Abstract

ProPhyle – a memory efficient BWT-based metagenomic classifier using k-mer propagation

Karel Břinda[1]*, Kamil Salikhov[1,2], Simone Pignotti[1], and Gregory Kucherov[1]

[1]*LIGM/CNRS Université Paris-Est, France*
[2]*Mechanics and Mathematics Department, Lomonosov Moscow State University, Russia*
***Corresponding author**: karel.brinda@univ-mlv.fr

Abstract

Metagenomics is a powerful approach to study genetic content of environmental samples that has been strongly promoted by NGS technologies. A way to improve the accuracy of metagenomic classification is to match the metagenome against a largest possible set of known genomic sequences. With many thousands of completed microbial genomes available today, modern metagenomic projects match their samples against genomic databases of tens of billions of bp.

To cope with increasingly large metagenomic projects, alignment-free methods have recently come into use [1, 2, 3, 4]. The most popular tool – Kraken [1] – performs metagenomic classification of NGS reads based on the analysis of shared k-mers between an input read and each genome from a pre-compiled database. Given a taxonomic tree involving the species of the reference database, Kraken "maps" each read to a node of the tree corresponding to the most specific taxon or clade for that read. Mapping is done by sliding through all k-mers of the read and determining, for each of them, the genomes of the database containing the k-mer. Based on the obtained counts and the tree topology, the algorithm assigns the read to the tree node "best explaining" the counts. Kraken provides an extremely rapid read classification, but its index suffers from two major limitations. First, Kraken's enormous memory consumption, due to a large hash table, does not allow one to perform classification other than on high-performance clusters. Second, each k-mer in the index is represented through its lowest common ancestor, which can result in an inaccurate classification.

We present Prophyle, a metagenomic classifier based on BWT-index. ProPhyle uses a classification algorithm similar to Kraken, but with an indexing strategy based on a bottom-up propagation of k-mers in the tree, assembling contigs at each node and matching using a standard full-text search. The obtained index occupies only a fraction of RAM compared to Kraken – 13 GB instead of 120 GB for index construction and 14 GB instead of 75 GB for index querying. The resulting index is also more expressive as it can, for every queried k-mer, retrieve a list of *all* genomes in which the k-mer occurs.

References

[1] Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46, 2014.

[2] Sasha K. Ames, David a. Hysom, Shea N. Gardner, G. Scott Lloyd, Maya B. Gokhale, and Jonathan E. Allen. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics*, 29(18):2253–2260, 2013.

[3] Jolanta Kawulok and Sebastian Deorowicz. CoMeta: Classification of Metagenomes Using k-mers. *PLOS ONE*, 10(4):e0121453, 2015.

[4] Daehwan Kim, Li Song, Florian P Breitwieser, and Steven L Salzberg. Centrifuge: rapid and sensitive classification of metagenomic sequences. *bioRxiv preprints*, 2016.