Practical universal *k*-mer sets for minimizer schemes

Dan DeBlasio[1], Fiyinfoluwa Gbosibo[2,3], Carl Kingsford[1], Guillaume Marçais[1]

[1]Computational Biology Department, School of Computer Science, Carnegie Mellon University
[2]Departments of Computer Science and Math, Lincoln University
[3]Internship in Biomedical Research, Informatics, and Computer Science (iBRIC) at University of Pittsburgh

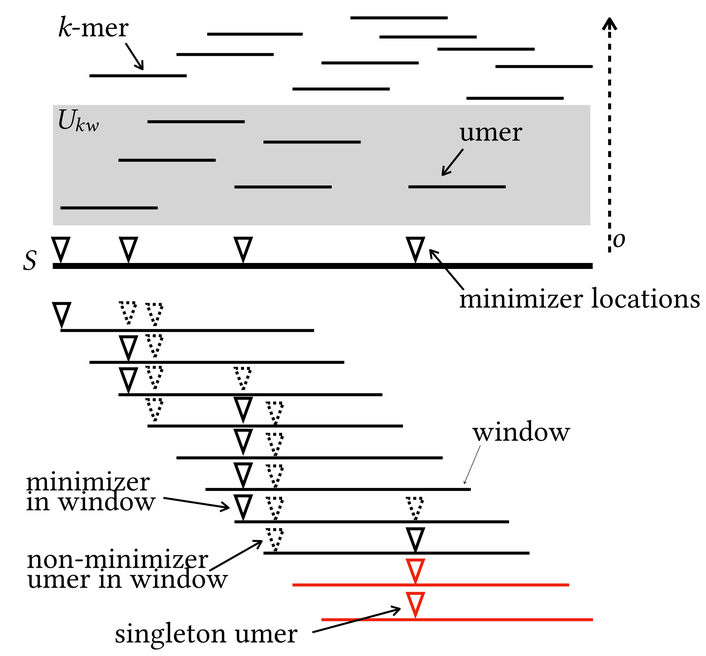
Universal *k*-mer Sets

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 sets can be used to find ordering for minimizer schemes that have better performance in general that lexicographic or random orders (Marçais, *et al.*, 2017).

Current methods for finding universal sets do not currently produce sets for the values of *k* and *w* used in practice.

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



re$M_u$val

Sets generated by naïve extension are not guaranteed to be optimal and some umers can be removed and still preserve the optimality.

Define a umer's minimum co-occurance number as:

$$M_u = \min_{\omega \in W_u} |\omega \cap U| \quad \forall u \in U$$

umers with co-occurace values above *1* always occur in windows with at least one other umer and are candidates for **re$M_u$val** (short for M_u guided removal).

THEOREM. Given universal *k*-mer set *U* and a umer $u \in U$ such that $M_u \geq 1$, the set $U \setminus u$ is also a universal set.

Optimal re$M_u$val

Not all umers with $M_u > 1$ can be removed from *U*, an integer linear program (ILP) can be constructed to find the subset of *k*-mers that can be removed to minimize the universal set.

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

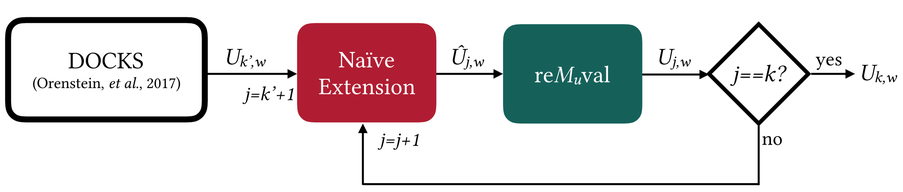
$$\text{subject to} \sum_{u \in \omega \cap U} y_u \geq 1 \quad \forall \omega \in W$$

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

This ILP is deceptively simple since the co-occurance behavior is encoded in the set of windows *W*.

Set Construction Procedure

Starting with a universal set $U_{k',w}$, iterate between naïve extension and re$M_u$val to find sets with the desired parameters



Sequence-specific re$M_u$val

In many cases, minimizer schemes are used on a single reference sequence.

Not all windows will appear in that sequence.

When constructing the ILP, *W* will be the subset of windows in the sequence.

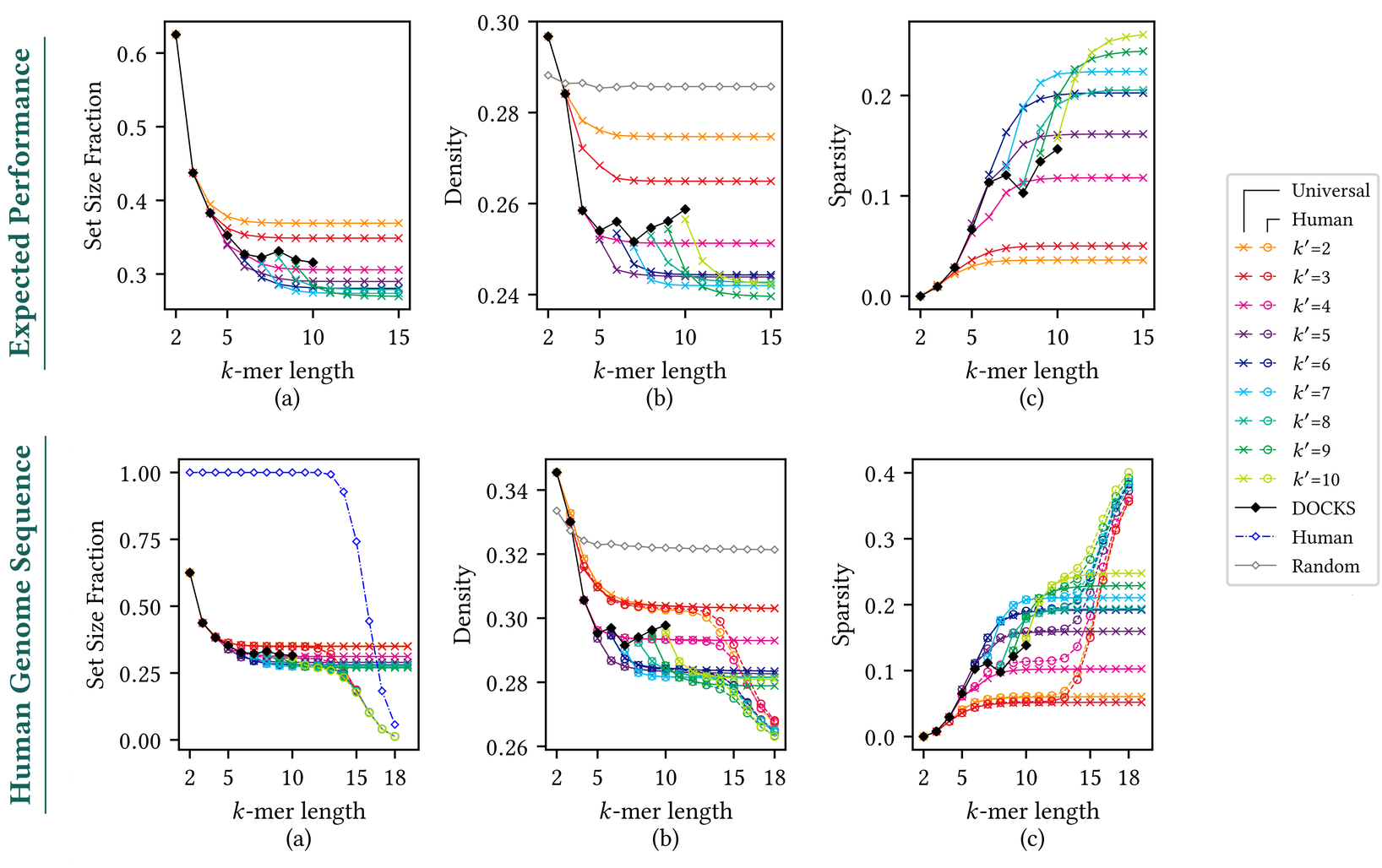
To appear at **ACM-BCB 2019**

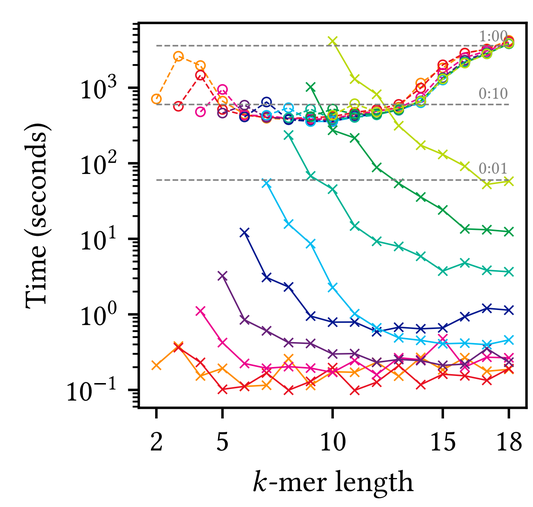
Preprint on bioRxiv

Evaluation

Comparison Metrics:

- **Size** – the number of *k*-mers chosen to be in the universal set, as the fraction of the entire set of *k*-mers.
- **Density** – using a minimizer set that is consistent with the universal set, and lexicographic for all umers
- **Sparsity** – the fraction of windows with only one umer.





References

Orenstein, Pellow, Marçais, Shamir, and Kingsford (2017) **Designing small universal k-mer hitting sets for improved analysis of high-throughput sequencing.** *PLOS Computational Biology* 13(10): e1005777. https://doi.org/10.1371/journal.pcbi.1005777

DeBlasio, Gbosibo, Kingsford, and Marçais. (2019) **Practical universal k-mer sets for minimizer schemes.** In the proceedings of the *10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (ACM-BCB '19). In Press.* https://doi.org/10.1145/3307339.3342144

Marçais, Pellow, Bork, Orenstein, Shamir, Kingsford (2017) **Improving the performance of minimizers and winnowing schemes.** *Bioinformatics (ISMB 2017)* 33(14):i110–i117. https://doi.org/10.1093/bioinformatics/btx235

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)