clustur: An R package for clustering features using sparse distances matrices

Running title: clustur

Gregory Johnson Jr., Sarah L. Westcott, 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 (needs to be under 50 words; currently 48)

The clustur R package implements the *de novo* clustering algorithms found in the mothur software package for assigning 16S rRNA gene sequences to operational taxonomic units (OTUs). Making these algorithms accessible through the R ecosystem will foster their further development, broader application, and integration within other R packages.

Announcement (needs to be ~500 words, currently 608)

Taxonomic classification of 16S rRNA gene sequences has been a persistent problem in microbial ecology studies because reference databases are incomplete (1). As an alternative, operational taxonomic units (OTUs) have been widely used for describing and comparing microbial communities. Although the biological interpretation is controversial, OTUs are typically defined as a group of sequences that are more than 97% similar or less than 3% dissimilar to each other (2). Methods for applying that definition has resulted in a sizable literature. Three general approaches have emerged for assigning sequences to OTUs. First, sequences can be clustered based on their similarity to each other. This approach has often been called "*de novo* clustering". Second, sequences can be clustered based on their similarity to reference sequences. This approach has often been called "closed reference clustering" or "phylotyping". Finally, a hybrid approach applies *de novo* clustering to sequences that are not sufficiently close to reference sequences when using closed reference clustering. This approach has been called "open reference clustering". We have described and compared these methods in great detail elsewhere (3–6). These methods are available through popular packages including mothur and QIIME (7, 8).

The clustur R package implements the *de novo* clustering algorithms implemented in mothur. The package name references its focus on clustering and the names of its predecessors DOTUR and mothur (7, 9). This package was developed to help address two issues. First, users would be able to more easily integrate the type of analysis that mothur specializes in with popular analysis and visualization packages within the R package ecosystem. Second, by making the code behind mothur's clustering functions accessible through the R language, we hope to encourage further development of the algorithms behind these functions and analyses based on the output of the functions. The clustur package implements hierarchical clustering algorithms including the furthest, nearest, unweighted (i.e. average), and weighted neighbor clustering algorithms and the OptiFit algorithm. Although functions implementing the hierarchical algorithms already exist within R, their implementations within clustur make use of a sparse input distance matrix and

output data for a single distance threshold. The benefits of censoring distances larger than the threshold and only outputting data for a single threshold include a smaller memory requirement and faster execution times. clustur makes use of the Rcpp R package to implement C++ code originally written for the mothur software package to preserve the speed of the functions.

Users can install the clustur package via CRAN or through the devtools package's install_github function. The primary input to clustur's functions is a sparse distance matrix and a count file. The sparse distance matrix is a data.table package object with two columns indicating the identifiers of the sequences being compared and a column with the distance between those sequences; data for comparisons with a distance larger than the desired threshold (e.g., 0.03) does not need to be included. The count file is a data.table package object indicating the number of times a sequence is found in each sample. The functions output a data.table object with two columns indicating the sample and OTU identifiers and a column indicating the number of times each OTU is found in each sample. Detailed vignettes are available within the package to teach users how to install the package, use its functions, and perform downstream analyses including analysis within the vegan and ggplot2 R packages.

Data availability

clustur is available through CRAN and developmental versions are available through the project's GitHub website (https://github.com/schlosslab/clustur). The package is available under the MIT open source license.

Acknowledgements

This project was supported, in part, by a grant from the US National Institutes of Health (U01CA264071) to PDS.

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. **STACKEBRANDT E**, **GOEBEL BM**. 1994. Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. International Journal of Systematic and Evolutionary Microbiology **44**:846–849. doi:10.1099/00207713-44-4-846.

3. **Schloss PD**, **Westcott SL**. 2011. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Applied and Environmental Microbiology **77**:3219–3226. doi:10.1128/aem.02810-10.

4. **Schloss PD**. 2016. Application of a database-independent approach to assess the quality of operational taxonomic unit picking methods. mSystems **1**. doi:10.1128/msystems.00027-16.

5. **Westcott SL**, **Schloss PD**. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ **3**:e1487. doi:10.7717/peerj.1487.

6. **Westcott SL**, **Schloss PD**. 2017. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. mSphere **2**. doi:10.1128/mspheredirect.00073-17.

7. **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.

8. **Bolyen E**, **Rideout JR**, **Dillon MR**, **Bokulich NA**, **Abnet CC**, **Al-Ghalith GA**, **Alexander H**, **Alm EJ**, **Arumugam M**, **Asnicar F**, **Bai Y**, **Bisanz JE**, **Bittinger K**, **Brejnrod A**, **Brislawn CJ**, **Brown CT**, **Callahan BJ**, **Caraballo-Rodríguez AM**, **Chase J**, **Cope EK**, **Da Silva R**, **Diener C**, **Dorrestein PC**, **Douglas GM**, **Durall DM**, **Duvallet C**, **Edwardson CF**, **Ernst M**, **Estaki M**, **Fouquier J**, **Gauglitz JM**, **Gibbons SM**, **Gibson DL**, **Gonzalez A**, **Gorlick K**, **Guo J**, **Hillmann B**, **Holmes S**, **Holste H**, **Huttenhower C**, **Huttley GA**, **Janssen S**, **Jarmusch AK**, **Jiang L**, **Kaehler BD**, **Kang KB**, **Keefe CR**, **Keim P**, **Kelley ST**, **Knights D**, **Koester I**, **Kosciolek T**, **Kreps J**, **Langille MGI**, **Lee J**, **Ley R**, **Liu Y-X**, **Loftfield E**, **Lozupone C**, **Maher M**, **Marotz C**, **Martin BD**, **McDonald D**, **McIver LJ**, **Melnik AV**, **Metcalf JL**, **Morgan SC**, **Morton JT**, **Naimey AT**, **Navas-Molina JA**, **Nothias LF**, **Orchanian SB**, **Pearson T**, **Peoples SL**, **Petras D**, **Preuss ML**, **Pruesse E**, **Rasmussen LB**, **Rivers A**, **Robeson MS**, **Rosenthal P**, **Segata N**, **Shaffer M**, **Shiffer A**, **Sinha R**, **Song SJ**, **Spear JR**, **Swafford AD**, **Thompson LR**, **Torres PJ**, **Trinh P**, **Tripathi A**, **Turnbaugh PJ**, **Ul-Hasan S**, **Hooft JJJ van der**, **Vargas F**, **Vázquez-Baeza Y**, **Vogtmann E**, **Hippel M von**, **Walters W**, **Wan Y**, **Wang M**, **Warren J**, **Weber KC**, **Williamson CHD**, **Willis AD**, **Xu ZZ**, **Zaneveld JR**, **Zhang Y**, **Zhu Q**, **Knight R**, **Caporaso JG**. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology **37**:852–857. doi:10.1038/s41587-019-0209-9.

9. **Schloss PD**, **Handelsman J**. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Applied and Environmental Microbiology **71**:1501–1506. doi:10.1128/aem.71.3.1501-1506.2005.