

Treemmer's pruning options

Fabrizio Menardo (fabrizio.menardo@swisstph.ch)

This is a short tutorial regarding the three pruning options implemented in Treemmer (*-lm*, *-mc* and *-lmc*).

General considerations

The three pruning options do not alter the basic algorithm of Treemmer. Treemmer identifies the leaf to prune as usual, however it prunes it only if this is not protected by the pruning options.

***-lm* option**

This option can be used to pass the path to the file containing the meta-information to Treemmer. Here an example:

```
Seq1,countryA
Seq2,countryB
Seq3,countryB
Seq4,countryA
Seq10,countryC
Seq11,countryC
Seq2,phenotypeA
Seq3,phenotypeB
Seq4,phenotypeA
Seq4,phenotypeB
Seq5,phenotypeB
...
...
...
```

Each row should have leaf name followed by a comma, followed by a string (meta-information tag). The meta-information tag can represent anything: geographic region, date of sampling, abundance, phenotypes etc. Commas are not allowed in the leaf name or in the meta-information tag.

Leaf names can appear more than once or not appear at all.

***-mc* option**

The *-lm* option alone does not modify the basic behavior of Treemmer, to use the information passed with the *-lm* option it is necessary to use the *-mc* option as well (or *-lmc*).

When the *-lm* and *-mc* options are passed, Treemmer identifies the leaf to prune following the standard algorithm, however before pruning it checks whether the leaf is present in the meta-information file. Let's make an example with the list of meta-information reported above and *-mc* argument = 10 (*-mc 10*).

Imagine that Treemmer selected Seq2 for pruning. Seq2 is present in the list of meta-information with two different meta-information tags: country B and phenotypeA. Treemmer will count the number of leaves in the tree with these two tags. If there are more than 10 leaves from countryB and more than 10 leaves with phenotypeA Treemmer will prune Seq2.

If there are less than 10 leaves from countryB or less than 10 leaves with phenotypeA Seq2 will not be pruned. Treemmer will continue its routine considering the other leaf in the selected leaves pair. If also the second leaf cannot be pruned, Treemmer will consider the pair with the second shortest distance. If also the leaves of the second pair are protected Treemmer will consider the third etc..

An example with the TB tree

Let's use the MTB tree trimmed to 95% of RTL for an example
(MTB_tree_trimmed_tree_RTL_0.95) (It is the tree in Fig 3 b in the manuscript)

This tree has 4919 leaves belonging to the following lineages:

Lineage	Number of leaves
L1	624
L2	570
L3	696
L4	2595
L7	7
Maf+animal_clade	427

The file MTB_meta_info_lineages contains the list of leaves in the tree with corresponding lineage.

We now want to run Treemmer maintaining at least 250 strains from each lineage (obviously for L7 we can have 7 at maximum).

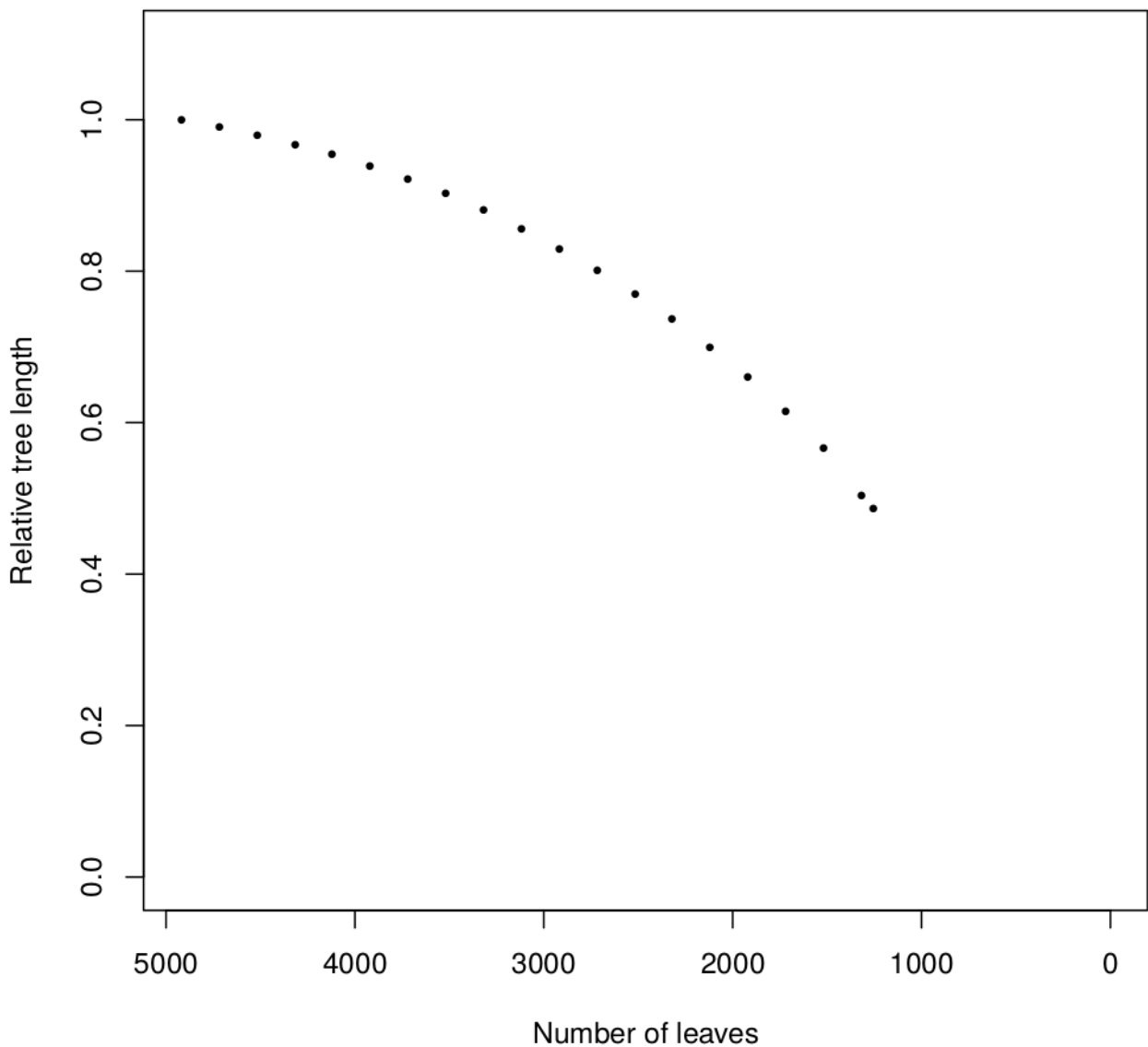
We run Treemmer as follow (you can use -r and -c options to speed up the run):

```
python Treemmer.py -lm MTB_meta_info_lineages -mc 250 MTB_tree_trimmed_tree_RTL_0.95
```

Treemmer prunes the tree and creates the RTL decay plot, but differently from a normal run it stop when all the remaining leaves are protected by the -lm and -mc options (this happened at 1257 leaves in this case), and outputs a warning message.

“WARNING: all remaining leaves are protected by the -lm option, outputting the results at current iteration”.

The resulting RTL decay plot (I used -r 200):



In combination with -X and -RTL options it is possible to obtain different results. If we want to obtain a tree with exactly 250 strains for lineage we can run

```
python Treemmer.py -lm MTB_meta_info_lineages -mc 250 MTB_tree_trimmed_tree_RTL_0.95
-X 10
```

Treemmer tries to prune the tree to ten leaves, however when the tree has around 1250 leaves all the remaining leaves are protected, and Treemmer outputs the current tree.

If we want to do a moderate pruning of the tree, say to 3000 leaves we can run:

```
python Treemmer.py -lm MTB_meta_info_lineages -mc 25 MTB_tree_trimmed_tree_RTL_0.95
-X 3000
```

In this case Treemmer outputs a tree with 3000 leaves and at least 250 leaves for each lineage

-lmc option

-lmc can be used in alternative to *-mc* . While using *-mc*, the minimum number of leaves to retain in the tree is the same for all meta-information tags, the *-lmc* option allows to set different “minimum numbers” for different meta-information tags.

To do this the argument of *-lmc* has to be the path to a file with the corresponding “*-mc* number” for each meta-information tag, here an example:

```
CountryA,5
CountryB,10
CountryC,25
PhenotypeA,100
PhenotypeB,100
...
...
...
```

Running Treemmer with the above example will always result in trees with at least 5 leaves from countryA, 10 from countryB, 25 from countryC, 100 leaves with PhenotypeA and 100 with phenotypeB.

An example with the TB tree

Let's consider again the MTB tree with 4919 leaves (MTB_tree_trimmed_tree_RTL_0.95).

Imagine that we are interested in the tree with all L2 strains, but we want to put it in a context together with a representative diversity of strains from the other lineages. We can achieve this with the *-lmc* list_meta_count

Here is the file list_meta_count:

```
L1,50
L2,1000
L3,50
L4,50
Maf,50
L7,50
```

With this setting Treemmer retains not less than 50 strains for L1, L3, L4, L7 and Maf and no less than 1000 for L2. Since L2 has only 570 strains in the original tree, this means that no L2 strain will be pruned, since L7 has only 570 strains in the original tree no L7 strains will be pruned.

```
python Treemmer.py -lm MTB_meta_info_lineages -lmc list_meta_count
MTB_tree_trimmed_tree_RTL_0.95 -X 10
```

At 777 leaves all remaining leaves are protected, Treemmer stops and outputs the current tree:

