Taxonomic Binning Script

Task

Produce overview tables and graphics of the microbial composition in the samples at different taxonomic levels.

Background

The taxonomic classification of Operational Taxonomic Units (OTUs) allows their combination to higher taxonomic levels for estimation of the taxonomic composition of samples. This is done by summing up the relative sequence abundances of all OTUs that share the same assignment at a given level. Of course, the quality of classification directly affects the output of taxonomic composition, especially for low taxonomic levels such as genera. There are many ways to classify OTUs to known taxonomies (all with different advantages and disadvantages), including the Bayesian classifier of RDP [1] and the Lowest Common Ancestor (LCA) used in SILVA [2]. Using multiple ones to refine classification is strongly recommended. Unfortunately, given that a large proportion of microorganisms have not yet been characterized, it is unavoidable that many OTUs originate from unknown species and thus present incomplete taxonomic classification. Furthermore, since the classification is performed using a small fragment of the 16S rRNA gene that may or may not have resolving capacity for all taxonomic levels and given the official taxonomic misclassification of many species, we recommend to rely on taxonomic binning at the family or higher taxonomic levels.

References

1. Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology, 73(16), 5261-5267.

2. Pruesse, E., Peplies, J., & Glöckner, F. O. (2012). SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics, 28(14), 1823-1829.

Input

The expected input file for this script is the OTU-table containing relative abundances and the taxonomic classification. If the complete Rhea pipeline is followed, the Normalization Script creates automatically a copy of the appropriate file for direct usage. The taxonomy string must be delimited by semicolons, with exactly 6 fields (left empty if not known).

	Sample1	Sample2	Sample3	Sample4	taxonomy
OTU_1	25.22243	45.84383	39.53622	62.86657	Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;;
OTU_2	32.54246	9.244332	29.61256	15.42465	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Rikenellaceae;Alistipes;
OTU_3	3.181451	6.775819	1.267464	3.097626	Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;;
OTU_4	3.019682	10.88161	2.822987	4.03707	Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;Clostridium XlVb;
OTU_5	5.068752	1.712846	2.074031	0.34277	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;;
OTU_6	17.57886	0.629723	5.213884	4.303669	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;Parabacteroides;
OTU_7	2.817471	20.25189	0.518508	0.228513	Bacteria;Proteobacteria;Deltaproteobacteria;Desulfovibrionales;Desulfovibrionaceae;;
OTU_8	3.639795	2.720403	2.189255	0.558588	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Porphyromonadaceae;;
OTU_9	2.588299	0.604534	1.339479	1.866193	Bacteria;Proteobacteria;Deltaproteobacteria;Desulfovibrionales;Desulfovibrionaceae;;
OTU_10	4.340793	1.335013	15.42561	7.274343	Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Bacteroidaceae;Bacteroides;

Output

The script is producing 6 tabular text files that contain the taxonomic composition of the samples at each level: Kingdom, Phylum, Class, Order, Family, and Genus. In addition, a combination of all levels in one file (tax.summary.all.tab) and a corresponding graphical display as a PDF file (taxonomic-overview.pdf) are offered. The summary table file is also copied wherever needed for subsequent statistical analysis.

	Sample1	Sample2	Sample3	Sample4
p__Bacteroidetes	52.00317	15.35963	48.32617	22.02857
p__Candidatus Saccharibacteria	0	0	0.047217	0.217281
p__Deferribacteres	0.173327	0.702073	0.136928	0.04623
p__Firmicutes	45.21864	78.82797	50.01653	76.09449
p__Proteobacteria	2.505819	3.96886	1.114311	1.460866
p__unknown_ Bacteria	0.099044	1.141465	0.358846	0.152559

Important Notes

The format of the taxonomic line is very important for correct taxonomic binning. If different classification methods were used for creation of the taxonomic assignment, special care must be taken to consistent taxa designation. Failure to do so will lead to either scripts errors or erroneous outputs. OTUs with incomplete classification are automatically assigned to the artificial classification “unknown” followed by the next parent level classification with a known name. For example, an OTU that was classified as:

“Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;;”

Will be assigned to “unknown_Ruminococcaceae” at the genus level. OTUs that share this classification are added together to build the relative abundance of the taxa (although in this particular case they may represent several unknown genera within *Ruminococcaceae*).

Common problems

- The path to the script is not set correctly
- The input file name is incorrect
- The input file is of different format (e.g. has no taxonomic classification column)
- The taxonomy line contains special characters (e.g. “”).
- The taxonomy line has incorrect format (e.g. missing semicolons)