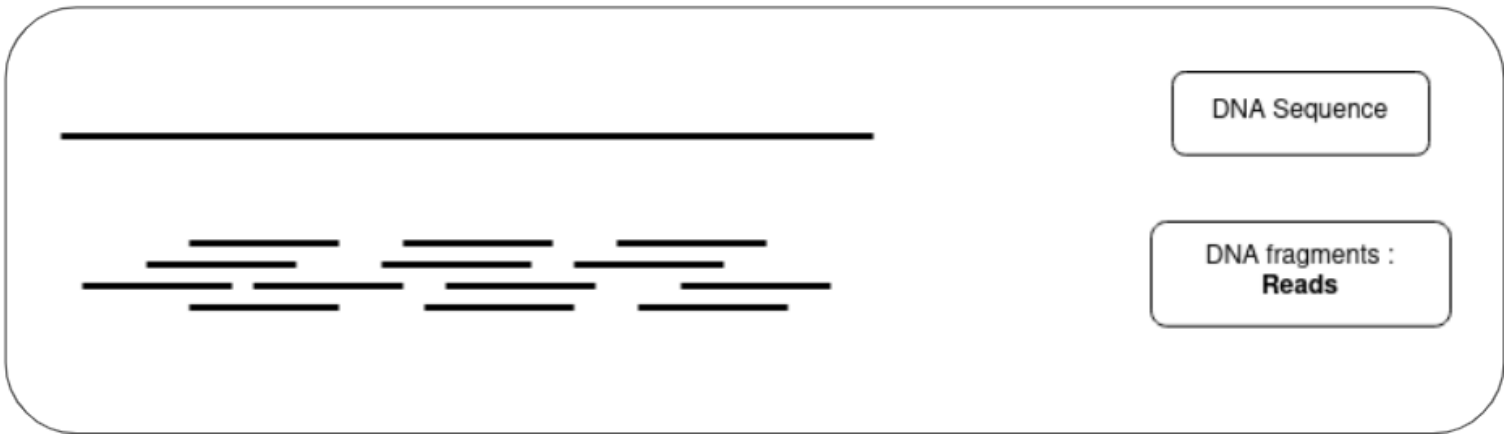
Kmer2Reads, an associative index for Third Generation Sequencing data

Léa Vandamme, Bastien Cazaux and Antoine Limasset

Univ. Lille, CNRS, UMR 9189 - CRIStAL, F-59000 Lille

SeqBIM
November 17, 2022

Genome sequencing

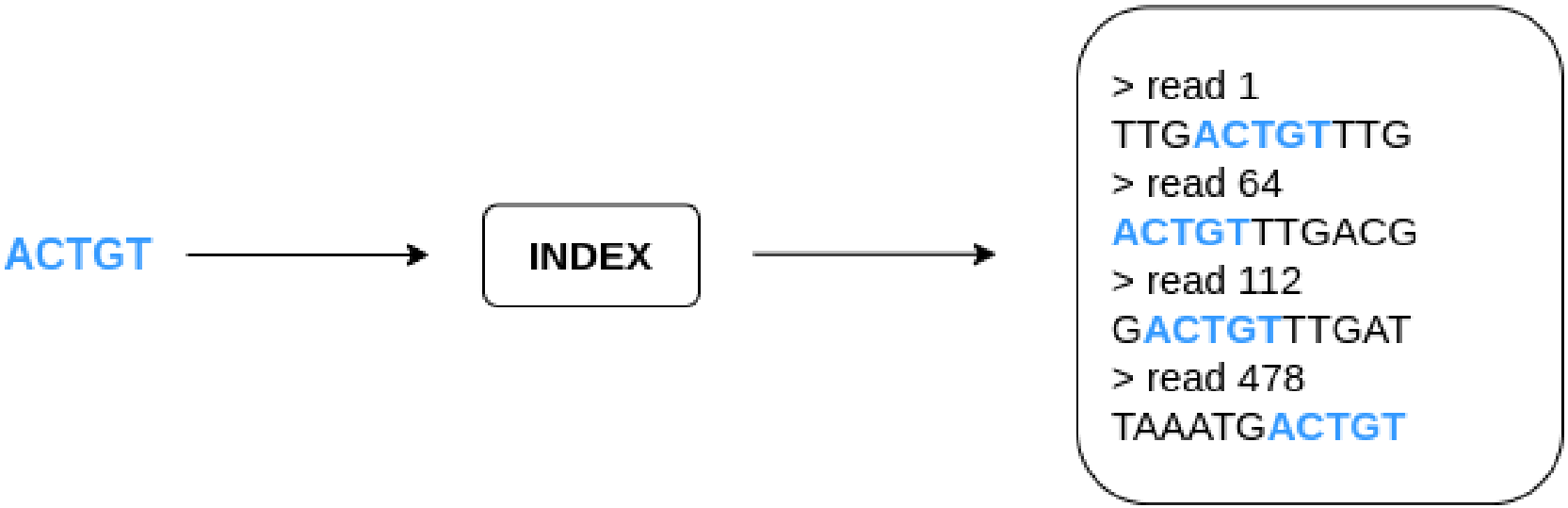


- Error rate : from 0.1% (HiFi) to 10% (ONT)
- Length : 10 - 100 kilo bases

Assembly

Human genome : 2022 - dozen of laboratories

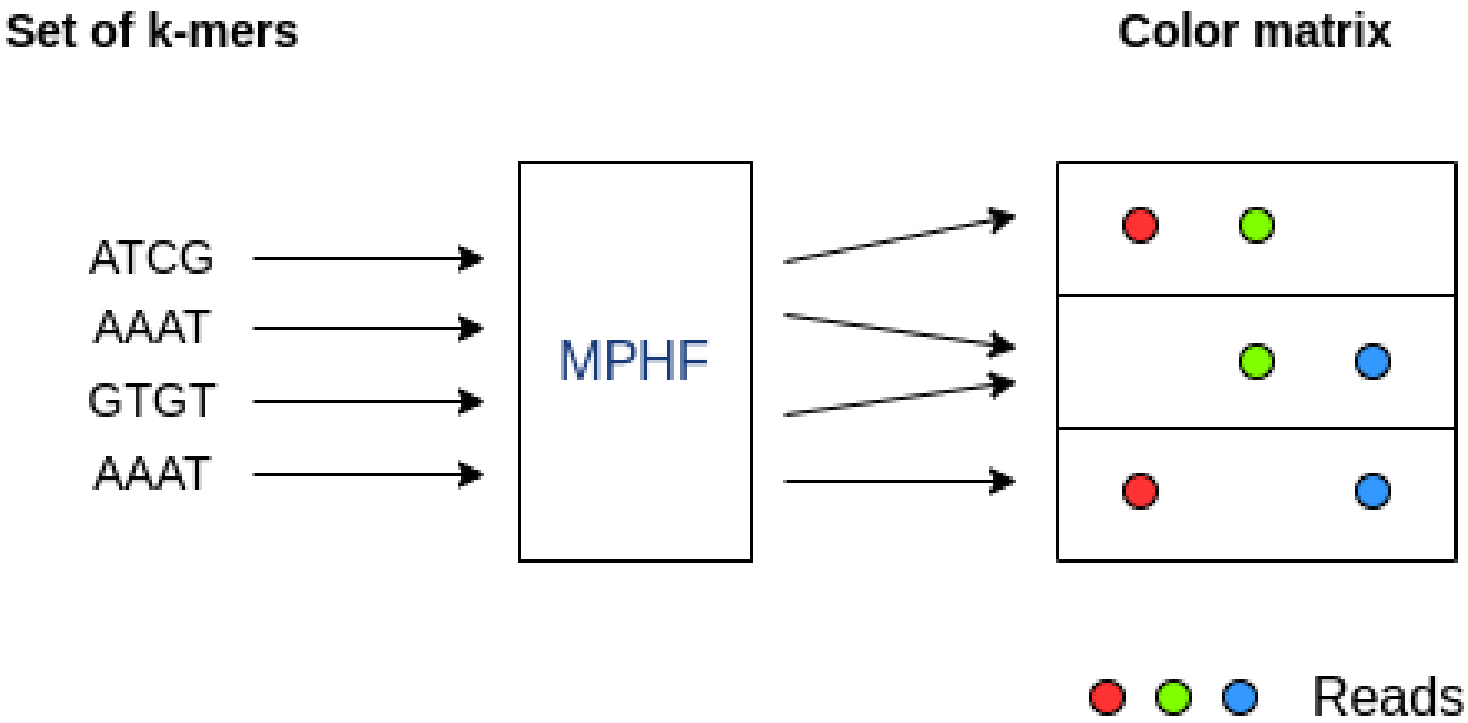
How to study reads ?



Check the **presence** and **locate** a sequence in a genome
Associate k-mers to reads in which they appear

State of the art : hashing methods

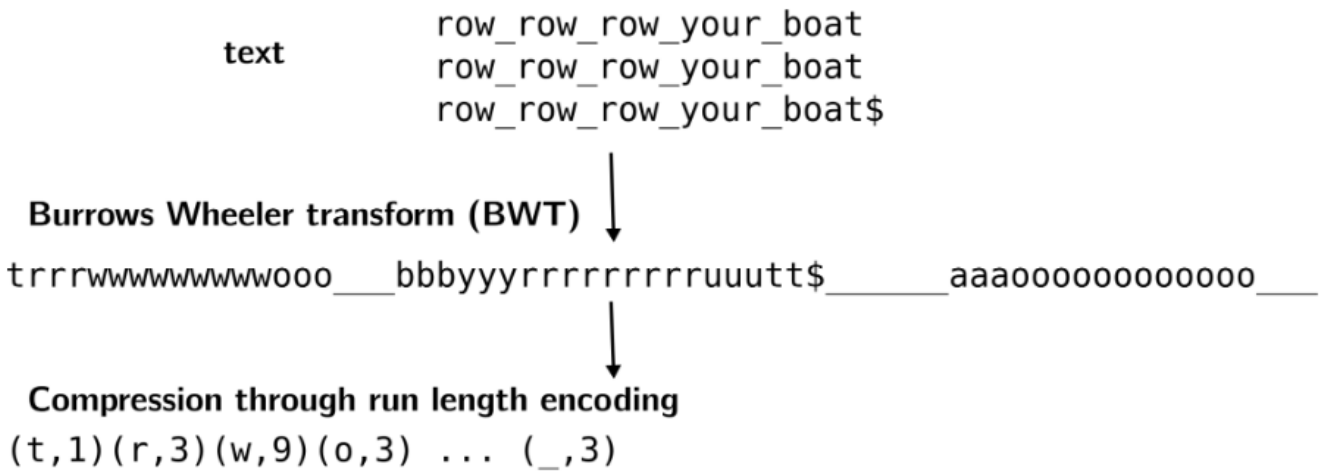
Based on hash tables and minimal hashing perfect functions.



Use example : BLight (Marchet, 2021), SRC (Marchet, 2020), Pufferfish (Almodaresi, 2018), HARC (Chandak, 2017)

State of the art : Full-text indexing (BWT)

Locate occurrences of patterns in a text



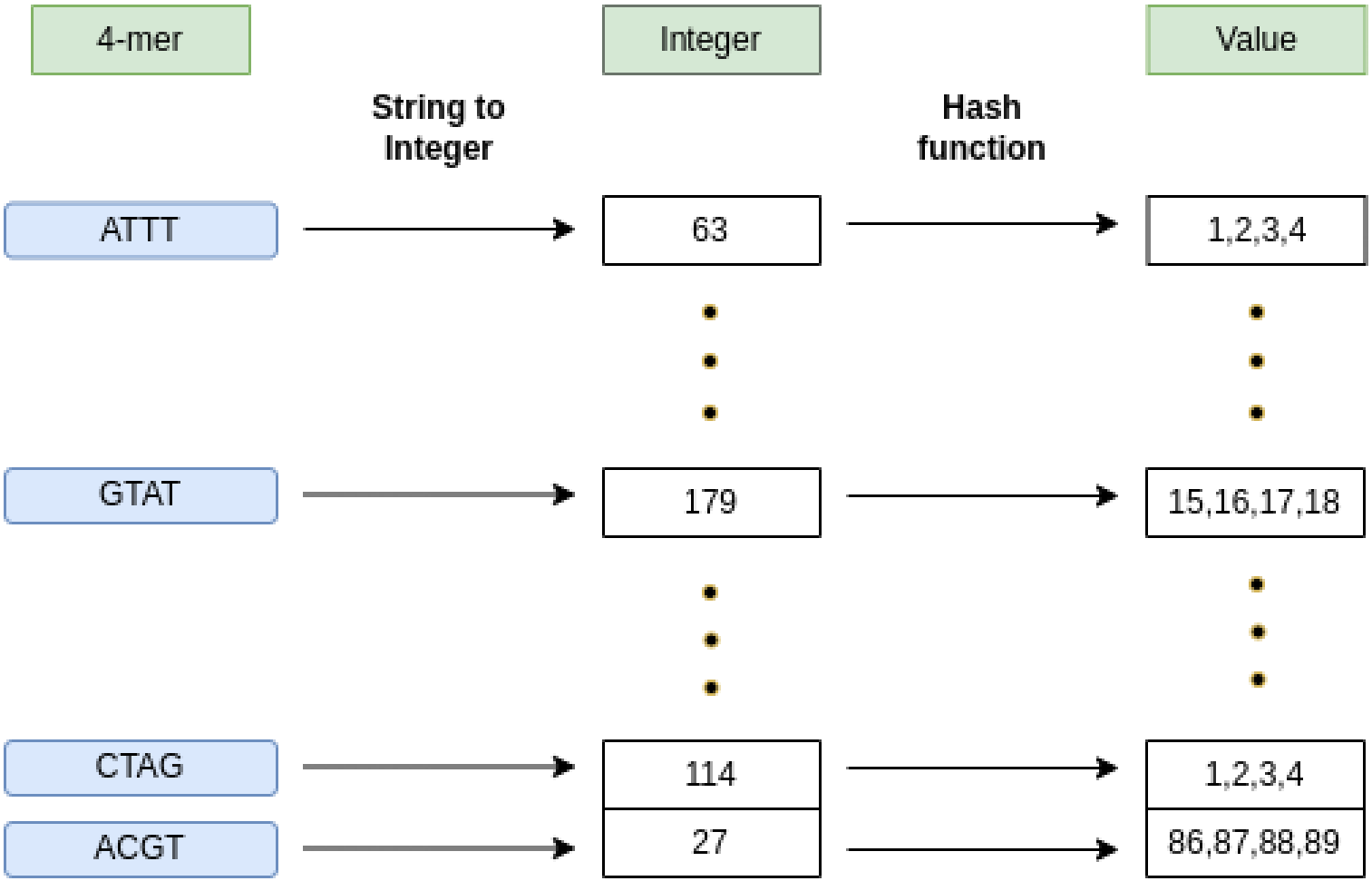
Variant of FM-index : r-index (Mun, 2020)

(Number of RLE optimization)

Goal : Efficient k-mer to read index

	Construction	Memory	Debit
Hashing methods (SRC)			
Full-text indexing (r-index)			~

Classical hashing index



Classic hash table & Robin Hood & MPHF

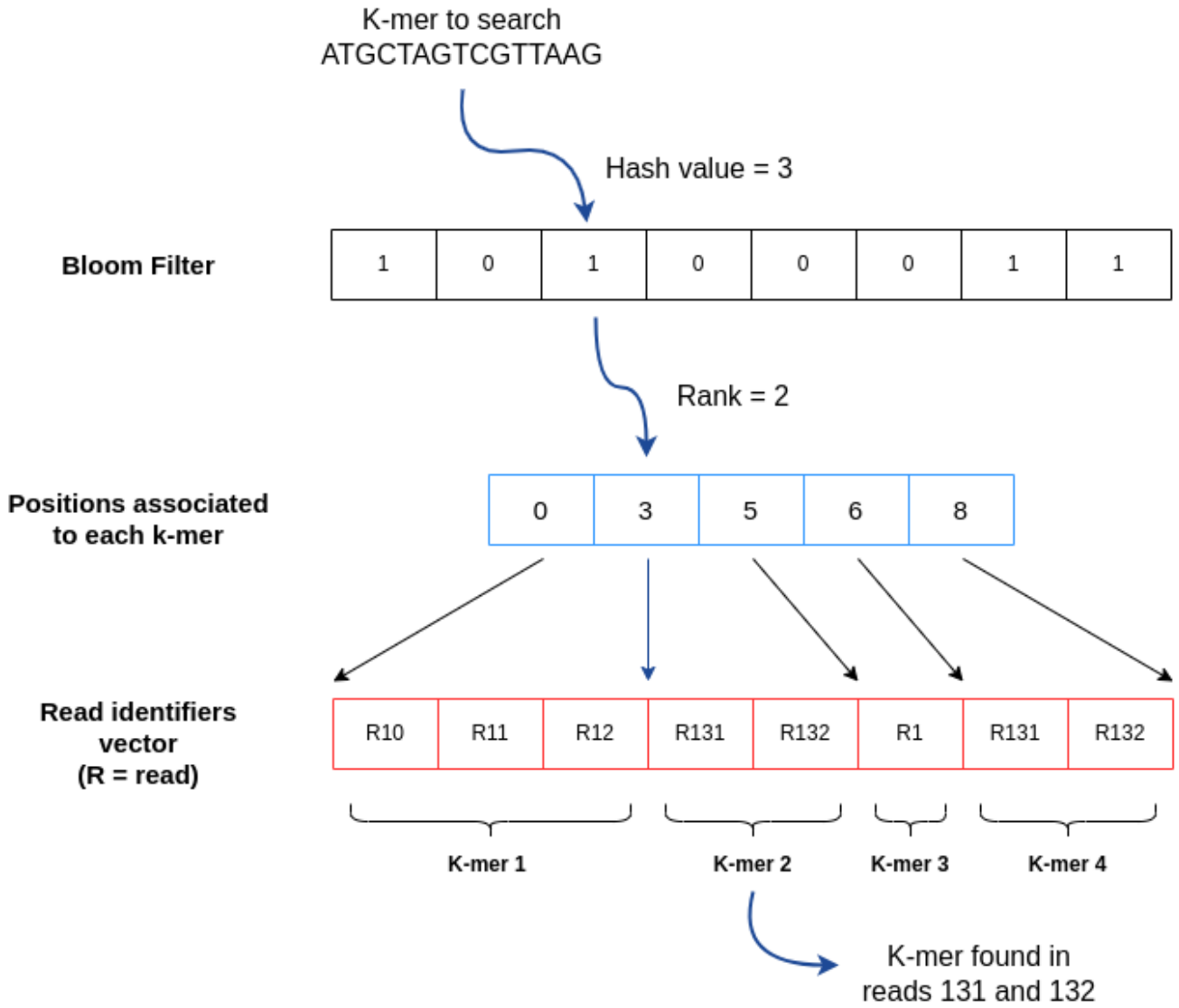
Limits

- Number of vector (1 per k-mer : billion for human genome)
- Vector size (4 bytes x hundreds of read id)

Solutions

- Unique vector
- Integer compression

K2r (K-mer to reads)



Compression

Compress the vector containing the reads.

Delta encoding : for sorted integer lists
[1000000, 1000005, 1000006, 1000010] = [1000000, +5, +1,+4]

- **TurboPFor**
High throughput, a few bit per identifier

`https://github.com/powturbo/TurboPFor-Integer-Compression`

Results : Construction

Input data

1. Mycoplasma (∼800Kb)
2. E.Coli (∼5Mb)
3. S.Cerevisiae (∼12Mb)

K-mer length : 15

Index creation

Error rate	Dataset	Memory (Mo)			CPU time (s)		
		r-index	SRC	K2r	r-index	SRC	K2r
0.1%	Mycoplasma	128.7	661	405.1	7.4	17.6	30.6
	E.Coli	1011.7	3113.5	1956.1	74.3	205	312.2
	S. Cerevisiae	623.3	7545	6986.8	97	630.5	801.1
1%	Mycoplasma	440	790.7	343.9	19.9	32.6	34.4
	E.Coli	3476.9	6084.2	2058.1	198.7	413.2	329.3
	S. Cerevisiae	8997.7	13664.9	7141.9	584.3	1138.1	823.4
10%	Mycoplasma	984.3	2037.6	323.1	45	95.3	37
	E.Coli	8373.6	13510	2147.8	467.6	467.6	323.2
	S. Cerevisiae	X	22587.3	9284.3	X	1816.6	827.5

Results : Query

Queried data : 31-mers, simulated from the corresponding genome.

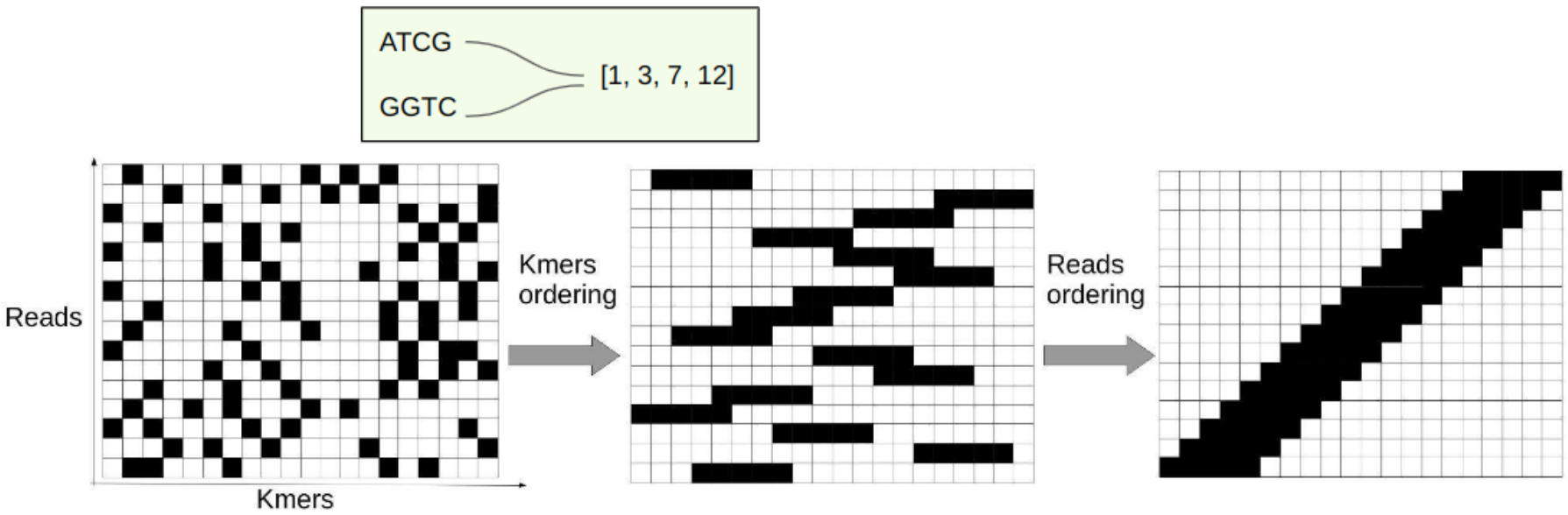
Index query

Error rate	Dataset	Memory (Mo)			CPU time (s)		
		r-index	SRC	K2r	r-index	SRC	K2r
0.1%	Mycoplasma	14	259.4	460.3	0.5	3.3	0.4
	E.Coli	98	2001.2	608.1	5.3	32.4	1
	S. Cerevisiae	116	5160.3	1278.4	23.5	158.4	3
1%	Mycoplasma	52	604.8	544.4	0.5	3.5	0.07
	E.Coli	459	4585.6	642.9	6.2	32	1
	S. Cerevisiae	1268	10510.3	1618.3	19.8	155.5	2.9
10%	Mycoplasma	248	677.3	560.3	0.5	3.9	0.1
	E.Coli	2256	12713.3	813.5	4.4	29.5	0.6
	S. Cerevisiae	X	22879.5	2800.5	X	90.2	2.6

Conclusion

	Construction	Memory	Debit
Hashing methods (SRC)			
Full-text indexing (r-index)			~
K2r		~	

Future improvements



- ➔ Less redundant vectors
- ➔ Improve delta encoding

Take home messages

Perspectives

- Sort k-mers & reads
- Super k-mers
- Full-text indexing

Take home messages

Objective

- Associate k-mers to reads
- In order to locate sequences in genomes

Benchmark

- Comparison with Full-text indexing and Hashing methods

Implementation

- Our new tool : K2r

Results

- Low query time (less than 3s to query on a 12Mb genome)
- Memory cost can be improved in case of low error rate