Kraken: ultrafast metagenomic sequence classification using exact alignments

Mohammadamin Edrisi (mae6)

February 7, 2019

Scientific question

The scientific for this paper is focused on *making a program, termed MetaPhlAn, that can both identify and utilize marker genes to measure relative abundance in a metagenomic sample.* Because of the importance of understanding the composition of a metagenomic sample to study disorders and the environment, many have undertaken different techniques to understand the relative abundance of a metagenomic sample. Namely, laboratory techniques that use 16S ribosomal RNA have been implemented, which do fairly well at characterizing the relative abundance of a sample with little information. There have also been other computational made to accomplish this goal, however most of these techniques contain costly time complexities.

Hypothesis tested

The hypothesis tested in this research is that *MetaPhlAn is able to uniquely identify multiple species within a metagenomic sample.* Their method for testing this hypothesis consists of comparing their program to other programs, and running their program on simulated and real, yet defined, metagenomic samples.

Methods

The methods section for this paper is quite extensive, taking up the later half of the paper. There are three inputs that go into MetaPhlAn to create the marker gene database: (1) The raw nucleotide sequences; (2) the CDS calls; and lastly, (3) the taxonomic classification of the genomes. Per species/genome, all of the CDS sequences are clustered together, then the representative seeds from the clustering, for all of the genomes, are clustered together. This is repeatedly performed, while climbing up the taxonomic tree, to identify signature sequences for all of the taxonomic ranks. The idea is to not include signature sequences from the parent clade, thus defining the idea of a "core gene". Next these core genes are further tested for uniqueness, thus getting rid of all multicopy genes (even like that of 16S rRNA- which is a multicopy gene).

Results

Kraken has been compared with four other methods existing at the time of publication, Megablast, PhymmBL, NBC, and MetaPhIAn. PhymmBL and NBC classify all sequences as accurately as possible, while Kraken and Megablast leave some sequences unclassified. MetaPhIAn only classifies a subset of reads that map to one of its marker genes; this is why the authors did not use MetaPhlAn in classification accuracy measurement and used its results to measure only the classification speed. Three database have been simulated for comparing the classification accurary and speed. Two measurement criteria are used here, precision and sensitivity. Precision refers to the proportion of correct classifications, out of the total number of classifications attempted. Sensitivity is the proportion of sequences assigned to the correct genus.

Key implications of the results

The genus-level precision of Kraken is the highest or among the highest precisions according to the results of simulated metagenomes, while it sensitivity is lower than those of the others. This shows that the use of exact k-mers yields to a higher precision for Kraken. The nonselective classifiers (PhymmBL and NBC) were able to achieve higher sensitivity than the selective ones (Kraken and Megablast) at a cost of a significantly lower precision.
The main motivation of the development of this method was to improve the run-time of the classification. Testing the methods on the simulated datasets showed that Kraken classifies much faster than ant other classifier. The only comparable classifier in terms of speed is MegaPhIAn which does not assign taxonomic labels to all the sequences.