A standard for taxonomic data storage and manipulation in R

Zachary S. L. Foster

Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, USA

Niklaus J. Grünwald (PI)

Horticultural Crops Research Laboratory, USDA-ARS, Corvallis, Oregon, USA

Abstract: The increasing use of high-throughput sequencing of environmental samples (soil, water, animal gut contents, etc.) has resulted in large data sets containing community taxonomic information. However, the hierarchical nature of taxonomic information makes manipulating these data difficult. We are developing the rOpenSci package `taxa` to provide a set of all-purpose classes and manipulation functions for taxonomic data. To make it a solid foundation for the community to build on, we want to make it faster, more robust, and work to integrate it with select packages that will speed adoption. We are also working on `metacoder`, an R package for visualization and analysis of community taxonomic data. `Metacoder` is currently being significantly refactored to use the classes provided by `taxa` and more functionality is being added specific to community taxonomic data analysis. An award to this project would help facilitate the continued development of standards for community taxonomic information in R, the development of `metacoder` as an example of an R package that works with these standards, and dissemination of these new tools through presentations at conferences, journal articles, and online blogs.

Problem statement

Large data sets containing community taxonomic information are becoming increasingly common, but the hierarchical nature of taxonomic information makes manipulating these data sets difficult and many R packages made for this have similar, yet unique, data structures and functions. These differences make it difficult to use multiple packages on the same data set and causes programmers to spend time creating and maintaining analogous solutions for the same low-level problems of data storage and manipulation, rather than focusing on novel functionality. These inefficiencies are likely to get worse as the cost of high-throughput sequencing decreases, the interest in microbiome research increases, and the number of associated R packages increases. To address this problem, we are developing what we hope will become a standard in R for querying, storing, manipulating, and plotting taxonomic information and any other application-specific data associated with it. Being awarded this fellowship will allow us to devote the time needed to bring this project to maturity and help integrate it into the community of R programmers working with taxonomic data. A common infrastructure will allow programmers to make focused, novel contributions without creating redundant functionality. A cohesive community of packages will encourage users to conduct analyses entirely within R, increasing the effectiveness of reproducibility tools like `packrat` and R Markdown.

Proposed activities and outcomes

Background on `taxa`: We are working on developing the rOpenSci package `taxa`, in collaboration with Scott Chamberlain, to provide a set of all-purpose classes and associated manipulation functions for taxonomic data, modeled after the `dplyr` data-manipulation philosophy. In addition to classes that represent taxa, classifications, and taxonomic trees, the `taxa` packages has a class that stores any number of application-specific tables, lists, or vectors mapped to a taxonomic tree so that manipulations to the taxonomy are also applied to the corresponding data and vice versa. For example, when filtering taxa, the corresponding user-supplied data can be optionally discarded or reassigned to supertaxa that pass the filter. Manipulation functions take into account the hierarchical relationships between taxa as well. For example, when filtering taxa or associated data, the supertaxa or subtaxa of taxa that meet some condition can be preserved or discarded. The `taxa` package also implements extremely flexible parsers to read data in from nearly any format containing taxonomic information.

Goals related to `taxa`: Although the `taxa` package is already useful, to make it a solid foundation for the community to build on, we want to make it faster, more robust, and work to integrate it with select packages

that will speed adoption. By testing with very large data sets, we want to determine current limits and bottlenecks so that we can identify parts of code to refactor or replace with C++ to increase speed. We also want to implement parallel processing to make better use of modern multi-core computers. Since we hope to make this a foundational package, we want to add extensive unit tests in addition to the many already implemented, to minimize bugs and their effects on dependent packages. Perhaps most importantly, we want to work with the maintainers of popular packages using taxonomic data to integrate `taxa` and jump start adoption. We have just started the process of integrating `taxa` into `taxize` (Chamberlain and Szöcs 2013), the primary package for downloading taxonomic data from public databases, and `metacoder`, our package for analysis and visualization of community taxonomic data. Once `taxize` and `metacoder` are refactored to use `taxa`, database searching/downloading and flexible visualization will be easy using the classes implemented by `taxa`, providing useful tools for a wide variety of research objectives and R packages. This useful, all-purpose functionality will incentivize others to adopt the `taxa` package as a standard.

Background on `metacoder`: The other related project we are working on is `metacoder`, an R package for visualization and analysis of community taxonomic data, with an emphasis on providing tools for microbiome research (Foster, Sharpton, and Grünwald 2017). `Metacoder` introduced a way of visualizing statistics distributed throughout a hierarchy (e.g. a taxonomic tree) by quantitatively mapping statistics to the color or size of nodes and edges in a tree. We call these "heat trees" and they have been enthusiastically received as an intuitive and information-dense alternative to stacked bar charts, which is the most common way taxon abundance in communities are currently plotted in publications. The `heat_tree` function in `metacoder` is highly flexible, yet only requires a few arguments to produce high-quality figures in most cases, since raw statistics are automatically mapped to color/size and the size range of elements displayed are optimized for each graph to strike a balance between avoiding overlaps and maximizing the apparent differences. `Metacoder` also provides a way to simulate PCR in R, which is important for selecting primers for metabarcoding research.

Goals related to `metacoder`: `Metacoder` is currently being significantly refactored to use the classes provided by `taxa` and more functionality is being added specific to community taxonomic data analysis. We are considering splitting out the low-level plotting abilities of `metacoder` into its own package, since these are universally useful for any hierarchical data, not just community taxon abundance data from high-throughput sequencing. In that case, `metacoder` would focus exclusively on aiding analysis of community taxonomic data, generally from high-throughput sequencing. One particularly useful high-level visualization technique we want to facilitate is a matrix of unlabeled heat trees that show which taxa are deferentially abundant between every pair-wise set of treatments accompanied by a larger, labeled tree that functions as a key. This can be used to show the results of thousands of statistical tests and the treatments/taxa they apply to in a graph that can fit in a publication, such as figure 5 in Foster, Sharpton, and Grünwald (2017). Currently, making these graphs is quite complicated, but we believe we can design a function that will make creating these graphs easy for users. We also want to implement pairs of parser and writer functions that handle the common formats used in metabarcoding research to make it easy for users to move data in to and out of R. These parsers will return `taxa` objects, so they will be useful for all packages that adopt the `taxa` package as a standard. They could also be used to convert between file formats, since a user could read from one format and write to another. The `taxa` classes are flexible enough that we should be able to losslessly read and write the same format, so this would provide a way to subset or otherwise modify large, complicated files, using the powerful manipulation functions provided by `taxa`.

Goals related to community building and outreach: Encouraging adoption of this framework by both users and developers is key to the success of this project. We hope to facilitate adoption by developers by offering to help maintainers adapt existing packages to `taxa`. Once a critical mass of packages have adopted `taxa`, we expect that it will naturally be adopted by new packages without our direct assistance. To attract users, we want to submit a series of posts on R-bloggers that cover specific uses of the framework and demonstrate its usefulness. We also plan to present this work at biological and computational conferences, including the UseR 2018 conference and the rOpenSci 2018 unconference. We have already submitted a workshop proposal for the 2018 11th International Congress of Plant Pathology joint conference on reproducible analysis of microbiome data in R, which will focus on applying these tools. We hope to publish a paper on the `taxa` package in the F1000 open access journal. Finally, if this framework is adopted by the community as we hope, we would also want to publish an article describing the resources available for all-R analyses of taxonomic data using this framework.

Tentative timeline

Months 1-3: Improvements to the `taxa` package

1. Add missing dplyr analog functions (e.g. summarise, group_by, etc)
2. Port slowest parts of the code to C++ to improve speed using `Rcpp` (Eddelbuettel et al. 2011)
3. Add ability to use multiple cores to improve speed using `RcppParallel`
4. Add exhaustive unit tests (test coverage of $> 95\%$) using `testthat` (Wickham 2011) and codecov

Months 4-8: Improvements to `metacoder`

1. Finish the transition to using `taxa`
2. Add function to easily create a heat tree matrix for pairwise comparison of treatments
3. For each major file format used in metabarcoding research, add a function to parse the file's contents into the classes provided by `taxa` and a corresponding writer to recreate the file format. When read from and written to the same format, the process should be lossless.
4. Add automatic overlap avoidance for labels
5. Add support for plotting categorical information
6. Add ability to plot multiple colors per node/edge automatically when multiple values per taxon are supplied. Nodes would be pie charts and edges would be stacked bar charts in this case.
7. Add ability to automatically query, download, and plot silhouettes representing `taxa` from the pylopics database in place of nodes using the rOpenSci `rphylopic` package
8. Add ability to highlight/delineate groups of `taxa` with shaded polygons
9. Add ability to plot user-supplied images in place of nodes
10. Add support for interactive plots using `plotly`
11. Update and enhance our website containing the manual and tutorials for `metacoder`

Months 9-12: Outreach and community building

1. Contact the maintainers of major packages using taxonomic information and offer to help adopt `taxa`
2. Finish refactoring of `taxize` to use `taxa`
3. Present at the 2018 rOpenSci unconference
4. Present at the 2018 useR! conference
5. Conduct a workshop and reproducible microbiome analysis in R using these tool at the 2018 11th International Congress of Plant Pathology
6. Write a series of blogs for R-bloggers demonstrating the usefulness of the `taxa`/`taxize`/`metacoder` framework
7. Publish a journal article on `taxa` in F1000Research, or another open-access journal
8. Publish a journal article on reproducible microbiome analysis in R in an open-access journal

High level budget

Item	Cost	Units	Total
Airtavel for the useR! conference: Portland, Oregon, USA to Brisbane, AU	1,400		1,400
Per diem lodging for the useR! conference	159	4 nights	636
Per diem meals for the useR! conference	98	5 days	490
Registration for the useR! conference	125		125
Travel and registration for other conferences	2,500		2,500
Graduate student salary	2,000	12 months	24,000
Graduate student fringe benefits	6,000		6,000
Grand total			35,151

References

Chamberlain, Scott A, and Eduard Szöcs. 2013. "Taxize: Taxonomic Search and Retrieval in R." *F1000Research* 2. Faculty of 1000 Ltd.

Eddelbuettel, Dirk, Romain François, J Allaire, John Chambers, Douglas Bates, and Kevin Ushey. 2011. "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software* 40 (8): 1–18.

Foster, Zachary SL, Thomas J Sharpton, and Niklaus J Grünwald. 2017. "Metacoder: An R Package for Visualization and Manipulation of Community Taxonomic Diversity Data." *PLoS Computational Biology* 13 (2). Public Library of Science: e1005404.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3 (1): 5–10.