Minimal perfect hash functions in large scale bioinformatics Problem

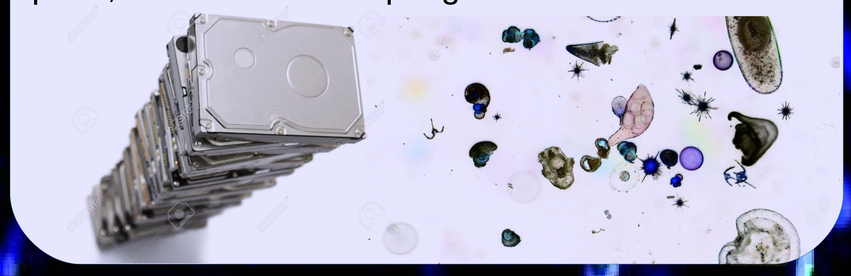
Antoine Limasset, Camille Marchet, Pierre Peterlongo, Lucie Bittner

Minimal perfect hash functions in large scale bioinformatics

Antoine Limasset, Camille Marchet and Pierre Peterlongo
INRIA/IRISA/GenScale, Campus de Beaulieu, 35042 Rennes Cedex, France
antoine.limasset@irisa.fr, camille.marchet@irisa.fr, pierre.peterlongo@irisa.fr

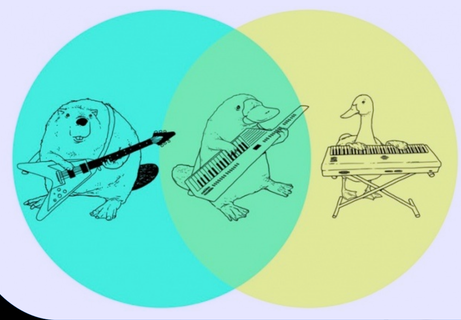
(Meta)Genomic Data

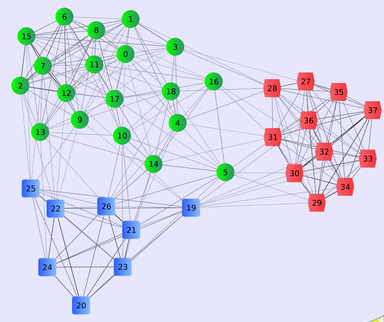
Billions of short sequences of hundreds of base pairs, from one or multiple genomes



Questions

Dataset comparison :
Detection of similar reads **inter** or **intra** datasets





Problem

Indexing huge set of elements

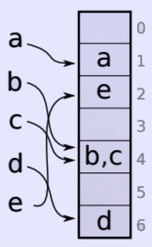
Hash functions

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

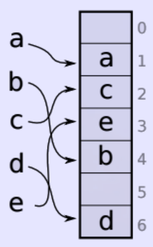
⟶ : hash function

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

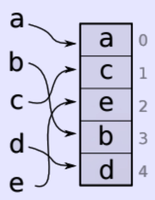
Classical hashing

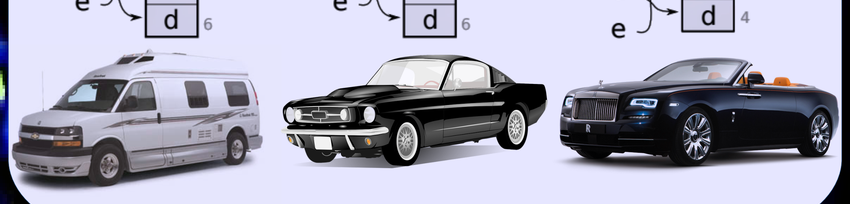


Perfect hashing (no collisions)



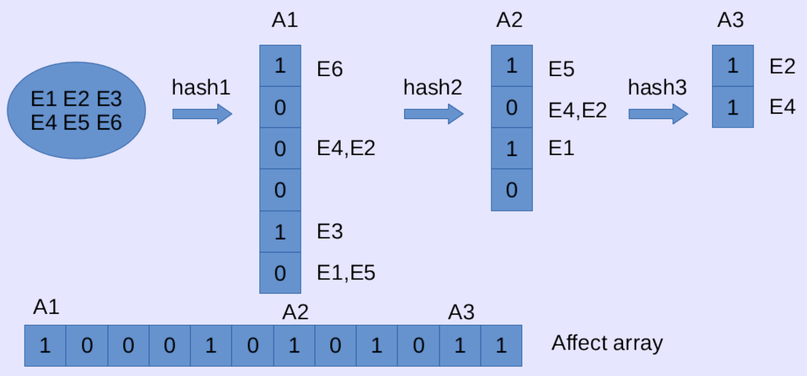
Minimal perfect hashing (no collisions, |image|=|input|)





BBhash library

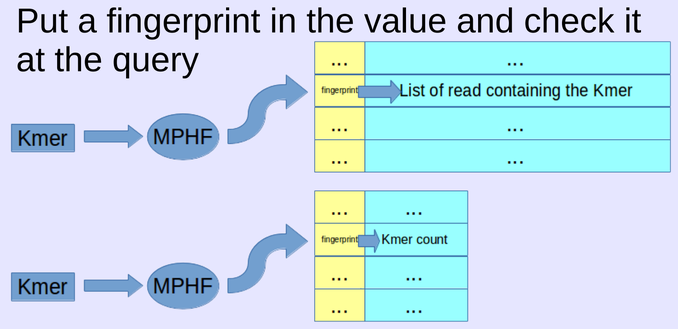
-Memory efficient (less than 3bits per key)
-Fast query (200ns)
-Fast to construct (even for billions elements)



BUT
-No membership operation
-A ‘stranger’ key can be associated to a value

Quasi-dictionary

Put a fingerprint in the value and check it at the query



False positive rate :

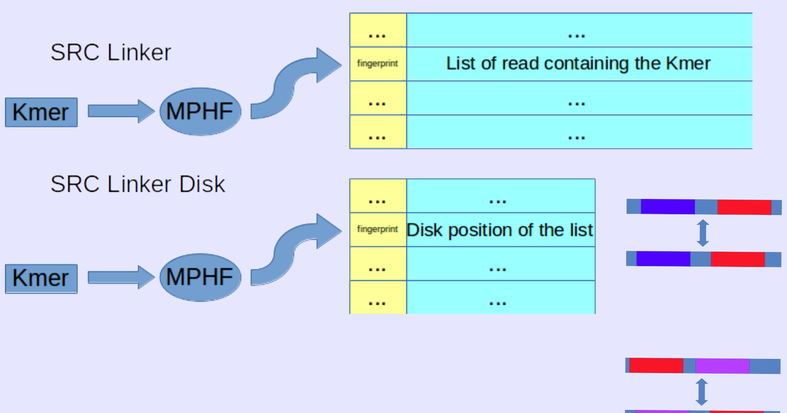
$$\frac{2^{(2*k-f)}-1}{2^{2*k}} \approx \frac{1}{2^f}$$

Memory consomption :
Bit/elements : 10 FP rate: $1/10^2$
Bit/elements : 20 FP rate: $1/10^5$
Bit/elements : 30 FP rate: $1/10^8$
...

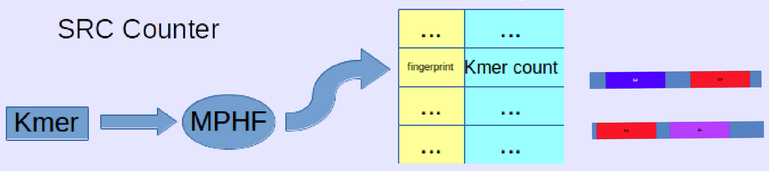
Short Read Connector tools

Given A and B sets of reads :

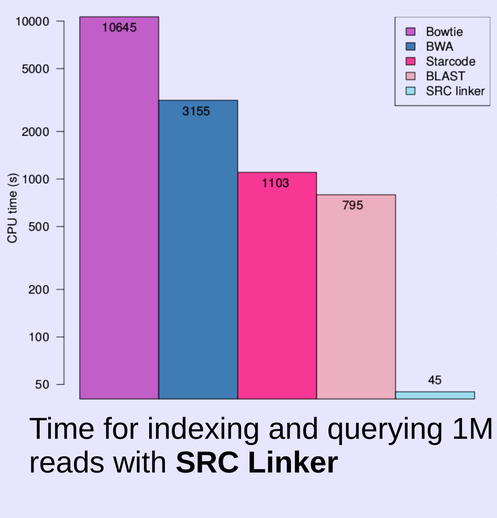
SRC Linker :
Output reads in A that has T Kmers that apear in set B.



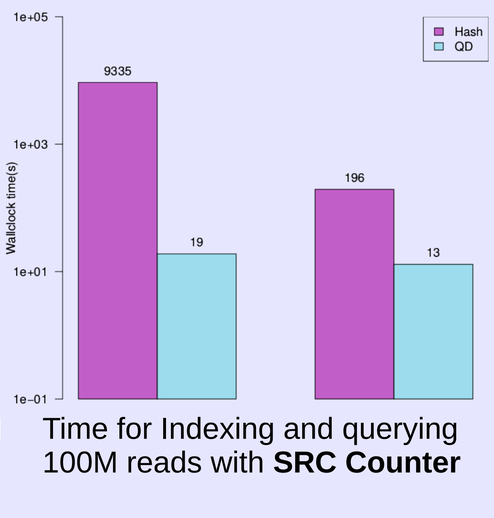
SRC Counter
Output reads in A that has T Kmers in common with the reads of B and estimate their coverage.



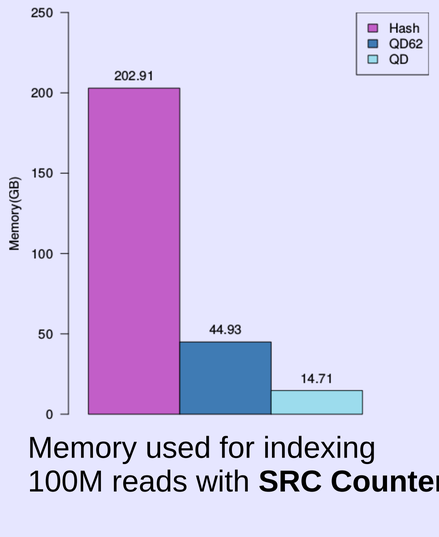
Results



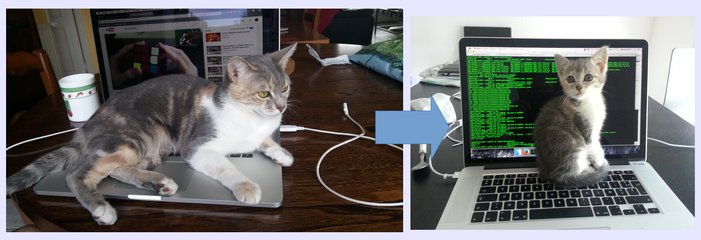
Time for indexing and querying 1M reads with **SRC Linker**



Time for Indexing and querying 100M reads with **SRC Counter**



Memory used for indexing 100M reads with **SRC Counter**



Less pressure on your machine !

Still have to assess the qualitative aspects of our methods ...

References

Bowtie2
Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." Nature methods 9.4 (2012): 357-359.

BWA
Li, Heng, and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform." Bioinformatics 25.14 (2009): 1754-1760.

Starcode
Zorita, Eduard, Pol Cuscó, and Guillaume J. Filion. "Starcode: sequence clustering based on all-pairs search." Bioinformatics 31.12 (2015): 1913-1919.

BLAST
Altschul, Stephen F., et al. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic acids research 25.17 (1997): 3389-3402.

SRC
http://arxiv.org/pdf/1605.08319.pdf

Links

Bbhash library :
github.com/rizkg/BBHash

Quasi-dictionary :
github.com/pierrepeterlongo/quasi_dictionary

Short Read Connector :
github.com/GATB/rconnector

QR code

Team