Treetime

By Marta Giovanetti

This document should provide you with the basic steps to build a dated time tree using Treetime.

Getting a ML Reference tree

First before we start, perhaps let take a minute to discuss how to create the input ML tree topology for TreeTime. I prefer to run ML phylogenies in IQTREE. If you have IQTEEE installed its as simple as:

```
iqtree2.1.2 -s file.fasta -fast -m GTR+G4+F -alrt 1000 -nt AUTO
```

Let's break this down. The first argument -s calls the alignment file into iqtree.
The second argument -m calls the nucleotide substitution model to be used, while the last argument -nt specify the number of cores to be used.

Installing treetime

```
pip install phylo-treetime
```

This should install treetime and all dependencies. If you don't have pip installed please do so before trying to install. Once installed you can run treetime -h to test.

```
$ treetime -h
usage: TreeTime: Maximum Likelihood Phylodynamics

positional arguments:
  {homoplasy,ancestral,mugration,clock,version}

optional arguments:
  -h, --help            show this help message and exit
  --tree TREE           Name of file containing the tree in newick, nexus, or
                        phylip format. If none is provided, treetime will
                        attempt to build a tree from the alignment using
                        fasttree, iqtree, or raxml (assuming they are
                        installed)
  --sequence-length SEQUENCE_LENGTH
                        length of the sequence, used to calculate expected
                        variation in branch length. Not required if alignment
                        is provided.
  --aln ALN             alignment file (fasta)
```

```
--vcf-reference VCF_REFERENCE
                      only for vcf input: fasta file of the sequence the VCF
                      was mapped to.
--dates DATES         csv file with dates for nodes with 'node_name, date'
                      where date is float (as in 2012.15)
--clock-filter CLOCK_FILTER
                      ignore tips that don't follow a loose clock, 'clock-
                      filter=number of interquartile ranges from
                      regression'. Default=3.0, set to 0 to switch off.
--reroot REROOT [REROOT ...]
                      Reroot the tree using root-to-tip regression. Valid
                      choices are 'min_dev', 'least-squares', and 'oldest'.
                      'least-squares' adjusts the root to minimize residuals
                      of the root-to-tip vs sampling time regression,
                      'min_dev' minimizes variance of root-to-tip distances.
                      Rerooting can be used with --covariation.
                      Alternatively, you can specify a node name or a list
                      of node names to be used as outgroup or use 'oldest'
                      to reroot to the oldest node. By default, TreeTime
                      will reroot using 'least-squares'. Use --keep-root to
                      keep the current root.
--keep-root           don't reroot the tree. Otherwise, reroot to minimize
                      the the residual of the regression of root-to-tip
                      distance and sampling time
--tip-slack TIP_SLACK
                      excess variance associated with terminal nodes
                      accounting for overdisperion of the molecular clock
--covariation         Account for covariation when estimating rates or
                      rerooting using root-to-tip regression, default False.
--gtr GTR             GTR model to use. '--gtr infer' will infer a model
                      from the data. Alternatively, specify the model type.
                      If the specified model requires additional options,
                      use '--gtr-params' to specify those.
--gtr-params GTR_PARAMS [GTR_PARAMS ...]
                      GTR parameters for the model specified by the --gtr
                      argument. The parameters should be feed as 'key=value'
                      list of parameters. Example: '--gtr K80 --gtr-params
```

```
                      kappa=0.2 pis=0.25,0.25,0.25,0.25'. See the exact
                      definitions of the parameters in the GTR creation
                      methods in treetime/nuc_models.py or
                      treetime/aa_models.py
--aa                  use aminoacid alphabet
--clock-rate CLOCK_RATE
                      if specified, the rate of the molecular clock won't be
                      optimized.
--clock-std-dev CLOCK_STD_DEV
                      standard deviation of the provided clock rate estimate
--branch-length-mode {auto,input,joint,marginal}
                      If set to 'input', the provided branch length will be
                      used without modification. Note that branch lengths
                      optimized by treetime are only accurate at short
                      evolutionary distances.
--confidence          estimate confidence intervals of divergence times.
--keep-polytomies     Don't resolve polytomies using temporal information.
--relax RELAX RELAX   use an autocorrelated molecular clock. Strength of the
                      gaussian priors on branch specific rate deviation and
                      the coupling of parent and offspring rates can be
                      specified e.g. as --relax 1.0 0.5. Values around 1.0
                      correspond to weak priors, larger values constrain
                      rate deviations more strongly. Coupling 0 (--relax 1.0
                      0) corresponds to an un-correlated clock.
--max-iter MAX_ITER   maximal number of iterations the inference cycle is
                      run. Note that for polytomy resolution and coalescence
                      models max_iter should be at least 2
--coalescent COALESCENT
                      coalescent time scale -- sensible values are on the
                      order of the average hamming distance of
                      contemporaneous sequences. In addition, 'opt'
                      'skyline' are valid options and estimate a constant
                      coalescent rate or a piecewise linear coalescent rate
                      history
--plot-tree PLOT_TREE
                      filename to save the plot to. Suffix will determine
                      format (choices pdf, png, svg, default=pdf)
```

```
  --plot-rtt PLOT_RTT   filename to save the plot to. Suffix will determine
                        format (choices pdf, png, svg, default=pdf)
  --tip-labels          add tip labels (default for small trees with <30
                        leaves)
  --no-tip-labels       don't show tip labels (default for small trees with
                        >=30 leaves)
  --keep-overhangs      do not fill terminal gaps
  --zero-based          zero based mutation indexing
  --report-ambiguous    include transitions involving ambiguous states
  --verbose VERBOSE     verbosity of output 0-6
  --outdir OUTDIR       directory to write the output to
```

Dating your tree

OK, now the first step would be to turn your ML tree topology into a dated topology (i.e. branches are in calander time). This is done with the base treetime function. Something like this:

```
treetime --aln <input.fasta> -- tree <input.nwk> -- dates <dates.csv>
```

We will add a couple of commands. First, we will specify a clock rate to use. We enforce a strict clock assumption on the data with a clock rate of 0.0008 mutations/site/year. Next, we will use TreeTime to root our phylogeny. We wish to root using the oldest sequences in the dataset. The code should look something like this:

```
treetime --aln ALIGNMENT.fasta --tree TREE.tree --dates METADATA.csv
```

And then you should see something like this:

```
$ treetime --aln 525only.fasta --tree 525only.nwk --dates Annotation.tsv --clock
-rate 0.0008 --reroot oldest


Attempting to parse dates...
        Using column 'Strain' as name. This needs match the taxon names in the t
ree!!
        Using column 'Date' as date.


0.00    -TreeAnc: set-up
```

```
114.42  WARNING: Previous versions of TreeTime (<0.7.0) RECONSTRUCTED sequences
        of tips at positions with AMBIGUOUS bases. This resulted in unexpected
        behavior is some cases and is no longer done by default. If you want to
        replace those ambiguous sites with their most likely state, rerun with
        `reconstruct_tip_states=True` or `--reconstruct-tip-states`.

196.26  TreeTime.reroot: with method or node: oldest

206.78  TreeTime.reroot: with method or node: oldest

251.99  ###TreeTime.run: INITIAL ROUND

389.30  TreeTime.reroot: with method or node: oldest

400.15  ###TreeTime.run: rerunning timetree after rerooting

547.52  ###TreeTime.run: ITERATION 1 out of 2 iterations

959.86  ###TreeTime.run: ITERATION 2 out of 2 iterations

1259.57 TreeTime: the following tips have been marked as outliers. Their date
        constraints were not used. Please remove them from the tree. Their dates
        have been reset:

1259.58 hCoV-19/India/TG-CCMB-BJ143/2021|EPI_ISL_1838719|2021-03-08, input date:
        2021.182191780822, apparent date: 2021.38

1259.58 hCoV-19/Turkey/HSGM-B4126/2021|EPI_ISL_1760146|2021-03-31, input date:
        2021.2452054794521, apparent date: 2021.52

1259.58 hCoV-19/Turkey/HSGM-B1702/2021|EPI_ISL_1678513|2021-03-29, input date:
        2021.2397260273972, apparent date: 2021.64

1259.58 hCoV-19/Turkey/HSGM-8429/2021|EPI_ISL_1403726|2021-03-03, input date:
        2021.168493150685, apparent date: 2022.11

1259.58 hCoV-19/Turkey/HSGM-11398/2021|EPI_ISL_2107334|2021-03-12, input date:
```

```
        2021.1931506849314, apparent date: 2022.11

1259.58 hCoV-19/Germany/BY-RKI-I-074298/2021|EPI_ISL_1640628|2021-03-07, input
        date: 2021.1794520547944, apparent date: 2021.57

1259.58 hCoV-19/Germany/BY-RKI-I-101273/2021|EPI_ISL_1846811|2021-03-13, input
        date: 2021.195890410959, apparent date: 2021.57

Inferred sequence evolution model (saved as 2021-06-08-0004_treetime/sequence_ev
olution_model.txt):
Substitution rate (mu): 1.0

Equilibrium frequencies (pi_i):
  A: 0.2958
  C: 0.171
  G: 0.1899
  T: 0.3333
  -: 0.01

Symmetrized rates from j->i (W_ij):
        A       C       G       T       -
  A     0       0.3282  1.0687  0.1404  15.174
  C     0.3282  0       0.2949  3.7921  20.0207
  G     1.0687  0.2949  0       1.1087  17.3364
  T     0.1404  3.7921  1.1087  0       31.8673
  -     15.174  20.0207 17.3364 31.8673 0

Actual rates from j->i (Q_ij):
        A       C       G       T       -
  A     0       0.0971  0.3162  0.0415  4.4892
  C     0.0561  0       0.0504  0.6484  3.4231
  G     0.2029  0.056   0       0.2105  3.2919
  T     0.0468  1.264   0.3696  0       10.6225
  -     0.1511  0.1994  0.1726  0.3173  0
```

```
Inferred sequence evolution model (saved as 2021-06-08-0004_treetime/molecular_c
lock.txt):

Root-Tip-Regression:

 --rate:       8.000e-04

 --r^2:        0.09


/Users/Mittenavoig/miniconda3/lib/python3.8/site-packages/Bio/Phylo/_utils.py:63
3: UserWarning: Matplotlib is currently using agg, which is a non-GUI backend, s
o cannot show the figure.

  plt.show()

--- saved tree as

        2021-06-08-0004_treetime/timetree.pdf


--- root-to-tip plot saved to

       2021-06-08-0004_treetime/root_to_tip_regression.pdf


--- alignment including ancestral nodes saved as

        2021-06-08-0004_treetime/ancestral_sequences.fasta


--- saved divergence times in

        2021-06-08-0004_treetime/dates.tsv


--- tree saved in nexus format as

        2021-06-08-0004_treetime/timetree.nexus


--- divergence tree saved in nexus format as

        2021-06-08-0004_treetime/divergence_tree.nexus


(base) MacBook-Air-di-Marta:Ed_Example Mittenavoig$
```

You can see that this runs pretty pretty slow.

With >20,000 taxa in the phylogeny TreeTime took nearly 3 and a half hours to run!!!

Next, we also see that several sequences violated the strict molecular clock assumption and either appear to be sampled before or after their actual sampling dates. These sequences/samples represent a problem. We will not be able to use this dated tree. In order to fix this, we can prune these taxa off of our original ML tree topology and try to re-run. In fact, we will continue to prune sequences off until TreeTime tells us its happy. Further, TreeTime tells us that it used a rate of 0.0008 and that the correlation of determination (r^2) is 0.77, which seems pretty good.

After closer examination of the TreeTime folder that was created we see the problem with the outlier sequences and why they need to be removed.

While there may be many ways to skin a cat, **I prefer to use the ape package in R** for this particular problem. The final R script (drop_tips_from_phylo_tree.R) is also contained within the folder. Let's have a look at the code:

```
# Set the working directory
setwd("")

# load libraries
library(ape)
library(treeio)

# load tree file
tree <- read.tree("525only.nwk")

# create vector list of seqIDs to drop from the tree
drop_tip <- c(hCoV-19/India/TG-CCMB-BJ143/2021|EPI_ISL_1838719|2021-03-08",
              "hCoV-19/Turkey/HSGM-B4126/2021|EPI_ISL_1760146|2021-03-31",
              "hCoV-19/Turkey/HSGM-B1702/2021|EPI_ISL_1678513|2021-03-29",
              "hCoV-19/Turkey/HSGM-8429/2021|EPI_ISL_1403726|2021-03-03",
              "hCoV-19/Turkey/HSGM-11398/2021|EPI_ISL_2107334|2021-03-12",
              "hCoV-19/Germany/BY-RKI-I-074298/2021|EPI_ISL_1640628|2021-03-07",
              "hCoV-19/Germany/BY-RKI-I-101273/2021|EPI_ISL_1846811|2021-03-13")

# now drop the tips from the tree
new_tree <- drop.tip(tree, drop_tip, trim.internal = TRUE)

# write the tree to file
write.tree(new_tree, file="new_tree.nwk", append = FALSE)
```

This produces a new ML tree topology but with the above mentioned taxa removed. Running TreeTime on this new_tree.nwk tree file will produce output akin to this:

```
treetime --aln 525only.fasta --tree new_tree.nwk --dates Annotation.tsv --clock-
rate 0.0008 --reroot oldest
```

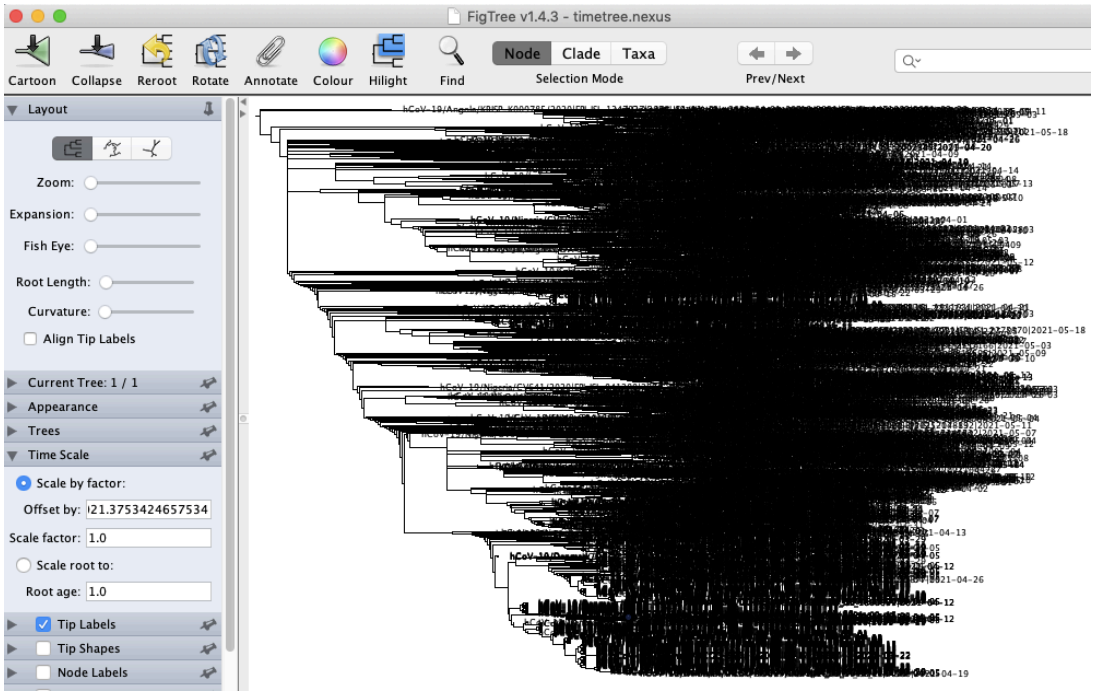
```
Attempting to parse dates...
       Using column 'Strain' as name. This needs match the taxon names in the t
ree!!
       Using column 'Date' as date.

0.00   -TreeAnc: set-up

72.48  WARNING: Previous versions of TreeTime (<0.7.0) RECONSTRUCTED sequences
of
       tips at positions with AMBIGUOUS bases. This resulted in unexpected
       behavior is some cases and is no longer done by default. If you want to
       replace those ambiguous sites with their most likely state, rerun with
       `reconstruct_tip_states=True` or `--reconstruct-tip-states`.

120.77 TreeTime.reroot: with method or node: oldest

126.48 TreeTime.reroot: with method or node: oldest

149.98 ###TreeTime.run: INITIAL ROUND

221.17 TreeTime.reroot: with method or node: oldest

226.34 ###TreeTime.run: rerunning timetree after rerooting

308.16 ###TreeTime.run: ITERATION 1 out of 2 iterations

560.19 ###TreeTime.run: ITERATION 2 out of 2 iterations

Inferred sequence evolution model (saved as 2021-06-08-0001_treetime/sequence_ev
olution_model.txt):
Substitution rate (mu): 1.0

Equilibrium frequencies (pi_i):
  A: 0.2958
  C: 0.171
  G: 0.1901
  T: 0.3332
```

```
  -: 0.01


Symmetrized rates from j->i (W_ij):
       A       C       G       T       -
  A    0       0.332   1.0495  0.1413  15.2328
  C    0.332   0       0.2833  3.8002  20.0162
  G    1.0495  0.2833  0       1.1215  17.4854
  T    0.1413  3.8002  1.1215  0       31.6637
  -    15.2328 20.0162 17.4854 31.6637 0


Actual rates from j->i (Q_ij):
       A       C       G       T       -
  A    0       0.0982  0.3104  0.0418  4.5055
  C    0.0568  0       0.0484  0.6499  3.4231
  G    0.1995  0.0538  0       0.2131  3.3232
  T    0.0471  1.2662  0.3737  0       10.5501
  -    0.1517  0.1993  0.1741  0.3153  0



Inferred sequence evolution model (saved as 2021-06-08-0001_treetime/molecular_c
lock.txt):
Root-Tip-Regression:
 --rate:       8.000e-04
 --r^2:        0.09


/Users/Mittenavoig/miniconda3/lib/python3.8/site-packages/Bio/Phylo/_utils.py:63
3: UserWarning: Matplotlib is currently using agg, which is a non-GUI backend, s
o cannot show the figure.
  plt.show()
--- saved tree as
        2021-06-08-0001_treetime/timetree.pdf


--- root-to-tip plot saved to
       2021-06-08-0001_treetime/root_to_tip_regression.pdf


--- alignment including ancestral nodes saved as
        2021-06-08-0001_treetime/ancestral_sequences.fasta
```

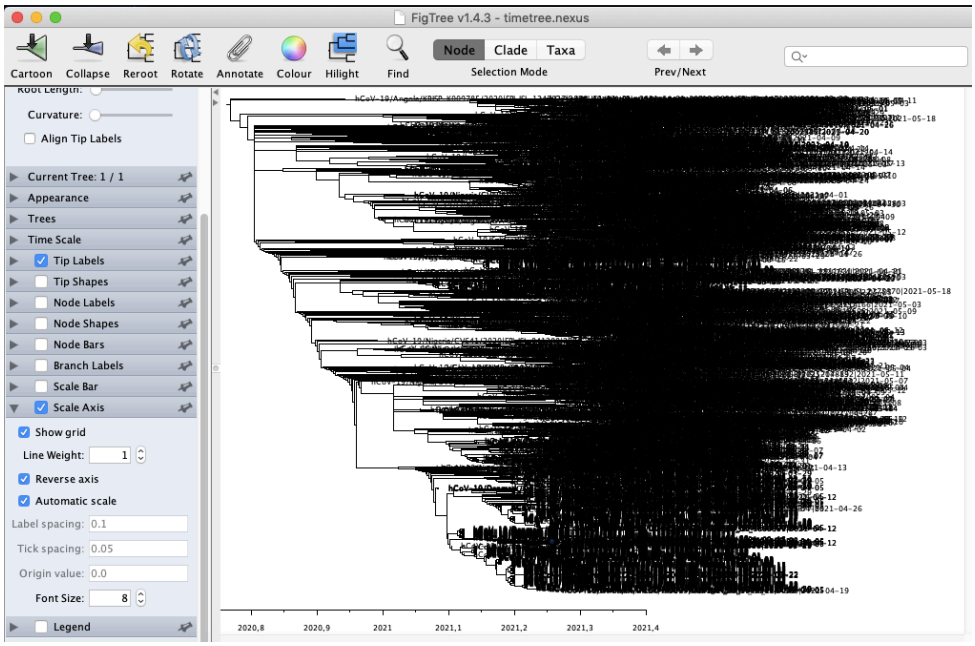
```
--- saved divergence times in
        2021-06-08-0001_treetime/dates.tsv

--- tree saved in nexus format as
        2021-06-08-0001_treetime/timetree.nexus

--- divergence tree saved in nexus format as
        2021-06-08-0001_treetime/divergence_tree.nexus
```

You will notice TreeTime successfully completed a correct analysis, so finally we will proceed analyzing the dated tree obtained using **Figtree.**
Since the most recent sampling date is "2021.3753424657534" you should use it to have a proper scale axis.

To do this: go to "Time scale" on the right side and put this date into the box as showed below:



Then uncheck the "Scale Bar" option and select the "Scale Axis" with "reverse axix". See below:



Now play with the command just seen to show this tree in a fancy way! Enjoy!!