Table of contents

Overview

The objectivity of this exercise is to help you understand how to perform Bayesian phylogenetic inferences using PhyloBayes. You will learn how to set models, check for convergence, and perform model comparison in PhyloBayes. Your will also learn how to run Maximum-likelihood analysis using mixture models in IQ-TREE.

Getting started

In this exercise, we will use PhyloBayes to analyze ten small protein sequence alignments simulated under LG+C60+F model using 4-taxon trees with various levels of long-branch attraction artefacts (Wang et al., 2017). The alignment simuLBA.C60.7.5.phy will be used as example throughout this tutorial, but you can use any of the 10 alignments.

The original alignments have >20,000 sites, which will be too long for this practical session. Therefore, we will first create a smaller alignment of 1,000 sites:

```
goalign sample sites -l 1000 -p -i simuLBA.C60.7.5.phy -o simuLBA.C60.7.5.1k.phy
```

Exercise 1: Tree inference

1.1: Basic analysis

To conduct a basic tree inference in PhyloBayes, run:

```
pb -s -cat -poisson -d simuLBA.C60.7.5.1k.phy simuLBA.C60.7.5.1k.CAT_F81.chain0
```

The “-d” option allows you to specify the data file. The “-s” option instructs the program to save all parameters sampled in the MCMC run in addition to tree topologies, so that we can later perform model comparison analyses. The “-cat” option turns on the CAT mixture model for equilibrium frequency profiles. The “-poisson” option sets the exchange rate matrix to Poisson (all exchangeabilities set to 1). The name of the run (“simuLBA.C60.7.5.1k.CAT_F81.chain0”) is and is specified at the end of the command.

To start another analysis under the same model, just change the name of the run:

```
pb -s -cat -poisson -d simuLBA.C60.7.5.1k.phy simuLBA.C60.7.5.1k.CAT_F81.chain1
```

1.2: Check convergence

Summary information of a MCMC run will be written in to the .trace file (e.g., simuLBA.C60.7.5.1k.CAT_F81.chain0.trace), whereas the trees will be written to the .treelist file (e.g., simuLBA.C60.7.5.1k.CAT_F81.chain0.treelist). We can check if the two runs have converged by using the tracecomp and bpcomp programs provided by PhyloBayes:

```
tracecomp -x 500 1 simuLBA.C60.7.5.1k.CAT_F81.chain0
simuLBA.C60.7.5.1k.CAT_F81.chain1
```

The tracecomp program will compare the two runs for a number of summary statistics and report the discrepancies and effective sizes. The “-x” option specifies the length of burn-in and the sampling frequency. In this case, the first 500 cycles will be discarded as burn-in and after that, every cycle will be sampled. Max. difference < 0.1 and effective size > 300 is considered good. Max. difference < 0.3 and effective size > 50 is considered acceptable.

```
bpcomp -x 500 1 simuLBA.C60.7.5.1k.CAT_F81.chain0
simuLBA.C60.7.5.1k.CAT_F81.chain1
```

The bpcomp program will compare the two runs for tree topologies, **generate a posterior consensus tree**, and report the max. and mean differences in bipartition frequencies. Max. difference < 0.1 is considered good and max. difference < 0.3 is be considered acceptable.

You can also download the trace files to your computer to check convergence using Tracer.

1.3: Alternative models

There is a number of models that you can try:

Equilibrium frequency	Exchange rate matrix	PhyloBayes option
CAT	GTR	-cat -gtr
CAT	LG	-cat -lg
CAT	Poisson	-cat -poisson

homogenous	GTR	-ncat 1 -gtr
homogenous	LG	-ncat 1 -lg
homogenous	Poisson	-ncat -poisson
C60	GTR	-catfix c60 -gtr
C10	GTR	-catfix c10 -gtr
C60	Poisson	-catfix c60
C10	Poisson	-catfix c10

To run a PhyloBayes analysis with one of these models, just replace the “-cat -poisson” in the command we used in 1.a with one of the options listed in the third column in the table above. Please also remember to change the name of the run!

Visualize and compare the trees inferred under different models using phylo.io. Are they the same?

1.4: PhyloBayes-MPI

PhyloBayes analyses can be really slow on empirical phylogenomic data sets, particularly when the CAT+GTR model is being used. To accelerate the analyses, we can use PhyloBayes-MPI to parallelize the computation:

```
mpirun -np 2 pb_mpi -cat -gtr -d simuLBA.C60.7.5.1k.phy
simuLBA.C60.7.5.1k.CAT_GTR.mpi.chain0
```

Here, pb_mpi is used instead of pb, and we asked for 2 threads with “mpirun -np 2”. Since we can only use two threads at most, you probably will not observe much

improvements in the runtime.

1.5: ML analysis using mixture models in IQ-TREE

Recently, several ML alternatives of the PhyloBayes-CAT model have been implemented in IQ-TREE. Here is how we can run the ML analysis using mixture models:

```
Iqtree -nt 2 -m LG+C60+F -bb 1000 -bnni -s simuLBA.C60.7.5.1k.phy -pre simuLBA.C60.7.5.1k.LG+C60+F
```

The LG+C60+F model was the model used to simulate these alignments. You can also use other model configurations, such as LG+C60, C60+F, and C60 (do you know the difference between them?).

You can also do a tree search using the homogenous model:

```
Iqtree -nt 2 -m LG+G+F -bb 1000 -bnni -s simuLBA.C60.7.5.1k.phy -pre simuLBA.C60.7.5.1k.LG+G+F
```

Again, you can compare the ML trees inferred under different models, and also the trees inferred by PhyloBayes in previous steps.

Exercise 2: Model comparison

2.1 Cross-Validation

To run cross-validation, we first have to generate pairs of learning and test data sets:

```
cvrep -nrep 1 -nfold 10 -d simuLBA.C60.7.5.1k.phy cvb
```

Cross-Validation analyses typically consist of 10 replicates, but due to the time limits we

will do just one replicate (-nrep 1). The “-nfold 10” option indicates that 90% and 10% of the original data set will be assigned to the learning and test data set, respectively. The name of the run is “cvb”.

Then we will run phylobayes analyses on the learning data set under different models using a fixed tree topology:

```
pb -T TREE -x 1 1100 -cat -gtr -d cvb0_learn.ali CAT_GTRcvb0_learn.ali

pb -T TREE -x 1 1100 -cat -poisson -d cvb0_learn.ali CAT_F81cvb0_learn.ali

pb -T TREE -x 1 1100 - gtr -d cvb0_learn.ali GTRcvb0_learn.ali
```

Please replace the “TREE” with the name of the tree file you would like to use. The “-x 1 1100” indicate that we would like to run 1100 cycles and save parameters per cycle.

When the runs are done, we can calculate likelihood values for the test data set under different models:

```
readcv -rep 0 -x 100 1 -cat -gtr cvb

readcv -rep 0 -x 100 1 -cat -poisson cvb

readcv -rep 0 -x 100 1 -gtr cvb
```

Here, the first 100 cycles will be discarded as burn-in (“-x 100 1”).

Lastly, we can compare between models:

```
sumcv -nrep 1 GTR CAT_F81 CAT_GTR cvb
```

2.2: Posterior predictive analysis

The PPA can be done using the ppred program provided by PhyloBayes:

```
ppred -x 500 1 1500 -div -comp -sat simuLBA.C60.7.5.1k.CAT_GTR.chain0
```

The options “-div”, “-comp”, and “-sat” tell the program to computes statistics for the following features: “biochemical specificity”, “compositional homogeneity”, and “saturation”, respectively.