TAX4FUN

Tax4Fun is an open-source R package that predicts the functional or metabolic capabilities of microbial communities based on 16S data samples. Tax4Fun is applicable to output as obtained through the SILVAngs web server (https://www.arb-silva.de/ngs/) or the application of QIIME (Caporaso et al., 2010) against the SILVA database (Quast et al., 2013). Further, the Tax4Fun package implements the MoP-Pro approach for whole metagenome shotgun sequence data (Kathrin Petra Aßhauer and Peter Meinicke, 2013). MoP-Pro implements a shortcut to estimate the metabolic profile of a metagenome. The taxonomic profile of the metagenome is linked to a set of pre-computed metabolic reference profiles. The combination of the taxonomic abundance estimates, obtained through the fast method Taxy-Pro (Klingenberg et al., 2013), and the metabolic reference profiles, based on the KEGG database, achieves an unrivaled speed of the metabolic profiling approach.

The R package is freely available at http://tax4fun.gobics.de/.

Installation

1) Checking system requirements

R (version 3.0.2 or higher)

2) Getting the most recent Tax4Fun package from http://tax4fun.gobics.de/ and the SILVA reference data http://tax4fun.gobics.de/ → SILVA Reference data (choose an appropriate SILVA version, e.g. 115)

3) Install R package according to your operating system

Manual Installation under Windows

1. Download a Tax4Fun_0.2.zip to somewhere on your local disk

2. Open the R-Gui (typically double-click the R icon on your desktop):

3. In the menu, click on Packages and select Install package(s) from local zip files...

4. Navigate to your local copy of Tax4Fun_0.2.zip and press Open.

5. Tax4Fun should be installed now - type the following (in R) to verify:

First `library(Tax4Fun)` (wait for execution), then `help(Tax4Fun)`.

Manual Installation under Linux

1. Download the Tax4Fun_0.2.tar.gz tarball to somewhere on your local disk.

2. Open a console, and navigate to the folder containing Tax4Fun_0.2.tar.gz.

3. Type `R CMD INSTALL Tax4Fun_0.2.tar.gz`

(Note: you probably need root privileges to do this)

4. Tax4Fun should be installed now - type the following in R to verify:

First `library(Tax4Fun)` (wait for execution), then `help(Tax4Fun)`.

Tax4Fun workflow

The whole prediction workflow consists of two parts, the generation of OTU tables using dedicated tools, e.g. QIIME (Caporaso et al., 2010) or SILVAngs (https://www.arb-silva.de/ngs/) , and the prediction of the functional and metabolic capabilities using Tax4Fun.

Generation of OTU tables

Details and examples for the generation of OTU tables can be found in section "*Prediction of OTU tables*".

Prediction of functional capabilities

1) Import OTU predictions

Tax4Fun supports different output formats of QIIME and SILVAngs. According to your output format, choose one of the following import functions:

`importQIIMEBiomData` – imports a single or multiple OTU tables in BIOM format e.g. generated by QIIME (see section *Prediction of OTU tables - QIIME - Step 5) Generate OTU table (BIOM format http://biom-format.org/)*)

`importQIIMEData` - imports a single or multiple OTU tables in txt format generated by QIIME (see section *Prediction of OTU tables - QIIME - Step 6) Convert BIOM to txt (optional; R can work with BIOM files)*)

`importSilvaNgsData` - imports a single or multiple OTU tables in csv format generated by SILVAngs (see section *Prediction of OTU tables - SILVAngs*)

Arguments:

`inputFiles` (required): a character vector with one or more character string(s) indicating the file location(s) of the BIOM/txt/csv formatted file(s).

Each of these import function takes a path either to a single input file or to a vector containing a set of input files. In the latter case, the all OTU tables are combined into a single input file for the functional or metabolic prediction.

2) Prediction of the functional or metabolic capabilities

`Tax4Fun` - predicts the functional or metabolic capabilities

Arguments:

`Tax4FunInput` (required): list containing the OTU table and sample names, e.g. imported with the functions `importQIIMEBiomData`, `importQIIMEData`, or `importSilvaNgsData`

`folderReferenceData` (required): a character vector with one character string indicating the folder location of the unzipped reference data. The reference data can be obtained from the Tax4Fun website http://tax4fun.gobics.de/ ("SILVA Reference data").

`fctProfiling` (optional): logical; if `TRUE` (default) the functional capabilities of microbial communities based on 16S data samples are computed using the pre-computed KEGG Ortholog reference profiles, and if `FALSE` the metabolic capabilities using the pre-computed KEGG Pathway reference profiles according to the MoP approach are computed.

`refProfile` (optional): an optional character string giving the method for pre-computing the functional reference profiles. This must be either "`UProC`" (default) or "`PAUDA`".

`shortReadMode` (optional): logical; if `TRUE` the functional reference profiles are computed based on 100 bp reads, and if `FALSE` (default) the reference profiles are computed based on 400 bp reads.

`normCopyNo` (optional): logical; if `TRUE` (default) the taxonomic profiles are normalized by the 16S rRNA gene copy number.

Prediction of OTU tables

Currently, Tax4Fun supports the output from the QIIME tool and the SILVAngs data analysis service for predicting the functional and metabolic profile based on amplicon data. In the following, you can find examples for the application of both tools. For further details, we refer to the official websites of QIIME (http://qiime.org/) and SILVAngs (https://www.arb-silva.de/ngs/).

QIIME

The Quantitative Insights Into Microbial Ecology (QIIME) is an open source software package for comparison and analysis of microbial communities (Caporaso et al., 2010).

The application of QIIME requires an installation of the software package (see http://qiime.org/). Further for the utilization of Tax4Fun the retrieval of taxonomic information has to be performed against the SILVA database. The supported SILVA database you can find at the Tax4Fun website (see http://gobics.de/kathrin/Tax4Fun/Tax4Fun.html - SILVA database for the application of QIIME).

1) Join appropriate-named sequences

```
cat soilA.fna soilB.fna soilC.fna > seqs.fna
cat *.fna > seqs.fna
```

Note: Sequences must be named appropriate. Underscores are allowed, but be careful.

Example:
3 soil samples: soil A, soil B, and soil C

Wrong: '>soil_A_seq1', '>soil_Aseq1', '>soilAseq1'
'>soil_A_seq1': header is interpreted to the first underscore. This example would generate an artificial sample called 'soil'.
'>soil_Aseq1': similar to '>soil_A_seq1'
'>soilAseq1': this example would generate an artificial sample called 'soilAseq1'. Each sequence would be treated as its own sample.

Correct: '>soilA_1' or 'soil.A_1'

The file should look like this:

```
>soilA_1
ACGTACGTACGTACGTACGTACGT
>soilA_2
ACGTACGTACGTACGTACGTACGT
…
>soilB_1
ACGTACGTACGTACGTACGTACGT
>soilB_2
ACGTACGTACGTACGTACGTACGT
…
>soilC_1
ACGTACGTACGTACGTACGTACGT
>soilC_2
ACGTACGTACGTACGTACGTACGT
```

```
…
```

2) Picking OTUs with uclust at 0.97 sequence similarity without no reverse strand matching:

```
pick_otus.py -i seqs.fna -o picked_otus
```

3) Obtaining reference sequences for each OTU

```
pick_rep_set.py -i picked_otus/seqs_otus.txt -f seqs.fna -o rep_set.fna -m
most_abundant
```

4) Retrieve taxonomic information for each OTU (with SILVA database in home directory)

single-core example:

```
assign_taxonomy.py -i rep_set.fna -b
~/SilvaSSURef_115_NR/SSURef_NR99_115_tax_silva_split.fasta -t
~/SilvaSSURef_115_NR/SSURef_NR99_115_tax_silva_split.taxonomy -m blast -o
rep_set_taxonomy
```

quad-core example:

```
parallel_assign_taxonomy_blast.py -i rep_set.fna -b
~/SilvaSSURef_115_NR/SSURef_NR99_115_tax_silva_split.fasta -t
~/SilvaSSURef_115_NR/SSURef_NR99_115_tax_silva_split.taxonomy -o
rep_set_taxonomy -O 4 -U start_parallel_jobs.py
```

5) Generate OTU table (BIOM format)

```
make_otu_table.py -i picked_otus/seqs_otus.txt -t
rep_set_taxonomy/rep_set_tax_assignments.txt -o otu_table.biom
```

6) Covert BIOM to txt (optional; R can work with BIOM files)

```
biom convert -i otu_table.biom -o otu_table.txt  --header-key taxonomy
-b
```

Tax4Fun supports both BIOM and txt format. For the import of OTU predictions in BIOM format use the `importQIIMEBiomData` and in txt format the `importQIIMEData` function.

SILVAngs

Especially, if you want to avoid a complex installation, you can use the SILVAngs web interface for rDNA-based microbial community analysis using next-generation sequencing (NGS) data approaches. The SILVAngs data analysis service is based on an automatic software pipeline and available at https://www.arb-silva.de/ngs/. For running SILVAngs see *SILVAngs user guide*.

After running the SILVAngs download your results using the option “download ARCHIVE ZIP”. Extract all files and choose the file “*projectName---ssu---fingerprint----Total---sim_93---tax_silva---td_20.csv*” in the directory “*results/ssu/tax_breakdown/fingerprint/*” as input file for Tax4Fun. Import the SILVAngs OTU prediction using the import function `importSilvaNgsData`.

MoP-Pro workflow

The whole MoP-Pro workflow consists of three steps, the prediction of Pfam protein domain families using UProC (http://uproc.gobics.de/) or CoMet-Universe (http://comet2.gobics.de/), taxonomic profiling using Taxy-Pro (version MoP-Pro Pfam 27) (Klingenberg et al., 2013), and the estimation of the metabolic capabilities using MoP-Pro.

1) Pfam prediction using UProC or CoMet-Universe webserver

UProC

Pfam predictions for input sequences can be carried out using the ultrafast protein classification (UProC) toolbox, which implements a novel algorithm ("Mosaic Matching") for large-scale sequence analysis. UProC is available in terms of an open source C library at: http://uproc.gobics.de/.

For prediction of the metabolic capabilities use the Pfam (version 27.0) as reference database and run the UProC engine (details see *UProC documentation*).

2) Taxonomic profiling using Taxy-Pro (version MoP-Pro Pfam 27)

Subsequent to the Pfam domain detection the taxonomic profile of the metagenome has to be calculated. This task can be performed using the Taxy-Pro toolbox for the Matlab/Octave programming environment. For the MoP-Pro workflow the Taxy-Pro toolbox (version MoP-Pro Pfam 27) has to be chosen. This Taxy-Toolbox you can find at the Tax4Fun website (http://gobics.de/kathrin/Tax4Fun/Tax4Fun.html).

With the output obtained from the UProC engine you can run Taxy-Pro. Note, both UProC and Taxy-Pro have to be executed in the same sequence length mode (long read mode or short read mode, for details see *UProC documentation*).

Taxy-Pro will put all output files in a directory. In this directory, you will find the taxonomic classifications at different taxonomic levels. For the metabolic profiling you have to select the file with suffix: "_MoPPro.csv".

3) Metabolic profiling using MoP-Pro

a) Import Taxy-Pro predictions

`importTaxyProData` – imports a single or multiple taxonomic profile in csv format generated by Taxy-Pro

Argument:

`inputFiles` (required): a character vector with one or more character string(s) indicating the file location(s) of the Taxy-Pro csv formatted file(s).

Each of these import function takes a path either to a single input file or to a vector containing a set of input files. In the latter case, all taxonomic profiles are combined into a single input file for metabolic prediction.

b) Prediction of the metabolic capabilities

`MoPPro` - predicts the metabolic capabilities for whole metagenome shotgun sequence data using the MoP-Pro approach

Argument:

`MoPProInput` (required): list containing the taxonomic profile and sample name(s), e.g. imported with the function `importTaxyProData`.

`folderReferenceData` (required): a character vector with one character string indicating the folder location of the unzipped reference data. The reference data can be obtained from the Tax4Fun website http://tax4fun.gobics.de/ ("SILVA Reference data").

Note: The CoMet-Universe webserver implements the complete MoP-Pro workflow

In case you want to avoid an installation of the standalone tools UProC, Taxy-Pro and the Tax4Fun R package you can use the CoMet-Universe web interface for Pfam protein domain detection, taxonomic and metabolic profiling. Beyond these metagenome analyses, the CoMet-Universe provides the possibility to compare a particular metagenome with more than thousand precomputed profiles from publicly available data sets or with previously uploaded data from the same user. Furthermore, CoMet-Universe provides interactive Krona charts for profile visualization. The CoMet-Universe webserver is available at: http://comet2.gobics.de/.

Literature

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Kathrin Petra Aßhauer, and Peter Meinicke (2013). On the estimation of metabolic profiles in metagenomics. (Göttingen: Schloss Dagstuhl - Leibniz-Zentrum für Informatik GmbH), pp. 1–13.

Klingenberg, H., Aßhauer, K.P., Lingner, T., and Meinicke, P. (2013). Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. Bioinforma. Oxf. Engl. *29*, 973–980.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. *41*, D590–596.