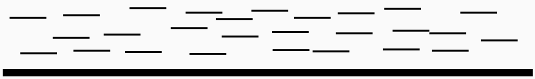
Generalization of the minimizers schemes

Guillaume Marçais, Dan DeBlasio, Carl Kingsford

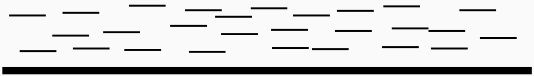
Carnegie Mellon University

Computing read overlaps

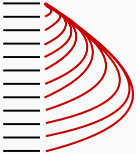


Roberts, *et al.* (2004). Reducing storage requirements for biological sequence comparison.

Computing read overlaps



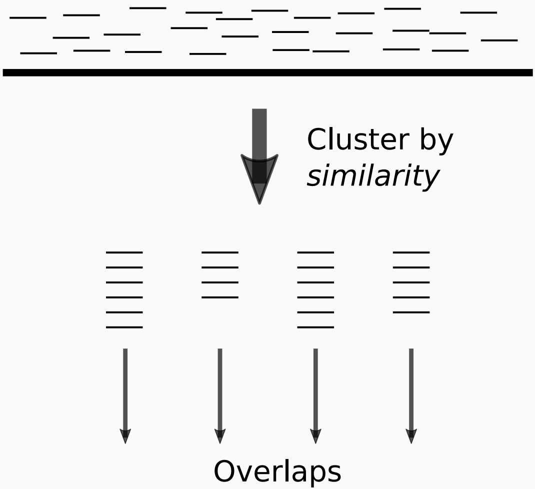
Roberts, *et al.* (2004). Reducing storage requirements for biological sequence comparison.



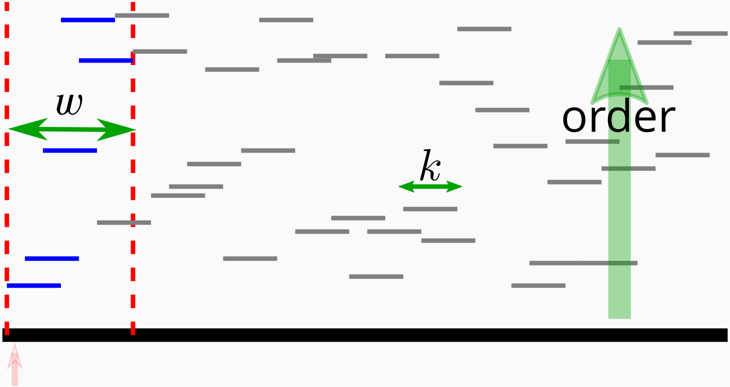
$O(n^2)$ alignments

Computing read overlaps

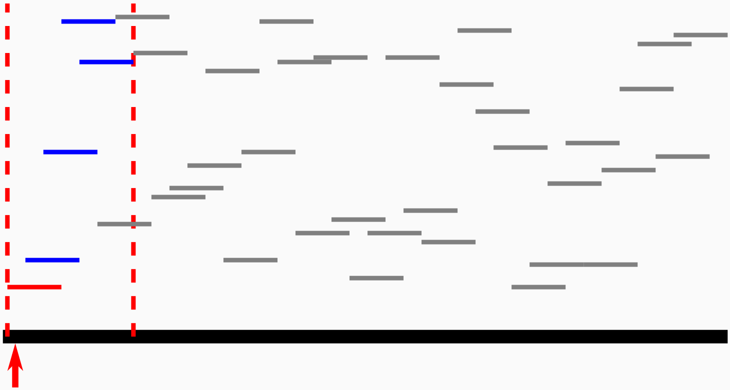
Roberts, *et al.* (2004). Reducing storage requirements for biological sequence comparison.



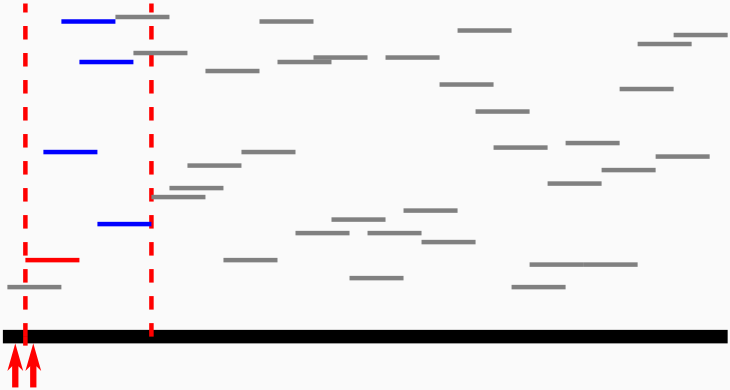
Computing minimizers



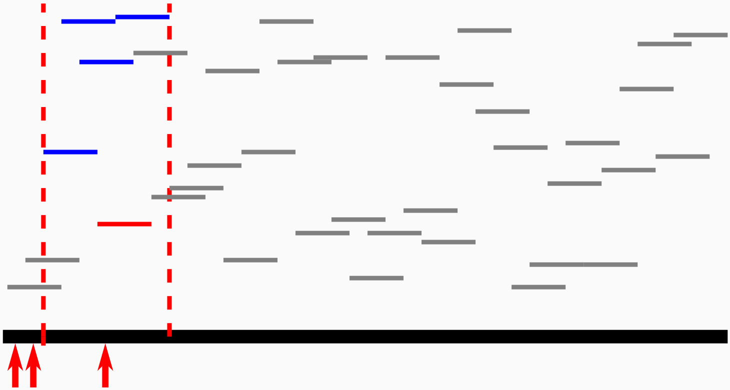
Computing minimizers



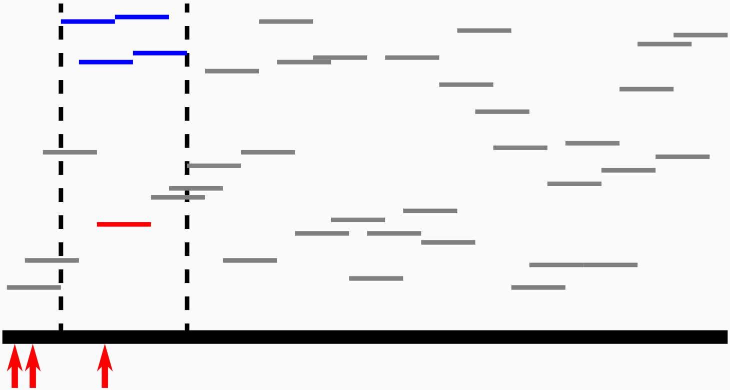
Computing minimizers



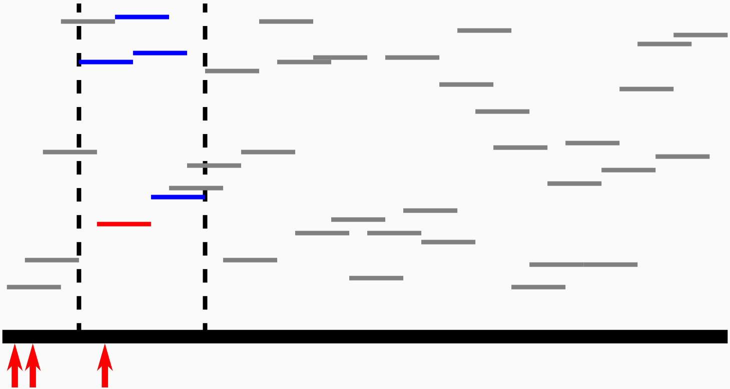
Computing minimizers



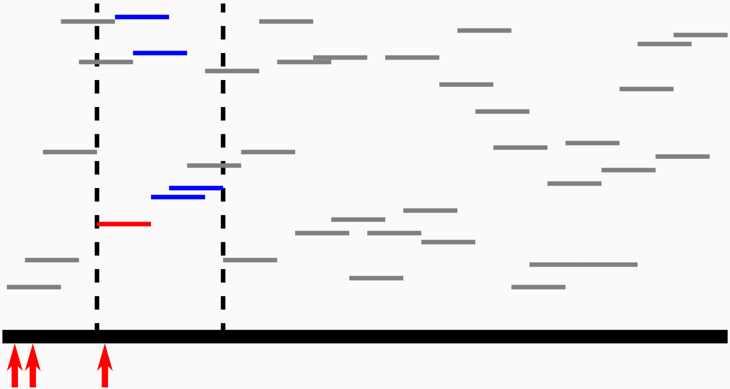
Computing minimizers



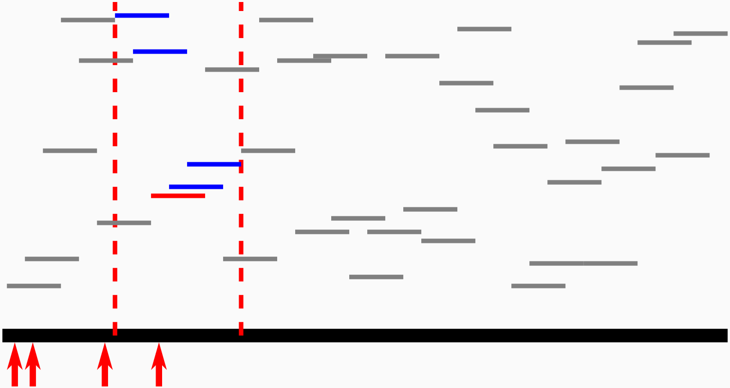
Computing minimizers



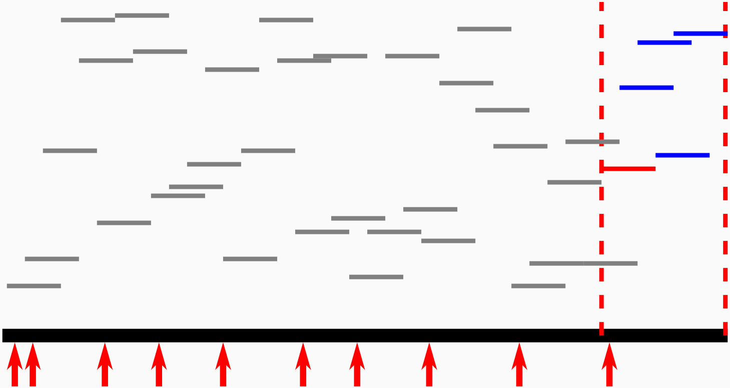
Computing minimizers



Computing minimizers



Computing minimizers



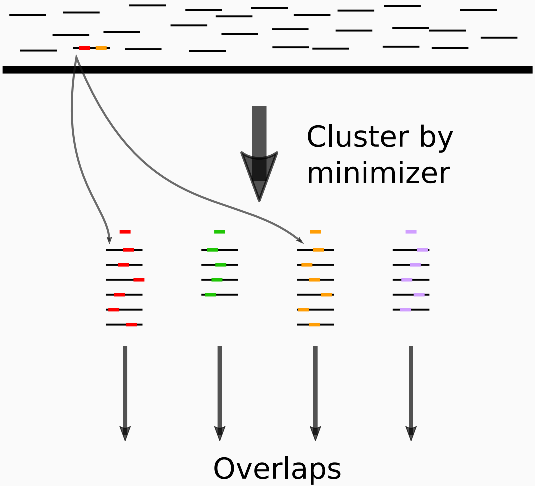
Minimizers definition and properties

Minimizers (k, w, o)
In each window of w consecutive k-mers, select the smallest k-mer according to order o.

1. **Uniform**: distance between selected k-mers is $\leq w$
2. **Deterministic**: two strings matching on w consecutive k-mers select the same minimizer

Computing read overlaps

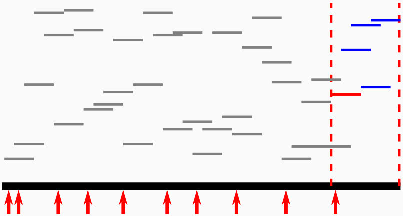
1. **Uniform**: no sequence ignored
2. **Deterministic**: reads with overlap in same bin



Many applications of minimizers

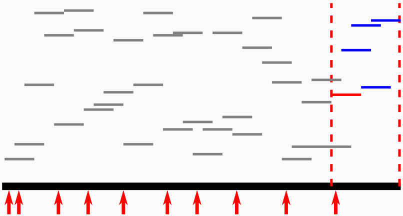
- **UMDOverlapper (Roberts, 2004)**: bin sequencing reads by shared minimizers to compute overlaps
- **MSPKmerCounter (Li, 2015), KMC2 (Deorowicz, 2015), Gerbil (Erber, 2017)**: bin input sequences based on minimizer to count *k*-mers in parallel
- **SparseAssembler (Ye, 2012), MSP (Li, 2013), DBGFM (Chikhi, 2014)**: reduce memory footprint of de Bruijn assembly graph with minimizers
- **SamSAMi (Grabowski, 2015)**: sparse suffix array with minimizers
- **MiniMap (Li, 2016), MashMap (Jain, 2017)**: sparse data structure for sequence alignment
- **Kraken (Wood, 2014)**: taxonomic sequence classifier

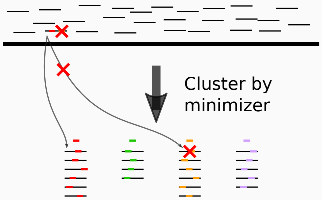
Improving minimizers by lowering density



Density
Density of a scheme is the expected proportion of selected k-mer in a random sequence:

$$d = \frac{\#\text{ of selected } k\text{-mers}}{\text{length of sequence}}$$





Density

Density of a scheme is the expected proportion of selected k-mer in a random sequence:

$$d = \frac{\#\text{ of selected } k\text{-mers}}{\text{length of sequence}}$$

Lower density

$\implies$ smaller bins

$\implies$ less computation

Minimizers density minimizing problem

For fixed k and w:

- Properties “Uniform” & “Deterministic” unaffected by order
- Density changes with ordering o
- Lower density $\implies$ sparser data structures and/or less computation
- Benefit existing and new applications

Density minimization problem
For fixed w, k, find k-mer **order** o giving the lowest expected **density**

Minimizers density minimizing problem

For fixed k and w:

- Properties “Uniform” & “Deterministic” unaffected by order
- Density changes with ordering o
- Lower density $\implies$ sparser data structures and/or less computation
- Benefit existing and new applications

Density minimization problem
For fixed w, k, find k-mer **order** o giving the lowest expected **density**

Density and density factor trivial bounds

Density

$$\underbrace{\frac{1}{w}}_{\text{Pick every other } w \text{ } k\text{-mer}} \leq d \leq \overbrace{1}^{\text{Pick every } k\text{-mer}}$$

$d =$ # of minimizers per base

Density and density factor trivial bounds

Density

$$\underbrace{\frac{1}{w}}_{\text{Pick every other } w \text{ } k\text{-mer}} \leq d \leq \overbrace{1}^{\text{Pick every } k\text{-mer}}$$

$d = $ # of minimizers per base

Density factor

$$1 + \frac{1}{w} \leq df = (w+1) \cdot d \leq w+1$$

$df \approx$ # of minimizers per window

Expected and bound on density

For an *idealized random* order o:

$$d = \frac{2}{w+1} \qquad df = 2$$

Expect ≈ 2 minimizers per window

For any order o:

$$d \geq \frac{1.5 + \frac{1}{2w}}{w+1} \quad df \geq 1.5 + \frac{1}{2w}$$

Requires ≥ 1.5 minimizers per window

Schleimer 2003, Roberts 2004

Expected and bound on density

For an *idealized random* order o:

$$d = \frac{2}{w+1} \qquad df = 2$$

Expect ≈ 2 minimizers per window

Not valid for $w \gg k$

For any order o:

$$d \geq \frac{1.5 + \frac{1}{2w}}{w+1} \quad df \geq 1.5 + \frac{1}{2w}$$

Requires ≥ 1.5 minimizers per window

Valid only for $w \gg k$

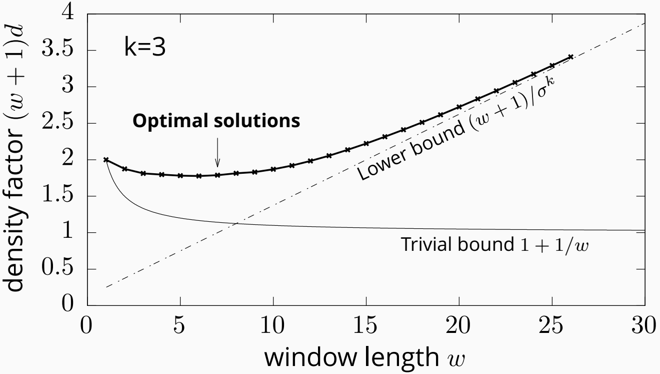
Schleimer 2003, Roberts 2004

Asymptotic behavior in k and w

What is the best ordering possible when:

- w is fixed and $k \to \infty$
- k is fixed and $w \to \infty$

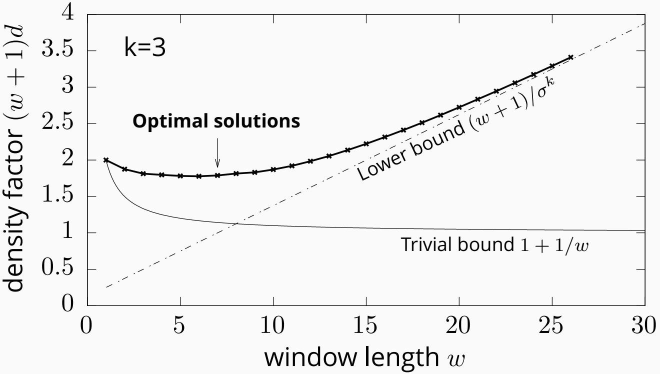
Asymptotic behavior in *w*



$$d \geq \frac{1}{\sigma^k}, \quad df \geq \frac{w+1}{\sigma^k}$$

Density factor is $\Omega(w)$, not constant

Asymptotic behavior in w



$$d \geq \frac{1}{\sigma^k}, \quad df \geq \frac{w+1}{\sigma^k}$$

Density factor is $\Omega(w)$, not constant

Asymptotic behavior in k

Asymptotically optimal minimizers schemes
There exists a sequence of orders $(o_k)_{k \in \mathbb{N}}$ which are asymptotically optimal:

$$d_{o_k} \xrightarrow[k \to \infty]{} \frac{1}{w} \qquad df_{o_k} \xrightarrow[k \to \infty]{} 1 + \frac{1}{w}$$

Optimal vertex cover of the de Bruijn graph (Lichiardopol 2006)
There exists a sequence of vertex cover V_k of DB_k which is asymptotically optimal in size:

$$|V_k| \xrightarrow[k\to\infty]{} \frac{\sigma^k}{2}$$

Optimal depathing of the de Bruijn graph
For a fixed w, there exists a sequence $(U_k)_{k\in\mathbb{N}}$ of sets of k-mers that covers every path of length w in DB_k such that

$$|U_k| \xrightarrow[k\to\infty]{} \frac{\sigma^k}{w}$$

Bound on density

For **all** k, w and order o:

$$d \geq \frac{1.5 + \frac{1}{2w} + \max\left(0, \lfloor \frac{k-w}{w} \rfloor\right)}{w + k}$$

Bound on density

For **all** k, w and order o:

$$d \geq \frac{1.5 + \frac{1}{2w} + \max\left(0, \lfloor \frac{k-w}{w} \rfloor\right)}{w + k}$$

$$df \geq 1 + \frac{1}{w} \qquad \text{for large } k$$

$$df \geq 1.5 + \frac{1}{2w} \qquad \text{for large } w$$

Density factor of minimizers

Asymptotic behavior of minimizers is fully characterized:

- Minimizers scheme is optimal for large k: $df \xrightarrow[k\to\infty]{} 1 + \frac{1}{w}$
- Minimizers scheme is not optimal for large w: $df = \Omega(w)$
- Better lower bound on d

Density factor of minimizers

Asymptotic behavior of minimizers is fully characterized:

- Minimizers scheme is optimal for large k: $df \xrightarrow[k\to\infty]{} 1 + \frac{1}{w}$
- Minimizers scheme is not optimal for large w: $df = \Omega(w)$
- Better lower bound on d

Good:

- First example of optimal minimizers scheme
- Constructive proof

Not good:

- Large k less interesting in practice
- Minimizers **don't** have **constant** density factor

Generalizing minimizers: local and forward schemes

Local scheme
Given $f : \Sigma^{w+k-1} \to [0, w-1]$, for each window ω, select k-mer at position $f(\omega)$.

Generalizing minimizers: local and forward schemes

Local scheme
Given $f : \Sigma^{w+k-1} \to [0, w-1]$, for each window ω, select k-mer at position $f(\omega)$.

Minimizers scheme with order o is a local scheme where $f = \arg\min_{i \in [0,w-1]} o(\omega[i : k])$

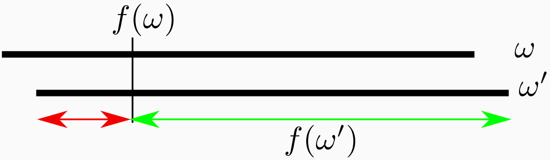
Generalizing minimizers: local and forward schemes

Local scheme
Given $f : \Sigma^{w+k-1} \to [0, w-1]$, for each window ω, select k-mer at position $f(\omega)$.

Minimizers scheme with order o is a local scheme where

$$f = \arg\min_{i \in [0, w-1]} o(\omega[i : k])$$

Forward scheme
Local scheme such that $f(\omega') \geq f(\omega) - 1$ if suffix of ω' equals prefix of ω



Local & forward as better minimizers schemes

$$\text{Minimizers} \subsetneq \text{Forward} \subsetneq \text{Local}$$

- Properties “Uniform” & “Deterministic” also satisfied
- Drop-in replacement for minimizers
- Potential for lower density

Density factor overview

Density factor df			
	$k \to \infty$	$w \to \infty$	
Scheme		Best	Bound
Minimizers			
Forward			
Local			

Density factor overview

Density factor *df*

Scheme	$k \to \infty$	$w \to \infty$ Best	$w \to \infty$ Bound
Minimizers	$1 + \frac{1}{w}$	$O(w)$	$\Omega(w)$
Forward			
Local			

Density factor overview

Density factor *df*			
	$k \to \infty$	$w \to \infty$	
Scheme		Best	Bound
Minimizers	$1 + \frac{1}{w}$	$O(w)$	$\Omega(w)$
Forward	$1 + \frac{1}{w}$	$O(\sqrt{w})$	$\sim 1.5 + \frac{1}{2w}$
Local			

Density factor overview

Density factor *df*			
	$k \to \infty$	$w \to \infty$	
Scheme		Best	Bound
Minimizers	$1 + \frac{1}{w}$	$O(w)$	$\Omega(w)$
Forward	$1 + \frac{1}{w}$	$O(\sqrt{w})$	$\sim 1.5 + \frac{1}{2w}$
Local	$1 + \frac{1}{w}$	$O(\sqrt{w})$	$1 + \frac{1}{w}$

Conclusion: the quest for constant density factor

- Minimizers schemes **can't** achieve **constant** density factor
- Local and forward schemes **may** achieve **constant** density factor

- Design of optimal orders or functions f still open

Carl Kingsford group:

Dan DeBlasio
Heewook Lee
Natalie Sauerwald
Cong Ma
Hongyu Zheng
Laura Tung
Postdoc position open

Carnegie Mellon University

GORDON AND BETTY MOORE FOUNDATION
GBMF4554

NSF
CCF-1256087
CCF-1319998

NATIONAL INSTITUTES OF HEALTH
R01HG007104
R01GM122935