Fractional Hitting Sets

Igor Martayan[1] Timothé Rouzé[2] Camille Marchet[2] Antoine Limasset[2]

[1]ENS Rennes, Univ. Rennes, France [2]Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, Lille, France

Sketching algorithms play an essential role in the analysis of large genomic datasets.
In particular, minimizers are widely used for partitioning *k*-mers from a sequence into buckets.

Sampling *k*-mers with minimizers

Minimizer

smallest *m*-mer of a *k*-mer according to some order (e.g. lexicographic)

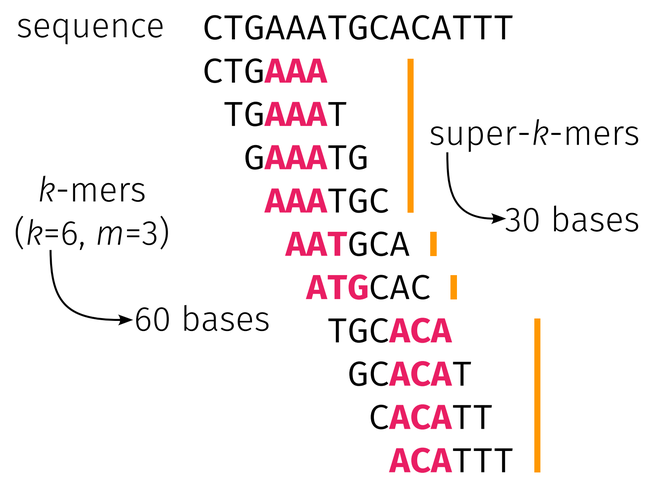
We typically order *m*-mers based on their hashes.

width parameter: $w = k - m + 1$

Super-*k*-mer

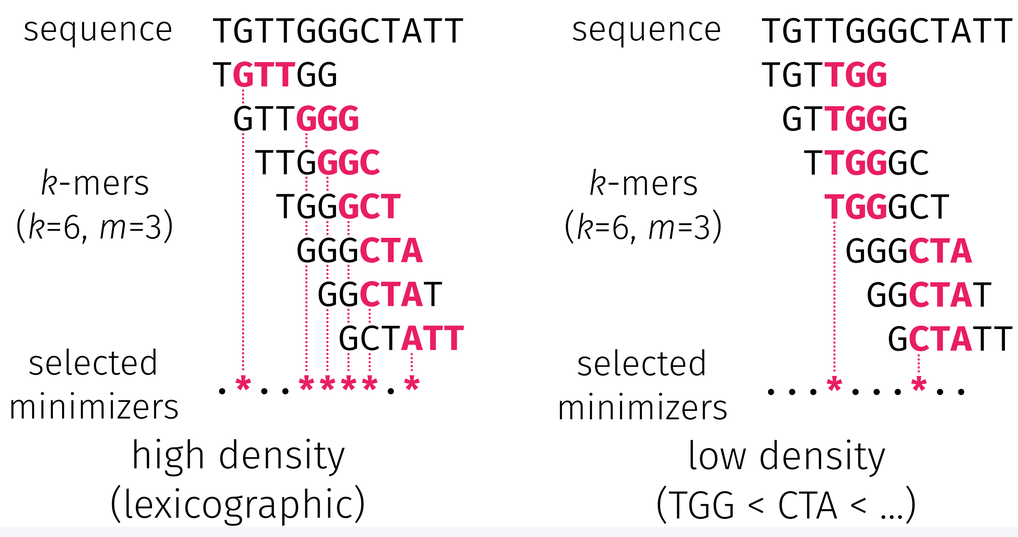
run of consecutive *k*-mers sharing the same minimizer

Super-*k*-mers provide a space-efficient representation of consecutive *k*-mers.



We use minimizers as a footprint for selecting super-*k*-mers.

Density of minimizer schemes & Universal Hitting Sets



lower density $\iff$ longer super-*k*-mers

Density

$$d = \frac{\#\text{selected minimizers}}{\#m\text{-mers}}$$

- optimal density: $1/w$
- expected density with a random order: $2/(w+1)$

Universal Hitting Set

set S of m-mers s.t. each run of w consecutive m-mers has $\geqslant 1$ element in S

minimizers form an UHS

In any UHS, $d \geqslant \dfrac{1.5}{w+1}$

Can we cross this lower bound by relaxing some constraints?

Fractional Hitting Sets

Instead of covering all *k*-mers, we cover a fraction f of them.

Fractional Hitting Set

set S of m-mers s.t. each run of w consecutive m-mers has $\geqslant 1$ element in S *with probability* $\geqslant f$

In practice, we sample minimizers with hashes smaller than

$$t = \left[1 - (1-f)^{1/w}\right] \cdot 4^m$$

We call them *small minimizers*.

Restricted density upper bound

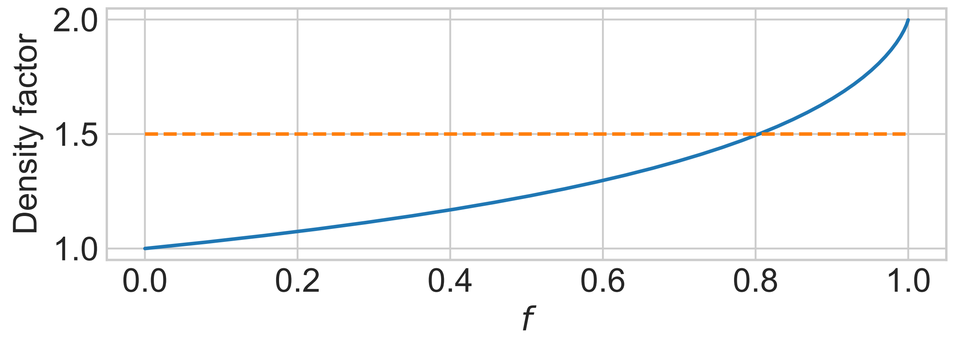
Given a covering fraction f, assuming $m > (3+\varepsilon)\log_4 w$, when restricting to *k*-mers containing small minimizers,

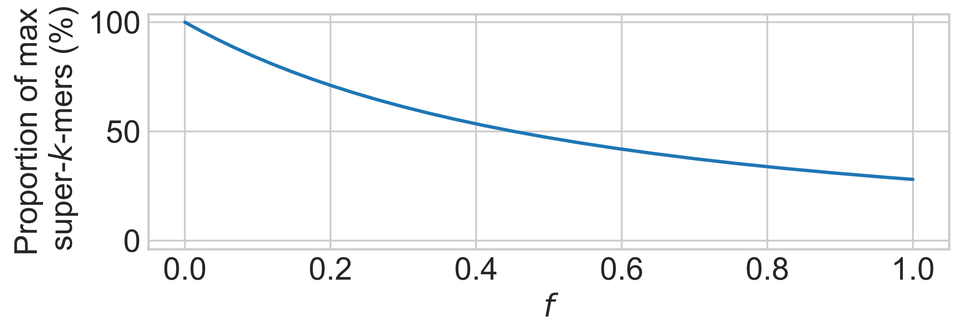
$$d \leqslant 2 \cdot \frac{f + (1-f)\ln(1-f)}{f^2(w+1)} + o(1/w)$$

Proportion of maximal super-*k*-mers

The average proportion of maximal super-*k*-mers is

$$\left[\left(1 - \frac{1}{w}\right)\frac{f}{1+f}\right]^2 + \frac{1 - f(1-2/w)}{1+f}$$





This approach can be combined with existing methods for building UHS: instead of sampling minimizers, we can sample elements from a UHS.

If a UHS U has density d_U, sampling minimizers from U leads to $d \leqslant f \cdot d_U + o(1/w)$

For practical applications of FHS, check out Timothé Rouzé's poster on SuperSampler!

[1] Timothé Rouzé, Igor Martayan, Camille Marchet, and Antoine Limasset. Fractional hitting sets for efficient and lightweight genomic data sketching. In *Workshop on Algorithms in Bioinformatics*, 2023.

[2] Hongyu Zheng, Carl Kingsford, and Guillaume Marçais. Improved design and analysis of practical minimizers. *Bioinformatics*, 36(Supplement_1):i119–i127, 2020.

igor.martayan@ens-rennes.fr