Physcraper: a python package for continual update of evolutionary estimates using the Open Tree of Life

1. Luna L. Sanchez Reyes

School of Natural Sciences, University of California, Merced

email: sanchez.reyes.luna@gmail.com

2. Martha Kandziora

School of Natural Sciences, University of California, Merced

email: martha.kandziora@mailbox.org

3. Emily Jane McTavish

School of Natural Sciences, University of California, Merced

email: ejmctavish@gmail.com

Correspondence address: Science and Engineering Building 1, University of California, Merced, 5200 N. Lake Rd, Merced CA 95343

Correspondence email: sanchez.reyes.luna@gmail.com, ejmctavish@gmail.com

Running title: Continually updated gene trees with Physcraper

Word count: 3340

Manuscript prepared for submission to Methods in Ecology and Evolution

Article type: Application

Abstract

1. Phylogenies are a key part of research in many areas of biology. Tools that automatize some parts of the process of phylogenetic reconstruction (mainly character matrix construction) have been developed for the advantage of both specialists in the field of phylogenetics and nonspecialists. However, interpretation of results, comparison with previously available phylogenetic hypotheses, and choosing of one phylogeny for downstream analyses and discussion still impose difficulties to one that is not a specialist either on phylogenetic methods or on a particular group of study.

2. Physcraper is an open-source, command-line Python program that automatizes the update of published phylogenies by enriching underlying gene alignments with public DNA sequence data, and linking taxonomic information across databases. This provides a framework for comparison of published phylogenies with their updated versions, by using the conflict Application Programming Interface (API) function of the Open Tree of Life project.

3. Physcraper can be used by the nonspecialist, as a tool to generate phylogenetic hypothesis based on already available expert phylogenetic knowledge. Phylogeneticists and group specialists will find it useful as a tool to facilitate dataset gathering and comparison of alternative phylogenetic hypotheses (topologies).

4. We hope that the Physcraper workflow demonstrates the benefits of opening results in phylogenetics, encouraging more researchers to strive for better data sharing practices. Physcraper can be used with any OS. Detailed instructions for installation and use are available at https://physcraper.readthedocs.

Keywords: cross-connectivity, gene tree, open science, open tree of life, phylogeny, public database, python, reproducibility, taxonomy, updated alignment

1 Introduction

Public biological databases provide an amazing resource for evolutionary estimation. By linking molecular data, available from databases such as the GenBank (Benson *et al.* 2000; Wheeler *et al.* 2000), to alignments and phylogenies, available in the TreeBASE repository (Piel *et al.* 2009) and the Open Tree of Life datastore (McTavish *et al.* 2015), we can place new biological data in an evolutionary context. However, connecting new sequence data with existing estimates of evolutionary relationships represents a challenge.

More than a decade ago, the National Center for Biodiversity Information (NCBI) molecular database, GenBank, released its version number 159 (April 15, 2007). With 72 million DNA sequences, it was estimated to have the potential to resolve evolutionary relationships of most of the 241 000 distinct taxa represented in it (about 98.05% of taxa in the NCBI taxonomy; Sanderson *et al.* 2008), which in turns covers about 10% of extant described biodiversity (taking a conservative estimate of extant diversity; Scott 2011; Federhen 2003). In comparison, the current GenBank release number 238 (June 15, 2020) has tripled in size, hosting data for more than 217 million DNA sequences (See GenBank's release website). Yet, publicly available estimates of phylogenetic relationships still cover only about 90 000 taxa [Hinchliff *et al.* (2015); CURRENT SYNTH TREE CITATION], covering less than one third of the taxonomic diversity with data available a decade ago. Some of this discrepance is because that many phylogenies are not publicly shared (Drew *et al.* 2013; Magee *et al.* 2014, @mctavish2018bioessay). However, most recently published large trees have been made available in recent years, indicating there is also a lag between the amount of new DNA data generated and the analysis of these data in a phylogenetic context.

Many useful computational tools have been developed in an effort to make sense of the large amount of data in public molecular databases, as well as private ones. Generally referred to as "pipelines", most of these tools were motivated by the genomics revolution, to help researchers to deal with the massive amount of data, and automatically identify clusters of homologs for genomic assembly. Notably, this homolog DNA clusters can be used as homology hypotheses (i.e., molecular alignments) to reconstruct phylogenetic relationships. Pipelines that automatize the assembly of DNA alignments from the GenBank database for phylogenetic reconstruction ("phylogenetic pipelines") such as PHYLOTA (Sanderson *et al.* 2008), PHLAWD (Smith *et al.* 2009), and

SUPERSMART (Antonelli *et al.* 2017), have been widely applied to study the evolutionary relationships among different organisms (TABLE?? maybe supplementary data), from a phylogeny of no more than 20 species of the family of barracuda fish (Sphyraenidae; Santini & Sorenson 2013), to a mega-phylogeny of almost 3 000 species of living ferns (Moniloformopses; Lehtonen 2011).

Automated sequence gathering pipelines have been an important incorporation to the field of phylogenetics in many ways, particularly because they represent a clear step towards reproducibility in the field. In contrast, most published phylogenies to date have been inferred using alignments that have been assembled and curated "by hand" (Morrison 2009). Automated approaches often suffer from errors in homology inference that could rapidly caught by human oversight. CITATION Many systemacists support a classic phylogenetics approach (few markers thoughtfully curated) over the genomics approach (a massive amount of DNA markers that will overcome potential errors in the alignment coming from a lack of human curation). Species tree reconstructions from multi-gene data sets taking into account the multispecies coalescent model are considered the gold standard for inferring species relationships [Song *et al.* (2012); ROJAS ET AL.]. It has also been suggested that manual curation of locus alignments produces better phylogenetic reconstructions and this has been demonstrated for genomic alignments (Fragoso-Martínez *et al.* 2017).

A way to incorporate the best of two worlds (massive amounts of newly released molecular data AND fine-grained curation from human experts) is be to rely on published manually curated homology hypotheses as "jump-start" alignments (Morrison 2006). This expert-curated alignments can be continuously enriched and updated by incorporating newly released data from public molecular databases.

As of April 2014, the TreeBASE repository hosted about 8 200 curated alignments, providing information on evolutionary relationships of almost 105 000 distinct taxa (see TreeBASE's website about). This database provides an untapped source of valuable expert knowledge with the potential to update phylogenetic relationships in several different regions of the tree of life.

The OpenTree tree repository (phylesystem; McTavish *et al.* 2015) automatically incorporates phylogenies uploaded into TreeBASE, and stores metadata linking the tree to its corresponding alignment repository in

TreeBASE. However, if there are multiple alignments, TreeBASE does not indicate how they were used to generate the tree. This provides a loose means of linking the tree with the exact alignment that generated it.

Often linking data in an original alignment with its corresponding phylogeny has to be done by a human curator. Moreover, different data repositories follow different systems for taxon and study identification, posing a real challenge to automatically link data from across databases that belong to the same taxon and study. OpenTree's metadata system incorporates taxon identifiers from a variety of taxonomies and repositories, including the NCBI taxonomy, GBIF, etc., providing a way to automatically link data from different databases.

Physcraper is a python pipeline that relies on the OpenTree metadata system to connect databases through taxon identification numbers. It's main functionality is to connect phylogenies stored in the OpenTree phylesystem, with alignments from TreeBASE and newly released DNA data from GenBank, in order to update previously known phylogenetic relationships in a continuous manner. Because of its design, it allows taking advantage of the many resources provided by the OpenTree. For example, it allows automatizing and standardizing the comparison of phylogenetic hypotheses with currently known relationships.

This is an effort to keep on directing ourselves towards a fully reproducible workflow in phylogenetics. And an effort to more effectively make big data connections for the advantage of phylogenetics and biology in general.

2 How does Physcraper work?

2.1 The input: a tree and an alignment

- In order to take advantage of the OpenTree tools, the input tree needs to be stored in the OpenTree phylesystem github.com/opentreeoflife/phylesystem. The main reason for this is that trees in phylesystem have a set of user defined characteristics that are essential for automatizing the phylogeny update process. The most relevant of these being the standardization of taxonomic names and the definition of ingroup and outgroup. Outgroup and ingroup taxa in the original tree are identified and tagged. This allows to automatically set the root for the updated tree on the final steps of the pipeline. Currently, only trees connected to a published study can be stored in the phylesystem (although there are plans to allow storage of unpublished trees). A user can choose from among the 1216 published trees supporting the resolved nodes of the synthetic tree in the OpenTree website (See OpenTree's website about). If the published tree you are interested in updating is not in there, you can upload it via OpenTree's curator tool (www.opentreeoflife.org/curator.

- The alignment should be a gene alignment that was used to generate the tree. The original alignments are usually stored in a public repository such as TreeBase (Piel *et al.* 2009; Vos *et al.* 2012), DRYAD (www.datadryad.org), or the journal were the tree was originally published. If the alignment is stored in TreeBase, `physcraper` can download it directly, either from the TreeBASE website (www.treebase.org) or through the TreeBASE GitHub repository (SuperTreeBASE; github.com/TreeBASE/supertreebase). If the alignment is on another repository, or constitutes personal data, a path to a local copy of the alignment has to be provided.
- A taxon name matching step is performed to verify that all taxon names on the tips of the tree are in the DNA character matrix and vice versa.
- A ".csv" file with the summary of taxon name matching is produced for the user.
- Unmatched taxon names are dropped from both the tree and alignment. Technically, just one matching name is needed to perform the searches. Please, see next section.
- A ".tre" file and a ".fas" file containing only the matched taxa are generated and saved in the `inputs` folder to be used in the following steps.

2.2 DNA sequence search and filtering

- Technically, any DNA molecular database can be used to search for new sequences. By default we rely on the GenBank database. The new sequence search can be performed on the remote database or in a local database.

- The next step is to identify the search taxon within the reference taxonomy. The search taxon will be used to constraint the DNA sequence search on the nucleotide database within that taxonomic group. Because we are using the NCBI nucleotide database, by default the reference taxonomy is the

NCBI taxonomy. The search taxon can be determined by the user. If none is provided, then the search taxon is identified as the Most Recent Common Ancestor (MRCA) of the matched taxa belonging to the ingroup in the OPenTree synthetic tree, that is also a named clade in the reference taxonomy. FIGURE RECOMMENDED. This is known as the Most Recent Common Ancestral Taxon (MRCAT; also referred in the literature as the Least Inclusive Common Ancestral Taxon - LICA) CITATION NEEDED. The MRCAT can be different from the phylogenetic MRCA when the latter is an unnamed clade in the reference taxonomy. To automatically identify the MRCAT of a group of taxon names, we make use of the OpenTree taxonomic tool Taxonomy-API-v3 (Rees & Cranston 2017).

Users can provide a search taxon that is either a more or a less inclusive clade relative to the ingroup of the original phylogeny. If the search taxon is more inclusive, the sequence search will be performed outside the MRCAT of the matched taxa, e.g., including all taxa within the family or the order that the ingroup belongs to. If the search taxon is a less inclusive clade, the users can focus on enriching a particular clade/region within the ingroup of the phylogeny.

- The Basic Local Alignment Search Tool, BLAST (Altschul *et al.* 1990, 1997) is used to identify similarity between DNA sequences within the search taxon in a nucleotide database, and the sequences on the checked alignment. The `blastn` function from the BLAST command line tools (Camacho *et al.* 2009) is used for local-database sequence searches. For remote-database searches, we modified the BioPython (Cock *et al.* 2009) BLAST function in the NCBIWWW module to accept an alternate BLAST URL. This is useful when a user has no access to the computer capacity needed to setup a local database, and a local blast database can be set up on a remote machine to BLAST without NCBI's required wait times.

- A pairwise BLAST search is performed. This means that each sequence in the alignment is BLASTed against DNA sequences in a nucleotide database constrained to the search taxon. Results from each one of these BLAST runs are recorded, and matched sequences are saved along with their corresponding identification numbers (accession numbers in the case of the GenBank database). This information will be used later to store the whole sequences in a dedicated library within the physcraper folder, allowing for secondary analyses to run significantly faster.

- Matched sequences will be discarded if the fall below a default e-value of 0.00001, and outside a default minimum and maximum length of 80% and 120%, respectively, of the average length of sequences in the checked alignment (gaps dropped). These parameters can be configured for each run. This filtering guarantees that whole genome sequences are not included, as they All accepted sequences are assigned an internal identifier, and are further filtered.

- New sequences that are identical to existing sequences, or to subsets of an existing sequence are discarded, unless they reperesent a different taxon than the existing sequence. Longer sequences belonging to the same taxon as the original sequence will be considered further for analysis.

- Among the filtered sequences, there are often several representatives per taxon. Although it can be useful to keep some of them, for example, to investigate monophyly within species, there can be hundreds of exemplar sequences per taxon for some markers. To control the number of sequences per taxon in downstream analyses, 5 sequences per taxon are chosen at random. This number is set by default but can be modified by the user.

- Reverse complement sequences are identified and translated.

- Users can choose to perform cycles of sequence similarity search iteratively, by blasting the newly found sequences until no new sequences are found. By default only one BLAST cycle is performed and only sequences in the checked alignment are blasted.

- Accepted sequences are then downloaded in full, and stored as a local database in a directory that is globally accessible (default to physcraper/taxonomy), so they are accessible for further runs.

- A fasta file containing all filtered and processed sequences resulting from the BLAST search is generated for the user, and is used as an input for alignment.

2.3 DNA sequence alignment

- The software MUSCLE (Edgar 2004) is used by default to perform sequence alignments.
- First, all new sequences are aligned using default MUSCLE options.
- Then, a MUSCLE profile alignment is performed, in which the original alignment is used as a template to align new sequences. This ensures that the final alignment follows the homology criteria established by the original alignment.
- The final alignment is not further processed automatically. So, we encourage users to check it either by eye and perform manual refinement or using any of the many tools for alignment processing, to eliminate columns with no information.
- Users may also use physcraper only to gather sequence matches, and apply their own preferred alignemnt and phylogenetic inference tools.

2.4 Tree reconstruction and comparison

- A gene tree is reconstructed for each alignment provided, using a Maximum Likelihood approach implemented with the software RAxML (Stamatakis 2014), using default settings such as a GTRCAT model of molecular evolution and 100 bootstrap replicates with default method. Currently only the number of bootsrap replicates can be modified by the user.
- The original tree is used as starting tree for the ML searches. It can also be set as a full topological constraint or not used at all, depending on the goals of the user.
- Bootstrap results are summarized with Dendropy (Sukumaran & Holder 2010).
- The final result is an updated phylogenetic hypothesis for each of the genes provided in the alignment.
- Tips on all trees generated by physcraper are defined by a taxon name space, capturing the NCBI accession information, as well as the taxon identifiers, allowing the user perform comparisons and conflict analyses.
- Robinson Foulds weighted and unweighted metrics between the tips in the input tree and those tips in the updated tree are calculated with Dendropy functions (Sukumaran & Holder 2010).
- Finally a conflict analysis is performed. This is basically a node by node comparison between the the

synthetic OpenTree and the original and updated tree individually. This is performed with OpenTree's conflict Application Programming Interface (Redelings & Holder 2017).

- For the conflict analysis to be meaningful, the root of the tree needs to be accurately defined.
- A suggested default rooting based on OpenTree's taxonomy is implemented for now. This approach uses the taxon labels for all the tips in the updated tree, pulls an inferred subtree from OpenTree's taxonomy and then applies the same rooting to the inferred updated tree. However, if the updated tree changes expectations from taxonomy, the root may no longer be appropriate. Automatic identification of a phylogenetic tree root is indeed a difficult problem that has not been solved yet. The best way right now is for users to define outgroups so trees are accurately rooted.

3 Examples

We will illustrate the utility of physcraper in here with two use-case scenarios. One in which the user is interested in a particular group. Another one in which the user is interested in a particular phylogeny. A tutorial as well as illustrated examples of commands for every step needed to perform a physcraper analysis are available elsewhere.

3.1 The hollies

A student is interested in the genus *Ilex*, the only extant clade within the family Aquifoliaceae, order Aquifoliales of flowering plants. The genus encompasses between 400-600 living species. A review of literature reveals that there are three published phylogenetic trees showing relationships within the hollies. The first one has been made available in TreeBASE as well as in the OpenTree phylesystem and is part of the synthetic tree. It samples 48 species. The second tree has not been made available anywhere, not even in the supplementary data of the original publication. The most recent one has been made available in the OpenTree Phylesystem and in the DRYAD repository. It is the best sampled yet, with 200 species. However, it has not been added to the syntehtic tree yet. This also makes it a perfect case to test the basic functionalities of physcraper: we know that the sequences of the most recently published tree have been made available on the GenBank database. Hence, we expect that updating the oldest tree will produce something very similar to the newest tree.

DESCRIBE RESULTS: SUMMARY OF NEW TAXA FOUND RELATIVE TO ORIGINAL TREE AND RELATIVE TO OpenTree RF DISTANCE INTERPRETATION HOW MUCH TIME THE BLAST RUN TOOK ML ESTIMATES OF UPDATED TREE VS ORIGINAL TREE

FIGURE: FACE TO FACE ORIGINAL VS UPDATED PHYLOGENY, IN RED NEW TAXA NOT IN OpenTree.

3.2 The Malvaceae

A postdoc started working with a new reserach group. They are interested in solving relationships among lineages of the Malvaceae, a family of flowering plants with almost 6 000 known species, containing the relatives of cacao, cotton, durian and okra. A review of the literature shows them that there are many phylogenetic trees encompassing some of the linegaes in the group. However, the head of the research group wants to use a particular marker they believe to be the best one to be able to solve the relationships in the group. They have been working in the alignment for long and they want to incorporate new data into the hypothesis of homology that they have been curating and that they trust.

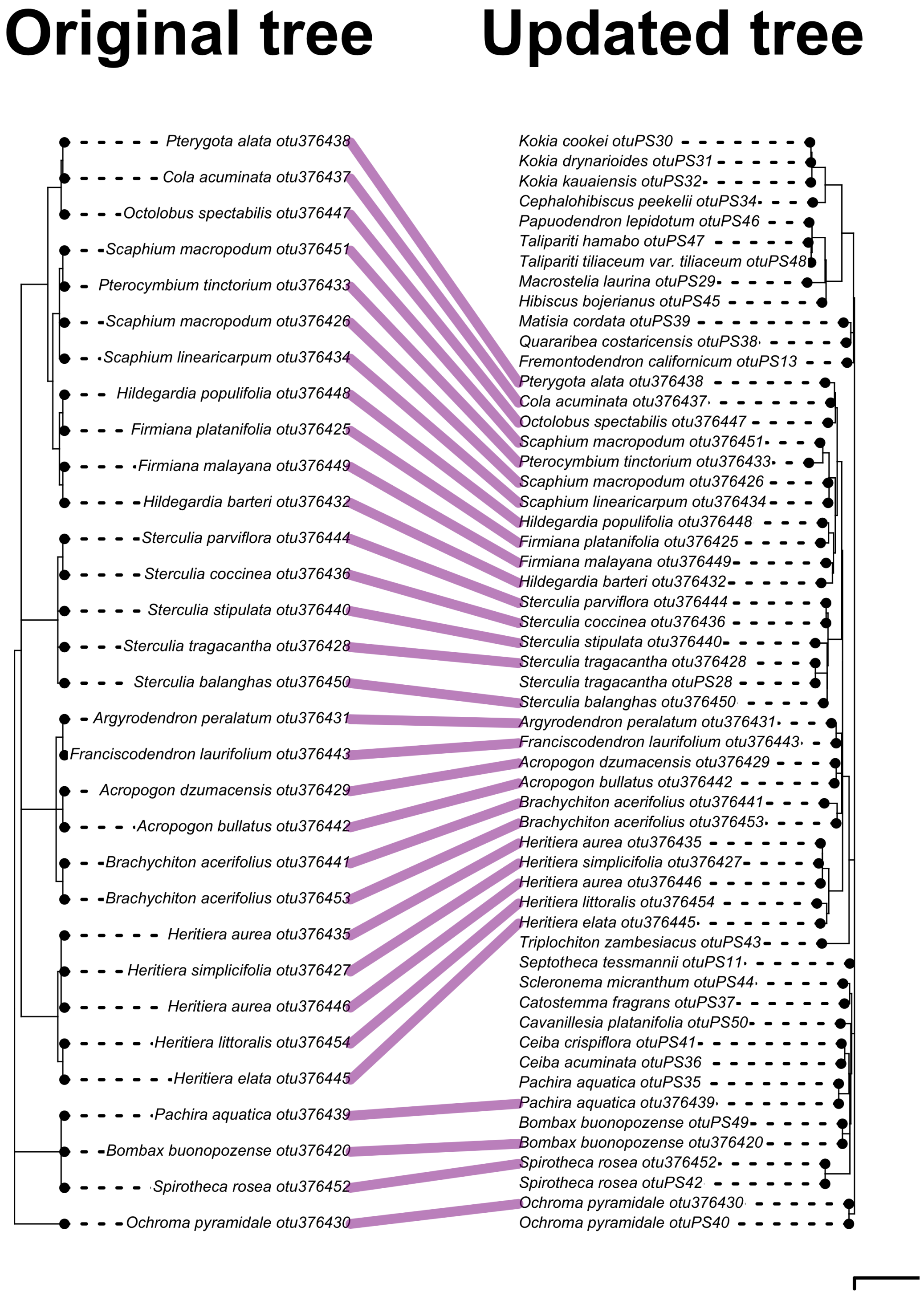
Original tree
Updated tree
Pterygota alata otu376438
Cola acuminata otu376437
Octolobus spectabilis otu376447
Scaphium macropodum otu376451
Pterocymbium tinctorium otu376433
Scaphium macropodum otu376426
Scaphium linearicarpum otu376434
Hildegardia populifolia otu376448
Firmiana platanifolia otu376425
Firmiana malayana otu376449
Hildegardia barteri otu376432
Sterculia parviflora otu376444
Sterculia coccinea otu376436
Sterculia stipulata otu376440
Sterculia tragacantha otu376428
Sterculia balanghas otu376450
Argyrodendron peralatum otu376431
Franciscodendron laurifolium otu376443
Acropogon dzumacensis otu376429
Acropogon bullatus otu376442
Brachychiton acerifolius otu376441
Brachychiton acerifolius otu376453
Heritiera aurea otu376435
Heritiera simplicifolia otu376427
Heritiera aurea otu376446
Heritiera littoralis otu376454
Heritiera elata otu376445
Pachira aquatica otu376439
Bombax buonopozense otu376420
Spirotheca rosea otu376452
Ochroma pyramidale otu376430
Kokia cookei otuPS30
Kokia drynarioides otuPS31
Kokia kauaiensis otuPS32
Cephalohibiscus peekelii otuPS34
Papuodendron lepidotum otuPS46
Talipariti hamabo otuPS47
Talipariti tiliaceum var. tiliaceum otuPS48
Macrostelia laurina otuPS29
Hibiscus bojerianus otuPS45
Matisia cordata otuPS39
Quararibea costaricensis otuPS38
Fremontodendron californicum otuPS13
Pterygota alata otu376438
Cola acuminata otu376437
Octolobus spectabilis otu376447
Scaphium macropodum otu376451
Pterocymbium tinctorium otu376433
Scaphium macropodum otu376426
Scaphium linearicarpum otu376434
Hildegardia populifolia otu376448
Firmiana platanifolia otu376425
Firmiana malayana otu376449
Hildegardia barteri otu376432
Sterculia parviflora otu376444
Sterculia coccinea otu376436
Sterculia stipulata otu376440
Sterculia tragacantha otu376428
Sterculia tragacantha otuPS28
Sterculia balanghas otu376450
Argyrodendron peralatum otu376431
Franciscodendron laurifolium otu376443
Acropogon dzumacensis otu376429
Acropogon bullatus otu376442
Brachychiton acerifolius otu376441
Brachychiton acerifolius otu376453
Heritiera aurea otu376435
Heritiera simplicifolia otu376427
Heritiera aurea otu376446
Heritiera littoralis otu376454
Heritiera elata otu376445
Triplochiton zambesiacus otuPS43
Septotheca tessmannii otuPS11
Scleronema micranthum otuPS44
Catostemma fragrans otuPS37
Cavanillesia platanifolia otuPS50
Ceiba crispiflora otuPS41
Ceiba acuminata otuPS36
Pachira aquatica otuPS35
Pachira aquatica otu376439
Bombax buonopozense otuPS49
Bombax buonopozense otu376420
Spirotheca rosea otu376452
Spirotheca rosea otuPS42
Ochroma pyramidale otu376430
Ochroma pyramidale otuPS40

Figure 1: Comparison of original tree and tree updated with Physcraper, family Malvaceae.

4 Discussion

Data repositories hold even more information than meets the eye. Besides the actual data, they are rich sources of metadata that can be used for the advantage of biology and science in general.

Many pipelines are making use of DNA data repositories in different ways. Most of them focus on efficient ways to mine the data – getting the most homologs. Some focus on accurate ways of mining the data – getting real and clean homologs. Others focus on refining the alignment. Most focus on generating full trees *de novo*, mainly for regions of the Tree of Life that have no phylogenetic assessment yet in published studies, but also for regions that have been already studied and that have phylogenetic data already. However, expert phylogenetic knowledge is also an important source of data in public and open repositories that is not being used to its full potential.

All these tools are key efforts for advancing towards reproducibility in phylogenetics, a field that has relied on processes which are somewhat artisanal. In here, we highlight the potential of taking advantage of this careful curation work in previous phylogenetic estimates. By taking sources of information available from data repositories and present a method to link data from different repositories, while leveraging on the knowledge and intuition of the expert community to build up our phylogenetic knowledge, we can use not only data accumulated in molecular data repositories, but phylogenetic knowledge accumulated in phylogenetic tree repositories. While not generating full phylogenies *de novo*, physcraper is still capable of generating new phylogenetic knowledge. It can also be combined with data from repositories other than molecular data. For example geographic locations (GBIF), fossils (PBDB), etc.

Physcraper has the potential to be applied for the advantage of the field to rapidly place newly discovered species phylogenetically (Webb *et al.* 2010), obtain trees for ecophylogenetic studies (Helmus & Ives 2012), help to systematize molecular databases, i.e., curate taxonomic assignations (San Mauro & Agorreta 2010), and rapidly generate custom species trees for downstream analyses (Stoltzfus *et al.* 2013).

5 Acknowledgements

Research was supported by the grant “Sustaining the Open Tree of Life”, National Science Foundation ABI No. 1759838. Computer time was provided by the Multi-Environment Research Computer for Exploration and Discovery (MERCED) cluster from the University of California, Merced (UCM), supported by the NSF Grant No. ACI-1429783.

We thank members of the “short bar” Science and Engineering Building 1, UCM, joint lab paper discussion meeting for valuable comments on a first version of this manuscript.

6 Authors’ Contributions

7 Data Avilability

8 References

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic acids research*, **25**, 3389–3402.

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

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. & Wheeler, D.L. (2000). GenBank. *Nucleic acids research*, **28**, 15–18.

Camacho, C., George, C., Vahram, A., Ning, M., Jason, P., Kevin, B. & Thomas, L. (2009). BLAST+: Architecture and applications. *BMC bioinformatics*, **10**, 421.

Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & others. (2009). Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Drew, B.T., Gazis, R., Cabezas, P., Swithers, K.S., Deng, J., Rodriguez, R., Katz, L.A., Crandall, K.A., Hibbett, D.S. & Soltis, D.E. (2013). Lost branches on the tree of life. *PLoS biology*, **11**.

Edgar, R.C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, **32**, 1792–1797.

Federhen, S. (2003). The taxonomy project. *The NCBI Handbook*.

Fragoso-Martínez, I., Salazar, G.A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Lemmon, E.M., Lemmon, A.R., Sazatornil, F. & Mendoza, C.G. (2017). A pilot study applying the plant anchored hybrid enrichment method to new world sages (salvia subgenus calosphace; lamiaceae). *Molecular Phylogenetics and Evolution*, **117**, 124–134.

Helmus, M.R. & Ives, A.R. (2012). Phylogenetic diversity–area curves. *Ecology*, **93**, S31–S43.

Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R. & others. (2015). Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences*, **112**, 12764–12769.

Lehtonen, S. (2011). Towards resolving the complete fern tree of life. *PLoS One*, **6**.

Magee, A.F., May, M.R. & Moore, B.R. (2014). The dawn of open access to phylogenetic data. *PLoS One*, **9**.

McTavish, E.J., Drew, B.T., Redelings, B. & Cranston, K.A. (2017). How and Why to Build a Unified Tree of Life. *BioEssays*, **39**. Retrieved April 10, 2018, from https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201700114

McTavish, E.J., Hinchliff, C.E., Allman, J.F., Brown, J.W., Cranston, K.A., Holder, M.T., Rees, J.A. & Smith, S.A. (2015). Phylesystem: A git-based data store for community-curated phylogenetic estimates. *Bioinformatics*, **31**, 2794–2800.

Morrison, D.A. (2006). Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany*, **19**, 479–539.

Morrison, D.A. (2009). Why would phylogeneticists ignore computerized sequence alignment? *Systematic biology*, **58**, 150–158.

Piel, W., Chan, L., Dominus, M., Ruan, J., Vos, R. & Tannen, V. (2009). Treebase v. 2: A database of phylogenetic knowledge. E-biosphere.

Redelings, B.D. & Holder, M.T. (2017). A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. *PeerJ*, **5**, e3058.

Rees, J.A. & Cranston, K. (2017). Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal*.

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**, 335–346.

San Mauro, D. & Agorreta, A. (2010). Molecular systematics: A synthesis of the common methods and the state of knowledge. *Cellular & Molecular Biology Letters*, **15**, 311.

Santini, F. & Sorenson, L. (2013). First molecular timetree of billfishes (istiophoriformes: Acanthomorpha) shows a late miocene radiation of marlins and allies. *Italian journal of zoology*, **80**, 481–489.

Scott, F. (2011). The ncbi taxonomy database. *Nucleic Acids Research*, **40**, D136–D14.

Smith, S.A., Beaulieu, J.M. & Donoghue, M.J. (2009). Mega-phylogeny approach for comparative biology: An alternative to supertree and supermatrix approaches. *BMC evolutionary biology*, **9**, 37.

Song, S., Liu, L., Edwards, S.V. & Wu, S. (2012). Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of Sciences*, **109**, 14942–14947.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stoltzfus, A., Lapp, H., Matasci, N., Deus, H., Sidlauskas, B., Zmasek, C.M., Vaidya, G., Pontelli, E., Cranston, K., Vos, R. & others. (2013). Phylotastic! Making tree-of-life knowledge accessible, reusable and convenient. *BMC bioinformatics*, **14**, 158.

Sukumaran, J. & Holder, M.T. (2010). DendroPy: A python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.

Vos, R.A., Balhoff, J.P., Caravas, J.A., Holder, M.T., Lapp, H., Maddison, W.P., Midford, P.E., Priyam, A., Sukumaran, J., Xia, X. & others. (2012). NeXML: Rich, extensible, and verifiable representation of comparative data and metadata. *Systematic biology*, **61**, 675–689.

Webb, C.O., Slik, J.F. & Triono, T. (2010). Biodiversity inventory and informatics in southeast asia. *Biodiversity and Conservation*, **19**, 955–972.

Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., Madden, T.L., Schuler, G.D., Tatusova, T.A. & Rapp, B.A. (2000). Database resources of the national center for biotechnology information. *Nucleic acids research*, **28**, 10–14.