Fractional Hitting Sets

Efficient and Lightweight Genomic Data Sketching

Timothé Rouzé, Igor Martayan, Camille Marchet & Antoine Limasset

September 5, 2023

Before we start

Timothé Rouzé

Antoine Limasset

Camille Marchet

slides

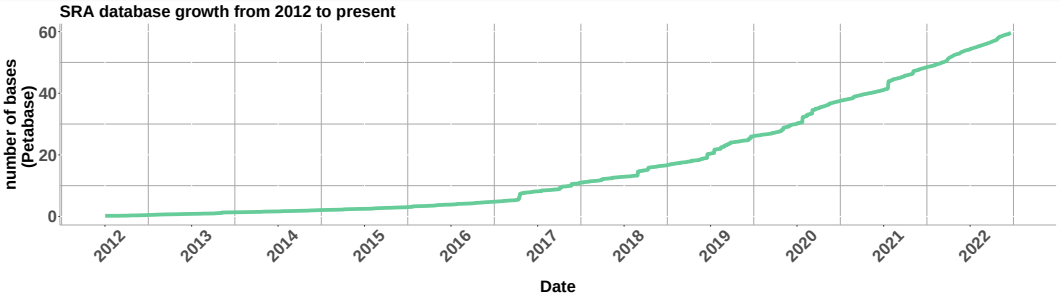
@IgorMartayan

@imartayan@genomic.social

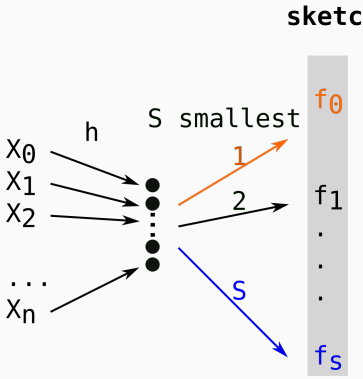
Outline

- Reminders on sketching & minimizers
- Fractional Hitting Sets
- SuperSampler, a sketching tool based on super-k-mers
- Experimental results
- Take home messages

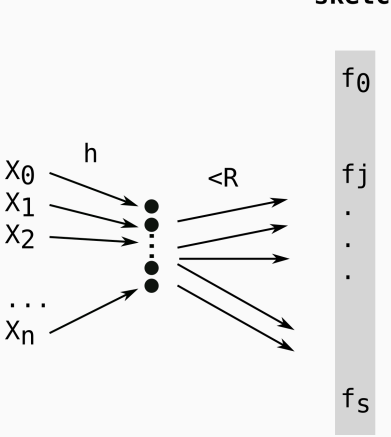
Bionformatician's Moore's law



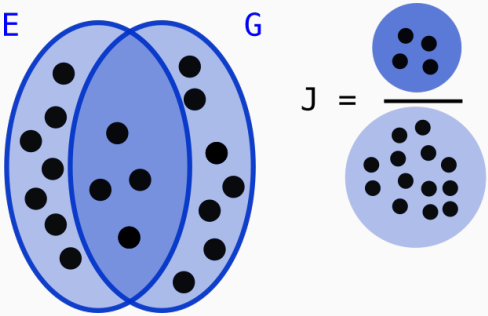
Bottom Minhash in MASH

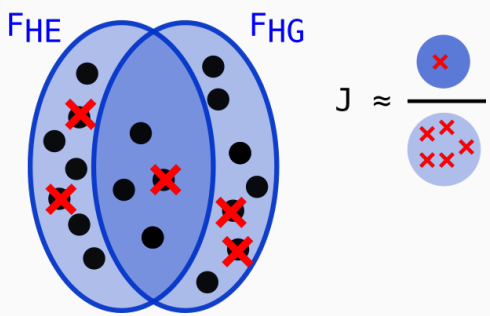


Scaled MinHash in Sourmash



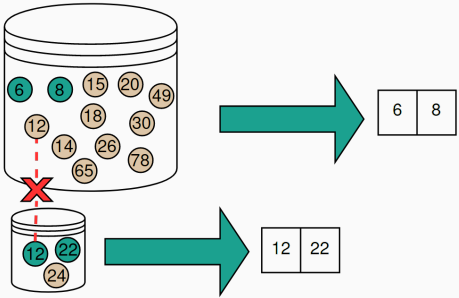
Approximating Jaccard Index



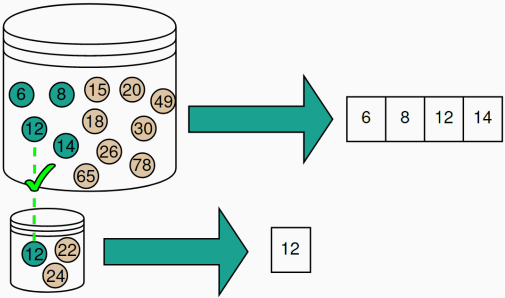


Fixed-size vs Scaled-size Sketching

Fixed size sketch



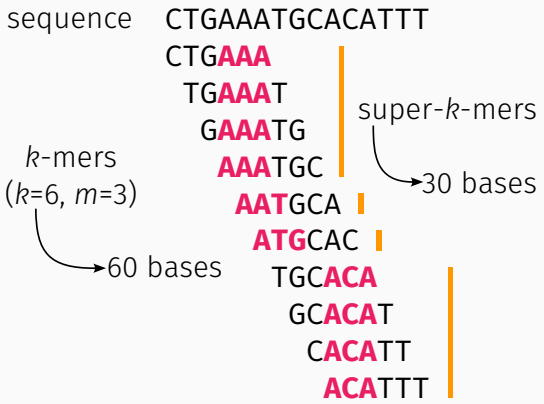
Scaled size sketch



Minimizer

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

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



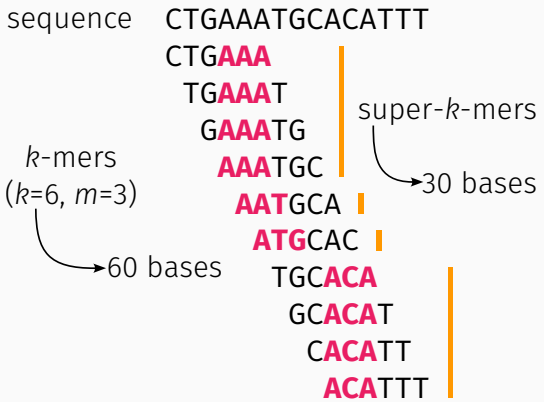
Minimizer

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

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

Super-k-mer

run of consecutive k-mers sharing the same minimizer



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

Density of minimizers

We want a sparse minimizer set

Density

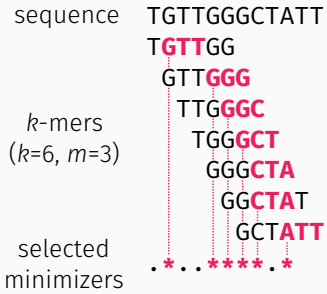
$$d = \frac{\#\text{selected minimizers}}{\#m\text{-mers}}$$

Density of minimizers

We want a sparse minimizer set

Density

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

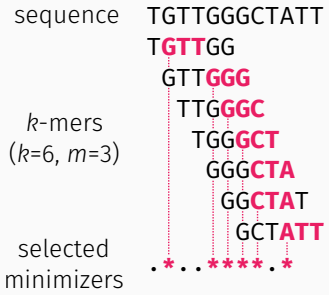


high density
(lexicographic order)

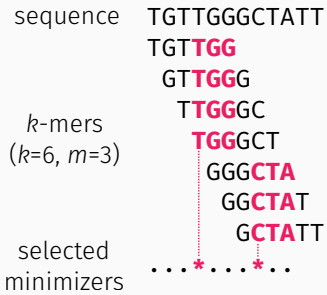
We want a sparse minimizer set

Density

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



high density
(lexicographic order)

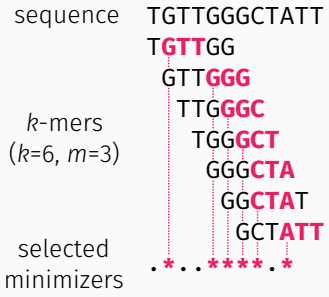


low density
(TGG < CTA < ...)

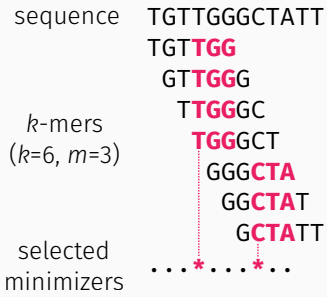
We want a sparse minimizer set

Density

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



high density
(lexicographic order)



low density
(TGG < CTA < ...)

low density $\Longleftrightarrow$ long super-k-mers

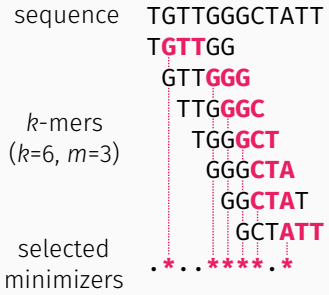
We want a sparse minimizer set

Density

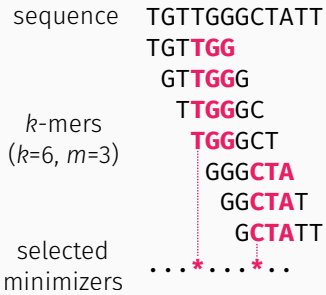
$$d = \frac{\#\text{selected minimizers}}{\#m\text{-mers}}$$

Optimal density: $d = 1/w$

When using a random order, the expected density is $\frac{2}{w+1}$



high density
(lexicographic order)



low density
(TGG < CTA < ...)

low density $\Longleftrightarrow$ long super-k-mers

Universal Hitting Set (UHS)

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

e.g. Decycling sets (Pellow & al., 2022), Miniception (Zheng & al., 2020)

Universal Hitting Set (UHS)

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

e.g. Decycling sets (Pellow & al., 2022), Miniception (Zheng & al., 2020)

Density lower bound

In any UHS, the density is $\geqslant \frac{1.5}{w+1}$ (i.e. the density factor is $\geqslant 1.5$)

Universal Hitting Set (UHS)

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

e.g. Decycling sets (Pellow & al., 2022), Miniception (Zheng & al., 2020)

Density lower bound

In any UHS, the density is $\geqslant \frac{1.5}{w+1}$ (i.e. the density factor is $\geqslant 1.5$)

Can we cross this lower bound by relaxing some constraints?

Fractional Hitting Sets

Fractional Hitting Sets

Instead of covering every k-mer, we cover a fraction f of them

Fractional Hitting Set (FHS)

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

Instead of covering every k-mer, we cover a fraction f of them

Fractional Hitting Set (FHS)

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

In practice, we select minimizers smaller than a certain threshold t

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

minimizers $\leqslant t$ are called small minimizers

Density upper bound

Given a covering fraction f, assuming $m > (3+\varepsilon)\log_4 w$,

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

Density upper bound

Given a covering fraction f, assuming $m > (3+\varepsilon)\log_4 w$,

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

$\oplus$ simple, consistent with known results for $f = 1$

Density upper bound

Given a covering fraction f, assuming $m > (3 + \varepsilon) \log_4 w$,

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

- $\oplus$ simple, consistent with known results for $f = 1$
- $\ominus$ not very meaningful as $f \to 0$
  (since most k-mers are not covered)

Density upper bound

Given a covering fraction f, assuming $m > (3+\varepsilon)\log_4 w$,

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

- $\oplus$ simple, consistent with known results for $f = 1$
- $\ominus$ not very meaningful as $f \to 0$
  (since most k-mers are not covered)

Is there a more meaningful metric?

Restricted density upper bound

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

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

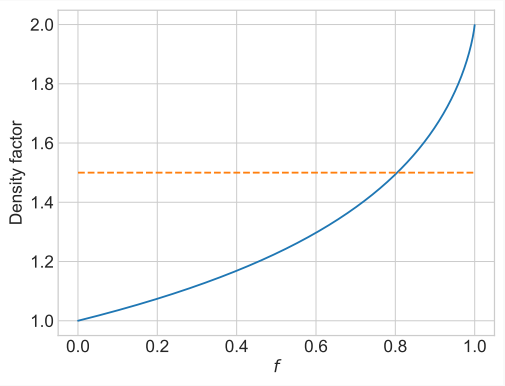
Restricted density upper bound for small minimizers

Restricted density upper bound

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

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

- below the $\frac{1.5}{w+1}$ barrier for $f \leqslant 0.8$
- approaches optimal density as $f \to 0$



Proportion of maximal super-k-mers

The average proportion of maximal super-k-mers is

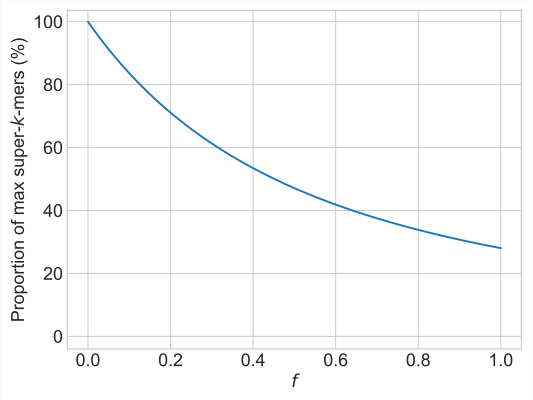
$$\left[\left(1-\frac{1}{w}\right)\frac{f}{1+f}\right]^2+\frac{1-f(1-2/w)}{1+f}$$

Proportion of maximal super-k-mers

The average proportion of maximal super-k-mers is

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

(for $w = 17$)

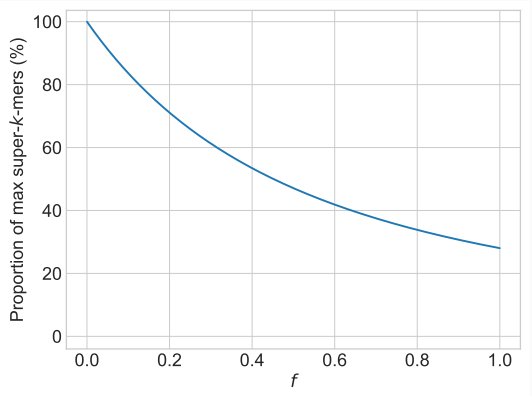


Proportion of maximal super-k-mers

The average proportion of maximal super-k-mers is

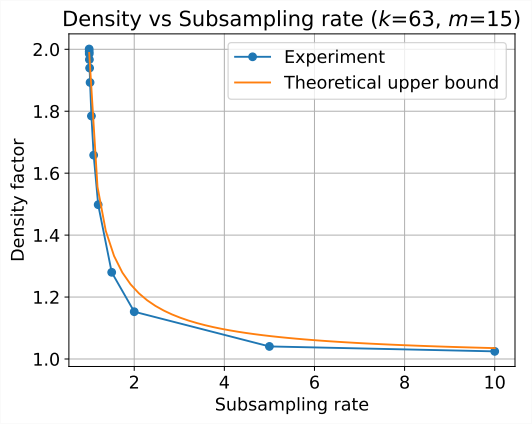
$$\left[\left(1-\frac{1}{w}\right)\frac{f}{1+f}\right]^2+\frac{1-f(1-2/w)}{1+f}$$

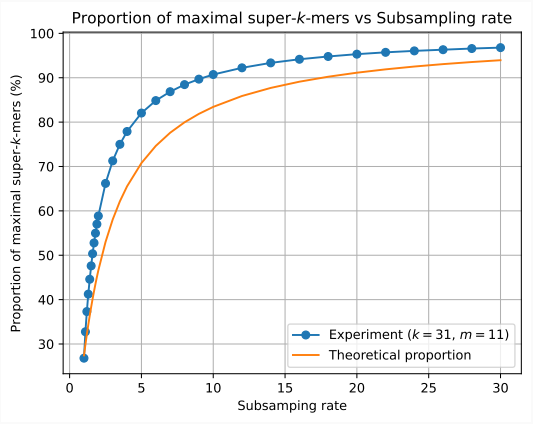
(for $w = 17$)



How accurate is it in practice?

Comparison with experimental results



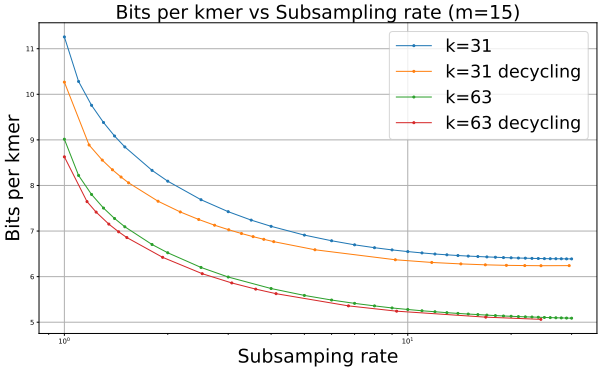


Combining FHS with existing UHS heuristics

Instead of applying a threshold on minimizers, we can:

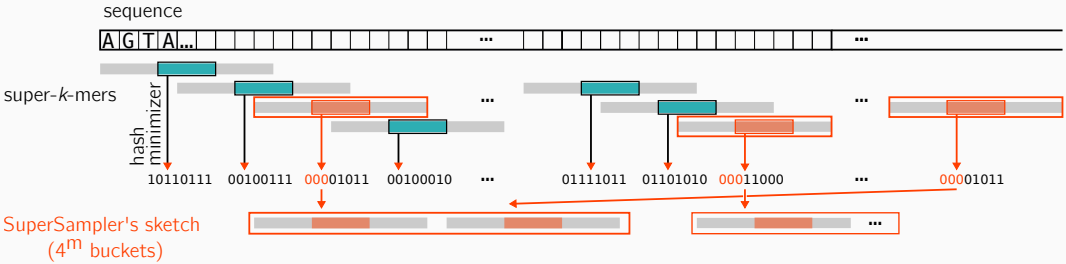
1. build a universal hitting set S (e.g. a decycling set)
2. sample elements from S (by hashing elements and applying a threshold)

Space usage in practice



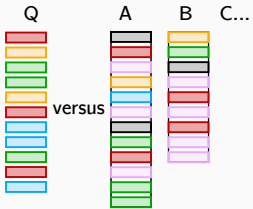
SuperSampler

SuperSampler's sketches



Sketch comparison

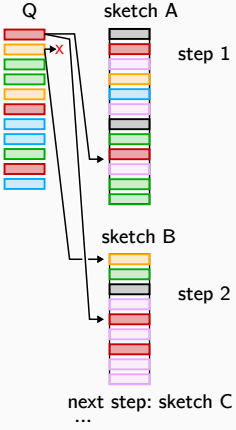
Sketches to compare



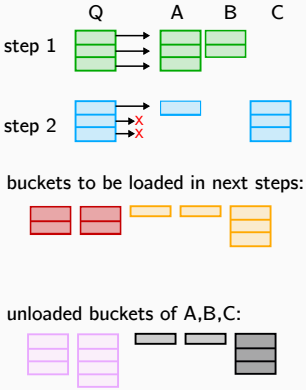
legend

fingerprint
find a match
no match found

Sourmash's Q vs all (A,B,C,...) comparison



SPSP's Q vs all (A,B,C,...) comparison

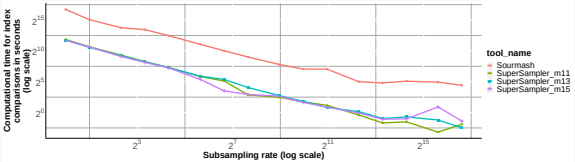


Experimental results vs Sourmash

Performance comparison on dissimilar data (RefSeq)

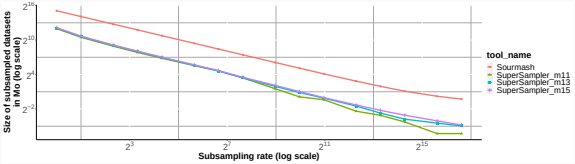
Computational time

≈ 40× faster



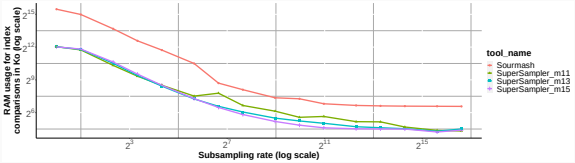
Disk usage

≈ 15× lighter



RAM usage

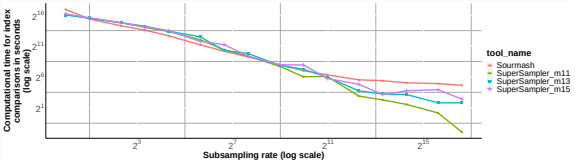
≈ 5× less RAM



Performance comparison on similar data (Salmonellas)

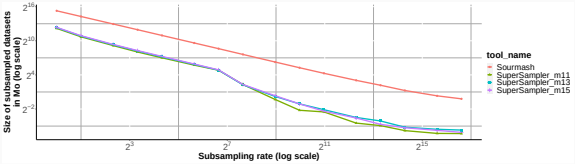
Computational time

similar w/ low sampling,
$\approx 2\times$ faster w/ high sampling



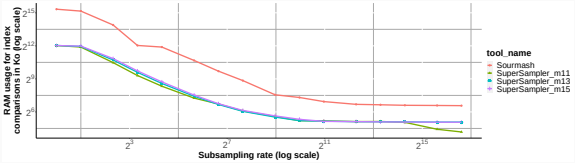
Disk usage

$\approx 40\times$ lighter



RAM usage

$\approx 5\times$ less RAM



Conclusion

Take home messages

Fractional Hitting Sets:

- unify UHS and sketching problems
- lead to lower density / longer super-k-mers
- can benefit from existing UHS building techniques

Super-k-mers:

- provide a space-efficient representation
- speed-up genome sketching & comparison

paper

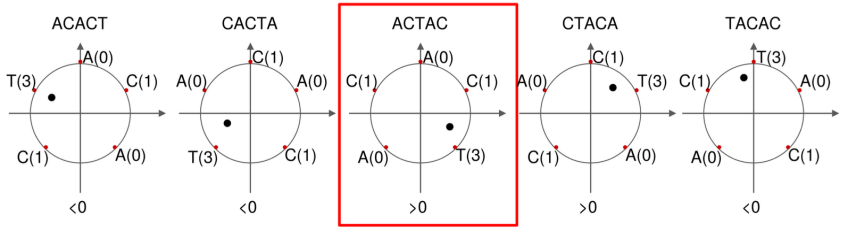
SuperSampler

Appendix

Decycling set

set S of m-mers whose removal make the De Bruijn graph acyclic

- if at least one m-mer is in S, take it in your UHS
- otherwise, use a random order to select a minimizer



Pellow & al., 2022

Density upper bound: sketch of the proof

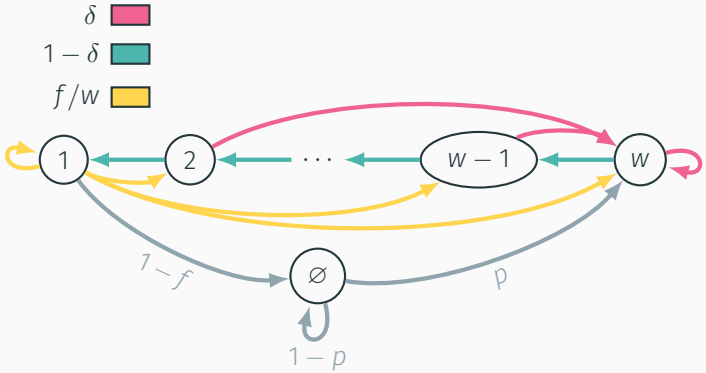
Key property (from Zheng & al., 2020)

Assuming $m > (3 + \varepsilon) \log_\sigma w$, the probability of having duplicate m-mers in a k-mer is negligible

We consider two consecutive k-mers,
the density is equal to the probability that they have different minimizers,
which is the expectation of $\frac{\#\text{small boundary } m\text{-mers}}{\#\text{small } m\text{-mers}}$.

The $\ln$ factor in the restricted density bound comes from a Taylor expansion.

Super-k-mers' Markov chain



- state i: small minimizer starts at i in the k-mer
- state $\varnothing$: no small minimizer in the k-mer