Improved computational techniques for k-mer-based metagenomic classification

Karel Břinda Maciej Sykulski Gregory Kucherov

Laboratoire d'Informatique Gaspard-Monge, Université Paris-Est&CNRS, France

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 as large set of known genomic sequences as possible. With many thousands of completed microbial genomes available today, modern metagenomic projects match their samples against genomic databases of tens of billions of bp [12].

Alignment-based classifiers [9] proceed by aligning metagenome sequences to each of the known genomes from the reference database, in order to use the best alignment score as an estimator of the phylogenetic "closeness" between the sequence and the genome. While this approach can be envisaged for small datasets (both metagenome and database) and is actually used in such software tools as MEGAN [4] or PHYMMBL [2] (see [9] for more), it is unfeasible on the scale of modern metagenomic projects. On the other hand, there exists a multitude of specialized tools for aligning NGS reads – BWA [7], NOVOALIGN (`http://www.novocraft.com/`), GEM [10], BOWTIE [6], just to mention a few popular ones – which perform alignment at a higher speed and are adjusted to specificities of NGS-produced sequences. Still, aligning multimillion read sets against thousands of genomes remains computationally difficult even with optimized ¡tools. Furthermore, read alignment algorithms are usually designed to compute high-scoring alignments only, and are often unable to report low-quality alignments. As a result, a large fraction of reads may remain unmapped [8].

To cope with increasingly large metagenomic projects, *alignment-free methods* have recently come into use. Those methods do not compute read alignments, thus do not come with benefits of these, such as gene identification. Two recently released tools – LMAT [1] and KRAKEN [12] – perform 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 database, those tools "map" each read to a node of the tree, thus reporting the most specific taxon or clade that the read gets associated with. Mapping is done by sliding through all k-mers occurring in the read and determining, for each of them, the genomes of the database containing the k-mer. Based on obtained counts and tree topology, algorithms [1, 12] assign the read to the tree node "best explaining" the counts. Further similar tools have been published during last months [11, 5].

In this work, we report on a computational improvement of methods [1, 12]. One source of improvement comes from using *spaced k-mers* rather than contiguous k-mers. Through a series a computational experiments, we show that this can significantly increase the accuracy of metagenomic classification of NGS reads [3]. In particular, we illustrate this by a series of large-scale metagenomic classification experiments with modified KRAKEN software [12] extended by the possibility of dealing with spaced seeds. Experiments have been performed on databases of size 3.3Gb to 4.1Gb and metagenomes (both simulated and real) of 10,000 to 50,000 reads.

We also present some other computational improvements, in particular a new indexing structure for the reference database: *tree of Bloom filters*. This data structure is currently being implemented in a new software tool, and we report on its developement.

References

[1] S. K. Ames, D. A. Hysom, et al. Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics*, 29(18):2253–2260, Sep 2013.

[2] A. Brady and S. Salzberg. PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat. Methods*, 8(5):367, May 2011.

[3] K. Brinda, M. Sykulski, and G. Kucherov. Spaced seeds improve k-mer-based metagenomic classification. *Bioinformatics*, July 2015. 10.1093/bioinformatics/btv419.

[4] D. H. Huson, S. Mitra, et al. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.*, 21(9):1552–1560, Sep 2011.

[5] J. Kawulok and S. Deorowicz. CoMeta: Classification of Metagenomes Using k-mers. *PLoS ONE*, 10(4):e0121453, 2015.

[6] B. Langmead, C. Trapnell, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

[7] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[8] Stinus Lindgreen, Karen L Adair, and Paul Gardner. An evaluation of the accuracy and speed of metagenome analysis tools. *bioRxiv*, 2015.

[9] S. S. Mande, M. H. Mohammed, et al. Classification of metagenomic sequences: methods and challenges. *Brief. Bioinformatics*, 13(6):669–681, Nov 2012.

[10] S. Marco-Sola, M. Sammeth, et al. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, 9(12):1185–1188, Dec 2012.

[11] R. Ounit, S. Wanamaker, T. J. Close, and S. Lonardi. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, 16:236, 2015.

[12] D. E. Wood and S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.*, 15(3):R46, 2014.