Exercise 9: Using MrModeltest to find evolutionary models for partitioned data

(BIO332 Phylogeny 2007. E.Willassen) PDF version

Mixed data

When we have assembled data from different sources, be it different genes, or gene sequences and morphological data, it would often intuitively feel reasonable to use different models for these different parts. We have already worked on a data set (Exercise5.nex) comprised by different genes and used the same model for the whole set in ML optimization. In this exercise we will run MrModeltest on two partitions and use the results to set up a Bayesian search with mixed models. MrModeltest works very much like Modeltest, but has fewer models defined in the batch file for PAUP*, which is called MrModelblock. For simplicity, let us divide the data in two groups: protein coding sites and tRNAs. We may define these character sets in a SETS block to PAUP*.

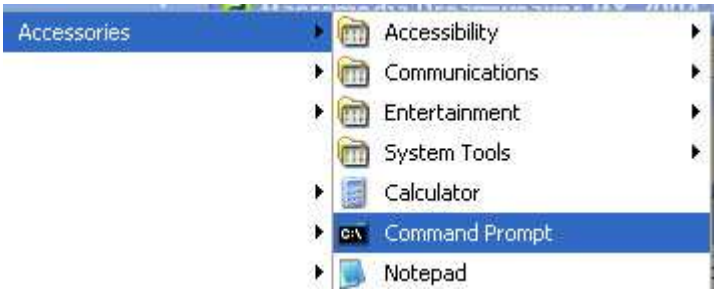
```
begin sets;
    charset div = 1 458;
    charset ND4 = 2-457; [! NB 458 + extra AA make stop codon]
    charset tRNA1 = 459-528; [tRNA_His]
    charset tRNA2 = 529-588; [tRNA_Ser]
    charset tRNA3 = 589-659; [tRNA_Leu]
    charset ND5 = 660-898;

    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;
```

Execute the data file with the character set definitions. Exclude all but the coding characters by executing the command include coding/only. Next, open MrModelblock and execute the file. When the run has finished, find the score file mrmodel.scores and rename it coding.scores. Then run MrModeltest.



MrModeltest may be handled exactly the same way a did with Modeltest (Exercise 8)

type mrmodeltest <coding.scores> mrmodeltest_co

When the run is finished, you may open mrmodeltes WordPad to study the results. Notice that MrModetes comparing models with hLRT, differing by the startir hierarchy:

```
ATTENTION: The choice based on hLRT can be sensitive for the specific
           hierarchy used. If selected models differ, user need to
           make the choice!

           Model selected by hLRT (default): GTR+I+G
           Model selected by hLRT2:          GTR+I+G
           Model selected by hLRT3:          GTR+I+G
           Model selected by hLRT4:          GTR+I+G
--


Model selected: GTR+I+G
  -lnL =       4718.2397
   K =         10
  Base frequencies:
    freqA =     0.3465
    freqC =     0.3529
    freqG =     0.0607
    freqT =     0.2399
  Substitution model:
    Rate matrix
    R(a) [A-C] =        0.9498
    R(b) [A-G] =       25.8090
    R(c) [A-T] =        0.8427
    R(d) [C-G] =        0.9090
    R(e) [C-T] =        9.8213
    R(f) [G-T] =        1.0000
  Among-site rate variation
    Proportion of invariable sites (I) = 0.2525
    Variable sites (G)
    Gamma distribution shape parameter = 0.9037
```

In this case, all hLRTs suggested the General Time Reversible (GTR or REV) with a proportion of invariable sites (I) and gamma correction (G) for rate heterogeneity over the sequence. Like in the case with Modeltest, we see that MrModeltest also provides some lines with a model definition that can be copied and pasted to a PAUP* block. The new thing is the MrBayes block:

```
[!
MrBayes settings for the best-fit model (GTR+I+G) selected by hLRT in MrModeltest 2.
]
BEGIN MRBAYES;
        Prset statefreqpr=dirichlet(1,1,1,1);
        Lset  nst=6  rates=invgamma;
END;
```

The second line (red arrow) with the lset command defines the model for MrBayes. The first line (blue arrow) defines the prior probabilities for the state (base) frequencies. If you look at the file named ML_models_in_MB3b4_and_PAUP.pdf, you will get more information about how to define nucleotide models for MrBayes.

We now have a model to use with MrBayes for the coding data. We follow the same procedure to find a model for the partition containing the three tRNAs. Testing with four hLRTs gives us the choice between the HKY+I and the GTR+G models. Akaike prefers the latter model, so we go for that:

```
BEGIN MRBAYES;
        Prset statefreqpr=dirichlet(1,1,1,1);
        Lset  nst=6  rates=gamma;
END;
```

In Exercise 10, we will prepare the data for MrBayes and use these models to infer a tree with posterior probabilities on branches.