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
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, images, etc.)

Classical hashing
Perfect hashing (no collision)
Minimal perfect hashing (no collision, image/output)

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

Quasi-dictionary
Put a fingerprint in the value and check it at the query
False positive rate:
\[ \frac{2^{2k-\ell} - 1}{2^{2k}} \]
Memory consumption:
Bit/elements: 10 FP rate: 1/10^0
Bit/elements: 20 FP rate: 1/10^5
Bit/elements: 30 FP rate: 1/10^8
... 

Quasi-dictionary

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

Given A and B sets of reads:

Short Read Connector tools
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
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...

Links
Bbhash library: https://github.com/rizkg/BBHash
Quasi-dictionary: https://github.com/pierrepeterlongo/quasi_dictionary
Short Read Connector: https://github.com/GATB/rconnector

References
Bowtie

BWA

Starcode

BLAST

STAR

Links
 SRC Counter: https://github.com/GATB/rconnector

QR code
Team