Boosting Backward Search Throughput for FM-Index using a Compressed Encoding

Jose M. Herruzo*, Sonia González-Navarro*, Pablo Ibañez†,
Victor Viñals†, Jesús Alastruey-Benedé† and Oscar Plata*

*Dept. Computer Architecture, University of Malaga †Dept. Comp. Science & Systems Engin., University of Zaragoza

{jmherruzo,sgn,oplata}@uma.es {imarin,victor,jalastru}@unizar.es

Contributions

We propose a new data-layout of the FM-index data structures, denoted as *split bit-vector encoding*. The goal is to store in a compact way, occupying a minimum number of cache blocks, all data needed to perform the searching process.

FM-index

Most relevant data structure is the Occ table which requires several GiBs of memory. A sampled Occ table, called SFM, stores one out of d values. The BWT is used to reconstruct the missing Occ values. Last-to-First or **LF() function** behaves similar to the rank() query. Some previous solutions use wavelet trees or look-up tables.

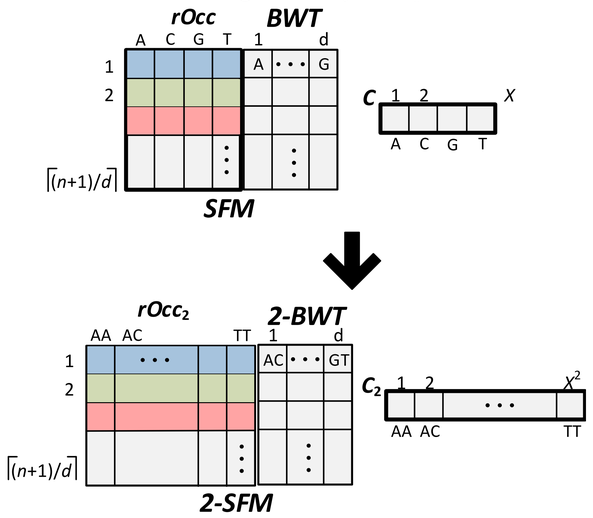
Algorithm BS: Backward Search

Input: $C[]$, $Occ[]$, Q query, n=$\|T\|$, m=$\|Q\|$
Ouput: (sp,ep): Interval pointers of Q in T

1: $sp = C[Q[m]]$, $ep = C[Q[m]+1]$
3: **for** i **from** m-1 **to** 1 **step** -1
4: $sp = LF(Q[i],sp)$
5: $ep = LF(Q[i],ep)$
6: **end for**
7: **return** (sp+1,ep)

K-step FM-index

Arranging the backward search in steps of k symbols at a time (instead of only one) improves data locality and reduces computing at the cost of increasing memory footprint for rOcc.



Computational Analysis

Memory Access Pattern: Due to the nature of the input queries and the Occ table, the search loop causes an access pattern not predictable and distributed along the whole table.

Search Throughput: To measure the computational performance we consider the *search throughput* (*ST*), defined as the number of LFMs (Last to First Mapping operation) performed per second.

Search Intensity: We define *search intensity* (*SI*) as the number of LFMs performed per transferred byte from memory, relating the number of operations to the required amount of data traffic.

Main References

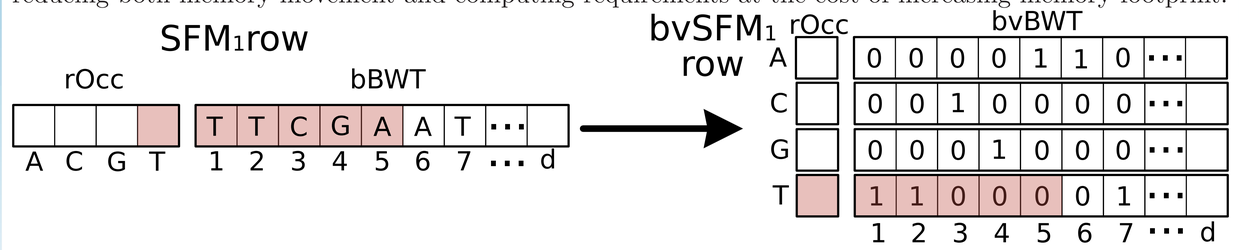
[1] Ferragina, P and Manzini, G. Opportunistic Data Structures with Applications In *41st Ann. Symp. on Foundations of Computer Science*, 2000.

[2] Chacon, Alejandro and Moure, Juan Carlos and Espinosa, Antonio and Hernandez, Porfidio n-step FM-index for Faster Pattern Matching In *Procedia Computer Science*, 2013.

[3] J.M. Herruzo, S. Navarro-González, P. Ibáñez, V. Viñals, J. Alastruey and O. Plata Accelerating Sequence Alignments Based on FM-Index Using the Intel KNL Processor In *IEEE/ACM Trans. on Comput. Biology and Bioinformatics*, 2019

Bit Vector Sampled FM-index (bvSFM)

Our new data layout, Split bit-vector, compacts all data needed to search k symbols in a single step (k-step), reducing both memory movement and computing requirements at the cost of increasing memory footprint.



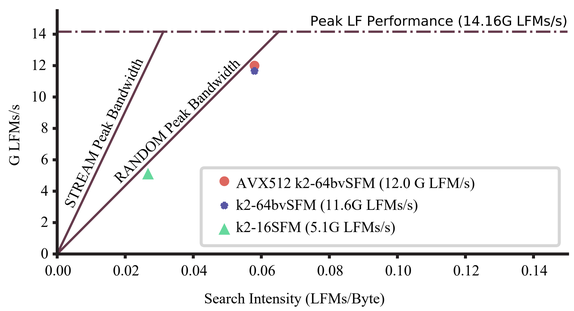
Version	Steps(k)	d	Occ entry size (B)	Occ size (GiB)	Search Intensity
SFM	1	32	24	3	0.0288
SFM	2	16	72	13.5	0.0274
bvSFM	2	64	12	12	0.0574

Methodology and Results

Experiments conducted on an Intel Xeon Phi 7210 processor (Knights Landing, or KNL), with 64 cores, 16 GiB of MCDRAM and 192 GiB of DDR4 DRAM, running Ubuntu 16.04.1 Linux.

	k	d	Theoretical Comp.	Theoretical BW	Exper.
SFM_k	1	32	5.04	6.33	4.29
SFM_k	1	192	2.15	6.44	1.69
SFM_k	2	16	8.87	6.02	5.06
SFM_k	2	128	3.38	6.38	3.16
$bvSFM_k$	2	64	14.16	12.61	11.6
$bvSFM_k$	2	96	8.76	12.70	5.91

Throughput shown on GLFM/s.



Implementation	Throughput (GLFMs/s)	Index size (GiB)
sdsl-lite library on KNL	0.455	1.25
2-Step FM-index on Kepler GTX Titan GPU	3.8	3
NVIDIA NVBIO on Tesla P100	2.7	0.23

Conclusions

Our optimized data structure packs all relevant data needed in a query step within a single cache block, minimizing the memory bandwidth demand.
Our proposal clearly outperforms previous versions. The performance obtained in the Intel Xeon Phi (KNL) architecture obtain a throughput about 3x faster than previous GPU implementations.

Acknowledgments

This work was supported in part by grants TIN2016-80920-R (AEI/ERDF, UE), TIN2016-76635-C2-1-R (AEI/ERDF, UE), University of Malaga and gaZ: T58_17R research group (Aragón Government and European Social Fund).