DR STEPHEN SMITH (Orcid ID : 0000-0003-2035-9531)

Article type : Application

Handling editor: Dr Natalie Cooper

PyPHLAWD: a python tool for phylogenetic dataset construction

Stephen A. Smith[1*] and Joseph F. Walker[1]

[1] Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, 48103

[*] corresponding author: eebsmith@umich.edu

Summary

1. Comprehensive phylogenetic trees are essential for many ecological and evolutionary studies. Researchers may use existing trees or construct their own. In order to infer new trees, researchers often rely on programs that construct datasets from publicly available molecular data.

2. Here, we present PyPHLAWD, a phylogenetically guided tool written in Python that creates molecular datasets for building trees. PyPHLAWD constructs clusters (putative orthologs) that may be used for downstream analyses and provides users with a set of easy to interpret results. PyPHLAWD can conduct both baited (analyses that require the identification of gene regions *a priori*) and clustering analyses (analyses that do not require *a priori* identification of gene regions).

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/2041-210X.13096

3. PyPHLAWD is extensible, flexible, open source, and available at https://github.com/FePhyFoFum/PyPHLAWD, with a detailed website outlining instructions and functionality at https://fephyfofum.github.io/PyPHLAWD/.

4. The utility of PyPHLAWD is highlighted here with an example workflow for the plant clade Dipsacales and may be applied to any clade with publicly available data on GenBank.

Key-words: phylogenetics, GenBank, sequences, tree of life, Python

Introduction

Large phylogenies have become important for addressing outstanding questions in comparative ecological and evolutionary studies. While resources are available that offer broadly sampled taxonomies and phylogenies (e.g., Open Tree of Life), for many reasons, researchers may want to process the available molecular data for the construction of targeted phylogenies.

Several programs have been developed to process publicly available GenBank data in order to create phylogenetic datasets including SUPERSMART (Antonelli et al. 2017), PHLAWD (Smith et al. 2009), Phylota (Sanderson et al. 2008), Phylogenerator (Pearse and Purvis 2013), among others. These programs can serve as complements/verification of each other, along with providing utility that offers greater flexibility in analyses. While automation can simplify many steps, recently, it has been shown that human intervention in many of the steps improves the quality of the resulting phylogeny (Beaulieu and O'Meara 2018; Smith and Brown 2018). Here, we describe PyPHLAWD, a Python package providing a diverse set of tools for constructing datasets using GenBank in a semi-automated fashion.

PyPHLAWD offers users flexibility throughout an analysis allowing for researchers to intervene at any step. The program implements both baited (Smith et al. 2009) and a tip-to-root clustering method (Smith and Brown 2018; Walker et al. 2018) for thorough and rapid orthology identification. PyPHLAWD is an extensible and flexible tool written in Python that can be easily incorporated into different workflows.

PyPHLAWD workflow

PyPHLAWD is an open source, freely available, set of scripts written in Python. The source code and instructions are available from https://github.com/FePhyFoFum/PyPHLAWD. PyPHLAWD requires a few, easy to install, dependencies including several programs from the `phyx` (Brown et al. 2017) package, `ncbi-blast+` (Altschul et al. 1997), `mcl` (Van Dongen 2000), `mafft` (Katoh and Standley 2013) and several python packages (`sqlite`, `networkx`, `clint`). For increased speed, PyPHLAWD can optionally be compiled with `cython`. Other programs, if installed, may be used such as `FastTree` (Price et al. 2010). All of these are freely available, and many through software repositories. Installation and examples are available from the website https://fephyfofum.github.io/PyPHLAWD/.

PyPHLAWD can be used unsupervised or with human intervention to investigate specific results. It relies on a pre-assembled database of sequences that may be created using `phlawd_db_maker` (https://github.com/blackrim/phlawd_db_maker) or downloaded from resources provided on the program website. PyPHLAWD allows a user to conduct both clustering analyses, without specifying the genes of interest, as well as a baited analysis.

A baited analysis in PyPHLAWD is done by running `setup_clade_bait.py` and providing a directory of *bait* sequence files that can be used to filter sequences from the database. A bait sequence file for a single gene should contain full length sequence for several species (e.g., 10-20) across the clade of interest. This was the standard analysis mode for PHLAWD (Smith et al. 2009) and can be useful when there are a few well-known genes sampled for a particular clade. Unlike PHLAWD, however, PyPHLAWD will conduct an analysis on each file (i.e., gene) in the bait directory instead of the user having to conduct an analysis on each gene separately.

In contrast to a baited analysis, a clustering analysis does not require a set of identified bait sequences. Instead, all genes available for the user specified taxonomic group will be identified and summarized as a user-friendly html file. During the process all-by-all BLAST analyses along with markov clustering analyses, MCL, are conducted to identify clusters. In PyPHLAWD, a clustering analysis is conducted by running `setup_clade.py`.

Whether baited or clustered, once gene regions are identified, the remaining analyses are the same (Fig 1). A set of directories representing the taxonomic hierarchy of the taxon of interest are created. Within each of these directories, the sequences representing the taxon of the directory are placed in a file. By default, whole genomes are excluded from analyses but are placed in a file in the relevant directory. Sequence similarity analyses are conducted from the tip to the root of the folder hierarchy. Because the sequence similarity analyses themselves do not impose a taxonomic constraint, only the order of analyses, it is unlikely that bias will be introduced by a potentially incorrect taxonomy. However, if there are egregious errors in the taxonomy, they may be edited within the database.

There are several parameters that can be edited to customize analyses. For example, sequences associated with misidentified taxa can be excluded by adding their taxonomic or sequence id to the files `bad_seqs.py` or `bad_taxa.py`. If the user wants to exclude sequences based on patterns in the sequence description (e.g., particular collections of sequences), these patterns can be added to the `exclude_patterns.py` or the `exclude_desc_patterns.py` files.

PyPHLAWD has a general configuration file, `conf.py`, that defines several parameters including user options like `smallest_size` meant to define the smallest sequences to exclude as well as several parameters related to the BLAST analyses (e.g., `length_limit`, `evalue_limit`, `evalue_limit_lc`, `perc_identity`). PyPHLAWD functions as a set of scripts instead of a python library installed at the

system level to encourage users that may be less familiar with library development to add functionality or edit scripts. Therefore, the `conf.py` file also requires that the installation directory of PyPHLAWD be specified (as this allows PyPHLAWD to locate the other scripts in the package).

Demonstration

We performed two demonstration analyses, a baited and a clustering analysis. Both analyses were performed on the plant clade Dipsacales. These runs were conducted on a 3.9Ghz quad-core laptop with 16 GB RAM and running Xubuntu Linux (kernel 4.16). For both analyses, we set `smallest_size = 400`, `length_limit = 0.55`, and `perc_identity = 20`. We used a copy of GenBank for the `pln` division built with `phlawd_db_maker` (website Smith et al. 2009) constructed on 04/11/18. For baited analyses, we used rbcL, trnL-trnF, ITS, and matK as bait (files available as part of the examples in the distributed source code). For clustering analyses, we conducted analyses with the above parameters.

Runs for similar methods

In order to compare the performance of PyPHLAWD, we also conducted analyses using PHLAWD, Phylota, SUPERSMART, and Phylogenerator. For PHLAWD, we used the parameters `mad = 0.05`, `coverage = 0.55`, and `identity = 0.2` using the same bait as the PyPHLAWD baited runs. We examined Phylota results as available on 05/01/18 using GenBank release 194. For SUPERSMART, we conducted runs, without divergence time estimation, on the plant clade Adoxaceae, within the Dipsacales, as runs conducted on the entire Dipsacales failed to successfully construct some elements of the tree. We, therefore, compared the Adoxaceae results from PyPHLAWD to those of SUPERSMART. Phylogenerator was tested using Dipsacales and searching for the genes "rbcL", "matK", "ITS" (alias: "internal transcribed spacer 1", fussy: false), and trnLF (alias: "trnL-trnF intergenic spacer", fussy: false). The Hawkeye sequence adjustment method was turned off and the minimum reference length was set to 400bp.

Results of verification

For the Dipsacales, there are 33,178 sequences available representing 864 taxa (as of 04/11/18).

Baited analysis

We constructed the bait sequences by manually downloading 20 representative sequences for each of the target genes on GenBank. The baited analysis took 52 seconds and PyPHLAWD found 299 taxa with rbcL, 333 with trnL-trnF, 553 with ITS, and 322 with matK. We constructed a phylogenetic dataset with these resulting gene regions that resulted in 641 taxa (173 taxa in Adoxaceae). We constructed a

phylogeny using RAxML v. 8.2.11 (Stamatakis 2014) with the `-f a` option generating 100 rapid bootstraps with GTR+G model of evolution and partitioned by gene region.

Clustered analysis

The clustered analysis took 5 minutes and 16 seconds and resulted in 684 taxa represented in 617 clusters. The largest clusters included ITS (559 taxa), trnLF (334 taxa), matK (333 taxa), and rbcL (299 taxa) with others containing between 268 and fewer taxa (Fig. 1). We constructed a phylogenetic dataset with the `find_good_clusters_for_concat.py` (with at least 20% taxon sampling and at least 20 taxa) script that identified 16 clusters representing 672 taxa (173 taxa in Adoxaceae). We constructed a phylogeny using RAxML v. 8.2.11 with the `-f a` option generating 100 rapid bootstraps with GTR+G model of evolution and partitioned by gene region. The `find_good_clusters_for_concat.py` script can construct a constraint tree. However, this was not used in the construction of the phylogeny presented here.

Results from other methods

Phylota recovered 536 ITS, 330 matK, 354 trnLF, and 242 rbcL taxa for Dipsacales with a concatenated dataset of 581 taxa (155 in Adoxaceae). PHLAWD baited analyses recovered 659 ITS, 329 matK, 349 trnLF, and 295 rbcL sequences for Dipsacales with a concatenated dataset of 731 taxa (201 in Adoxaceae). Run-times were between 16 and 40 seconds. SUPERSMART recovered for 147 taxa for Adoxaceae with a run-time of ~4 minutes to perform the taxize, align, orthologize, and bbmerge steps. The total run time for Phylogenerator was between 32 minutes and 1.3 hours depending on whether alignment and reconstruction was done and resulted in a phylogeny consisting of 627 tips and 347 matK, 298 rbcL, 325 trnLF, and 566 ITS sequences found.

Discussion

PyPHLAWD is a package that can be used for constructing phylogenetic datasets with molecular data. In addition to conducting baited analyses, as in PHLAWD, it also allows for clustering sequences without bait. PyPHLAWD constructs clusters rapidly by only conducting all-by-all comparisons for the most nested taxonomic units (e.g., genera) and then merging these clusters using BLAST. Additionally, PyPHLAWD is written as a set of scripts to allow for more flexibility for the user.

Comparison to other methods

We compared the results of baited and clustering analyses to those from PHLAWD, Phylota, SUPERSMART, and Phylogenerator. PyPHLAWD baited analyses are different than PHLAWD analyses in that PyPHLAWD uses the BLAST algorithm instead of the Smith-Waterman algorithm for sequence homology analyses. As a result, the specific parameters used in each analysis are not comparable.

Nonetheless, given that these parameter values are typical for many analyses, the *results* are comparable. PyPHLAWD and PHLAWD baited results were similar with differences reflecting the differences between BLAST and Smith-Waterman and the specific cutoffs used (e.g., sequence length, overlap). PHLAWD recovered slightly fewer sequences of rbcL and matK and more sequences of trnLF and ITS. In both cases, where more sequences were recovered, it was the result of PHLAWD allowing the retrieval of shorter sequences. Including short sequences in the analysis can increase error in alignment and so are excluded from PyPHLAWD by default. The phylogenies produced from the PyPHLAWD datasets had fewer long branches (i.e., from alignment error) than those from PHLAWD.

Instead of comparing all of the clusters found by Phylota, we compared the results for ITS, matK, rbcL, and trnLF from the PyPHLAWD clustering analyses to the equivalent clusters available through Phylota. There were several differences that were due to the differences in the GenBank releases used as Phylota covers release 194 while PyPHLAWD used release 224. PyPHLAWD also uses a more stringent sequence length cutoff. For example, PyPHLAWD recovered 559 taxa with ITS and Phylota, which includes much smaller sequences (e.g., 154 bp the case of ITS), recovered 536 taxa. Similar differences were found for other gene regions.

Phylogenerator and PyPHLAWD baited analyses differ in the method of orthologous sequence detection. PyPHLAWD recognizes sequences based on sequence similarity, whereas Phylogenerator uses sequence label matching. Phylogenerator requires the user to specify the alias of a sequence and relies on proper curation of deposited samples. It, however, does not require a local database and therefore downloads sequences at the time the user runs the program. This difference in the time required for sequence retrieval helps explain the overall differences in runtime. Otherwise, the results were comparable with fewer taxa per gene in certain cases and more in others.

SUPERSMART provides a comprehensive package for the construction of trees from GenBank data. Despite some similarities, there are several differences between SUPERSMART and PyPHLAWD. First, SUPERSMART is run as a virtual machine, allowing for fewer steps in installing dependencies. However, this comes at the cost of flexibility. There are also methodological differences. For example, SUPERSMART relies on initial clusters constructed from Phylota that it then verifies with BLAST. PyPHLAWD uses a database created by the user that may be much more up to date. SUPERSMART conducts all analyses required for dataset construction, phylogenetic reconstruction, and divergence time estimation, presenting the user with the final results. Intermediate outputs are obfuscated to the user. PyPHLAWD expects human intervention at any stage. Each output can be used for other analyses. For example, PyPHLAWD provides the clustered genes, aligned and unaligned, in folders based on taxonomy. The detailed summary of the output given by PyPHLAWD gives users greater ease to conduct an array of downstream analyses on the clades.

We do not aim to criticize other programs and procedures that may be useful and available to the community. Instead, we feel that these programs are complimentary and may serve different purposes for diverse goals and analyses.

On baited vs. clustering analyses

As shown here, it is possible for baited and clustering analyses to produce different results with PyPHLAWD. There are several reasons for this. For example, the behavior of the baited analysis may change depending on the quality, size, and taxonomic coverage of the user identified bait files. Alternatively, clustering runs will only be dependent on the available sequences in the source database. A user should choose the analysis that is the most reasonable given their goals. When common gene regions for a clade are unknown, clustering analyses are preferable.

Authors' contributions

JFW and SAS developed the procedure. SAS wrote the code. JFW and SAS conducted the analyses and wrote the manuscript.

Acknowledgments

We are grateful to early adopters and testers.

Data accessibility

All code is open source and available from at Smith and Walker (2018), doi:10.5281/zenodo.1400789, and https://github.com/FePhyFoFum/PyPHLAWD. A web manual is also available at https://fephyfofum.github.io/PyPHLAWD/.

Supporting Information

SAS was supported by NSF 1458466 and 1207915. JFW was supported by NSF 1354048.

References

Altschul, Stephen F, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. 1997. “Gapped Blast and Psi-Blast: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research* 25 (17). Oxford University Press: 3389–3402.

Antonelli, Alexandre, Hannes Hettling, Fabien L Condamine, Karin Vos, R Henrik Nilsson, Michael J Sanderson, Hervé Sauquet, et al. 2017. “Toward a Self-Updating Platform for

Estimating Rates of Speciation and Migration, Ages, and Relationships of Taxa." *Systematic Biology* 66 (2). Oxford University Press: 152–66.

Beaulieu, Jeremy M., and Brian C. O'Meara. 2018. "Can We Build It? Yes We Can, but Should We Use It? Assessing the Quality and Value of a Very Large Phylogeny of Campanulid Angiosperms." *American Journal of Botany* 105 (3): 417–32. doi:10.1002/ajb2.1020.

Brown, Joseph W, Joseph F Walker, and Stephen A Smith. 2017. "Phyx: Phylogenetic Tools for Unix." *Bioinformatics* 33 (12). Oxford University Press: 1886–8.

Katoh, Kazutaka, and Daron M Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30 (4). Society for Molecular Biology; Evolution: 772–80.

Pearse, William D, and Andy Purvis. 2013. "PhyloGenerator: An Automated Phylogeny Generation Tool for Ecologists." *Methods in Ecology and Evolution* 4 (7). Wiley Online Library: 692–98.

Price, Morgan N, Paramvir S Dehal, and Adam P Arkin. 2010. "FastTree 2–Approximately Maximum-Likelihood Trees for Large Alignments." *PloS One* 5 (3). Public Library of Science: e9490.

Sanderson, Michael J, Darren Boss, Duhong Chen, Karen A Cranston, and Andre Wehe. 2008. "The Phylota Browser: Processing Genbank for Molecular Phylogenetics Research." *Systematic Biology* 57 (3). Taylor & Francis: 335–46.

Smith, Stephen A, and Joseph F. Walker. 2018. "PyPHLAWD: Release V.1.0." doi:10.5281/zenodo.1400789.

Smith, Stephen A, Jeremy M Beaulieu, and Michael J Donoghue. 2009. "Mega-Phylogeny Approach for Comparative Biology: An Alternative to Supertree and Supermatrix Approaches." *BMC Evolutionary Biology* 9 (1). BioMed Central: 37.

Smith, Stephen A., and Joseph W. Brown. 2018. "Constructing a Broadly Inclusive Seed Plant Phylogeny." *American Journal of Botany* 105 (3): 302–14. doi:10.1002/ajb2.1019.

Stamatakis, Alexandros. 2014. "RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies." *Bioinformatics* 30 (9). Oxford University Press: 1312–3.

Van Dongen, Stijn Marinus. 2000. “Graph Clustering by Flow Simulation.” PhD thesis.

Walker, Joseph F., Ya Yang, Tao Feng, Alfonso Timoneda, Jessica Mikenas, Vera Hutchison, Caroline Edwards, et al. 2018. “From Cacti to Carnivores: Improved Phylotranscriptomic Sampling and Hierarchical Homology Inference Provide Further Insight into the Evolution of Caryophyllales.” *American Journal of Botany* 105 (3): 446–62. doi:10.1002/ajb2.1069.

Figure Captions

Figure 1. PyPHLAWD procedure with an example from the plant clade Dipsacales. The taxonomy on the left illustrates the clustering order and the folder structure generated from PyPHLAWD. The baited analysis panel demonstrates that sequences will be compared to bait sequences. Clustering is conducted from tip-to-root as shown in Adoxaceae. Supermatrices may be constructed (though not required) from the resulting clusters or bait sequences. A phylogeny may be built from these supermatrices (right panel) or from individual gene regions.