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

Gaëtan BENOIT[1], Pierre PETERLONGO[1], Dominique LAVENIER[1] and Claire LEMAITRE[1]

[1] Inria Rennes Bretagne Atlantique, UMR 6074 Irisa, Genscale, 263 Avenue Général Leclerc, Campus de Beaulieu, 35042, Rennes, Cedex, France

Auteur à contacter : gaetan.benoit@inria.fr

## Abstract

Comparative metagenomics aims to provide high-level information based on DNA material sequenced from different environments. The purpose is mainly to estimate proximity between two or more environmental sites at the genomic level. One way to estimate similarity is to count the number of similar DNA fragments. From a computational point of view, the problem is thus to calculate the intersections between datasets of reads. Resorting to traditional methods such as all-versus-all sequence alignment is not possible on current metagenomic projects. For instance, the Tara Oceans project involves hundreds of datasets of more than 100M reads each.

Maillet et al. defined the following heuristic in their method called Commet[1]. Two reads are considered similar if they share t non-overlapping kmers (words of length k). This method is currently the fastest but still does not scale on Tara Oceans samples.

To tackle this issue, we introduce a new method, called Simka[2], which computes the similarity between two datasets based on their shared kmers. To scale on large metagenomic projects, we use the GATB library[3] which provides a kmer counting tool able to count the kmers of N datasets simultaneously. Counting kmers also offers new possibilities such as filtering low frequency kmers, which potentially contain sequencing errors. Simka also provides efficiently new similarity functions. The first is based on Bray Curtis, a well-know similarity function in ecology, which informed about species abundance. The second computes the Jaccard similarity between the datasets and thus informed about presence and absence of species.

Simka was tested and compared to the state of the art on 21 Tara Oceans samples. This shows that our kmer-based similarity function is very close to the read-based ones. Regarding sample proximity, different methods identify the same clusters of datasets. The fastest method of the state of the art required a few weeks to compute all the intersections whereas Simka took only 4 hours.

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

**[2] Simka: fast kmer-based method for estimating the similarity between numerous metagenomic datasets.** G. Benoit, P. Peterlongo, D. Lavenier, C. Lemaitre. Hal – Inria, 2015

**[3] 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