phylotypr: An R package for classifying DNA sequences

Running title: phylotypr

Patrick D. Schloss†

†To whom correspondence should be addressed

pschloss@umich.edu

Department of Microbiology & Immunology

University of Michigan

Ann Arbor, MI 48109

Software Announcement

Abstract

The phylotypr R package implements the popular naive Bayesian classification algorithm that is frequently used to classify 16S rRNA and other gene sequences to taxonomic lineages. A companion data package, phylotyprrefdata, also provides numerous versions of taxonomic databases from the Ribosomal Database Project, SILVA, and greengenes.

Announcement

Since it was published in 2007, the naive Bayesian classifier has been the most popular and performant tool for classifying 16S rRNA gene sequences (1). The method calculates the probability distributions of k-mers (typically 8-mers) across a reference collection and within each genus represented in the collection. These probabilities are used within a pseudo-bootstrapping procedure to classify unknown sequences and assign a confidence score to that classification. The confidence scores are used to prune the Linnaean taxonomy to the deepest possible taxonomic level with sufficient confidence (typically 80%). The algorithm was been made available by the original developers as an application coded in Java; a wrapper for the original code was available in QIIME (2). A C++ version has been available in mothur and a Python version in QIIME2 (3, 4). Until March 2023, users could classify sequences with an online interface at the Ribosomal Database Project (RDP); this interface is no longer available. The RDP developers continue to update their code and the database through their GitHub and Sourceforge-based repositories (5).

Considering the growing popularity of the R programming language among microbial ecologists (6–10), I developed an R-based version of the algorithm that is available as the phylotypr package. Users can install phylotypr via CRAN or through the devtools package's install_github function. Classification consists of two steps. First, the reference sequences and taxonomies are used to calculate kmer-based probabilities with the build_kmer_database function. Users can specify their desired kmer size when generating the database. These probabilities can be recalculated for each R session or saved as an R data file. Their calculation can be completed within several seconds. Second, user-supplied sequences can be classified using the reference database with the classify_sequence function. Accessory filter_taxonomy and print_taxonomy functions allow the user to output the results of their classifications using a minimum confidence score threshold. A detailed vignette is available within the phylotypr package that demonstrates how to install the package, use its functions, and parallelize its performance using the furrr package. The R-based execution time is comparable to or faster than that found when using the

classify.seqs mothur function written in C++.

Many microbial ecologists have benefited from training the algorithm using reference sequences and taxonomies curated by the RDP as well as other providers including greengenes and SILVA (5, 11–16). For demonstration purposes, phylotypr includes a small reference database using version 9 of the RDP's reference. A companion data package, phylotyprrefdata, is available on GitHub and can be installed using the install_github function from the devtools package. The current version of the data package (v0.1.0) includes all publicly available versions of the references from each of the RDP, greengenes, and SILVA references. Because of the size of the package (150 MB), it is too large to post to CRAN. I plan to make regular updates to the data package as new versions of databases become available. Users can also provide their own reference data to classify genes other than the 16S or 18S rRNA gene to to improve classification of lineages that are poorly represented in the references.

Data availability

phylotypr is available through CRAN and developmental versions are available through the project's GitHub website (https://github.com/mothur/phylotypr). A pkgdown version of the documentation is hosted at (https://mothur.org/phylotypr). The phylotyprref data package is available through the project's GitHub website (https://github.com/mothur/phylotyprref). The phylotypr package is available under the GNU General Public License (v3) and the phylotyprref package is available under the MIT open source license.

Acknowledgements

phylotypr was developed as a series of videos on the Riffomonas YouTube channel (https://www.youtube.com/playlist?list=PLmNrK_nkqBpIZlWa3yGEc2-wX7An2kpCL). I am grateful to the viewers of the Riffomonas YouTube channel for their questions, suggestions, and encouragement throughout its development.

References

1. **Wang Q**, **Garrity GM**, **Tiedje JM**, **Cole JR**. 2007. Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology **73**:5261–5267. doi:10.1128/aem.00062-07.

2. **Caporaso JG**, **Kuczynski J**, **Stombaugh J**, **Bittinger K**, **Bushman FD**, **Costello EK**, **Fierer N**, **Peña AG**, **Goodrich JK**, **Gordon JI**, **Huttley GA**, **Kelley ST**, **Knights D**, **Koenig JE**, **Ley RE**, **Lozupone CA**, **McDonald D**, **Muegge BD**, **Pirrung M**, **Reeder J**, **Sevinsky JR**, **Turnbaugh PJ**, **Walters WA**, **Widmann J**, **Yatsunenko T**, **Zaneveld J**, **Knight R**. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods **7**:335–336. doi:10.1038/nmeth.f.303.

3. **Schloss PD**, **Westcott SL**, **Ryabin T**, **Hall JR**, **Hartmann M**, **Hollister EB**, **Lesniewski RA**, **Oakley BB**, **Parks DH**, **Robinson CJ**, **Sahl JW**, **Stres B**, **Thallinger GG**, **Van Horn DJ**, **Weber CF**. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology **75**:7537–7541. doi:10.1128/aem.01541-09.

70 4. **Bokulich NA**, **Kaehler BD**, **Rideout JR**, **Dillon M**, **Bolyen E**, **Knight R**, **Huttley GA**, **Gregory Caporaso J**. 2018. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome **6**. doi:10.1186/s40168-018-0470-z.

5. **Wang Q**, **Cole JR**. 2024. Updated RDP taxonomy and RDP classifier for more accurate taxonomic classification. Microbiology Resource Announcements **13**. doi:10.1128/mra.01063-23.

6. **Liu C**, **Cui Y**, **Li X**, **Yao M**. 2020. Microeco: An r package for data mining in microbial community ecology. FEMS Microbiology Ecology **97**. doi:10.1093/femsec/fiaa255.

7. **Buttigieg PL**, **Ramette A**. 2014. A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses. FEMS Microbiology Ecology **90**:543–550. doi:10.1111/1574-6941.12437.

8. **McMurdie PJ**, **Holmes S**. 2013. Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. PLoS ONE **8**:e61217. doi:10.1371/journal.pone.0061217.

9. **Callahan BJ**, **McMurdie PJ**, **Rosen MJ**, **Han AW**, **Johnson AJA**, **Holmes SP**. 2016. DADA2: High-resolution sample inference from illumina amplicon data. Nature Methods **13**:581–583. doi:10.1038/nmeth.3869.

10. **Dixon P**. 2003. VEGAN, a package of r functions for community ecology. Journal of Vegetation Science **14**:927–930. doi:10.1111/j.1654-1103.2003.tb02228.x.

11. **DeSantis TZ**, **Hugenholtz P**, **Larsen N**, **Rojas M**, **Brodie EL**, **Keller K**, **Huber T**, **Dalevi D**, **Hu P**, **Andersen GL**. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology **72**:5069–5072. doi:10.1128/aem.03006-05.

12. **McDonald D**, **Price MN**, **Goodrich J**, **Nawrocki EP**, **DeSantis TZ**, **Probst A**, **Andersen GL**, **Knight R**, **Hugenholtz P**. 2011. An improved greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME Journal **6**:610–618. doi:10.1038/ismej.2011.139.

13. **McDonald D**, **Jiang Y**, **Balaban M**, **Cantrell K**, **Zhu Q**, **Gonzalez A**, **Morton JT**, **Nicolaou G**, **Parks DH**, **Karst SM**, **Albertsen M**, **Hugenholtz P**, **DeSantis T**, **Song SJ**, **Bartko A**, **Havulinna AS**, **Jousilahti P**, **Cheng S**, **Inouye M**, **Niiranen T**, **Jain M**, **Salomaa V**, **Lahti L**, **Mirarab S**, **Knight R**. 2023. Greengenes2 unifies microbial data in a single reference tree. Nature Biotechnology **42**:715–718. doi:10.1038/s41587-023-01845-1.

14. **Yilmaz P**, **Parfrey LW**, **Yarza P**, **Gerken J**, **Pruesse E**, **Quast C**, **Schweer T**, **Peplies J**, **Ludwig W**, **Glöckner FO**. 2013. The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. Nucleic Acids Research **42**:D643–D648. doi:10.1093/nar/gkt1209.

15. **Werner JJ**, **Koren O**, **Hugenholtz P**, **DeSantis TZ**, **Walters WA**, **Caporaso JG**, **Angenent LT**, **Knight R**, **Ley RE**. 2011. Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. The ISME Journal **6**:94–103. doi:10.1038/ismej.2011.82.

16. **Quast C**, **Pruesse E**, **Yilmaz P**, **Gerken J**, **Schweer T**, **Yarza P**, **Peplies J**, **Glöckner FO**. 2012. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research **41**:D590–D596. doi:10.1093/nar/gks1219.