Abstract

Identification of bacterial strains using Oxford Nanopore sequencing

Grégoire Siekaniec[1,2]*

[1]*Univ Rennes, Inria, CNRS, IRISA, Rennes F-35000, France*
[2] *STLO, INRAE, Agrocampus Ouest, Rennes, France*
***Corresponding author**: gregoire.siekaniec@inria.fr

Abstract

The bacterial taxonomic assignation from sequencing data is usually based on few ubiquitous genes (RNA16S, ITS, MLST). However, due to the close genomic proximity of strains of a same species, this approach does not allow to distinguish them. Thanks to the reduction in sequencing costs, it is now possible to consider routine sequencing and identification based on whole bacterial genomes. Most current taxonomic assignation software based on whole genome only support short reads data. In contrast, our project is based on the Oxford Nanopore technology (MinION device) generating long DNA sequences. Using ONT means tackle with high error level (about 6% on raw uncorrected reads). Main existing software dealing with long reads stop at the species level, which are very efficient but do not fully exploit the potential of long reads. We want to show using the *Streptococcus thermophilus* species, that consideration of the whole genome combined with the use of long reads generally enables rapid distinction between one strain and another.

Our method is based on an efficient storing of the known genome sequences of strains. We have chosen spaced seeds [1] for this task, which are more sensitive than standard kmer indexes. Spaced seeds are kmers with defined positions that are not taken into account during matching. Then, with the help of Téo Lemane (PhD student in our lab) we extended to spaced seeds the HowDeSBT structure explained and implemented in [2] in order to obtain a small index allowing to store the genomes. Once the index is built, the second issue is the query part that assigns reads to strains. First, reads are selected based on a quality filter, to limit the error rate. Then, the data structure is requested with the seed matches extracted from the reads, each read is assigned to its potential strains and a presence/absence matrix (strains x reads) is created. Finally, the identification step was performed as an optimization issue, by looking for the minimum number of strains that can explain all the reads from the strains x reads matrix. Our implementation used the answer set programming (ASP) framework.

References

[1] Laurent Noé, Best hits of 11110110111: model-free selection and parameter-free sensitivity calculation of spaced seeds, Algorithms for Molecular Biology, volume 12, issue 1, 2017.

[2] Robert S. Harris, and Paul Medvedev. Improved representation of sequence Bloom trees. Bioinformatics, volume 36, issue 3, pages 721–727, February 2020, btz662