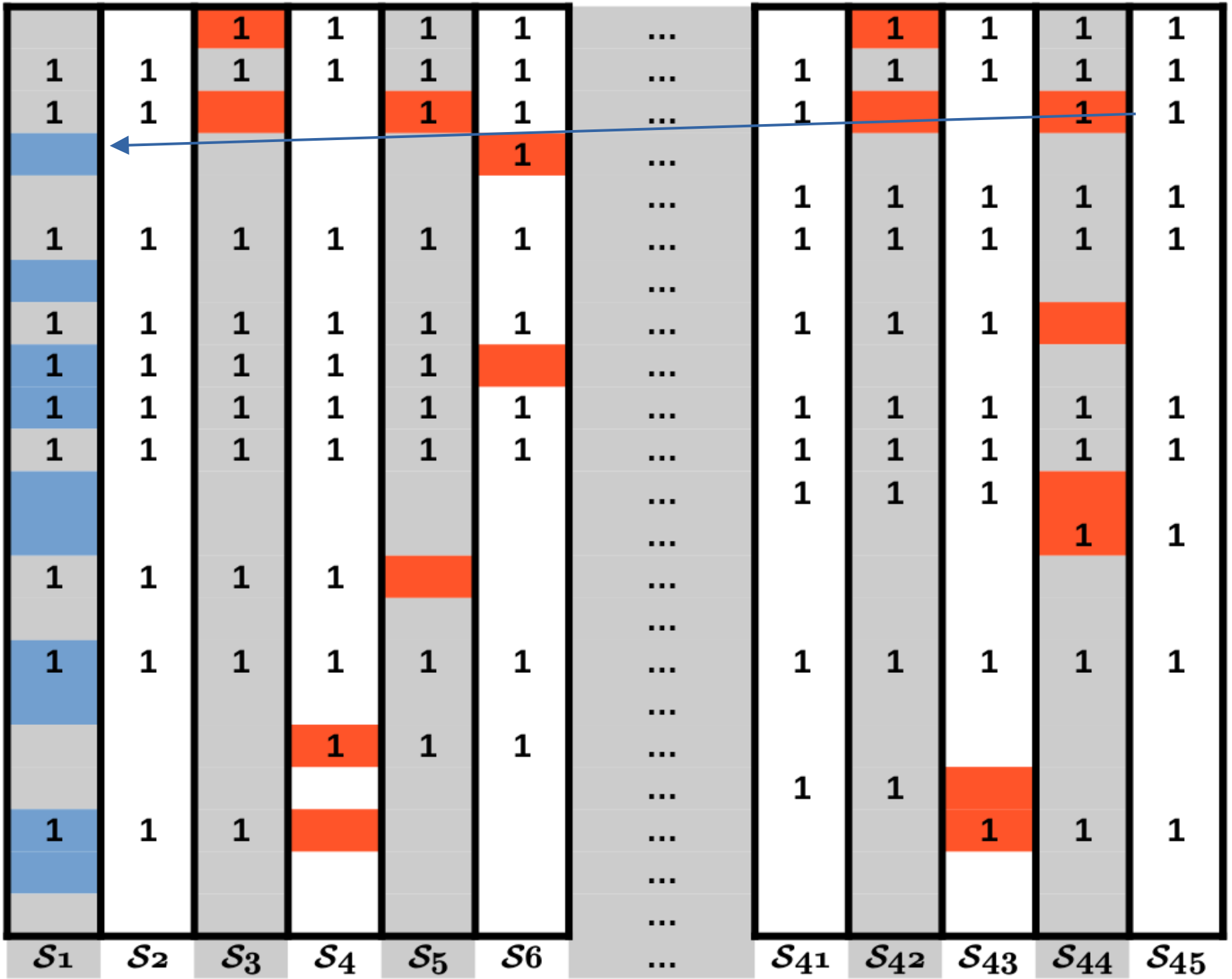
Number of bit changes upon line skips :

$$x = \sum_{i=1}^{|\mathcal{S}_U|-1} |\mathcal{M}_{i,|\mathcal{S}|} - \mathcal{M}_{i+1,1}|$$

$\delta \geq 8$

$x = 10$

Matrix of sketches



- $2\delta + x$ bit changes in RLE
- $x \leq 2s - 1$

So at most $2\delta + 2s - 1$ bit changes.

Size of output :

$$|\mathcal{S}_U| + |RLE(\mathcal{M})|$$

$$64s \leq |\mathcal{S}_U| \leq 64(s + \delta)$$

$$(1 + 2\delta + x)(|bit| + |int|) \leq 64(\delta + s)$$

Size of input :

$$\geq 64 \cdot |\mathcal{S}| \cdot s$$

- 64 bits per hash
- s hashes per sketch
- $|\mathcal{S}|$ sketches in $\mathcal{S}$

Compression ratio r :

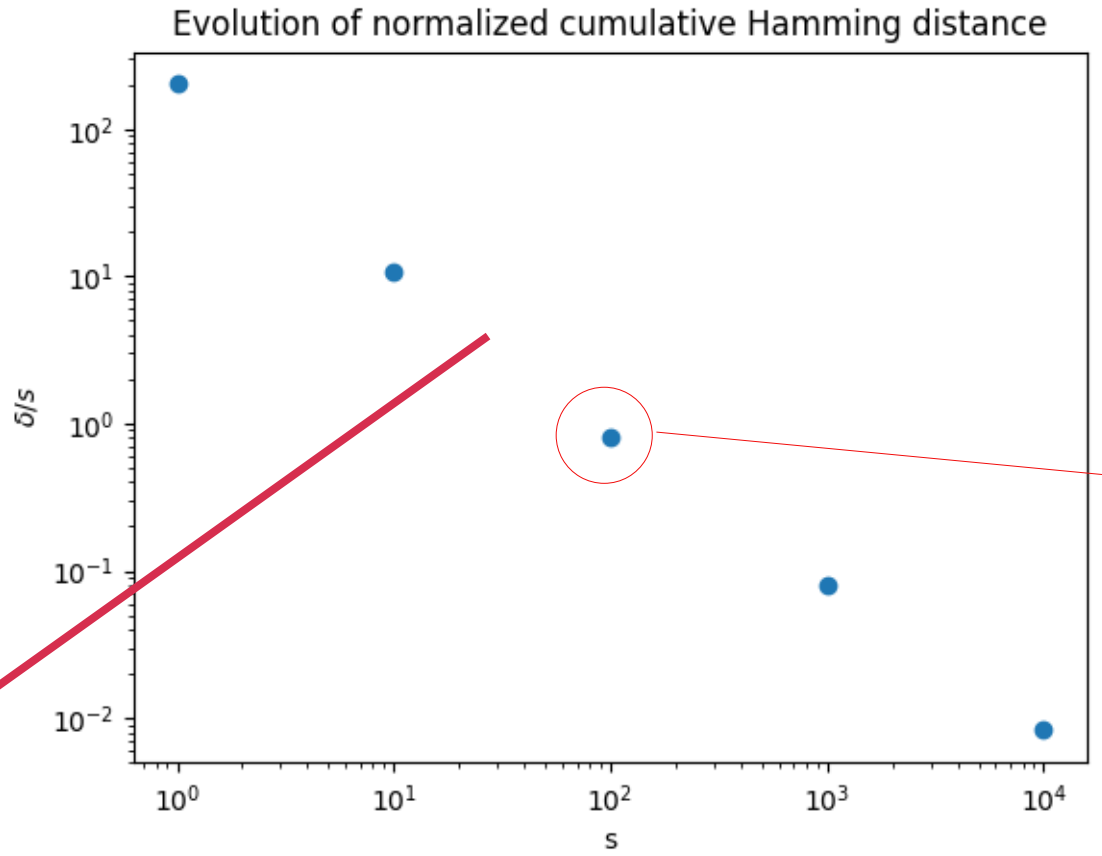
$$r = \frac{|Input|}{|Output|}$$

$$\geq \frac{64 \cdot |\mathcal{S}| \cdot s}{2 \cdot 64(\delta + s)}$$

$$\geq \frac{|\mathcal{S}|}{2(1 + \frac{\delta}{s})}$$

Experimental evaluation

- Genomes to sketch : neisseria_gonorrhoeae__01.tar.xz in part 54 - https://zenodo.org/records/15367750
- 4000 genomes



$$\frac{\delta}{s} \leq 1$$

$$r \geq \frac{4000}{4} \geq 1000$$

!!!

s inversely proportional to $\frac{\delta}{s}$

δ constant ??

$$\delta = \sum_{i=0}^{|\mathcal{S}|-1} \frac{1}{2} d_{Hamming}(S_i, S_{i+1})$$

$$= \sum_{i=0}^{|\mathcal{S}|-1} (s - |S_i \cap S_{i+1}|)$$

$$= |\mathcal{S}|s - \sum_{i=0}^{|\mathcal{S}|-1} |S_i \cap S_{i+1}|$$

$$s \cdot J(G_i, G_{i+1}) \leqslant \mathbb{E}(|S_i \cap S_{i+1}|)$$

Jaccard index

Linear in s

$$\mathbb{E}(\delta) \leqslant s \left(|\mathcal{S}| - \sum_{i=0}^{|\mathcal{S}|-1} J(G_i, G_{i+1}) \right)$$

* a linear lower bound probably exists
(should look at mash paper)

hamming(*u*, *v*, *w*=*None*) [source]

Compute the Hamming distance between two 1-D arrays.

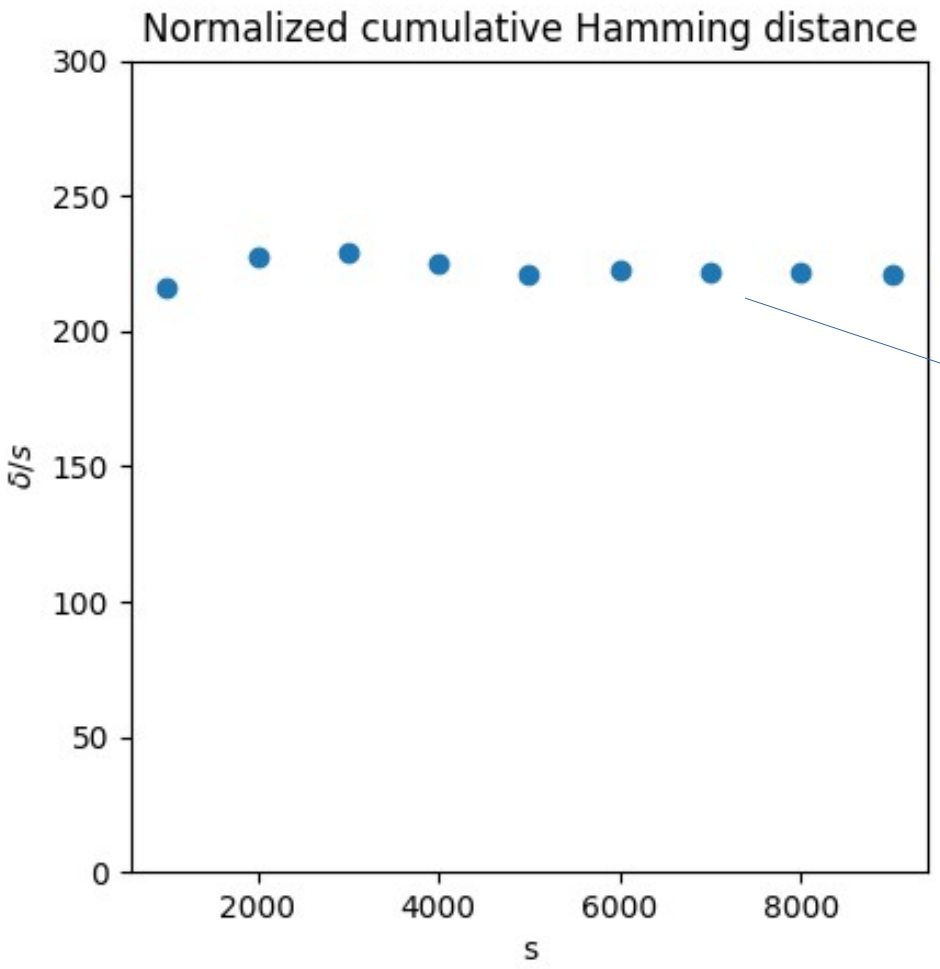
The Hamming distance between 1-D arrays *u* and *v*, is simply the proportion of disagreeing components in *u* and *v*. If *u* and *v* are boolean vectors, the Hamming distance is

$$\frac{c_{01} + c_{10}}{n}$$

where c_{ij} is the number of occurrences of $\mathtt{u[k]} = i$ and $\mathtt{v[k]} = j$ for $k < n$.

Normalized by vector size

Results : s in [1000, 9000]



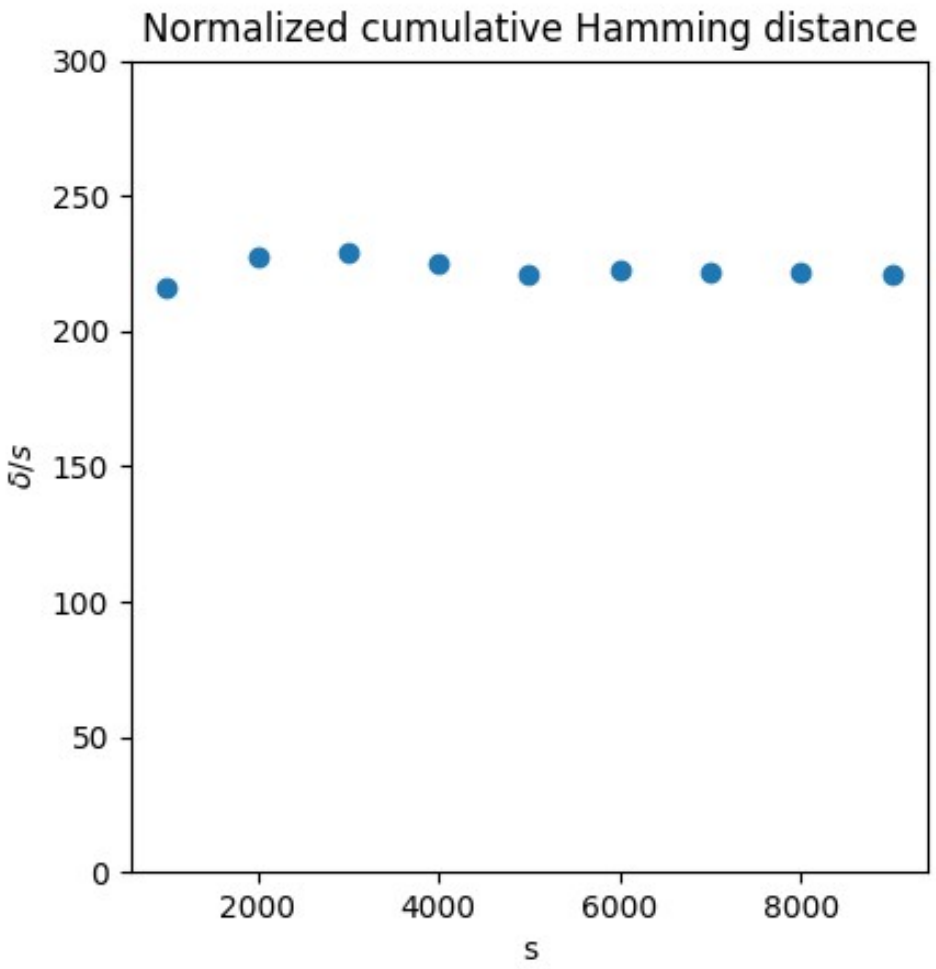
converges

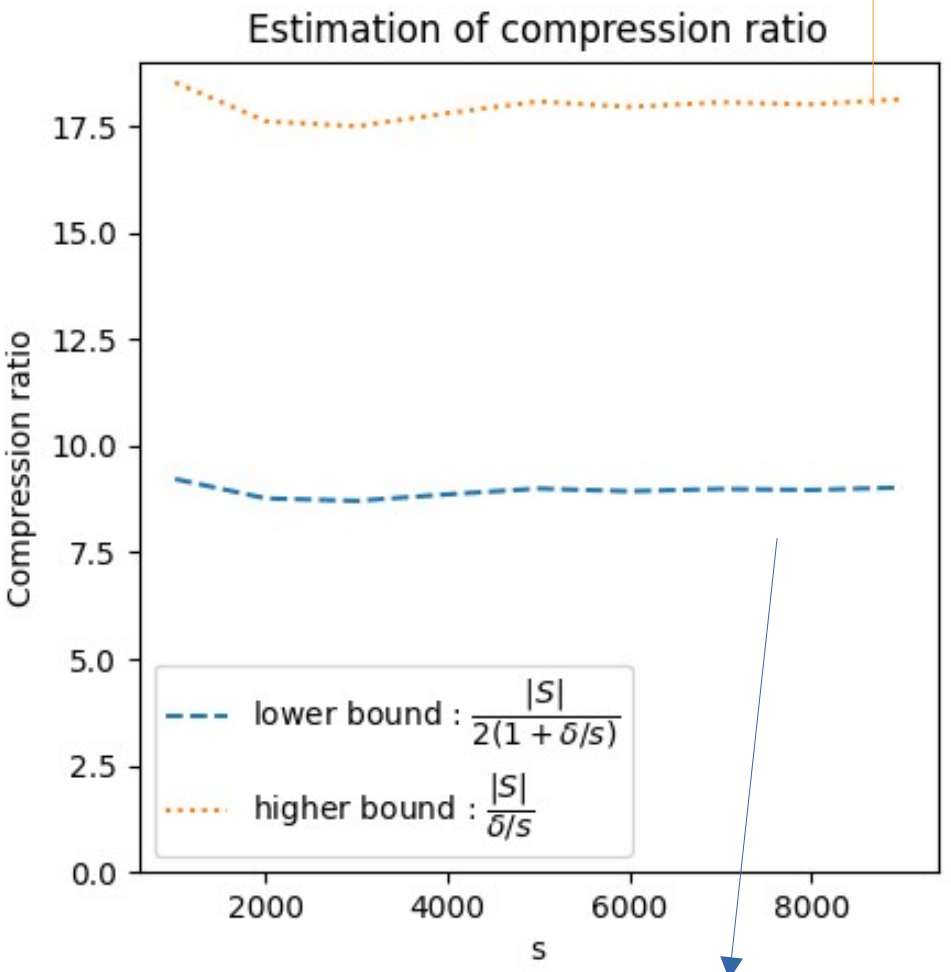
Rough estimation of average Jaccard index

$$\frac{1}{|\mathcal{S}|}\sum_{i=0}^{|\mathcal{S}|-1} J(G_i, G_{i+1}) \approx 0.95$$

Results : s in [1000, 9000]

If the union of all sketches is well-compressed

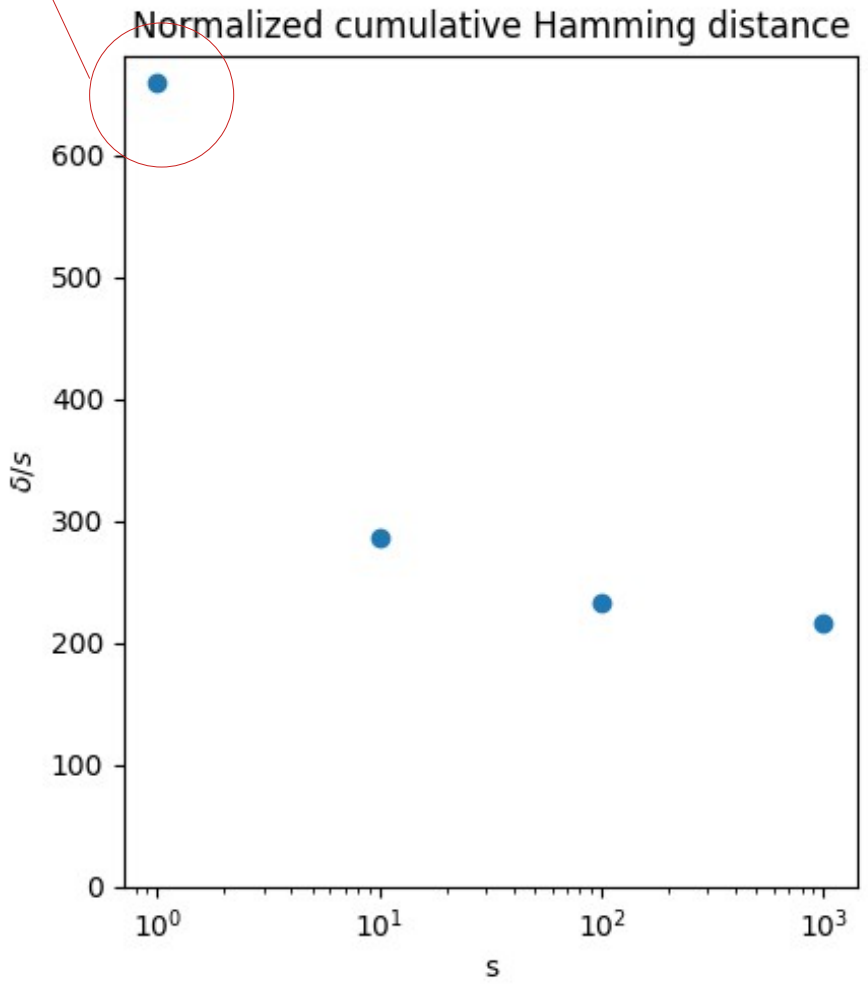


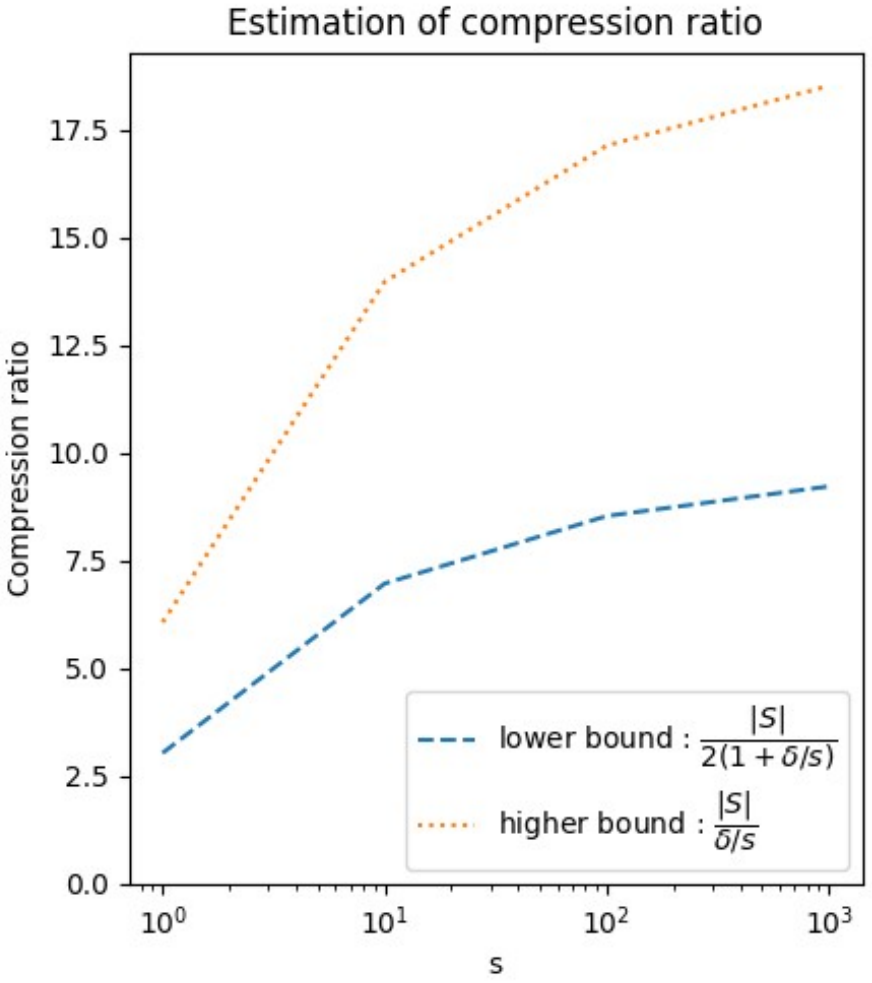


If the union of all sketches is not compressed

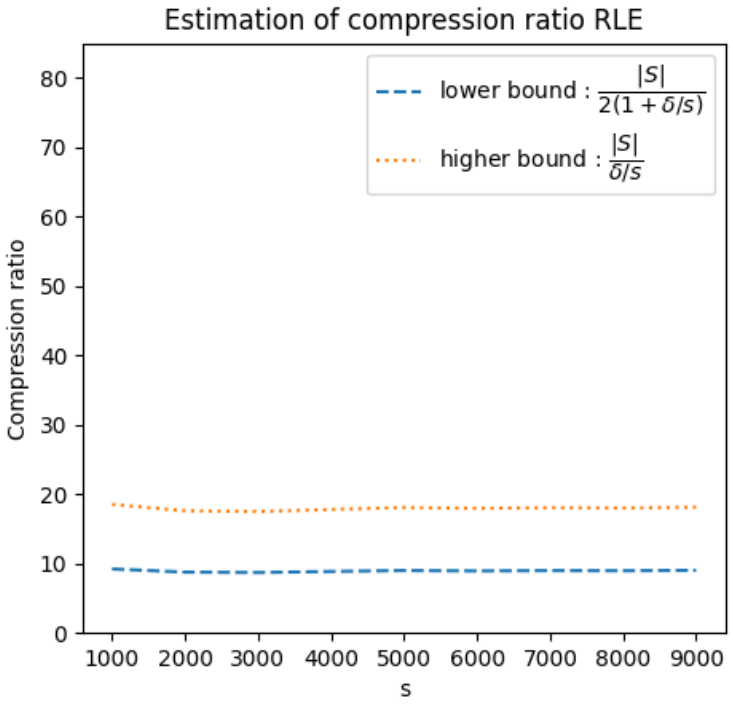
4 different minimal values, one change every 6 genomes

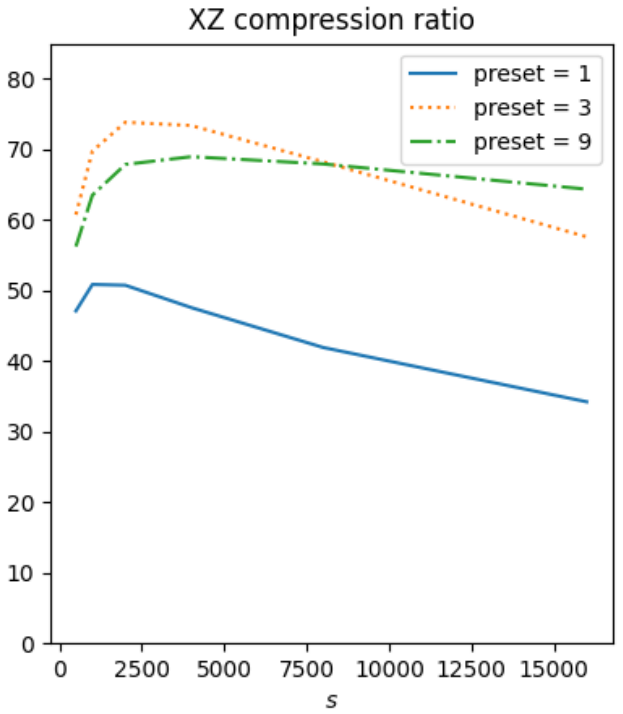
Results : s in [1, 1000] (logscale)

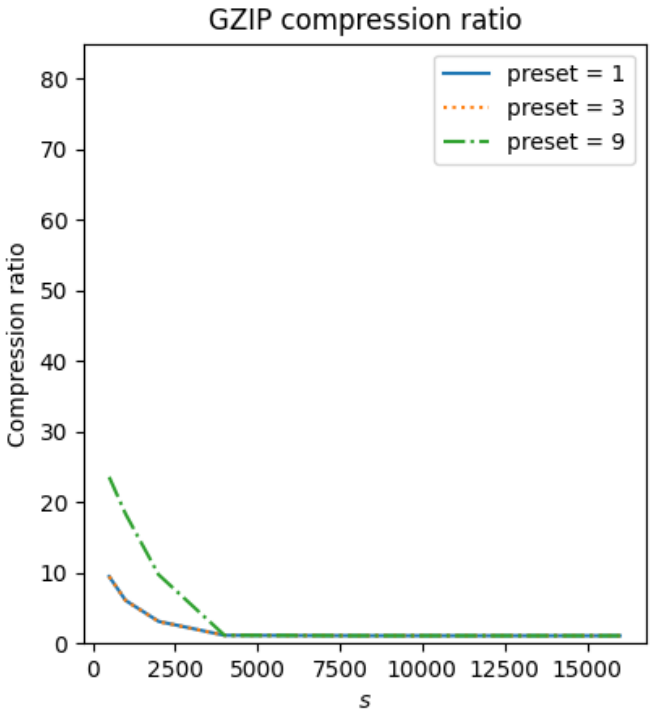




RLE vs XZ vs GZIP compression







* beware of x axis : not exactly the same range for RLE and XZ or GZIP

Conclusions :

- Seems to be a non-optimal order
- Reasonable compression rate, but lower than XZ (on par or better than GZIP though)

To do :

- Look into biological model of $\frac{\delta}{s}$
- Compress $\mathcal{S}_U$ (Elias Fano ?)
- Improve algorithm instead of RLE