Integrative Biology 200A University of California, Berkeley
"PRINCIPLES OF PHYLOGENETICS" Spring 2012
Updated by Michael Landis

Lab 9: Maximum Likelihood and Modeltest

In this lab we're going to use PAUP* to find a phylogeny using molecular data and Maximum Likelihood as the optimality criterion. The computer evaluates the likelihood of each tree, including topology and branch lengths, one at a time. It calculates the probability of each base pair changing in such a way as to generate the states observed at the tips of the branches based on the tree and a set of parameters describing how the bases change with time. The likelihood of a data set for a given tree is the product of these probabilities for all the base pairs. The computer chooses the topology and branch lengths that produce the highest likelihood for the data set. So what parameters of nucleotide change do we use and what values do we give them? This is called the model of nucleotide change and today we will pick a model using ModelTest.

There are an infinite number of possible models. Many have been implemented in various programs, many have been suggested and never implemented, and even more have never been conceived. Today we are only going to deal with a few models that are implemented in PAUP* and evaluated by ModelTest.

A model is considered nested within another model if its parameters are a limited set of the parameters in the other model. For example the Jukes-Cantor model, which assumes that every nucleotide has the same rate of change to any other, is nested within the Kimura two-parameter model, which assumes different transition and transversion rates. A model without any invariant sites would be nested within one with some percentage of invariant sites. Any two models are not necessarily nested.

Adding parameters to a model always increases the maximum likelihood of the data. However, if a model has too many parameters, then maximum likelihood becomes unreliable. Therefore to accept a new parameter into your model it must produce a significant increase in the likelihood. How do you tell if a difference in likelihood is significant? Well, I'm sure you'll be shocked to learn that there is a formula. It is called the Likelihood Ratio Test (LRT). For a given model with likelihood, L1, nested within another model with likelihood, L2, with n less parameters:

$$C\ 2\ (\text{chi squared}) = 2\ * (\ln(L2) - \ln(L1)) \text{ with n degrees of freedom.}$$

You can use this equation to pick the most inclusive model that cannot be significantly improved on. The only drawback of this equation is that you cannot use it to compare different trees, because different trees are not different models – they are more like alternative parameter values. Therefore, you have to compare the different models on a single tree, and which tree to compare them on may not be obvious. Luckily, you tend to get similar results as long as you use a reasonable tree.

Models of Nucleotide Change

The Transition Matrix

The transition matrix (not as in transition/transversion) is a matrix showing the instantaneous stochastic rate of change between any two nucleotides. It can be used to calculate the chance of one nucleotide changing into another on a branch with a given length. The most unrestrained matrix would look like this:

	A	C	G	T
A	-a-b-g	a	b	g
C	d	-d-e-z	e	z
G	h	q	-h-q-i	i
T	k	l	m	-k-l-m

As you can see, the diagonals are all negative as each nucleotide will be changing away from itself at any instant, so that each row adds up to 0. Furthermore, the average rate of change of all the off diagonals is normalized to 1, so that you can eliminate another parameter for a total of 11 parameters.

On the other hand the Kimura two parameter model would look like this:

	A	C	G	T
A	-a-2b	b	a	b
C	b	-a-2b	b	a
G	a	b	-a-2b	b
T	b	a	b	-a-2b

Here there are two parameters, transition and transversion rate, which can be reduced to just one by normalizing the matrix.

Most programs, PAUP* included, can only calculate matrices with reversible models. This means that change has an equal probability of happening in either direction on a branch. Thus trees can be evaluated as unrooted networks, making the computationally-intensive likelihood calculations much easier. If you used an irreversible model then you could assign a root without the use of an outgroup, although I don't know how reliable an estimate that would be. For a model to be reversible it must be true that:

pX RX>Y=pY RY>X

where RX>Y is the instantaneous rate of change from nucleotide X to nucleotide Y, and pX is the equilibrium frequency of nucleotide X. The equilibrium frequency is the frequency of that nucleotide if the substitution process is allowed to run forever, and can be considered another parameter. Thus any model in which RX>Y=pY rXY, will be reversible. So the General Time Reversible (GTR) matrix looks like:

	A	C	G	T
A	-	pC rAC	pG rAG	pT rAT
C	pA rAC	-	pG rCG	pT rCT
G	pA rAG	pC rCG	-	pT rGT
T	pA rAT	pC rCT	pG rGT	-

with the diagonal filled in appropriately. The sum of the equilibrium frequencies for all four bases must equal one, so that there are three equilibrium frequency parameters. Furthermore, one of the rate parameters can be eliminated by normalizing the matrix, leaving eight parameters total.

Some special cases of the GTR that are commonly used (and that might be familiar from last week's lab on distance methods) are:

- JC : Jukes and Cantor (1969) - All nucleotide substitutions are equal and all base frequencies are equal. This is the most restricted (=specific) model of substitution because it assumes all changes are equal.
- F81 : Felsenstein (1981) - All nucleotide substitutions are equal, base frequencies allowed to vary.
- K2P : Kimura two-parameter model, Kimura (1980) - Two nucleotide substitutions types are allowed, those between transitions and transversions. Base frequencies are assumed equal.
- HKY85: Hasegawa-Kishino-Yano (1985) - Two nucleotide substitutions types are allowed, those between transitions and transversions. Base frequencies are allowed to vary.

Proportion of Invariable Sites (I)

This is a model that assumes some proportion of the sites, pi, can not change. Thus it makes two calculations for each base pair. First it calculates the chance, li, that that base pair would have the observed distribution that it does if it could not change. This will be 1, if it is the same in all taxa, or 0, if there are any differences among the taxa. It then calculates the probability, lv, that it would have the observed distribution if it could change, using the transition matrix and the tree. Then it calculates the overall likelihood for that base as:

$$l = pi\ li + (1\text{-}pi)\ lv$$

Among-site rate variation (Γ, also abbreviated G by ModelTest)

Under the null hypothesis, all sites are assumed to have equal rates of substitution. One way of relaxing this assumption is to allow the rates at different sites to be drawn from a gamma distribution (with the mean value across all sites within a class, such as A-T, represented in the substitution matrix). The gamma distribution is used because the

shape of the curve (α = shape parameter) changes dramatically depending on the parameter values of the distribution.

This calculation is done essentially the same way as it is for invariable sites. The likelihood is calculated for each value of the gamma distribution for each base pair and added together. In practice this is only done for a few values of the gamma distribution, as there are an infinite number of possible values for the gamma distribution and each likelihood calculation is computationally burdensome. This serves as a good approximation of a true gamma distribution.

Choosing a Model Using ModelTest

ModelTest is an extension for PAUP* by Posada and Crandall, which is freely available at http://darwin.uvigo.es/software/jmodeltest.html. It uses PAUP* to calculate the likelihoods of several different models. The Modeltest program chooses among the models using two different criteria. The first is the LRT that we discussed above. The other is the Akaike Information Criterion (AIC), which makes slightly different calculations to compare the models, but the principle of comparison is basically the same. Each criterion produces a different model choice, although they often agree.

1. Download the Nexus file of Cephalopod COI genes that we've been using from http://ib.berkeley.edu/courses/ib200a/cephalopod_COI_Clustalw.nex or the sylalbus page, whichever is easier. Save it to a folder that you make on your desktop. Copy the folder Applications>IB200 >Modeltest3.7 folder into this folder.
2. Rename "Modeltest3.7 folder" to "Modeltest3.7".
3. Open PAUP*.
4. Execute the sequence file in PAUP*:
   Execute cephalopod_COI.nex ;
5. Execute the Modeltest PAUP block:
   Execute Modeltest3.7/paupblock/modelblockPAUPb10 ;
6. PAUP* will now run, while it evaluates the different models. This may take a few minutes. When it is done it will stop running and say that it is completed.
7. Now, go to the desktop, and open yourfoldername>Modeltest3.7 folder>paupblock. You will find a file **model.scores** in the paupblock folder. Rename this file, but make sure it still ends in .scores then copy it, and paste it into yourfoldername>Modeltest3.7>bin
8. In order to run Modeltest, we will need to use the command prompt. Go to the Start menu and choose All Programs>Accessories>Command Prompt
9. Once the command prompt is open, you need to move into the bin directory where you just pasted your .scores file. There are a couple of ways to do this, but one of the easiest is to type "cd" then open the Modeltest3.7 folder and drag-and-drop the bin folder onto the command prompt window. This is almost perfect, unfortunately you need to delete the "" marks that Windows inserts before you press return.
10. Ok, now we are ready to go. Type the following at the command line:

modeltest3.7 < filename.scores > anotherfilename.modeltest
You should replace filename with whatever name you used above and anotherfilename with any other name you'd like. Press return. Once the data is processed and a new prompt (C: … >) appears, you can close the command prompt.

11. A new output file with the name you gave it will have appeared in your bin folder.
12. Open this file in a text editor. (Select program from a list>Notepad) Now, we will get to see which models Modeltest suggests based on several different model-testing criteria. Both are ways to pick the most inclusive model that can not be significantly improved on. Scroll down until you see the banner indicating the results from the Hierarchical Likelihood Ratio Tests:

```
-----------------------------------------------------------------
*                                                               *
*       HIERARCHICAL LIKELIHOD RATIO TESTS (hLRTs)              *
*                                                               *
-----------------------------------------------------------------
```

Then keep going until you see:

Model selected:
followed by a type of model. Make a note of which model Modeltest chose under this criterion. If you get the same answer as me, the model will be

Model selected: GTR+I+G
Which stands for a General Time Reversible model with a certain proportion of Invariant sites and a Gamma distribution of changes.

13. Keep scrolling down until you see the results under the Akaike Information Criterion:

```
-----------------------------------------------------------------
*                                                               *
*          AKAIKE INFORMATION CRITERION (AIC)                   *
*                                                               *
-----------------------------------------------------------------
```

Right under it you should see

Model selected:
followed by a type of model. What model did Modeltest choose under each criterion? Are they the same? Go ahead a peruse some of the other statistics in the file, and see what other models Modeltest looked at and rejected.

Finding a Maximum Likelihood Tree in PAUP*

Fixed Parameter Values

First let's use the parameter values chosen by Modeltest.

1. In the Modeltest output file you will find a PAUP block that can be inserted directly into the Nexus file. It starts
   BEGIN PAUP;
   and ends with
   END;
   This block changes the Likelihood Settings (Lset), by setting the base frequencies at equilibrium (Base), the number of substitution types (Nst), the rate matrix of instantaneous substitution rates (Rmat), the among site rate variation (Rates), the shape of the gamma distribution (Shape), and the proportion of invariant sites (Pinvar).
2. Copy the PAUP block from the text file. Edit your Nexus file in PAUP* (or a text editor) and paste the PAUP block from Modeltest directly into it. It can go after any END; statement.
3. Execute the newly-edited sequence file in PAUP* again.
4. Set the search criterion to maximum likelihood:
   set criterion = likelihood;
5. Then do a heuristic search
   hs;

Fit the Parameter Values Along with Finding the Tree

It is also possible for PAUP* to search for the parameter values at the same time as it searches for the best tree using the model - but not the parameter values- chosen by Modeltest. The following Lset command will require PAUP to estimate the base frequencies, rate matrix, shape of the gamma distribution, and proportion of invariant sites:

Lset Base=estimate Nst=6 Rmat=estimate Rates=gamma Shape=estimate Pinvar=estimate;

If you ran this, are you still waiting? Yeah, this takes way too long. Just stop it. Why do you think it takes so much longer? If you did let it finish would the best tree have a higher or lower likelihood than with the fixed parameter values? What are the advantages of each method?

If you try to estimate all of these at once, it will take an extremely long time (it is even possible, with this much uncertainty, that the search will never converge and go on forever.) If you want to try this, it is probably better to try just one or two parameters at a time. It will still probably take a while. Likelihood searches are pretty slow.