Exercise 8: Using ModelTest to find an evolutionary model for your data

(BIO332 Phylogeny 2007. E.Willassen) PDF version

The Modelblock batch file

Recall from Exercise 7 how we used PAUP* to estimate the log likelihoods (-lnL) of trees from the data, given a particular model of evolution, which in that case was the HKY85. One part of the ModelTest package is a nexus format batch file that, when executed in PAUP* produces -lnL estimates of a particular ("reasonably good") tree under 56 different models of DNA evolution. You should find this batch file called modelblockPAUPb10.nex. Open it with WordPad and try to understand what it does. Firstly, PAUP* is told to write a log file which is more or less exactly a saved version of what you see in the buffer window when PAUP* is working. It names the log file modelfil.log and it may be wise to remember that the replace option of the command will overwrite a file with the same name from a previous run (red arrow). On the next line (blue arrow), the Dset command tells PAUP* to use Jukes-Cantor distance (JC) and the Minimum Evolution (ME) criterion in a distance based tree calculation. If some branch length comes out negative as a result of the calculations, that value will be set to zero instead. Finally, a neighbour joining tree (NJ) is calculated and stored in memory without being printed to the screen (green arrow).

```
BEGIN PAUP;
    log file= modelfit.log replace;
    DSet distance=JC objective=ME base=equal rates=equal pinv=0
    subst=all negbrlen=setzero;
    NJ showtree=no breakties=random;
End;
```

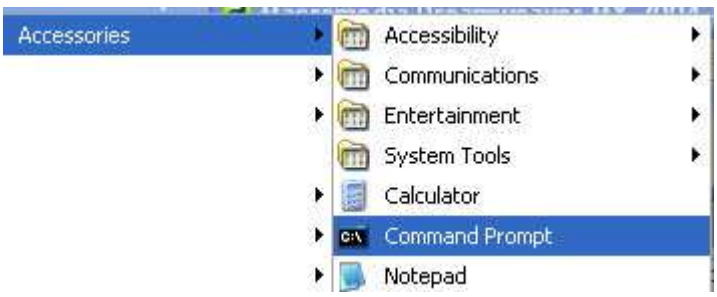
The next step (see below) is that log likelihood scores are calculated for the tree (just the first one, if the NJ results in more than one tree) under the 56 different model definitions. Notice, for example, how the model is set to HKY85 in model 13 of 56 (grey arrow). You may recognize the setting of this model from Exercise 7. PAUP* is instructed to append the likelihood scores from each model estimate to a file called model.scores (black arrow).

```
[!
** Model 12 of 56 * Calculating K80+I+G **]
lscores 1/ nst=2 base=equal tratio=est rates=gamma shape=est pinv=est
scorefile=model.scores append;

[!
** Model 13 of 56 * Calculating HKY **]
lscores 1/ nst=2 base=est tratio=est rates=equal pinv=0
scorefile=model.scores append;
```

When the batch file has finished to execute, we may use the results in model.scores to test for the best model.

Running ModelTest



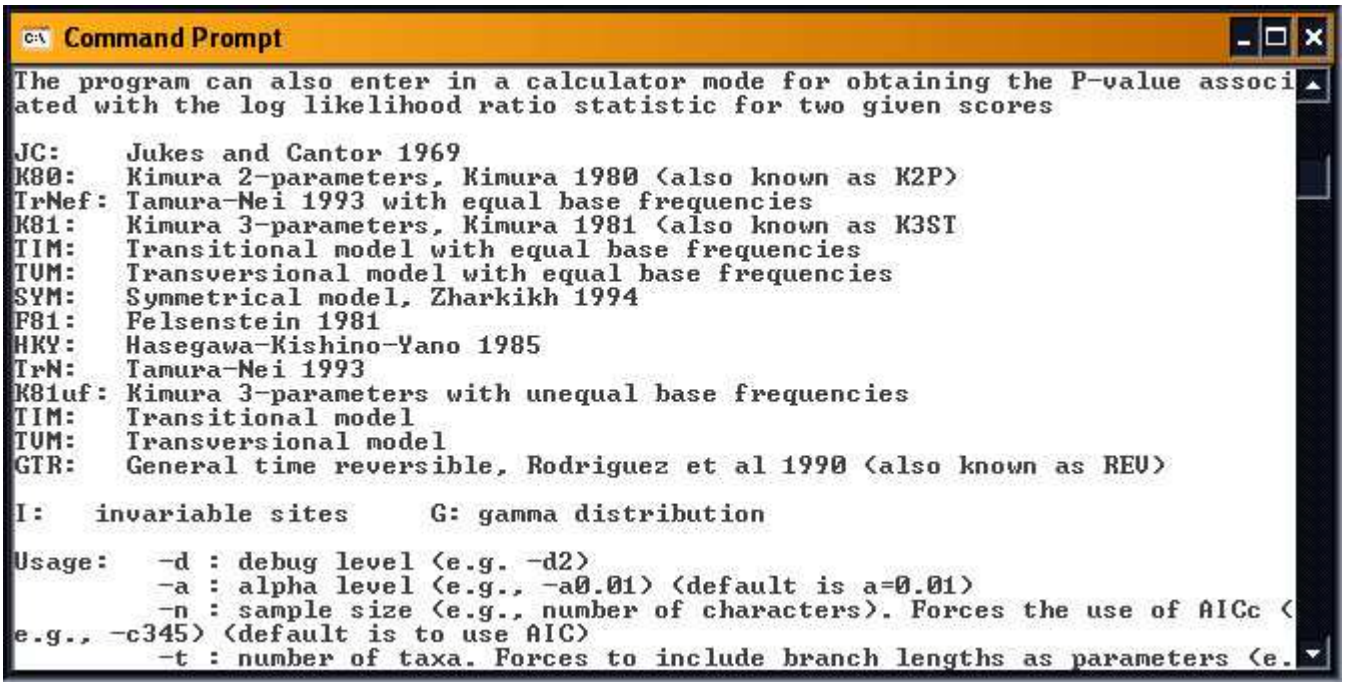
In order to run Modeltest on a Windows computer, y
command prompt window via start>program>acces
manouver to the folder where Modeltest is situated b
(change directory) DOS command. Make sure that y
modeltest exe file is called exactly (may be modelte
the model.scores file from your last run with PAUP*
folder.

type modeltest <model.scores> modeltest.txt

When the run is finished, you may open modeltest.t
study the results. Before you do that, execute the cc

-?

ModelTest responds by displaying (see below) explanations for the abbreviated model names that you will see when you open the modeltest.txt file.



Interpreting results from ModelTest

When you open modeltest.txt, you see the results from the HIERARCHICAL LIKELIHOOD RATIO TESTS (hLRTs) beginning on top of the page. It shows how one model is selected or discarded based on p-values. Hence, we see that TVM+G model is significantly better than the TVM null model ($p<0.000001$). On the other hand, by adding the additional parameter I, which is the proportion of invariable sites, the likelihood does not improve significantly:

```
Equal rates among sites
  Null model = TVM                          -lnL0 = 5938.5615
  Alternative model = TVM+G               -lnL1 = 5709.6323
  2(lnL1-lnL0) =  457.8584                    df = 1
  Using mixed chi-square distribution
  P-value = <0.000001
No Invariable sites
  Null model = TVM+G                        -lnL0 = 5709.6323
  Alternative model = TVM+I+G             -lnL1 = 5709.6323
  2(lnL1-lnL0) =    0.0000                    df = 1
  Using mixed chi-square distribution
  P-value = >0.999999
```

Accordingly, we would select the TVM+G as a model for this data set. By studying the parameter estimates (below) from this batch operation, we may get a better understanding of the characteristics of the TVM model: Firstly, notice that the nucleotide frequencies are unequal. Next, we see that the rate matrix has identical (high) substitution rates for AG and CT mutations, i.e. transitions. This is why it makes sense to call TVM a transitional model with unequal base frequencies.

```
Model selected: TVM+G
 -lnL  =       5709.6323
  K    =       8
  Base frequencies:
    freqA =     0.3581
    freqC =     0.3186
    freqG =     0.0846
    freqT =     0.2387
  Substitution model:
    Rate matrix
    R(a) [A-C] =        3.9989
    R(b) [A-G] =       40.5786
    R(c) [A-T] =        3.4119
    R(d) [C-G] =        2.3908
    R(e) [C-T] =       40.5786
    R(f) [G-T] =        1.0000
  Among-site rate variation
    Proportion of invariable sites = 0
    Variable sites (G)
     Gamma distribution shape parameter =      0.3752
```

One could choose to use this model with its parameter estimates directly to search for Maximum Likelihood trees. Notice how this is expressed in the Lset command to PAUP* with parameter values:

```
BEGIN PAUP;
Lset Base=(0.3581 0.3186 0.0846) Nst=6 Rmat=(3.9989 40.5786 3.4119 2.3908 40.5786) Rates=gamma
Shape=0.3752 Pinvar=0;
END;
```

Nst=6 means six types of substitutions and Rmat defines the rates of those, scaled to the rate GT, which is set to 1. The options Rates and Shape define the rates as unequal over the sites, - they are gamma distributed with an alpha parameter value of 0.37. Do you remember if this alpha value implies little or much rate variability over the sequence sites?

Recall from Exercise 7 that we could also define the model in a more general way and reestimate parameter values, either directly or by iterative optimization of each parameter. We can use rate classes in PAUP* to make sure that the new estimates of rates comply with the TVM model. In the expression below, we see that the CT rate har been set equal to the AG rate Rate(b):

```
Lset Base=est Nst=6 Rmat=est rclass=(a b c d b e) Rates=gamma Shape=est Pinvar=0;
```

You could now repeat what you did in Exercise 7, but this time by using the TVM+G model rather than HKY. First, optimize parameter values, then do a ML tree search with these values fixed. This would be similar to the Lset syntax in the results from Modeltest, only now with your new parameter values instead:

```
Lset Base=(freqA freqC freqG freqT) Nst=6 Rmat=(rAC rAG rAT rCG rCT) Rates=gamma Shape=alphavalue
Pinvar=0;
```

Alternative to hLRT: AIC

Using the Akaike Information Criterion with this data set, will suggest that the Tamura-Nei93 model with unequal base frequencies and gamma correction is the best model. It may be confusing to face the obligation to choose between model types when different criteria give diverging results. One should bear in mind that AIC gives penalties for increasing numbers of parameters.

* MODEL SELECTION UNCERTAINTY : Akaike Weights

Model	-lnL	K	AIC	delta	weight
TrN+G	5710.5513	6	11433.1025	0.0000	0.2463
HKY+G	5711.9385	5	11433.8770	0.7744	0.1672
TIM+G	5710.4355	7	11434.8711	1.7686	0.1017
TrN+I+G	5710.5513	7	11435.1025	2.0000	0.0906
TVM+G	5709.6323	8	11435.2646	2.1621	0.0836
K81uf+G	5711.8125	6	11435.6250	2.5225	0.0698
GTR+G	5708.9224	9	11435.8447	2.7422	0.0625
HKY+I+G	5711.9385	6	11435.8770	2.7744	0.0615
TIM+I+G	5710.4360	8	11436.8721	3.7695	0.0374
TVM+I+G	5709.6323	9	11437.2646	4.1621	0.0307
K81uf+I+G	5711.8125	7	11437.6250	4.5225	0.0257