

SortMeRNA 2: ribosomal RNA classification for taxonomic assignation

E. Kopylova, L. Noé, P. Pericard, M. Salson, H. Touzet
LIFL (CNRS, University of Lille) and Inria Lille

The application of next-generation sequencing technologies to RNA or DNA directly extracted from a community of organisms yields a mixture of nucleotide fragments. The task to distinguish amongst these and to further categorize the families of ribosomal RNAs is an important step for examining the phylogenetic classification of the constituting species.

In this perspective, we have developed SortMeRNA, a software designed to filter reads from metatranscriptomic or metagenomic data. It is capable of handling large sets of reads and sorting out all fragments matching to a database of annotated ribosomal RNA sequences from the three-domain system with high sensitivity and a low running time. The first version of SortMeRNA was released in 2012, and published in [1]. It has identified users in multiple research laboratories worldwide, and has been integrated into two computational pipelines [2, 3].

We now make available a new version, SortMeRNA2, that brings several helpful improvements. Most importantly, it can now perform alignments with sequences from the ribosomal RNA database, which allows the user to study the taxonomic content of the microbial sample. For that, we have developed an alignment strategy based on approximate seeds and seed extension with a Longest Increasing Subsequence. SortMeRNA2 also applies statistical analysis to evaluate the significance of an alignment, based on the E-value, which confers a great accuracy to the program. Currently, sequencing platforms such as Illumina, 454, Ion Torrent and PacBio are all supported and the results are output into SAM and Blast-like human readable formats.

SortMeRNA2 is available for download from our official website: <http://bioinfo.lifl.fr/RNA/sortmerna/>. Ongoing development of SortMeRNA2 will provide support for amplicon sequencing and is under integration to the QIIME pipeline (Quantitative Insights Into Microbial Ecology, [4]).

References

- [1] SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Kopylova E., Noé L. and Touzet H., *Bioinformatics* (2012), doi: 10.1093/bioinformatics/bts611
- [2] A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. M. Leimena et al., *BMC genomics* (2013)
- [3] Metagenome survey of a multispecies and algae-associated biofilm reveals key elements of bacterial-algae interactions in photobioreactors. I. Krohn-Molt et al. *Applied and environmental microbiology* (2013)
- [4] QIIME allows analysis of high-throughput community sequencing data. J. G. Caporaso et al. *Nature Methods* (2010)