Simka: fast kmer-based method for estimating the similarity between numerous metagenomic datasets

Gaëtan Benoit, Pierre Peterlongo, Dominique Lavenier, Claire Lemaitre

Simka: fast kmer-based method for estimating the similarity between numerous metagenomic datasets

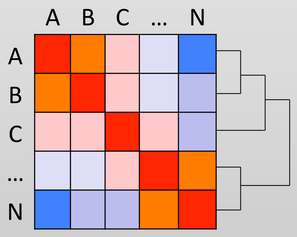
Gaëtan Benoit[1], Pierre Peterlongo[1], Dominique Lavenier[1], Claire Lemaitre[1]
[1] Inria/IRISA GenScale, Campus de Beaulieu, 35042 Rennes cedex, France.
gaetan.benoit@inria.fr, claire.lemaitre@inria.fr, pierre.peterlongo@inria.fr

Metagenomic

A liter of sea water:
- Hundred of millions of species
- > 90% unknown species

Comparative metagenomics

N metagenomic samples
N > 1000 in Tara Oceans project



Similarity heatmap N^2
with hierarchical clustering

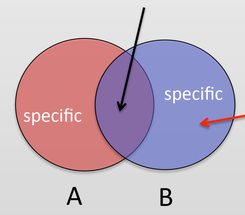
N read-sets
Usually **>100M** reads
per dataset

Similarity between 2 read-sets

Idea: similarity is given by the **size of the intersection**

Intersection = number of **similar reads**

Similar: their alignment score > 90%



100M reads

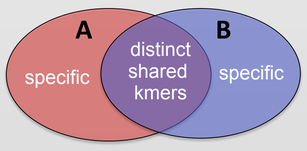
✖ Blast-like approach required **months** of computation for a single intersection

Commet [1] (state of the art)
- Heuristic: Two read are similar if they share some **kmers**
- Computes one intersection in **few hours**
- **Still does not scale on large metagenomic projects**
  - **N(N-1) intersections to compute**

Kmer-based similarity functions

- We have **very fast** methods for **indexing** and **querying** kmers
- A read-sets is now view as a **set of its kmers**
- We define new pairwize similarity measures based on **shared kmers**

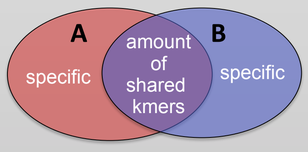
1) Presence / Absence of kmers



Similarity based on presence/absence of species

$$Jaccard(A,B) = \frac{DistinctKmers(A \cap B)}{DistinctKmers(A \cup B)}$$

2) Abundance of shared kmers



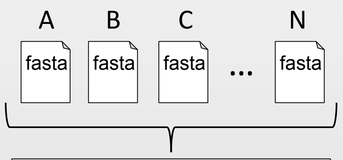
Similarity based on abundance of species

$$AbundanceJaccard(A,B) = \frac{\sum_{w \in A \cap B} N_A(w) + N_B(w)}{\sum_{w \in A \cup B} N_A(w) + N_B(w)}$$

w: kmer, *Ni(w)*: abundance of *w* in read-set *i*

Method

To scale on N large datasets, we count the kmers of N datasets simultaneously.



GATB[2] - DSK
multi-dataset kmer counting

Kmer's abundances ***v***

	A	B	C	...	N
ACGTAT	0	4	52	...	0

For each kmer, we get a vector of its abundances in N datasets

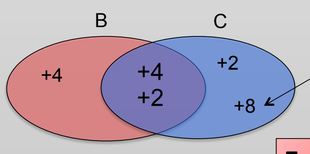
Is kmer solid ? (optional)

Yes

Update intersections

Kmer shared by B and C

	A	B	C
ACC	0	4	2



Kmer specific to C

	A	B	C
TGC	0	0	8

Execution time repartition:
- Counting kmers O(N): 75%
- Updating intersections **O(N²)**: **25%**

Estimating similarity is now a problem of counting kmers

Filtering sequencing errors

Kmer solid:
A kmer that contains no sequencing errors

Statistics:
80% of distinct kmers appear **only one time** and **only in a single dataset**

Noisy data:
It is hard to distinguish erroneous kmers from unique genomic kmers

We defined a solidity definition that can saved some unique kmers:
Given a solidity threshold ***t*** (***t***=2), a kmer ***w*** is solid if:

abundance(***w***) in at least one dataset >= ***t***

Examples on 4 datasets:

A	B	C	D
2	0	1	0

✓

A	B	C	D
52	0	1	1

✓

A	B	C	D
0	1	1	0

✗

Statistics on solid 31-mers from 21 Tara Oceans samples:
- Distinct: 18G (15% of all distinct kmers)
- Abundance: 106G (49% of all kmers)

Datasets

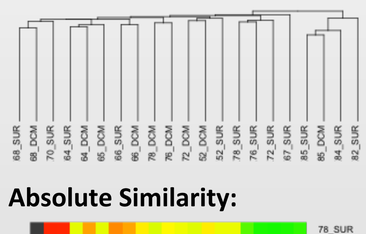
Simka was tested and compared to **Commet**[1] (state of the art) on **21 Tara Oceans samples**:
- 3G reads
- 400 GB of data
- 219G 31-mers
- 123G distinct 31-mers

Alignment based methods

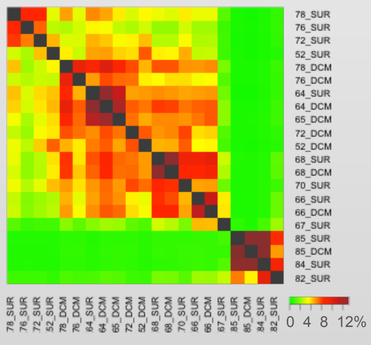
Years

Presence / Absence of kmers

Relative similarity:



Absolute Similarity:

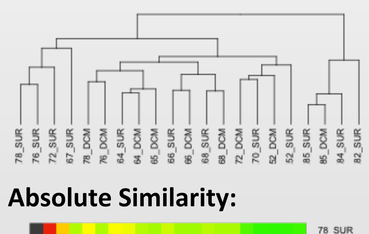


5 Hours

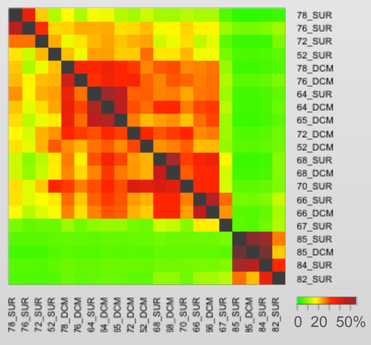
■ Can differs from Commet because not based on species's abundance
✓ Simka is the first method to provide quickly information about presence and absence of species

Abundance with error filter

Relative similarity:



Absolute Similarity:

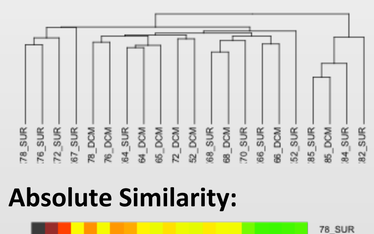


5 Hours

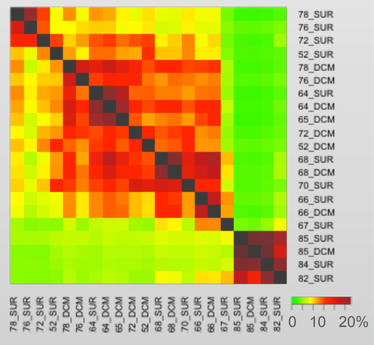
✓ Clustering less noisy and more readable
✓ Close to Commet while using only 18% of distinct kmers (half of the datasets)

Abundance of shared kmers

Relative similarity:



Absolute Similarity:

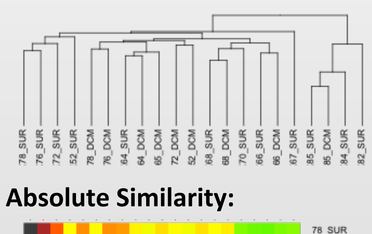


5 Hours

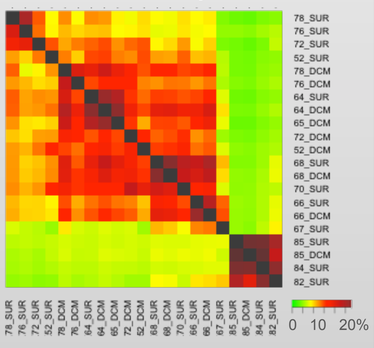
✓ Similarity and clustering close to read-based methods

Commet (state of the art)

Relative similarity:



Absolute Similarity:



Weeks

✓ Read-based
✗ Too slow on large metagenomic projects

Perspectives

- Filtering input reads
- Provide similarity confidence intervals with bootstrap
- Selecting discriminative kmers

Simka

☆ New similarity functions based on **shared kmers**
  ☆ Based on **abundance** and **presence/absence** of species
  ☆ Results close to read-based methods
☆ Fast and low memory thanks to the **GATB library [2]**

References

[1] COMMET: comparing and combining multiple metagenomic datasets
N. Maillet, G. Collet, T. Vannier, D. Lavenier, P. Peterlongo
IEEE BIBM, 2014

[2] GATB: Genome Assembly & Analysis Tool Box
E. Drezen, G. Rizk, R. Chikhi, C. Deltel, C. Lemaitre, P. Peterlongo, D. Lavenier
10.1093/*Bioinformatics*/btu406, 2014
https://gatb.inria.fr/

Funding: ANR Hydrogen, ANR-14-CE23-0001