

Journal Entry #4

Yining Ding

15/10/2020

Objective:

The purpose of this journal is to draw down the goal of the R package I am going to develop in this course. Note that some minor details might change, but the general goal will remain the same.

Procedure

1. Wrote down the outline of the R package.
2. Search my idea on Google. Thought about what kind of new features my package will provide compared to existing packages.
3. Define the targeted audiences
4. Assumptions of the package
5. Think about the visualization my package will provide.
6. Number of functions my package will provide to the users.
7. Dependencies of my package
8. Name of my package and why.
9. Create a public GitHub repository and share the link
10. Share a brief overview of my package and the GitHub link on the discussion board.

Results

1. Provide an outline of the R package I would like to develop

I would like to develop a R package that provides visualizations of phylogenetic trees with multiple sequence alignment on the side through the R Shiny APP.

2. Are there similar packages that have already been developed? Do a search. What aspect of data analysis does your R package permit to perform, that has not been done to date?

Yes, there are many similar packages that provides phylogenetic tree visualizations on shiny apps. For example, Interactively view and subset phylogenetic trees But there is no package that includes the multiple sequence alignment (Bradley, 2019). I believe a MSA visaulization on the side will give the users more information about the phylogenetic tree, such as on which are the conserved amino acids between the homologs etc.

3. Who is the targeted audience of your package?

The targeted audiences are those who wish to have a quick overview of their tree and the msa. The users will not be able to adjust and create a publication level phylogenetic tree with msa.

4. What are some of the assumptions of your package?

I would assume that users provide their msa file and also the tree txt file (in certain format, I haven't decided yet but I think I will include the phylip format for now).

5. What type of a visual output might help perform the exploratory analysis of your package?

I would use ggplot and ggtree to provide basic visualization of the phylogenetic tree, and the msa function in the package to create a basic multiple sequence alignment visualization. On the tree, users will be able to adjust/rotate clades by selecting a certain node which I believe helps with tree analysis the most. Coloring and font size adjustment will also be available for users to mark some of the important homologs/species for example. If I have time, I will include a notes section where users can take notes on their trees and use it as a notebook.

6. Approximately how many functions would be available to user?

Approximately 6 functions will be available to the users. * Create tree and msa * Clade rotation on selected nodes * Change color on one node * Change color on one clade * Change font size on one node * Change font size on one clade * etc. (if time permits)

7. What other packages would your package depend on?

My package will be depending on ape (Paradis and Schliep, 2019), ggtree (Guangchuan et al, 2017), and Rphylip (Felsenstein, 2013).

8. What is the name of your package? Be sure to pick a meaningful name, with points discussed in Lecture 4. Describe the name you selected.

- phyMSAViewer
- “phy” describes the phylogenetic tree.
- “Viewer” describes that the users will be able to view the phylogenetic tree along with its multiple sequence alignment

9. Create a GitHub repository with the package name. Share the link. Your project will be housed in this repository.

phyMSAViewer. GitHub link: <https://github.com/helen307/phyMSAViewer>

Conclusions

From this assessment, I developed a general idea about the type of R package I am going to create in the next month. I will create a R package with visualization through R shiny app to provide the users with phylogenetic tree visualization with multiple sequence alignment. The extra feature, MSA could inform the users which amino acid or how many amino acid are conserved between two homologs for example. Details have not developed yet, but this general idea will not change. I searched on Google about my idea and I found many similar visualization applications, but none of them included MSA on the side. I named my package: phyMSAViewer. The name is not complicated, it directly informs the user that they can visualize a phylogenetic tree and a msa at the same time. I updated my GitHub name and also created a new repository: <https://github.com/helen307/phyMSAViewer>. I shared this link on the discussion board.

Next Steps

Timeline of package development:

- 1st week (10/24-10/25): I will review lecture 5 and lecture 6 to set up my package on the weekends.
- 2nd week (10/26-11/1): Then I plan to discover the dependency packages and try to come up with basic functions. On the shiny app, I could upload files and see the results of the tree and msa.
- 3rd week (11/2-11/8): I will add functions to the visualization, such as colors, rotations, etc.
- Initial submission: 11/11
- 4th week (11/9-11/15): Prepare presentation (due 11/17)

References

1. Bradley, T. (2019). Subsetting Phylogenetic Trees - goonR blog. Retrieved October 23, 2020, from <https://tbradley1013.github.io/2018/06/19/subsetting-phylogenetic-trees/>
2. Guangchuang Yu, David Smith, Huachen Zhu, Yi Guan, Tommy Tsan-Yuk Lam. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* 2017, 8(1):28-36
3. Guangchuang Yu, Tommy Tsan-Yuk Lam, Huachen Zhu, Yi Guan. Two methods for mapping and visualizing associated data on phylogeny using ggtree. *Molecular Biology and Evolution* 2018, 35(2):3041-3043. doi: 10.1093/molbev/msy194
4. Felsenstein, J. (2013) PHYLIP (Phylogeny Inference Package) version 3.695. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
5. Paradis E. & Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35: 526-528.