Genome analysis

Treephyler: fast taxonomic profiling of metagenomes

Fabian Schreiber1,2,3,∗, Peter Gumrich1, Rolf Daniel4 and Peter Meinicke1
1Abteilung Bioinformatik, Institut für Mikrobiologie und Genetik, Georg-August-Universität Göttingen, Goldschmidtstrasse 1, 37077 Göttingen, 2Department of Earth- and Environmental Sciences, 3GeoBioCenterLMU, Ludwig-Maximilians-Universität München, Richard-Wagner-Strasse 10, 80333 München and 4Abteilung Genomische und Angewandte Mikrobiologie, Institut für Mikrobiologie und Genetik, Georg-August-Universität, Giselaichstrasse 8, 37077 Göttingen, Germany

Associate Editor: Joaquín Dopazo

ABSTRACT

Summary: Assessment of phylogenetic diversity is a key element to the analysis of microbial communities. Tools are needed to handle next-generation sequencing data and to cope with the computational complexity of large-scale studies. Here, we present Treephyler, a tool for fast taxonomic profiling of metagenomes. Treephyler was evaluated on real metagenome to assess its performance in comparison to previous approaches for taxonomic profiling. Results indicate that Treephyler is in terms of speed and accuracy prepared for next-generation sequencing techniques and large-scale analysis.

Availability: Treephyler is implemented in Perl; it is portable to all platforms and applicable to both nucleotide and protein input data. Treephyler is freely available for download at http://www.gobics.de/fabian/treephyler.php

Contact: fschrei@gwdg.de

Received on December 18, 2009; revised on January 25, 2010; accepted on February 16, 2010

1 INTRODUCTION

Beyond the analysis of single species genomes of culturable organisms, metagenomics currently opens a new view on the exploration of microbial communities. Progress in sequencing technology enables broader and deeper genomic sampling of the biosphere which in turn puts new challenges for sequence analysis methods. Problems arise from the sheer mass and the short length of sequencing reads. Usually only a small fraction of reads can be assembled due to the phylogenetic diversity in the samples. In the first instance, large-scale analysis of short metagenomic sequencing reads has to provide an estimate of the phylogenetic distribution of the sample. Taxonomic profiling achieves this task by assigning sequencing reads to phylogenetic categories. The most common methods are based on homology to known genes.

The classical ‘gold standard’ approach to taxonomic profiling in metagenomics is focused on the 16S rRNA gene and relies on a sufficient number of sequences of that gene in metagenomic sequence data. Usually the number of reads containing sufficiently long stretches of 16S rRNA is small. Therefore, several researchers perform deep sequencing of that particular gene [see e.g. (Hamady and Knight, 2009)]. Although this approach efficiently overcomes the sparseness of 16S rRNA in metagenomic samples, the sequence data support taxonomic profiling only, without any explicit information about the functional inventory of microbial communities. Furthermore, 16S analysis does not apply to metatranscriptomics, an increasingly important approach to direct measurement of the metabolic activity of microbial communities. Another way to cope with the small proportion of 16S rRNA in metagenomic data is to extend the set of marker genes to particular protein coding genes (Wu and Eisen, 2008). In von Mering et al. (2007), a set of 31 marker genes for metagenome analysis was proposed. This principle has been further extended in Krause et al. (2008) where all PFAM protein domains are used as potential markers. Their tool CARMA searches metagenomic sequences for PFAM domains and classifies them on the basis of phylogenetic trees built from the metagenome and reference sequences. Although computationally demanding for large-scale metagenome analysis, the CARMA approach shows the potential of a dual use of PFAM domain assignments which not only provides a basis for taxonomic profiling but also for functional profiling as well. In principle, also BLAST-based analysis [MEGAN (Huyn et al., 2007), MG-RAST (Meyer et al., 2008)] can achieve both kinds of profiling at the same time because the detected homologies may provide information about functional and taxonomic relations. However, the known shortcomings of BLAST-based analysis in metagenomics include the requirement of a sufficient sequence length and the existence of close homologues in the reference database. In contrast to homology-based approaches, several methods pursue the direct classification of the DNA signature of single reads [PhyloPythia (McHardy et al., 2007), TACOA (Diaz et al., 2009), Phymm (Brady and Salzberg, 2009)]. While previous methods showed a rapidly decreasing classification performance for read lengths <1000 bp, more recent approaches also seem to perform reasonably well on short reads. Here, we present a new tool for community profiling in metagenomics and metatranscriptomics which is based on PFAM domain assignments. Previous methods like the CARMA approach are limited to small-scale analysis due to computational expense of homology search and tree inference. Here, we propose an approach which combines ultra-fast PFAM domain prediction as obtained from the UFO web server (Meinicke, 2009) with an efficient phylogenetic method based on fast tree inferences using approximate maximum likelihood trees (Price et al., 2009).

2 METHODS

Our algorithm offers fast taxonomic profiling to investigate the community structure of metagenomes. Based on PFAM predictions, e.g. by UFO, pre-calculated profile Hidden Markov Models of all PFAM families are used to
screen matching sequencing reads for significant hits. Reads are classified using a phylogenetic tree. For each PFAM family with a sufficient number of newly assigned sequences, approximate-maximum likelihood trees of the PFAM database sequences and the matching reads are computed using FastTree, which combines the speed of minimum-evolution methods with the accuracy of maximum likelihood methods. Once trees are computed, Treephyler uses the algorithm of (Nguyen et al., 2006) to classify reads according to the phylogenetic placement in the tree (see also Treephyler web site). Treephyler offers an efficient way to balance the computation load on multi-core computers or computer clusters. By this, the runtime only depends on the computation of the largest trees. Similar to CARMA, Treephyler only computes trees for PFAM families with less than 3000 (assigned + reference) sequences.

3 RESULTS

The glacial ice dataset (Simon et al., 2009) was taken as a reference because of its relatively short read length (~200 bp), the availability of results from a 16S analysis and the moderate sample size (~0.2 Gbp). We analysed the glacial ice dataset to assess the performance of Treephyler in comparison with the tree-based tool CARMA and the signature-based tool Phymm and the 16S RNA reference analysis. The analysis was conducted on a single 2.4 GHz dual-core CPU AMD Opteron with 16 Gb RAM.

For runtime comparison, we randomly selected 1% of the glacial ice dataset to allow the comparison with CARMA. Both Treephyler and Phymm analysed the reduced dataset in ~25 min, while it took CARMA 168 h to complete the analysis. On the full dataset, Treephyler needed only 12 h, while Phymm needed 30 h. The estimated runtime for CARMA is 696 h. The runtime of UFO for the reduced and the full dataset was 22 s and ~30 m, respectively. Results on the full dataset of Treephyler and CARMA [taken from (Simon et al., 2009)] are in good agreement with the 16S analysis, except for the phyla Bacteroidetes (Phymm: 3%, CARMA: 8%, Treephyler: 9%, 16S: 30%) and Gammaproteobacteria (P: 21%, C: 24%, T: 24%, 16S: 41%), where all three methods differ from the 16S analysis (see Fig. 1). This may be the consequence of an uneven taxon sampling of PFAM. Remarkably, Phymm also disagreed on the phyla Firmicutes (P: 11%, 16S: 0%) and Gammaproteobacteria (P: 21%, 16S: 5%). Test data and additional results for the class level are available at the Treephyler web site.

4 CONCLUSION

We introduced Treephyler, a new tool for fast taxonomic profiling of metagenomes. We evaluated our method on real metagenomic data by comparison with previous approaches for taxonomic profiling. We could show a close correspondence between the predicted profiles of Treephyler and CARMA, while computational speed was increased by orders of magnitude. While speed is not necessarily an essential requirement in genome analysis, the increase of metagenomic sequence data urges for particularly efficient techniques, which also work with limited computational resources. Therefore, the approach we propose here is well prepared for next-generation sequencing technologies and large-scale studies like the exploration of the human microbiome.

Funding: DFG (German Research Foundation) Priority Program SPP1174 ‘Deep Metazoan Phylogeny’ (Project Wo896/6-1).

Conflicts of Interest: none declared.

REFERENCES