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: 3246

Manuscript prepared for submission to Methods in Ecology and Evolution

Article type: Application

1 Abstract

1. Phylogenies are a key part of research in all 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 making use of public DNA sequence data and taxonomic information, providing a framework for comparison of published phylogenies with their updated versions.

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 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 much about a certain biological group.***

4. Physcraper implements node by node/topology comparison of the the original and the updated trees using the conflict API of OToL, and summarizes differences.

5. We hope the physcraper workflow demonstrates the benefits of opening results in phylogenetics and encourages researchers to strive for better 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.

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

2 Introduction

Phylogenies are important.

Generating phylogenies is not easy and it is largely artisanal. Although many efforts to automatize the process have been done, and the community is using those more and more, automatization of phylogenetic reconstruction is still not a widespread practice and among other benefits, it might be key for adoption of better reproducibility practices in the phylogenetics community. ***paragraph better to end discussion???***

The process of phylogenetic reconstruction implies many steps (that I generalize to the following):

1. Obtention of molecular or morphological character data – get DNA from some organisms and sequence it, or get it from an online nucleotide data repository, such as GenBank (Benson *et al.* 2000; Wheeler *et al.* 2000).
2. Assemble a hypothesis of homology – Create a matrix of your character data, by aligning the sequences, in the case of molecular data. Make sure thay are paralogs!
3. Analyse this hypothesis of homology to infer phylogenetic relationships among the organisms you are studying – Use different available programs to infer molecular evolution, trees and times of divergence.
4. Discuss the inferred relationships in the context of previous hypothesis, the biology and biogeography of the organisms, etc. – Answer the question, *is this phylogenetic solution fair/reasonable?*

Each of these steps require different types of specialized training: in the field, in the lab, in front of a computer, discussions with experts in the methods, and/or in the biological group of study. All of these steps also require considerable amounts of time for training and implementation.

In the past decade, various studies have developed solutions to automatize the first and second steps, by creating pipelines that mine already available molecular data from the GenBank repository (Benson *et al.* 2000; Wheeler *et al.* 2000), to obtain homologous characters that can be used for phylogenetic reconstruction. These tools have been presented as aid for the nonspecialist to decrease some of the difficulties in the generation of phylogenetic knowledge. However, they are not that often used as so, suggesting that there are

still difficulties for the nonspecialist. The phylogenetic community has some reserves towards these tools, too. Mainly because they sometimes act as a black box. However, automatizing the assembly of the character data set is a crucial step towards reproducibility for a task that was otherwise primarily artisanal and hence largely non-reproducible.

Even if it is hard to obtain phylogenies, we invest copious amounts of time and energy in generating them. Issues such as food security, global warming, global health are crucial to solve and phylogenies might help. There is a lot of phylogenetic knowledge already available in published peer-reviewed studies. In this sense, the non-specialists (and also the specialist) face a new problem: how do I choose the best phylogeny.

Public phylogenies can be updated with the ever increasing amount of genetic data that is available on GenBank (Benson *et al.* 2000; Wheeler *et al.* 2000).

We present a way to automatize and standardize the comparison of phylogenetic hypotheses and to allow reproducibility of this last step of the research process.

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 on the focal group of study.

Concerns I think people have about 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 construction from zero? Yes.

The goal of Physcraper is to build upon previous phylogenetic knowledge, allowing a direct comparison

between existing phylogenies and phylogenies that are constructed using new genetic data retrieved from a public nucleotide database (i.e., GenBank (Benson *et al.* 2000; Wheeler *et al.* 2000)).

To achieve this, Physcraper uses the Open Tree of Life phylesystem and connects it to the TreeBase database, to (1) get the original DNA data set 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 default 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.

3 How does Physcraper work?

3.1 The input: a study tree and an alignment

- The study tree is a published phylogenetic tree stored in the OToL database, phylesystem (McTavish *et al.* 2015). 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 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 next steps of the pipeline. A user can choose from the 'r rotl::tol_about()$num_source_trees' published trees supporting the resolved node of the synthetic tree in the OToL website (<>). If the tree you are interested in updating is not in there, you can upload it via OToL's curator tool (<https://tree.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 (http://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 (https://treebase.org/) or through the TreeBASE GitHub repository (SuperTreeBASE; https://github.com/TreeBASE/supertreebase). If the alignment is on another repository, or provided personally by the owner, a copy of it has to be downloaded by the user, and it's local path has to be provided as an argument.
- 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.

3.2 DNA sequence search and cleaning

- 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 provided 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 tree, that is also a named clade in the reference taxonomy. This is known as the Most Recent Common Ancestral Taxon (MRCAT; also referred in the literature as the Least Inclusive Common Ancestral Taxon - LICA). 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 OToL taxonomy tool (https:

//github.com/OpenTreeOfLife/germinator/wiki/Taxonomy-API-v3#mrca).

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); altschul1997gapped] is used to identify similarity between DNA sequences within the search taxon in a nucleotide database, and the accepted sequences on the alignment. The blastn function from the BLAST command line tools (Camacho *et al.* 2009) is used for local-database searches. A modified biopython blast function is used for web-based searches.

- The DNA sequence similarity search can be done on a local database that is easily setup by the user. In this case, the blastn function is used to performs the similarity search (Camacho *et al.* 2009).

- The search can also be performed remotely, on the NCBI database. In this case, the bioPython BLAST function was modified to accepts is used to perform the similarity search.

- A pairwise alignment-against-all 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 (accesion 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 below an e-value, percentage similarity, and outside a minimum and maximum length threshold are discarded. ***REPORT THE DEFAULT VALUES AND DESCRIBE WHAT THEY MEAN*** This filtering leaves out genomic sequences. All acepted sequences are asigned an

internal identifier, and are further filtered.

- Because the original alignments usually lack database id numbers, a filtering step is needed. Accepted sequences that belong to the same taxon of the query sequence, and that are either identical or shorter than the original sequence are discarded. Only longer sequences belonging to the same taxon as the orignal sequence will be considered further for analysis.

- Among the remaining filtered sequences, there are usually several exemplars per taxon. Although it can be useful to keep some of them to, for example, 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 a more "cycles" of sequence similarity search, by blasting the newly found sequences. This can be done iteratively, but by default only sequences in the alignment are blasted. ***Is there an argument to control the number of cycles of blast searches with new sequences?***

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

- A fasta file containing all filtered and processed sequences resulting from the BLAST search is generated for the user.

3.3 DNA sequence alignment

- The software MUSCLE (Edgar 2004) is implemented to perform 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. 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.

3.4 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) with 100 classic rapid bootstrap (Felsenstein 1985) replicates by default. The number of bootsrap replicates can be modified by the user. Other type of bootstrap that I think is not yet incorporated into physcraper is the Transfer Bootstrap Expectation (TBE) recently proposed in Lemoine *et al.* (2018).
- Bootstrap results are summarized with Dendropy ADD CITATION
- 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, allowing to perform comparisons and conflict analyses.
- Robinson Foulds weighted and unweighted metrics ARE CALCULATED WITH DENDROPY TOO.
- Describe what a conflict analysis is: Node by node comparison of the resulting clades compared to CITE REDELINGS AND HOLDER (**???** and holder)
- For the conflict analysis to be meaningful, the root of the tree ineeds to be accurately defined.
- A SUGGESTED DEFAULT ROOTING BASED ON THE OPEN TREE TAXONOMY is implemented for now. DESCRIBE HOW IT WORKS. SAY THAT IT IS A PROBLEM. Automatic rooting is not that smart yet. The best way right now is for users to define outgroups so trees are better rooted.
- Currently, the root is determined by finding the parent node of the sequences that do not belong to the ingroup/ search taxon. This ensures a correct rooting of the tree even when the search taxon is more inclusive than the ingroup.
- Conflict information can only be generated in the context of the whole Open Tree of Life. Otherwise, it is not really possible to get conflict data. ***- One way to compare two independent phylogenetic***

trees is to compare them both to the synthetic OToL and then measure how well they do against each other

4 Examples

4.1 The hollies

The genus *Ilex* is the only extant clade within the family Aquifoliaceae, order Aquifoliales of flowering plants. It encompasses between 400-600 living species. A review of litterature shows that there are three published phylogenetic trees, showing relationships within the hollies. The first one has been made available both on OToL phylesystem and synth tree, and on treeBASE, it samples 48 species. The second has not been made available anywhere, not even in supplementary data of the journal. ***Contact authors? They seem old school, probably do not wanna share their data.*** The most recent one has been made available in the OToL Phylesystem and DRYAD. It is the best sampled yet, with 200 species. However, it has not been added to the syntehtic tree yet. This 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 (Benson *et al.* 2000; Wheeler *et al.* 2000). Updating the oldest tree, we should get something very similar to the newest tree.

4.2 The Ascomycota

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 study_year title

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

Both trees from this study have NA tips.

Let's check what one of the trees looks like:

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:

```
physcraper_run.py -s pg_238 -t tree109 -o ../physcraper_example/pg_238
```

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:

```
physcraper_run.py -s pg_238 -t tree109 -a treebase_alns/pg_238tree109.aln -as "nexus" -o pg_238
```

4.3 Testudines example

Phylogeny of the Testudines 6 tips from Crawford *et al.* (2012) There is just one tree in OToL. There is just one alignment on treebase with all the 1 145 loci.

```
physcraper_run.py -s pg_2573 -t tree5959 -tb -db ~/branchinecta/local_blast_db/ -o pg_2573
```

5 Discussion

Data repositories hold more information than meets the eye. Besides the actual data, they have other types of information that can be used for the advantage of science.

Usually, initial ideas about the data are changed by analyses. We expect that this new ideas on the data can be registered on data bases, exposing new comers to expert understanding about the data.

There are many tools that 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 refinement of 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.

All these tools are great efforts for advancing towards reproducibility in phylogenetics, a field that has been largely recognised as somewhat artisanal. We propose adding focus to other sources of information available from data repositories. Taking advantage of public DNA data bases have been the main focus. However, phylogenetic knowledge is also accumulating fast in public and open repositories. In this way, the physcraper pipeline can be complemented with other tools that have been developed for other purposes.

We emphasize that physcraper takes advantage of the knowledge and intuition of the expert community to build upon phylogenetic knowledge, using not only data accumulated in DNA repositories, but phylogenetic knowledge accumulated in tree repositories. This might help generate new phylogenetic data. But physcraper does not seek to generate full phylogenies *de novo*.

Describe again statistics to compare phylogenies provided by physcraper via OpenTreeOfLife. Mention statistics provided by other tools: PhyloExplorer (Ranwez *et al.* 2009). Compare and discuss.

How is physcraper already useful: - to mine targeted sequences, in this way it is similar to baited analyses from PHLAWD and pyPHLAWD. Phylota does not do baited analyses, I think, only clustered analyses. - Finding

How can it be used for the advantage of the field: - rapid phylogenetic placing of newly discovered species, as mentioned in Webb *et al.* (2010) - obtain trees for ecophylogenetic studies, as mentioned in Helmus & Ives (2012) - one day could be used to sistematize nucleotide databases, such as Genbank (Benson *et al.* 2000; Wheeler *et al.* 2000), as mentioned in San Mauro & Agorreta (2010), i.e., curate ncbi taxonomic assignations. - allows to generate custom species trees for downstream analyses, as mentioned in Stoltzfus *et al.* (2013)

Things that physcraper does not do: - analyse the whole GenBank database (Benson *et al.* 2000; Wheeler *et al.* 2000) to find homolog regions suitable to reconstruct phylogenies, as mentioned in Antonelli *et al.* (2017). There are already some very good tools that do that. - provide basic statistics on data availability to assemble molecular datasets, as mentioned by Ranwez *et al.* (2009). Phyloexplorer does this? - it is not a tree repo, as phylota is, mentioned in Deepak *et al.* (2014)

6 Acknowledgements

We acknowledge contributions from

The University of California, Merced cluster, MERCED (Multi-Environment Research Computer for Exploration and Discovery) supported by the National Science Foundation (Grant No. ACI-1429783).

7 Authors' Contributions

8 Data Avilability

9 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.

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.

Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winker, K. & Glenn, T.C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology letters*, **8**, 783–786.

Deepak, A., Fernández-Baca, D., Tirthapura, S., Sanderson, M.J. & McMahon, M.M. (2014). EvoMiner: Frequent subtree mining in phylogenetic databases. *Knowledge and Information Systems*, **41**, 559–590.

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

Felsenstein, J. (1985). Confidence intervals on phylogenetics: An approach using bootstrap. *Evolution*, **39**, 783–791.

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

Lemoine, F., Entfellner, J.-B.D., Wilkinson, E., Correia, D., Felipe, M.D., De Oliveira, T. & Gascuel, O. (2018). Renewing felsenstein's phylogenetic bootstrap in the era of big data. *Nature*, **556**, 452–456.

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.

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

Ranwez, V., Clairon, N., Delsuc, F., Pourali, S., Auberval, N., Diser, S. & Berry, V. (2009). PhyloExplorer: A web server to validate, explore and query phylogenetic trees. *BMC evolutionary biology*, **9**, 108.

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.

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

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.

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.