Assignment 11_2024

Your name:
Your email address:

Answer all the questions in red in the provided boxes.

Assignments:

Exercise 1:

We will set-up and analyze a MrBayes analysis of the intein and extein sequences from the Yeast vma1 sequence alignment (see lab 9). The files were saved in Nexus format "as selection" from seaview, and the following MrBayes block was added to each:

```
begin mrbayes;
lset nst=6 rates=gamma;
mcmcp filename=analysis_Extein;
mcmcp samplefreq=50 printfreq=50 diagnfreq=500;
mcmcp ngen=20000;
mcmcp savebrlens=yes;
end;
```

nst6 denotes the GTR model. I only chose the gamma distribution (not "gamma+invariant sites") because the two parameters have very similar effects in describing ASRV. (More sites that do not change either increase "I" or lower alpha.)
Note that these files are different form the analysis described in the lecture 18, where the intein and extein sequences were saved as partitions in a single file. All files that we will use today are in this zip file. Download this file onto your computer. (You will need to extract the file before you can transfer them to Xanadu using filezilla. Alternative below.)

Start filezilla and connect to **transfer.cam.uchc.edu**. (use you username and password)
Create a lab10 directory and transfer the files Yeast_vma1_extein_aligned0gen.nxs and Yeast_vma1_intein_aligned0gen.nxs into that directory.

If you want to explore submitting jobs to the queue, also transfer the shell scripts (ending in .sh) and the nexus files that include mcmc, autoclose, sump and sumt commands.
Connect to xanadu in terminal, (see here for details).
ssh mcb3421usrXX@xanadu-submit-ext.cam.uchc.edu

Go to a compute node:

```
srun --pty -p mcbstudent --qos=mcbstudent --mem=2G bash
curl -O https://j.p.gogarten.uconn.edu/mcb3421_2024/labs/Lab11.zip
unzip Lab11.zip
```

This creates a directory called Lab11 and puts all the files into it.

```
cd Lab11
module load MrBayes/3.2.7
```

Check that the nexus files are in the directory (using **more** to is a good choice).

```
more *.nxs
```

Read through the file, with special attention to the MrBayes block.
Which commands in the MrBayes blocks are different between the 0gen and the 10000000gen files? What might these different commands accomplish?

Start MrBayes by typing

mb

then type

```
execute Yeast_vma1_intein_aligned0gen.nxs
showmodel
```

Read through the output after each command.

Type

```
mcmc
```

to start the chains.
Every 50 generations the program prints a line to the screen and every 500 lines it prints the

```
Average standard deviation of split frequencies
```

This value *should be* below 0.01 to indicate the two parallel runs obtain similar results. If the value is above, run the chain for another couple 10000 generations.

Note: As this is a "get to know" exercise, we don't need to take the .01 too seriously. When you are done (do not do more than 40000 generations - you can download the parameter and treefiles from a longer runs below)
type "**no**" at the Continue with analysis? (yes/no): prompt.

After the run is finished, the " **sump** " command will plot the logL vs. generation number, that allows to determine the necessary burnin (you want to discard those samples as "burnin" where the -logL is still rising steadily).

To see the whole logL curve, you need to set the burnin fraction to .02 . (type help sump at the mb command line). **`sump burninfrac=.02`**

At the start of the run, the likelihood rapidly increases by orders of magnitude. If the first samples are included in the plot, one really cannot see if the likelihood values fluctuate around a constant
value. You can exclude the first couple of samples by specifying a `burninfrac`. The new version of MrBayes uses a burnin of 25% by default.

Repeat the sump command using different values for `burninfrac`. Select the value that no longer reveals an upswing at the beginning.

Summarize the trees sampled after the burnin using the sumt command:
`sumt burninfrac =.25`
, where you need to substitute '.25' with the number you obtained in the previous step of the exercise.
This command creates a consensus tree, and shows posterior probabilities of the clades. You can take a look at the tree **on the screen** (scroll up to see the bipartition table, and different version of the tree).

1) Which value did you use for the burnin/burninfraction?

2) Which branch in the tree is the longest? (check the branch length table, then look up the branch in the bipartition table)

3) How long is it? (use the mean value)

4) What is the measure? (Check the label at the scale bar for the last tree image)

5) How big is the shape parameters for the intein? What is the 95% credibility interval?

6) What is the probability for a nucleotide to be an A? What is the probability to be a C?

7) Can you explain in a few words, why is it important to exclude a 'burnin' from our analyses?

Exercise 2:

Check if tracer is already installed on your computer. If not, install the latest version from the github repository on your computer.
Start Tracer on your computer. Drag the four .p files analysis_Intein2 and analysis_Extein2

into the Trace file window (upper left). (Use the ones in this zip file, not the ones form the analyses in Exercise 1, these overwrote the .p and .t files from the analyses with 1,000,000 generations)
These files were generated by submitting the script "shell_for_YeastIntein_student.sh" and "shell_for_YeastExtein_student" to the queue using the sbatch command: **sbatch shell_for_YeastIntein_student.sh**
To check, if the process is running use **squeue -u mcb3421usr40**
A running script would result in an output like this:
`3566182 mcbstuden Intein mcb3421u R 14:22 1 xanadu-68`
The R stands for running. If the queue is busy you might be stuck in PD instead of R.
You also can use filezilla and check the Intein.out or Extein.out files.
The script and the nexus files with the corresponding MrBayes blocks are in the zip archive. Read through them!

Select the 2 runs for the intein (in the trace file window upper left), select trace on the upper right above the plot window, select LnL (log likelihood), uncheck Draw line plot (below the plot) to make the image less busy.
Repeat for the two extein runs.
Note that 10% burnin is already removed, or grayed out (checkmark below the plot window).
Select alpha (the shape parameter), select marginal density on top, and histogram in the display options.
What are the shape parameters estimates (check μ estimates on top).
Write down mean and HPD interval below. Repeat for the two extein run:
alpha (and HPR) for the intein:
alpha (and HPR) for the extein:

Switch to pi(A) and pi(T).
Is the frequency of As and Ts different for inteins and exteins? (if you select all 4 runs on the left, turn on the legend so you know which is which)

The tree length for each of the sampled trees (in substitutions per site) is sampled under TL.
How long are the two trees on average?
mean intein: mean extein:

How many more substitutions per site occurred overall in the intein and compared to the extein?

Is the different significant? Use the High Probability Density intervals to decide on the significance.

Which other parameter estimates are very different for intein and exteins?

Finished? Don’t forget to email your worksheet.