Alignment and phylogeny tutorial

Rayna Hamilton

April 24, 2025

This writeup is meant to serve as a brief overview of the process I normally follow when generating phylogenies for the Alexander lab as well as the basic theory behind the process. Scripts used throughout the tutorial will be located in this Github repository as well as in the /proj/omics/alexander/rhamilton/2025-phylogeny-tutorial directory on Poseidon.

Preparing your input data

Let's assume that you have already decided on a sequence of interest and target set of organisms. You will need all sequences in one FASTA format file for alignment. If you are working with protein-coding DNA, you should also decide whether to use amino acid or nucleotide data. In general, protein sequences evolve more slowly and will allow for reliable alignment of more distantly-related organisms, but nucleotides (particularly rapidly-evolving introns) can allow you to differentiate between more similar organisms.

If your sequences are all about the same degree of relatedness to each other and you do not know which was the first to diverge evolutionarily, then include at least one outgroup sequence that is more distantly related to other members of the tree, but still similar enough to align well. This will be used to root the tree. Tree rooting is important because most phylogenetic methods do not determine the root, that is, the common ancestor of all taxa within the tree. Indicating a sequence or group of sequences which we know diverged first from the main lineage allows us to visualize the direction of evolutionary time. Here I am working with a set of ammonium transporters with share a common evolutionary origin across the tree of life (including eukaryotic and prokaryotic transporters). Most of my sequences are eukaryotic, but I included a pair of Bacteria ammonium transporters to root the tree. You do not need to go for this wide of an evolutionary range if your sequences are not as taxonomically diverse, e.g. a non-*Pseudomonas* Gammaproteobacteria would be acceptable for making a phylogeny of *Pseudomonas* members.

The tutorial relies on IQTree for phylogeny as well as Clustal Omega or Muscle for alignment. If you do not have these installed, you can make a conda environment using the sample environment yaml file:

```
Conda env create -n phylogeny --file=sample_env.yaml
```

Sequence alignment

Decide whether you would like to use Muscle or Clustal omega for alignment. I have found that, with default parameters, Muscle is more likely to make gaps when it is not confident that sequence regions are truly homologous, while clustal is more likely to align nonsynonymous bases. Modify the relevant .sh script to point to your conda environment name and sequence input/output filenames.

Submit the alignment job to the slurm job scheduler:

sbatch 1_align_muscle.sh

or

sbatch 1_align_clustal.sh

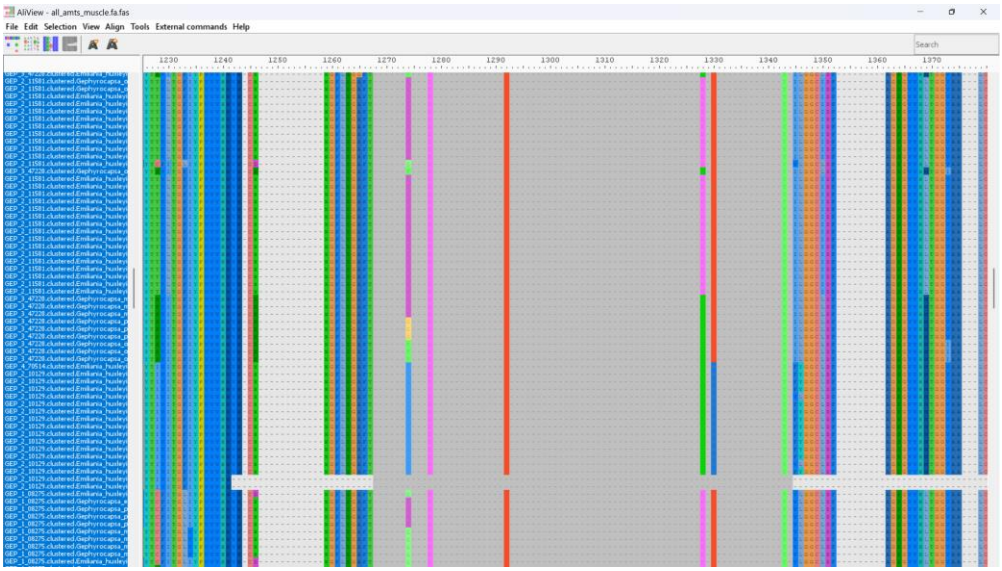
Your output alignment will be in aligned fasta (.fas) format. For smaller datasets, it is recommended to manually examine the alignment and remove low-quality regions. Download the output file to using scp or sftp.

Install a graphic alignment editor such as AliView, available at https://ormbunkar.se/aliview/

Assuming that you are not working with organisms as highly diverged as this set, your raw alignment probably will not look this messy. In general, regions of an alignment which are almost entirely gaps contain very little information for phylogenetic programs. Additionally, phylogenetic methods rely on the assumption that all bases in a column evolved from the same ancestral base. As such, poorly-aligned regions where it is not clear that any real homology is present will worsen phylogenetic accuracy. We should manually remove these regions or, if you can see how the bases should align, realign these regions specifically.

You may find yourself removing chunks at the start and end of the alignment until a region shared across all sequence members is found. In this case there is a high degree of evolutionary divergence so I will be removing and deleting quite a bit.

Example of a region that you may want to just remove (in Aliview, highlight the relevant regions then Edit -> Delete Selection. Be sure to highlight all sequence rows or your columns will get out of sync):



If some sequences seem like they are not aligning with any others or are substantially shorter than the others, you can remove these at your discretion.

After some deletion and re-alignment of chunks (Align -> Realign selected block in Aliview), my alignment looks like this:

Rename and re-upload your clean alignment using scp or an sftp connection.

Phylogeny

Adjust the 2_iqtree.sh script to indicate your conda environment and filename. Be sure to replace any spaces, colon, square bracket or comma characters in your input as iqtree will cut off sequences ids or replace these characters. To replace spaces with underscores and remove everything else, something like the following will work:

```
cat my_clean_alignment.fa.fas | sed 's/ /_/g' | tr -d "[" | tr -d "]" | tr -d ":" | tr -d "," > my_really_clean_alignment.fa.fas
```

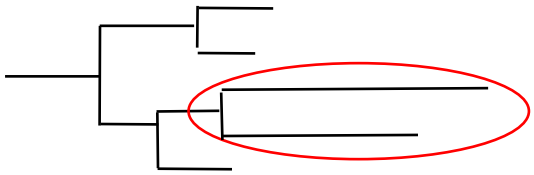
Submit your phylogeny job to slurm:

```
sbatch 2_iqtree.sh
```

The .treefile output produced by iqtree contains the NEWICK format output. You can visualise and customize the appearance of this tree using the Interactive Tree of Life (https://itol.embl.de/). If you are in the Alexander lab, you will probably want to sign in to our account and make your own workspace (ask Harriet!). Once signed in, click Tree upload in the top-left corner and paste in or upload your treefile.

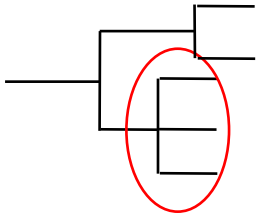
Some warning signs to look out for:

Long-branch attraction:



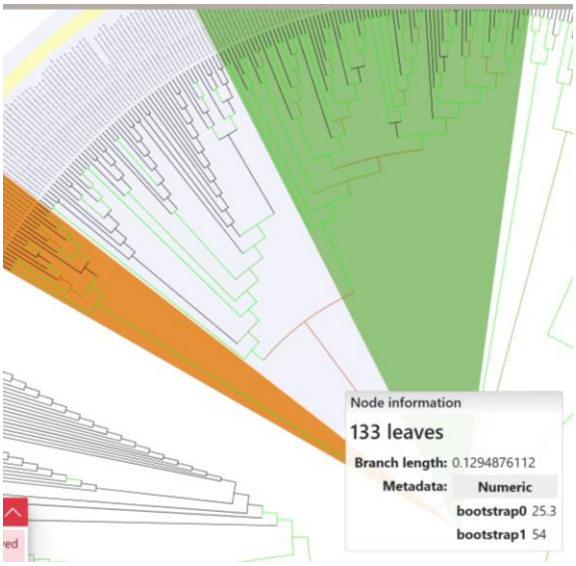
Branch lengths in a tree represent evolutionary distance, or how diverged sequences are from each other. Sequences that are very different from other sequences in the tree will often group together regardless of whether they are similar to each other. This is fine if the grouping is taxonomically meaningful (e.g. the grouping of Pseudomonadota and Cyanobacteriota species in the below tree because they are both bacteria in an otherwise eukaryotic tree). However, it is cause for concern if you do not expect the sequences to group together. Unexpected long branches can be indicative of issues with your dataset, such as poor alignment.

Unresolved nodes:



Multiple branches diverging from the same ancestral node at the same time indicates that your data is not sufficient to determine the order that these sequences diverged from a common ancestor, or that different bootstrap replicates often yielded different results. This is sometimes unavoidable if two sequences are just very similar, but if it is a common occurrence throughout your tree you should consider re-aligning your input data, removing poorly-aligned regions or selecting a different gene to align.

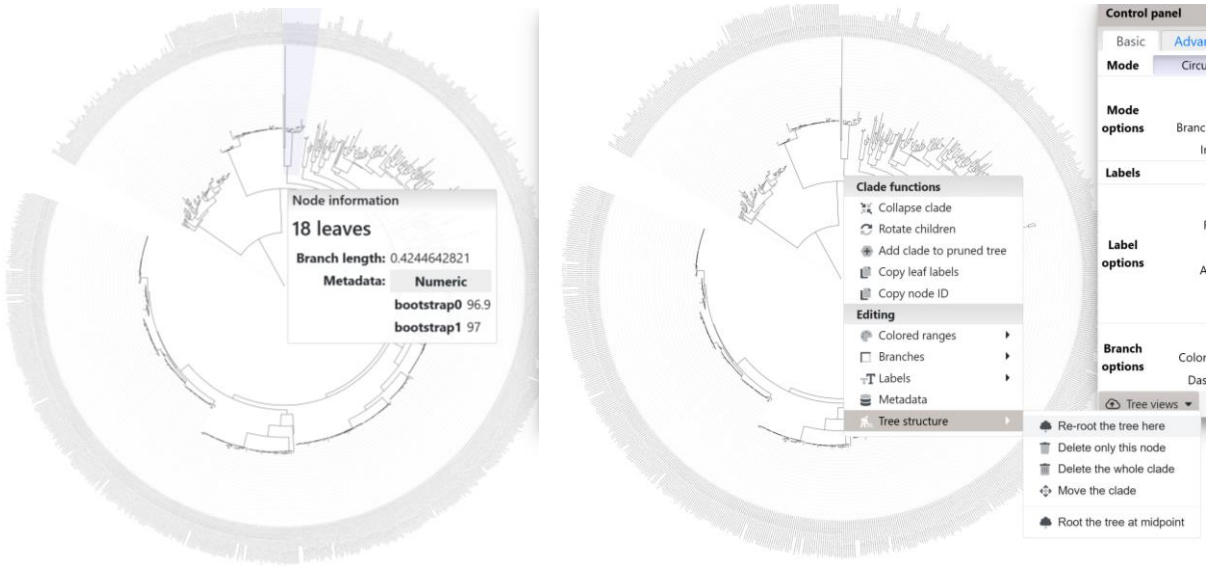
Very low bootstrap values:



Bootstrap values are produced by phylogenetic methods such as maximum likelihood and represent the robustness of a branch, that is, how frequently the sequences in that branch were grouped together when making multiple phylogenies with different subsets of the alignment. These range from 0 to 100, with 100 indicating that all bootstrap replicates yielded the same grouping. Lower bootstrap values indicate less robust branches that are not strongly supported by your data. As above, if your tree has many such branches you should re-evaluate your input data.

Assuming that your tree does not have any of the above issues, let's move on to iTOL visualization.

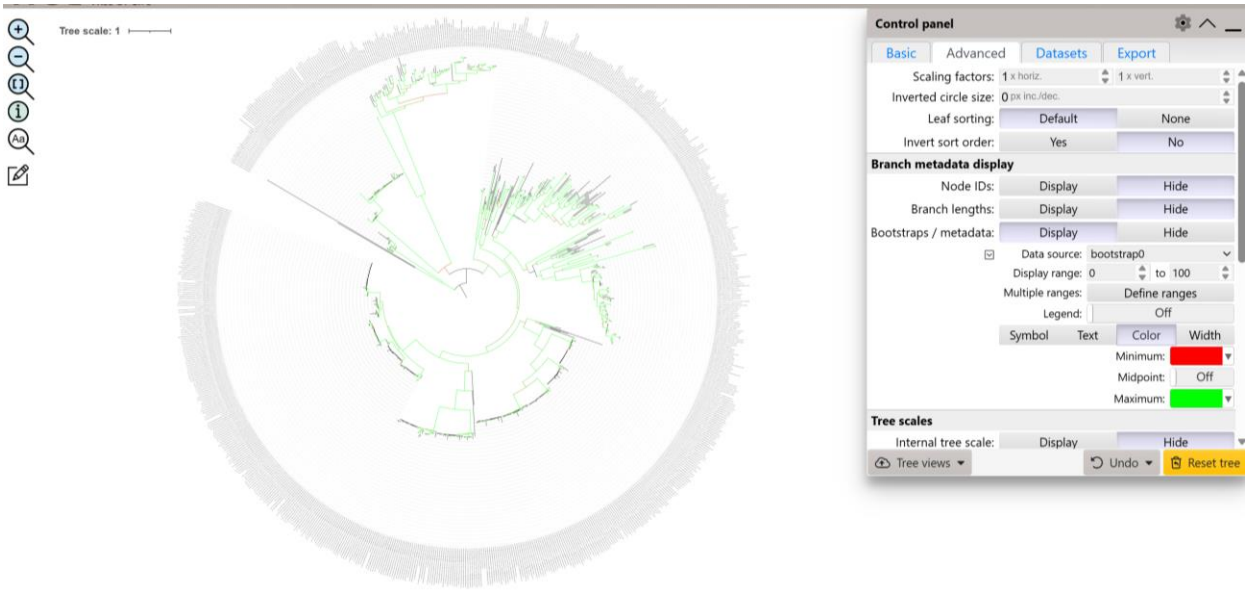
Locate the branch containing the sequence or group of sequences you designated as your outgroup and click it, then click Tree structure -> Re-root the tree here.



Now you can begin customizing your tree. You may, for example, want to switch between rectangular and circular mode in the basic tab, or display bootstrap values in the advanced tab. Given that the tree I am working with is too large to see the individual bootstrap numbers, I am going to display the values as a colour scale.

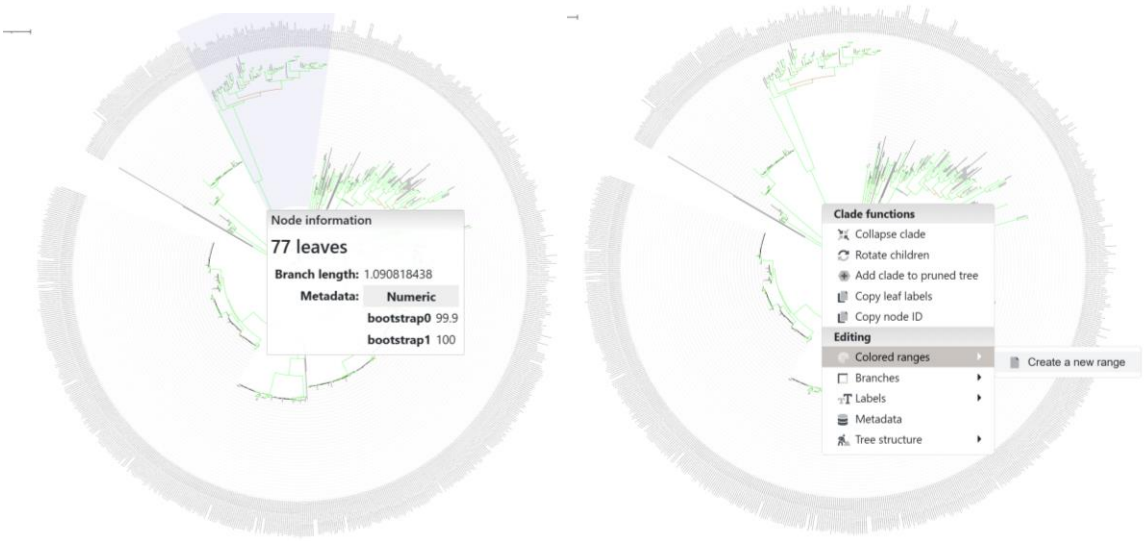
Within control panel:

Advanced -> Bootstraps / metadata -> Display

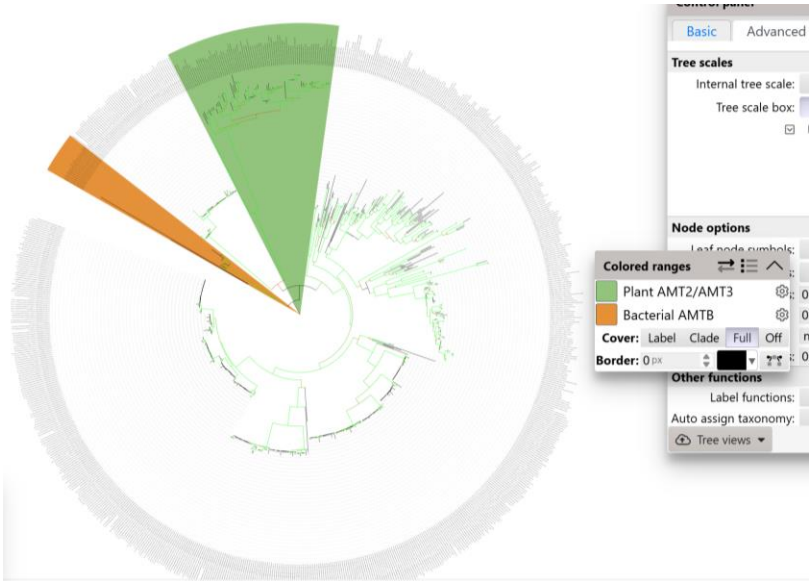


A full description of iTOL customization methods is available at https://itol.embl.de/help.cgi.

A big part of using iTOL to make publication-quality trees is producing annotations describing attributes of your dataset e.g. taxonomic group, location of isolation, genome completeness. You can colour regions of a tree manually by highlighting branches and clicking Colored ranges -> Create a new range and selecting your desired colour. Here I have indicated that all the plant AMT2/AMT3 transporters occupy a single branch.



After colouring branches, you can play around with the colored ranges menu to adjust branch colours, names and how much of the clade is covered.



This process is, however, often insufficient for datasets where the information is not categorical or does not perfectly co-occur with the tree branches. As such, you may instead find it more practical to upload iTOL-format annotations. The Tree annotation section in the iTOL help page (https://itol.embl.de/help.cgi#annot) has a variety of sample input files available for different datatypes. There is also an iTOL excel plugin (https://itoleditor.letunic.com/). I have also added the sample make_itol_annotation.py script which generates an annotation file from csvs of sequence-to-category and category-to-colour information. To add an annotation file to iTOL, use the Datasets -> Upload annotation file. Please be aware that your sequence names in the treefile must match the annotation file exactly to be mapped. After addition of datasets, my final ammonium transporter tree looks like this: