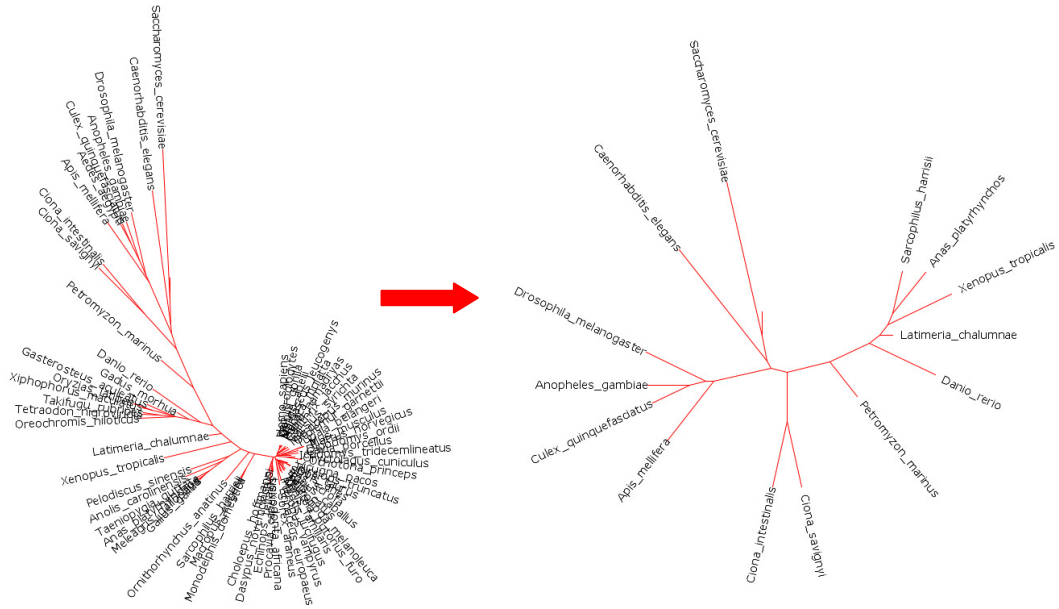
PhyloPrune User Manual

February 2015

1. Introduction

Pruning is a method for the simplification of a given tree. On the left, you can see the original tree and on the right the simplified pruned one, which holds the most representative leaves of the original:



PhyloPrune is an algorithm that automatically prunes large phylogenetic trees (>10,000 leaves) by removing leaves with low information content, fast, via the use of lossless heuristic methods and progressive patristic distance calculations.

The algorithm is written in PHP5 and it can run in command line or through the web.

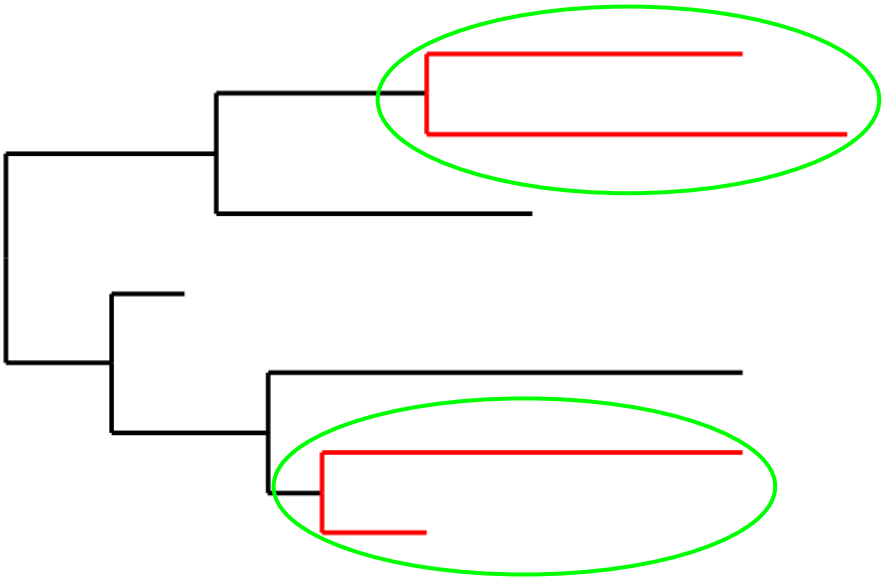
Source code and documentation for PhyloPrune are freely available for download at http://michalopoulos.net/phyloprune/

2. Algorithm

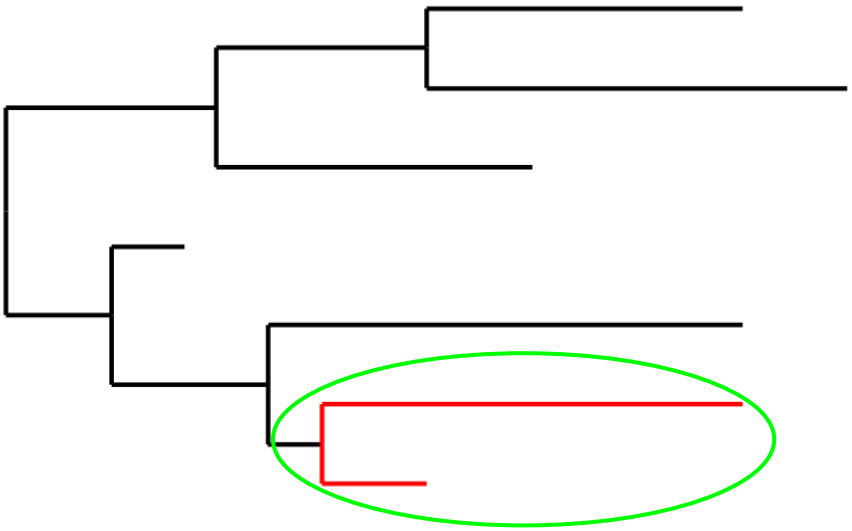
PhyloPrune algorithm parses a weighted phylogenetic tree file in Newick format to search for the pair of leaves with the minimum patristic distance and then to prune the leaf of that pair, whose average patristic distance to all leaves is maximum, producing a pruned Newick file. This procedure repeats until a set number of iterations is met. Below we explain the heuristic methods we used in the algorithm using an example tree.

2.1. Minimum patristic distance

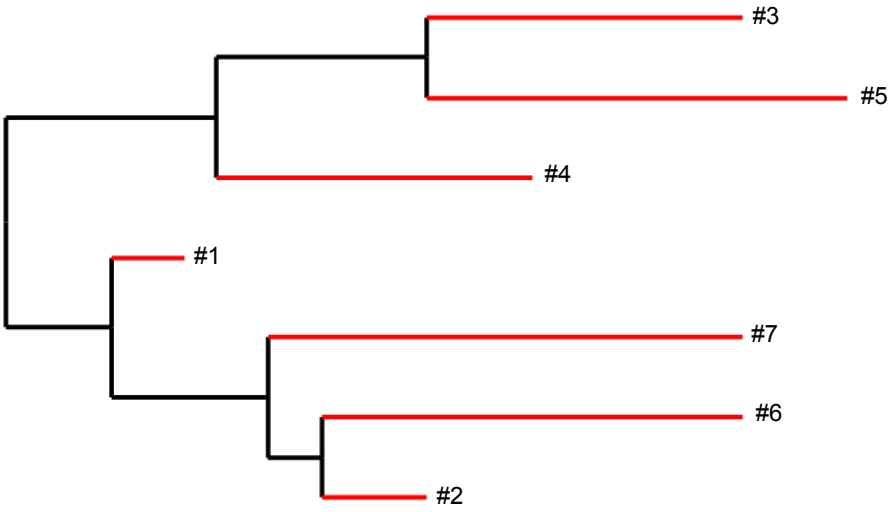
- We first calculate the patristic distance of leaves that share the same immediate parental node (cup-like leaf pairs):



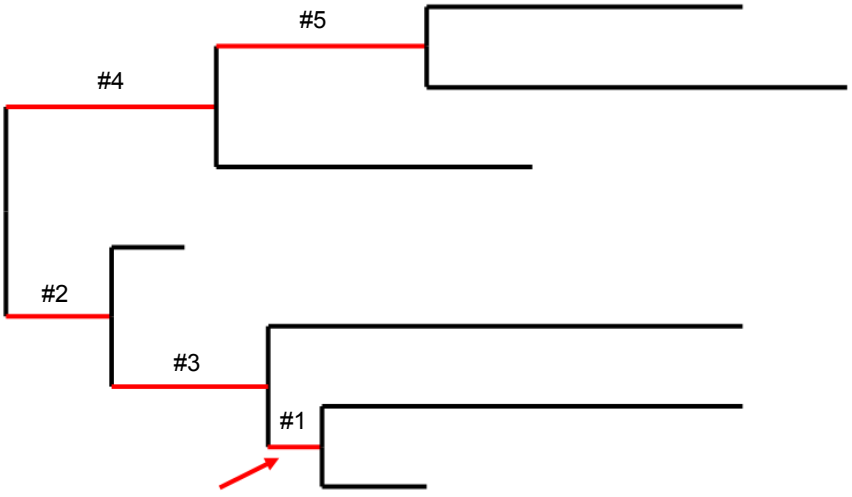
- We identify the pair with the minimum patristic distance and we set this distance as the initial cutoff:



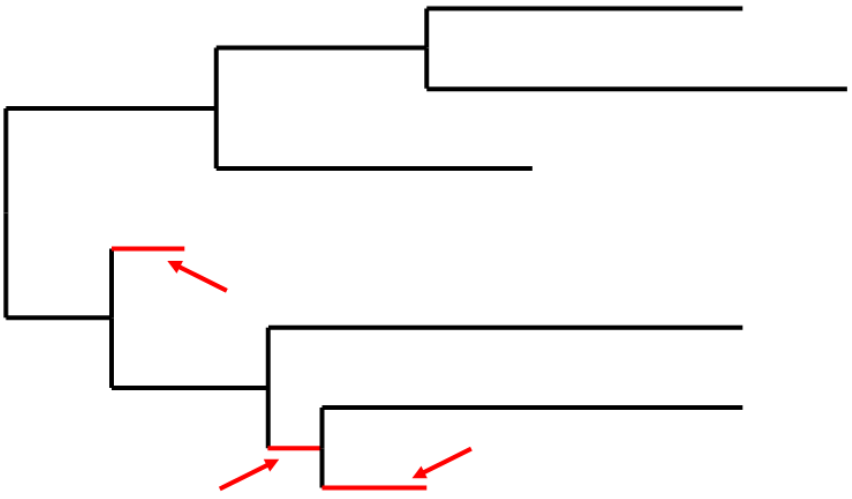
- We sort the length of the branches of each leaf to its immediate parental node:



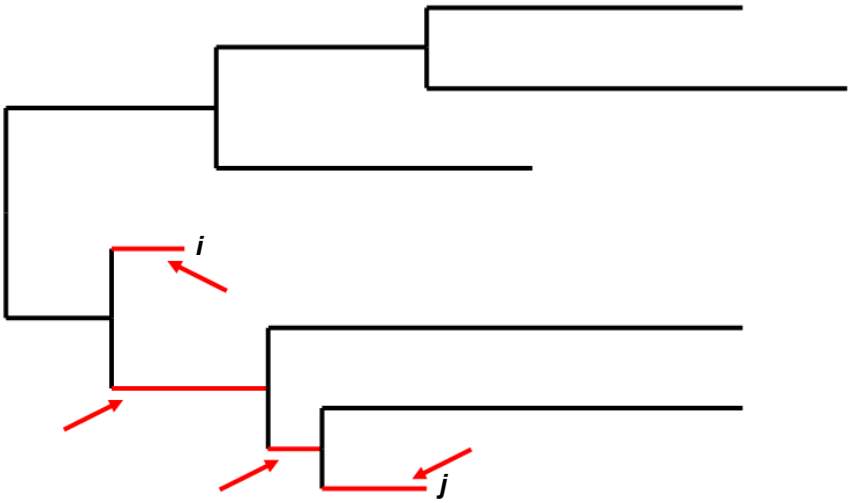
- We also identify the shortest branch between inner nodes:



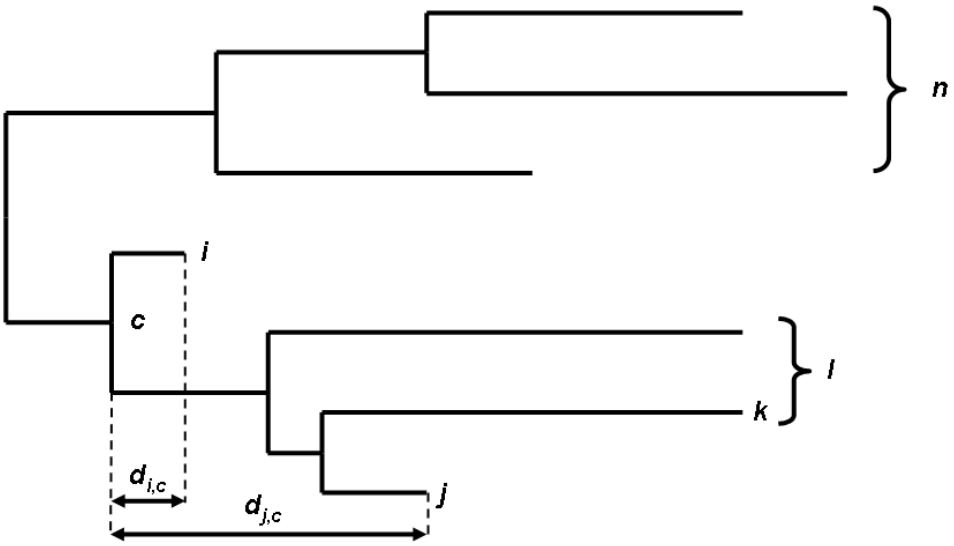
- We calculate the “pseudopatristic” distance of leaves by summing the shortest inner node branch length with the branch length of each leaf to its immediate parental node:



- If the “pseudopatristic” distance is lower than the cutoff, the real patristic distance is calculated and it becomes the new cutoff, if it is smaller than the cutoff. The procedure ends when the pair of leaves (*i* and *j*) with the minimum patristic distance is found:



2.2. Maximum distance



We now need to find the total distance of the leaves *i* and *j* whose patristic distance is minimum, to the rest of the leaves. Since we are calculating the patristic distances progressively, we do not have all patristic distances from leaves *i* and *j* to the rest of the leaves in order to sum and compare them. Thus, we calculate the difference $Diff_{i,j}$ of the total distance of leaves *i* and *j* with the rest of the tree leaves, using the following equation:

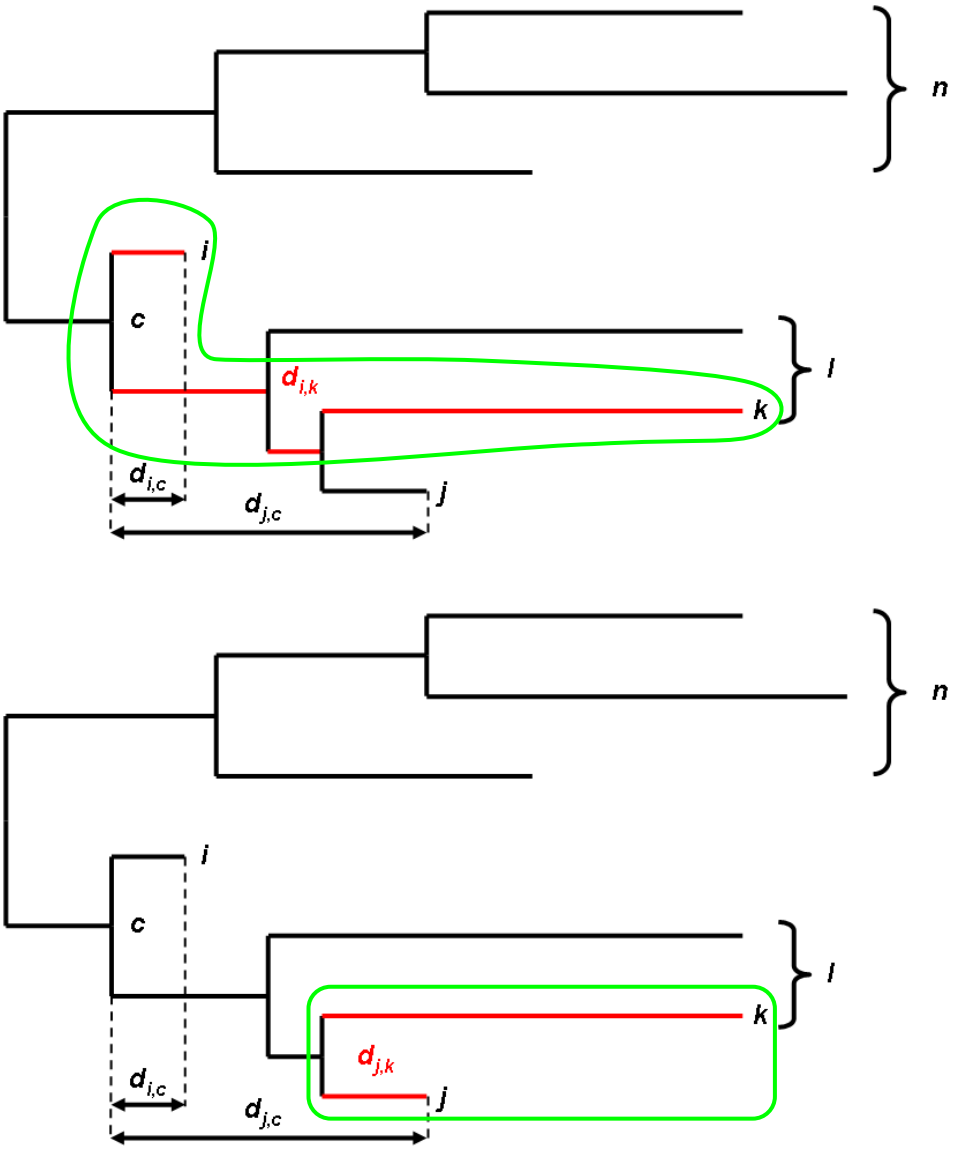
$$Diff_{i,j} = \left(n \cdot d_{j,c} + \sum_{k=1}^{l} d_{j,k} \right) - \left(n \cdot d_{i,c} + \sum_{k=1}^{l} d_{i,k} \right)$$

where,

- *c* is the lowest common ancestor (LCA) shared by *i* and *j*
- *n* is the number of leaves outside *c*
- $d_{i,c}$ and $d_{j,c}$ are the distances from *c* to *i* and *j,* respectively
- *k* is a leaf inside *c*
- *l* is the number of leaves inside *c*
- $d_{i,k}$ and $d_{j,k}$ are the patristic distances from *i* and *j* to *k,* respectively

If $Diff_{i,j}$ equals 0 (the total distance of leaf *i* with the rest of the tree leaves is equal to the total distance of leaf *j* with the rest of the tree leaves), the leaf with the maximum number of internal nodes between itself and the LCA, is pruned.

The patristic distances $d_{i,k}$ and $d_{j,k}$ are displayed below:



3. Installing PHP

To test whether you have PHP installed in your computer, type in a command line prompt:

```
php -v
```

or

```
php5 -v
```

If you do not get something like the following output:

```
PHP 5.4.8 (cli) (built: Oct 16 2012 22:30:23)
Copyright (c) 1997-2012 The PHP Group
Zend Engine v2.4.0, Copyright (c) 1998-2012 Zend Technologies
```

you must install PHP, as follows:

- **Windows:**

Download PHP for Windows from http://php.net/downloads.php.
To run PHP without using the full path every time, you may set the windows command path by following the next steps:

  - Click the **Start** button  , right-click **Computer**, and then click **Properties**.
  - Click **Advanced system settings.**  If you're prompted for an administrator password or confirmation, type the password or provide confirmation.
  - In the **Advanced** tab, click the **Environment Variables** button.
  - Finally, in the Environment Variables window, highlight the **Path** variable in the **Systems Variable** section and click the **Edit** button. Add the PHP path (e.g. C:\php5). Separate each directory with a semicolon (;).

- **Linux:**

Refer to your distro's package manager on how to install PHP.

- **Mac OS X:**

The Mac OS X operating system comes pre-installed with the libraries needed to run PHP programs.

4. Automated pruning

To get the automated pruning help message, type:

```
php PhyloPrune.php
```

or:

```
php PhyloPrune.php -h
```

The help lines are as follows:

```
PhyloPrune: An automated tree pruning algorithm retaining the most
representative leaves

Usage:

php Programs\PhyloPrune.php -i <inputfile> -o <outputfile>
[ -f <n> -l <m> -s <s> -d ]

Options:
        -i <inputfile>  Input file in Newick format
```

Input the Newick file you wish to prune by typing the path/name of the file.

```
        -o <outputfile> Output pruned tree in Newick format
```

Type the desired location of your output file(s) by typing path/name.

```
        -f <integer>    Maximum number of leaves after pruning(default: All)
```

Type the number of leaves where pruning starts from.

```
        -l <integer>    Minimum number of leaves after pruning (default: 2)
```

Type the number of leaves where pruning ends. This number must not be lower than 2 or higher than the number of leaves where pruning starts from.

```
        -s <integer>    Step (default: 1)
```

Type the number of pruning events every which you want to see Newick outputs.

```
        -d              Write output in a single file
```

With this option you get all the pruned Newick trees in a single (digest) file.

```
        -h              This message
```

Below, there are some different runs, using a tree which consists of 68 species whose genome is available on Ensembl (if you are unaware of the total number of leaves, you should run `leafcounter.php`, a program that returns the total number of leaves in a Newick tree).

```
php PhyloPrune.php -i ensembl68.new -o ensembl_pruned.new
```

PhyloPrune algorithm reads the Newick file `ensembl68.new` and it creates a series of pruned files (`ensembl_pruned_67.new, ensembl_pruned_66.new, ensembl_pruned_65.new,…`). Notice that we did not use any other parameters, thus the program will use its default values: It prunes all the leaves up to the last two, creating *n*-2 Newick files where *n* is the total number of leaves of the original tree (in this case it creates 68-2=66 files). If you want a single file containing all 66 trees, you have to use the `-d` flag, as follows:

```
php PhyloPrune.php -i ensembl68.new -o ensembl_pruned.new -d
```

In the next example, we will focus on a small pruning window:

```
php PhyloPrune.php -i ensembl68.new -o ensembl_pruned.new -f 40 -l 10 -s 5 -
d
```

where PhyloPrune will output pruned Newick files starting from the pruned Newick tree with 40 leaves (`-f 40`) to the pruned tree with 10 leaves (`-l 10`) every 5 pruning events (`-s 5`). Those trees will have 40, 35, 30, 25, 20, 15 and 10 leaves. To open a single file with many Newick trees, we recommend using Archaeopteryx: https://sites.google.com/site/cmzmasek/home/software/archaeopteryx

Finally, PhyloPrune algorithm produces the list of the pruned leaves up to the last pruning event with the extension `.list`. For the example above, the `ensembl68.new.list` is as follows:

```
Homo_sapiens
Macaca_mulatta
Pan_troglodytes
Gorilla_gorilla
Pongo_abelii
Ailuropoda_melanoleuca
Papio_hamadryas
Gallus_gallus
Canis_familiaris
Callithrix_jacchus
Bos_taurus
Meleagris_gallopavo
Rattus_norvegicus
Ovis_aries
Tursiops_truncatus
Mustela_putorius_furo
Macropus_eugenii
Vicugna_pacos
Dasypus_novemcinctus
Equus_caballus
Otolemur_garnettii
Procavia_capensis
Monodelphis_domestica
Tarsius_syrichta
Pteropus_vampyrus
Microcebus_murinus
Felis_catus
Myotis_lucifugus
Nomascus_leucogenys
Loxodonta_africana
Choloepus_hoffmanni
Ochotona_princeps
Taeniopygia_guttata
Ictidomys_tridecemlineatus
Tupaia_belangeri
Oryzias_latipes
Oryctolagus_cuniculus
Erinaceus_europaeus
Tetraodon_nigroviridis
Cavia_porcellus
Sorex_araneus
Dipodomys_ordii
Echinops_telfairi
Aedes_aegypti
Mus_musculus
Xiphophorus_maculatus
Gasterosteus_aculeatus
Takifugu_rubripes
Sus_scrofa
Oreochromis_niloticus
Gadus_morhua
Anolis_carolinensis
Ornithorhynchus_anatinus
Pelodiscus_sinensis
Sarcophilus_harrisii
Culex_quinquefasciatus
Anas_platyrhynchos
Xenopus_tropicalis
```

5. Manual pruning

To get the manual pruning help message, type:

```
php Manual_prune.php
```

or:

```
php Manual_prune.php -h
```

The help lines are as follows:

```
Manual tree pruning algorithm
Usage:

php Manual_prune.php -n <newickfile> -l <listfile> -d|-r

Options:
        -n <newickfile> Tree file in Newick format
```

Input the Newick file you wish to prune by typing the path/name of the file.

```
        -l <listfile>   List with leaves
```

Type the path/name of the file with selected leaf names. Leaf names must be separated by comma, space, tab, or new line characters.

```
        -d              Delete leaves from list
```

With this option, you prune the leaves which are on the list file, from the Newick tree.

```
        -r              Retain leaves from list
```

With this option, you retain the leaves which are on the list file, from the Newick tree and you prune all the remaining leaves.

```
        -h              This message
```

In the following runs, we use the same tree from Ensembl as before, and a list of selected leaves (`list.txt`).

```
php Manual_prune.php -n ensembl68.new -l list.txt -d
```

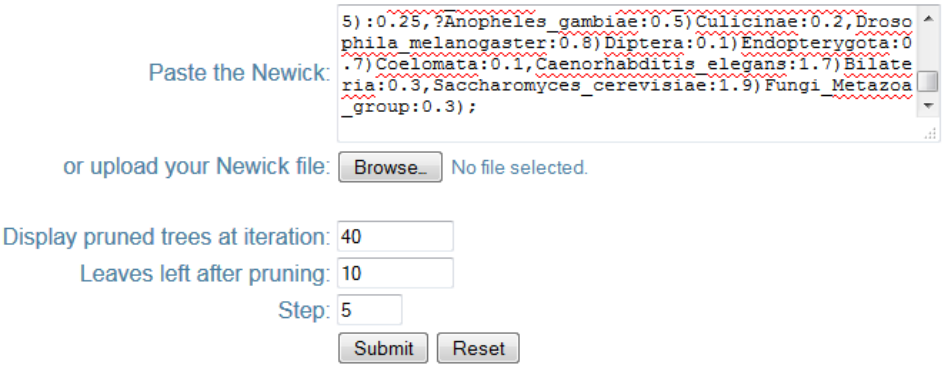
Manual_prune algorithm reads the Newick file `ensembl68.new` and the list file `list.txt`. With the `-d` option, all the leaves in the list are pruned. If you want the leaves from the list to be retained and all the others pruned, you have to use `-r` option as follows.

```
php Manual_prune.php -n ensembl68.new -l list.txt -r
```

6. Web installation

In order to install the Web interface on your computer, make sure that the PHP module is loaded on your Web server, copy PhyloPrune directory into your web directory and visit it with a browser. Currently, the Web interface has size limit of 1000 leaves, which you can increase by editing `maxtreesizewebonly.txt` file.

The submission form for PhyloPrune Web interface is as follows:



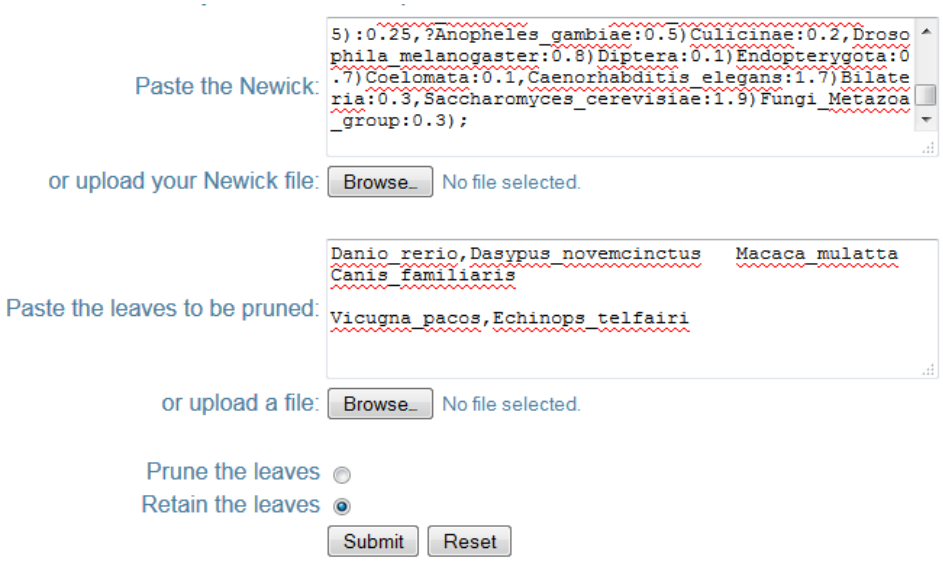
- **Newick text area:** In this box input strictly one Newick-formatted phylogenetic tree,
  or
- **Upload file:** Upload a single Newick-formatted tree from a file.

- **Display pruned trees at iteration:** Start showing the trees at user's preferred iteration. Default selection shows all the pruned trees.

- **Leaves left after pruning:** Type the number of leaves where pruning ends. This number must not be lower than 2 or higher than the number of leaves where pruning starts from.

- **Step:** Type the number of pruning events every which you want to see Newick outputs.

The submission form for Manual pruning Web interface is as follows:



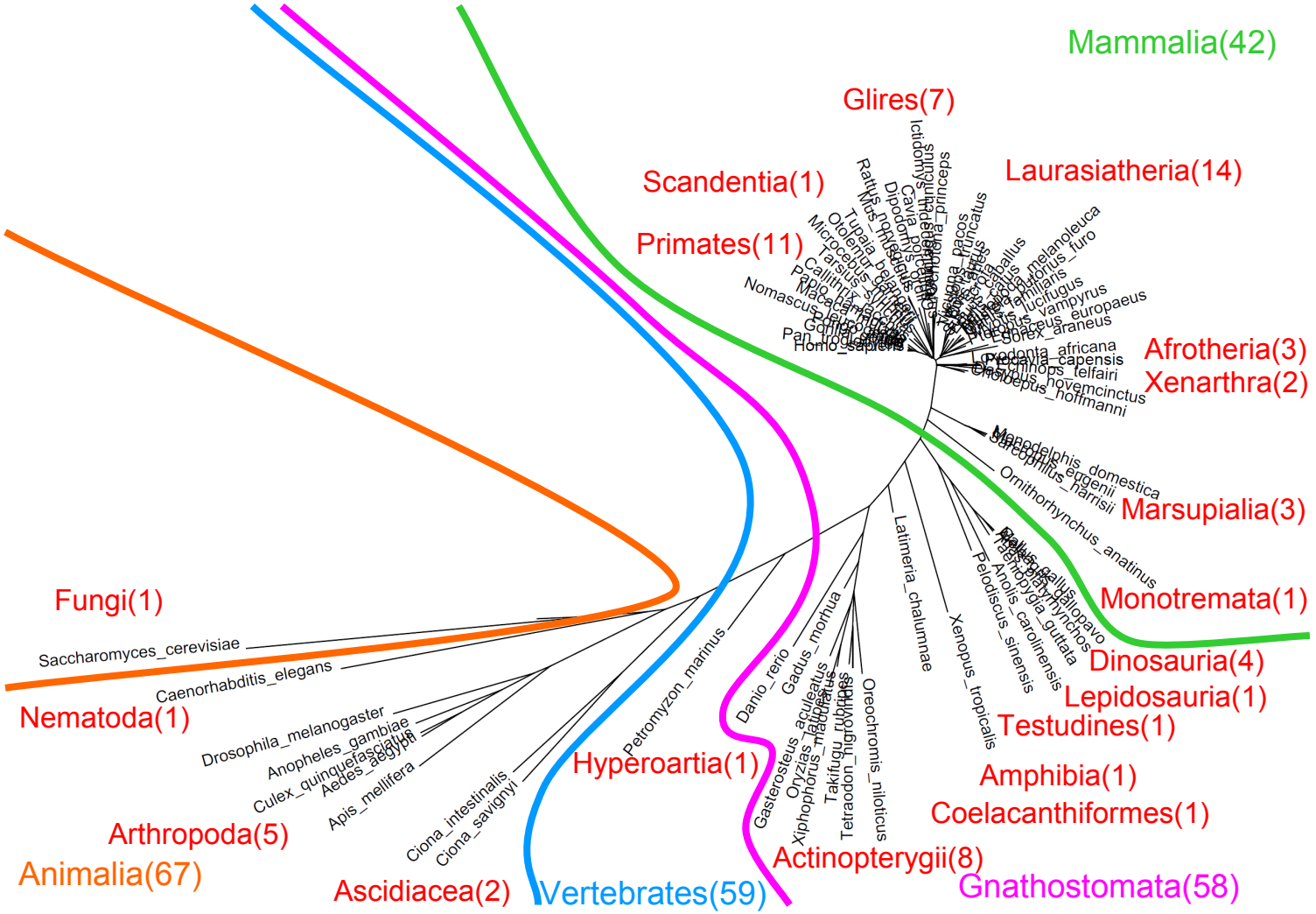
- **Newick text area:** In this box input strictly one Newick-formatted phylogenetic tree.
  or
- **Upload file:** Upload a single Newick-formatted tree from a file.

- **Leaves to be pruned text area:** In this box input the leaves to prune/retain,
  or
- **Upload file:** Upload the leaves to prune/retain from a file.

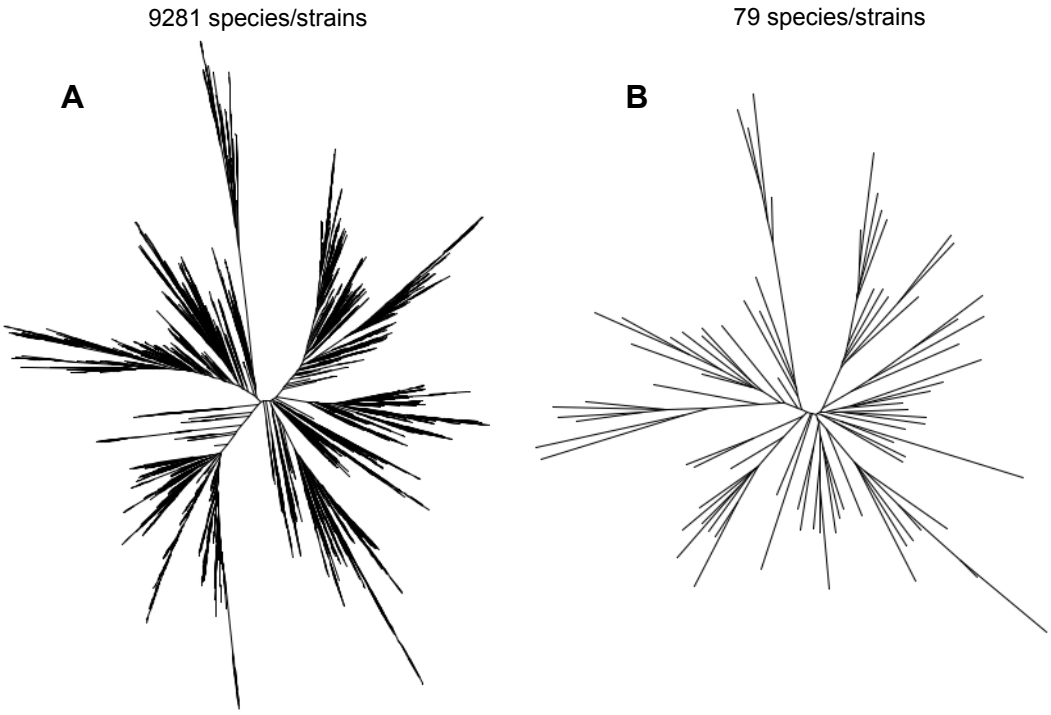
- **Prune/Retain:** Select whether to prune or retain the leaves you choose.

The users can also check the integrity of their tree and get the number, names of the leaves at “Check your tree” by uploading or copying their tree it in the designated area.

7. *Supplementary figures*



Supplementary Fig. 1. Taxonomic classification of the original Ensembl tree which consists of 68 species whose genome is available in Ensembl.



Supplementary Fig. 2. A) The LTPs108_SSU original tree which is a phylogenetic reconstruction of 9281 bacterial species or strains based on 16S rRNA gene sequence alignment. B) The LTPs108_SSU pruned tree with 79 leaves of which 76 represent distinct bacterial families.