

## Exercise 10: Setting up a mixed nucleotide model for MrBayes search

(BIO332 Phylogeny 2007. E.Willassen) [PDF version](#)

### The nexus data file for MrBayes

See if you can find an example file in the MrBayes folder called **KIM.NEX**. It shows how different data types can be input from the same file, as long as they are properly defined in terms of data (symbol) types and positions in the alignment:

```
#NEXUS
[This is an example of a model using a complex partitioned model, including
analyzed using a codon model.]

[Data from Kim, S., K. M. Kjer, and C. N. Duckett. 2003. Comparison between
morphological-based phylogenies of galerucine/alticine leaf beetles
(Coleoptera: Chrysomelidae). Insect Syst. Evol. 34:53-64.]

begin data;
  dimensions ntax=27 nchar=1742;
  format datatype=mixed(rna:1-516,dna:517-1398,protein:1399-1692,standard
matrix
-
```

If your starting point is a nexus file with a taxon and a characters block, execute the the data in PAUP\* and save it with the option `format= nex`.

In this exercise we use the data file originally from Exercise 5 that has later been modified to contain a PAUP\* SETS block as follows:

```
begin sets;
  charset div = 1 458;
  charset ND4 = 2-457; [! NB 458 + extra AA make stop codon]
  charset tRNA1 = 459-528; [tRNA_His]
  charset tRNA2 = 529-588; [tRNA_Ser]
  charset tRNA3 = 589-659; [tRNA_Leu]
  charset ND5 = 660-898;

  charset 1st = 2-457\3 660-898\3;
  charset 2nd = 3-457\3 661-898\3;
  charset 3rd = 4-457\3 662-898\3;

  charset coding = 2-457 660-898; ←
  charset trna = 459-659;

end;
```

Open the file with a text editor (WordPad). Mark out and copy the full content of the SETS block, including the expressions **Begin** and **end**; . Paste the copied text in after the SETS block and edit the second "begin sets" so that it reads `begin mrbayes` instead.

In our data for PAUP\*, we defined a character partition called **div** consisting of two positions only (see above). We want to exclude those characters from the phylogeny estimates with MrBayes and need to do a little trick to do that. First, we assign the two positions as members of the character set coding (upper red arrow). Then, we define character partitions that sum up to include *all* the characters in the matrix. Finally, we can exclude those two characters. **MrBayes will return an error message unless the partitions are defined with all characters included.** Notice also how PAUP\* language "`charpart genegroup = 1:coding, 2:trna`" translates to "`partition genegroup=2:coding, trna`" in MrBayesian. We can define many different partitions in a similar way. The `set`

partition command makes one particular partition active in a run, in this example the partition **genegroup**.

```

charset coding = 1-458 660-898;
charset trna = 459-659;

partition genegroup = 2:coding,trna;
exclude div;
set partition = genegroup;

lset applyto = (all) nucmodel=4by4;
lset applyto = (all) nst=6;
lset applyto = (1) rates=invgamma ngammacat=4;
lset applyto = (2) rates=gamma ngammacat=4;

prset applyto = (all) statefreqpr=dirichlet(1,1,1,1);

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

```

### Setting models and priors

With partitions set, it is time to set the models with the **lset** command. Our use of MrModeltest in Exercise 9 suggested variants of the GTR model for both partitions. We therefore apply **4by4** nucleotide models with six substitution rates for both. It can be expressed either with **applyto=(all)** or **applyto=(1,2)**. In the latter case, the numbers refer to the first and second part of the partitions. We additionally set up an I+G specification for the coding data with **rates=invgamma**. The tRNAs will be modeled with gamma rates only. In both cases, we define four categories for the gamma rates.

We simply use the results from MrModeltest to define the prior probabilities ( **prset** ) for the state frequencies. With **lset** and **prset** commands in place, we could save the file as "myfile.nex" (without closing it) and test run it: Start MrBayes. On the prompt **>**, type execute **myfile.nex** and see what happens. If the settings load OK, you can start the sampling by executing the command **MCMC**. Just run a few generations for now and stop the execution by pressing the **Ctrl** and **C** keys. Open the MrBayes folder and notice the files that were produced by the short run: Unless you specified otherwise, the parameter file was named **myfile.nex.p** and the tree file **myfile.nex.t**. Open the first in Excel, and the second in WordPad to see their structure. If you kept **myfile.nex** open in WordPad, you can continue editing and / or correcting commands that did not execute well.

Take the opportunity to get acquainted with the **unlink** command (green arrow above). It is important in runs with partitioned data. You may get a better idea of what this command does if you do a small experiment: First, make this command line passive by enclosing it in brackets [ ]. Execute the file with one or two sampled generations, and open the parameter file (with extension **.p**) in Excel. Notice that there is one column for each parameter sample:

D	E	F	G	H	I	J	K	L
r(A<->C){all}	r(A<->G){all}	r(A<->T){all}	r(C<->G){all}	r(C<->T){all}	r(G<->T){all}	pi(A){all}	pi(C){all}	pi(G){a
0.166667	0.166667	0.166667	0.166667	0.166667	0.166667	0.25	0.25	
0.162167	0.14542	0.128264	0.170264	0.236489	0.157396	0.25	0.25	
0.134897	0.182751	0.10453	0.190777	0.24517	0.141875	0.272447	0.304494	0.1

Next, reopen the input file and remove the brackets. Run MrBayes for a few generations again, this time with unlinked parameters. Then look at the new parameter file. Now we see separate estimates for each partition. The columns marked with red arrows below are results of the **unlink revmat(all)** command:

D	E	F	G	H	I	J	K	L
r(A<->C){1}	r(A<->G){1}	r(A<->T){1}	r(C<->G){1}	r(C<->T){1}	r(G<->T){1}	r(A<->C){2}	r(A<->G){2}	r(A<->
0.166667	0.166667	0.166667	0.166667	0.166667	0.166667	0.166667	0.166667	0.
0.161425	0.163435	0.144017	0.171881	0.207287	0.151955	0.137399	0.176536	0.
0.142103	0.193952	0.156863	0.110404	0.235121	0.161556	0.140356	0.173515	0.

With the command **prset ratepr = variable**, we allow for the substitution rates to be variable over the partitions.

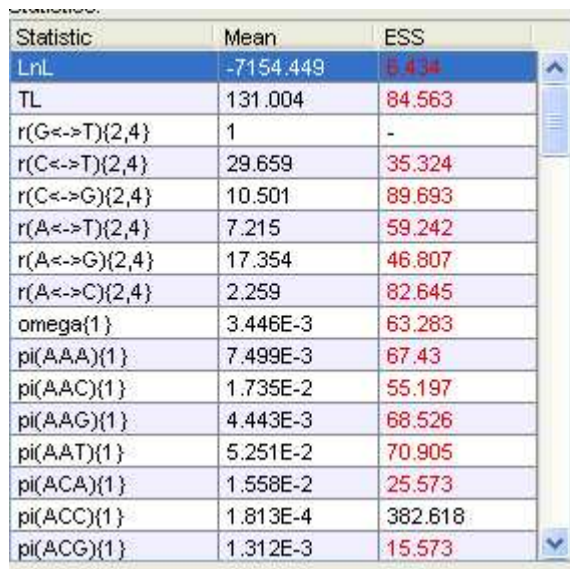
## Setting up the MCMC

A simple mcmc command could look like this: `mcmc ngen=1000000 samplefreq=100 printfreq=100 savebrlens=yes;`. It will tell MrBayes to run 1 mill generations (`ngen`). The parameter estimates will be sampled (`samplefreq`) every hundredth generation and print to screen (`printfreq`) with the same frequency. The trees in the tree file will be saved with branch lengths. If you execute this line of commands, you will see that MrBayes runs two (`nruns=2`) simultaneous runs, each with three hot and one cold chain (`nchains=4`). These settings serves us well for now, but you may consider changing those settings by putting these commands on the mcmc line with other number values. To make sure that the chains mix well and explore the landscape, we may sometimes also consider if the default temperature setting, `temp=0.2`. Study the manual to read what effects a change could imply on the mixing and acceptance rates for moves in the cold chain.

Before we start a serious run that may take days and weeks with more complicated data, we need to consider how to deal with the emerging critical question: Did we sample enough data from the parameter space to get a representative picture of the posterior distribution? In other word, did we run MrBayes through a sufficient number of generations? You could take some time to try out different procedures to make that judgment:

The "Average standard deviation of split frequencies" (SDSF) is a diagnostic tool that indicates convergence between the independent runs when approaching zero. It is reported by default every 1000 generation. We may use `SDSF=0.01` as a rule of thumb for convergent runs. Unless we gave MrBayes the command `set autoclose=on`, it will run the specified number of generations, then stop and give you the option of running an additional number of generations. Your answer `yes` or `no` may be based on the current value of SDFS. If it is lower than your stop rule, you discontinue the run. You can mak this step automatic by adding the following on the mcmc command line: `mcmcdiagn=yes stoprule=yes stopval=0.01`.

Also, recall from a previous [chapter](#) on this CD how the programme [Tracer](#) can be used to calculate effective sample size of all the parameters in the Bayesian reconstruction. Consult instructions there on how to finally use the SUMT command with burnin to make a consensus tree with branch lengths and posterior probabilities.



Statistic	Mean	ESS
LnL	-7154.449	5.434
TL	131.004	84.563
r(G<->T){2,4}	1	-
r(C<->T){2,4}	29.659	35.324
r(C<->G){2,4}	10.501	89.693
r(A<->T){2,4}	7.215	59.242
r(A<->G){2,4}	17.354	46.807
r(A<->C){2,4}	2.259	82.645
omega{1}	3.446E-3	63.283
pi(AAA){1}	7.499E-3	67.43
pi(AAC){1}	1.735E-2	55.197
pi(AAG){1}	4.443E-3	68.526
pi(AAT){1}	5.251E-2	70.905
pi(ACA){1}	1.558E-2	25.573
pi(ACC){1}	1.813E-4	382.618
pi(ACG){1}	1.312E-3	15.573