Compressibility of the union of sketches

Elias-Fano and RLE

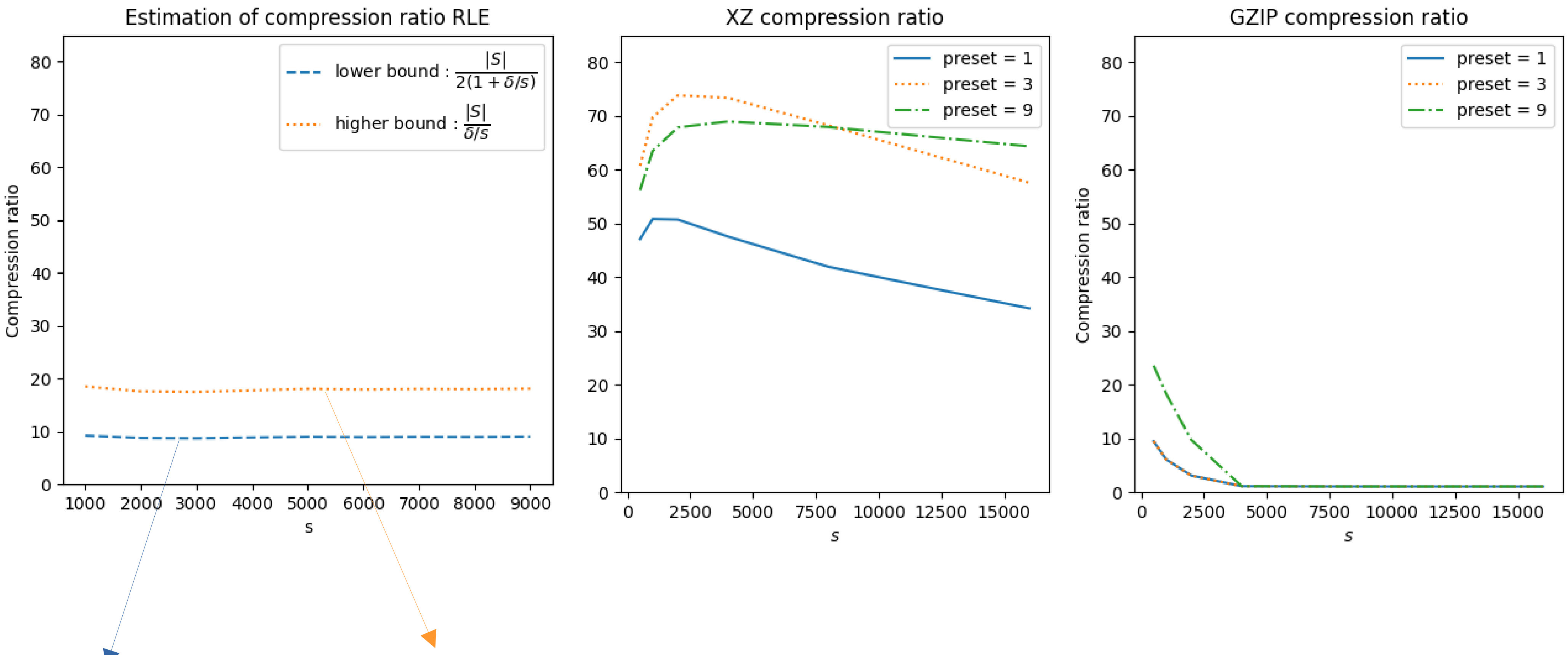
Marie Picard

–

Supervisors : Karel Břinda, Leo Ackermann

December 4, 2025

Reminder of last time's results



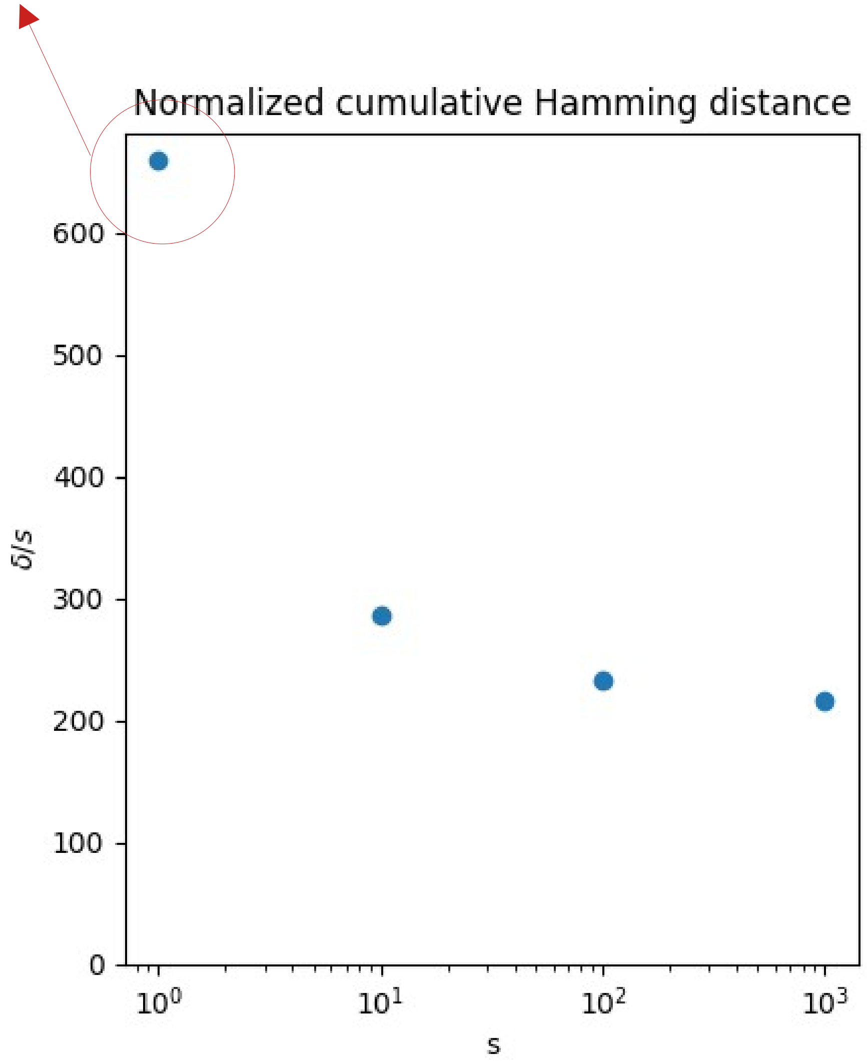
If the union of all sketches is not compressed

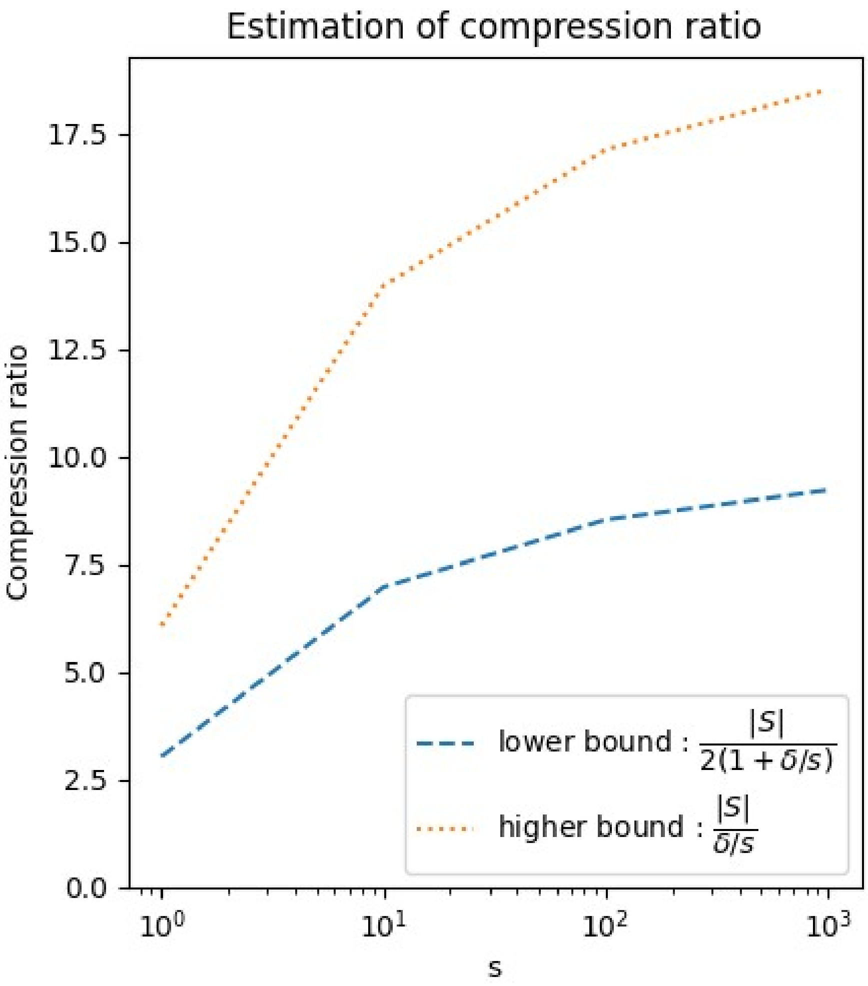
If the union of all sketches is well-compressed

*Neisseria gonorrhoeae 01, 4000 genomes

Reminder of last time's results

4 different minimal values, one change every 6 genomes





*Neisseria gonorrhoeae 01, 4000 genomes

What can we take away ?

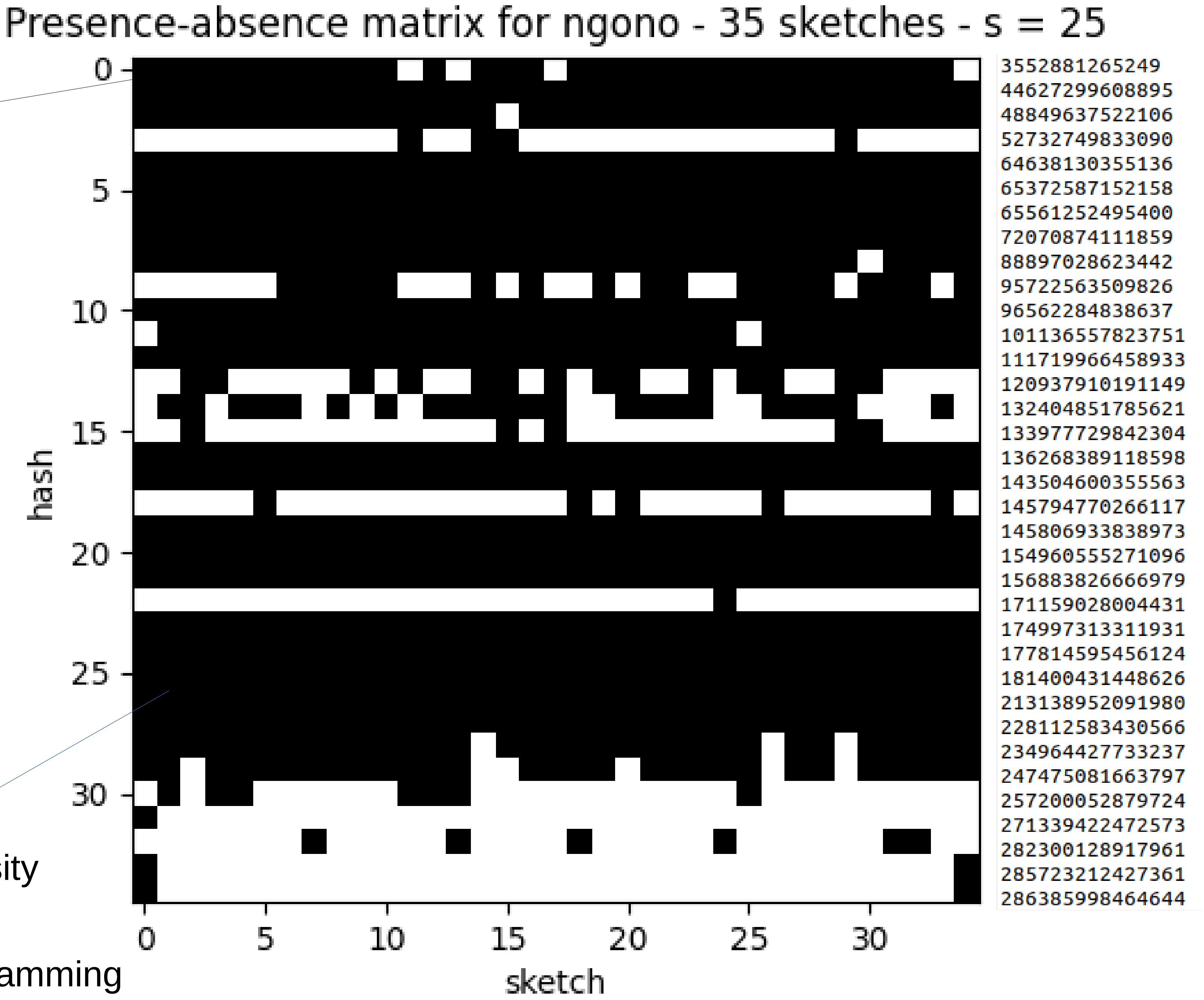
- Factor 2 depending on how well the union of sketches is compressed
- For the binary matrix :
  - Better compression scheme than RLE
  - Is the order optimal ?

What can we take away ?

- Factor 2 depending on how well the union of sketches is compressed
- For the binary matrix :
  - Better compression scheme than RLE
  - Is the order optimal ?

Binary matrix representation

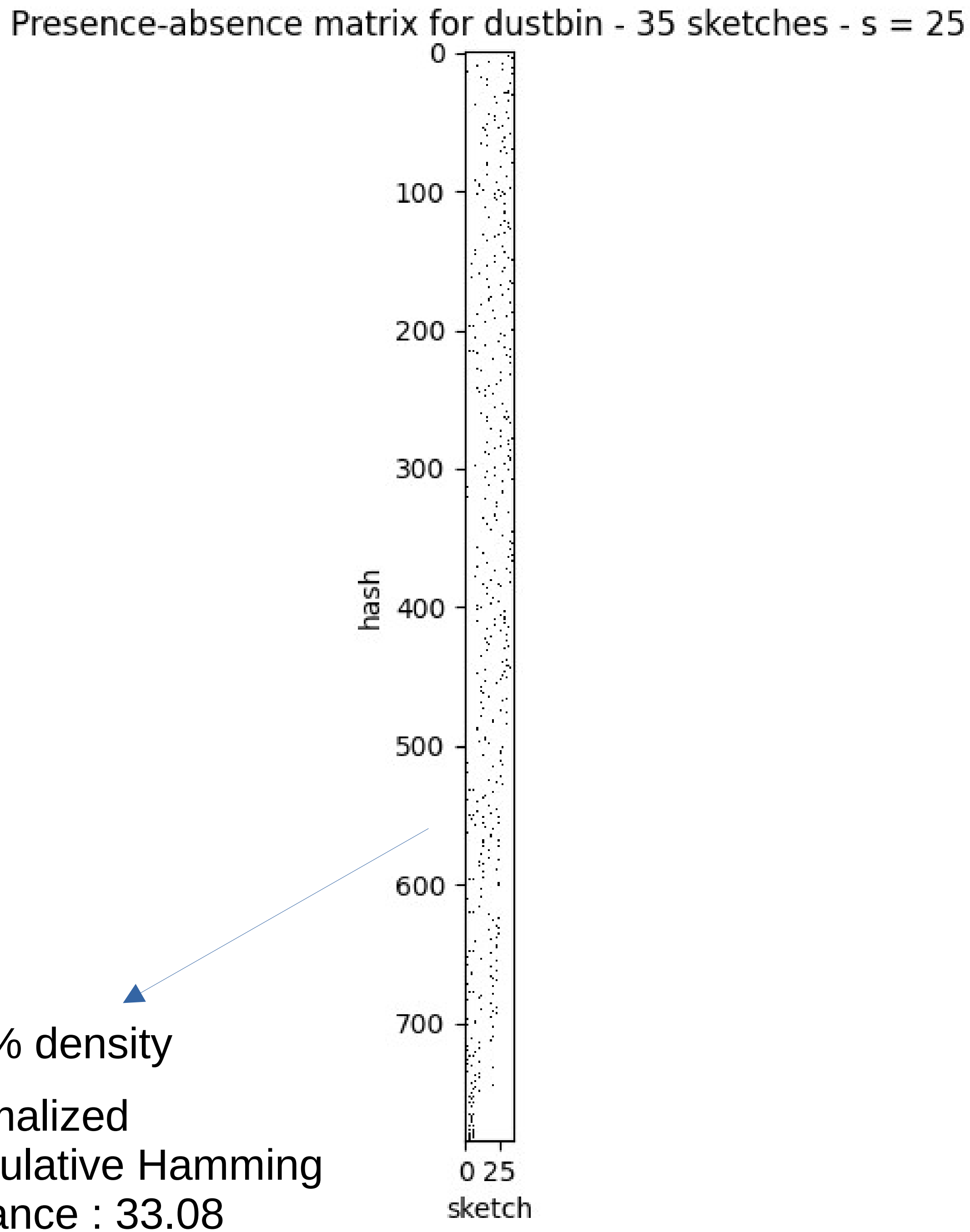
- |S| = 35 (number of genomes sketched)
- s = 25 (number of hashes per sketch)
- Types of genomes sketched :
  - Neisseria gonorrhoeae (part 54, n°01)
  - Dustbin (part 24, n°23)
- Phylogenetic order



Black = presence

> 70% density

normalized cumulative Hamming distance : 2.56

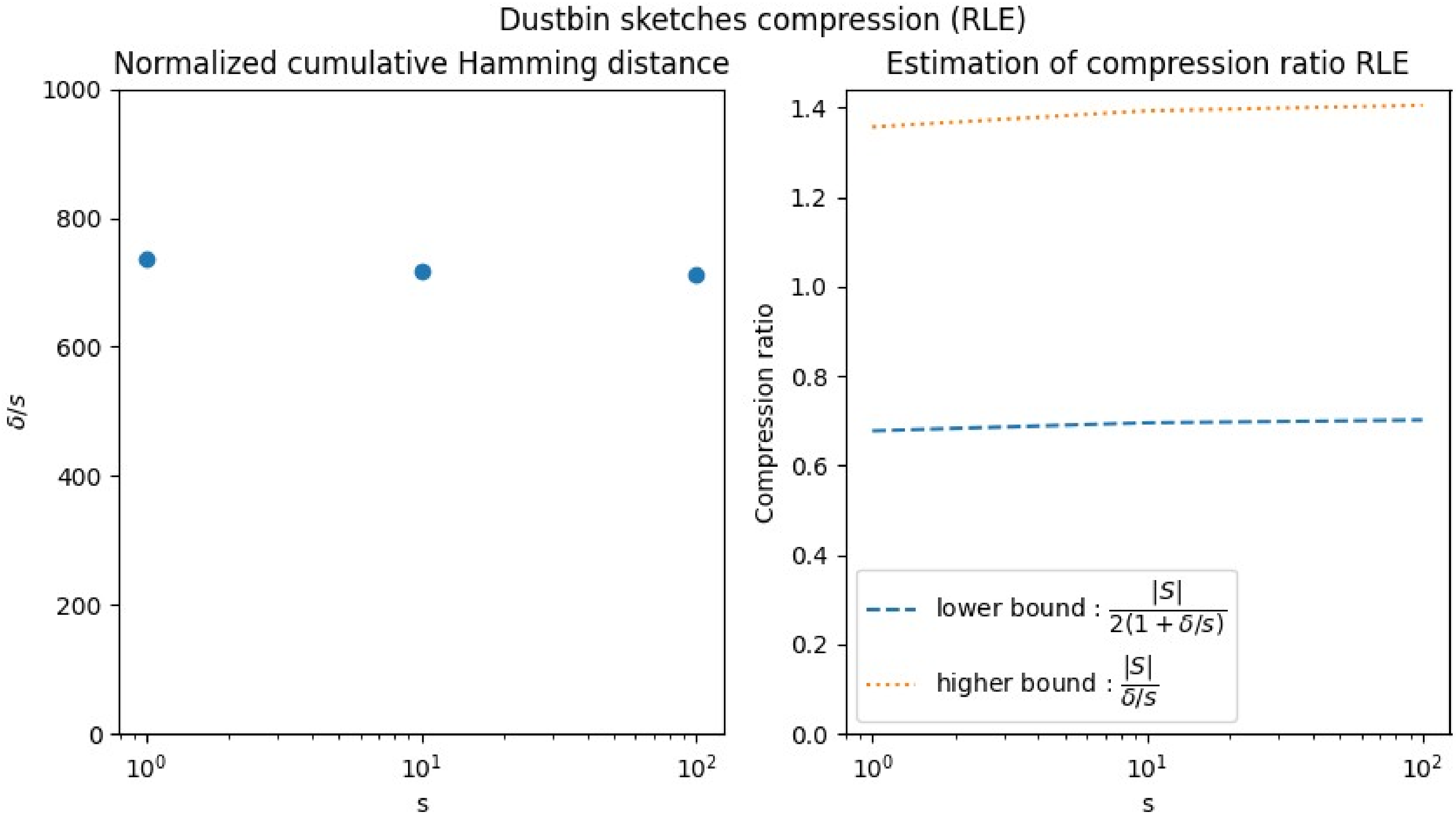


~ 3% density

normalized
cumulative Hamming
distance : 33.08

Compressibility of dustbin

- |S| = 1000 (number of genomes sketched)
- s in [1, 100] (still computing for s = 1000)
- Type of genomes sketched :
    - Dustbin (part 24, n°23)


Dustbin sketches compression (RLE)
Normalized cumulative Hamming distance
δ/s
1000
800
600
400
200
0
10^0
10^1
10^2
s
Estimation of compression ratio RLE
Compression ratio
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
lower bound : $\frac{|S|}{2(1+\delta/s)}$
higher bound : $\frac{|S|}{\delta/s}$
10^0
10^1
10^2
s

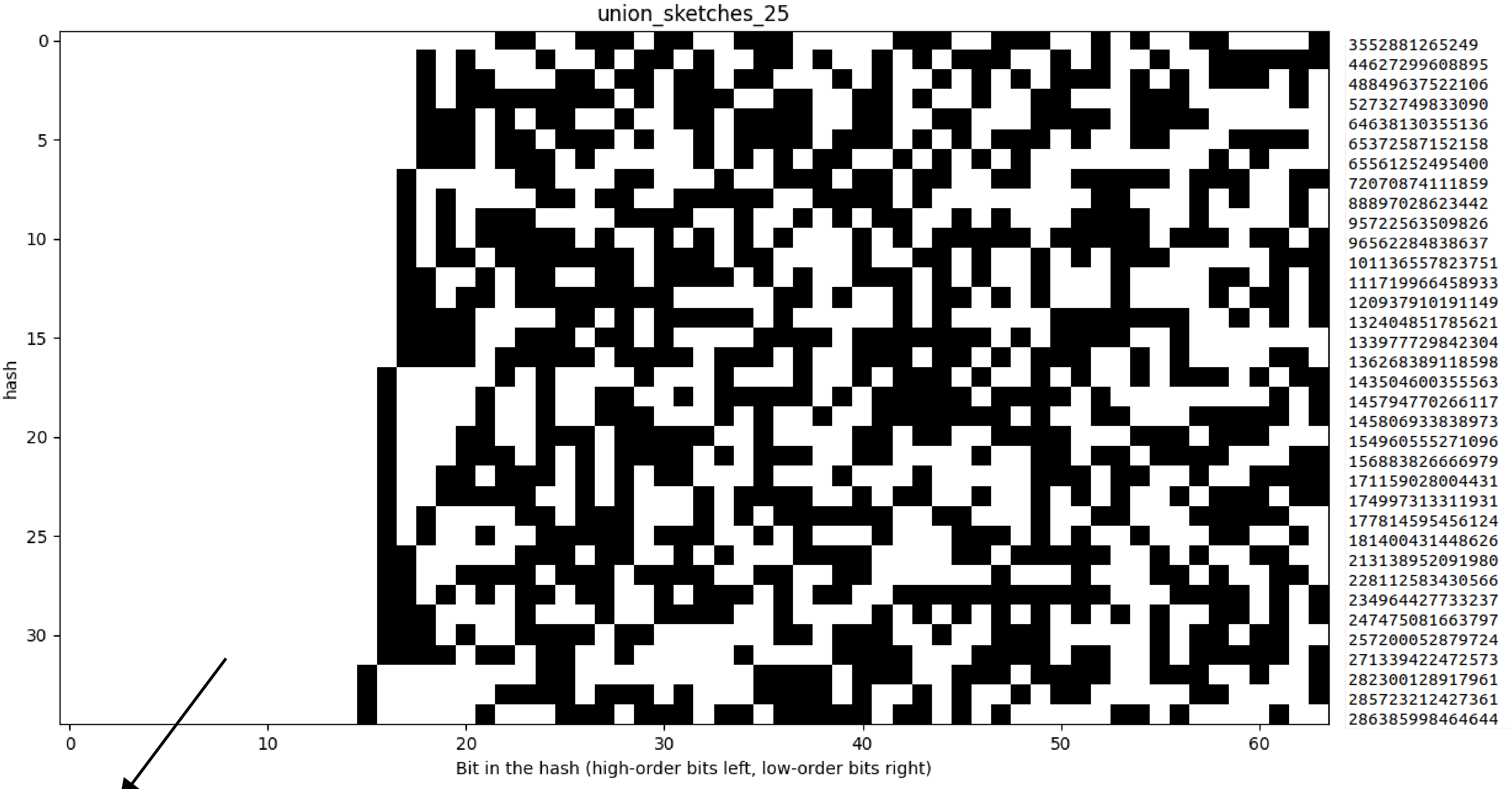
What can we take away ?

- Factor 2 depending on how well the union of sketches is compressed
- For the binary matrix :
  - Better compression scheme than RLE
  - Is the order optimal ?

Representing union of sketches as a binary matrix

- |S| = 35 (number of genomes sketched)
- s = 25 (number of hashes per sketch)
- Type of genomes sketched :
  - Neisseria gonorrhoeae (part 54, n°01)
- Phylogenetic order

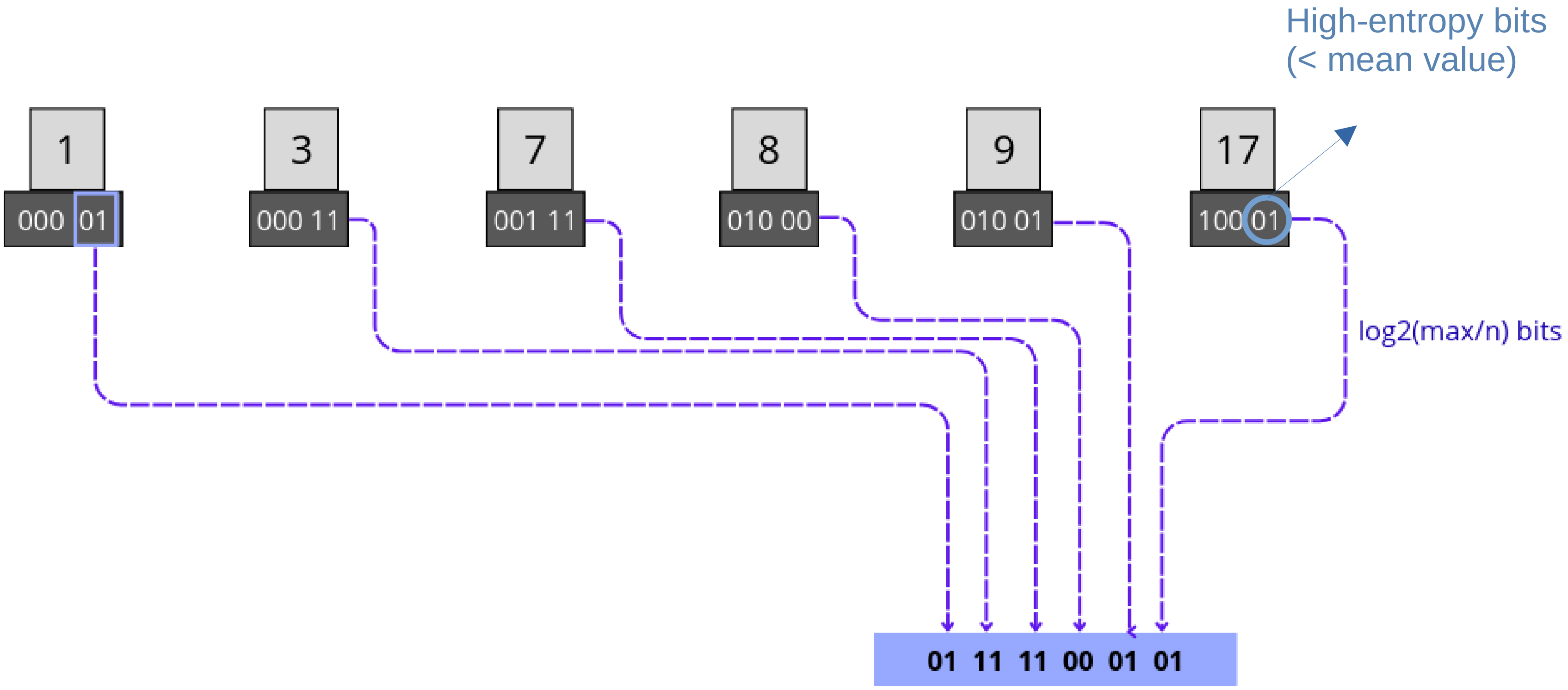
Union of hashes as a binary matrix



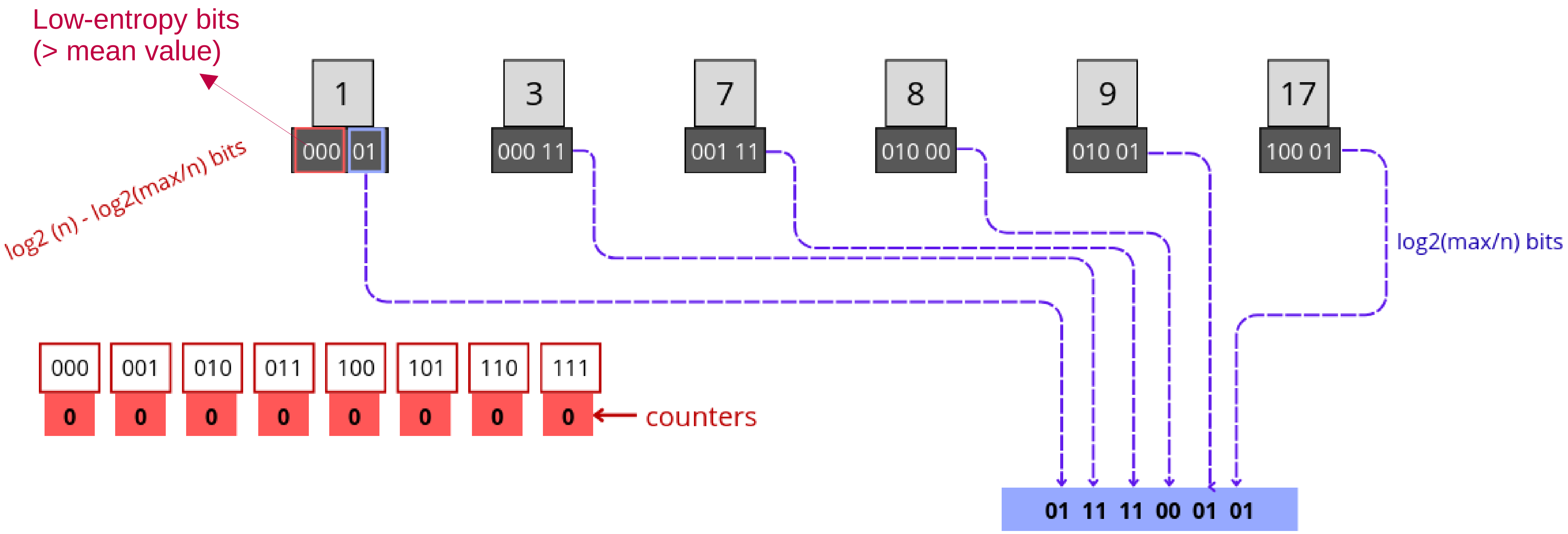
```
3552881265249
44627299608895
48849637522106
52732749833090
64638130355136
65372587152158
65561252495400
72070874111859
88897028623442
95722563509826
96562284838637
101136557823751
111719966458933
120937910191149
132404851785621
133977729842304
136268389118598
143504600355563
145794770266117
145806933838973
154960555271096
156883826666979
171159028004431
174997313311931
177814595456124
181400431448626
213138952091980
228112583430566
234964427733237
247475081663797
257200052879724
271339422472573
282300128917961
285723212427361
286385998464644
```

High-ordered bits are well-compressible, low-order bits not (RLE)

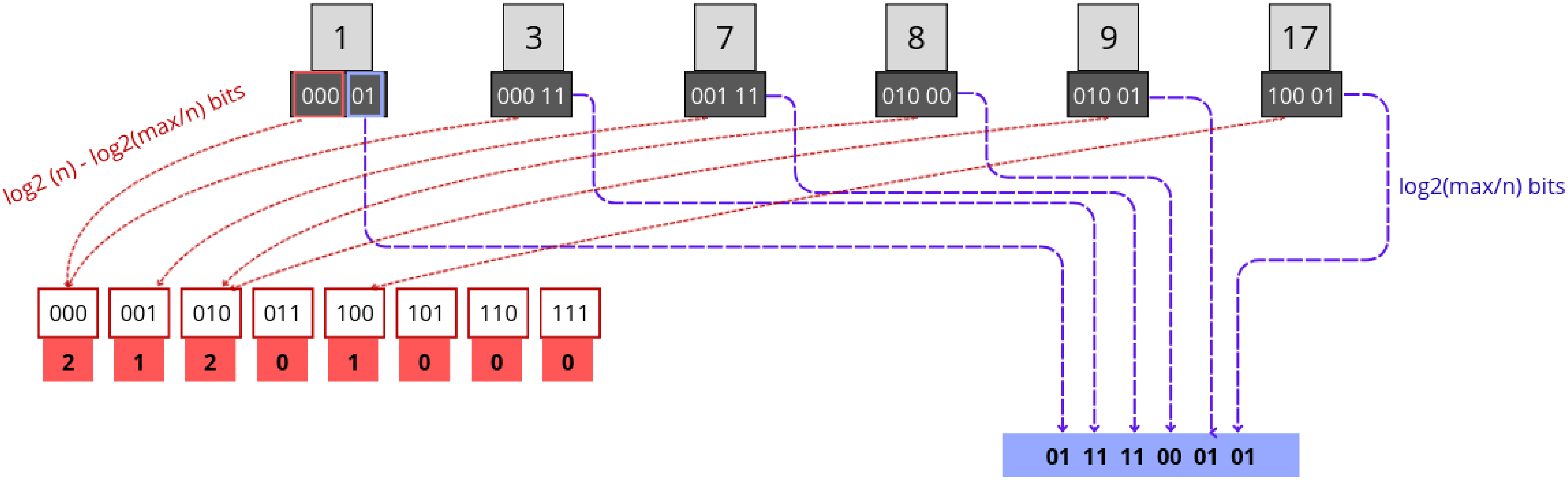
Elias-Fano compression



Elias-Fano compression



Elias-Fano compression



1	3	7	8	9	17
000 01	000 11	001 11	010 00	010 01	100 01

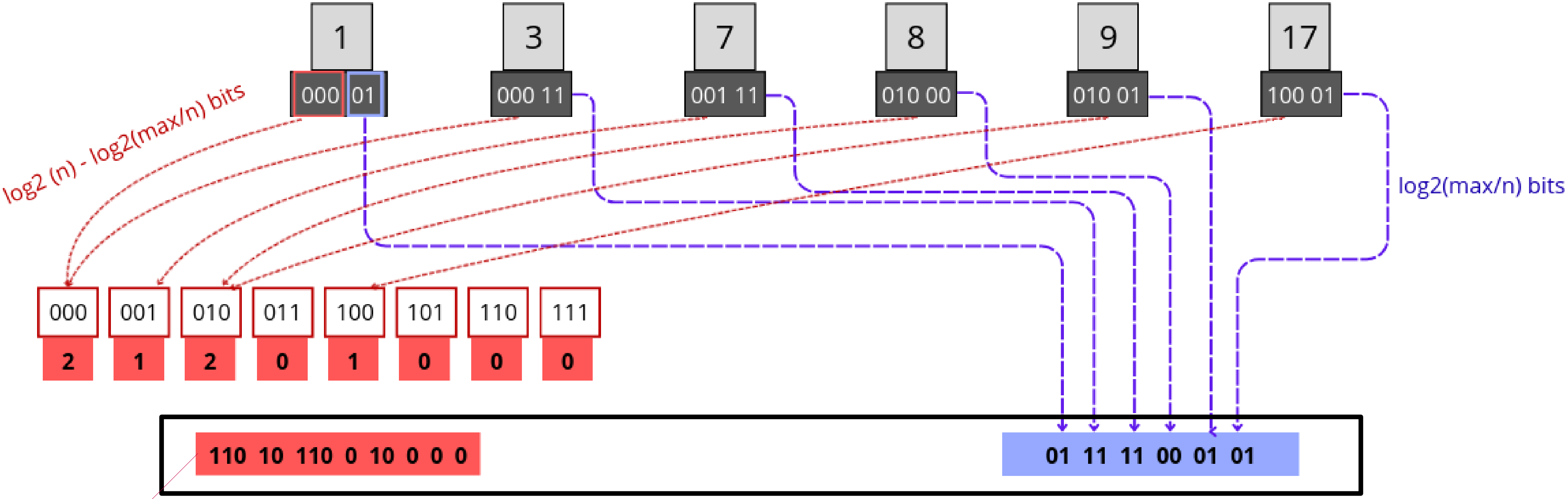
log2 (n) - log2(max/n) bits

log2(max/n) bits

000	001	010	011	100	101	110	111
2	1	2	0	1	0	0	0

01 11 11 00 01 01

Elias-Fano compression

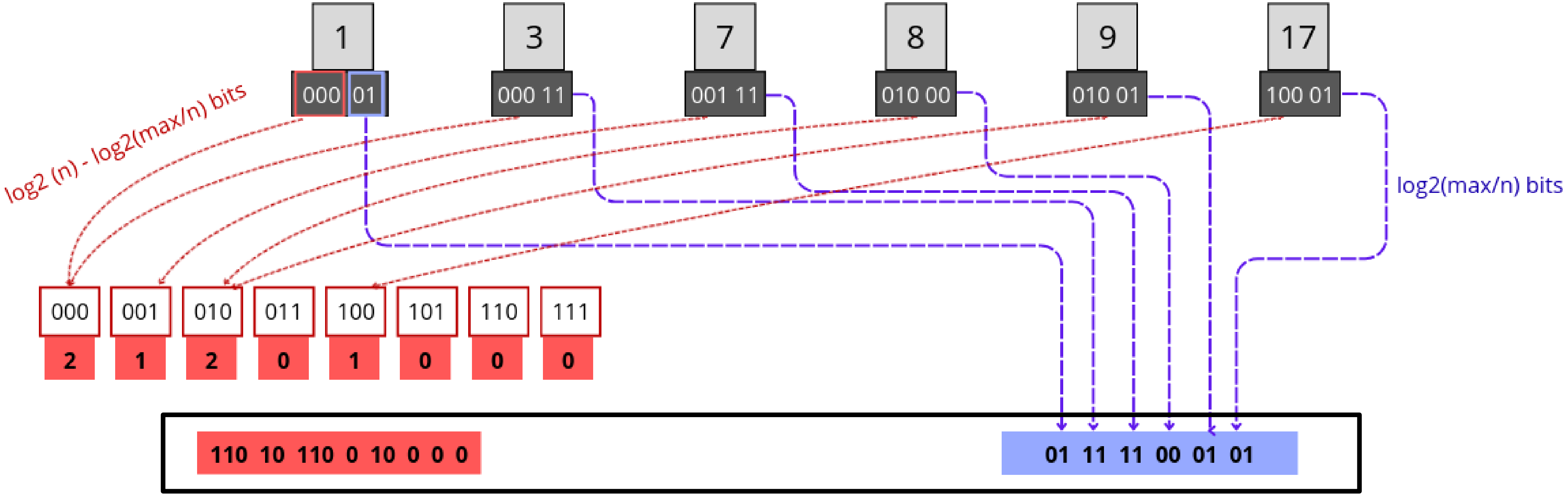


000	001	010	011	100	101	110	111
2	1	2	0	1	0	0	0

Final compressed representation

Unary encoding of
the counters

Elias-Fano compression

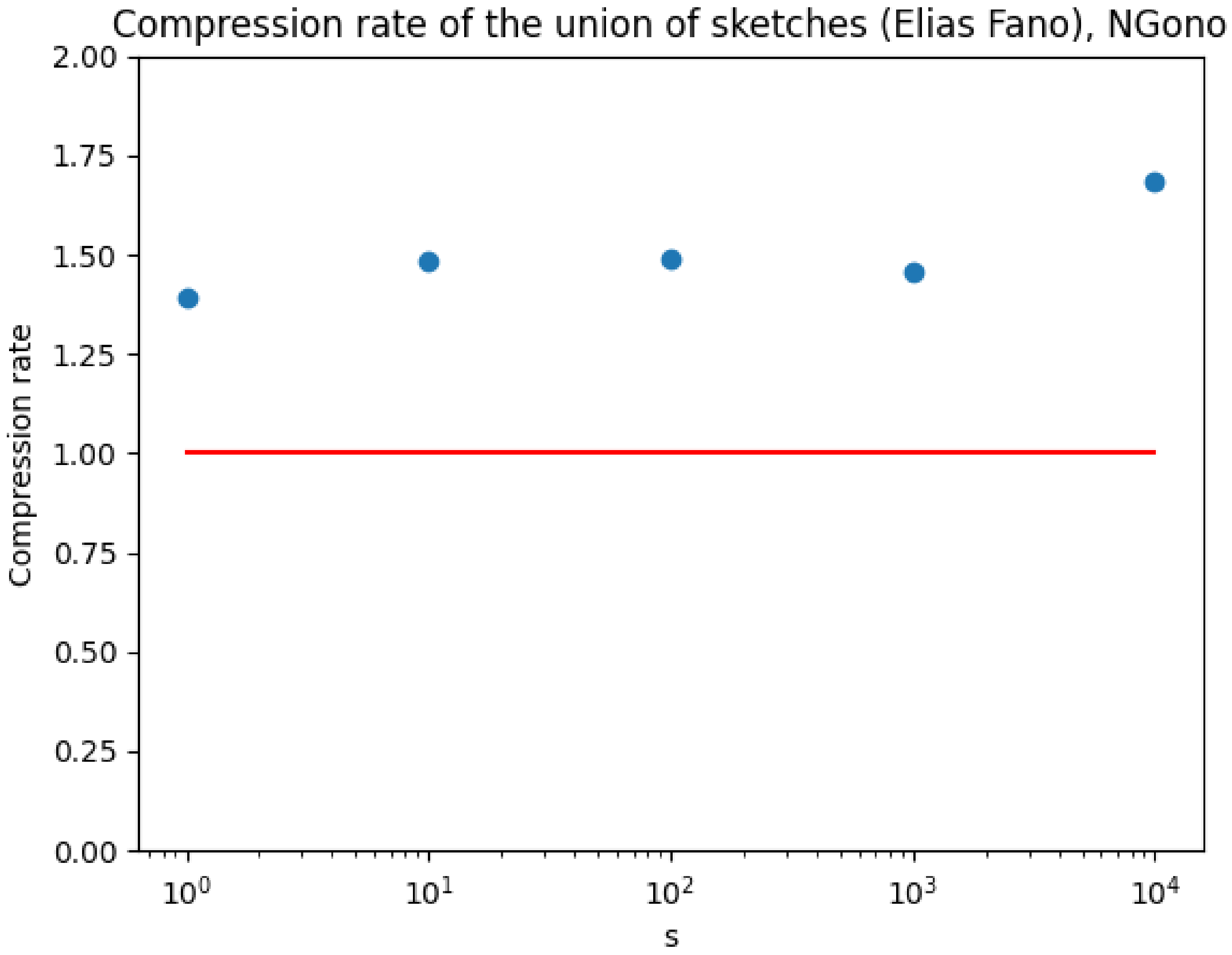


Final compressed representation

$$2n + n(\lceil \log_2(m) \rceil - \lceil \log_2(n) \rceil) \text{ bits}$$

Elias Fano compression

- |S| = 4000 (number of genomes sketched)
- s in [1, 10000] (log scale)
- Type of genomes sketched :
    - neisseria_gonorrhoeae_01 (part 54)


Compression rate of the union of sketches (Elias Fano), NGono
Compression rate
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
10^0
10^1
10^2
10^3
10^4
s

Conclusion

- Results :
  - Reasonable compression rate with EF
  - Phylogenetic order helps – but a suborder could be considered

- Todo :
  - Look at the total compression rate (RLE + EF)
  - Look into new compression schemes for the union