Gene trees species tree estimation

We will use a full Bayesian framework for species tree estimation. The statistical approach is called *BEAST2 (see Heled and Drummond, 2010; Mol Biol Evol). You will need the following software available : BEAST2, Tracer and FigTree. We will infer the species tree of clownfish that is most probable given the multi-locus sequence data available in the seven genes that we used in the practical.

1 Setting up and running the analyses

BEAST2 uses a specific *XML* format, but there is a software called BEAUti to easily create the input data file.

1. open BEAUti and install the package StarBeast2 (menu `File` → `Manage packages`). You might have to restart BEAUti
2. in BEAUti, select `File` → `Template` and choose `StarBeast2`
3. import successively the six DNA regions for the clownfish (drop the files or click the '+' at the bottom of the window). It can read directly the `fasta` files, but make sure that it load the data as **nucleotides**
4. the DNA regions are all from different chromosomes and we can expect recombination to happen between them. We should thus unlink the topologies for each of these regions (column `Tree` in the `Partitions` tab)
5. set `Taxon sets` to define the different species that we have. You can use the `Guess` button for this, click the `split on character` and enter "." and take group 2
6. leave the `Gene Ploidy` level to 2.0 for the nuclear DNA and to 1.0 for the mtDNA
7. leave the `Population Model` to the default value, which is called `Analytical Population Size Integration`
8. set the `Site Model` for each DNA region to what was estimated for the PhyML analyses. Take particular care to the following parameters:
   - do not estimate the substitution rate. It should be normalized to 1 as explained in the first part of the lecture
   - if the model uses a Gamma distribution, set the `Gamma Category Count` to 4 and estimate the `Shape` parameter
   - if the model uses a propostion of invariant sites, change this proportion to 0.5 and click the `estimate` button on the right of the window.
9. set the `Clock Model` by assuming a strict clock (for simplicity). We will not estimate the rate for each DNA region
10. leave the `Priors` and `MCMC` to the default values
11. save the BEAST2 XML file, but leave the BEAUti software open.
12. run BEAST2

Once the analysis is done, check the log file created by BEAST2 in the software Tracer. If the MCMC run looks fine, summarize the sample of plausible trees using the software TreeAnnotator, which is found in the BEAST2 folder.

1. set the burnin based on the Tracer output
2. set the posterior probability limit to 0.5 (we don't want clades that appear in less than 50% of the trees)
3. select the "Maximum clade credibility tree" and "Median heights" for node heights.
4. use the species trees as input.
5. visualize the resulting species tree using FigTree or Seaview.

Do the same for each of the gene trees.

2 Questions

- is the species obtained by *BEAST2 the same as the PhyML trees? Why?
- which gene trees are incongruent with the species tree?