The Journal of Open Source Software

Physcraper: A Python package for continually updated gene trees

Luna L. Sanchez Reyes[1] **and Emily Jane McTavish**[1]

1 University of California, Merced **2** Institution 2

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

1. Phylogenies are still a key part of research in all areas of biology. The nonspecialist can use one of the many tools available to automatize phylogenetic reconstruction. However, interpretation of results, comparison with previously available phylogenetic hypotheses, and choosing of one phylogeny for discussion still impose difficulties to one that is not a specialist on phylogenetic methods or on a particular group of study.
2. Physcraper is an open-source Python program that automatizes the update of public phylogenies by making use of public DNA sequence data and taxonomic information.
3. Physcraper can be used by the nonspecialist, as a tool to generate phylogenetic hypothesis based on public expert phylogentic knowledge. The phylogenetics and group specialist will find it useful as a tool to facilitate comparison of alternative phylogenetic hypotheses (topologies). ***Is physcraper intended for the nonspecialist?? We have two types of nonspecialists: the ones that do not know about phylogenetic methods and the ones that might know about phylogenetic methods but do not know of a certain biological group.***
4. Physcraper implements node by node comparison of the the original and the updated trees using the conflict API of OToL.
5. We hope the physcraper workflow encourages researchers on their data sharing practices.
6. Physcraper can be used with any OS. Detailed instructions for installation and use are available at https://github.com/McTavishLab/physcraper.

Intro

There is a lot of phylogenetic knowledge already available. Some of it has been made public at OToL. Most or all public phylogenies have the potential to be updated with the ever increasing amount of genetic data that is available on GenBank. There are now various methods to automatize use of GenBank data for phylogenetic reconstruction, which are generally intended to facilitate the use of phylogenetic knowledge for the nonspecialist. However, none of them provide a means to compare with previously available expert knowledge on a particular group.

A key aspect of the standard phylogenetic workflow is comparison with already existing phylogenetic hypotheses and with phylogenies that are considered "best" by experts not only in phylogenetics, but also experts of the focal group of study.

It is well known that GenBank holds enormous amounts of genetic data, and it continues to grow. A lot of this genetic data has the potential to be used to reconstruct the phylogenetic history of various organisms (Sanderson, Boss, Chen, Cranston, & Wehe, 2008). Pipelines that harness this potential have been available for over a decade now, such as the Phylota brower, and PHLAWD. New ones keep on being developed, such as SUPERSMART and the upgraded version of PHLAWD, PyPHLAWD. Notably large phylogenies have been constructed using some of these tools, Some other have not been used that much. So, how well accepted is this approach in the community?

Concerns with these tools: Errors in identification of sequences Little control along the process Too much of a black box?

Most of these phylogenies are being constructed by people learning about the methods, so they want to know what is going on.

The pipelines are so powerful and they will give you an answer, but there is no way to assess if it is better than previous answers, it just assumes it is better because it used more data.

All these pipelines start tree consruction from zero?

The goal of Physcraper is to build upon previous phylogenetic knowledge, allowing a direct comparison of existing phylogenies to phylogenies that are constructed using new genetic data from GenBank

To achieve this, Physcraper uses the Open Tree of Life phylesystem and connects it to the TreeBase database, to (1) get the original DNA dataset matrices (alignments) that produced a phylogeny that was published and then made available in the OToL database, (2) use this DNA alignments as a starting point to get new genetic data belonging to the focal group of study, to (3) finally update the phylogenetic relationships in the group.

A less automated workflow is one in which the alignments that generated the published phylogeny are stored in other public database (such as DRYAD) or elsewhere (the users computer), and are provided by the users.

The original tree is by defualt used as starting tree for the phylogenetic searches, but it can also be set as a full topological constraint or not used at all, depending on the goals of the user.

Physcraper implements node by node comparison of the the original and the updated trees, using the conflict API of OToL.

Physcraper overview

Examples

Imagine you are starting to work on a new biological group X. You have not much of an idea about its phylogenetic relationships, you are a newly established researcher, and the group is not anything any of your collaborators have worked on before. A good idea is to start an intensive litt review on the phylogenetics of the group. Rapidly, you find out there are 5 different phylogeneis, that used different markers, and that the papers, published at different times, do not discuss which phylogeny is the one accepted by the expert community on X. You might need to go to the annual conference of X, and even then, you might only find different and contrasting opinions. Somewhere along these months or even years doing this task, you looked into the the OToL database. You found in there some or all the published trees of X, along with a tree that has been deemed the best tree by curators and ideally experts on X?

Let's be more specific now about our X group and say it is the Ascomycota. The best tree currently available in OToL was published by Schoch et al. (2009). The first step, is to get the Open Tree of Life study id. There are some options to do this: - You can go to the Open Tree of Life website and browse until you find it, or - you can get the study id using R tools: - By using the treebase ID of the study (which is not fully exposed on the Treebase website home page of the study, so you have to really look it up manually):

```
rotl::studies_find_studies(property = "treebaseId", value = "S2137")
##   study_ids n_trees         tree_ids candidate
## 1    pg_238       2 tree109, tree110
##   study_year title
## 1       2009
##                                  study_doi
## 1 http://dx.doi.org/10.1093/sysbio/syp020
```

- By using the name of the focal clade of study (but this behaved very differently):

```
rotl::studies_find_studies(property="ot:focalCladeOTTTaxonName", value="Ascomycota
```

Once we have the study id, we can gather the trees published on that study:

```
rotl::get_tree_ids(rotl::get_study_meta("pg_238"))
## [1] "tree109" "tree110"
rotl::candidate_for_synth(rotl::get_study_meta("pg_238"))
## NULL
my_trees <- rotl::get_study("pg_238")
```

Both trees from this study have 434 tips.

Let's check what the trees look like:

```
ape::plotPhylo(my_trees[[1]])
```

1. Download the alignment from Treebase If you are on the Treebase home page of the study, you can navigate to the matrix tab, and manually download the alignments that were used to reconstruct the trees reported on the study that were also uploaded to Treebase and to the Open Tree of Life repository. To make this task easier, you can use a command to download everything into your working folder:

```
auto_scrape.py -s pg_238 -t tree109 -o ../physcraper_example/treebase_alns
```

In this example, all alignments posted on Treebase were used to reconstruct both trees.

1. With the study id and the alignment files saved locally, we can do a physcraper run with the command:

```
opentree_scrape.py -s pg_238 -t tree109 -a treebase_alns/pg_238-LSU-M3956.nex -as
```

Tools on a similar track:

Tools that do similar things: pyPhlawd (Smith & Walker, 2019) SUPERSMART (Antonelli et al., 2017)

Acknowledgements

We acknowledge contributions from

References

Antonelli, A., Hettling, H., Condamine, F. L., Vos, K., Nilsson, R. H., Sanderson, M. J., Sauquet, H., et al. (2017). Toward a self-updating platform for estimating rates of speciation and migration, ages, and relationships of taxa. *Systematic Biology*, *66*(2), 152–166.

Sanderson, M. J., Boss, D., Chen, D., Cranston, K. A., & Wehe, A. (2008). The PhyLoTA Browser: Processing GenBank for Molecular Phylogenetics Research. *Systematic Biology*, *57*(3), 335–346. doi:10.1080/10635150802158688

Schoch, C. L., Sung, G.-H., López-Giráldez, F., Townsend, J. P., Miadlikowska, J., Hofstetter, V., Robbertse, B., et al. (2009). The ascomycota tree of life: A phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Systematic biology*, *58*(2), 224–239.

Smith, S. A., & Walker, J. F. (2019). PyPHLAWD: A python tool for phylogenetic dataset construction. *Methods in Ecology and Evolution*, *10*(1), 104–108.