



HAL
open science

Minimal perfect hash functions in large scale bioinformatics Problem

Antoine Limasset, Camille Marchet, Pierre Peterlongo, Lucie Bittner

► **To cite this version:**

Antoine Limasset, Camille Marchet, Pierre Peterlongo, Lucie Bittner. Minimal perfect hash functions in large scale bioinformatics Problem. JOBIM 2016, Jun 2016, Lyon, France. hal-01341718

HAL Id: hal-01341718

<https://hal.science/hal-01341718>

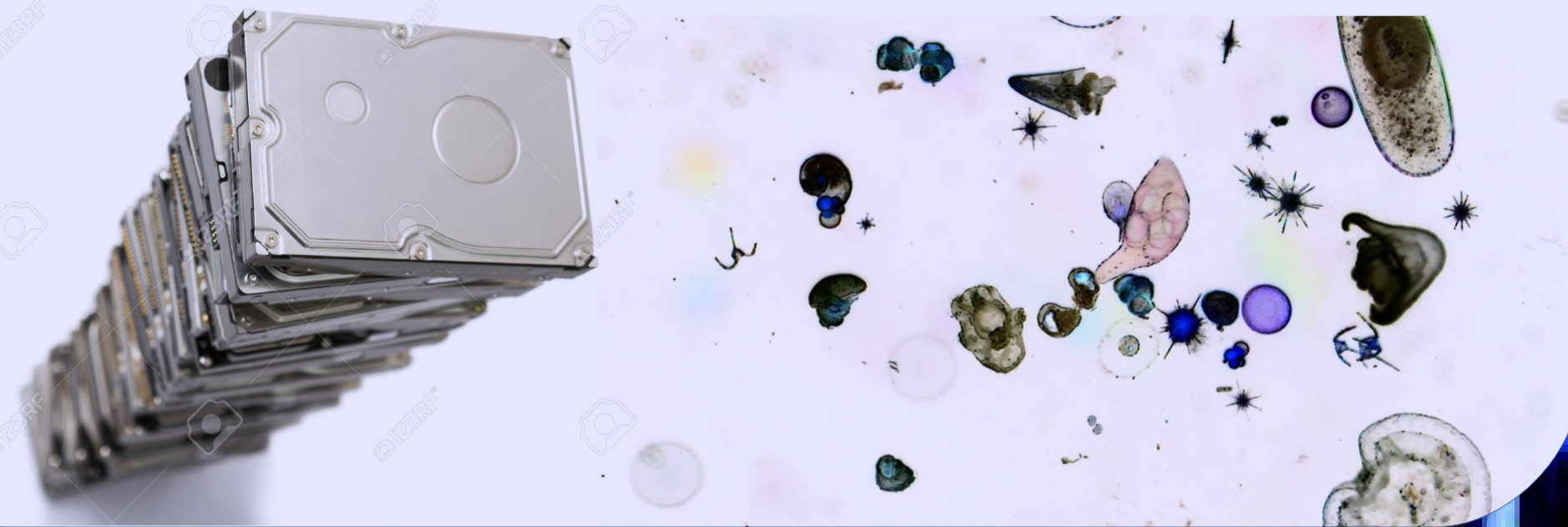
Submitted on 4 Jul 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

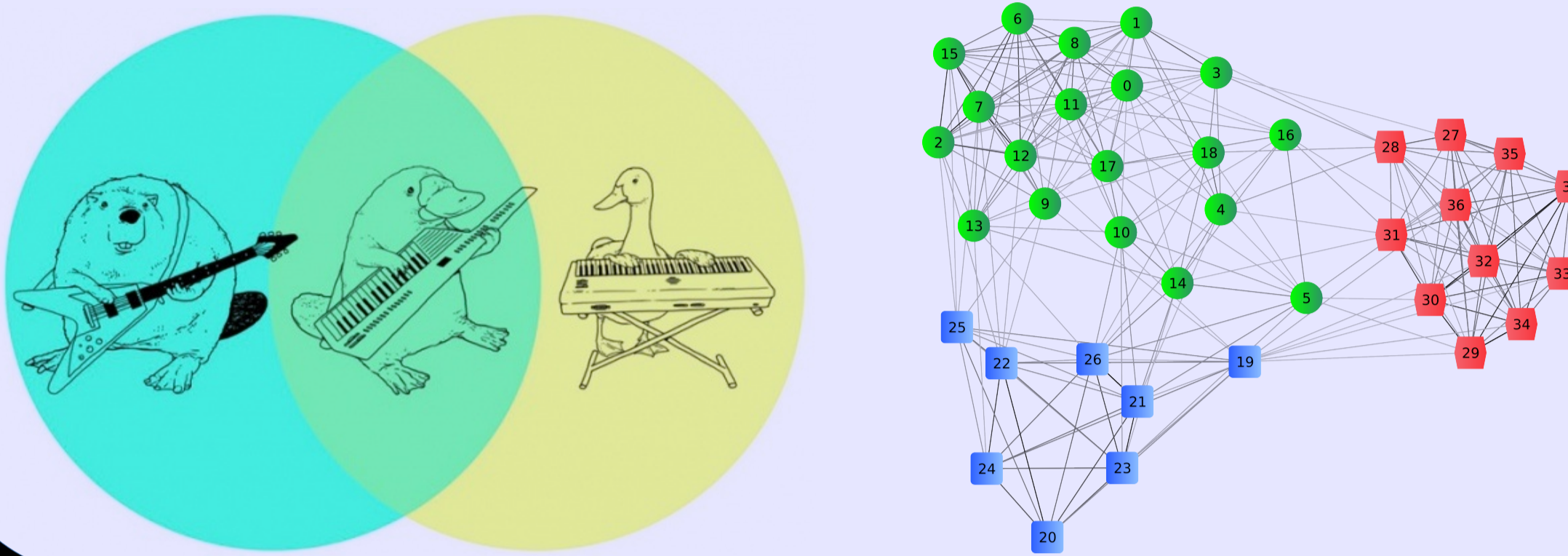
(Meta)Genomic Data

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



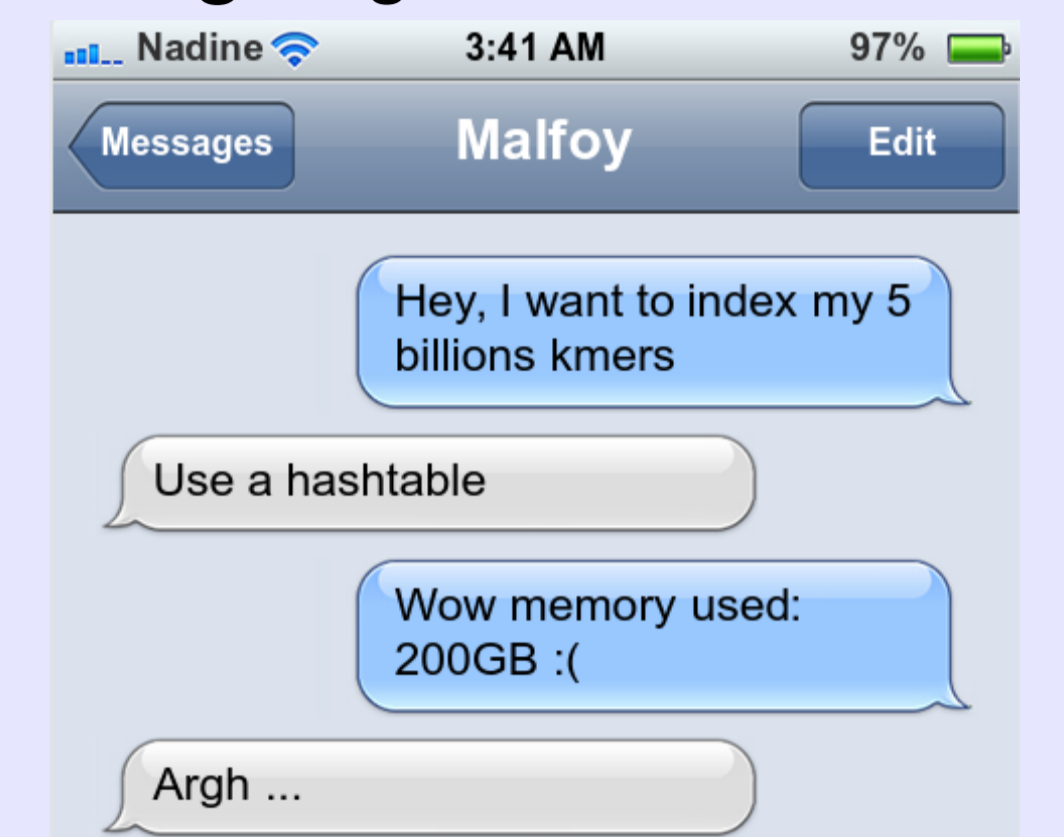
Questions

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



Problem

Indexing huge set of elements

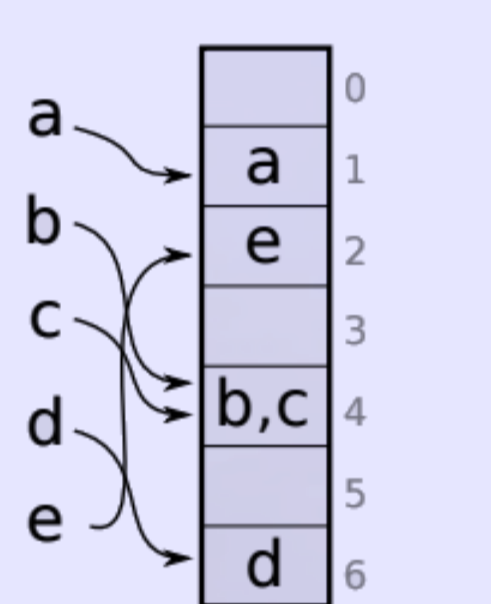


Hash functions

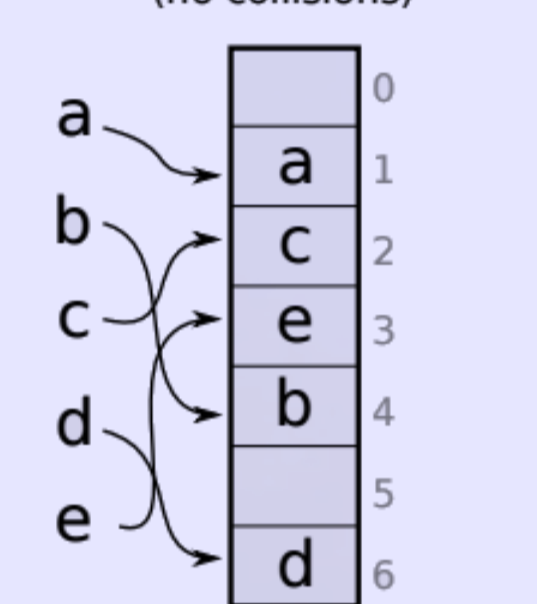
a,b,c,d,e : hashable elements (e.g. strings, integers, etc..)

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

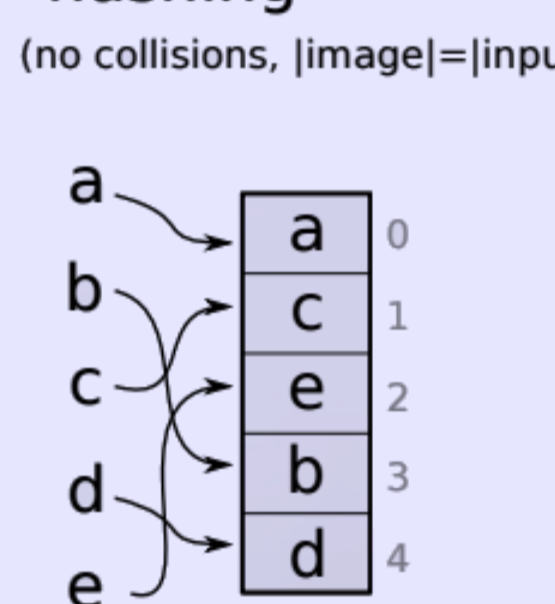
Classical hashing



Perfect hashing (no collisions)

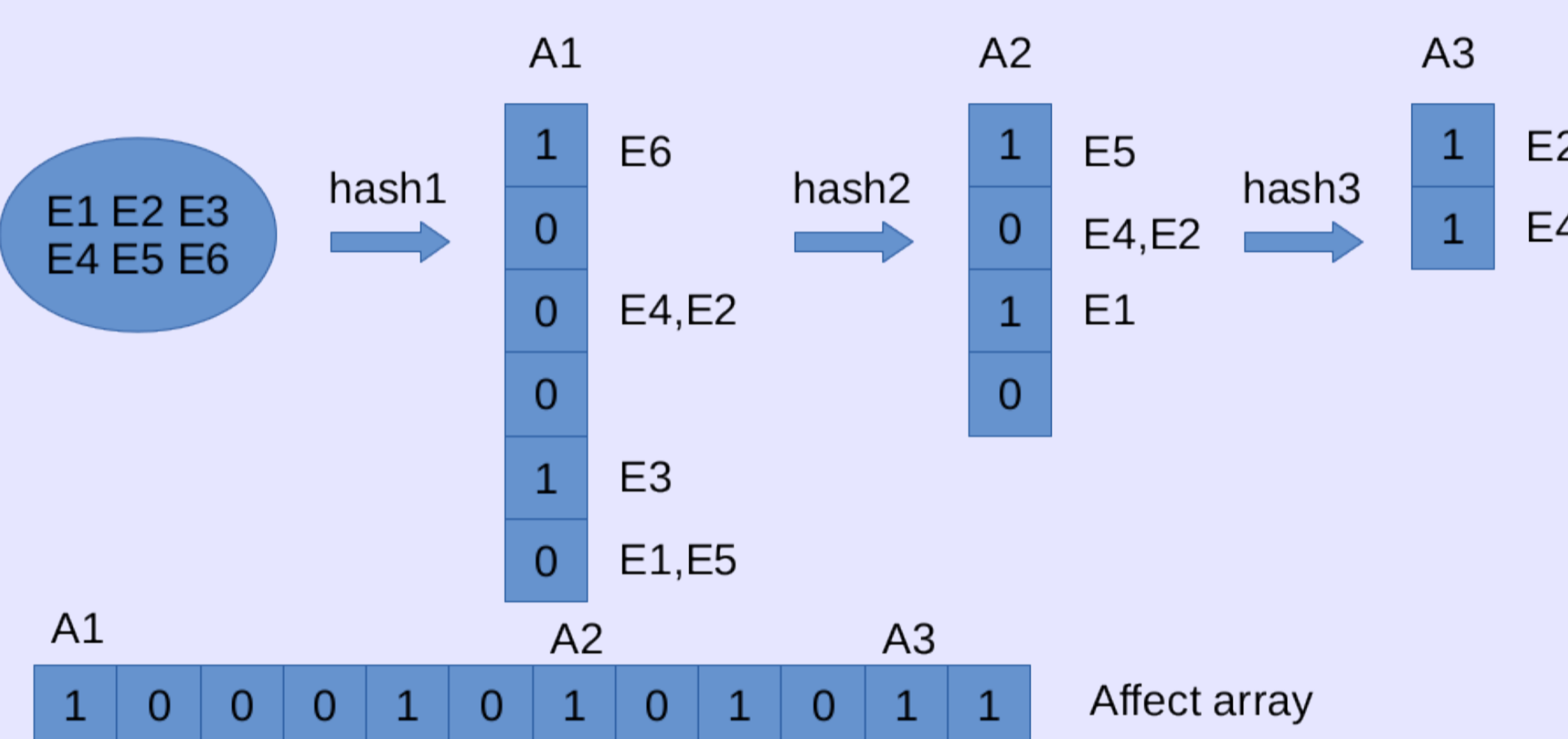


Minimal perfect hashing (no collisions, |image|=|input|)



BBhash library

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

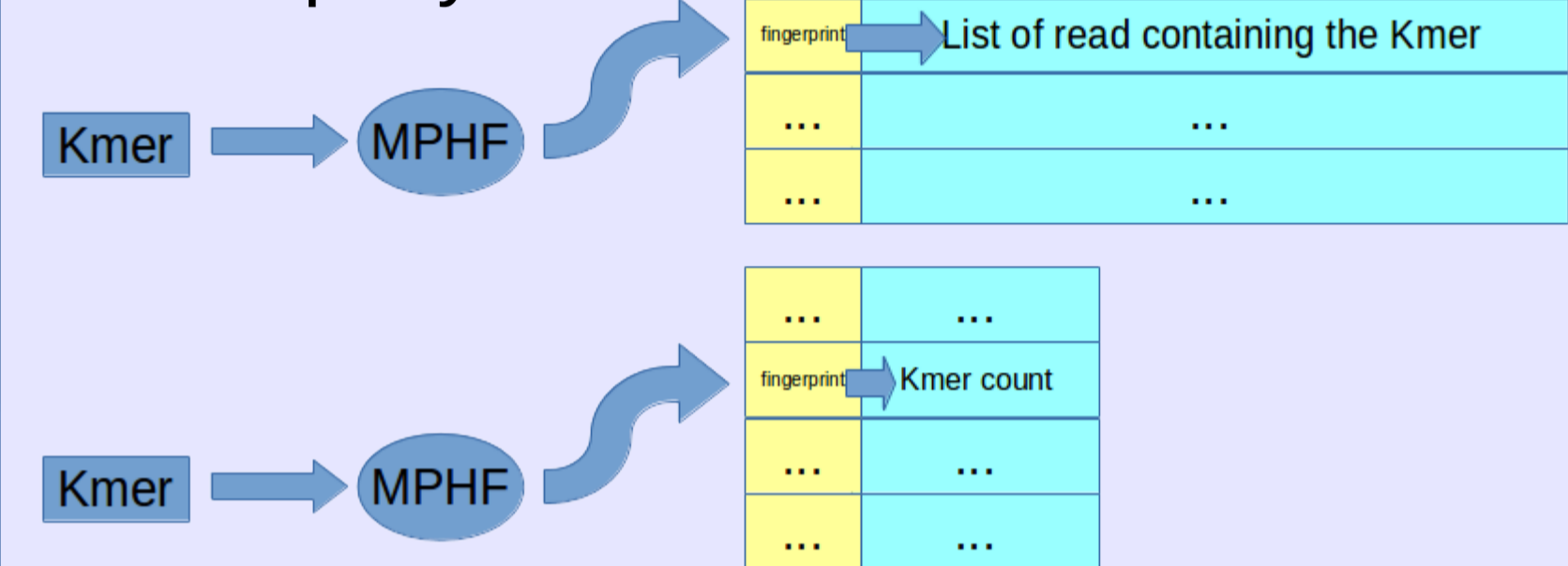


BUT

- 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 consumption :

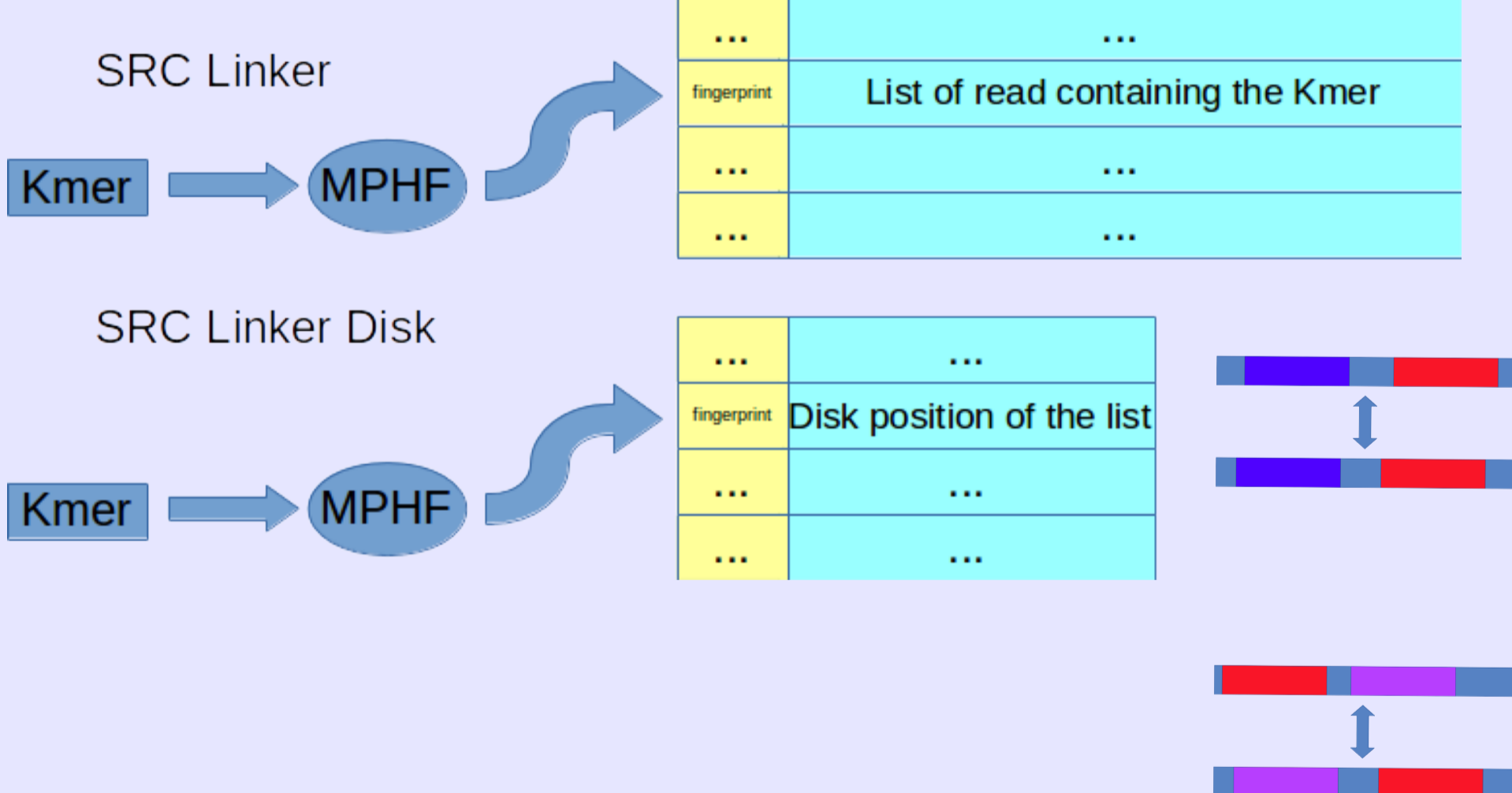
- Bit/elements : 10 FP rate: 1/10²
- Bit/elements : 20 FP rate: 1/10⁵
- Bit/elements : 30 FP rate: 1/10⁸

Short Read Connector tools

Given A and B sets of reads :

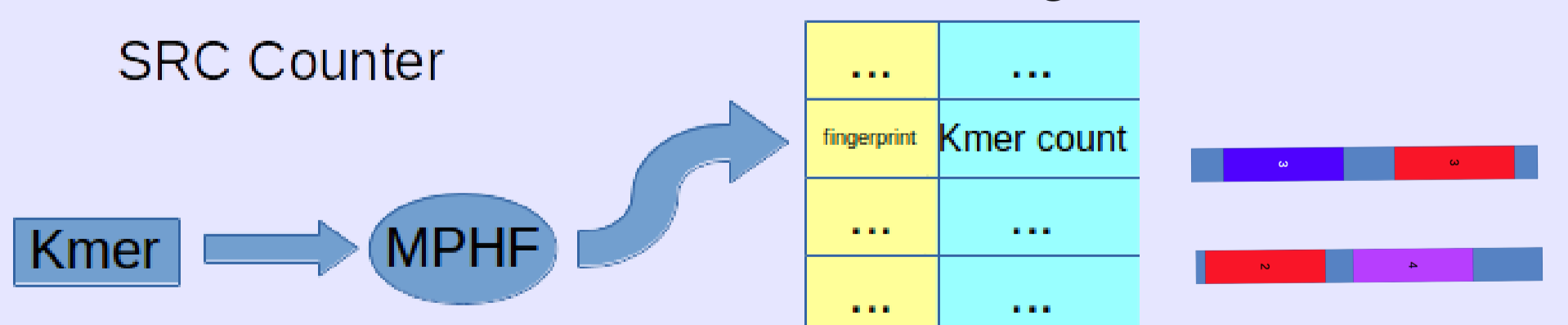
SRC Linker :

Output reads in A that has T Kmers that appear in set B.

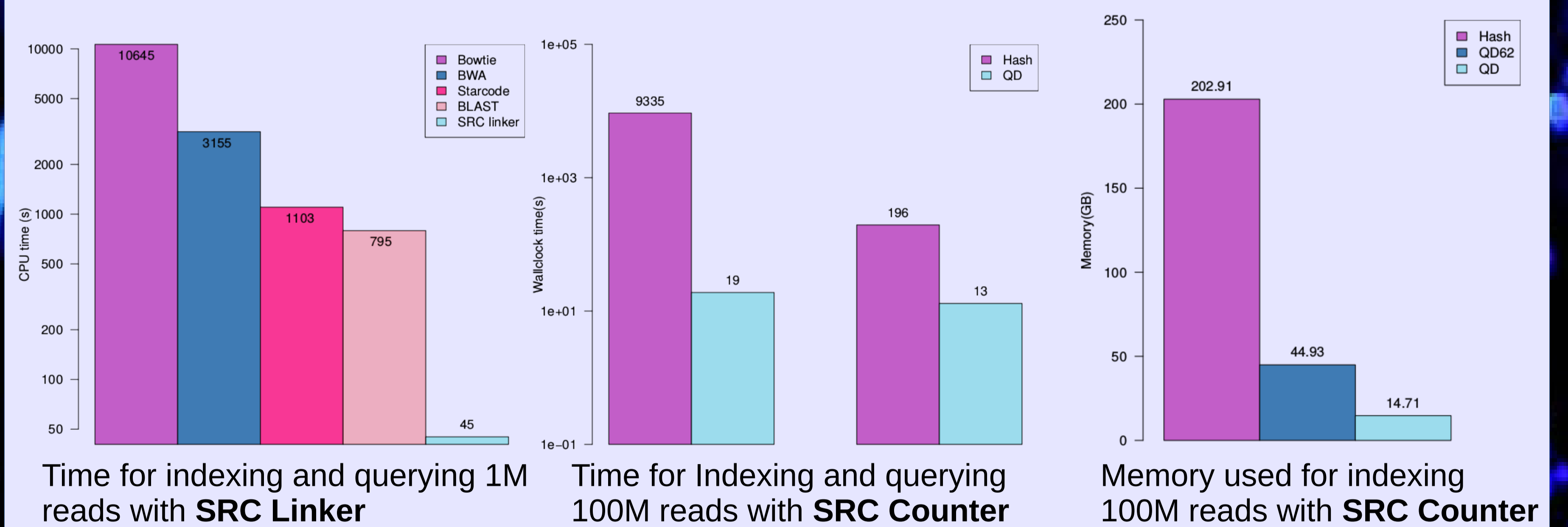


SRC Counter

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



Results



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**
Zotile, 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

