Likelihood and Modeltest Lab

In this lab we're going to use *PAUP** to find a phylogeny for a group of genes using 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 nucleotides 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, Λ_1, nested within another model with likelihood, Λ_2, with n less parameters:

$$X^2 \text{ (chi squared)} = 2 * (ln(\Lambda_2) - ln(\Lambda_1)) \quad \text{with n degrees of freedom.}$$

You can use this equation to pick the most inclusive model that can not be significantly improved on. The only drawback of this equation is that you can not 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	$-\alpha-\beta-\gamma$	α	β	γ
C	δ	$-\delta-\varepsilon-\zeta$	ε	ζ
G	η	θ	$-\eta-\theta-\iota$	ι
T	κ	λ	μ	$-\kappa-\lambda-\mu$

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	$-\alpha-2\beta$	β	α	β
C	β	$-\alpha-2\beta$	β	α
G	α	β	$-\alpha-2\beta$	β
T	β	α	β	$-\alpha-2\beta$

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, which greatly eases the calculations. If you used an unreversible 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:

$$\pi_X R_{X>Y} = \pi_Y R_{Y>X}$$

, where $R_{X>Y}$ is the instantaneous rate of change from nucleotide X to nucleotide Y, and π_X 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 $R_{X>Y}=\pi_Y r_{XY}$, will be reversible. So the General Time Reversible (GTR) matrix looks like:

	A	C	G	T
A	–	$\pi_C r_{AC}$	$\pi_G r_{AG}$	$\pi_T r_{AT}$
C	$\pi_A r_{AC}$	–	$\pi_G r_{CG}$	$\pi_T r_{CT}$
G	$\pi_A r_{AG}$	$\pi_C r_{CG}$	–	$\pi_T r_{GT}$
T	$\pi_A r_{AT}$	$\pi_C r_{CT}$	$\pi_G r_{GT}$	–

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 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, p_i, can not change. Thus it makes two calculations for each base pair. First it calculates the chance, λ_i, 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, λ_v, 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:

$$\lambda = p_i \lambda_i + (1\text{-}p_i) \lambda_v$$

Among-site rate variation (Γ)

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/modeltest.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. Save it to a folder that you make on your desktop. Copy the folder Applications>IB200 >Modeltest3.7 folder into this folder.

2) Open *PAUP**.

3) Execute your sequence file in *PAUP**.

4) Execute the file Modeltest3.7 folder >paupblock >**modelblockPAUPb10** in *PAUP**.

*PAUP** will now run, while it evaluates the different models. When it is done it will stop running and say that it is completed. You will find a file **model.scores** in the paupblock folder.

5) Rename this file with a .scores suffix still.

6) Run the program Modeltest3.7 folder >bin >**Modeltest3.7.macX**

7) Click the button next to File for the input file (on the left) and select your '.scores' file.

8) Repeat the process for the output file (on the right). Navigate to a place where you want to send this file to and name the file. Select it and hit OK.

You will now find the output file where you told *Modeltest* to save it

5) Open this file in a text editor. What model did it choose under each criterion? Are they the same? What do all these other statistics mean? (Maybe you should ask me about that one.)

Finding a Maximum Likelihood Tree in *PAUP**

Fixed Parameter Values

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

1) Open your sequence file in *PAUP** again.

2) Pull down the **Analysis** menu and select **Likelihood**.

3) Pull down the **Analysis** menu and select **Likelihood Settings**.

If you got the same results as me *Modeltest* chose a GTR+I+Γ model with certain values for the parameters. For simplicity let's use the LRT model for all are settings, although the literature probably shows a bias to the AIC..

4) Select the GTR model from the **Substitution Model** page.

5) Click the button next to **Set to** in the GTR box. Set the values to those that *Modeltest* selected.

6) Select the **Base Frequencies** page from the page pull down menu.

7) Click **Set to** and type in the values that *Modeltest* found.

8) Select **Among-site Rate Variation** from the page pull down menu.

9) Select **Proportion of invariable sites: Set to:** and type in the number from *Modeltest.*

10) Select **Gamma distribution** and then **Set to:.** Type in the number from *Modeltest*.

11) Hit **OK**

12) Pull down the **Analysis** menu and select **Heuristic Search**. Hit **Search**.

This may take a while. When it's done hit **Close** and check out your trees, make sure to look at a phylogram, so that you can see the branch lengths.

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*.

1) Pull down the **Analysis** menu and select **Likelihood Settings**.

2) Select the GTR model from the **Substitution Model** page.

3) Click the button next to **Estimate** in the GTR box.

4) Select the **Base Frequencies** page from the pull down menu.

5) Click **Estimate**.

6) Select **Among-site Rate Variation** from the page pull down menu.

7) Select **Proportion of invariable sites: Estimate.**

8) Select **Gamma distribution** and then **Estimate.**

9) Hit **OK**

10) Pull down the **Analysis** menu and select **Heuristic Search**. Hit **Search**.

This will take even longer, so wait a second.

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?

A Really Big Shortcut

So I kind of cheated you guys to make you see how to set the model using *PAUP**. There is a much quicker way to set the likelihood model and parameter values chosen by *Modeltest*. 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**;. Just copy this from the text file. **Edit** your Nexus file in *PAUP** and paste the PAUP block from *Modeltest* directly into it. Now **Execute** the nexus file. Make sure that *PAUP** read the PAUP block by looking at the **Likelihood Settings**. Now do the **Heuristic search**.