

Practical universal k -mer sets for minimizer schemes

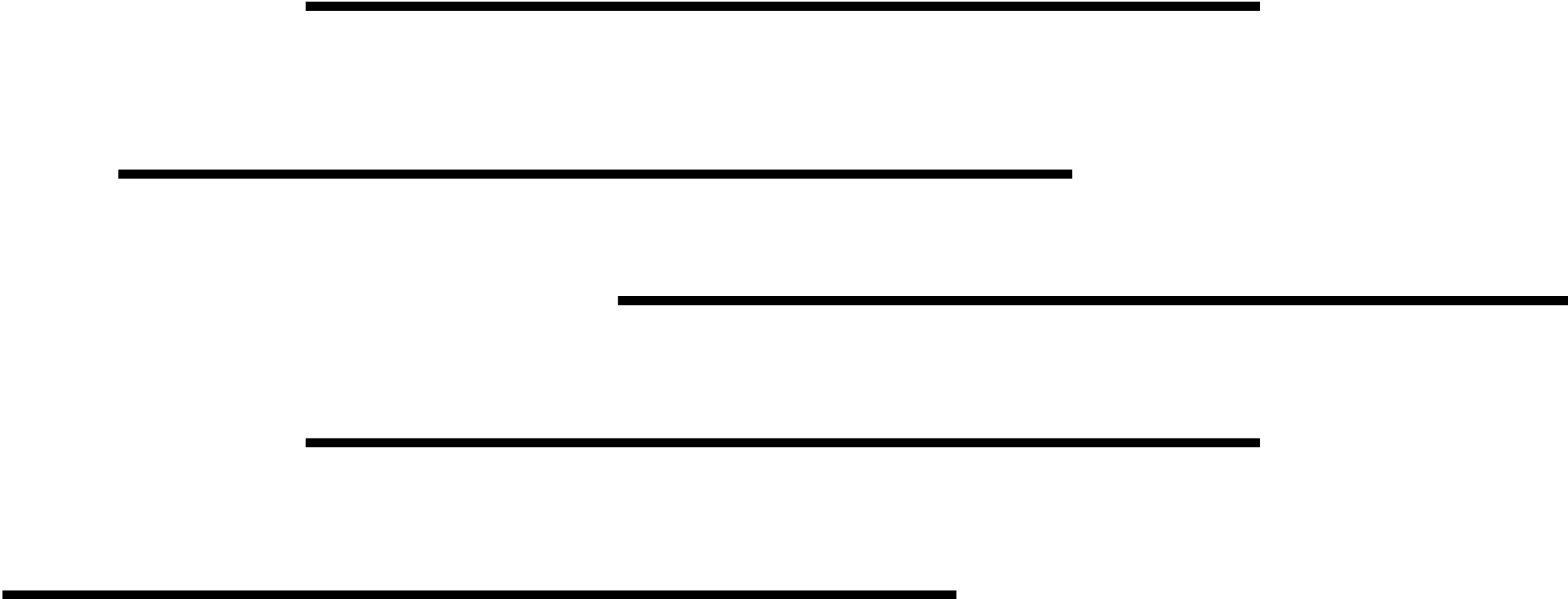
Dan DeBlasio, Fiyin Gbosibo, Carl Kingsford and Guillaume Marçais

slides: dandeblasio.com/bcb2019



Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



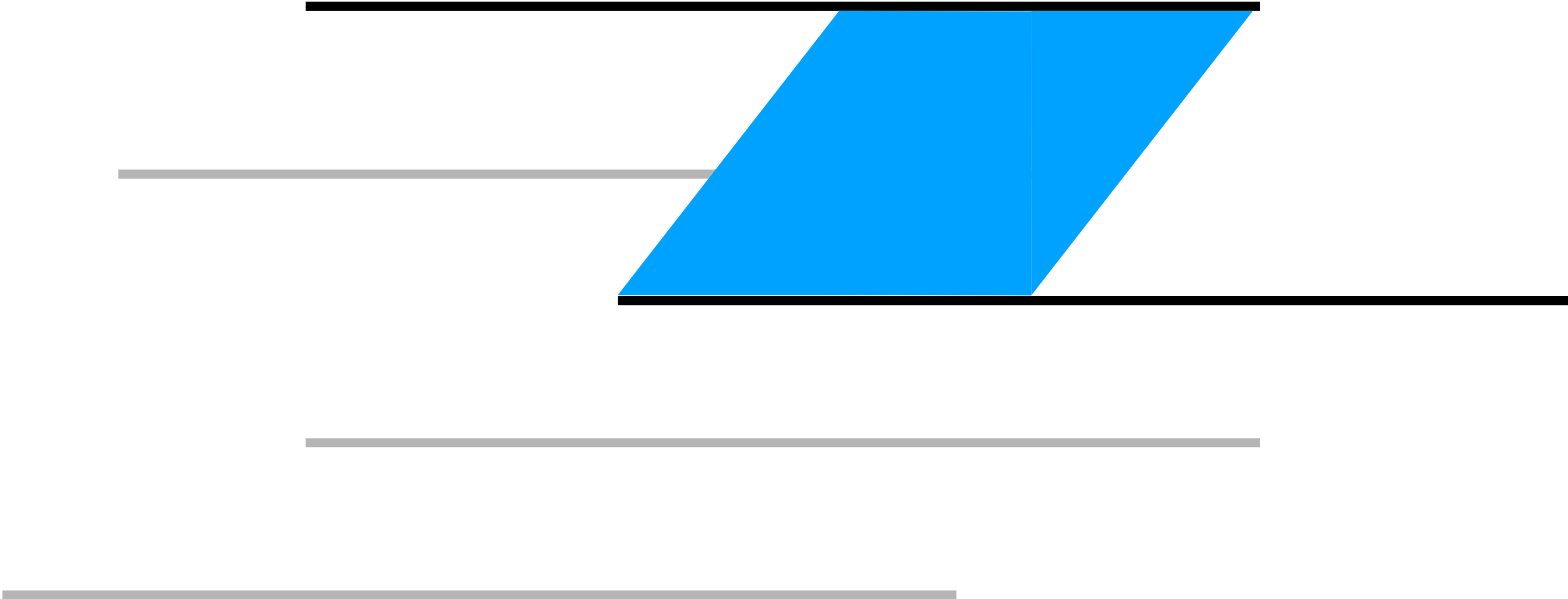
Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



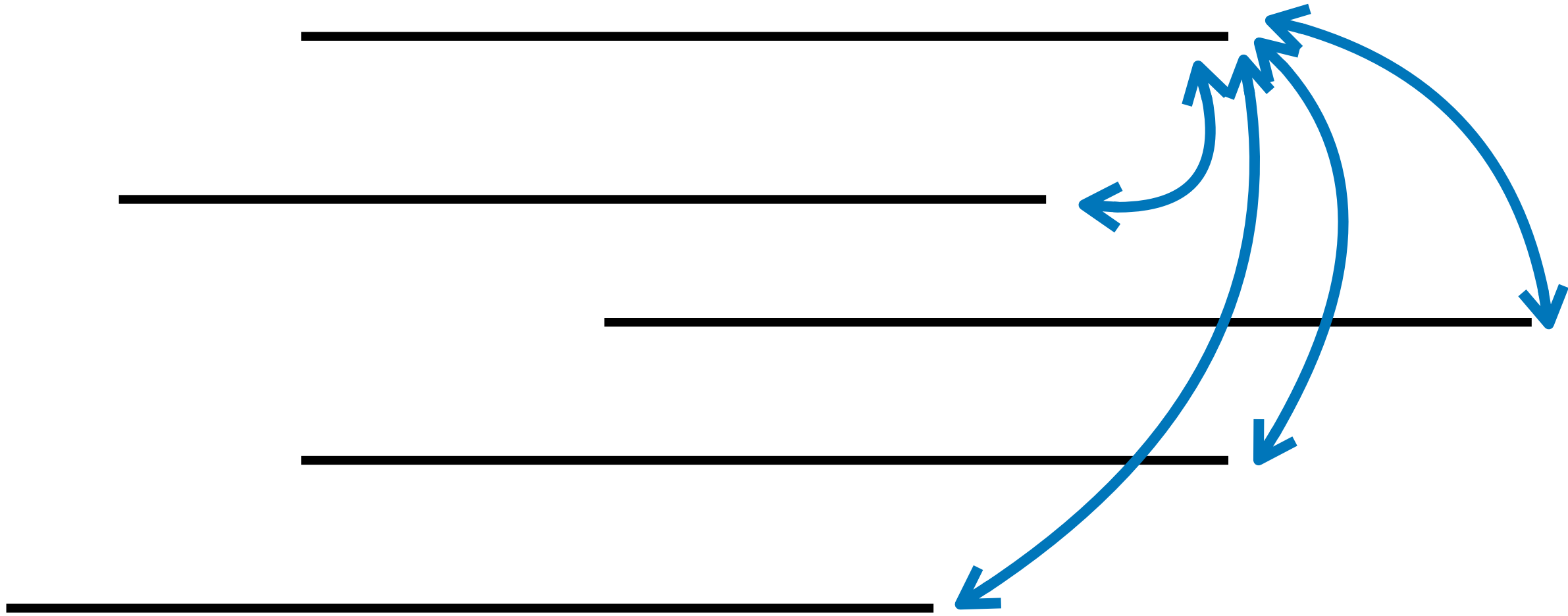
Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



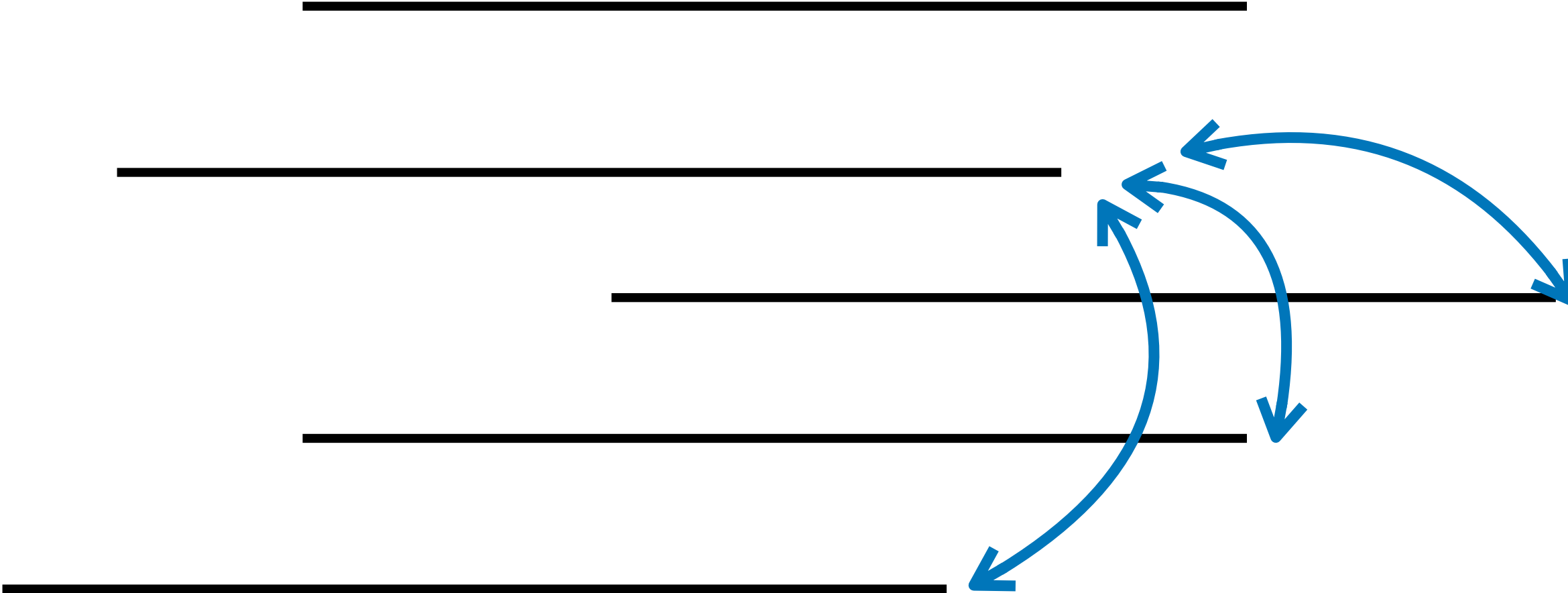
Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



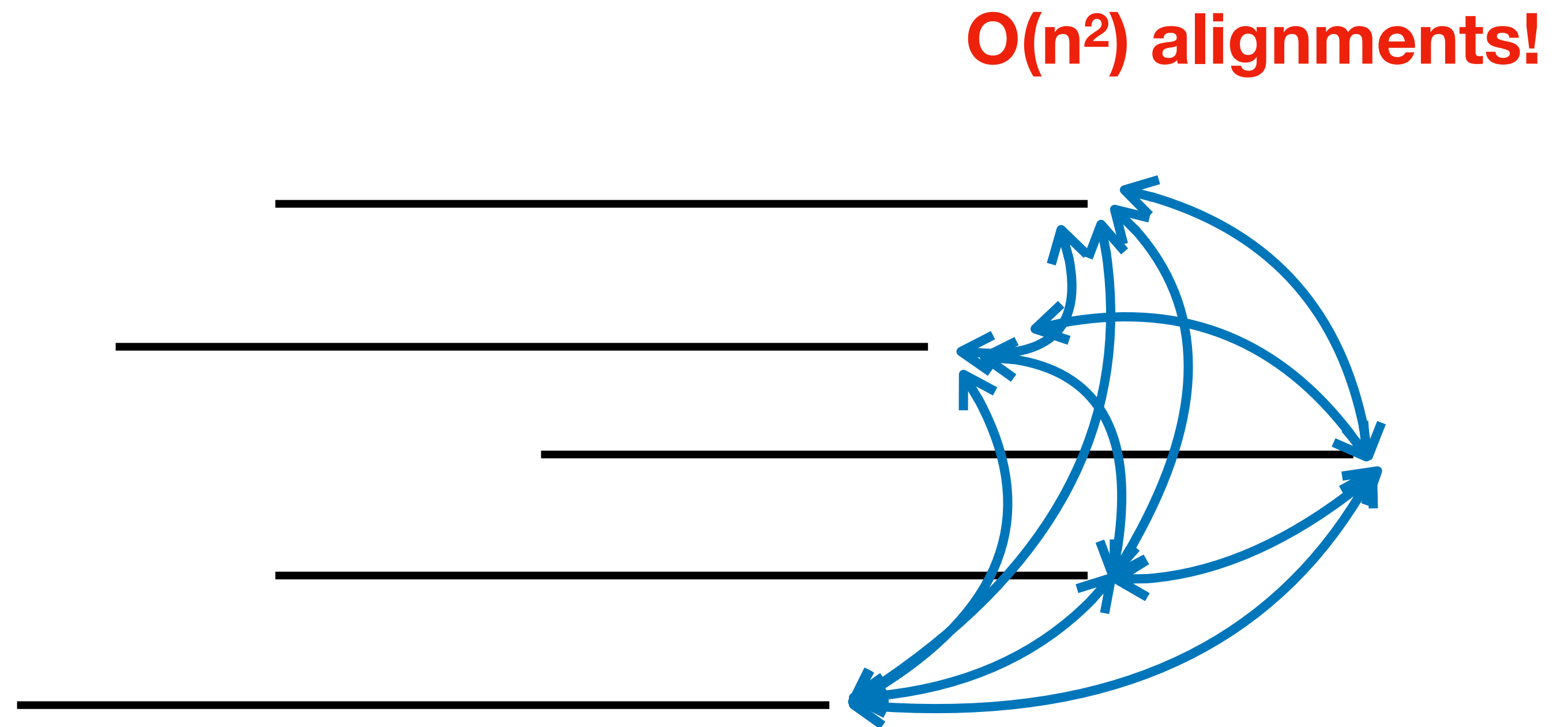
Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



Minimizer Schemes

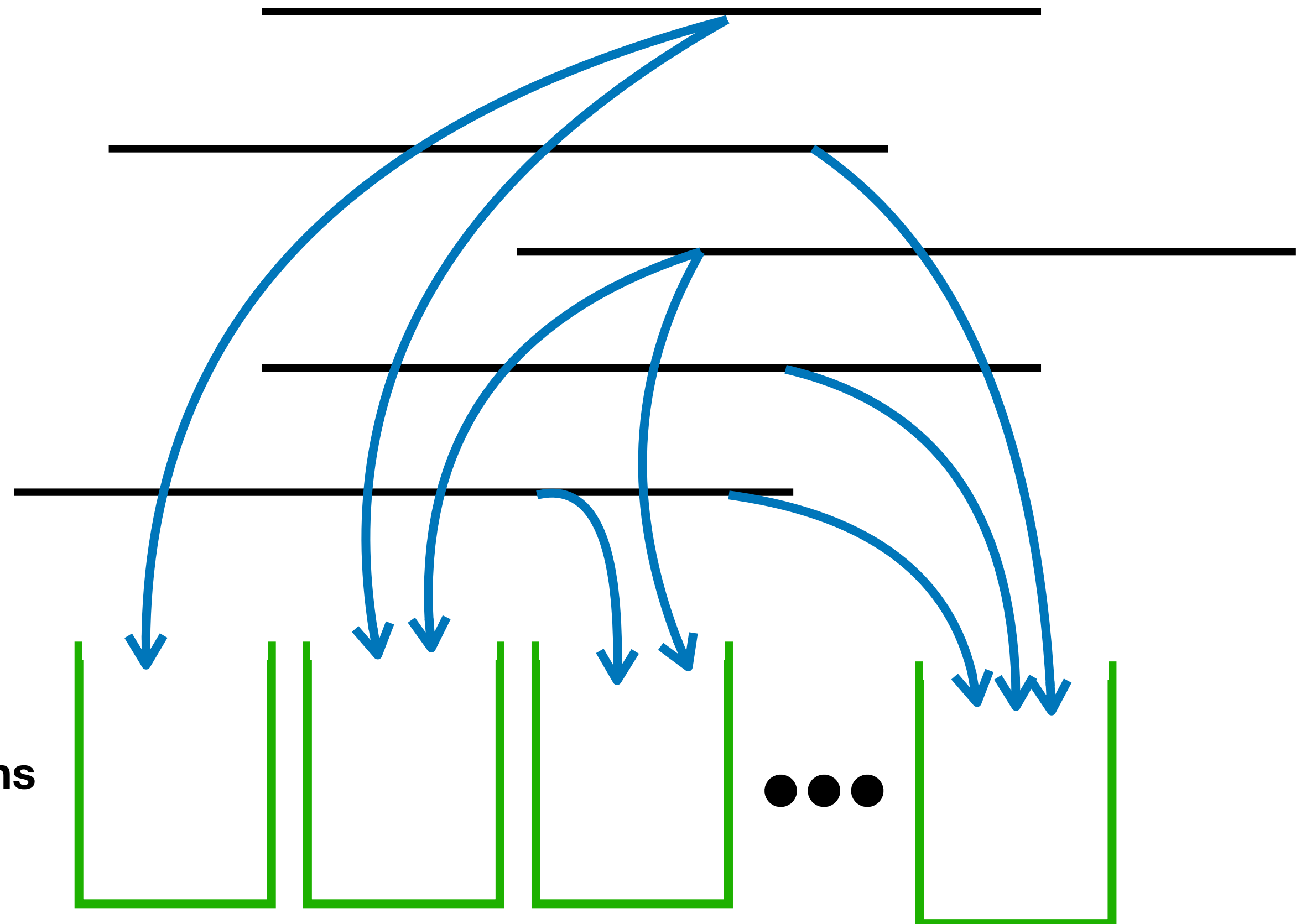
Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation



Minimizer Schemes

Roberts, *et al.* (2004) introduced minimizer schemes as a way to decrease the time needed for sequence overlap computation

Only compare within bins

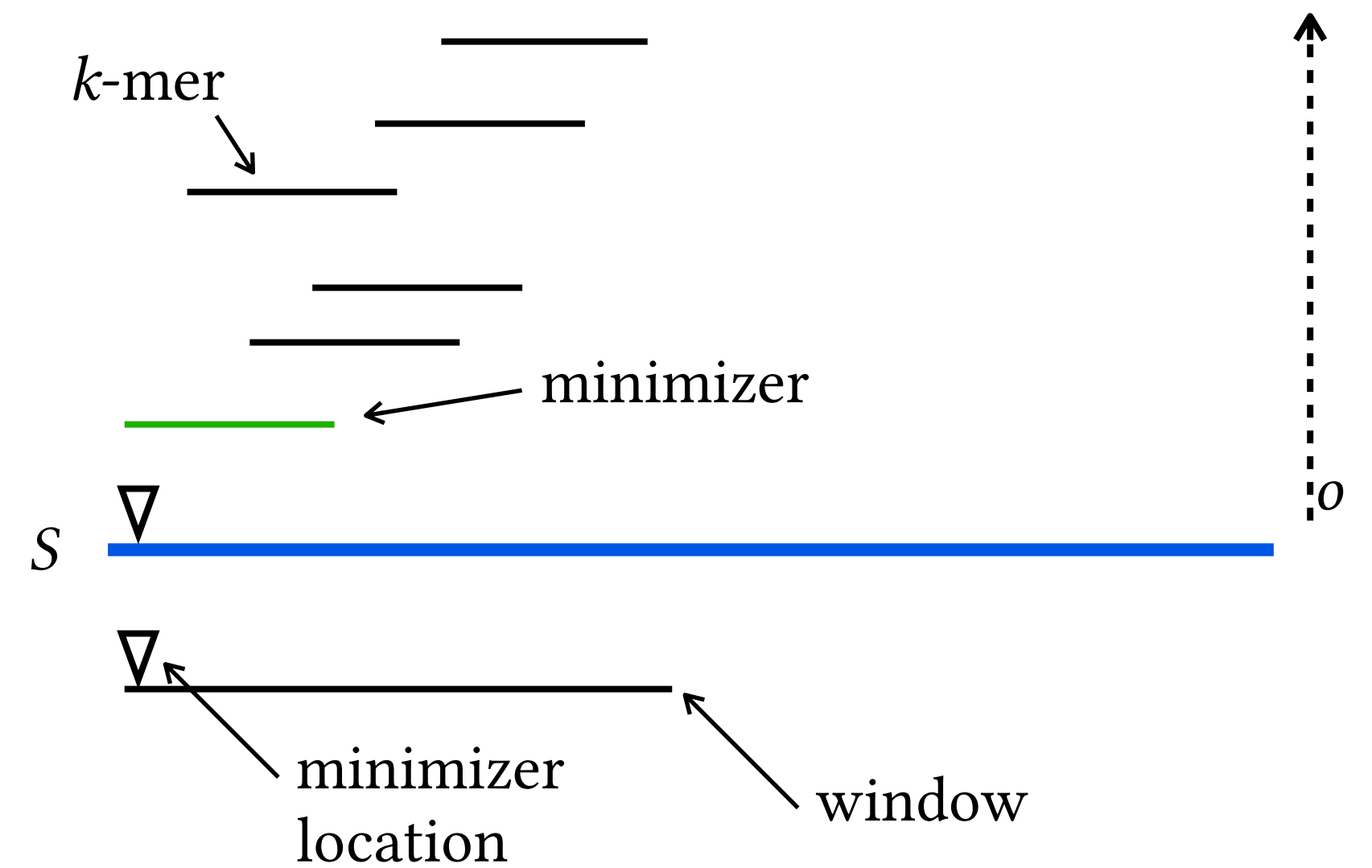


Minimizer Schemes

- Minimizer schemes have two special properties:
 - two sequences with a long exact match must select the same k -mers
 - there are no large gap between selected k -mers
- Use in k -mer counting, *de Bruijn* graph construction, data structure sparsification, etc.

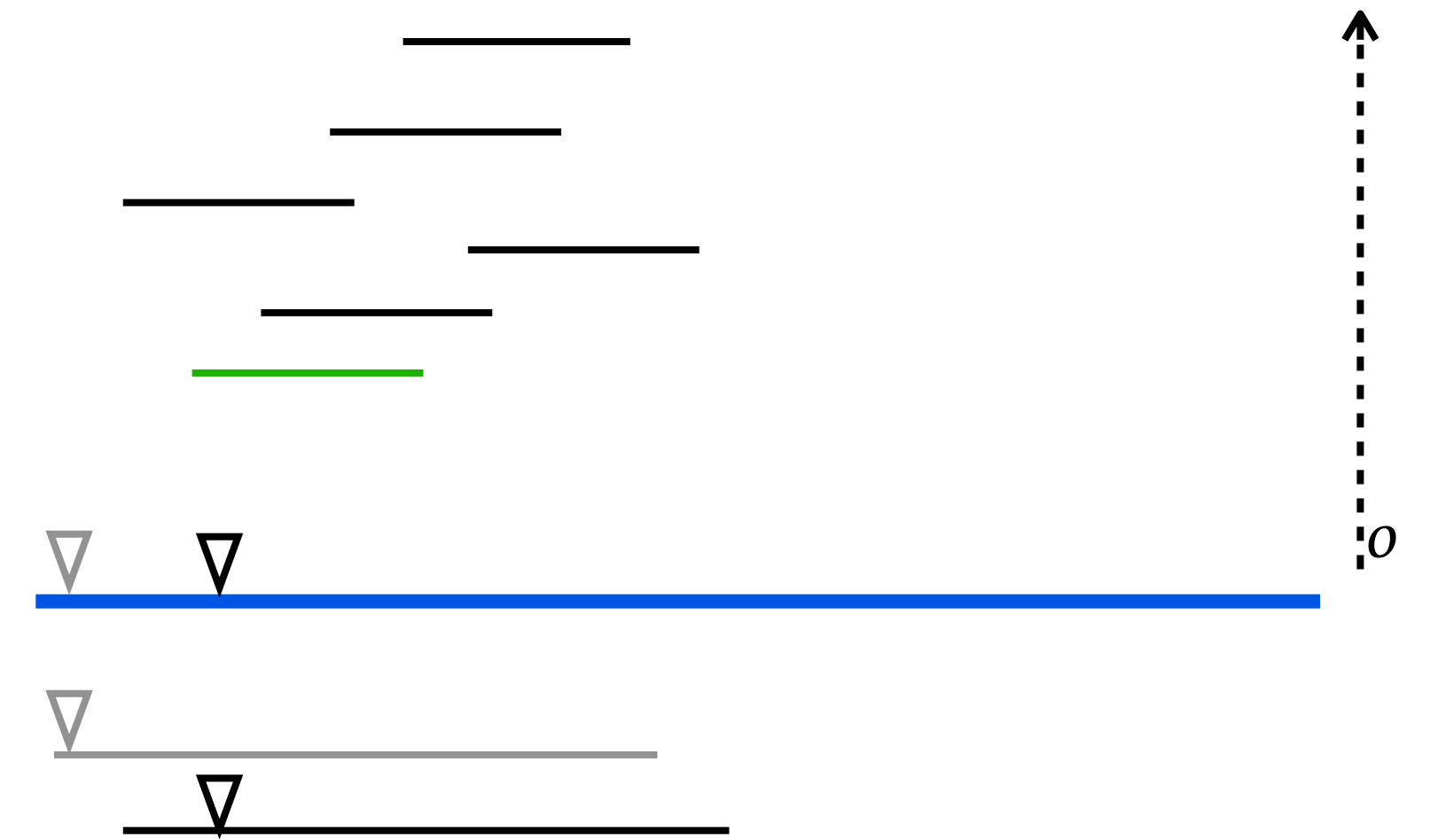
Minimizer Schemes

For a windows of w consecutive k -mers from a sequence S , a minimizer scheme selects the minimum according to an ordering o as a representative



Minimizer Schemes

For a windows of w consecutive k -mers from a sequence S , a minimizer scheme selects the minimum according to an ordering o as a representative



Minimizer Schemes

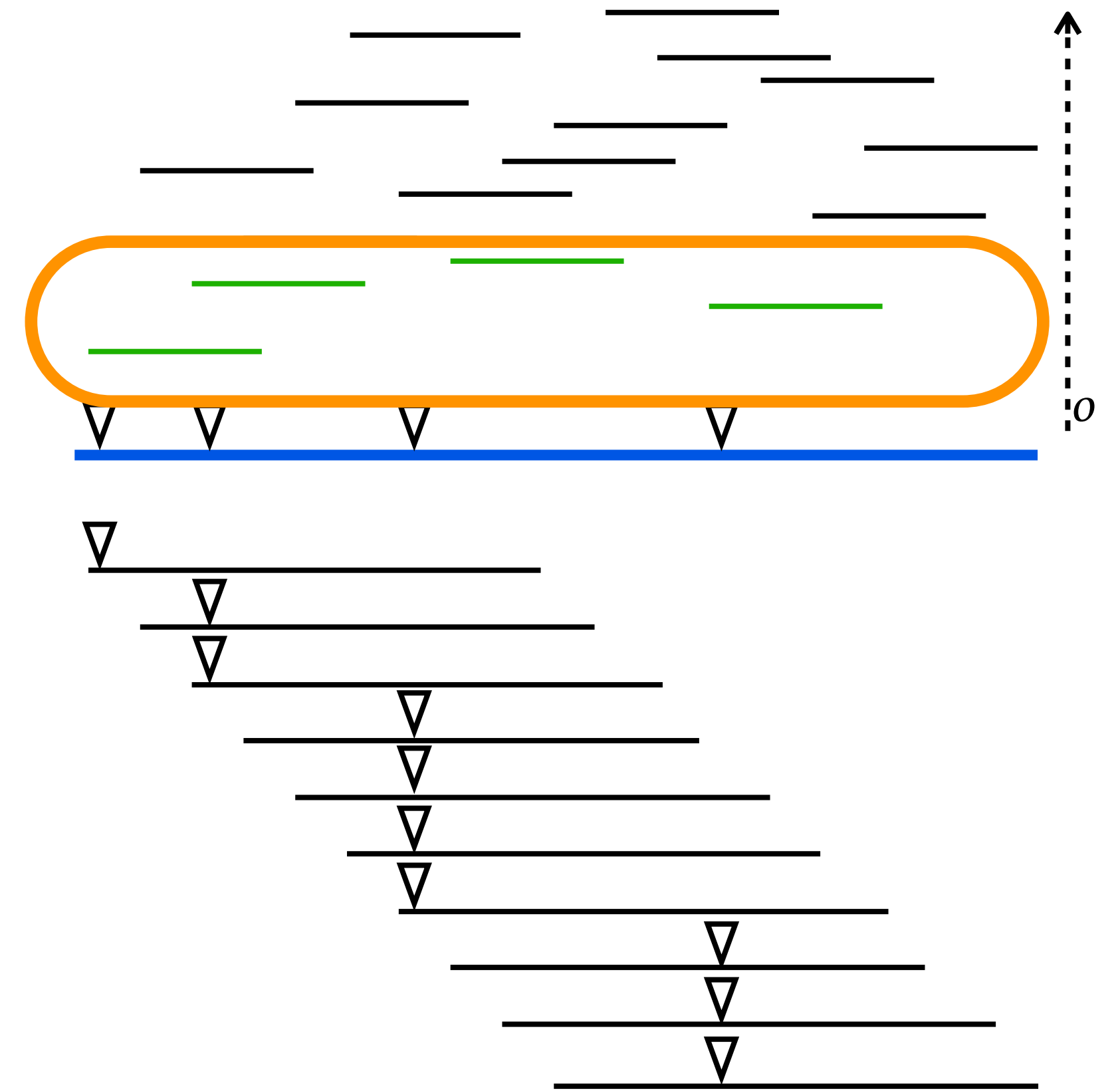
For a windows of w consecutive k -mers from a sequence S , a minimizer scheme selects the minimum according to an ordering o as a representative



Minimizer Schemes

- Changing the ordering used can greatly impact the number of unique minimizers
- Can we find an order that minimizes the number of minimizer locations

Only some k -mers are used as minimizers

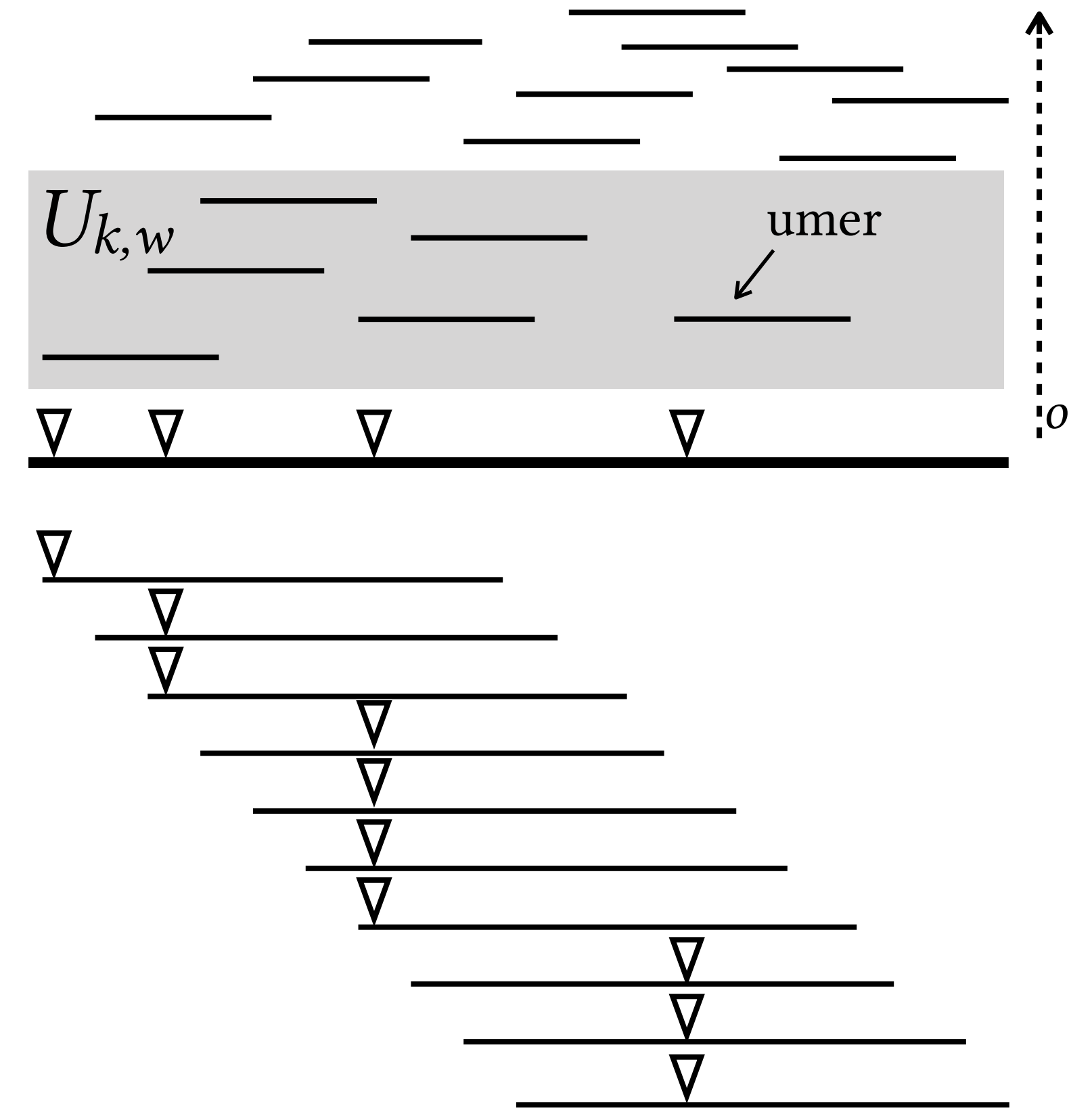


Universal k -mer Set

A universal k -mer set $U_{k,w} \subseteq \Sigma^k$ is a set of k -mers such that any window of w consecutive k -mers must contain at least one element from the set

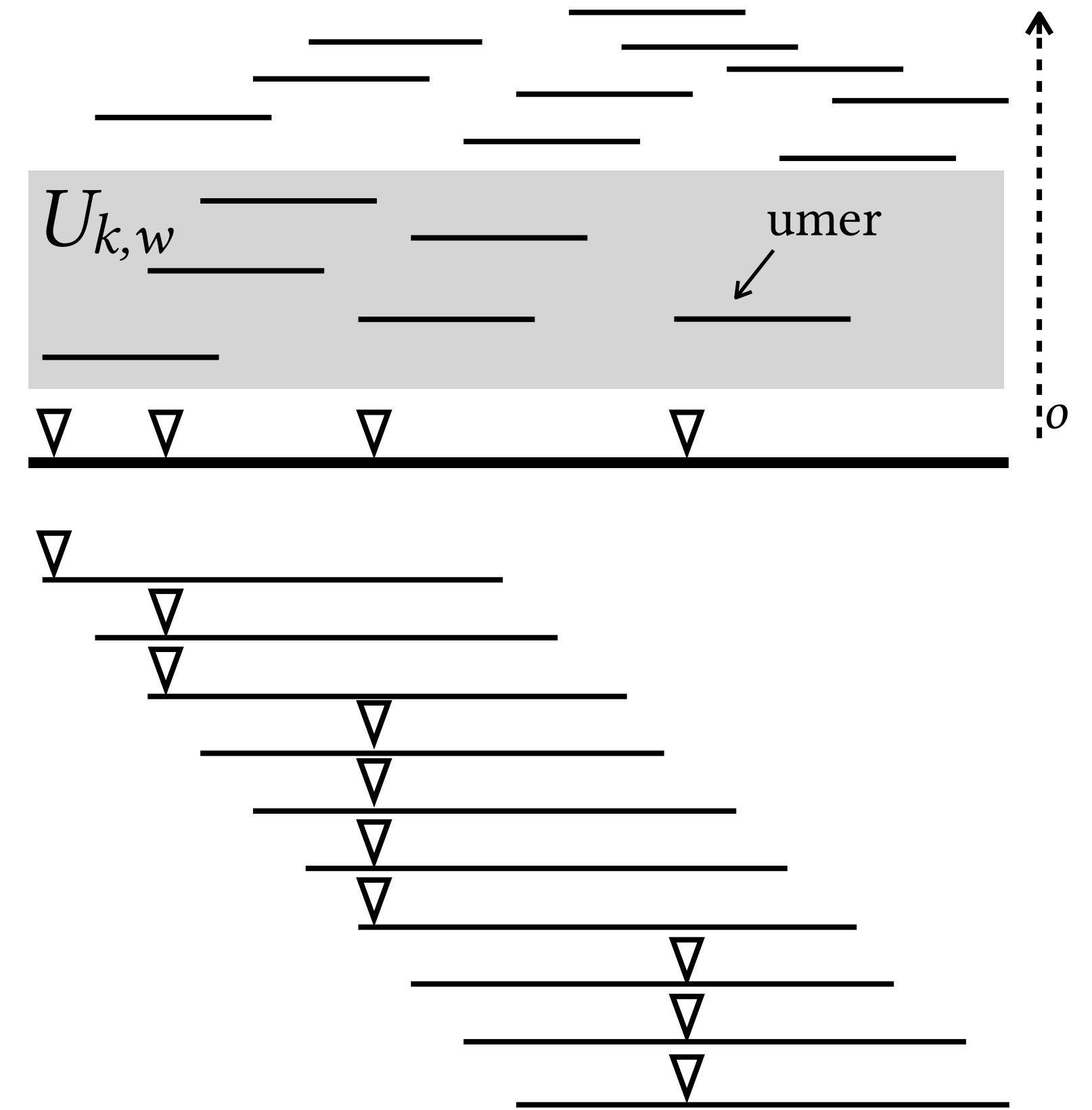
Universal k -mer Set and Minimizer Ordering

- A universal k -mer set induces a family of compatible orderings
- Orderings based on universal sets have better performance than lexicographic or random orders (Marçais, et al., 2017)



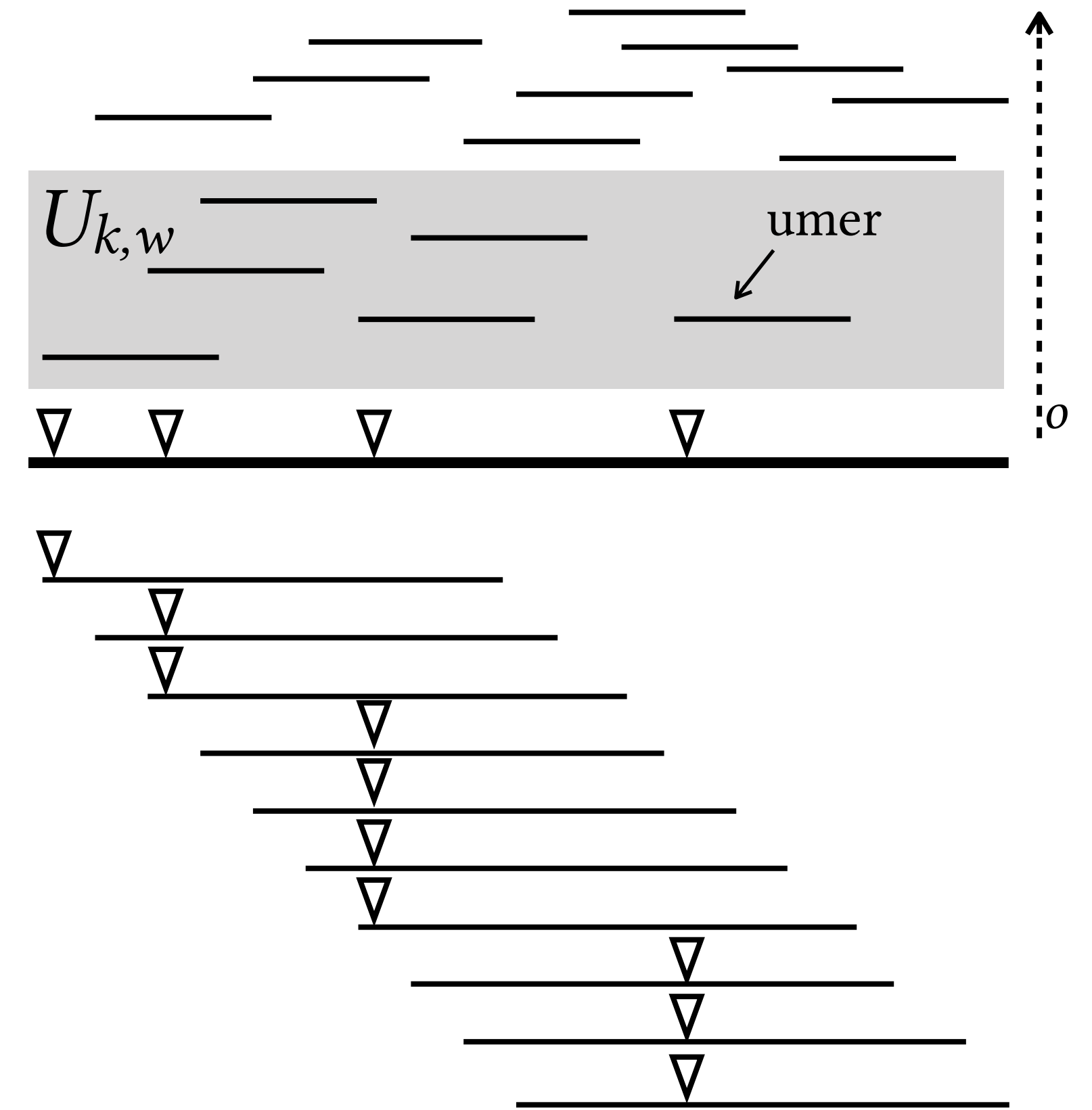
Universal k -mer Set and Minimizer Ordering

- A universal k -mer set induces a family of compatible orderings
- Orderings based on universal sets have better **performance** than lexicographic or random orders (Marçais, et al., 2017)



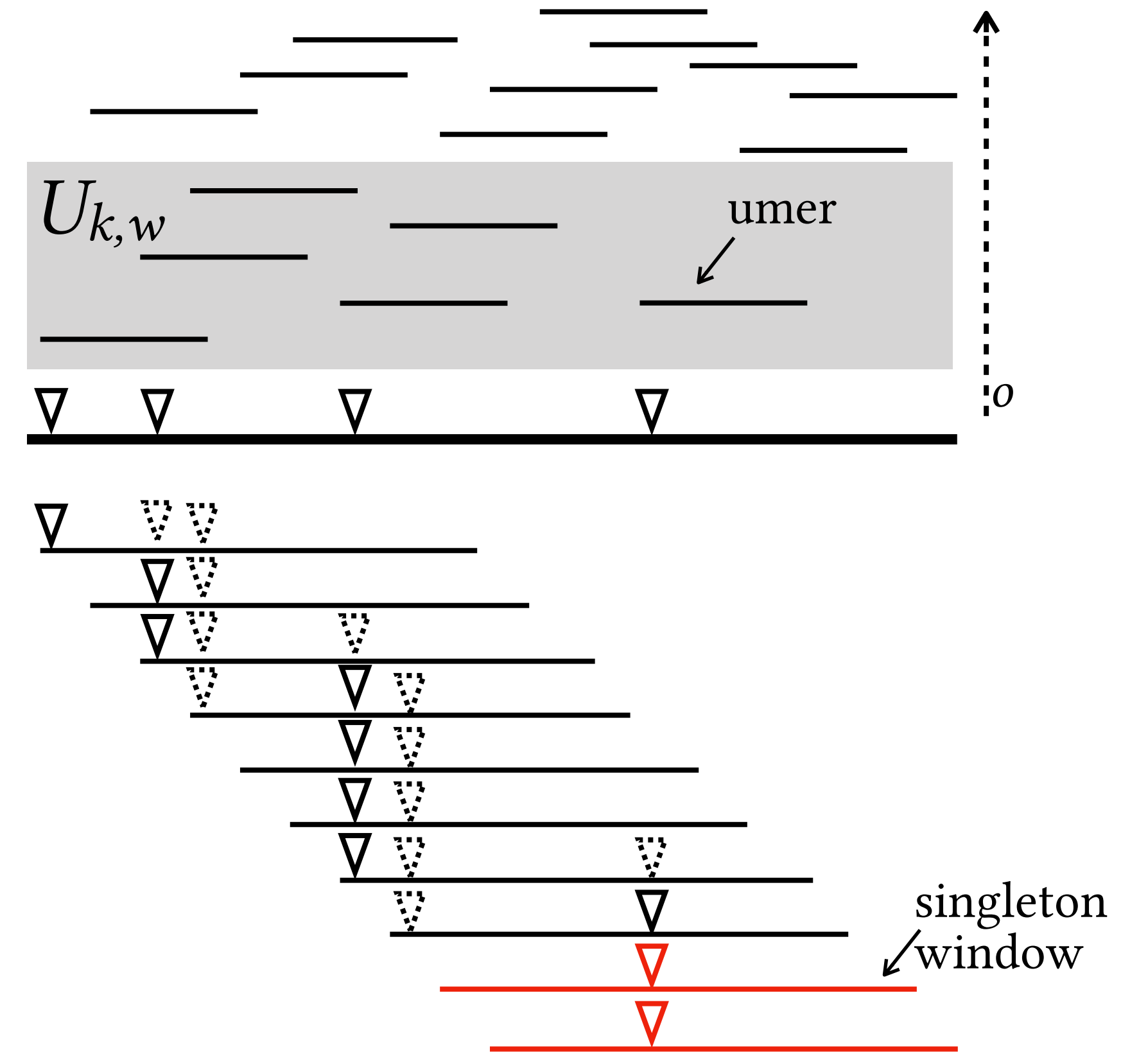
Universal k -mer Set and Minimizer Ordering

- Set Size
 - Fraction of all k -mers in the universal set
- Density
 - Normalized count of minimizer locations in S



Universal k -mer Set and Minimizer Ordering

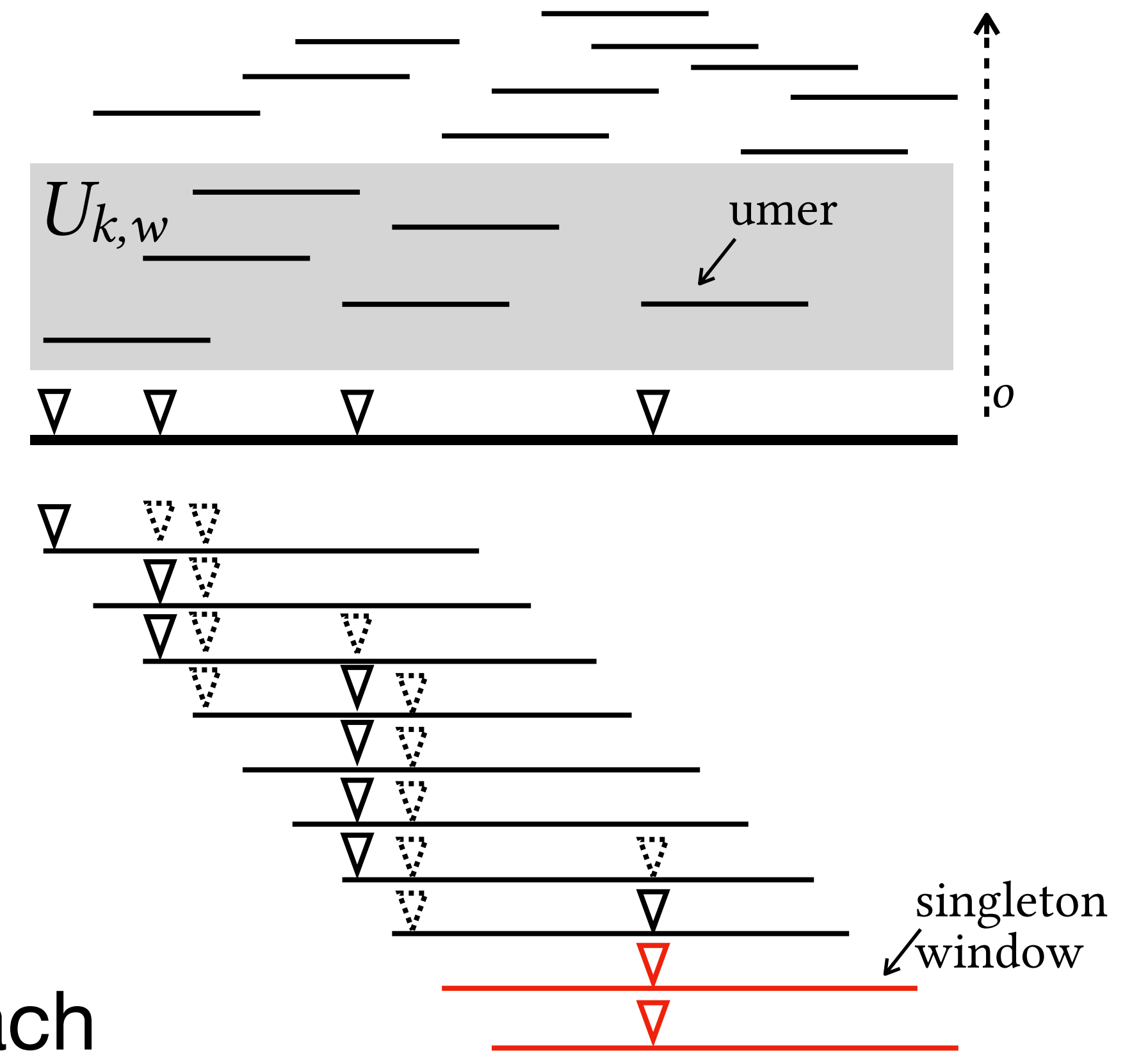
- Set Size
 - Fraction of all k -mers in the universal set
- Density
 - Normalized count of minimizer locations in S
- Sparsity
 - Normalized count of windows in S with only one umer (universal k -mer)



Universal k -mer Set and Minimizer Ordering

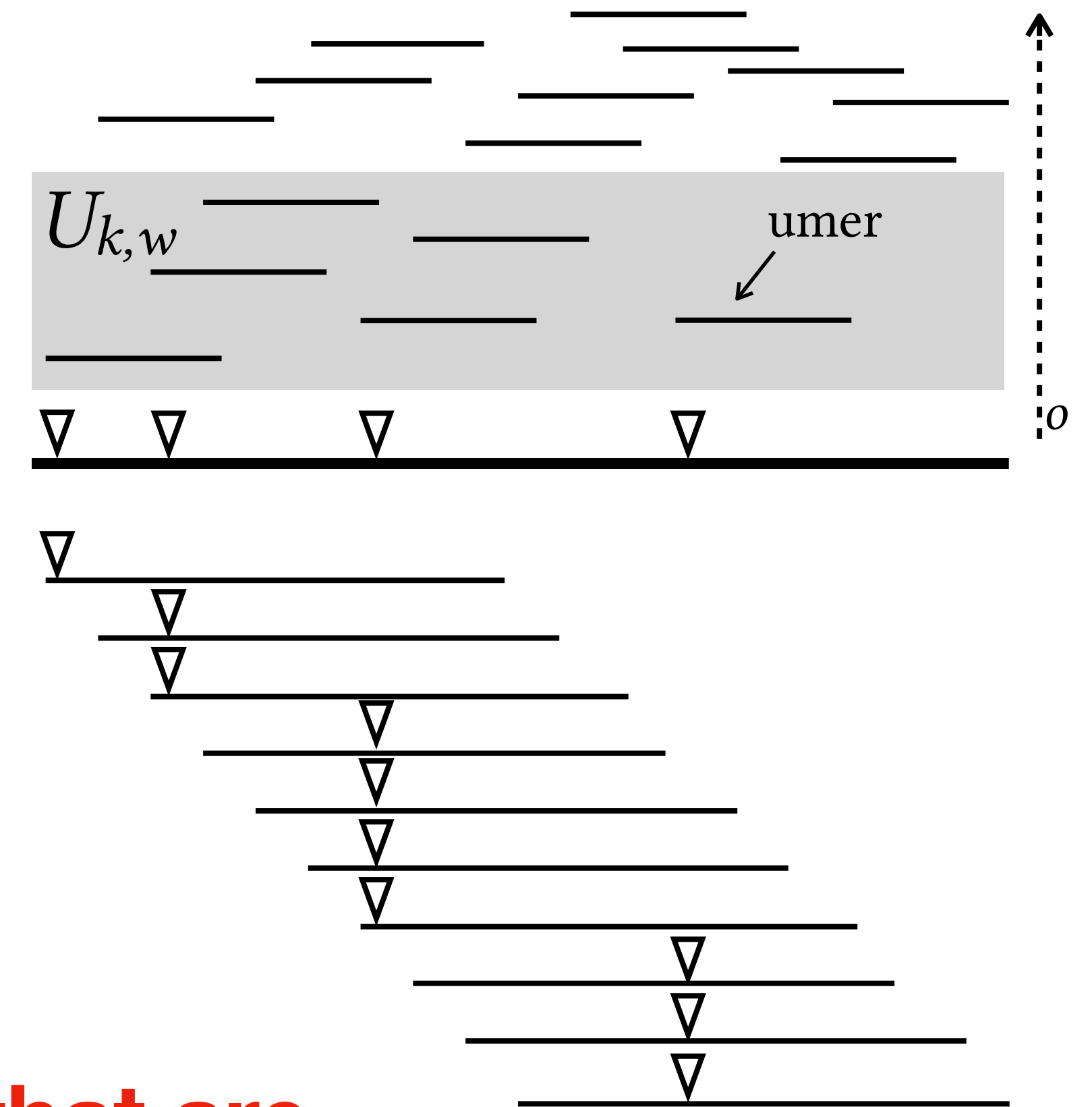
- Set Size
 - Fraction of all k -mers in the universal set
- **Expected** Density
 - Normalized count of minimizer locations in B_L
- **Expected** Sparsity
 - Normalized count of windows in B_L with only one umer (universal k -mer)

B_L is the **de Bruijn** sequence of order L , it contains each window exactly once



Universal k -mer Set and Minimizer Ordering

- A universal k -mer set induces a family of compatible orderings
- Orderings based on universal sets have better performance than lexicographic or random orders (Marçais, et al., 2017)
- Current methods cannot construct sets for values of k and w used in practice



Can we construct universal k -mer sets that are practical for use in minimizer schemes?

Universal k -mer Set Extension

The naïve extension $U_{k,w} \cdot \Sigma$ of a universal set $U_{k,w}$ is universal

create $|\Sigma|$ new $(k+1)$ -mers from each k -mer
by concatenating each character from Σ to the end

Example:

ACCTG $\in U_{k,w} \rightarrow$

{ACCTGA, ACCTGC, ACCTGT, ACCTGG} $\in U_{k,w} \cdot \Sigma$

Universal k -mer Set Extension

The naïve extension $U_{k,w} \cdot \Sigma$ of a universal set $U_{k,w}$ is universal

The sparsity of $U_{k,w} \cdot \Sigma$ is equal to that of $U_{k,w}$

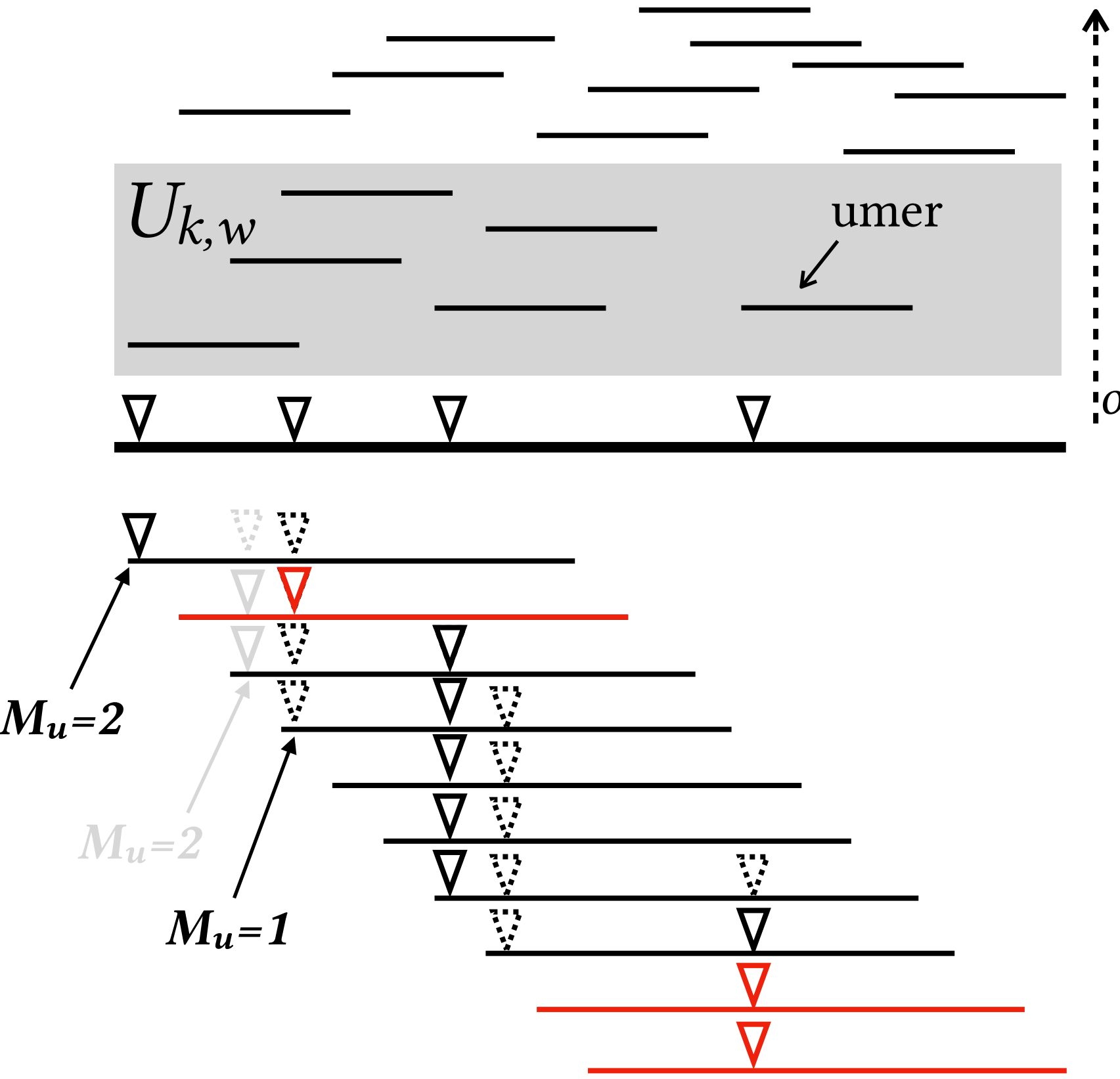
The density of a compatible order for $U_{k,w} \cdot \Sigma$ is less than or equal to the density of a compatible order for $U_{k,w}$ if the orderings are compatible with each other

M_u and $\text{re}M_u\text{val}$

- the minimum co-occurrence count for $u \in U$

$$M_u = \min_{\omega \in W_u} |\omega \cup U|$$

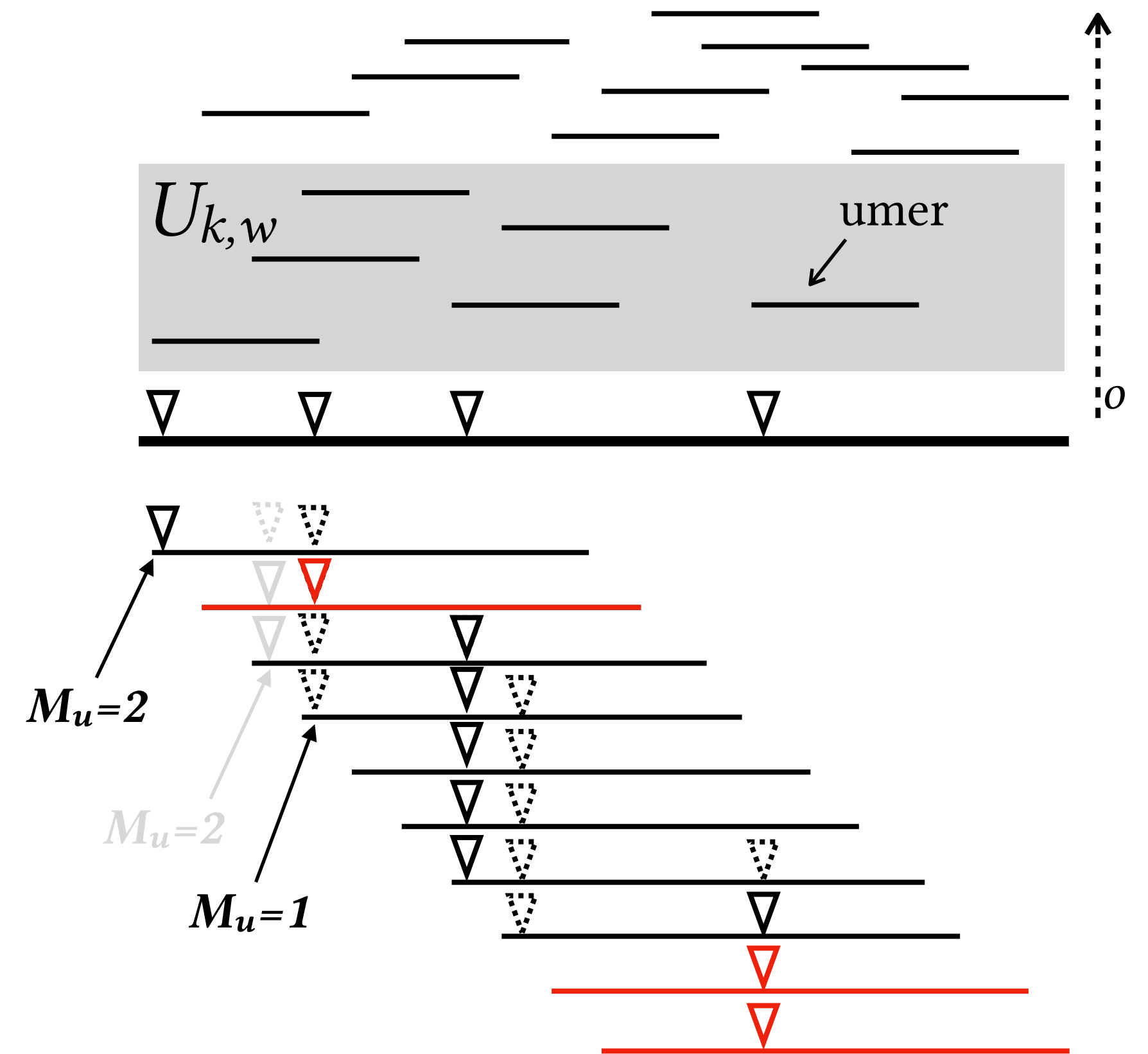
- For any $u \in U$ such that $M_u > 1$, $U \setminus u$ is universal



M_u and re M_u val

- the minimum co-occurrence count for $u \in U$

$$M_u = \min_{\omega \in W_u} |\omega \cup U|$$
- For any $u \in U$ such that $M_u > 1$, $U \setminus u$ is universal
- The universal set after the removal of u has:
 - smaller size, and
 - higher (possibly equal) sparsity



Optimal re*M_u*val

- Not all umers with $M_u > 1$ can be removed from U ,
- Integer linear programming (ILP) is used to find the minimum number of k -mers to retain
- The ILP is deceptively simple

$$\text{minimize } \sum_{u \in U} y_u$$

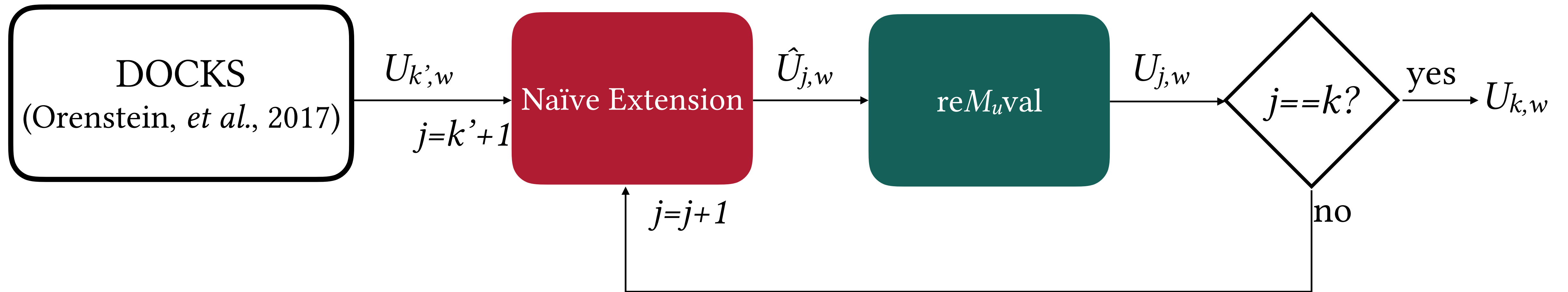
$$\text{subject to } \sum_{u \in \omega \cap U} y_u \geq 1 \quad \forall \omega \in W$$

$$y_u \in \{0,1\} \quad \forall u \in U$$

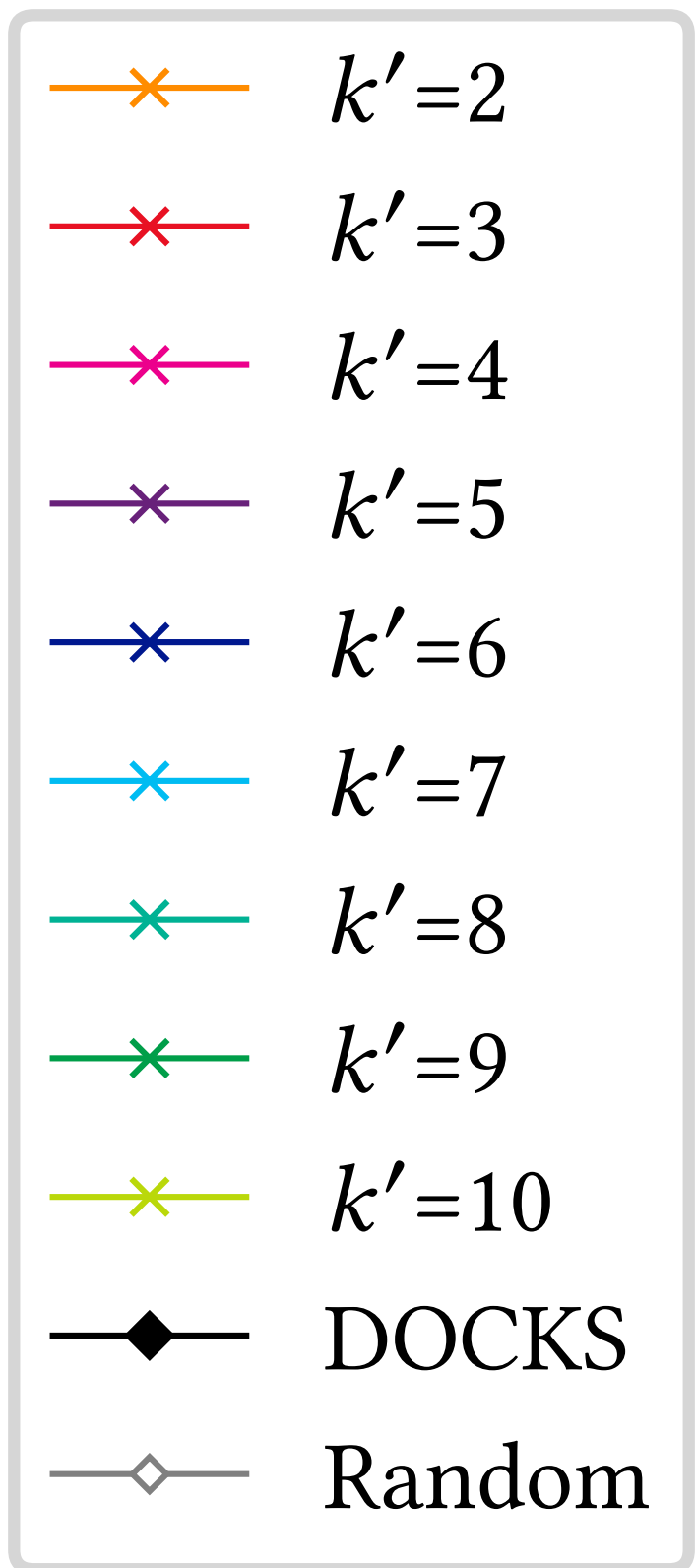
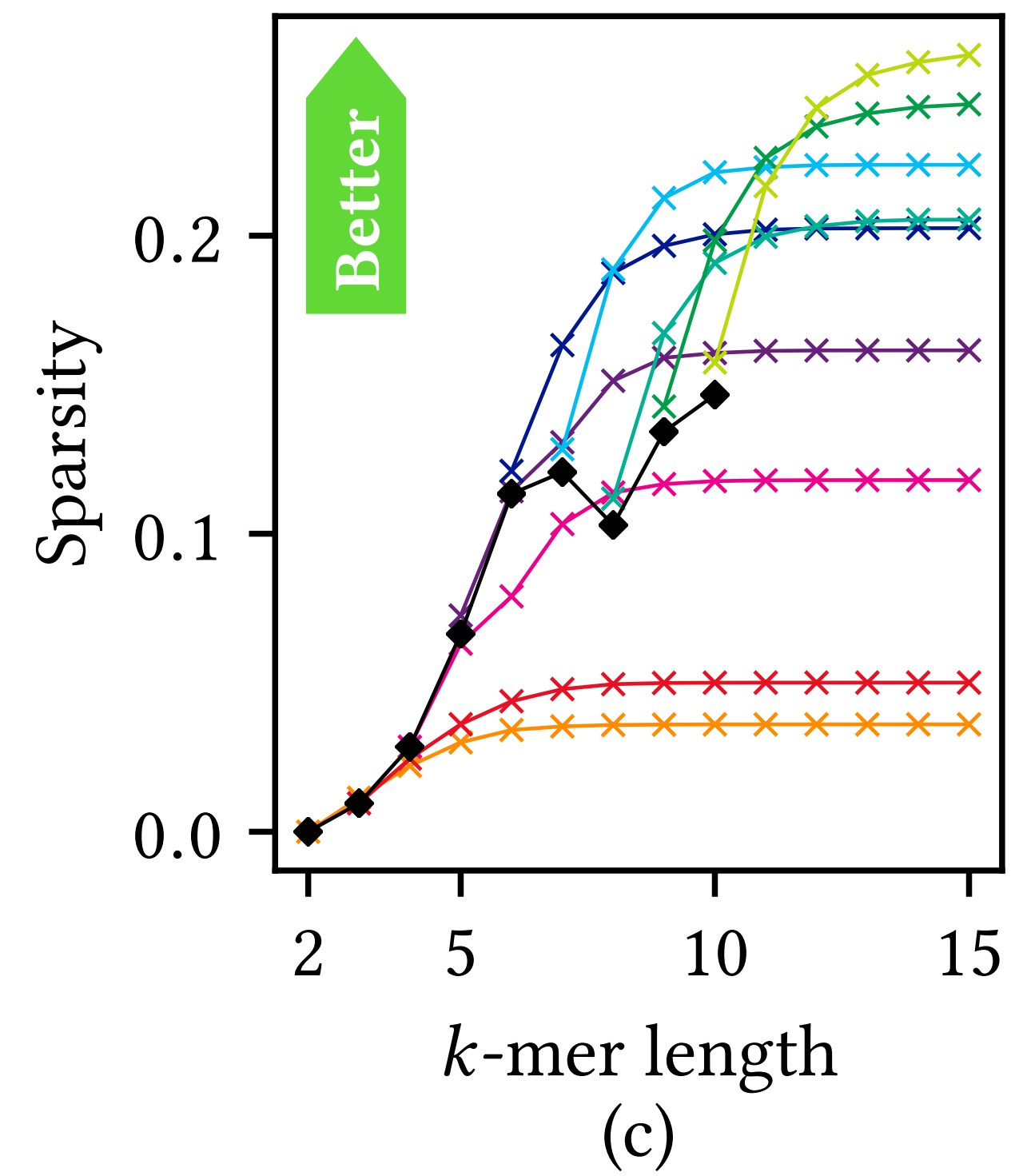
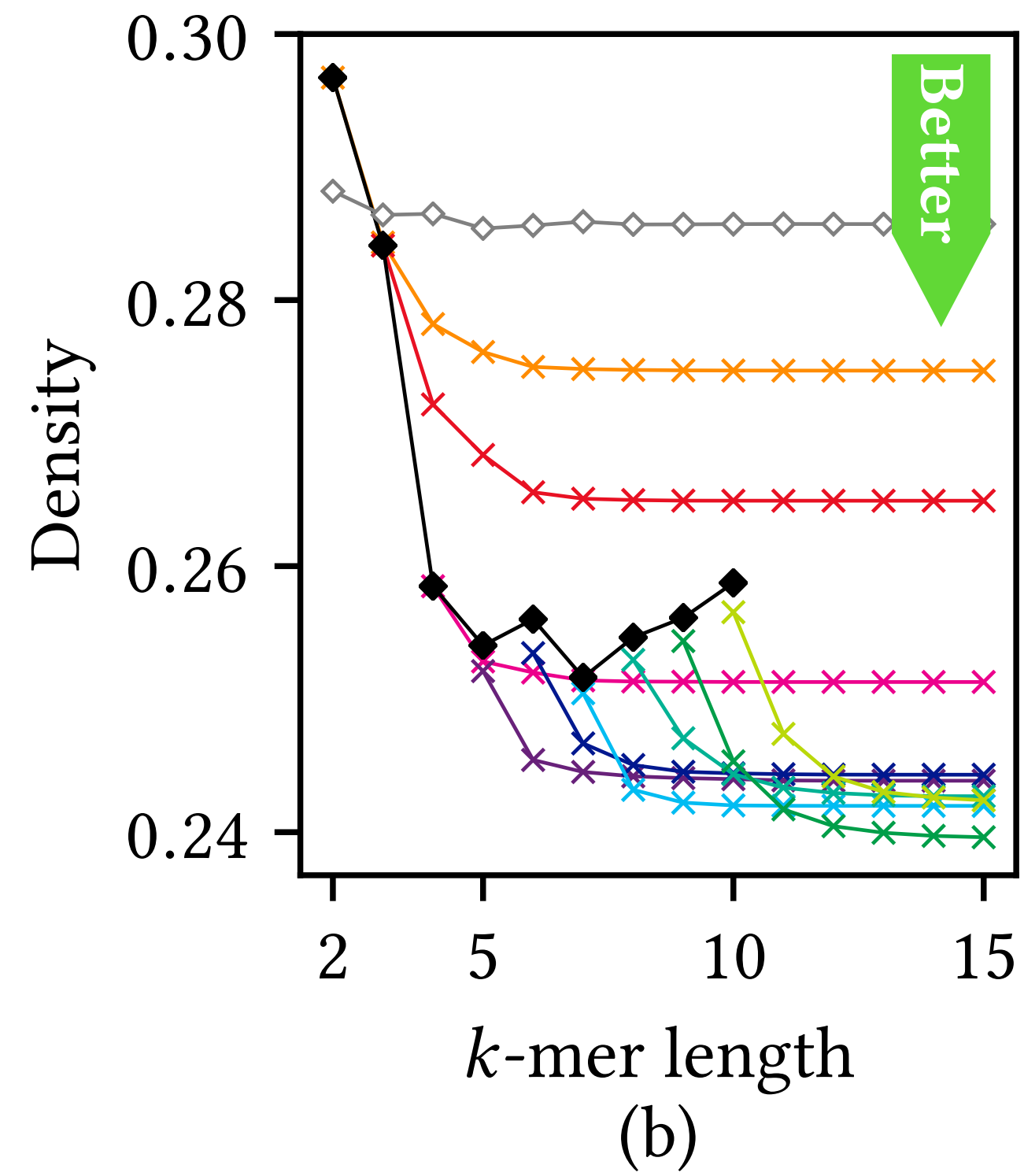
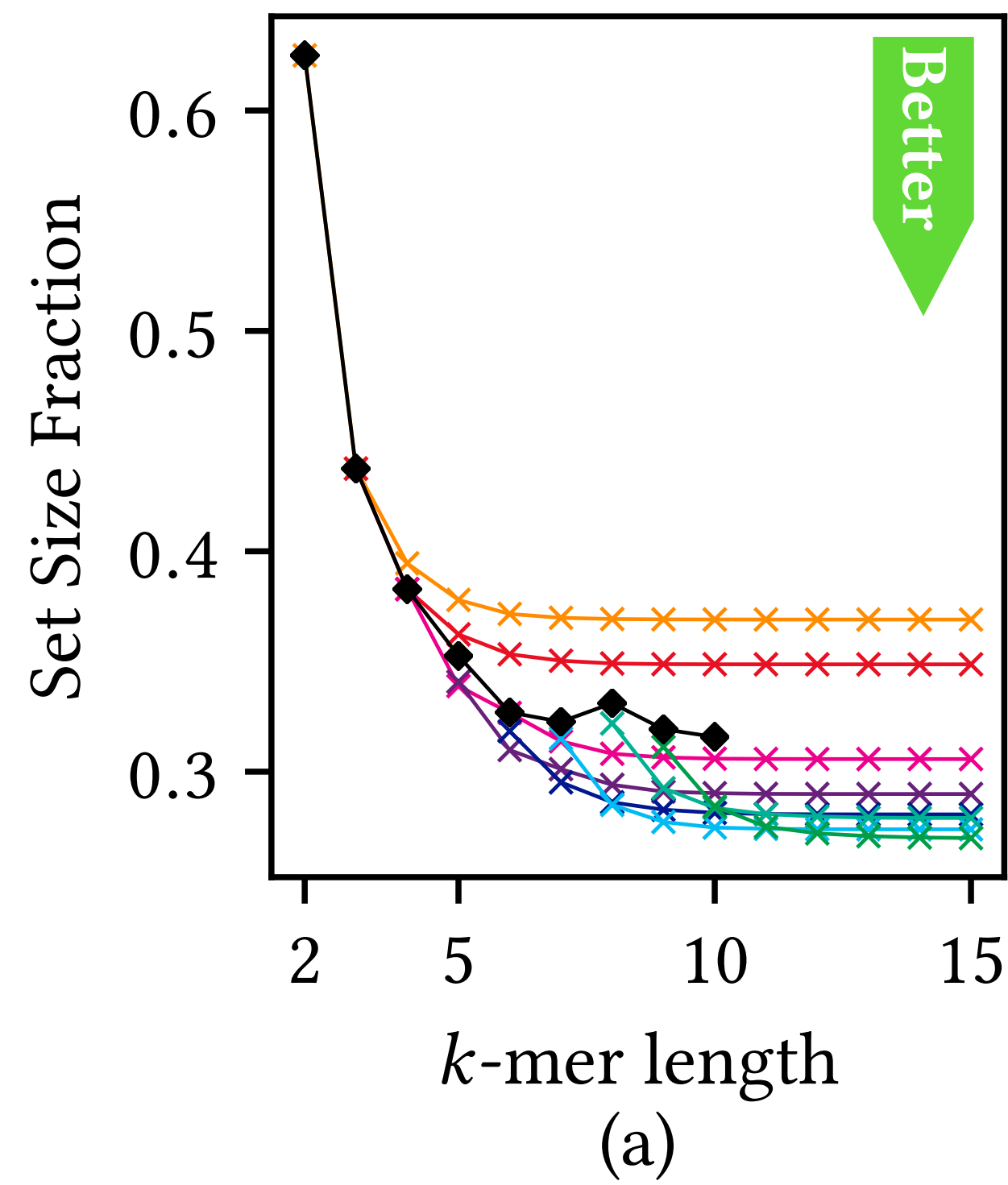
All of the umer co-occurrence information is encoded in W



Practical Universal k -mer Set Construction



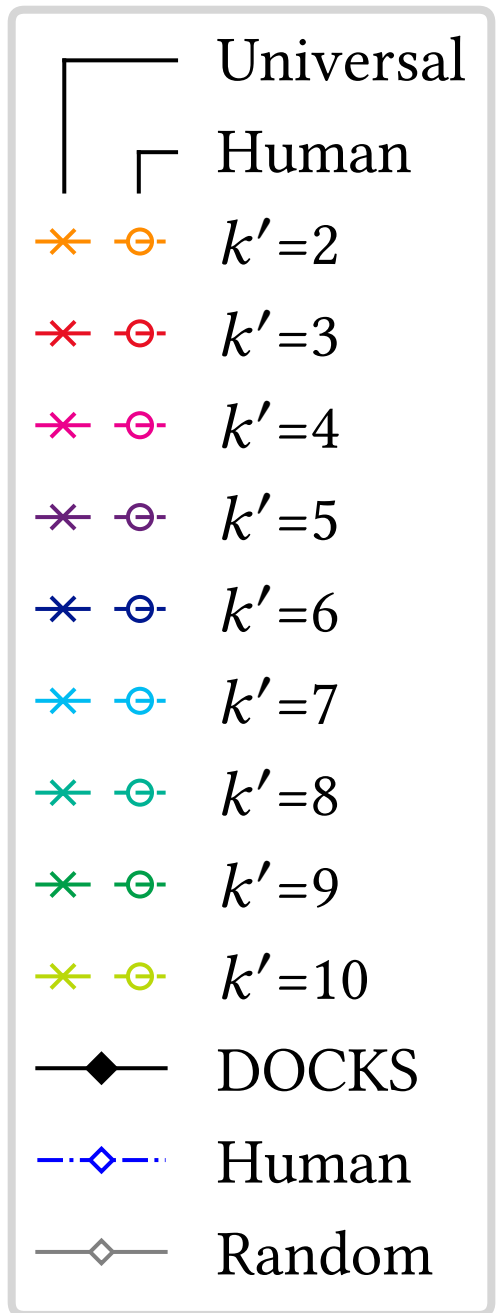
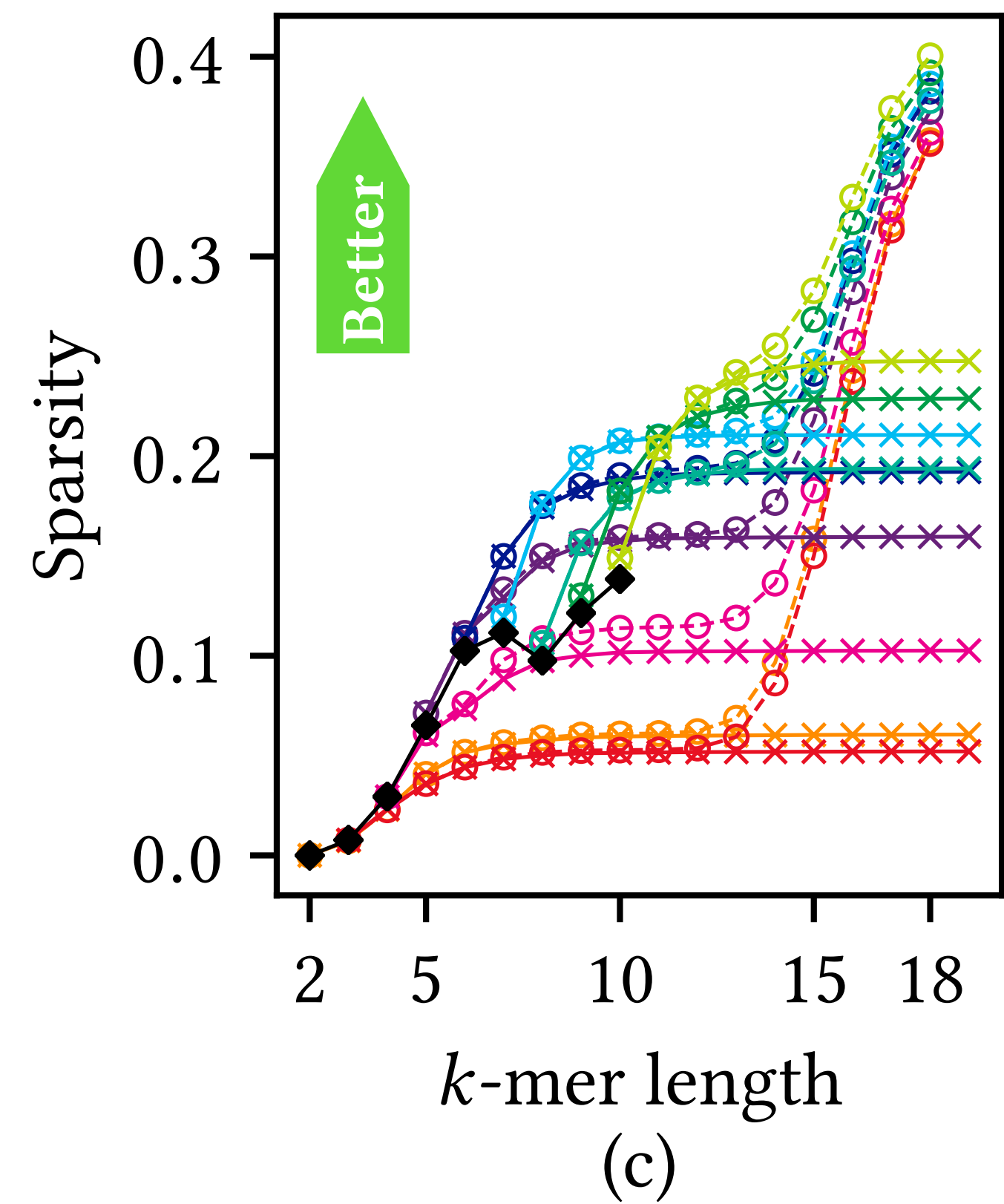
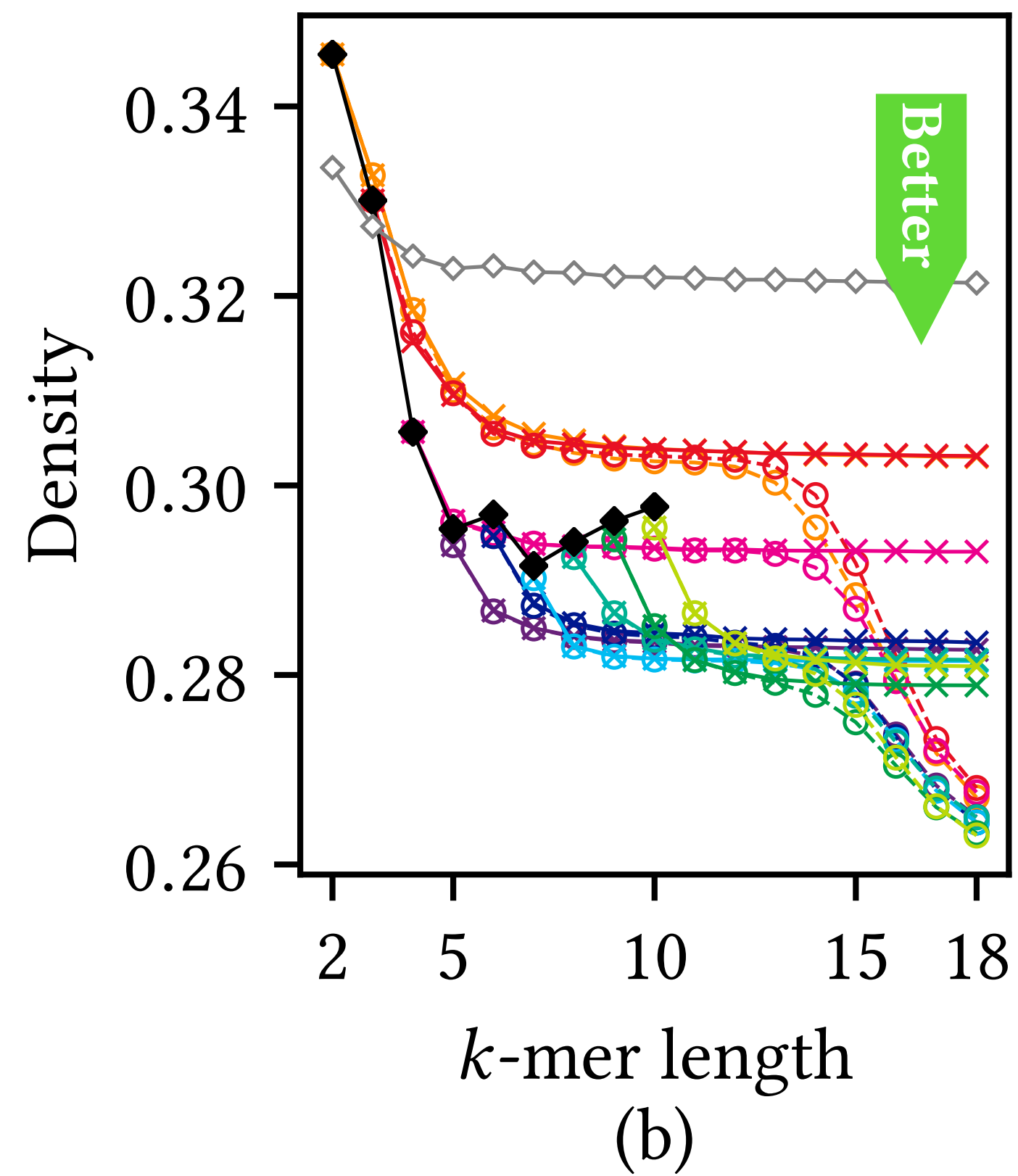
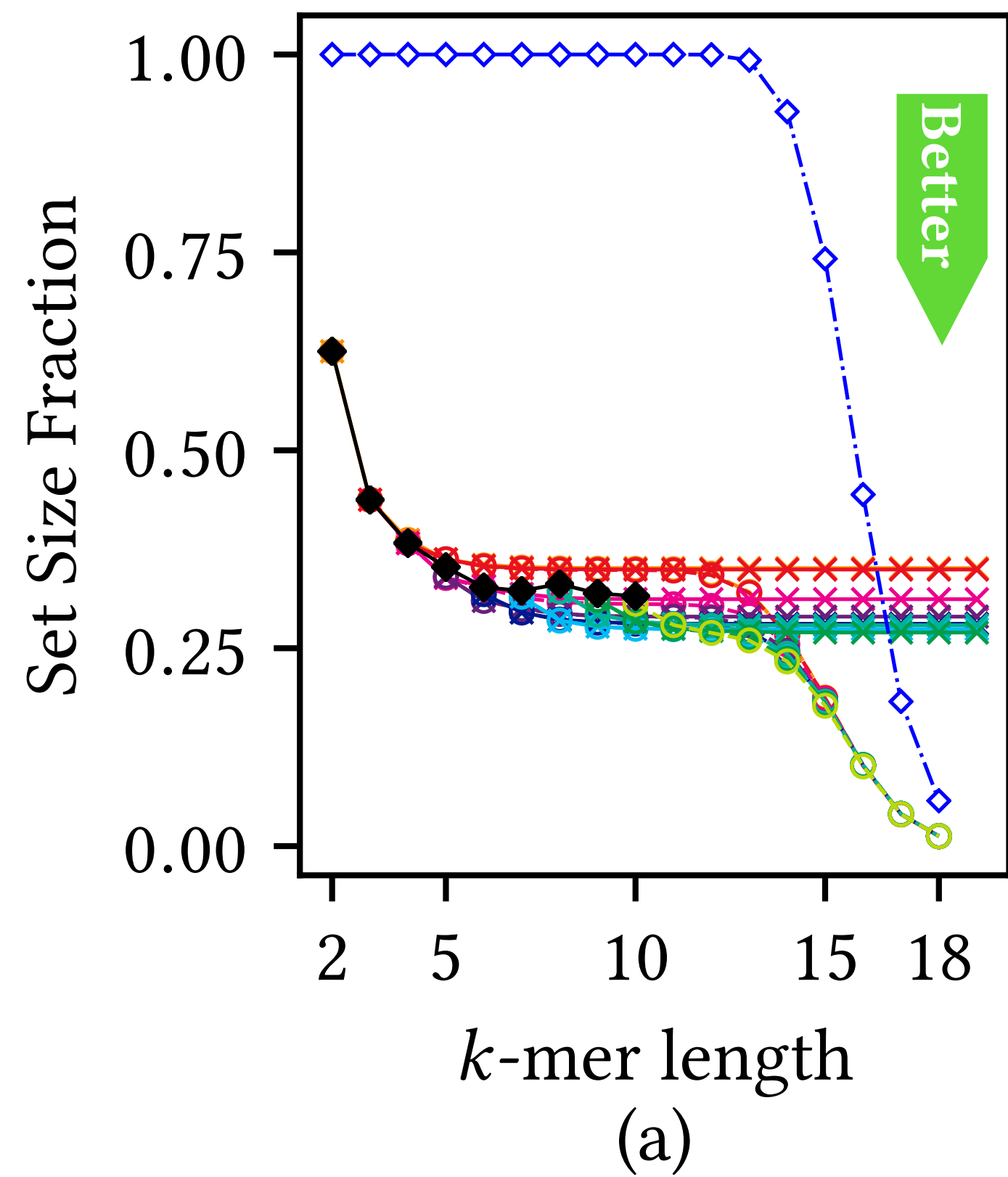
Improvements Over DOCKS Sets



Sequence-specific re M_u val

- Often minimizer schemes are used on a single reference sequence
- Not all windows will appear in that sequence
- The set of windows for re M_u val W can be limited to those from the reference

Improvement Over General k -mer Sets



Summary

- Universal k -mer sets can be constructed for use as in for minimizer scheme orderings for large k
- Reference-specific universal sets have better performance
- Open question if similar methods can be used to extend in w

Practical universal k -mer sets for minimizer schemes

Dan DeBlasio, Fiyin Gbosibo, Carl Kingsford and Guillaume Marçais

dandeblasio.com



slides: dandeblasio.com/bcb2019
code: github.com/Kingsford-Group/removal

Thanks

Kingsford group members

iBRIC administration, especially David Boone

Funding

Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative (GBMF4554)

US National Institutes of Health (R01GM122935)

US National Science Foundation (CCF- 1256087, CCF-1319998)

