Exercise 7: Working with likelihoods

(BIO332 Phylogeny 2007. E.Willassen) PDF version

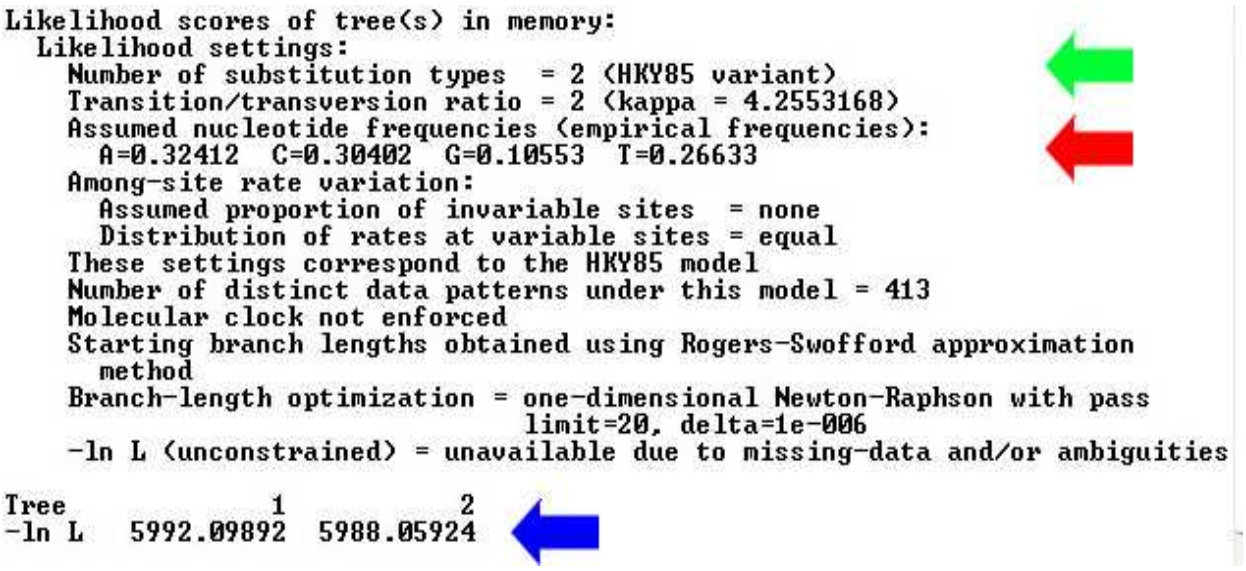
Computing the likelihood of trees

If you already finished Exercise 6, you have a file called ex5_p_tre.tre containing two equally parsimonious trees resulting from the datafile Exerise5.nex. Now, reopen this data file and execute it in PAUP*. Execute the command basefreq. You will see that PAUP* computes base frequencies for each sequence and finally a mean value for the frequency of nucleotides A, C, G, T. This is what we call the empirical base frequencies.

Then, load your treefile into memory by writing: gettrees file=ex5_p_tre.tre. Notice the response from PAUP* in the log window.

```
Processing TREES block from file "ex5_p_tre.tre":
   Keeping: trees from file (replacing trees in memory)
   2 trees read from file
   Time used = 0.00 sec
```

PAUP* has a command that computes the likelihood of trees in memory. Now, execute this command: lscore. There are several points to notice in the screen printout. The first thing to notice is that PAUP* automatically uses the HKY85 model if you didn't specify otherwise (green arrow). You see that that the assumed frequencies of nucleotides are the empirical frequencies, that is they have been computed directly from the data (red arrow).



The lscore command finally gives you the log likelihoods for the trees and we see that the tree with the best likelihood is tree number 2 (blue arrow).

Finally, you should notice that the ti/tv ratio has been set to 2, that all the sites in the data are assumed to be variable, and that the rates of change are assumed to be equal over all the sites. These are important additional assumptions of the model which may affect the outcome of the likelihood calculation.

Optimizing a likelihood model

Switch from parsimony to likelihood by by executing this command: set crit=l;. When setting a likelihood model, we usually do that in order to estimate ML trees from the data. The problem is that it may be computationally very expensive to optimize all parameter values of the model and ML trees simultaneously. It may be better to optimize at least some of the parameters in advance of the ML tree search. We can do that from a "reasonably good" tree, for example like we did above with the trees from the maximum parsimony search.

Find the file called ML_models_in_MB3b4_and_PAUP.txt and open it in Notepad. Find out how the HKY(85) model is specified in PAUP*:

lset basefreq=estimate nst=2 rates=equal tratio=estimate;

When the HKY is written like this, we may estimate all the parameter values of the model. However, in the example above, we actually used fixed parameter values for the base frequencies and the transision/transversion ratio. We would obtain the above default setting by writing HKY as follows:

lset basefreq=emp nst=2 rates=equal tratio=2; You probably gather that this specifices empirical base frequencies (rather than estimated frequencies) and that the tr/tv ratio is set to 2.

Now, check out what happens with the likelihoods of the trees in memory if you estimate the parameter values instead of using preset values. Type lset basefreq=estimate nst=2 rates=equal tratio=estimate; and then lscore;

```
Likelihood scores of tree(s) in memory:
  Likelihood settings:
    Number of substitution types   = 2 (HKY85 variant)
    Transition/transversion ratio estimated via ML
    Nucleotide frequencies estimated via ML
    Among-site rate variation:
      Assumed proportion of invariable sites   = none
      Distribution of rates at variable sites = equal
    These settings correspond to the HKY85 model
    Number of distinct data patterns under this model = 413
    Molecular clock not enforced
    Starting branch lengths obtained using Rogers-Swofford approximation
      method
    Branch-length optimization = one-dimensional Newton-Raphson with pass
                                 limit=20, delta=1e-006
    -ln L (unconstrained) = unavailable due to missing-data and/or ambiguities

Tree                    1           2
--------------------------------------
-ln L          5986.46731  5981.72023
Base frequencies:
  A              0.315806    0.313924
  C              0.292611    0.292607
  G              0.103578    0.103941
  T              0.288005    0.289528
Ti/tv:
  exp. ratio     2.375563    2.405404
  kappa          4.944714    4.986246

Time used to compute likelihoods = 0.77 sec
```

You see that the tree likelihoods have improved, the estimated base frequencies are slightly different from the empirical ones, and the ts/tv ratio was estimated to be slightly higher than 2. Say we wanted to use these parameter estimates for base frequencies and the tr/tv ratio in a search for a ML tree. Then write: lset basefreq=prev nst=2 rates=equal tratio=prev;. "Prev" stands for "previous" and refers to the values in memory. With the parameter values set, we can proceed by executing a simple heuristic search under the ML criterion. Recall from Exercise 6 how we can specify differnt ways to do the search, but for now, simply write: hs;

When the search has finished, write: describetree/ plot=ph; The option behind the slash plots the tree as a phylogram.

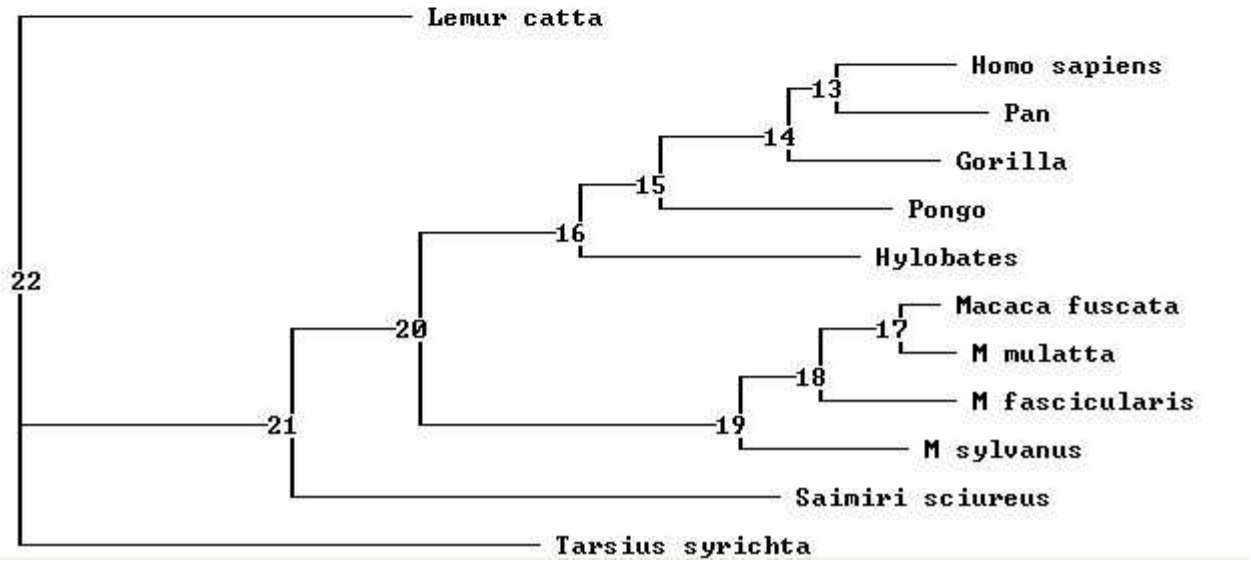
```
Heuristic search completed
   Total number of rearrangements tried = 546
   Score of best tree(s) found = 5981.72023
   Number of trees retained = 1
   Time used = 2.45 sec

Tree description:

  Unrooted tree(s) rooted using outgroup method
  Optimality criterion = likelihood
    Likelihood settings:
      Number of substitution types   = 2 (HKY85 variant)
      Transition/transversion ratio = 2.4054 (kappa = 4.9862458)
      Assumed nucleotide frequencies (set by user):
        A=0.31392  C=0.29261  G=0.10394  T=0.28953
      Among-site rate variation:
        Assumed proportion of invariable sites   = none
        Distribution of rates at variable sites = equal
      These settings correspond to the HKY85 model
      Number of distinct data patterns under this model = 413
      Molecular clock not enforced
      Starting branch lengths obtained using Rogers-Swofford approximation
        method
      Branch-length optimization = one-dimensional Newton-Raphson with pass
                                   limit=20, delta=1e-006
      -ln L (unconstrained) = unavailable due to missing-data and/or
                              ambiguities

Tree number 1 (rooted using default outgroup)

-Ln likelihood = 5981.72023
```



We save the tree like we did in the case of parsimony (Exercise 5). However, since this time we calculated, not a cladogram, but a phylogram, we also want to save the branch length information. Make sure you understand how to do that: savetrees file= ex5_ml_tre.tre brlens=yes;

We have now used the ML criterion to search for the best tree. However, the HKY85 model may not be the best model for this data set, and we should explore the possibility that other models may fit the data better. One way of doing that is to use the program ModelTest (Exercise 8).