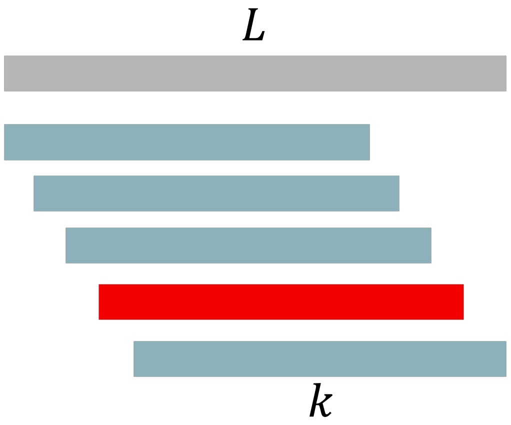
Minimizers and one question about de Bruijn graphs

Gregory Kucherov (CNRS/Univ Gustave Eiffel)

Minimizers: definition

- Consider alphabet A, integers $L > k > 0$, and a linear order on A^k. For $s \in A^L$, the *minimizer* of s is the smallest substring of s of length k



Minimizers in a string

$L = 6, k = 3$
lexicographic order

```
a c t t a g t t g g a a c a a a a a c t
a c t t a g   t g g a a c   a a a a c t
  c t t a g t   g g a a c a
    t t a g t t   g a a c a a
      t a g t t g   a a c a a a
        a g t t g g   a c a a a a
          g t t g g a   c a a a a a
            t t g g a a   a a a a a c
```

- Order can be specified by a hash function $h\colon \Sigma^k \to \mathbb{N}$

References

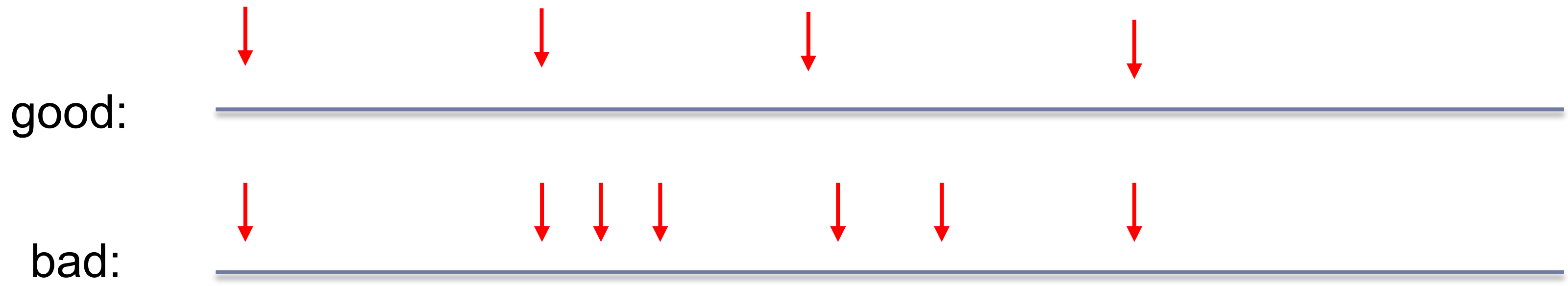
- Credits:
  - Schleimer et al. *Winnowing: local algorithms for document fingerprinting*, SIGMOD Int Conf on Management of Data, 2003
  - Roberts et al. *Reducing storage requirements for biological sequence comparison*, Bioinformatics, 2004
- Applications:
  - L-mer processing:
    - clustering similar L-mers (*locality-sensitive hashing)*
    - L-mer counting [KMC 2015, MSPKmerCounter 2015]
    - metagenomic classification [Kraken 2014]
  - sampling k-mers in a genomic sequence to be used as *seeds* for similarity search
    - read mapping/alignment and assembly [minimap, miniasm 2016, 2018, MashMap 2018], mapping to variation graphs [V-MAP 2019]
    - genome assembly [BCALM 2016 …]
  - stringology tasks: sparse suffix array [SamSAMi 2015]
  - … and more

Sampling

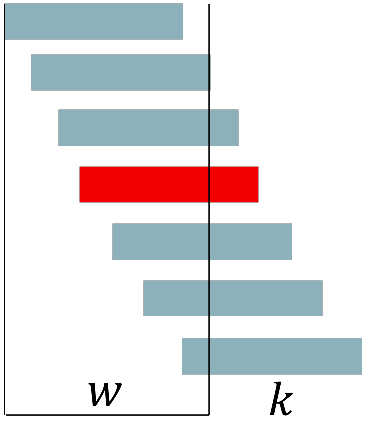
- General goal: sample positions such that
  - consecutive positions cannot be too far away from each other (each L-window contains a position)
  - identical L-windows have the same relative sampled positions
  - positions are distributed as sparsely as possible along the string

Density of minimizers

- We are interested in sparsely distributed minimizers



- $w = L - k + 1$: window of starting positions
- *density* of minimizers : expected density on i.i.d. random sequence ($n \to \infty$)
- [Marçais et al. 17] Given k, w, the density of minimizers equals the density of minimizers on any de Bruijn sequence of order $w + k$



Which order to choose?

- [Schleimer et al. 03, Roberts et al. 04] Assuming that every k-mer from among $w + 1$ consecutive k-mers has equal chance to be minimal, the density of minimizers is $2/(w + 1)$
- lexicographical order performs worse than that
- [Orenstein et al. 17] Expected density of minimizers for $m = w$ can be made below $1.8/(w + 1)$
- [Schleimer et al. 03] Lower bound: $1.5/(w + 1)$

Local selection schemes [Zheng *et al*. 2020]

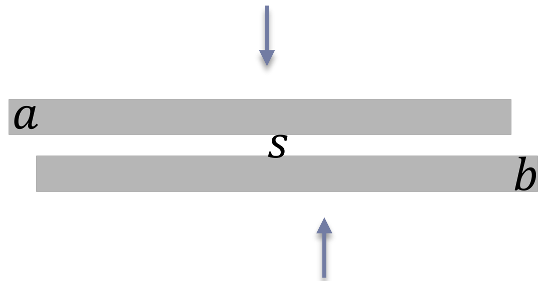
- Selecting a position from among w consecutive positions does not have to be based on an order of k-mers starting at these positions

Local selection schemes [Zheng *et al*. 2020]

- Selecting a position from among w consecutive positions does not have to be based on an order of k-mers starting at these positions
- *Local selection scheme (LSS)*: $f : A^w \rightarrow [1..w]$

Local selection schemes [Zheng *et al*. 2020]

- Selecting a position from among w consecutive positions does not have to be based on an order of k-mers starting at these positions
- *Local selection scheme (LSS)*: $f: A^w \to [1..w]$
- *Forward LSS*: $\forall s \in A^{w-1}, a, b \in A: f(as) \le f(sb) + 1$



Local selection schemes [Zheng *et al*. 2020]

- Selecting a position from among w consecutive positions does not have to be based on an order of k-mers starting at these positions
- *Local selection scheme (LSS)*: $f: A^w \rightarrow [1..w]$
- *Forward LSS*: $\forall s \in A^{w-1}, a, b \in A: f(as) \leq f(sb) + 1$
- *Density on string s*: fraction of selected positions for all windows $s[i..i+w-1]$
- [Zheng *et al*. 20] Density on a random i.i.d. string = density on a de Bruijn string of order $2w-1$ (general) or $w+1$ (forward)

Local selection schemes [Zheng *et al*. 2020]

- Selecting a position from among w consecutive positions does not have to be based on an order of k-mers starting at these positions
- *Local selection scheme (LSS)*: $f : A^w \to [1..w]$
- *Forward LSS*: $\forall s \in A^{w-1}, a, b \in A: f(as) \leq f(sb) + 1$
- *Density on string s*: fraction of selected positions for all windows $s[i..i + w - 1]$
- [Zheng *et al*. 20] Density on a random i.i.d. string = density on a de Bruijn string of order $2w - 1$ (general) or $w + 1$ (forward)
- [Zheng *et al*. 20] There is a forward LSS with density $O(\log w / w)$

Lexicographically smallest rotation LSS

- $f(s[1..w])$ = starting position of the lexicographically smallest circular shift of $s[1..w]$

Lexicographically smallest rotation LSS

- $f(s[1..w])$ = starting position of the lexicographically smallest circular shift of $s[1..w]$
- this is *not* a forward LSS

```
  ↓
001000100
 010001001
   ↑
```

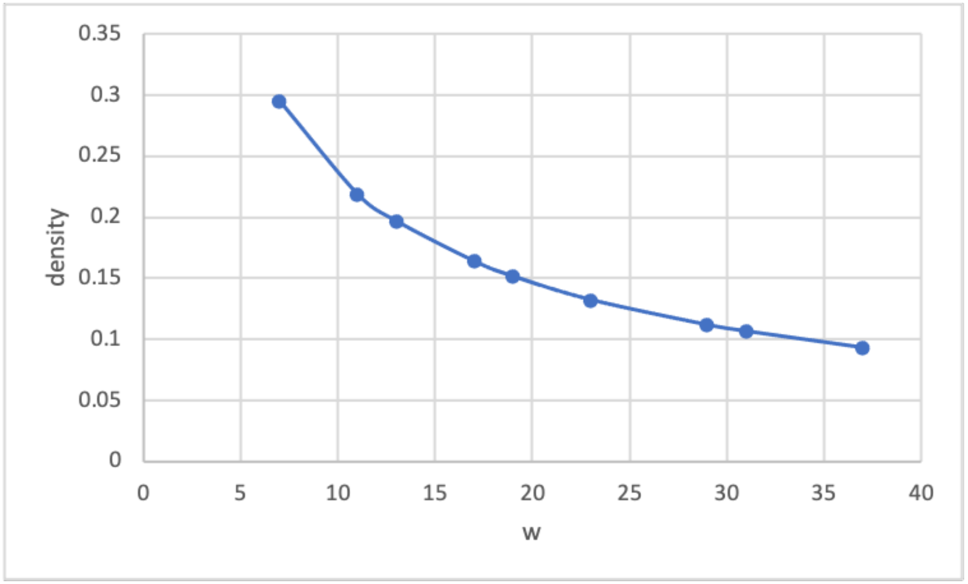
Lexicographically smallest rotation LSS

- $f(s[1..w])$ = starting position of the lexicographically smallest circular shift of $s[1..w]$
- this is *not* a forward LSS

```
001000100
 010001001
```

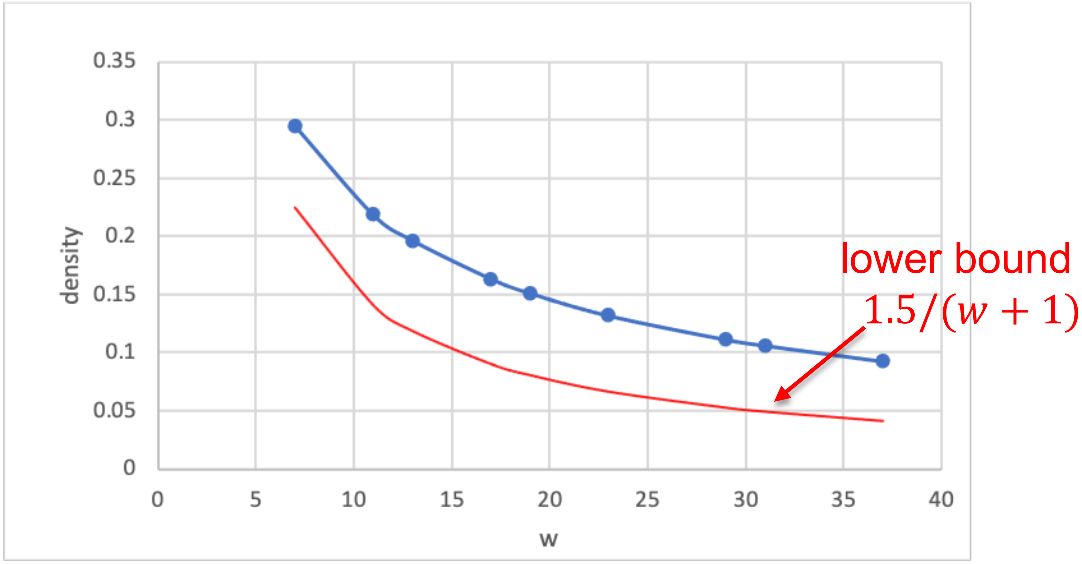
Question: what is the density produced by this LSS?

Lexicographically smallest rotation: experiment



Density of selected positions by lexicographic smallest rotation scheme on binary alphabet

Lexicographically smallest rotation: experiment



Density of selected positions by lexicographic smallest rotation scheme on binary alphabet

Questions

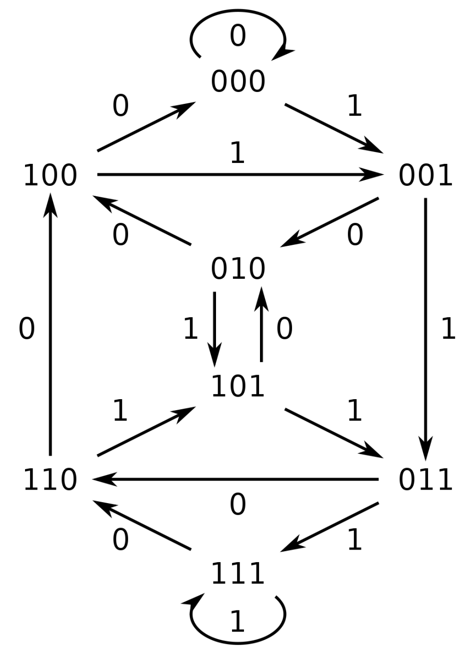
- What is the asymptotic density produced by the smallest rotation scheme? Is it $O(\frac{1}{w})$?
- What about other (better?) schemes?
- What about forward schemes? Is $O(\frac{\log w}{w})$ the tight bound? Can we resolve the constant factor?

de Bruijn graph framework

- The number of conjugacy classes is

$$C(w) = \frac{1}{w}\sum_{d|w}\phi(\frac{w}{d})2^d = \frac{2^w}{w}(1+o(1))$$

  where ϕ is Euler's totient function



- [Mykkelveit 72] There exists an unavoidable subset $S \subseteq A^w$ with $|S| = C(w)$

  (cf also [Champarnaud *et al*. 04])
- Equivalently, the *decycling number* of a de Bruijn graph is $C(w)$
- We need more than breaking all cycles
- …

Thanks!