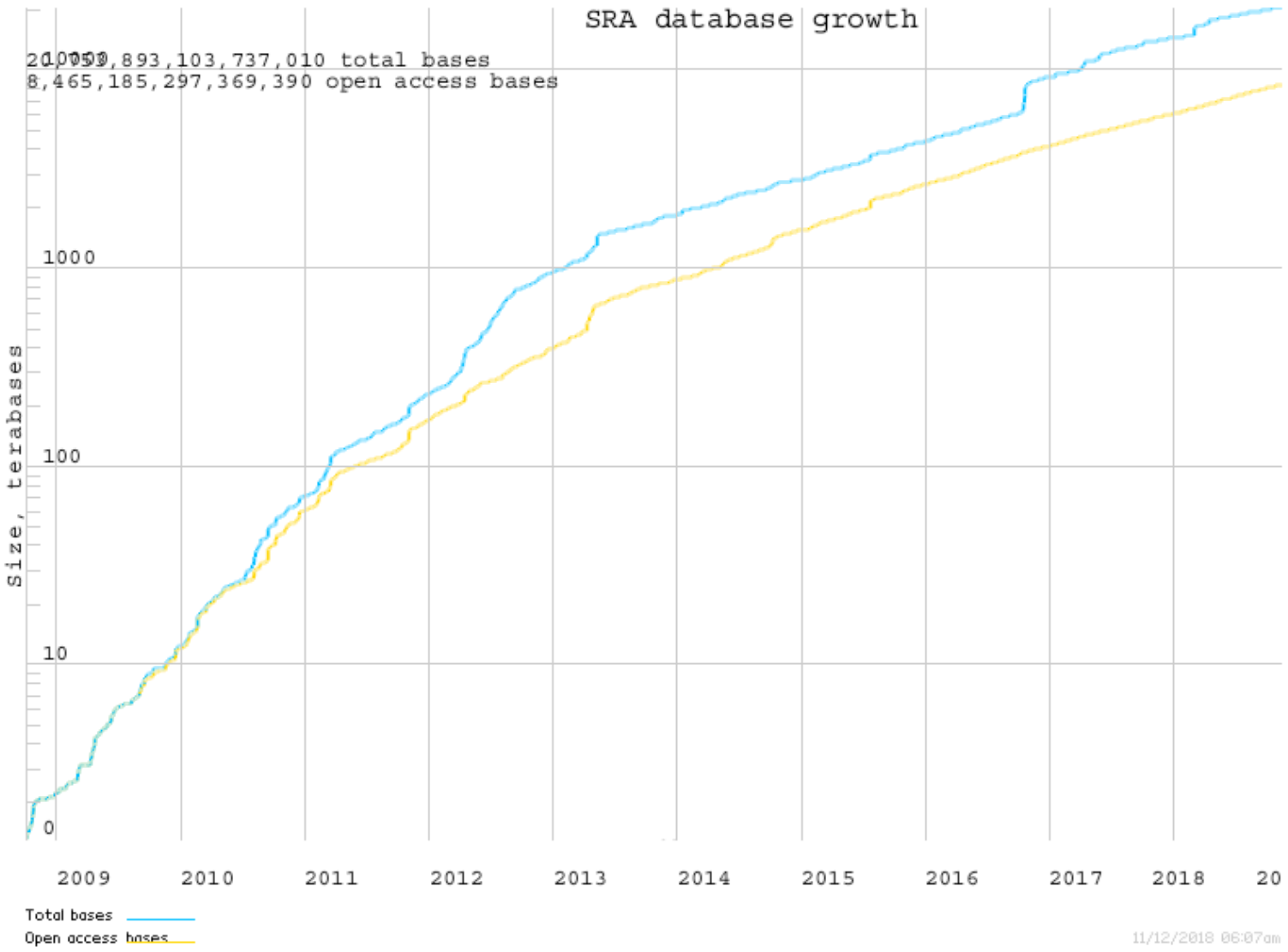
Indexing de Bruijn graph with minimizers

Antoine Limasset

Bonsai Team, CRIStAL, Université de Lille, CNRS

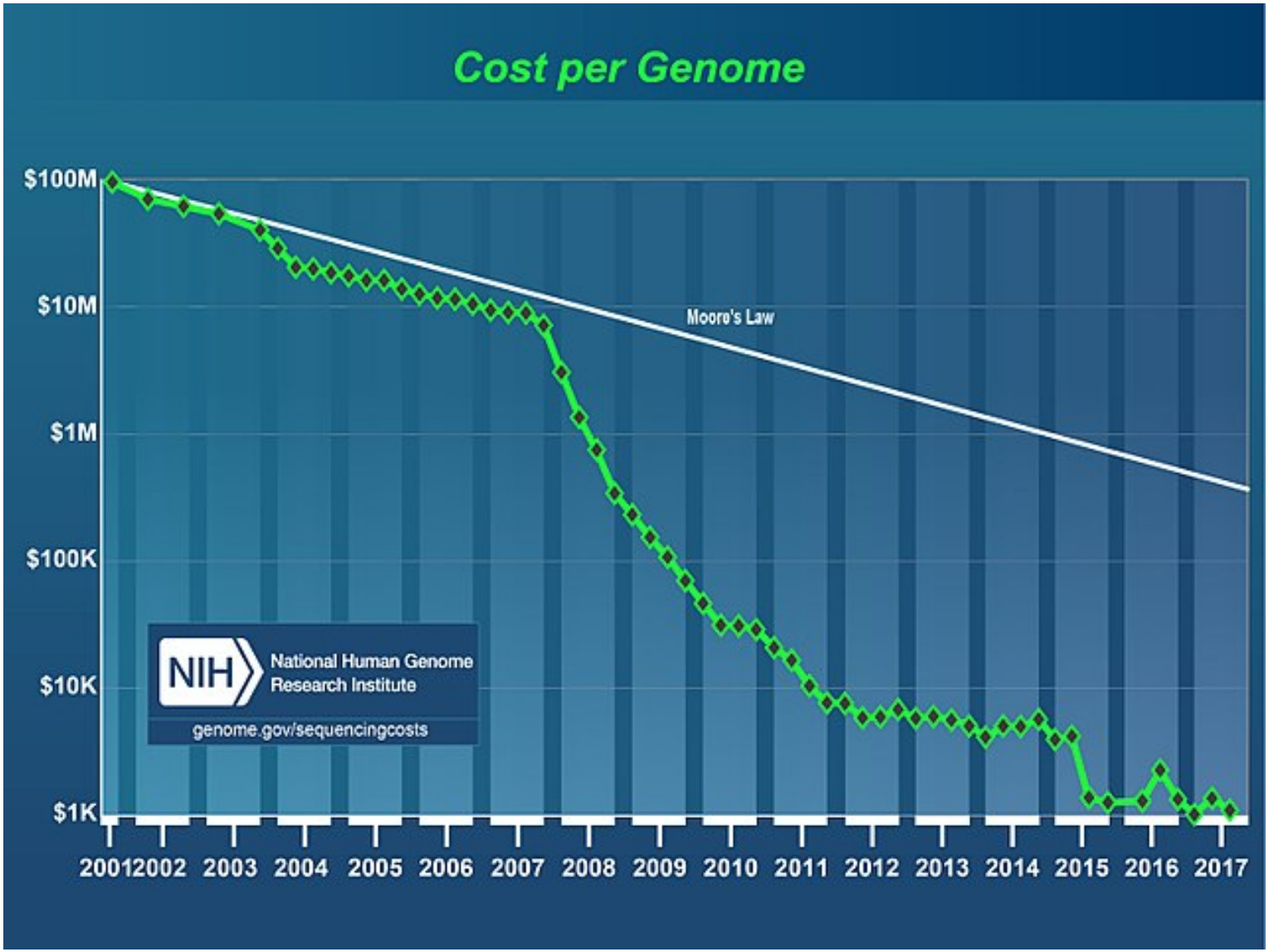
November, 2018

Data Deluge



NovaSeq: 1TB/day

Decreasing cost



100$ Human genome incoming

Omni-Genomic ?

Can biologists sequence the genomes of all the plants and the animals in the world, including this greater bird of paradise in Indonesia? TIM LAMAN/NATIONAL GEOGRAPHIC CREATIVE

Biologists propose to sequence the DNA of all life on Earth

By Elizabeth Pennisi | Feb. 24, 2017 , 1:15 PM

Can we work with it ?

Kmer/word associative indexing

- CATGCTAGCATACG-> Found at position 987,654
- AAGTTACGTACGAT-> Present in dataset "Nadine12"
- TTCGATTCGGTGGG-> Seen 666 times

Fundamental problem

- Sequence similarity (BLAST)
- Overlap detection (Minimap)
- Genome comparison (Mummer)
- Variant calling (Cortex)
- Quantification (Kallisto)
- Assembly (SPAdes)
- ...

Genome Size

Species	*T2 phage*	*Escherichia coli*	*Drosophila melanogaster*	*Homo sapiens*	*Paris japonica*
Genome Size	170,000 bp	4.6 million bp	130 million bp	3.2 billion bp	150 billion bp
Common Name	Virus	Bacteria	Fruit fly	Human	Canopy Plant

From `http://ib.bioninja.com.au`

- Pangenome
- Meta-Genome
- Environmental meta-genome

Scaling problem

How to index 10^{10}, 10^{11}, 10^{12} kmers ?

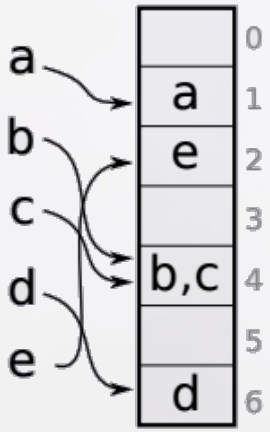
Hash functions

a,b,c,d,e : hashable elements (e.g. strings, integers, etc..)

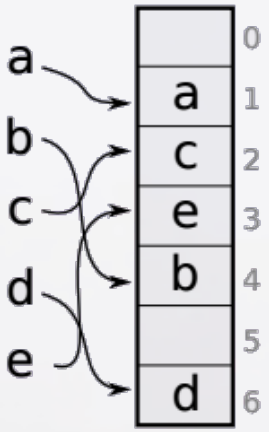
→ : hash function

: image [0;m] of hash function
(e.g. indices of buckets in a hash table)

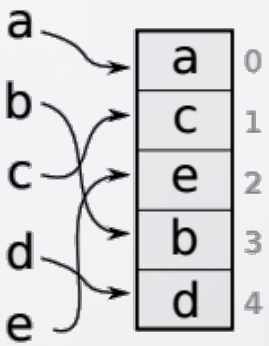
Classical hashing



Perfect hashing
(no collisions)



Minimal perfect hashing
(no collisions, |image|=|input|)



Constructing a perfect hash means finding the function (→).
Note, storing the function occupies much less space than input elements.

BBhash[1] library

Method	Query time (ns)	MPHF size (bits/key)	Const. time(s)	Const. memory (bits/key)
BBhash	216	3.7	35	4.3
EMPHF	246	2.9	2,642	247.1
EMPHF HEM	581	3.5	489	258.4
CHD	1037	2.6	1,146	176.0
Sux4J	252	3.3	1,418	18.10

Achieved the construction of a trillion key MPHF

[1] Limasset, Antoine, et al. "Fast and scalable minimal perfect hashing for massive key sets." SEA(2017).

Alien problem

MPHF have undefined behavior on non-indexed keys (alien keys)

BBhash accept aliens

- Key -> [Value]
- CATGCTAGCATACG -> Found at position 987,654
- AAGTTACGTACGAT -> Present in dataset "Nadine12"
- TTCGATTCGGTGGG -> Seen 666 times

TGTGTGTGTGTGTGTGTG -> Present in dataset "Nadine12"

Classic solution

Keep the original key

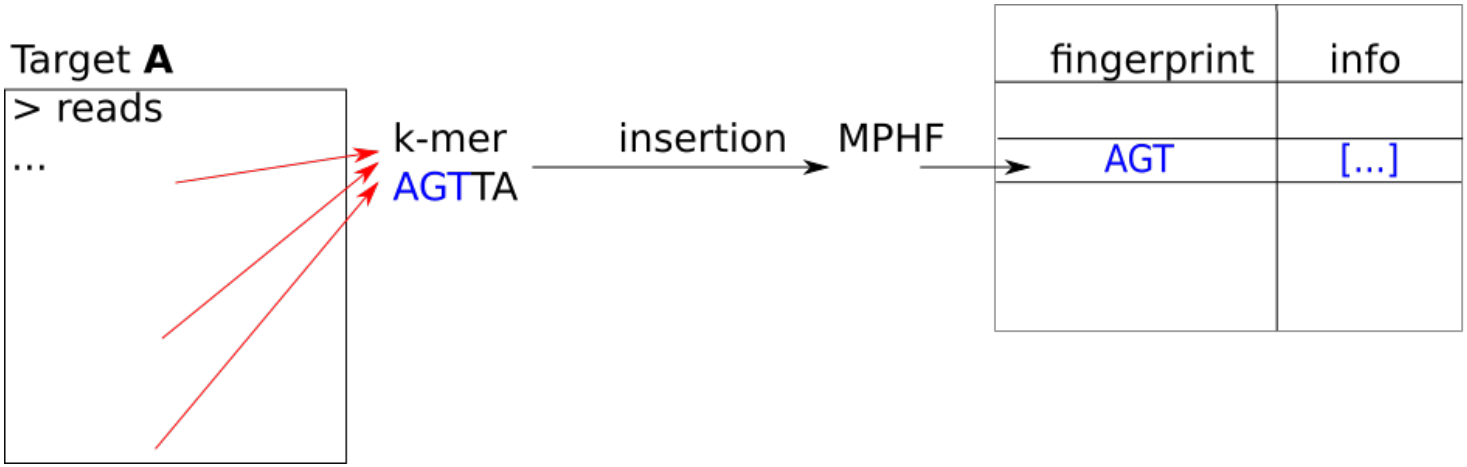
- Key -> [Key,Value]
- CGTCGTCGT-> [AAGTTACGTACGAT,Seen 666 times]
  Alien key detected !

Memory cost per key

MPHF: half a byte
32mer: 4 bytes
64mer: 8 bytes

25 GB for a human genome 1.2 TB for *P. Japonica*

Quasi-dictionnary



SRC-Linker[a]

- Key -> [Fingerprint,Value]

A fingerprint of f bits mean a false positive rate of $\approx 1/2^f$

[a] Marchet, Camille, et al. "A resource-frugal probabilistic dictionary and applications in bioinformatics." Discrete Applied Mathematics (2018).

Default value $f = 12$

Represent ≈ 2 bytes per kmer for a false positive rate of 0.02%

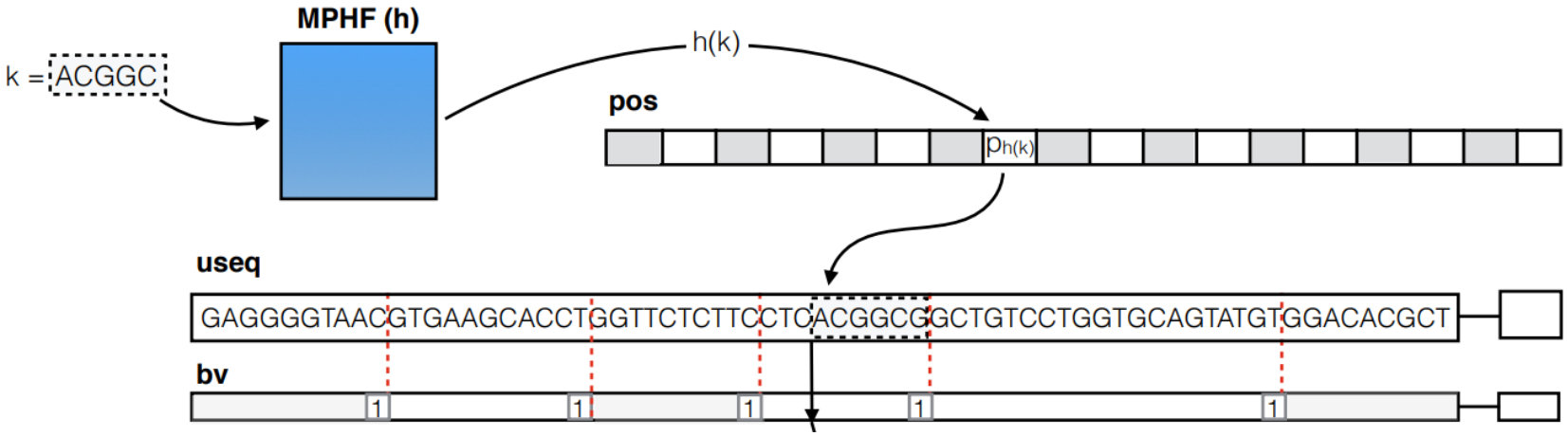
Using a reference graph

De Bruijn graph reference

A compacted de Bruijn graph can store efficiently a set of kmer (May be as low as 2 3 bit per kmer)

CATGCATGACTGACTGCTGCATCGTAGCTCGATCGTCAGTC
Represent 30 11mer with 41 nucleotide (>3 bits per kmer)

Pufferfish[2]



Reference graph encoding

- Key -> [Position in the graph,Value]

Achieved a rate memory usage of 12.5 GB for a human genome (35 bit per kmer)

[2] Almodaresi, Fatemeh, et al. "A space and time-efficient index for the compacted colored de Bruijn graph." Bioinformatics 34.13 (2018): i169-i177.

Partition

Pufferfish memory bottleneck

Position field $\approx log2(genome_Size)$

Position field of partitioned graph $\approx log2(\frac{genome_Size}{number_partition})$

Using minimizer

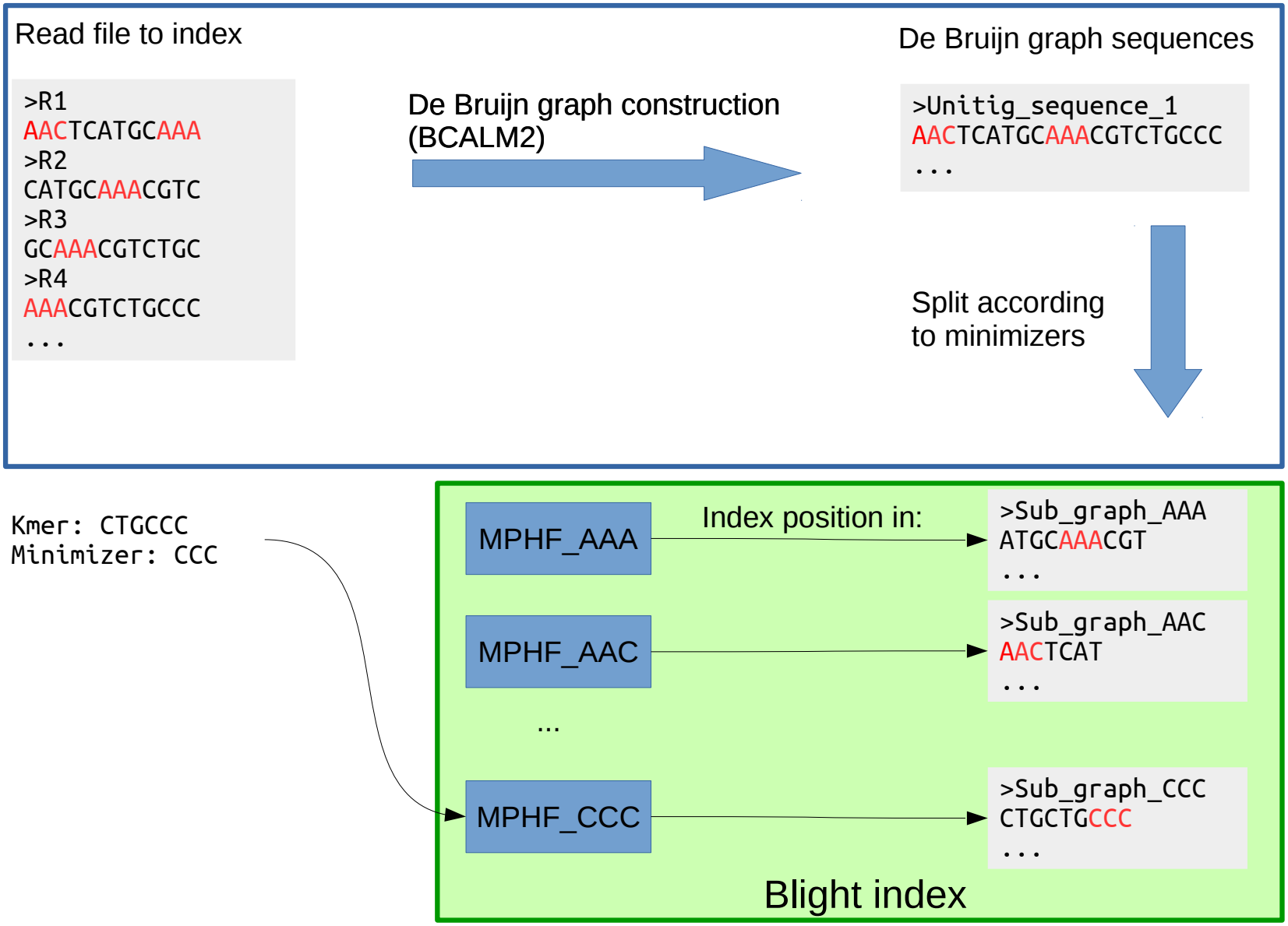
Partition the kmers according to their minimizer

Index each partition separately

Various advantages

- Parallel construction
- Cache coherence during query

Blight


Read file to index
>R1
AACTCATGCAAA
>R2
CATGCAAACGTC
>R3
GCAAACGTCTGC
>R4
AAACGTCTGCCC
...
De Bruijn graph construction (BCALM2)
De Bruijn graph sequences
>Unitig_sequence_1
AACTCATGCAAACGTCTGCCC
...
Split according to minimizers
Kmer: CTGCCC
Minimizer: CCC
MPHF_AAA
Index position in:
>Sub_graph_AAA
ATGCAAACGT
...
MPHF_AAC
>Sub_graph_AAC
AACTCAT
...
...
MPHF_CCC
>Sub_graph_CCC
CTGCTGCCC
...
Blight index

Memory result

Minimizer size	Graph sequences	Positions	Total
8	10	12	26
10	12	9	25
12	13	6	24
Pufferfish			35

Pufferfish used 12.5 GB for the human Genome
Blight objective: 8 GB

Time result

Whole human genome

- Construction time: 3,064
- Query time 311
- Pufferfish construction time: 4,248
- Pufferfish query time: 1,331

Objectives

Efficient AND user-friendly library

- Single header to include
- Serialization (index saved on disk)
- Results on largest genome, pangenome, metaGenome

Optimization

- Direct spitted graph construction
- Successive positive query (50 sec on human genome)
- Specialized minimizer scheme

The end