AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

| Computational Biology | Announcement

clustur: an R package for clustering features using sparse distance matrices

Gregory Johnson,[1] Sarah L. Westcott,[1] Patrick D. Schloss[1]

AUTHOR AFFILIATION See affiliation list on p. 2.

ABSTRACT 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.

KEYWORDS bioinformatics, software, hierarchical clustering, operational taxonomic unit, bacterial taxonomy, microbiome

Taxonomic classification of 16S rRNA gene sequences has been a persistent challenge 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 their 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: *de novo* clustering, closed reference clustering or phylotyping, and open reference clustering (3–9). These methods are available through popular packages, including mothur and QIIME2 (10, 11).

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 (10, 12). 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 OptiClust algorithm. Functions implementing the hierarchical algorithms already exist within R; however, 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 (4). 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

Editor Catherine Putonti, Loyola University Chicago, Chicago, Illinois, USA

Address correspondence to Patrick D. Schloss, pschloss@umich.edu.

The authors declare no conflict of interest.

See the funding table on p. 2.

Received 14 November 2024
Accepted 9 January 2025
Published 30 January 2025

distance larger than the desired threshold (e.g., 0.03) do 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 cluster functions output two data.table objects. The first one has two columns indicating the sequences and OTU identifiers. The second displays the abundance of each sequence in each OTU. This has identical functionality to the cluster and make.shared functions from mothur. 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.

ACKNOWLEDGMENTS

This project was supported, in part, by a grant from the U.S. National Institutes of Health (U01CA264071) to P.D.S.

AUTHOR AFFILIATION

[1]Department of Microbiology & Immunology, University of Michigan, Ann Arbor, Michigan, USA

AUTHOR ORCIDs

Patrick D. Schloss http://orcid.org/0000-0002-6935-4275

FUNDING

Funder	Grant(s)	Author(s)
HHS | National Institutes of Health (NIH)	U01CA264071	Patrick D. Schloss

AUTHOR CONTRIBUTIONS

Gregory Johnson, Methodology, Resources, Software, Writing – review and editing | Sarah L. Westcott, Methodology, Resources, Software, Writing – review and editing | Patrick D. Schloss, Conceptualization, Project administration, Resources, Software, Supervision, Writing – original draft, Writing – review and editing

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 GNU General Public License (v3.0).

REFERENCES

1. Wang Q, Garrity GM, Tiedje JM, Cole JR. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267. https://doi.org/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. Int J Syst Evol Microbiol 44:846–849. https://doi.org/10.1099/00207713-44-4-846
3. Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, Xu Z, Ursell LK, Lauber C, Zhou H, Song SJ, Huntley J, Ackermann GL, Berg-Lyons D, Holmes S, Caporaso JG, Knight R. 2013. Advancing our understanding of the human microbiome using QIIME, p 371–444. In Microbial metagenomics, metatranscriptomics, and metaproteomics. Elsevier.
4. Schloss PD, Westcott SL. 2011. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol 77:3219–3226. https://doi.org/10.1128/AEM.02810-10
5. Schloss PD. 2016. Application of a database-independent approach to assess the quality of operational taxonomic unit picking methods. mSystems 1. https://doi.org/10.1128/mSystems.00027-16
6. 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. https://doi.org/10.7717/peerj.1487
7. Westcott SL, Schloss PD. 2017. OptiClust, an improved method for assigning amplicon-based sequence data to operational taxonomic units. mSphere 2:e00073-17. https://doi.org/10.1128/mSphereDirect.00073-17
8. Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahé F, He Y, Zhou H-W, Rognes T, Caporaso JG, Knight R. 2016. Open-source sequence clustering methods improve the state of the art. mSystems 1. https://doi.org/10.1128/mSystems.00003-15

9. He Y, Caporaso JG, Jiang X-T, Sheng H-F, Huse SM, Rideout JR, Edgar RC, Kopylova E, Walters WA, Knight R, Zhou H-W. 2015. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. Microbiome 3. https://doi.org/10.1186/s40168-015-0081-x

10. 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. Appl Environ Microbiol 75:7537–7541. https://doi.org/10.1128/AEM.01541-09

11. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852–857. https://doi.org/10.1038/s41587-019-0209-9

12. Schloss PD, Handelsman J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. Appl Environ Microbiol 71:1501–1506. https://doi.org/10.1128/AEM.71.3.1501-1506.2005