Space-efficient Indexing of Spaced Seeds for Accurate Overlap Computation of Raw Optical Mapping Data

Riku Walve[1] Simon J. Puglisi[1] Leena Salmela[1]

Helsinki Institute for Information Technology HIIT
Department of Computer Science, University of Helsinki

February 4, 2020

Overview

- Spaced (l, k)-mer index (from Salmela et al.)
- Compressing the index
- Space-efficient construction of the compressed index
- Overlap computation (using Valouev et al.)
- Results

Optical Mapping

- Restriction enzyme cuts DNA at specific cut sites
- Lengths between cuts are measured to form restriction maps (Rmaps)
- Rmaps are analogous to genomic reads

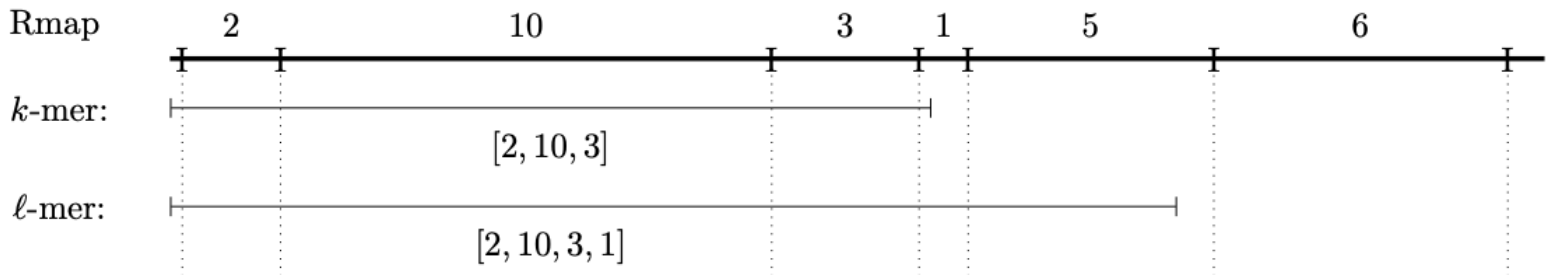
G = ccGAAaatttGAActctggcGAA
R = [0, 2, 5, 6]

Optical Mapping

- Rmaps are strings of numbers (fragment lengths)
- The strings have errors (missing/additional cut sites, sizing errors)
- We want to index the Rmaps without indexing the errors

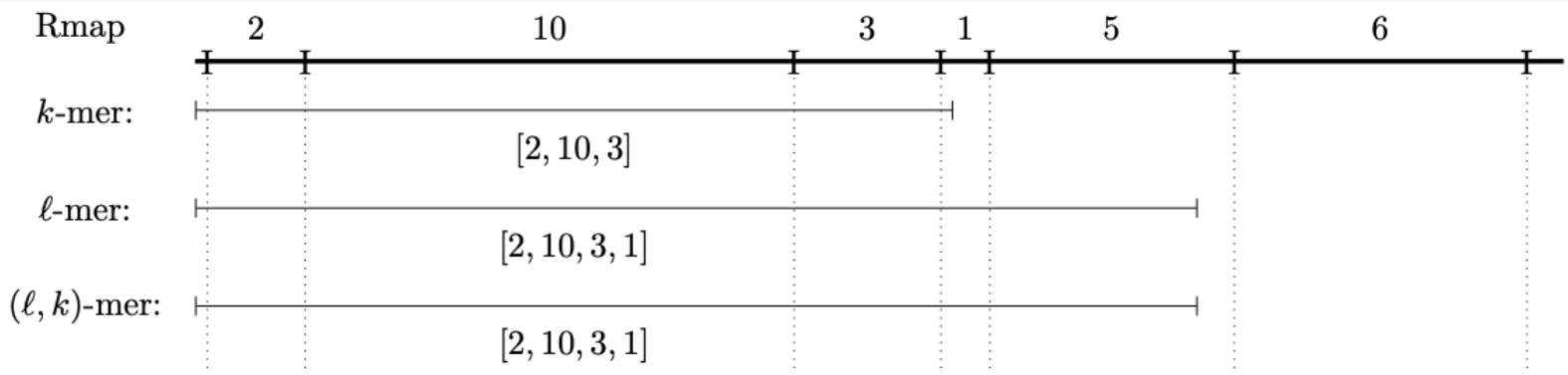
l-mers and k-mers

- k-mers are k-length substrings
- l-mers are maximal substrings with sum less than l
- l-mers are better at capturing underlying information in optical mapping data



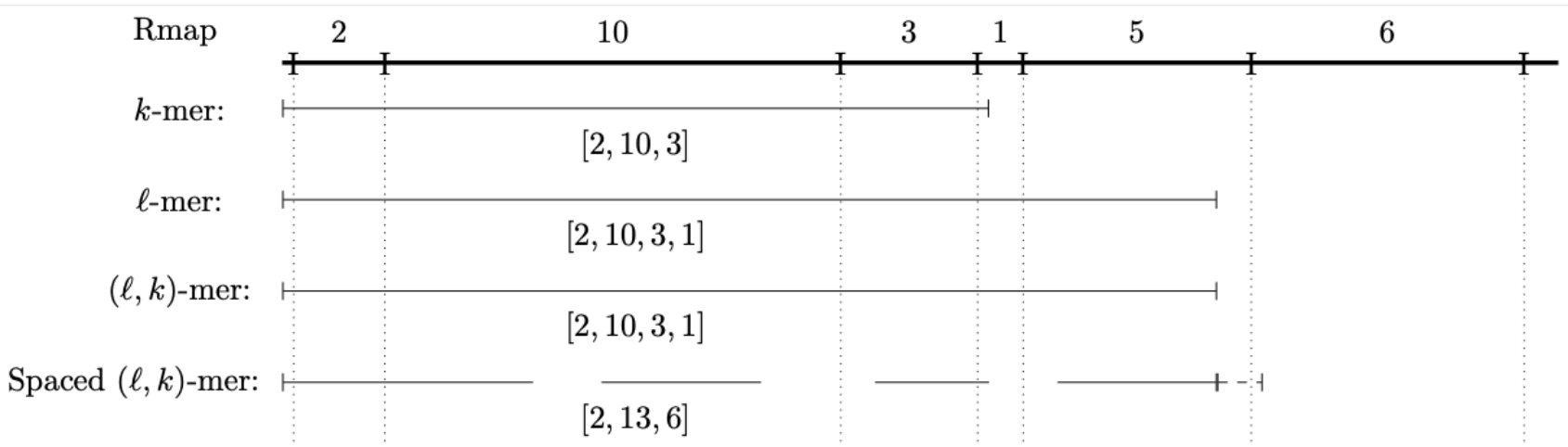
(l, k)-mers

- (l, k)-mers are l-mers with at least k fragments
- Lengths of (l, k)-mers are more predictable



Spaced (l, k)-mers

- Thanks to the predictable length, (l, k)-mers can be extended to a spaced (gapped) variant
- Spaced (l, k)-mers are (l, k)-mers with skipped positions
- Binary pattern marking skipping positions
- Skipped positions are added to the next seen value to keep the sum at l



Spaced (l, k)-mer index

- Maps spaced (l, k)-mer to lists of Rmap occurences
- Originally for error correction
- Can find similar Rmaps by looking at the lists

Merging similar spaced (l, k)-mers

- As we quantize fragment lengths, we introduce rounding errors
- We merge lists from off-by-one spaced (l, k)-mers together up to some threshold

$$s_1 = [0, 2, 4, 6]$$
$$s_2 = [0, 2, 5, 6]$$
$$s_3 = [0, 2, 6, 6]$$
$$s_4 = [0, 2, 6, 7]$$

Compression

Two distinct compression problems

- Compressing the dictionary
- Compressing the lists

Compression - the dictionary

- Take the set of spaced (l, k)-mers and construct a minimal perfect hashing function
- MPHF maps each spaced (l, k)-mer uniquely to first natural numbers
- Concatenate the occurence lists in the order of the MPHF
- Store the cumulative lengths of the lists as a sparse bitvector

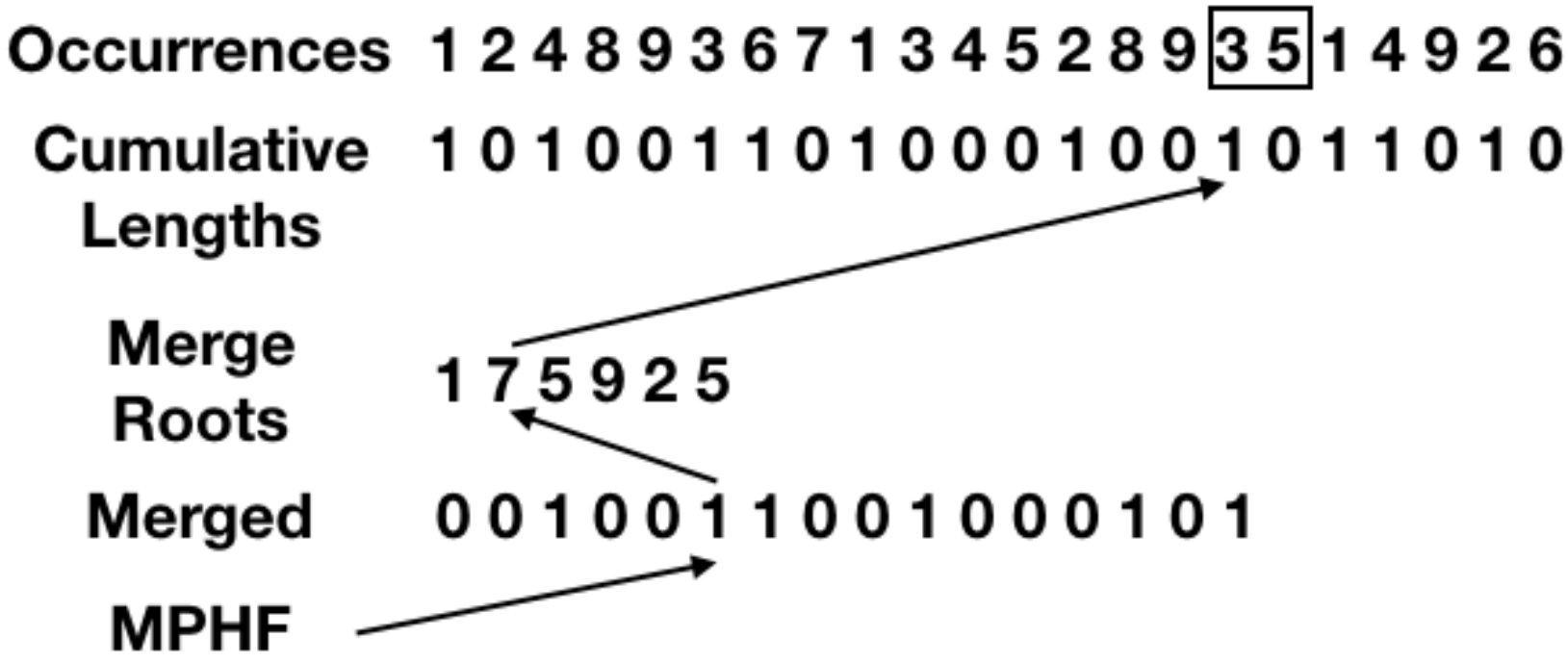
Compression - the lists

- Store the occurence lists as encoded differences
- Choice of integer coding here is arbitrary, we use VByte

Compression - merges

One additional problem: merged spaced (l, k)-mers pointing to same list

- Use a bitvector marking merged spaced (l, k)-mers
- Use an array pointing to root spaced (l, k)-mer
- Rank on bitvector to get index from array to find merge root



Construction

- Full uncompressed index is never required in memory
- Always use compressed structures to construct next step

Construction - Dictionary

- Collect all keys by filling buffer
- Sort keys on disk using multi-way merge
- Construct MPHF on disk

Construction - Merges

- Read a spaced (l, k)-mer in to memory and modify one fragment length
- If the modified value is in the index, mark one as merged
- Keep track of merge sizes to keep merges under threshold

Construction - Lists

- Partition the set based on MPHF indices
- Use MPHF to collect lists for each partition
- Compress and write disk

Overlap computation

- Valouev et al. presented a dynamic programming solution
- Sort of like the Smith-Waterman of optical maps
- We use our index to find candidates for overlaps to speed up the computation.

Results - Datasets

Data set	Genome size (Mbp)	Number of Rmaps
Ecoli1	4.6	2000
Ecoli2	4.6	129 819
Ecoli3	4.6	272
Human	3234.8	1 582 942

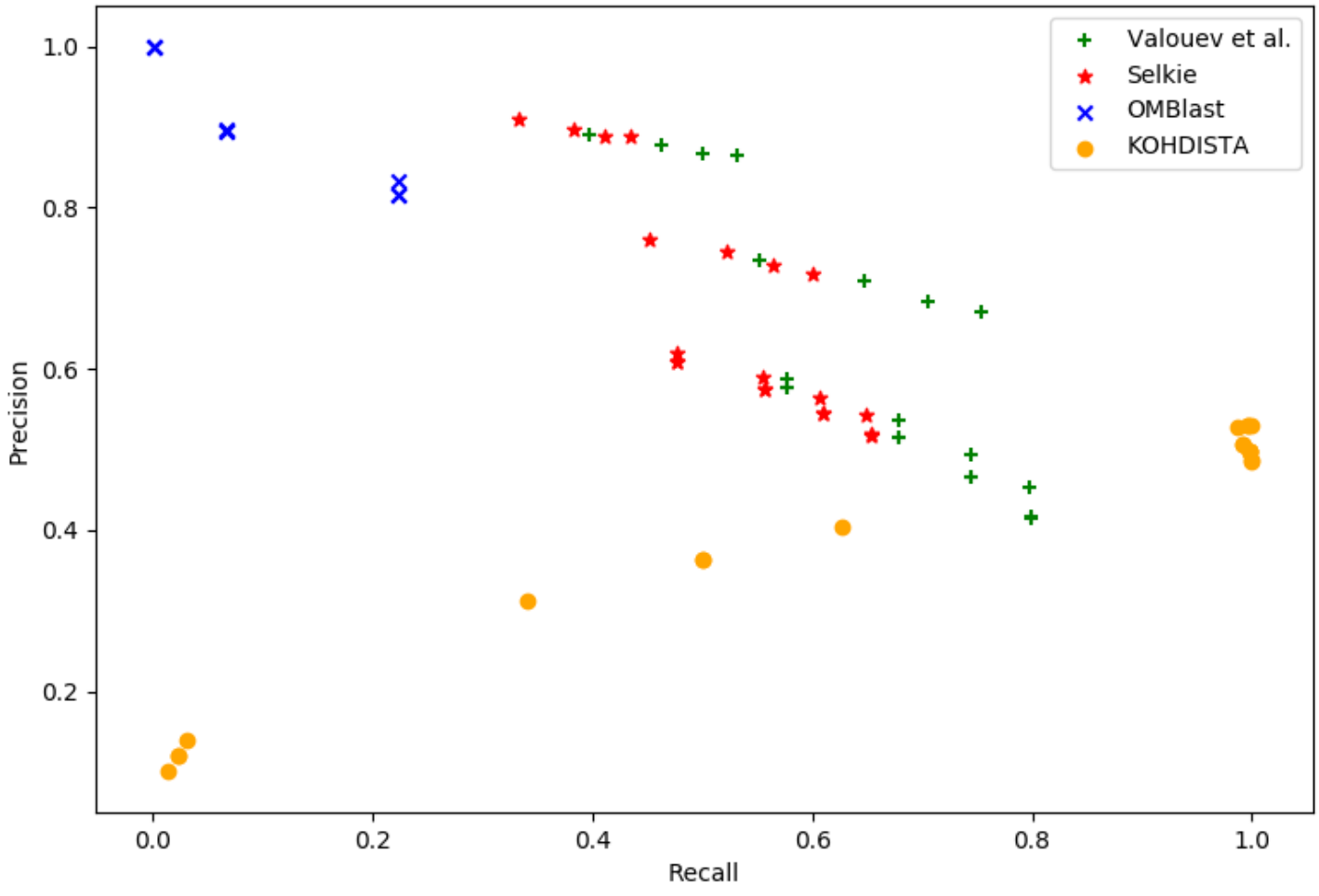
Results - Compression

	Ecoli1	Ecoli2	Human
MPHF (MB)	0.3	8.8	84.8
Merge structures (MB)	2.0	154.3	1456.2
Occurrence lists (MB)	3.6	368.6	3110.0
Total (MB)	5.9	531.7	4651.0
Uncompressed (MB)	46.2	1808.4	16302.7

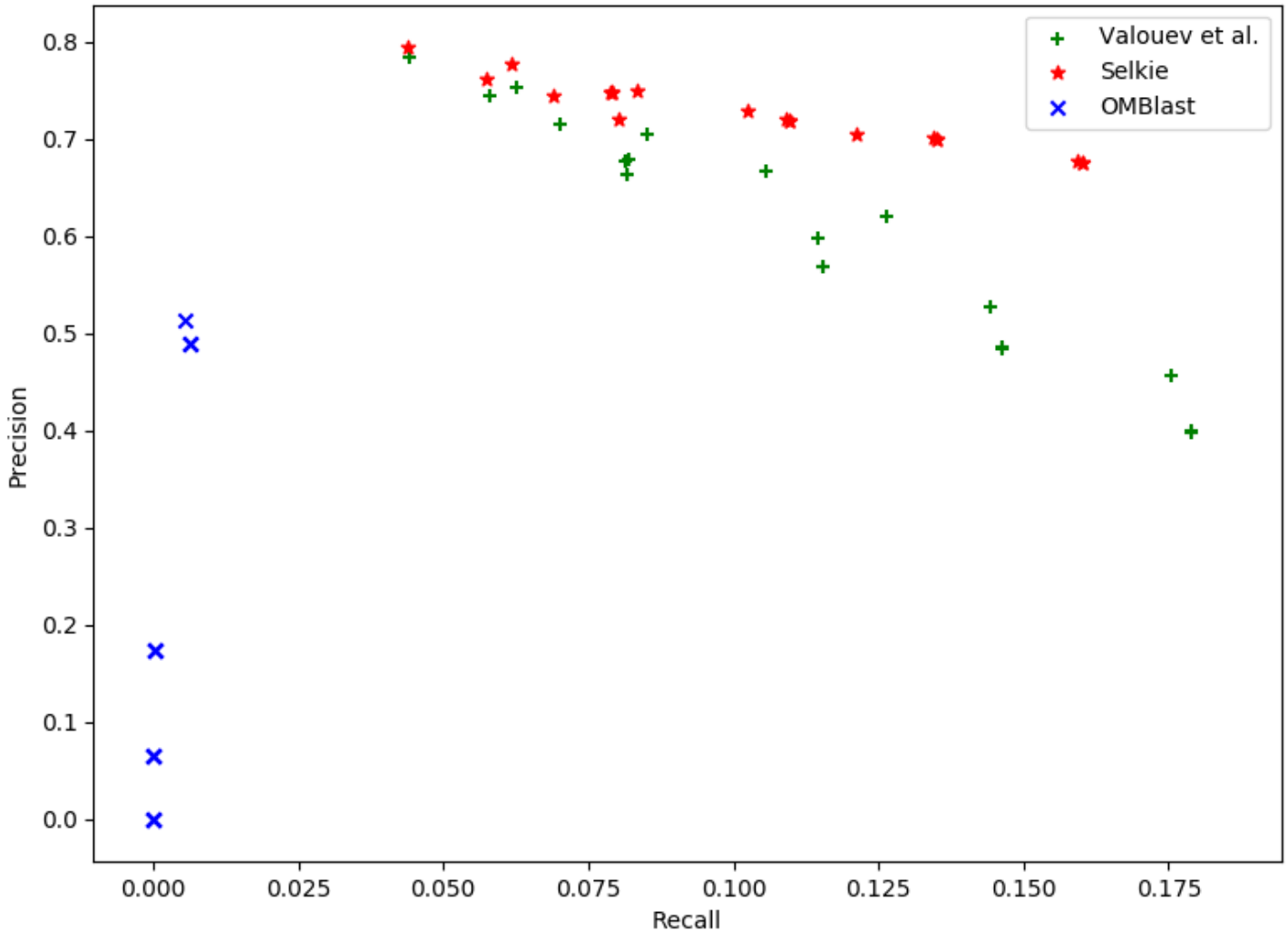
Results - Construction

Data set	Method	Runtime	Peak memory usage
Ecoli1	Uncompressed	1 min 39 s	210.71 MB
	Compressed	23 s	1.16 MB
Ecoli2	Uncompressed	39 min 5 s	8.83 GB
	Compressed	36 min 40 s	2.10 GB
Ecoli3	Uncompressed	6 s	1.16 MB
	Compressed	2 s	1.16 MB
Human	Uncompressed	8 h 59 min	84.05 GB
	Compressed	3 h 50 min	21.04 GB

Results - Overlaps - Ecoli3



Results - Overlaps - Ecoli1



Thanks

Available at `github.com/rikuu/selkie`
Paper available in the future