Evaluation framework for the taxonomic classification of metagenomic sequences

Dmitrij Turaev, Michael Sommer, Thomas Rattei

Department of Computational Systems Biology, University of Vienna, Althanstraße 14, 1090 Vienna, Austria;
`dmitrij.turaev@univie.ac.at`

Metagenomics offers access to microbial communities, which are difficult or impossible to study under laboratory conditions [1]. It is on its way to establish itself as a common laboratory method in many contexts [2]. This was possible due to a number of novel sequencing technologies – next-generation sequencing technologies, coming up during the last decade and allowing for cheap and quick sequencing of nucleic acids [3]. The lengths of resulting sequence reads range between less than hundred and around thousand nucleotides. Low sequencing coverage, which is typical for the genomes of most of the community members, considerably limits their assembly. The short lengths of single reads and most contigs complicate one of the primary data analysis tasks: the taxonomic classification of metagenomic sequences. Thereby, sequence similarity or sequence composition can be utilized to determine from which organism a particular sequence is derived. However, no single generally accepted method exists. On the contrary, there is a substantial and still growing number of bioinformatics tools for this task [4]. Not always is the latest tool the greatest one, and many tools do not provide optimal classification results in particular cases. For the researcher it is important to choose the analysis tool wisely to prevent poor classification results. For example, if a sequence can't be classified correctly, it is more reasonable that it remains unclassified than if it is wrongly classified. However these choices are often made deliberately, as a rich empirical basis of evidence – apart from the researchers own expert knowledge – is lacking.

To address this problem and facilitate a well-founded tool selection, we propose a computational framework allowing for the evaluation of taxonomic different classification methods in terms of their predictive performance. The framework allows the simulation of defined metagenomic communities, which are then classified by multiple tools. If possible, the test data are separated from the training data and reference sequence databases, so that the classification is unbiased. The outcome of this classification step can be evaluated by counting the false and true positives and the false and true negatives and by calculating the sensitivity and selectivity of the single tools. The framework is written in Python. It is modular, allowing for an easy integration of new tools.

We are going to present the classification results of several simulated metagenomes, differing with respect to multiple parameters: the taxonomic complexity (number of different species represented), the taxonomic "novelty" (a lack of close relatives in training data or reference databases) of particular organisms, the sequencing technology and the associated read length. Is noteworthy that different tools can deliver substantially differing results on the same dataset. For example, our results show that in case of taxonomic novelty some tools, which may work well otherwise, turn out to be completely unreliable. This is especially true for composition-based classification algorithms. Similarity-based approaches, on the other hand, mostly suffer from incomplete or erroneous databases.

While most tools for taxonomic classification are best suited for prokaryotic genomes, we have also investigated the applicability of such tools to eukaryotic genomes, viral genomes and plasmids.

A future aim of the study is to test whether an integrative scoring of the results of multiple tools would have the potential to increase the prediction confidence.

References

[1] P. Hugenholtz et al. (1998) Hugenholtz, P., Goebel, B.M., Pace, N.R., 1998. Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity. J Bacteriol 180, 4765–4774.

[2] T. Thomas et al. (2012) Metagenomics - a guide from sampling to data analysis. Microb Inform Exp 2, 3.

[3] M. L. Metzker (2010) Sequencing technologies - the next generation. Nat. Rev. Genet. 11, 31–46.

[4] S. S. Mande et al. (2012) Classification of metagenomic sequences: methods and challenges. Brief. Bioinformatics 13, 669–681.