A Fast alignment-free whole-genome distance calculator with applications in post-assembly bin validation and annotation of Metagenomic samples

Gleb Goussarov [1,2], Ilse Cleenwerck [2], Mohamed Mysara [1], Natalie Leys [1], Pieter Monsieurs [1,3], Guillaume Tahon[2], Aurélien Carlier[2], Peter Vandamme[2] and Rob Van Houdt [1]

[1] *Microbiology Unit, Belgian Nuclear Research Centre (SCK•CEN), Mol, 2400, Belgium*

[2] *Laboratory of Microbiology and BCCM/LMG Bacteria Collection, Faculty of Sciences, Ghent University, Ghent, 9000, Belgium.*

[3] *Unit Health, Flemish Institute for Technological Research (VITO), Mol, 2400, Belgium.*

E-mail: ***ggoussar@sckcen.be***

Abstract

In the last decade, improvements in sequencing technology have enabled affordable whole genome and metagenome sequencing. In cases where previously unsequenced organisms are contained with the sample, alignment-based methods can fail to provide satisfactory results for both assembly and annotation. In order to address this, we are developing a pipeline focused on de-novo Metagenome assembly. In this pipeline, we aim to reconstruct the genomes contained with a sample based on their bacterial type. Here, we present a method which is able to conclude whether two whole genome genomes belong to the same bacterial species and type. Our alignment-free methods relies on oligonucleotide counting and has shown excellent performance for bacterial identification on a dataset of individually assembled genomes, demonstrating its ability to quickly compare thousands of genomes to each other. This method could be applied near the end of a metagenome assembly pipeline as way to both validate and guide contig binning.