Formalization of sketches compression

Input

- A list $[S_1, ..., S_n]$ where S_i is an integer set representing a genome sketch
- A phylogenetic tree T where the leaves are $S_1, ..., S_n$ when read left to right

Output

A losslessly compressed representation of this list[a]

[a] to be specified

Hypothesis

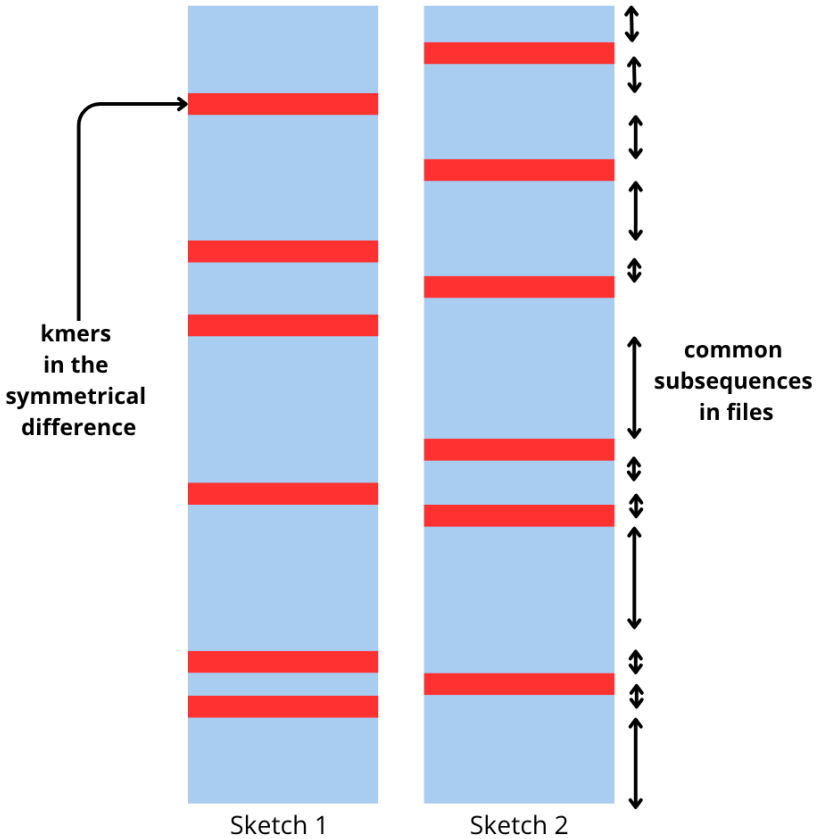
infinite computational resources

Left-to-right compression : gzip

Output

$[S_1, ..., S_n]$ gzip-ed

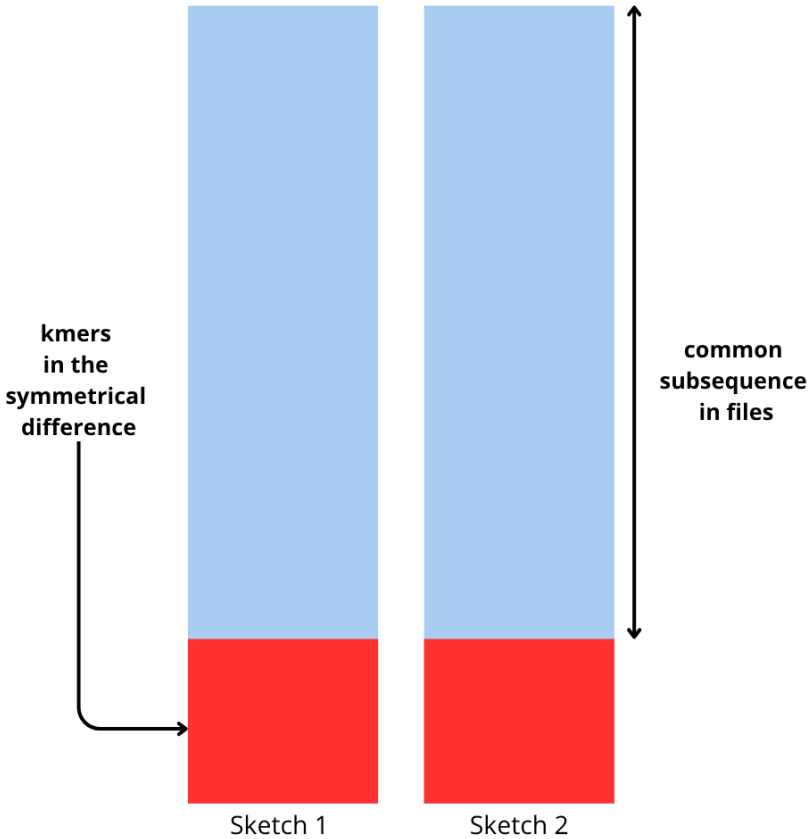
gzip : 2 sets



s	sketch size
S_i	sketch $i, i \in \{1, 2\}$
x	k-mers in $S_i \setminus (S_1 \cap S_2)$

$2x$	common subsequences
$\frac{s}{2x}$	avg common subsequence lgth
$3x$	subsequences to encode

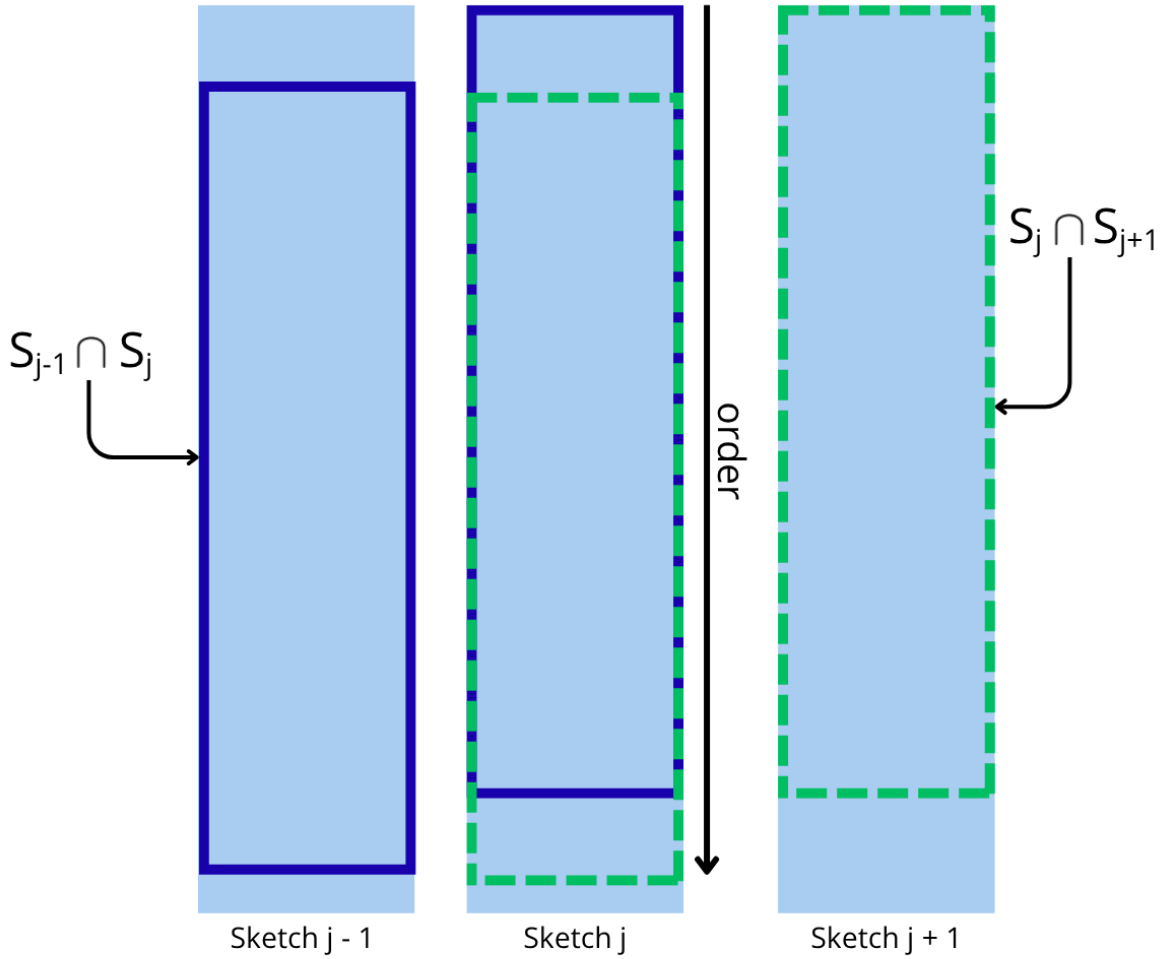
gzip : 2 sets



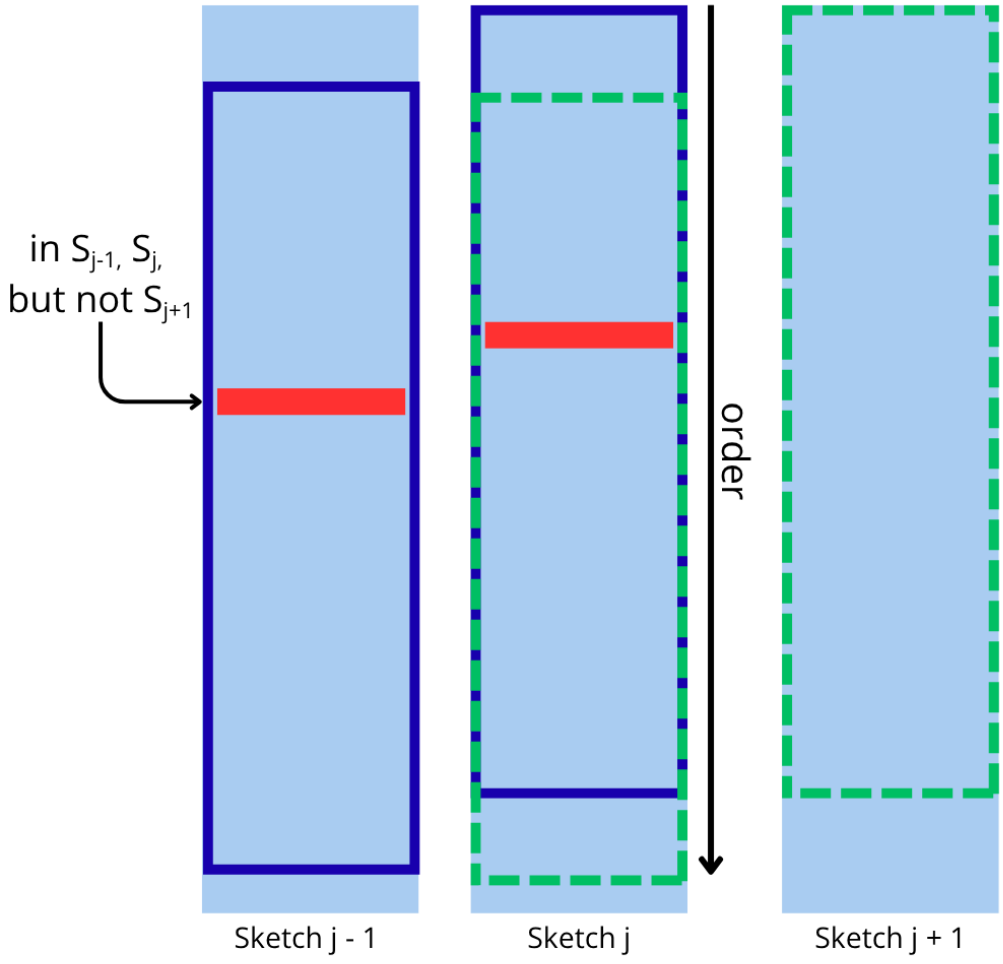
s	sketch size
S_i	sketch $i, i \in \{1, 2\}$
x	k-mers in $S_i \setminus (S_1 \cap S_2)$

1	common subsequence
$s - x$	avg common subsequence lgth
$x + 1$	subsequences to encode

gzip : n sets



gzip : n sets



Bottom-up compression

Output

a list $[S'_1, ..., S'_{2n-1}]$ of $2n-1$ integer sets such that

$$\forall N \in T, S'_N = I_N \setminus I_{parent(N)}$$

where $I(N)$ is inductively defined by

- $\forall L \in Leaves(T), I_L = S_L$
- $\forall N(left, right) \in T, I_N = I_{left} \cap I_{right}$

Notation

For each node N in T, we write S_N (resp. I_N) $S_{i(N)}$ (resp. $I_{i(N)}$) where $i(N)$ is the number of N in the prefix ordering of T

Bottom-up compression