Abstract

SeqBIM

kmtricks: modular k-mer count matrix and Bloom filter construction for large read collections

Téo Lemane[1]*, Rayan Chikhi[2], Pierre Peterlongo[1]

[1] *Univ. Rennes, Inria, CNRS, IRISA, Rennes, France*
[2] *Institut Pasteur, CNRS, Paris, France*
***Corresponding author**: teo.lemane@inria.fr*

Abstract

The exponential growth of sequencing data repositories prompts the development of algorithms enabling to query those repositories with any sequence of interest (similarly to Internet search engines). Despite recent intensive developments (see [1] for a detailed review), even the latest tools cannot be used to screen across large collections of sequencing experiments. The fundamental need is a compact data-structure able to assign any queried k-mers, to the list of genomic file(s) (either sequencing experiment or assembled genomes) where this k-mer occurs.

In this context, we present a novel strategy to construct a one-hash Bloom filter which is the basic data structure involved in HowDe-SBT [2], one of the state-of-the-art k-mer indexer. Our method uses minimizers in order to partition and parallelize computations. It directly counts hash values (instead of k-mers) and outputs a matrix, in which each column can be seen as a one-hash Bloom filter corresponding to one data set.

This method improves the efficiency of Bloom filter construction in terms of time and memory footprint. The matrix structure also enables to leverage information across samples in order to recover some rare k-mers usually considered as errors. We will present the algorithmic foundations, current results, and possible future works on query improvements by taking advantage of the better data locality provided by the partitioned hash space.

References

[1] Camille Marchet, Christina Boucher, Simon Puglisi, Paul Medvedev, Mikaël Salson, and Rayan Chikhi. Data structures based on k-mers for querying large collections of sequencing datasets. *bioRxiv*, page 866756, dec 2019.

[2] Robert S Harris and Paul Medvedev. Improved representation of sequence Bloom trees. *Bioinformatics*, 2019.