Bacterial strains identification using Oxford Nanopore sequencing

Grégoire Siekaniec, Emeline Roux, Eric Guédon, Jacques Nicolas

Bacterial strains identification using Oxford Nanopore sequencing

Grégoire Siekaniec[1,2], Emeline Roux[1,2,3], Eric Guedon[2] and Jacques Nicolas[1]

[1] Univ Rennes, Inria, CNRS, IRISA, Rennes F-35000, France
[2] STLO, INRAE, Agrocampus Ouest, Rennes, France
[3] CALBINOTOX, Université de Lorraine, F-54000, Nancy, France

Corresponding Author: gregoire.siekaniec@inria.fr

1 Introduction

The bacterial taxonomic assignation from sequencing data is usually based on few ubiquitous genes (RNA16S, ITS, MLST). However, due to the close proximity at the genomic level of bacterial strains of a same species, these conventional techniques do not allow the strains to be distinguished from one another. 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 inputs. In contrast, our project is based on the Oxford Nanopore technology (MinION device) generating long DNA sequences.

The challenge is to tackle with a relatively high error level (about 7% on raw uncorrected reads) and show that whole genome data and long reads allow to quickly distinguish one species from another and even to go down to the strain level. The few existing software dealing with long reads stop at the species level probably due to the use of too short signatures (such as minimizers in Kraken [1]), which are very efficient but do not fully exploit the potential of long reads. The identification of complex mixture of different bacteria (metagenomic samples) poses the additional problem of separating the fragments specific to each microorganism.

2 Methods

The first problem is to store efficiently the known genome sequences within a structure that allows to retrieve the possible origins of a given sequenced fragment with errors. We have chosen spaced seeds [2] for this task, which are more sensitive than standard kmer indexes. The spaced seeds are kmers with fixed length "gaps", which are defined positions that are not taken into account during matching. Spaced seeds are thus suitable for handling mismatch errors but not sufficient for indel errors,which are frequent in long reads. For this reason, we also experiment with an extension of spaced seed, the indel seed [3]. To obtain a small index allowing to store a large quantity of genomes, the structure of the Bloom filter tree explained and implemented in [4] was employed and modified to use spaced and indel seeds instead of kmers.

Once the index built, the second issue is the query part that assigns reads to species/strains by the following steps : 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 being assigned to a taxonomy level with respect to a majority vote if one is interested by the identification of the main species in the sample. The read may be assigned to several strains with respect to a threshold of recognized seeds if one studies a mixture of bacteria and is interested by their relative abundance.

References

[1] Derrick E. Wood, and Steven L. Stalzberg. *Kraken: ultrafast metagenomic sequence classification using exact alignments*. Genome Biology, volume 15, 2014, 10.1186/gb-2014-15-3-r46.

[2] 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.

[3] Denise Mak, Yevgeniy Gelfand and, Gary Benson. *Indel seeds for homology search.* Bioinformatics, volume 22, issues 14, pages e341-e349, July 2006, btl263.

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