 Name:………………………………………………………..

**Molecular phylogeny and evolution**

**Course 17 – 23 January 2019**

**Course Manual**

*Juan I. MONTOYA BURGOS*

Laboratory of Molecular Phylogeny and Evolution in Vertebrates

Department of Genetics & Evolution

University of Geneva

**General informations**

**IMPORTANT**:

* **For Mac or Linux users**: Please start by creating a new folder named "Phylo\_course" in your personal computer, either at the root (C:) or on your desktop.
* **For Windows users**: Please start by creating a new folder named "Phylo\_course" in the desktop of your personal computer, for example C:/Users/*YourUserName*/Desktop/Phylo\_course
* Please also check if the path contains special characters or blanks... and delete them as this can create problems.
* Keep the **names of taxa the same** throughout the whole workshop – it should be **maximum 10-character** names for all names.

**Move the softwares that you installed before the course, in the newly created "Phylo\_course" folder**

**For Windows users:**

* In order to connect and exchange files with a distant server you need to have an SSH and telnet client. If you don't have one, you will need to install one. The software PuTTY is a good and easy SSH and telnet client. In addition, you will need to install PSCP (an SCP client, i.e. command-line secure file copy) which works with PuTTY.
* The two softwares:  
  putty.exe  
  pscp.exe  
  can be downloaded here: http://www.putty.org/
* **save them at the root (C:)**

The exercises are designed to be compatible with Mac OsX, PC Windows or Linux systems. The datasets have been uploaded into the server BAOBAB, on the following folder:

/home/montoya/MolPhylEvol/Students/

To download the course manual into your computer, use the following command, but ***note that the period (.) at the end of these command is very important !***

-From YOUR Linux/Mac:

*scp YourUserName@baobab.unige.ch:/home/montoya/MolPhylEvol/Students/Exercises.docx .*

-From YOUR PC (Windows) with pscp.exe:

*pscp.exe -scp YourUserName@baobab.unige.ch:/home/montoya/MolPhylEvol/Students/Exercises.docx* ***.***

Note to Windows users: If you do this command and you cannot see the file in "C:", substitute the dot (.) at the end of the previous command by the address of your Desktop.

**Contents**

**Thursday**

1. **Blast: sequence identification 4**
2. **Sequence alignment with Seaview 5**
   1. Manual sequence alignment 5
   2. Automated sequence alignment 6
3. **Sequence analysis with MEGA 7**
4. **Selection of the model of sequence evolution 8**
   1. Selecting the best model for nucleotide sequences, with MEGA 8
   2. Selecting the best model for amino acids sequences, with MEGA 9

**Friday**

1. **Phylogenetic inference methods part I: Distance methods 9**
   1. Distance method: Neighbor Joining (NJ) 10
2. **Displaying and working with trees 10**

6.1 Visualization of phylogenetic trees with FIGTREE 10

1. **Phylogenetic inference methods part II: Maximum Likelihood 10**
   1. ML inference using DNA, with RAxML 10
   2. ML inference using amino acid sequences, with RAxML 12

**Monday**

1. **Phylogenetic inference methods part III: Bayesian Inference 12**
   1. Bayesian inference using DNA, with MrBayes 12
      1. DNA alignments 12
      2. *Optional: additional practical aspects of MrBayes 15*
2. **Multigene analyses with partitions 16**
   1. DNA data: MrBayes analysis with partitions 17
   2. Amino acid data: RAxML analysis with partitions 18
3. **Statistical tests based on phylogenetic trees 20**

10.1 Bootstrap: statistical support of the nodes 20

10.2 ML Clock test, introduction to Topological Tests 21

**Tuesday**

**11. Dating phylogenetic trees with BEAST 22**

**12. Topological test based on Multi-gene trees 24**

**Wednesday**

**13. Analysis of the BEAST results 26**

**14. Analyses and interpretations of results based on multi-gene data 27**

**15. Final synthesis, discussion and conclusion 27**

**16. Optional: Analysis of selection pressure using Paml, at BAOBAB 27**

16.1 Omega calculation for each branch of the tree 27

16.2Search for codons under positive selection: Sites model 30

17. ANNEX 1: Model definitions for MrBayes 32

**Thursday**

**1. Blast: sequence identification**

You will find the sequence to identify as a ".txt" file in the BAOBAB server at the place:

/home/montoya/MolPhylEvol/Students/Sequences/

First choose one of the three available sequences and write down its number here: .............

(filenames: Sequence1.txt, Sequence2.txt or Sequence3.txt).

Then, you will have to copy the sequence file into your computer, within the "Phylo\_course" folder. To do this, open a Terminal command session and reach the location in which the folder "Phylo\_course" is.

Then type (or copy - paste) the following commands: be careful and **include the spaces**, the **final dot** and **replace “yourfilename”** by the actual name of your file:

* Linux+Mac:

scp username@baobab.unige.ch:/home/montoya/MolPhylEvol/Students/Sequences/*yourfilename* Phylo\_course/.

* Windows:

C:\pscp.exe -scp

username@baobab.unige.ch:/home/montoya/MolPhylEvol/Students/Sequences/*yourfilename* Phylo\_course\.

***NOTE: The period (.) at the end of these commands is very important****, as it tells the command to paste the files in the same folder the terminal window is in. You can also change it by the path of anther folder.*

Make sure that your sequence has an open reading frame: Check out the six possible reading frames with the program “Translate” at the EXPASY website: <http://web.expasy.org/translate/>. Select the “Compact” output format, and once you have found the correct frame, copy the correct amino acid sequence by selecting it with your mouse and proceed to the identification of your sequence by Blast search.

Identify your proteic sequence using Blast on a protein database:

* Use NCBI’s Blast: <http://www.ncbi.nlm.nih.gov/BLAST/>
* In "Basic BLAST", select "protein blast"
* Paste your sequence
* Select the database "non-redundant protein sequences (nr)"
* Blast it!

Questions**:**

**✔** Which gene is this?

**✔** What is its function?

**✔** From which organism does this sequence comes from?

Find out the CDS (nucleotidic coding sequence) of this protein and save it in FASTA format. To do so:

1) Click on the accession number of your best Blast hit and browse the details until you find the nucleotidic coding sequence link and click on [CDS].

2) A new bar will appear at the bottom of the window, under [Display], choose [Fasta]

3) At the top right of the new window, in the [Send] menu, choose [Coding Sequence] and Format [FASTA Nucleotide] and then save the sequence in your directory, with the name of your choice:

Write down the name of your sequence here:...................................................................

**2. Sequence alignment with Seaview**

You have been asked to install Seaview in your computer! If you haven't done it, install it now!!! (http://pbil.univ-lyon1.fr/software/seaview.html ;or run sudo apt-get install seaview if on Linux).

**2.1 Manual alignment**

Go and get the alignment already available for your sequence (corresponding number) that you will find in the server Baobab. The available files are: Sequence1alignCFTR.fas, Sequence2alignCTTNBP2.fas, Sequence3alignMET.fas.

Copy the alignment into your local folder.

Linux+Mac:

scp [username@baobab.unige.ch](mailto:username@baobab.unige.ch):/home/montoya/MolPhylEvol/Students/Alignments/YourAlignment.fas .

Windows:

C:\pscp.exe -scp username@baobab.unige.ch:/home/montoya/MolPhylEvol/Students/Alignments/YourAlignment.fas .

***NOTE: The period (.) at the end of these commands is very important****, as it tells the command to paste the files in the same folder the terminal window is in. You can also change it by the path of anther folder.*

Then, align your sequence against the other sequences of the alignment. For this, use Seaview and open the alignment in Fasta format (for opening your Fasta file use the option "show: all files"). Then open your sequence in Fasta format too.

Click on your sequence, copy it ("edit" menu, "copy selected seqs"), then move to your alignment window and paste it ("edit" menu, "paste alignment data"). Your new sequence will be located at the bottom of your alignment.

IMPORTANT: rename your sequence according to the same model as the other sequences in the alignment (with 10 characters maximum). To do so, select only your sequence (unselect all others), go to the "edit" menu and select "Rename sequence". **Save this alignment!**

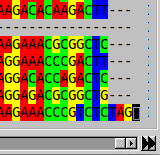
Align your sequence manually using the space bar to generate gaps if needed. To delete a gap, use the deletion back arrow key. WARNING: don’t delete parts of the sequences. You don’t need to modify the existing alignment. Just align your sequence in the alignment. Make sure you keep the reading frame!

**2.2 Automated sequence alignment**

You may want to try an automatic alignment program. Two such programs are available in Seaview: ClustalO and Muscle (you can select them in the "Align" menu and "alignment options"). A good option for this exercise is to perform an automatic alignment that aligns your new sequence against an existing alignment without modifying it. This is called a "profile aligment". To do so with Seaview, you must first create a species group including all the already-aligned sequences. For this, select all the species except your new species. Then go to the "Species" menu, click on "Create group" and give it a name. Then select your new species and, in the "Align" menu, choose "Profile alignment" and choose as profile the name of the species group you just created. Choose the Muscle algorithm and start the automatic alignment. A new window will appear displaying the parameters of the program. When finished, click on "OK" and the new alignment will be displayed on the main Seaview window.

**IMPORTANT: check the new alignment from start to end!!!**

*Homology of sites is set during the alignment process. Crap in, crap out!*

You will probably notice that there is an extra codon at the end of **your** sequence. This is the STOP codon that has been deleted from the other sequences. You need to delete it as well from your new sequence. For this, select all sequences and go to the "Props" menu, and check the "Allow sequence editing" option. Now, go to the end of your sequence, and click on the site *after* the last letter of your sequence, so that it looks like this:

...and then delete the last three bases (NOTE: this must also delete the last three places on the other sequences).

Save this alignment!

***Good to know***: there are several automated sequence alignment software, some of which have an online interface:

Mafft: http://mafft.cbrc.jp/alignment/server/

Muscle: http://www.ebi.ac.uk/Tools/msa/muscle/

Clustal Omega: http://www.ebi.ac.uk/Tools/msa/clustalo/

Question: **✔**Do you notice something special on the final alignment like a sequence shorter than others?

Go to the menu "File", choose "Save as" and save your alignment in the following formats:

- FASTA format - Phylip format - Nexus format

***Warning***: some programs generate an ambiguous Nexus format. Check if your Nexus file has been properly generated. To do this, open your Nexus formated file with a Text editor and check if the data type is correct (DNA or protein) and check if the alignment is organized in an interleave or non-interleave manner (you may need to correct/add the command "interleave=no" or "interleave=yes").

Write down your alignment names: Nucleotide Fasta: ................................................  
 Nucleotide Phylip:................................................  
 Nucleotide Nexus:................................................

**3. Sequence analysis with MEGA.**

*Explore and understand your dataset !*

Start MEGA.

Open your nucleotide alignment file in Fasta format. Choose the "Analyze" mode. Remember that your gene is a protein coding gene and the Open Reading Frame starts at position 1.

Nucleotide frequency

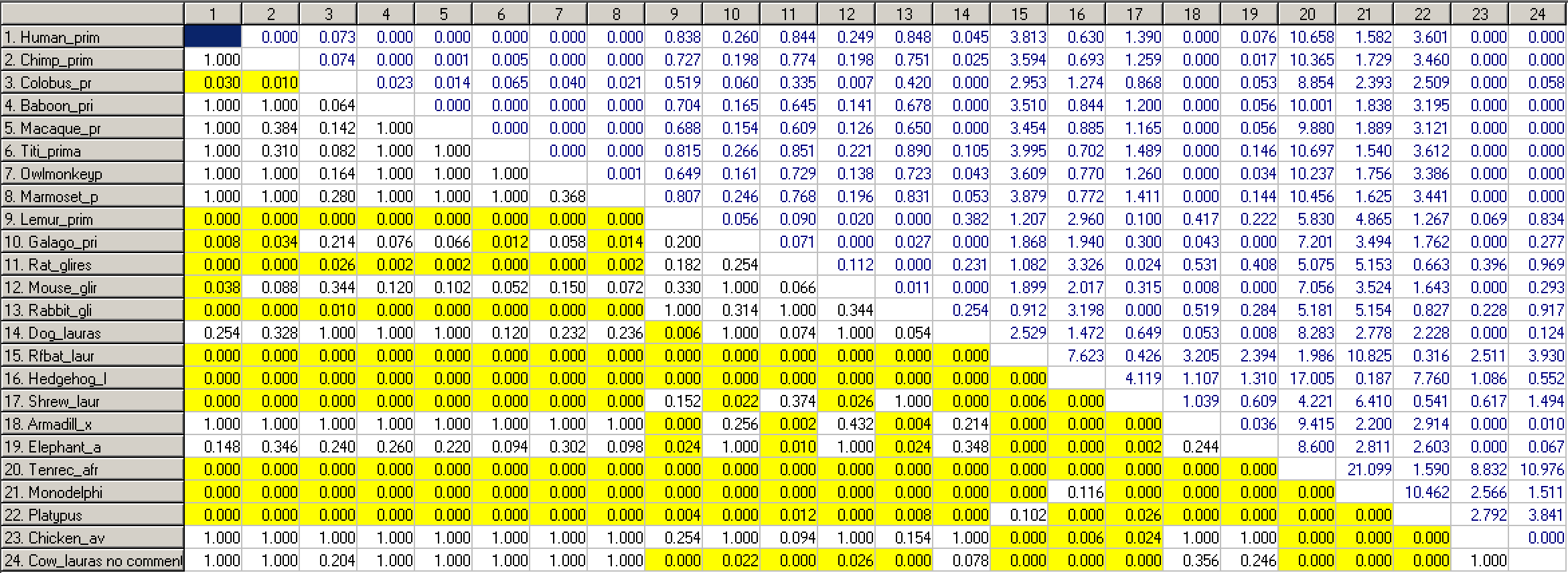
* Are the nucleotides A, T, G and C equally frequent?

To answer to this question go to the "Models" menu and choose "Compute Nucleotide Composition". The results will be saved as an .xls file that you can open with Excel. Save this file for your report.

* Is the bias in nucleotide frequency the same for all the species of the data set?

To answer to this question compute the Disparity Index (ID). The Disparity Index measures the departure from base compositional stationarity for a pair of sequences. It corresponds to the observed compositional difference between two sequences compared to the compositional difference that would be expected under homogeneity. The ID equals 0 when the homogeneity assumption is satisfied. The significance of a given ID value is then assessed using a Monte Carlo approach (Kumar & Gadagkar, 2001).

To calculate the ID, in the menu bar, choose the menu “Models”, and the option “Disparity Index Test of Pattern Heterogeneity”. Use the default parameters values and compute the analysis. The result looks like this:



Disparity Index (DI)

 **P-value**

Save these results to a file. The null hypothesis is homogeneity between two sequences. The P-value is the probability of being wrong when rejecting the null hypothesis.

Estimation of the Transition/Transversion ratio

* What is the ratio of Transitions (Ts) over Transversions (Tv): Ts/Tv?

In the menu bar, choose the menu “Models”, and:

in MEGA 5: the option “Estimate Transition/Transversion Bias (ML)”

in MEGA 6: the oprtion “Compute MCL Transition/Transversion bias”

in MEGA 7: both options are available and possible.

Use the default parameter values and compute the analysis. An error message may appear but it’s a bug so click on “continue application” and it will continue. Save the Results.

Variable evolutionary rates among sites

* Are all the sites of the alignment evolving at the same rate?

In the menu bar, choose the menu “Rates” and the option “Estimate Position-by-Position Rates (ML)”. Use the default parameter values and compute the analysis. Save the results (as an Excel file). Use this Excel output to check if all sites evolve at the same rate (= constant or homogeneous evolutionary rate across sites). Prepare a graph showing the estimated real rates for the 60 first alignment positions.

Put this graph in your report, with comments.

**4. Selection of the model of sequence evolution that fits the best to your data.**

**4.1 Selecting the best model for nucleotide (DNA) sequences, with MEGA**

Start MEGA. Open your nucleotide alignment file in Fasta format. In the menu bar, choose the menu “Models” and the option “Find Best DNA/Protein Models (ML)”. Use the default parameters and start the computation (about 1-10 min). Save the complete table of results (look for the option to save it as an Excel file).

Questions:

* Which is the best model according to the Akaike information criterion (AIC) or the Bayesian Information Criterion (BIC)?

AIC: ……………………………………. BIC: …………………………..………..

The AIC is a measure of the relative quality of a statistical model, for a given set of data. As such, AIC provides a means for model selection. When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting. Both BIC and AIC resolve this problem by introducing a penalty term for the number of parameters in the model; the penalty term is larger in BIC than in AIC.

The AIC formula is: mathit{AIC} = 2k - 2\ln(L)

Where k is the number of free parameters in the statistical model, and *L* is the maximized value of the likelihood function for the estimated model.

The BIC formula is: -2 \cdot \ln{p(x|M)}} \approx \mathrm{BIC} = {-2 \cdot \ln{\hat L} + k \ln(n) }. \ 

Where hat L= the maximized value of the likelihood function of the model http://upload.wikimedia.org/math/6/9/6/69691c7bdcc3ce6d5d8a1361f22d04ac.png; and (x|M)= the marginal likelihood of the observed data given the model http://upload.wikimedia.org/math/6/9/6/69691c7bdcc3ce6d5d8a1361f22d04ac.png;

* What are the parameter values of the best model according to BIC?

……………………………………………………………………………………………

**4.2 Selecting the best model for amino acid sequences, with MEGA**

You need first to translate your nucleotide alignment into an amino acid alignment!

Translating your nucleotide alignment.

Re-start or start Seaview. Open your nucleotide alignment file in Fasta format. Go to “Props”, and click on “View as Proteins”. Save this amino acid alignment in Fasta format and Nexus format. For this, go to the menu “File” then “Save prot aligmnt”, and choose the desired formats.

Name of the amino acid alignment in Fasta format: …………………………………………...

Name of the amino acid alignment in Nexus format: ……………………………….................

Check if the Nexus format has been properly generated. Open your Nexus format file with a Text editor and check if the data type is correct (DNA or protein) and check if the alignment is organized in an interleave or non-interleave manner (you may need to correct/add the command "interleave=no" or "interleave=yes").

Finding the best model of evolution for the amino acid alignment

Start MEGA. Open your amino acid alignment file in Fasta format, remember to choose “Protein Sequences” in the Input Data window. In the menu-bar, choose the menu “Models” and the option “Find Best DNA/Protein Models (ML)”. Use the default parameters and start the computation (about 10-20 min). Press on the icon that says “XL” to export your data into EXCEL format (.xls).

Questions:

* Which is the best model?: …………………………………………………………………..
* What are the parameter values? : ………………………………………………………...

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**Friday**

**5. Phylogenetic inference methods part I: distance methods**

**5.1 Distance method: Neighbor Joining (NJ)**

Work with your nucleotide alignment and then with your amino acid alignment.

Start MEGA. In the menu bar, choose the menu “Phylogeny”. Then select the option “Construct/Test Neighbor-Joining Tree”. In the Options Summary window, choose the best options and parameters for your dataset (substitution model, rates among sites, …). Compute the analysis.

In the TreeExplorer window, you can re-root your tree (=choose a new outgroup) by selecting the “Place Root on Branch” button on the left bar. Use *Chicken* as outgroup.

Save your tree to a file (in the “File” menu choose “Export current tree (Newick)”) and write here the name of this file: ……………………………..............

Do the same but with your amino acid alignment. Save your amino acid based tree to a file and write here the name of this file: ……………………………………………

IMPORTANT: Make sure that in these and every other tree and alignments you produce, the taxon names in the tree correspond perfectly to the taxon names in the alignment (MEGA can change the names on the trees).

Questions:

* Describe your nucleotide NJ tree by checking the lineages that have been correctly inferred and the ones that appear as polyphyletic: ……………………………………..
* Describe your amino acid NJ tree; any difference with your nucl. NJ tree? ………….

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**6. Displaying and working with trees**

**6.1 Visualization of phylogenetic trees with FIGTREE**

Several programs are available to display trees, to change outgroup, to swap branches, and to improve the graphical output. We will use Figtree because it is available for all OS (Windows, Mac OS X and Linux) and because it accepts the two main tree formats: Newick and Nexus.

Objective of the exercise: Open with Figtree the tree file(s) obtained with the NJ method and prepare an appropriate figure of your phylogenie(s). You can play with the various options.

Use *Chicken* as outgroup.

Save your nice and artistic tree(s) for your report.

**7. Phylogenetic inference methods part II: maximum likelihood**

**7.1 ML inference using DNA, with RAxML in BAOBAB**

* Copy your nucleotide alignment in **PHYLIP** format into your folder in **BAOBAB**
  + Linux+Mac:

scp YourAlignment.phy username@baobab.unige.ch:~

* + Windows:

C:\pscp.exe -scp YourAlignment.phy username@baobab.unige.ch:~

* Log into your **BAOBAB** account
  + Linux+Mac:

ssh username@baobab.unige.ch

* + Windows: using Putty

C:\putty.exe -ssh username@baobab.unige.ch

* Create a folder for all your Maximum Likelihood analyses called “ML\_analyses”

mkdir ML\_analyses

* Move your alignment into the newly created folder

mv YourAlignment.phy ML\_analyses

* Copy the example RAxML batch file into this file

cp /home/montoya/MolPhylEvol/Students/Batch\_Files/RAXML\_batchfile.sh ML\_analyses

* Navigate into the newly created folder

cd ML\_analyses

* Open the RAXML\_batchfile.sh file using the command:

nano RAXML\_batchfile.sh

and check that the batch file has all the correct filenames in the RAxML command: choose the Alignment file name, the best model to use and don't forget to change the output name: the "-n" option at the end of the RAxML command). To make a change, simply use the arrows of your keyboard and move to the desired place.

**MODELS for RAXML**

For DNA data use:    GTRGAMMA  or GTRGAMMAI                   
For AA data use:     PROTCATmatrixName[F]  or PROTGAMMAmatrixName[F]  or  
                     PROTCATImatrixName[F] or PROTGAMMAImatrixName[F]     
The AA substitution matrix can be one of the following:   
 DAYHOFF, DCMUT, JTT, MTREV, WAG, RTREV, CPREV, VT, BLOSUM62, MTMAM, LG, MTART, MTZOA, PMB, HIVB, HIVW, JTTDCMUT, FLU, LG4M, LG4X, DUMMY, DUMMY2, PROT\_FILE, GTR\_UNLINKED, GTR  
  
With the optional "F" appendix you can specify if you want to use empirical base frequencies

The commands should look like the original batch file, but with your own file names and output names (check the original if you suspect something important was deleted).

***Important*:** Best model should be written in UPPERCASE

***Important*:** write down the name of your best DNA RAxML output analysis

RAxML DNA analysis output name:……………......................................................

Check also if the name of the reservation is completed and correct:

#SBATCH --reservation=.... (complete the reservation name)

Once you finished editing this file, press Ctrl+X, then "Y" to save changes.

* Execute RAxML by running the batch file

sbatch RAXML\_batchfile.sh

Now you can visualize the status of your job by typing squeue. Under the status column (“ST”) you should be able to see if your job is pending (“PD”) or running (“R”), and you will see your job ID number. Once your job is finished, it will not be visible on this list anymore.

If you need to cancel a job, type: scancel *JOB\_ID\_NUMBER*

A summary of your search job, together with the information on the best independent analysis is given in the file ending by “…\_info” (inside your ML\_analyses folder). Please write down the information or download and save your “…\_info” file (you will need this information to answer the questions below).

Display the best tree by copying the content of the file “RAxML\_bestTree…” corresponding to the best analysis and by pasting it in Figtree.

Questions:

* What is the likelihood score of your ML tree? ….............................………………….....
* What are the estimated values of the parameters of your model? ……….................…

…………………………………………………………………………………………………………………………………………..

Save the results for your report.

**7.2 ML inference using amino acid sequences, with RAxML**

Do the same exercise as above but with your amino acid dataset. You will have to change several options in the RAXML\_batchfile.sh file, so make a copy named RAXML\_batchfile-aa.sh:

cp RAXML\_batchfile.sh RAXML\_batchfile-aa.sh

Change the “best model” option (-m), the input filename (-s), and the output name (-n) in the RAxML command accordingly.

nano RAXML\_batchfile-aa.sh

***Important:*** write down the name of your best amino acid RAxML analysis.

RAxML amino acid analysis output name:………………………………………………………………………………………………

Questions:

* What is the likelihood score of your ML tree? ……………………………………………
* Describe the differences between your RAxML DNA tree and your RAxML amino acid tree: ……………………………………………………………………………………………..

Save the results for your report.

\_\_\_\_\_\_\_\_\_\_\_\_\_

**Monday**

**8. Phylogenetic inference part III: Bayesian Inference (BI)**

**8.1 Bayesian Inference using DNA, with MrBayes (on BAOBAB)**

**8.1.1 DNA alignment**

Nucleotide data

MrBayes works only with NEXUS format. The commands and options of the program have to be included as a command block at the end of the alignment file.

Modify your nucleotide alignment file in Nexus format to add the block of information that the program will use concerning the best model of sequence evolution.

Check that your “begin data” info has all the following commands:

begin data;

dimensions ntax=#oftaxa nchar=#ofcharacters;

format gap=- datatype=DNA interleave=yes/no missing=? matchchar=. ;

The two last elements (missing=? matchchar=.)are optional.

The commands for MrBayes should look like this:

....[the last sequence]GGATGATGAT;

BEGIN MRBAYES;

set autoclose=yes nowarn=yes;

Charset alignment\_name = 1- ;

Partition Dummy = 1:alignment\_name;

Set partition = Dummy;

Lset applyto=(1) nst=6 rates=invgamma;

Prset applyto=(1) revmatpr=Dirichlet(1.0,1.0,1.0,1.0,1.0,1.0) statefreqpr=Dirichlet(1.0,1.0,1.0,1.0) shapepr=Uniform(0.1,50.0) pinvarpr=Uniform(0.0,1.0);

mcmcp ngen=1000000 nchains=2 samplefreq=100 printfreq=10;

propset ExtTBR$prob=0;

mcmc;

sump;

sumt;

end;

- Open your Nexus alignment file with a text editor. Copy carefully this block at the end of your Nexus alignment file, just before the last line which contains “end ;”. You will have to modify the number of sites of your alignment in the line “Charset alignment\_name = 1 - ………..” . The number of sites of your alignment is given at the beginning of the Nexus file. If necessary, you will have to modify the parameters of the model of sequence evolution.

The commands “nst=6 rates=invgamma” correspond to the model GTR + G + I. So if your best model is that one, you don’t need to modify the command bloc. If your best model is:

1. GTR + G: you have to change “rates=invgamma” to “rates=gamma”
2. GTR: you have to delete “rates=invgamma”
3. HKY + G: you have to set “nst=2” and “rates=gamma”
4. Another model: see ANNEX 1 (at the end of this manual)

- Verify that the gaps are encoded as “-“ and not as “?” in your alignment (if they are encoded as “?”, then change in the “begin data” block the “missing=-“ command by “missing=?”).

- You can define your value for burnin by using: sump burnin=100; sumt burnin=100;

- Save your edited file.

Indications for using MrBayes on BAOBAB

* Before performing your search in BAOBAB, test if your command block is correct by starting a search in your local MrBayes program. To do so, start your local MrBayes and execute your file by typing the command “execute” followed by the file name (a simple way to get file names into the command line without having to type a long path is by using drag and drop, but in some OS it doesn't work). If your search starts without an error message, you can stop the search and go to BAOBAB. Else, find the error in your command block.
* Copy your \*.nex alignment file (YourNEXUSFile.nex) with the command block into your folder in **BAOBAB**
  + Linux+Mac:

scp YourNEXUSFile.nex username@baobab.unige.ch:~

* + Windows:

C:\pscp.exe -scp YourNEXUSFile.nex username@baobab.unige.ch:~

* Create a new folder in **BAOBAB** for bayesian analyses called “Bayesian\_analyses”

mkdir Bayesian\_analyses

* Move your \*.nex file into the newly created folder, then navigate into the folder

mv YourNEXUSFile.nex Bayesian\_analyses

cd Bayesian\_analyses

* Create a new folder in “Bayesian\_analyses” for the DNA analysis of your gene called “BI\_DNA”

mkdir BI\_DNA

* Move your \*.nex file into the newly created folder, then navigate into it

mv YourNEXUSFile.nex BI\_DNA

cd BI\_DNA

* Copy the batch file for MrBayes into this folder:

cp /home/montoya/MolPhylEvol/Students/Batch\_Files/MB\_batchfile.sh .

(the period “.” at the end of this command is very important)

* Edit the batch file for MrBayes, using the command: nano MB\_batchfile.sh

Change the input file (after “srun mb” put the name of your nexus file). Also, be sure that at the beginning of the batch file you have lines that look like this:

#SBATCH --time=15:00:00 (15h is a good time estimation)

#SBATCH --reservation=.... (complete the reservation name)

Again, once you've made all the necessary changes, press Ctrl+X, then “Y” to save the changes.

* Submit job to cluster

sbatch MB\_batchfile.sh

Follow your job's progress with the squeue command. Once finished, the resulting MrBayes outputs will be located in your folder named “BI\_DNA” (in BAOBAB).

Post-processing MrBayes outputs in your personal computer

You need to copy all the output files into your personal computer. To this aim, you should open a Terminal command session in your computer and go to the location of "Phylo\_course" (cd Phylo\_course).

Create a new folder named "Bayesian\_analyses" (mkdir Bayesian\_analyses).

Then navigate into that folder (cd Bayesian\_analyses).

Then copy all the outputs of MrBayes **from BAOBAB** into a new folder in your personal computer, in the local folder "Bayesian\_analyses":

Linux+Mac

scp -r username@baobab.unige.ch:~/Bayesian\_analyses/BI\_DNA *YourNewFolderName(forExample:myBI\_DNA)*

Windows

C:\pscp.exe -scp -r username@baobab.unige.ch:~/Bayesian\_analyses/BI\_DNA *YourNewFolderName(forExample:myBI\_DNA)*

The resulting MrBayes consensus tree is now in the file that has been created with the extension “.con”. To display the MrBayes consensus tree:

- start the program Figtree and open the file “.con”.

- If needed, reroot your tree (click on the branch of your outgroup and select the “Reroot” icone).

- To display the posterior probabilities, select the option “Node Labels” and open the menu. In this menu you will choose to display your posterior probabilities or "prob".

- you can customize your tree by adding colors, scale bars, etc…

Save your MrBayes tree for your report.

Questions:

✔ Do the “average standard deviation of split frequencies” value is below 0.01?  
This value is an estimate of the convergence between the different runs. You will find this value in the file with the extension “.mcmc”.

✔ What are the estimated means of the values of the model's parameters?  
This information is available in the file ending with “.pstat “ or can be found by using the command " sump burnin=*number\_of\_burn-in\_samples* relburnin=no*"* in the MrBayes window.

✔ Which groups are not well supported by posterior probabilities?

✔ Describe the differences between this tree and the trees previously obtained with

other methods of phylogenetic inference used in this course.

\_ \_ \_ \_ \_ \_ \_

***8.1.2 Optional: Additional information about practical aspects of MrBayes***

***1. Indications for performing an analysis on MrBayes locally.******This is just for your information, no need to run it !***

*Open MrBayes (Dos window) and execute your file [execute file name].*

*Launch a standard analysis with the [mcmc ngen=200000 nchains=3 samplefreq=100] command. This corresponds to 1000 prints of the posterior probabilities distribution.*

*Once the 100’000 generations are done, the program gives the “standard deviation of split frequencies” value. If this value is under 0.01, you can stop the analysis (continue the analysis? NO), if not continue the analysis (continue the analysis? YES) for another 10’000 generations.*

*To visualize the evolution of your analysis and to better define the number of samples to eliminate as a burn-in, open the output file* ***.p*** *with Excel. Do a graph using the first two columns (number of samples and lnL). For a summary (concensus) of all trees sampled, type the [sumt] command followed by the number of samplings to eliminate as a burn-in [burnin = number] (use the inferred number from the Excel graph or then, by default, we eliminate the first quarter (25%) of the samples).*

***2. How to run a MrBayes analysis with partitions by codon position. This is just for your information, no need to do it !***

*To refine your inference analysis, you can consider that different parts (partitions) of your gene may have evolved differently. To take this consideration into account, you can use different models or different parameter values of the same model for the different parts of your gene (or multi-gene dataset). Here we will consider as different partitions the 1rst, 2nd and 3rd codon positions as they are known to evolve under different constrains.*

*You need to declare the different partitions in the "MrBayes" command block. You also need to declare the different models to use for your different partitions or, alternatively, you need to declare a single model but with different parameter values for each partition (the current exercise corresponds to this last option). For this, you will need to modify the file containing the nucleotide alignment in Nexus format that you just used for the non-partitioned search mode. So, make a copy of the alignment used in the previous exercise and rename it, save it into a new folder for the codon analysis, and edit it with the nano command. Follow this example:*

*cp YourNEXUSfile.nex YourNEXUSfile\_codons.nex*

*mkdir BI\_DNA\_123*

*mv YourNEXUSfile\_codons.nex BI\_DNA\_123*

*cd BI\_DNA\_123*

*nano YourNEXUSfile\_codons.nex*

*Here is an example of the "MrBayes" block you must have at the end of your Nexus file:*

*BEGIN MRBAYES;*

*set autoclose=yes nowarn=yes;*

*charset pos1 = 1-.\3;*

*charset pos2 = 2-.\3;*

*charset pos3 = 3-.\3;*

*partition by\_codon = 3: pos1, pos2, pos3;*

*set partition = by\_codon;*

*Lset applyto=(all) nst=6 rates=invgamma; # That's GTR+G+I, change to your best model*

*unlink statefreq=(all) revmat=(all) shape=(all) pinvar=(all);*

*Prset applyto = (all) ratepr=variable;*

*mcmcp ngen=1000000 nchains=2 samplefreq=100 printfreq=10;*

*propset ExtTBR$prob=0;*

*mcmc;*

*sump;*

*sumt;*

*END;*

* *Change model with your own best model*
* *Before submitting your search to the* ***BAOBAB*** *queue, test if your command block is correct by running a search directly from the terminal. To do so, run the following commands:*

*/opt/modules/Modules/default/bin/add.modules #Type "y", and press enter, then logout and log back into your account*

*~~module load openmpi/gcc~~ should be replace by: /opt/mrbayes/mb*

*~~module load mrbayes~~*

*~~mb~~*

* *At this stage, you should have started MrBayes interactively. Execute your newly created file in the interactive MrBayes:*

*MrBayes > execute YourNEXUSfile\_codons.nex*

* *If your search starts without an error message, you can stop the search and delete the files it produced. Otherwise, find the error in your command block.*
* *Once your interactive version is running smoothly, you must submit the job to the* ***BAOBAB*** *queue using a batch file. Luckily, we made it for the previous analysis, so copy and paste that same file into your current folder*

*cp ../MB\_batchfile.sh . #Remember the period “.” at the end!*

* *Now you just have to edit the batch file so that it submits the new file. Once again use nano and change the line “srun mb YourNEXUSfile.nex” to “srun mb YourNEXUSfile\_codons.nex”. Then submit the job:*

*sbatch MB\_batchfile.sh*

*Once all the analyses are done, transfer the final consensus tree to your computer and proceed as done in the previous exercise.*

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

9. Multigene analyses with partitions (Monday afternoon)

Preparation of the concatenated alignment for the Multigene analysis with partitions (3 concatenated genes)

First you will need to generate a "concatenated" alignment, where each single-gene alignment will be merged to one another (be sure to keep the correspondance between the taxon and the sequence). We will use a simple tool included in Seaview.

In addition to your gene, you need to get the alignments of the two other genes used in this course. Ask other students. Once you have the three single-gene alignments, open them with Seaview. Check carefully if taxa have the same name in all the alignments (they don't need to be in the same order). Also, note down the length of each alignment: **Warning,** in Seaview, the alignment position is given at the top of the alignment on the left side, noted "pos:...." (it is NOT the last number given on the top right side).

Alignment 1 Gene1 name:....................................... Alignment length:..........................

Alignment 2 Gene2 name: ....................................... Alignment length:..........................

Alignment 3 Gene3 name: ....................................... Alignment length:..........................

Then close the alignment number 3.

Click on your alignment number 1 (it will be the first gene in the concatenated alignment). Go the "File" menu, choose "concatenate" and use the concatenation option "by name" (do not use the option "add gaps"). Then close the alignment number 2. Open alignment number 3 and repeat the same procedure. Then close alignment number 3. The alignment number 1 will be the concatenated alignment having the order:

Alignment1 (gene1: .................); Alignment2 (gene2: ..............); Alignment3 (gene3: .............)

Check by eye the concatenated alignment from start to end !

\* Save the concatenated alignment with a NEW NAME !!!!!! in Fasta format and then in NEXUS format. Check if the NEXUS format has been correctly generated (open the file with a Text editor and check if the data type is correct and if the interleave mode is correct).

**9.1 DNA data: MrBayes multigene analysis with partitions**

Open the nucleotide concatenated alignment in Nexus format, check that the “begin data” block resembles the example in **8.1.1**, and for adding the MrBayes command block at the end of the file use the following example:

BEGIN MRBAYES;

set autoclose=yes nowarn=yes;

charset gene1 = 1- ;

charset gene2 = - ;

charset gene3 = - ;

partition by\_genes = 3: gene1, gene2, gene3;

set partition = by\_genes;

Lset applyto=(all) nst=6 rates=invgamma;

unlink statefreq=(all) revmat=(all) shape=(all) pinvar=(all);

Prset applyto = (all) ratepr=variable;

mcmcp ngen=1000000 nchains=2 samplefreq=100 printfreq=10;

propset ExtTBR$prob=0;

mcmc;

sump;

sumt;

end;

- If you have a block that says “BEGIN NOTES; TEXT TAXON=24 TEXT='no comment';”, you can remove it.

- You can define your value for burnin by using: sump burnin=100; sumt burnin=100;

- Before performing your search in **BAOBAB**, test if your command block is correct by starting a search in your local MrBayes program.

- Then proceed as in previous MrBayes analyses.

* Upload and run this new file with MB\_batchfile.sh, as previously done (see point **8.1.1**)
* Hint: in BAOBAB, in your directory “Bayesian\_analyses”, create a new directory named “BI\_DNA\_3genes” for the outputs of this analysis.

Questions

✔ What are the estimated means of the values of the model's parameters?

This information is available in the file ending with “.pstat “ or can be found by using the command " sump burnin=*number\_of\_burn-in\_samples"* in the MrBayes window.

The parameter values are given for each partition; the partition number is in "{..}".

✔ Describe the differences between this MrBayes multi-gene based tree and previously obtained trees.

**9.2 Amino acid data: RAxML multigene analysis with partitions**

**Sequence translation**

You need first to translate your concatenated alignment. Proceed as done previously with Seaview (Exercise 4.2).

Name of the concatenated amino acid alignment in Phylip format:………..…………………….

Name of the concatenated amino acid alignment in Nexus format:………..…………...............

Check if the Nexus format has been properly generated. Open your Nexus format file with a Text editor and check if the data type is correct (DNA or protein) and check if the alignment is organized in an interleave or non-interleave manner (you may need to correct the command "interleave=no" or "interleave=yes").

You also need to retrieve the order of your genes in the concatenated alignment file and calculate the length of each of your genes in amino acids:

Length in amino acids = Length in nucleotides /

Gene order: …………………………………………………………………………...

Gene 1 length: ……………..…….. , from position 1 to ………………………….

Gene 2 length: …………..……….. , from position …………….. to ……………..

Gene 3 length: ……………………. , from position …………….. to ……………..

**RAxML analysis**

Using RAxML, you have not only the “Fixed” models as in MrBayes, but you also have a number of models that use empirical base frequencies, and you can include evolutionary rate heterogeneity among sites with a GAMMA distribution (with or without an Invariant extra category) or with an approximation to GAMMA which is called CAT.

- Models with fixed transition rates: *DAYHOFF, DCMUT, JTT, MTREV, WAG, RTREV, CPREV, VT, BLOSUM62, MTMAM, LG*

- Models with empirical base frequencies drawn from the alignment: *DAYHOFFF, DCMUTF, JTTF, MTREVF, WAGF, RTREVF, CPREVF, VTF, BLOSUM62F, MTMAMF, LGF*

- Incorporation of rate heterogeneity among sites: here we may also use the CAT approximation of rate heterogeneity.

To perform a partitioned analysis with RAxML, you need:

* your sequence alignment file in **Phylip** format (here with your amino acid alignment)
* a plain text file describing the partitions and, if needed, the different models to use.

Here are examples of plain text files with the information about partitions and models.

**For amino acid alignments**

Example 1: you can provide different substitution models for each partition:

JTT, gene1 = 1-500

WAGF, gene2 = 501-800

**For adding Gamma or GammaI, for amino acid or DNA:**

You can specify PROTGAMMAIWAG (or for DNA: GTRGAMMAI) in the options available in BAOBAB, the only thing that will be extracted from the string selected on BAOBAB is the model of rate heterogeneity you specify.

**For DNA alignments**

Reminder: RAxML only uses the GTR model of substitution probabilities (which can be combined with GAMMA, GAMMAI, CAT, CATI).

Example 1: two genes

DNA, gene1 = 1-500

DNA, gene2 = 501-1000

Example 2: intron-exon partition

DNA, introns = 1-200, 800-1,000

DNA, exons = 201-799

Example 3: Assign distinct models to the different codon positions:

DNA, gene1codon1 = 1-500\3

DNA, gene1codon2 = 2-500\3

DNA, gene1codon3 = 3-500\3

Example 3: If you only need a distinct model for the 3rd codon position:

gene1codon1andcodon2 = 1-500\3, 2-500\3

gene1codon3 = 3-500\3

**Combination of DNA and amino acids**

DNA, gene1 = 1-500  
JTTF, gene2 = 501-1000

**Good to know**: You cannot assign different models of rate heterogeneity to different partitions, i.e. it will be either CAT or GAMMA over the entire dataset.

To run the concatenated RAxML analysis in **BAOBAB**

* Copy your concatenated alignment in **PHYLIP** format, and the partitions file into your folder in **BAOBAB**
  + Linux+Mac:

scp Concat\_AA.phy Partitions.txt username@baobab.unige.ch:~

* + Windows:

C:\pscp.exe -scp Concat\_AA.phy Partitions.txt username@baobab.unige.ch:~

* Log into your **BAOBAB** account
* Move your alignment and your partition file into the “ML\_analyses” folder

mv Concat\_AA.phy Partitions.txt ML\_analyses

* Navigate into the “ML\_analyses” folder: cd ML\_analyses
* Make a copy of the batch file under a new name

cp RAXML\_batchfile.sh RAXML\_batchfile-partit.sh

* Using nano, modify the following #SBATCH values:

#SBATCH –-ntasks=**16**

#SBATCH --time=**2-00:00:00**

Also make sure that the “srun” command looks like this, with your own file names and models:

srun raxmlHPC-MPI-AVX -s YourConcatenatedAminoAcidAlignment.phy -q YourGenePartitionFile.txt -m YourBestModel -p 234666 -c 4 -f d -i 10 -# 2 -n YourConcatenatedAminoAcidOutput

Explanation about how the models (substitution matrices) need to be declared for your partitioned analysis: for each partition, the model to be used is declared in the "YourGenePartitionFile.txt" but the information about GAMMA or GAMMAI needs to be declared in the command line (included in the batch file), under the variable "-m", where you will declare one possible substitution matrix (for example JTT) and the indication(s) of GAMMA or GAMMAI. The program will take the substitution matrix (and "+F" if needed) declared in ""YourGenePartitionFile.txt", while it will take only the information about GAMMA or GAMMAI from what has been declared in the command line under "-m".

Example of a "YourGenePartitionFile.txt":

JTT, gene1 = 501-1000

JTTF, gene2 = 1001-2000

WAGF, gene3 = 2001-3000

Example of a command line (in the batch file), under the "-m" variable:

-m PROTGAMMAJTT

(*or* -m PROTGAMMAIJTT *or* -m PROTJTT).

* Once the batch file is ready, submit your job to the **BAOBAB** queue.

sbatch RAXML\_batchfile-partit.sh

* Visualize the status of your job, to see if it has finished

squeue

Question

✔ Describe the differences between this MrBayes multi-gene amino acid based tree and previously obtained trees.

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**10. Statistical tests based on phylogenetic trees** (Monday afternoon)

**10.1 Bootstrap: statistical support of the nodes**

ML analysis with bootstrap with RAxML on **BAOBAB**

* Log into **BAOBAB**
* Navigate into the ML\_analyses folder, which should still have your nucleotide alignment in **PHYLIP** format

cd ML\_analyses

* Make a copy of the RAxML batch file that you should have in this folder, and open it with nano for editing:

cp RAXML\_batchfile.sh RAXML\_batchfile-bootstrap.sh

nano RAXML\_batchfile-bootstrap.sh

* Make sure that the srun command looks like the one below, with your alignment name, correct model, and an output name that identifies the job as a bootstrap search.

srun raxmlHPC-MPI-AVX -s YourAlignment.phy -m YourBestModel -p 234666 **-b 3457345** -c 4 -f d **-# 100** -n YourOutputName-Boot

* Notice the new “-b” option and the change in “-#”, and be sure that at the beginning, the line starting with “#SBATCH --time” reads:

#SBATCH –-time=1-00:00:00

And that the one starting with “#SBATCH --ntasks” says:

#SBATCH --ntasks=16

* Sumbit your job to the **BAOBAB** queue.

sbatch RAXML\_batchfile-bootstrap.sh

* If there are no error messages, you should have a file titled “RAxML\_bootstraps...”. Now locate your best ML tree file, which should be on the same folder, and whose name starts with “RAxML\_bestTree...” and was produced on exercise 7.1. Summarize the bootstrap data:

/opt/RAxML/raxmlHPC-AVX -s YourAlignment.phy -t RAxML\_bestTree.YourOutputName -z RAxML\_bootstrap.YourOutputName-Boot -m YourBestModel -p 234666 -c 4 -f b -n YourOutputName-BootTrees

* Your final bootstrapped tree will be in a file titled “RaxML\_bipartitions...”.

Questions:

✔ Which are the phylogenetic relationships that are poorly supported?

✔ Do the bootstrap supports correspond to the values of posterior probabilities obtained with the Bayesian analysis? Which one of these two statistics gives systematically higher supports?

**10.2 ML Clock Test**

Check if the sequences evolve in a clock-like manner in your dataset (test for a global homogeneous clock-like evolution).

Start MEGA.

Open your nucleotide alignment in Fasta format. In the menu bar, under the “Clocks” menu, choose “Test Molecular Clock (ML)”.

Choose the appropriate parameters and input your best tree (for instance your RAxML tree).

Question:

✔ Are the gene sequences evolving clock-like across taxa?

**Introduction to Topological tests: SH test -----> See point 12**

Topological tests can tell us if a given dataset can significantly reject an alternative topology or, on the contrary, if the alternative topology is not significantly "worse" than the best topology (the ML topology).

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**Tuesday**

**11. Dating phylogenetic trees with BEAST**

Classical phylogenetic analyses like the ones we have done so far do not include the dimension of time. The length of branches and the divergence between species are quantified in substitutions per site.

The pace at which these substitutions occur will depend on both the time of divergence between lineages, and the rate of evolution of the lineages (in other words, it depends on how long has the lineage been evolving for, and how fast this evolution has been). However, until we add information on the age of the lineages we are studying, we are not able to separate between age and evolutionary rate.

To access this information, we can time-calibrate the phylogeny. To do this, we use fossil evidence that can tell us roughly when some lineages emerged. After identifying several of these “calibration points” within our phylogeny, we use a program like BEAST to determine, across our tree, how much of the evolution is due to time. This will result in a time-calibrated phylogeny, in which each node has the most likely date of emergence.

For this study, we will use the following three calibration points:

1. BH: Split between (baboon, macaque, colobus) and (human, chimp)

20-25 MYA (based on primate fossils; Yoder and Yang 2000)

1. EU: Split between metatherians (e.g. marsupials) and eutherians

143-178 MYA (based on fossil evidence and molecular phylogeny; Philipps et al. 2009, Luo et al. 2011)

1. TH: Emergence of the first monotreme-like mammals

190.8-199.3 MYA (based on a time-calibrated morphological phylogeny; Luo et al. 2011)

In order to time-calibrate our data, we will need 2 input files:

1. Our three amino acid datasets, each as an independent file, in FASTA format
2. Our best tree (from the amino acid concatenated analysis, for example), in NEXUS format

***\*\*\*\*\*\*VERY IMPORTANT: the names of ALL species must match in your three alignments and in your tree\*\*\*\*\*\****

**Preparation of the XML file with BEAUTi (in the BEAST package)**

Next, we must produce an input file for **BEAST**. This file will be in the XML format, and we will produce it through an application called **BEAUti**.

**BEAUti** is part of the suite of programs from **BEAST** and was installed together with BEAST.

In your computer, open **BEAUti** (it is in the same folder as BEAST) and import your alignments and your tree.

File → Import Data…

\*\*\*If you have trouble importing your tree into **BEAUti**, open it first in **figtree**, root it, and then click File → Export Trees…, and re-save it in NEXUS format. Then, try importing it again into **BEAUti**\*\*\*

* Tell **BEAUti** that you want to use the imported tree for the analysis.
  + On the “Trees” tab in **BEAUti**, and click on “User-specified starting tree”
    - On the left column, change the name of the tree to “best\_tree”
  + On the “Partitions” tab, the “Partition Tree” column, ensure that all sequences have “best\_tree”.
* Allow for each alignment to have their own model of sequence evolution and molecular clock model.
  + On the “Partitions” tab, select all alignments and click on “Unlink Subst. Models”, and “Unlink Clock Models”.
  + Under the “Site Model” and “Clock Model” columns choose the respective name of each alignment (it should be done by default).
* Set the best amino acid evolution model as you and your colleagues calculated
  + On the “Sites” tab.
* Set the random local clock model for each of the alignments.
  + On the “Clocks” tab.

Now we can set our calibration points, go to the “Taxa” tab.

For each calibration point, press the “+” button, to make a taxon set including all the species emerging from the calibration point, and press the “→” to confirm it. Name them according to the calibration point they represent.

* Tell **BEAUti** that these groups must be monophyletic.
  + Beside the name of the calibration point there is a column that says “Mono?”.
    - Make a checkmark on all calibration points
* Set a starting age for the Bayesian process.
  + In the “Age” column input, for each taxon set, the most recent age of their age range.
* Set the date prior. For this example we will set uniform priors. This means that the most recent common ancestor could have lived with equal probability any time during the time bracket given.
  + Go to “Priors” tab, and for each of the “tmrca()” parameters, click on the “Prior” field, select “Uniform” for the three genes, put the most recent age from the age range as the lower boundary, and the oldest age as the upper boundary. By default, the ages are in million years.

Finally, we must set the parameters for the Bayesian function. This is done in the “MCMC” tab. Make a chain of 1M steps, and have it echo the state to screen and save the parameter every 1000 steps. Remember to give a prefix for all your output files (in “File name stem:”).

Now you can generate the BEAST file (the button on the lower left corner).

* Copy your \*.xml file (YourBEASTFile.xml) into your folder in **BAOBAB**
  + Linux+Mac:

scp YourBEASTFile.xml username@baobab.unige.ch:~

* + Windows:

C:\pscp.exe -scp YourBEASTFile.xml username@baobab.unige.ch:~

* Create a new folder in **BAOBAB** named BEAST\_analysis, move your XML file into it, and copy the BEAST batchfile into it.

mkdir BEAST\_analysis

mv YourBEASTFile.xml BEAST\_analysis

cp /home/montoya/MolPhylEvol/Students/Batch\_Files/BEAST\_batchfile.sh BEAST\_analysis

* Edit the batchfile using the command: nano BEAST\_batchfile.sh

and put your XML file name as an input to BEAST, in last line:

beast Your\_XML\_File.xml

Check also if the name of the reservation is completed and correct:

#SBATCH --reservation=.... (complete the reservation name)

* Now you can launch your analysis!

sbatch BEAST\_batchfile.sh

This analysis will run for ~9 hours; you will collect and analyse the results tomorrow.

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**12. Topological test based on Multigene trees**

Topological tests can tell us if a given dataset can significantly reject an alternative topology or, on the contrary, if the alternative topology is not significantly "worse" than the best topology (the ML topology).

Preparation of the alternative topologies:

The aim of this exercise is to provide different credible topologies retracing the relationships between the major lineages of mammals and to describe them in Nexus format. These alternative topologies will be evaluated against your best hypothesis (best topology) by using a Topological Test, the SH test (Shimodaira-Hasegawa test). In general, alternative topologies are selected among the resulting topologies obtained by the analysis of other datasets (other genes, morphological characters, ...).

Retrieve (download from BAOBAB) and examine the results of the multigene analysis done with RAxML and based on the 3-genes Amino Acid alignment (you generated this result in point **10.2.1)**. If the topology of your multi-gene tree corresponds to one of the three hypotheses about the root of the eutherian mammals, you will compared it to the other two alternative hypotheses. If the topology of your multi-gene tree do not correspond to any of the 3 rooting hypotheses, you will compared your best topology with the three alternative hypothesis about the root of the eutherian mammals.

You will prepare the alternative topologies as follows:

To create a new topology from a current tree file in Newick format:

With MEGA.

Start MEGA. In the menu bar, choose the “User Tree” menu. Open an existing tree that includes all the taxa you need and that has taxa names of 10 characters maximum, that is, corresponding to the Phylip format (in this exercise, use your best 3-genes Amino Acid RAxML tree). You should use strictly bifurcating trees (no polytomies) so use as your best tree the RaxML tree (and not your BI tree that may have polytomies). To open and edit a user the tree, select the option "Edit/draw tree manually" in the menu "User tree".

Using the pointer, select the branch to be moved and place it at its new position. Make all the changes you need and then save your new tree with an explicit file name (in Newick format).

**Name of the file with the alternative topology n°1:.............................................................**

**Name of the file with the alternative topology n°2:.............................................................**

(if needed: Name of the file with the alternative topology n°3:…………………………...........…)

**Topological tests: SH test implemented in the RAxML package**

Preparation of a file containing all the alternative topologies

Gather all your alternative topologies in a single file by using a text editor as described below

IMPORTANT! Please note here the order in which you place your alternative topologies:

Alternative topology 1: .……………………………................................

Alternative topology 2: .…………………………….................................

(if needed: Alternative topology 3: .……………………………................................)

With a text editor, open the file of your Alternative topology 1 in Newick format.

WARNING: If you use a tree file with branch lengths, negative branch length should be avoided (this can happen with NJ) so replace the negative values by “0.000”.

Keep this file open and open the file of the Alternative topology 2 with a text editor. Copy and paste this second topology in your first file, just after your first topology.

Paste the remaining alternative topologies the same way.

Save your file containing the alternative topologies. **Name of the file**: .....................................

**Important**: - You will need to use an alignment file in Phylip format, so the species names should be 10 characters long followed by 2 empty spaces before the start of the sequence.

- check if the species names are identical in all the files (alignment file and file with the topologies)

Topology test with **RAxML, the SH test**

We will use the SH test (Shimodaira-Hasegawa test), as implemented in RAxML. You don’t need to perform partitioned analyses for the topological tests (it’s possible but a bit complicated).

* Copy the file with the alternative topologies to be tested to your folder in **BAOBAB**
  + Linux+Mac:

scp TreesToBeTested.nwk username@baobab.unige.ch:~

* + Windows:

C:\pscp.exe -scp TreesToBeTested.nwk username@baobab.unige.ch:~

* Log into your **BAOBAB** account
* Move the file with the alternative topologies to be tested into the folder “ML\_analyses”

mv TreesToBeTested.nwk ML\_analyses

* Navigate to said folder: cd ML\_analyses
* Make sure that you have the following files on this folder:
  1. Your alignment in **PHYLIP** format (should be there from previous exercises)
  2. Your best 3-genes Amino Acid RAxML tree (should also be there, starts with “RAxML\_bestTree…”) (you generated this result in point **9.2)**
  3. The file you just uploaded with the set of alternative topologies to be tested
* Run the SH-test

/opt/RAxML/raxmlHPC-AVX -s YourAlignment.phy -m GTRGAMMA -p 234666 -c 4 -f h -t YourBestTree.nwk -z TreesToBeTested.nwk -n YourOutputName-SHTEST

Write down here YourOutputName: ……………………………………………………

The results of the SH test are in the file “RAxML\_info.RAxML”. At the bottom of this file you will find the details of this test and conclusions about rejection or not at three different significance levels: 5%, 2%, and 1%.

Questions:

✔ What are the conclusions of the SH test based on the 3-genes amino acid alignment concerning the different comparisons?

✔ Based on your data set, which topologies are significantly rejected?

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**Wednesday**

**13. Analysis of the BEAST results**

Once your analysis is finished (~9 hours), you can recover your data and visualize with the **Tracer** program how the parameter estimations progressed. Copy all the results files (\*.log, \*.ops, and \*.trees files) into a designated folder in your computer.

Launch **Tracer** an open your \*log file. What do you see?

* Are the likelihood values distributed more or less normaly? Is there a single mean?
  + Press on the “Trace” tab, and see how the likelihood values progressed through the chains. Do you see large changes? Do you agree with the suggested burn-in, or do you think more has to be burnt-in? Include an image of the trace for your report.
* How do the distributions for each of the three “tmrca()” parameters look like?

Now we can also create a consensus tree where we will see the estimated dates for all nodes in our phylogeny. For this we will use a program included in the **BEAST** suite, called “**treeannotator**”. Launch **treeannotator** and load your \*.trees file obtained from the BEAST analysis. This file contains all the trees sampled in the **BEAST** analysis. Specify as burnin the first 100,000 states (10%), and set the node heights at “Mean heights”. Before clicking “Run”, give a name to your output file (it will be a tree in NEXUS format), you can leave the other values in their default. Once you’ve ran it, open the program again, and build a new consensus tree using your own burn-in value. Don’t forget to name it differently!

Now visualize your trees using **Figtree**. In order to see the node ages, activate “Node Labels”, and make them display “heights”. In order to place node bars indicating the 95% highest posterior density (HPD) intervals, activate “Node Bars” and make them display the “height 95% HPD”. Save an image of each tree for your final report.

Questions:

* Do you see any difference in topology between your trees? Any difference in the node ages?
* In “Node Labels”, select “posterior”. This number is the posterior probability of each node, based on your data. How do these numbers change between your trees with different burn-ins?

**14. Analyses and interpretations of results based on multi-gene data**

You have time to assemble and analyze the results you obtained with the multi-gene data. Compare your results across the methods and data type. You can also compare and discuss your results with other participants.

**15. Final synthesis, discussion and conclusion**

Please organize your results in a summary Excel table, given at the end of this manual. We will then collect the results of all participants to discuss together about the pro and cons of the methods and approaches. We will see if we can have a convincing answer to the question about the root of Eutherian Mammals.

\_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_ \_

**16. Optional: Analysis of selection pressure in primates using Paml**

The Paml package is not really destined for inferring phylogenies as it is very sophisticated and thus too slow, but it is used for numerous analyses based on Maximum Likelihood (ML) procedures. It includes several programs of which the most important are:

-Baseml: analysis of non-coding sequences

-Codeml: analysis of coding sequences

We are restricting our analysis to primates because:

a) we want to reduce the amount of time that the analysis takes, and

b) an analysis of selective pressure is more case sensitive when only closely related species are analyzed.

Building an alignment concerning primates only

Open the DNA alignment of your gene using Seaview. Delete all non-primate sequences; for this, select their names, then go to the menu “Edit” and click on the option “Delete Sequences”. There should be 10 sequences remaining.

Verify gaps, they should all be marked by “-“ (and not “~” or “?”). Save the alignment as “Phylip” format (menu “File”, option “Save As”).

Creation of a tree containing primates only

Open the file containing your best tree with Text Editor. Eliminate all taxa names and branch lengths that do not concern primates. The tree must stay unrooted (= trichotomy at the base).

Example:(taxa1, taxa 2,((taxa3,taxa4), (taxa5, taxa6))).

Verify that the number of opened brackets is equal to the number of closed brackets!

Finally save the tree in Newick format under a new name and reopen it with Figtree to check if everything is correct and if the topology corresponds to your expectations.

Alternatively: you can used MEGA to generate your primate tree. Start MEGA, open your primates alignment and generate an NJ tree. Edit the NJ tree so that is corresponds to your best primate topology. Save this tree in Newick format.

Alternatively: You can find an example of the primate tree (“primates\_tree.tre”) in BAOBAB (folder /home/montoya/MolPhylEvol/Students).

**16.1 Omega (dN/dS) calculation for each branch of the tree = Analysis 1**

1. To perform this analysis with Paml you need:

\* An alignment file in Phylip format of a protein coding sequence with the Open Reading Frame (ORF) starting at position 1 of the alignment and without the final STOP codon.

\* A tree file with the same set of taxon and names as in the alignment file

\* A control file with the set of commands for the program (Codeml, Baseml, …)

1. The Phylip format that is recognized by Paml needs to be slightly modified in the following way:

\* open your alignment file in Phylip format with a text editor, go to the first line and place the cursor just after the number of sites of the alignment (end of first line). Add a space and then an “I”.

Example: “ 10 5025 I”

This exercise can be performed on **A) your local computer** or in **B) BAOBAB**, up to you!

**A) On your local computer**: Download and install Paml on your computer:

<http://abacus.gene.ucl.ac.uk/software/paml.html#download>

Explore the content of the Paml folder and check the location of the “bin” sub-folder which contains the executable files and the control files (.ctl).

\* Save your modified alignment file in the “bin” folder located within the PAML folder.

In the “bin” folder of Paml, you will find the “codeml.ctl” control file.

**B) In BAOBAB (first part):** You need to copy into your personal computer a file named “codeml.ctl” that you will find in BAOBAB. Open a Terminal command session and reach the location in which the folder Phylo\_course\_2015 folder is, and copy the “codeml.ctl” file in it:

* + Linux+Mac:

scp username@baobab.unige.ch:/home/montoya/MolPhylEvol/Students/codeml.ctl Phylo\_course\_2015/.

* + Windows:

C:\pscp.exe -scp

username@baobab.unige.ch:/home/montoya/MolPhylEvol/Students/codeml.ctl Phylo\_course\_2015/.

**In both cases, A and B:**

1. Edit the control file “codeml.ctl” according to the search you would like to perform. Open the control file “codeml.ctl” with a text editor. A list of commands will appear that can be used for different search options, all of them using the maximum likelihood (ML) method. Lines starting with a star (\*) are not considered by the program; these lines contain explanations or a list of available options.

\* Specify the name of your alignment file (1st line).

\* Specify the name of your tree file (2nd line). Codeml will estimate the different parameters that have been selected based on a given topology. Use therefore the tree that seems to be the best one (here: your primates-only tree).

\* Specify the name of the output file = outfile (3rd line).

\* Indicate that your alignment consists of coding sequences (7th line):

Seqtype=1

\* Choose frequency parameters for each codon (line 8). The options are:

Codon Freq = 0 (same frequency for each codon: 1/61)

Codon Freq = 1 (calculation according to the mean frequency of each nucleotide)

Codon Freq = 2 (calculation according to the mean frequency of each nucleotide at each codon position)

Codon Freq = 3 (list of frequencies given by the user)

Use “CodonFreq=2”

\* To calculate one omega value (omega=dN/dS) for each branch of the tree as asked in the first analysis you have to set in the 13th line:

Model = 1

\* To calculate only one omega value for all the codons in the alignment (= for the entire gene), as asked in the first analysis, you have to set in the 14th line:

Nssites=0

\* To include values for standard errors (25th line):

getSE=1

1. Save the control file “codeml.ctl” and close it.

If you have a local version of Paml:

Test your control file by starting the analysis using the locally installed PAML (in your computer). You have to go to the directory where Codeml is located.

For Windows users: Go to the “Démarrer” menu of windows, choose “Programmes”, select “Accessoires” and “Command Prompt”. Then move to the location of Codeml; to do so, enter the command “cd” followed by a nonbreaking space and the path description for the Codeml program. Example: "cd C:\ProgramFiles\Paml3.15\bin\".

Then, enter the command “Codeml” to start the analysis.

In case the program is interrupted and the indication “*Species* number?” appears on the screen you have to open your tree file with text editor and remove all numbers that are not part of sequence names as well as all colons “:”. The reason for this is that the program cannot cope with branch lengths numbers and other clade numbers that are present in your file. Save the tree file once you have removed all disruptive factors and restart the analysis.

If there is no problem, the Codeml program will run for about 20 min.

**B) In** **BAOBAB (continuation)**

* Copy your alignment in **PHYLIP** format, your tree file and your control file into your folder in **BAOBAB**
  + Linux+Mac:

scp YourAlignment YourTreeFile YourControlFile username@baobab.unige.ch:~

* + Windows:

C:\pscp.exe -scp YourAlignment YourTreeFile YourControlFile username@baobab.unige.ch:~

* Log into your **BAOBAB** account
* Create a new folder for each of your PAML analyses titled “PAML\_analysis1”, “PAML\_analysis2”, and “PAML\_analysis3” (number 2 and 3 will be used for the next exercise)

mkdir PAML\_analysis1 PAML\_analysis2 PAML\_analysis3

* Copy the alignment and the tree file into all three folders

cp YourAlignment YourTreeFile PAML\_analysis1

cp YourAlignment YourTreeFile PAML\_analysis2

cp YourAlignment YourTreeFile PAML\_analysis3

* Move your control file into the folder for your first analysis, “PAML\_analysis1”

mv YourControlFile PAML\_analysis1

* Navigate into that folder

cd PAML\_analysis1

* Execute the control file for your first PAML analysis

/opt/paml/bin/codeml YourControlFile

Once the search is finished, copy all the output files into a new folder of your personal computer. To this aim, you should open a Terminal command session in your computer and go to the location of "Phylo\_course\_2015" (cd Phylo\_course\_2015).

Create a new folder named "PAML\_analyses" (mkdir PAML\_analyses). Then navigate into that folder (cd PAML\_analyses).

Then copy all the outputs of Paml **from BAOBAB** into a new folder, in the local folder "PAML\_analyses", in your personal computer:

Linux+Mac:

scp -r username@baobab.unige.ch:~/PAML\_analysis1 *YourNewFolderName(forExample:myPAML\_analysis1)*

Windows:

C:\pscp.exe -scp -r username@baobab.unige.ch:~/PAML\_analysis1 *YourNewFolderName(forExample:myPAML\_analysis1)*

**In both cases, A and B:**

Open then the file “.rst” that contains the ancestral reconstructed sequences. Go to line entitled “ tree with node labels for Rod Page’s TreeView” (it’s approximately the 7th one). There you will find the topology with the code for each node of the tree, in Newick format. Copy this codified tree, open the Figtree program and paste this tree into the Figtree window. Click on “Node labels” and you will see the numbers for each node of the tree. You need to know the node numbers to interpret the outfile results.

Open the file “outfile” with a text editor and scroll down until you find a list with omega values for each branch of the tree.

Optional questions:

✔ Put this list in your report.

✔ Which branch shows the biggest and which one the smallest omega value?

✔ Are some lineages characterized by an increasing/decreasing trend of omega with time ?

To check that, you can plot the trees with branch lengths representing only dS (dS tree) or representing only dN (dN tree) by copying the respective lines at the end of the file and pasting them into the Figtree window.

✔ When comparing the dS and the dN trees, which branch shows the biggest change in length as compared to the others?

**16.2 Search for codons under positive selection: Site models *=* Analysis 2**

Here we will calculate the likelihood score (lnL) of our tree using two different models of sequence evolution, called M1a and M2a.

Model M1a: includes two types of sites, those with omega < 1 and those with omega = 1. It does not take into consideration positive selection (omega > 1) and is therefore called the “nearly neutral” model.

Model M2a: includes three types of sites, those with omega < 1, with omega =1 and with omega > 1. It is called the “positive selection” model.

The likelihood scores established according to these two models will be compared using a likelihood ratio test to determine whether or not positive selection is to be invoked to better explain our data.

Only if the “positive selection” model (M2a) turns out to fit our data significantly better, one can search the codons that evolve under positive selection.

Searching for codons under positive selection:

Modify the control file used on the previous exercise, to produce two additional control files (one for each model):

1. Choose a new name for the output file (3d line)
2. Use the same omega value for the whole tree (13th line):

Model = 0

1. define the 2 models that will be used in the analysis (14th line):

Nssites = 1 (for the nearly neutral model)

Nssites = 2 (for the positive selection model)

Save the control files as “control-nn.ctl” (for the nearly neutral models), and “control-ps.ctl” (for the positive selection model) on their respective PAML\_analysis folders.

**In the local version of Paml**: In the Command Prompt window, type “codeml” and the analysis will start. Duration of the search: about 30 min.

Once finished, the program will have generated nine different files. Create a new folder and copy all files into this folder.

**In BAOBAB**: Proceed as before, uploading each control file to a different folder, and navigating into that folder to execute the “/opt/paml/bin/codeml YourControlFile” command.

Optional questions:

✔ What is the likelihood score for the “nearly neutral” model (M1a)?

✔ What is the likelihood score for the “positive selection” model (M2a)?

✔ Does positive selection play a role? Do a likelihood ratio test with the lnL scores of the 2 models. What is the value of df? What is the Chi-square value? What is the associated

*P*-value? (Find a website to calculate the *P*-value for Chi-square tests).

✔ If the model M2a is significantly better than the model M1a is it possible to identify codons

under positive selection pressure using Bayes Empirical Bayes (BEB) analysis? Which codons

exactly are these?

**17. ANNEX 1: Model definitions for MrBayes**

[JC] lset nst=1 rates=equal; prset statefreqpr=fixed(equal);

[F81] lset nst=1 rates=equal; prset statefreqpr=dirichlet(1,1,1,1);

[JC+I] lset nst=1 rates=propinv; prset statefreqpr=fixed(equal) pinvarpr=uniform(0,1);

[F81+I] lset nst=1 rates=propinv; prset statefreqpr=dirichlet(1,1,1,1) pinvarpr=uniform(0,1);

[JC+G] lset nst=1 rates=gamma; prset statefreqpr=fixed(equal) shapepr=uniform(0.1,50);

[F81+G] lset nst=1 rates=gamma; prset statefreqpr=dirichlet(1,1,1,1) shapepr=uniform(0.1,50);

[JC+G+I] lset nst=1 rates=invgamma; prset statefreqpr=fixed(equal) shapepr=uniform(0.1,50) pinvarpr=uniform(0,1);

[F81+G+I] lset nst=1 rates=invgamma; prset statefreqpr=dirichlet(1,1,1,1) shapepr=uniform(0.1,50) pinvarpr=uniform(0,1);

[K80] lset nst=2 rates=equal; prset statefreqpr=fixed(equal) tratiopr=beta(1,1);

[HKY] lset nst=2 rates=equal; prset statefreqpr=dirichlet(1,1,1,1) tratiopr=beta(1,1);

[K80+I] lset nst=2 rates=propinv; prset statefreqpr=fixed(equal) tratiopr=beta(1,1) pinvarpr=uniform(0,1);

[HKY+I] lset nst=2 rates=propinv; prset statefreqpr=dirichlet(1,1,1,1) tratiopr=beta(1,1) pinvarpr=uniform(0,1);

[K80+G] lset nst=2 rates=gamma; prset statefreqpr=fixed(equal) tratiopr=beta(1,1) shapepr=uniform(0.1,50);

[HKY+G] lset nst=2 rates=gamma; prset statefreqpr=dirichlet(1,1,1,1) tratiopr=beta(1,1) shapepr=uniform(0.1,50);

[K80+G+I] lset nst=2 rates=invgamma; prset statefreqpr=fixed(equal) tratiopr=beta(1,1) shapepr=uniform(0.1,50) pinvarpr=uniform(0,1);

[HKY+G+I] lset nst=2 rates=invgamma; prset statefreqpr=dirichlet(1,1,1,1) tratiopr=beta(1,1) shapepr=uniform(0.1,50) pinvarpr=uniform(0,1);

[SYM] lset nst=6 rates=equal; prset revmatpr=dirichlet(1,1,1,1,1,1) statefreqpr=fixed(equal);

[GTR] lset nst=6 rates=equal; prset revmatpr=dirichlet(1,1,1,1,1,1) statefreqpr=dirichlet(1,1,1,1);

[SYM+I] lset nst=6 rates=propinv; prset revmatpr=dirichlet(1,1,1,1,1,1) statefreqpr=fixed(equal) pinvarpr=uniform(0,1);

[GTR+I] lset nst=6 rates=propinv; prset revmatpr=dirichlet(1,1,1,1,1,1) statefreqpr=dirichlet(1,1,1,1) pinvarpr=uniform(0,1);

[SYM+G] lset nst=6 rates=gamma; prset revmatpr=dirichlet(1,1,1,1,1,1) statefreqpr=fixed(equal) shapepr=uniform(0.1,50);

[GTR+G] lset nst=6 rates=gamma; prset revmatpr=dirichlet(1,1,1,1,1,1) statefreqpr=dirichlet(1,1,1,1)

shapepr=uniform(0.1,50);

[SYM+G+I] lset nst=6 rates=invgamma; prset revmatpr=dirichlet(1,1,1,1,1,1) statefreqpr=fixed(equal) shapepr=uniform(0.1,50) pinvarpr=uniform(0,1);

[GTR+G+I] lset nst=6 rates=invgamma; prset revmatpr=dirichlet(1,1,1,1,1,1) statefreqpr=dirichlet(1,1,1,1) shapepr=uniform(0.1,50) pinvarpr=uniform(0,1);

