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.archives-ouvertes.fr/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.
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

Problem
Quasi-dictionary
- False positive rate:
  - Bit/elements: 10 FP rate: 1/10^0
  - Bit/elements: 20 FP rate: 1/10^5
  - Bit/elements: 30 FP rate: 1/10^8
  ...

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

References
Bowtie2

BWA

Starcode

BLAST

SRC

Links
Bbhash library:
github.com/rizkg/BBHash

Quasi-dictionary:
github.com/pierrepeterlongo/quasi_dictionary

Short Read Connector:
github.com/GATB/rconnector

Results
Less pressure on your machine!

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