Lightweight compositional analysis of metagenomes with sourmash gather

This manuscript ([permalink](#)) was automatically generated from [dib-lab/2020-paper-sourmash-gather@dfca6e9](#) on September 30, 2020.

Authors

- **Luiz Irber**
  ⓘD [0000-0003-4371-9659](#) · ⬡ [luizirber](#) · 🐦 [luizirber](#)
  Graduate Group in Computer Science, UC Davis; Department of Population Health and Reproduction, UC Davis · Funded by Grant XXXXXXXX

- **C. Titus Brown**
  ⓘD [0000-0001-6001-2677](#) · ⬡ [ctb](#)
  Department of Population Health and Reproduction, UC Davis

Abstract

Here we describe an extension of MinHash that permits accurate compositional analysis of metagenomes with low memory and disk requirements.

Results

Scaled MinHash accurately estimates containment

- scaled minhash supports similarity and containment
- compares well with others
- supports large-scale sketching of genbank

maybe split into two: definition, and then benchmarking.

second results section would be, "Scaled minhash has good performance…"

xx How much is missed figure; Poisson calculations?

Scaled MinHash sketches support efficient indexing for large-scale containment queries

Efficient indexing of scaled minhash signatures is cool.

- hierarchical and inverted indices (SBT and LCA)
- supports efficient containment and similarity queries

Metagenome sketches can be accurately decomposed into constituent genomes by a greedy algorithm, 'gather'

Greedy decomposition of metagenome sketches by k-mer containment is accurate

Greedy decomposition of metagenomes by k-mer containment (gather) is cool.

- outline algorithm
- compare conceptually vs least/lowest common ancestor approaches; combinatorial
- showcase some examples on synthetic data

Taxonomic profiling based on 'gather' is accurate

constituent gather is cool.

- CAMI results
- suggests gather/greedy decomposition is pretty good

Discussion

Scaled MinHash offers benefits, drawbacks vs regular MinHash

Combine theoretical discussion with practical discussion of benefits/drawbacks.

Gather works surprisingly well and matches simple data structures

gather is a straightforward algorithm.

easy to take advantage of other data structures b/c "just k-mers".

SBT, LCA implementations.

xx can we guess at places where gather would break?

Taxonomy results are excellent.

Discuss vs LCA.

reference the LCA-has-limits/k-mers saturate paper

mix and match taxonomies is easy b/c we anchor to genomes.

Algorithm is simple, computational performance is great

Performant implementation in sourmash

Database types work well

"online" approaches

Some limitations of gather and database types (equal results can be hard to detect efficiently with current SBT implementation)

Scaled minhash has limitations vs regular minhash

virus, etc. (could go in first discussion section, but also deserves to be highlighted)

References