Improving protein-based metagenomic reads classification

Giovanni Marques de Castro, Francisco Pereira Lobo

Universidade Federal de Minas Gerais

Abstract

The most used software for the classification of metagenomic reads have the option for the user to build a custom database, which is expected to be updated and encompassing as much taxa as possible. The proposal of KAIJU is that amino-acids are more conserved and using them increase the sensibility, especially from samples of extreme environments. However, the database proposed by them that includes the widest range of taxa is the nr_euk. For metagenomic studies that focuses on fungal organims, the results obtained can be much worse than using a nucleotide search, missing most of the reads and having an increased misclassification. This problem arises because fungal genomes are deposited in GenBank without their protein annotation, this is reflected in the source database used by KAIJU, the NR, not containing the peptides from those genomes. After downloading 2027 fungal genomes from GenBank that passed some filters, only 794 had their protein file available to download (*protein.faa.gz in the GenBank ftp of the genome). This means that 60% of the fungal genomes do not have their proteomes available. To generate a protein database that is broad and includes information from those fungal genomes, genomes from bacteria, viruses, archaea and from those downloaded fungal genomes were translated in their six frames. Due to RAM limitation, only peptide sequences with over 60 amino acids were kept and indexed using KAIJU. This is an strategy similar to tblastx, but much faster as it uses KAIJU. In a preliminar result with a metagenomic sample, CENTRIFUGE, a nucleotide based classifier, using a database with the same downloaded genomes, classified over 600.000 reads as Lasiodiplodia. On the other hand, KAIJU using the nr_euk classified only a bit over 5000 reads as Lasiodiplodia and over 150.000 reads as Diplodia, both Fungi in the Botryosphaeriaceae family. Nonetheless, when using the database of six frames translated genomes, KAIJU was able to classify around 600.000 reads as Lasiodiplodia, while almost no reads as Diplodia, close to the result returned from CENTRIFUGE. Lasiodiplodia is an example of a gender of Fungi without a single predicted proteome, moreover the single genome available have over 97% completeness when analyzing its quality with BUSCO, so it should be well assembled by this parameter. With this strategy to build a protein database, not only KAIJU, but other protein-based classifiers could use it, as the NR lacks over half of the potential proteins from Fungi in which the genomic information is already public available.

Funding: