Tree Building and Phylogenetic Analysis

Contents

- 4.2.1b: Tree building
  - 4.2.1b.1: Preparation
  - 4.2.1b.2: Alignment
  - 4.2.1b.3: Model selection
  - 4.2.1b.4: Tree building
  - 4.2.1b.5: Tree visualisation
  - 4.2.1b.6: FastTree

4.2.1b: Tree building

In this practical we will be analysing and interpreting best fitting substitution models, bootstrap values and finding the best ways to visualise trees. We will also be extracting useful metadata and annotating this on to phylogenetic trees.

4.2.1b.1: Preparation

Start by making a folder for the analysis in your home directory. Call this `YOURNAME_Phylogenetics_analysis` replacing YOURNAME with your first name (no spaces or special characters except _ please!)

You'll then need to copy the fasta file `nig-af2-seqs.fasta` we'll use for this practical from the `Home/RAGE-workshop-2024/day4/Phylogenetics/Data` folder to the folder you just made.

4.2.1b.2: Alignment

To make trees, sequences first need to be aligned! We will do this using mafft.

Make sure you're in your `YOURNAME_Phylogenetics_analysis` directory you just made.

Task 1

Enter the conda environment by typing `conda activate MADDOG_backup`

Write the command `mafft nig-af2-seqs.fasta > nig-af2-seqs_aligned.fasta`

This will create the file `nig-af2-seqs_aligned.fasta` in your folder - this contains the sequences from `nig-af2-seqs.fasta` aligned by the FFT-NS-2 algorithm.

Task 2

It's generally a good idea to take a quick manual look at the alignment file before proceeding. You can open the fasta file in an alignment viewer for a full look, but for now just click to open the text file.

Task 3

How many sequences does this fasta file contain?

You can either click on the file to open it in a new window and then use ctrl + f to find the '>' character

OR

You can use commands!

Hint: You can use `grep` to search the file, and search for the `>` character, of which each will denote a sequence, and `wc -l` to count lines. Use `|` to pipe one into the other.

4.2.1b.3: Model selection

Finding the substitution model that best fits the data is very important. There are various tools to estimate the best substitution model but for now, we're going to use the inbuilt modeltest function in IQTREE.

About the program - IQTREE

This program allows you to perform Maximum Likelihood phylogenetic analysis. It uses efficient algorithms to explore the tree space, allowing very large matrices to be analyzed with reliable results (hundreds / thousands of sequences). It allows estimating the evolutionary model (ModelFinder module) followed by the phylogenetic analysis and implements support measures to evaluate the reliability of the groupings (Bootstrap, Ultrafast Bootstrap Approximation and probabilistic contrasts). The program can be downloaded and run locally, or on online servers such as: http://iqtree.cibiv.univie.ac.at/ https://www.phylo.org/ https://www.hiv.lanl.gov/content/sequence/IQTREE/iqtree.html You can find many basic and advanced tutorials in http://www.iqtree.org/doc/

Task 4

In the terminal, make sure you're still in the correct directory and the conda environment.

Type: `iqtree -h` to see all the command options for IQTREE, and note any that may be useful.

You may notice the `-m` argument to specify a model. You can run `-m TEST` when making a tree to find the best substitution model and then make a tree using that model. Alternatively, you can use `-m TESTONLY` to find the best fitting substitution model without making a tree.

Other important arguments we'll be using include:

`-s` to specify the name of the alignment file,always required by IQ-TREE to work.

`-B` to specify the number of replicates for ultrafast bootstrap in IQ-TREE v2. Slightly less accurate than regular bootstrapping, but MUCH faster.

`-b` to specify the number of replicates for regular bootstrapping. Either this or `-B` can be used, not both.

To start with, we'll run the modeltest without running a tree.

Task 5

In the terminal, type `iqtree -s nig-af2-seqs_aligned.fasta -m TESTONLY`

What is the best fitting model according to corrected akaike information criteria? Make a note of this!

Hint: This will be on the terminal screen, you may have to scroll up a little to under where each model was tested.

Task 6

What do the parameters of the best fitting model mean?

4.2.1b.4: Tree building

Now we know the best fitting model, we can make our tree!

Task 7

In the terminal write `iqtree -s nig-af2-seqs_aligned.fasta -m BEST_FITTING_MODEL -B 1000 -redo` replacing `BEST_FITTING_MODEL` with the model you found in your last step.

This command means we are building this tree with 1000 ultrafast bootstrap replicates. The `-redo` argument is necessary to overwrite the files made in the last step and allow this to run.

You'll notice a lot of new files appear in your directory. The ones we're most interested in are the `.contree` file, which we'll use to visualise our tree, and the `.log` file which contains all the information that appeared on the terminal window.

4.2.1b.5: Tree visualisation

Now you've built a tree, it's time to visualise it. We'll be using FigTree. FigTree is designed as a graphical viewer of phylogenetic trees and as a program for producing publication-ready figures. It can be downloaded from: http://tree.bio.ed.ac.uk/software/figtree/

We have installed FigTree on your SSDs. To access it you'll need to leave the MADDOG environment by typing `conda deactivate` and enter the visualization environment by typing `conda activate visualization`

Task 8

Open figtree by typing `Figtree` into the terminal and pressing enter.

Click File → Open → select the file `nig-af2-seqs_aligned.contree` that you just downloaded.

When prompted, enter `bootstrap` as the label name.

Your tree will appear on the screen! However, it needs a bit of manipulation to be in its most useful form.

Along the side tabs, go to Trees. Root the tree at the midpoint and order the nodes by increasing order.

Task 9

What initial observations can you make from the tree topology?

Task 10

Now go to Node labels, and display the bootstrap values.

What is the purpose of bootstrap values?

Task 11

It often isn't useful to see ALL the bootstrap values - we just want to know which branches have strong support and which don't. Therefore, we can hide the lower values and only show values of interest.

To do this, click `ctrl+f`. A search bar will open at the top of the figure. Search for bootstrap values less than 90. Click `Node` in selection mode, then click the `annotate` paperclip. This will bring up a box. Select `bootstrap` from the drop down list, leave the second box blank, and apply.

This will remove any bootstrap values less than 90.

Look at the bootstrap values on the tree now. What do they tell us?

Task 12

Spend some time now playing about with the settings in FigTree to make an aesthetically pleasing tree! Make sure to keep it open when you're done, as we'll be going back to this.

4.2.1b.6: FastTree

IQTree isn't the only tool we can use for tree building. While this dataset was quite small, and therefore ran quite quickly, datasets can quickly become more complex and take a long time! FastTree (http://www.microbesonline.org/fasttree/#Usage) is much faster, and therefore may be a good first step to assess rough tree topology.

Task 13

We're going to be using the same dataset. The standard usage of FastTree is `FastTree -gtr -nt alignment_file > tree_file`. This uses the GTR model. The output will be a newick tree.

In the terminal, make sure you're in the folder you created that has your phylogenetic data in and in the visualization conda environment.

Write `FastTree -gtr -nt nig-af2-seqs_aligned.fasta > nig-af2-seqs_aligned.nwk` and press enter.

Task 14

When the program has finished, open your newick tree in FigTree. Compare this to the IQTree tree.

Do they differ?