Sequence bioinformatics, computer exercise XIII

University of Potsdam, WS 2023/2024, Stefanie Hartmann

Give it all you've got! And since the mere presence of smartphones reduces available cognitive capacity (`https://www.journals.uchicago.edu/doi/full/10.1086/691462`), I suggest you put your cell phone out of sight and into your bag for the duration of this computer lab.

Inferring phylogenies II

Objective

Today you'll compute a ML tree, using the same sequences from which you computed a NJ tree last week. The computation of the tree won't take too much time, and the main goal for today's lab is to read, analyze, and interpret today's and last week's trees.

1 Locally installed software you'll need

You'll again need a tree viewer, e.g., FigTree, available for all operating systems at `https://github.com/rambaut/figtree/releases`. If you wish to view the alignment again, you'll also need an alignment viewer (e.g., JalView).

2 Logging into the cluster and preparing for the analyses

☞ Log into the university's computing cluster as described in the PDF of the first week and change to your directory for this class on the /work partition. Create a new directory for today's exercise (e.g., lab13) and change into it.

☞ You'll use the RAxML software today. Load the corresponding module with the command `module load bio/RAxML/8.2.12-foss-2020b-pthreads-avx2`

☞ Start an interactive session: `srun --cpus-per-task=1 --mem=5G -p teaching --pty bash`
Please coordinate with your neighbors so not everyone starts the interactive session at the same time.

3 The data

Please use for today's exercise the same alignment in phylip format that you used last week.

4 Computing & viewing a Maximum likelihood tree

RAxML (Randomized Axelerated Maximum Likelihood) was written by Alexandros Stamatakis. The software is fast enough to be of use for large data sets. Look at a summary of the available options by calling the program (here: raxmlHPC) with the `-h` option. You can see that RAxML is a very powerful analysis package that can be fine-tuned for many different types of analysis. Much more information and an extensive manual in PDF format is available at the author's webpage (currently `https://cme.h-its.org/exelixis/web/software/raxml/`).

In contrast to the NJ analysis using PHYLIP, RAxML can do in a single step the analysis of a maximum likelihood tree, the generation and evaluation of bootstrap data sets, and the mapping of the support values onto the best ML tree. You can root the tree with the cyanobacterial sequences.

Here is a sample RAxML call and what it specifies:

```
raxmlHPC -T 1 -#10 -s tocs.phy -n tocs -m PROTGAMMAWAG -x 1357 -p 12345 -f a -o pro_AAQ,pro_CAE
```

-fa compute the best-scoring ML tree, also do a bootstrap analysis program run

-#10 compute 10 bootstrap data sets

-T if the parallelized version of RAxML has been installed (which is the case here), you can use this option to specify the number of CPUs you want to use for the analysis. Since the srun session was called for just one CPU, one is what you'll use here

-s use as input the alignment file in phylip format (adjust the name if necessary)

-n use the stem "tocs" in all generated output files (or adjust if you want it called something else)

-o allows an outgroup to be specified. Use the cyanobacterial sequences.

-m specifies the particular model of amino acid substitution to be used for the analysis. (In this case, the 'WAG' substitution matrix is used with a gamma model of rate heterogeneity.)

-x expects a random seed for the bootstrapping analysis

-p expects a random seed for the parsimony analysis (the search, per default, starts with a parsimony tree)

☞ Execute the RAxML code as shown above, adjusting for file names if necessary.

☞ The call generates five files, take a look at each of them with less:

- RAxML_info.tocs: information about the model and algorithm used, how RAxML was called, about progress of the run, and some results. There are a lot of details here; for now just get an overview of what kind of information is provided here
- RAxML_bestTree.tocs: the best-scoring ML tree of the original alignment
- RAxML_bootstrap.tocs: trees for all 20 bootstrap replicates
- RAxML_bipartitions.tocs: the best-scoring ML tree of the original alignment with confidence values from 0 to 100 drawn on it.
- RAxML_bipartitionsBranchLabels.tocs: the same information as in the previous file, but in a slightly different format (support values are given as branch labels).

☞ Use `scp` to copy the file RAxML_bipartitions.tocs to your local machine and open it in FigTree.

☞ When you open the bipartitions file, you'll be asked what to call the bootstrap values. Leave the generic 'label' or change it to 'bootstraps'.

☞ You can select the bootstrap values to be shown by selecting either in the 'Node Labels' or in the 'Branch Labels' menu the option 'Display: label'.

☞ Which groups are well supported, which are not?

☞ Also add the NJ tree from last week in a separate window, so you have them side by side. For easier comparison, you can change for both of them the display in the 'Trees'-Menu to 'order nodes: decreasing'.

☞ Is the ML tree the same as the NJ tree from last week? If not, where are the differences?

☞ Name one pair of genes on the gene trees that are paralogs, one pair that are orthologs, and one pair that are homologs.

☞ Optional: Adjust the display to your liking and export the image as a PDF so you'll be able to look at it later, without opening FigTree.

5 Exiting the terminal

Log out of the interactive session with `exit`.

Then log out of the current session to the cluster, type `exit` again.

Finally, to cleanly close the terminal window, type `exit` yet again.

Do not close a terminal window with the mouse, especially not if an ssh session is still active.

6 Analysis & interpretation questions

6.1 Species tree

Knowing the species relationships is important for the interpretation of a gene tree. However, and as discussed in lecture, the relationships of species can't necessarily be inferred from a gene tree. You therefore need to obtain the species' relationships independently.

☞ Look up the relationships of the eukaryotic species from which there are sequences in the computed gene tree. One way to do this is to use NCBI's "Taxonomy Common Tree":

- direct your browser to `https://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi`
- at the top, in the box next to "Enter name or id" enter the name of the first species name and click on the "Add" button. Repeat with the next one. You'll find a list of the species name in the BLAST exercise.
- once you have added all species names, you can view the tree online. You can also export the tree in phylip format: Click on the gray button 'text tree' and set it to 'phylip tree'. Click on "save as" to download the file.
- open the file in FigTree. The 'labels' now correspond to the names of lineages: make them visible by selecting the check box of the "Node Labels" menu.

☞ Which of the species are green algae? red algae? angiosperms? monocots? dicots? Which species are part of the "Viridiplantae" (green plants)?

☞ Does this species tree contain any polytomies? If so, what does this mean?

6.2 Species tree vs gene tree

☞ There are four major clades in the two gene trees you computed. They should roughly be the same between the two gene trees, although the exact relationships of the sequences within each of these clades might differ:

- clade 1 contains sequences from plants and green algae (i.e., viridiplantae)
- clade 2 also contains sequences from viridiplantae (and for labelling purposes, it doesn't matter which one you call clade 1 and which one clade 2.)
- clade 3 contains sequences from red algae
- clade 4 contains sequences from cyanobacteria

☞ Re-draw (by hand) the ML and NJ trees you computed, reducing it to the four clades mentioned above. For example (((clade1,clade2),clade3),clade4). Are the relationships of these four clades different in the two different gene trees? What are the bootstrap values for these clades?

Reducing the species tree to these three taxonomic lineages, we get ((viridiplantae, red algae), cyanobacteria).

☞ For each of the ML and the NJ tree (if they're different), interpret the gene tree in the context of the species tree: In which taxonomic lineages (i.e., on the species tree) did gene duplications or losses occur?

☞ If there are differences between the ML and NJ tree: summarize the differences.

☞ The alignment I provided only contains the conserved functional domain of the sequences, as determined by Pfam. Do you think that using only part of the gene sequence for computing a phylogeny is a good idea in general and in this specific case? Justify your answer.

☞ In the alignment I provided, I excluded one of the sequences (that was part of previous analyses) because it is a pseudogenes. What is a "pseudogene", and why should a pseudogene not be included in a phylogeny?

☞ The sequences you analyzed during the last several computer labs are all members of the Toc75/Oep80 gene family. In plants and algae, these are part of the plastid import machinery and trace their common origin back to a cyanobacterial gene sequence. Much more information about the biology and evolution of these genes is available in a recent publication (`https://doi.org/10.1016/j.isci.2020.100896`) that also includes a phylogeny. Check it out if you like! And compare their tree to the ones you computed. Which of the clades on your phylogenies corresponds to the Toc75 sequences, and which to the Oep80 sequences?

➜ Don't forget to take your flash drive with you when you leave!