EVALUATION AND COMPARISON OF DIFFERENT TAXONOMIC CLASSIFICATION APPROACHES OF METAGENOMIC SEQUENCES

Dmitrij Turaev, Michael Sonntag, Thomas Rattei
Department of Computational Systems Biology, University of Vienna, Althanstr. 14, 1090 Vienna, Austria

Background. Metagenomics is the sequencing and analysis of genetic material directly from environmental samples. It allowed for unbiased studies of unculturable bacterial species – which are thought to constitute the overwhelming majority of bacteria [1] – while preserving the natural community composition and the environmental context. This revolutionized disciplines like the environmental and the medical microbiology and provided fascinating insights into nature [2]. New faster and cheaper sequencing technologies like 454 pyrosequencing, Illumina (Solexa) sequencing and SOLiD sequencing (reviewed in [3]) made deep sequencing performable and granted access even to low abundancy species.
A crucial part of the data analysis is the taxonomic classification of metagenomic sequences which do not contain taxonomic marker genes. Three sources of information can be exploited to achieve this: sequence similarity information, compositional peculiarities of the nucleotide sequence and phylogenetic relationships between sequences. The obtained taxonomic profile of a community can serve as a fingerprint for community or habitat comparison and provides information about the ecological function of this community. A number of programs was developed in the recent years to address this task.

Study objectives. Although the performances of these programs were initially investigated in the respective manuscripts, no broad independent evaluation was performed for a large number of use cases until now. We expected that the performance of the programs would differ depending on the implemented algorithm and on the characteristics of the studied metagenome. Therefore we performed a small-scale evaluation in a practical student course. Metagenomes were simulated using the program MetaSim [4], the obtained multifasta files were analyzed using different classification programs and the results were evaluated. Unpublished genomic data was utilized to simulate cases of phylogenetic “novelty”. This comparison showed that the classification success varied considerably between different programs and that no single program delivered completely satisfactory results. For this reason we decided to create a framework allowing for such simulations in an exhaustive way which would ease the determination of the best-suited program in each single use case.

Methods/Results. Our software aims for a comprehensive evaluation of the taxonomic classification performance of different available programs. It is implemented in Python and designed as a flexible framework of four modules, which can be run in succession or independently. The first part is the simulation of a metagenome using MetaSim according to a user-defined taxon profile, which produces a multifasta file of simulated short reads. The second part is the application of different programs performing the taxonomic classification of these reads. The sensitivity and specificity of the classfication are assessed in the third step, the evaluation. The fourth step is the comparison between the outcomes to summarize the results and to give a recommendation for the program to use.
The list of tools which can be utilized in step two, the program application, comprises so far two exemplary programs, Carma3 [5] and Phymm/PhymmBL [6]. It can be easily extended by program-specific wrapper functions, passing a multifasta file and the necessary parameters as input and accepting the prediction result as output. This result file is parsed and converted to a standardized file format which serves as input for the third step, the evaluation. To simulate cases of phylogenetic novelty, where no close relatives exist for the species from which reads are obtained, is has to be assured that the species (and their close relatives) which are present in the test data set are removed from the training data set; this ist the most challenging part as it is program-specific. The first three out of the four modules are implemented by now; the list of programs utilized in step two will be extended and the evaluation module will be implemented in near future.

Conclusion. Our evaluation framework will allow for an easy and comprehensive comparison of

taxonomic classification programs for metagenomic sequences. It will be easy to maintain and to extend for new programs, and will simplify the choice for scientists how to perform the taxonomic classification of their particular metagenome – an analysis method which, we think, will be even more important in the future. This work is very much a work-in-progress yet.

Bibliography

1: Hugenholtz, P., Goebel, BM. and Pace, NR., Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity, 1998
2: Venter, JC., Remington, K., Heidelberg, JF. et al., Environmental genome shotgun sequencing of the Sargasso Sea, 2004
3: Metzker, ML., Sequencing technologies - the next generation, 2010
4: Richter, DC., Ott, F., Auch, AF. et al., MetaSim: a sequencing simulator for genomics and metagenomics, 2008
5: Gerlach, W. and Stoye, J., Taxonomic classification of metagenomic shotgun sequences with CARMA3, 2011
6: Brady, A. and Salzberg, SL., Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models, 2009