The Journal of Open Source Software

phylosmith: an R-package for reproducible and efficient microbiome analysis with phyloseq-objects

Schuyler D. Smith[1]

1 Department of Bioinformatics and Computational Biology, Iowa State University

DOI: 10.21105/joss.01442

Software

- Review
- Repository
- Archive

Submitted: 03 May 2019
Published: 20 June 2019

Summary

This paper presents `phylosmith`, an `R`-package that enables reproducible and efficient analysis of microbiome data with `phyloseq-class` objects by providing robust and efficient functions. `phylosmith` utilizes the standardized data format of `phyloseq` and `R` object accession methods to provide functions with simple and intuitive input arguments.

The functions provided in `phylosmith` have been divided into 3 categories.

Data Wrangling

These functions include operations that will either return a transformed version of the input `phyloseq` object, a subset version, or extracted data on set parameters. In some cases the functions are practical rewrites of ones already available from `phyloseq` with additional features or a more efficient implementation with `data.table` for large datasets. Various functions include finding shared or exclusive taxa by treatments, agglomeration, factor handling, and filtering.

Graphs

The graphs are designed to serve as a quick and easy way to visualise data for analysis and to provide a foundation for figures for publishing. Graphics include ordinations, phylogeny profiles, and co-occurrence networks. All images are produced as `ggplot` objects (Wickham (2016)), allowing for the image to be altered and additional layers given to tailor the graphic as desired. Additionally, the code for producing the graphs is readily accessible, allowing for the code to be reused and tailored to fit needs, providing a foundation to start from. The most novel, for the field of microbiome research, is the implementation of a t-SNE ordination. Most studies have used PCA or NMDS, which can suffer from converging to a local minima on large datasets, t-SNE is designed for large datasets and is not susceptible to these same limitations.

Calculations

As of publication of this paper, the functions in this section all pertain to calculating and analyzing the Spearman rank co-occurrence. The routine was written in efficient C++ code and interfaced with R using the `Rcpp` API (Eddelbuettel et al. (2011)). The resulting co-occurrence table matches that produced by the `cor()` function in the `R stats` package, but is calculated much faster on a single thread, with a multi-threading options implemented as well.

Need

Adoption of data-standards enable data that are readily available for sharing and also the creation and implementation of tools for reproducible research. It is commonly said that in the age of big-data, biologists are required to have computational proficiency and literacy (Carey, 2018). It seems reasonable that there should be a large onus on bioinformaticians to create accessible and practical tools that enable the biologists.

For the field of microbiome research, a formulaic approach to analysis has developed as commonplace. A generic study will incorporate some combination and implementation of the same resulting figures; ordination, profile bar-chart, heatmap, network, etc (Huttenhower et al., 2012), (Turnbaugh et al., 2009), (Arumugam et al., 2011). Each new scientist, often from a biology, microbiology, ecology, or environmental science background, is required to learn how to produce these analyses and figures. A lot of time is spent learning how to generate these plots. Even more time is spent learning how to process data; which can easily be done incorrectly without being apparently obvious (i.e., incorrect logical subsetting, factor levels set incorrectly, or even reordering of samples due to string sorting methods), leading to incorrect results and conclusions.

For microbiome researchers using the `R` statistical programming language (R-Core & others, 2013), a data-standard has been available in `phyloseq` (McMurdie, 2013). `phyloseq` provides an S4-class object that contains a count table, taxa table, and associated metadata, along with a phylogenetic tree slot and reference sequence slot. For beginning and intermediate, users of `R`, S4 objects can be a barrier, as they require an additional layer of accession methods compared to the base S3 objects. `phyloseq` offers several functions for handling its objects, as well as functions for producing some common figures, but is by no means a complete toolset. Additionally, when the authors originally wrote `phyloseq`, advanced tools such as the `data.table` package (Dowle et al., 2014) were not practically available and thus had not been implemented within the program.

Providing tools for reproducible and efficient research can help microbiome researchers to focus more effort on answering biological questions. Providing simple implementations of tools, such as t-SNE (Maaten & Hinton, 2008), can increase the acceptance and adoption of new techniques in a field that is hesitant to do so. The importance of these tools should not be overlooked for the importance of science and understanding as a whole.

Acknowledgement

This work was supported by the National Science Foundation Directorate of Biological Sciences under awards DEB 1737758 and DEB 1737765.

References

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., et al. (2011). Enterotypes of the human gut microbiome. *nature*, *473*(7346), 174.

Carey, J. A., Maureen A. AND Papin. (2018, January). Ten simple rules for biologists learning to program. *PLOS Computational Biology*. Public Library of Science. doi:10.1371/journal.pcbi.1005871

Dowle, M., Short, T., Lianoglou, S., Saporta, R., Srinivasan, A., & Antonyan, E. (2014). Data. Table: Extension of data. Frame.

Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., et al. (2011). Rcpp: Seamless r and c++ integration. *Journal of Statistical Software*, *40*(8), 1–18.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., et al. (2012). Structure, function and diversity of the healthy human microbiome. *nature*, *486*(7402), 207. doi:10.1038/nature11234

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(Nov), 2579–2605.

McMurdie, S., Paul J. AND Holmes. (2013). Phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE*, *8*(4), 1–11. doi:10.1371/journal.pone.0061217

R-Core, & others. (2013). R: A language and environment for statistical computing.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., et al. (2009). A core gut microbiome in obese and lean twins. *nature*, *457*(7228), 480.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer.