lab7_questions_and_maj_answers

Author: Mohammed Al-Jaff
Genomics & Bioinformatics: lab 7

Question 1: Which files do IQ-TREE output? Explain briefly what each of them is.

Ans: [SOURCE: http://www.iqtree.org/doc/iqtree-doc.pdf and some biostar forum threads]

According to the docs for IQ-tree the following files outputted bye the program contain/represent:

```
.ckp.gz
```

The 'checkpoint' file which IQ-tree saves to disk periodically throughout the run so that, in case of any run interruption, the program doesn't not have to start again from scratch. Apparently, this file also acts as a safeguard against unnecessarially rerunning the same analysis twice because iq-tree will be aware of its perchance and notify the user that the inputed analysis has already been performed.

--

```
.iqtree
```

Contains the report of the reconstruction run/analysis by IQ-tree. The idea here is that this is a human readable (and understandable) report detailing what has be opted for by the user in the phylogenetic tree reconstruction instructions to IQ-tree and the resulting analysis done by the program. Included here is basically metrics for

--

```
.log
```

Contains basically the same information that was outputted in STDOUT in the command line.

```
.model.gz
```

Contains the log-likelihood values of for all the performed on the 280 or so (substitution) models. Note that this is the compressed file indicated by the .gz extension.

```
.mldist
```

Contains a distance matrix with elements beeping the pairwise 'maximum likelhood' distance between each sequences from the best found model.

The "tree files" all contain tree's in NEWICK" notation [which is a nice parenthesis-and_comma-based approach to representing trees in a linear script]. The Maximum-likelihood tree is in the *.treefile*, the so-called consensus tree is in the *.contree* file. Lastly, in the *.bionj* file there is a tree arrived at through a version of the neighbourhood-joining algorithm. Depictions of the consensus tree and the maximum likelihood trees are also in the *.iqtree* report file.

```
.splits.nex
```

Some type of matrix with results from the bootstrapping tree reconstructions procedure. I think the branch support values in our consensus tress com from here somehow.

Question 2 & 3: Which model did ModelFinder choose? From all the criteria calculated by this software, which was used to determine the best-fitting model? & Briefly explain the best-fitting model.

Ans:

ModelFinder found that the substitution model TIM2+F+I+G4 was 'best fitting' to the cytB multiple alignment given. The criteria used to select the best-fit model seems to be the 'BIC' measure/metric [Bayesian information criteria]

From what is think I know currently; we should be thinking of substation models not only dealing with mutation 'rates' between biases but also make assumptions on base frequencies of the occurrences of each base. [perhaps this gets used as the prior in getting at a bayesian value?].

The TIM2 part of this substitution model indicates that we are assuming i) unequal base occurrence probabilities, ii) that we have only 2 transversion mutation rates [ie purine-to-purin does not have to be equal to pyrimidine-pyrimidine mutation rate] and iii) unequal transition rates [ie purine-pyrimidine mutation rates].

The +F part i think indicates that we use the "empirical" base occurrence probabilities that we see directly from the given sequences in our multiple alignment.

The +I+G4 means that the substitution rates can be position dependent, ie various regions of the sequences need not have the same rate of substitutions. The distribution of how much the rates can vary is assumed to be gamma distributed.

Question 4: Now look at both your Maximum Likelihood tree and Consensus Tree. Are they the same? If not, where do they differ?

Ans: Structurally(topologically?), both trees are the same [minus the visualisation of outgroup clad] apart from the branch support values which seem to differ a bit between them.

Question 5: In both trees you can see a number at the base of each branch. That is the number of iterations that supported that branching during bootstrapping. Which is your least supported branch? What does that mean to your question?

Ans:

In the maximum-likelihood tree, the least supported branch is "inside" the 'lizard' clad splitting the iguanas from the rest of the lizards [minus geckos which slept on the part node]. For our project question, this doesn't alter the conclusion. The conclusion here being that salamanders are more related to frogs than salamanders, because of their mutual common ancestor. This agrees with the current consensus phylogeny.

This is the same situation in the consensus tree but where the least supported branch splits up the lizards species in another fashion. In any case, this least supported branching does not influences the outcome of the question because we have the same overall "conclusion" that salamanders have a closer common ancestor to frogS than they do with lizards.
