Filogeografía

Posgrado en Ciencias – Biología, Universidad del Valle
Profesores Crawford, Cerón y Cárdenas
22-27 de junio, 2009

Lab 4
Selecting model of DNA sequence evolution using *jModeltest* 0.1.1
23 June 2009

The goal for today's lab is to select a model of DNA sequence evolution which we will use in the next lab to infer a phylogenetic tree using the criterion of maximum likelihood (ML). To infer a ML phylogeny, we require both the data and a model. For DNA sequence data, we need a model of molecular evolution. Our goal is to choose a model sufficiently complex to describe adequately the data in hand, yet not so complex that it assumes more models that the data can adequately support. We require a computer software application that will try to find those parameter values (not only the parameters θ_i of our evolutionary model, but also the topology τ_i and branch lengths v_i) that maximize the quantity Prob.(X | τ_i, v_i, θ_i), where X = the data. Calculating the parameter values is relatively rapid, as we shall see. Finding the optimal tree or trees, however, can be a very slow process depending on the number of samples in your data set. For example, when searching for the ML tree (or trees) using *PAUP**, a data set of 20 samples may take minutes, while 100 samples may take days, depending on the computer power available. Students should be aware that there are many other software packages available to calculate ML trees, most of which are free. An updated and extensive listing of software programs for phylogenetics and population genetics is posted at http://evolution.genetics.washington.edu/phylip/software.html.

To choose an appropriate model of evolution, we will use *jModeltest* version 0.1.1, *por el famoso* Profesor David Posada. You may apply this lab to your data, GenBank data, or we can provide you with some data on marine gastropods, if you like.

Software

jModeltest software is available here:

`http://darwin.uvigo.es/software/jmodeltest.html`

Citations

Posada D (2008) jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* **25**, 1253-1256.

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution* **14**, 685-695.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**, 696-704.

Additional references:

Burnham KP, Anderson DR (2001) Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research* **28**, 111-120.

Burnham KP, Anderson DR (2004) Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research* 33: 261-304. *(A detailed comparison of model selection methods, applicable to any field, not just phylogenetics.)*

Huelsenbeck JP, Larget B, Alfaro ME (2004) Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular Biology and Evolution* **21**, 1123-1133.

Posada D and Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of the AIC and Bayesian approaches over likelihood ratio tests. *Systematic Biology* 53: 793-808. *(For a richer understanding of the issues of model selection, David Posada suggests we read this paper.)*

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*. 17: 368-376. *(First viable implementation of ML phylogenetics, along with a proposal to use a likelihood ratio test to evaluate the molecular clock hypothesis.)*

I. Keep the documentation handy

For every analysis, I recommend first locating and opening the documentation. Inside the folder "jModelTest 0.1 package" look for another folder called "doc". Inside that folder, locate and open a file called "jModelTest.0.1.1.pdf" and keep this document handy at all times. You can add files or folders here, but do not change the positions or names of any files or folder than came with the software! The Java program will be looking for certain files in certain folders with certain names, so be careful.

II. Preparing your infile

Make a new folder for today's lab. To help you keep your data and output files organized, I recommend that for each individual analysis you make a new and appropriately labeled folder. Locate your best, most complete NEXUS file containing all your samples, including outgroups (when you have them), and all your genes, and place a copy of this in your new folder. When choosing the best evolutionary model using *Modeltest*, I recommend that you do **not** include the outgroup. Today we have only mitochondrial genes, but when you have both mitochondrial and nuclear genes, I recommend that when using *PAUP** you analyze the two genomes separately for ML analysis. If you have multiple mitochondrial genes, I recommend that your **first** analyze a single data set that includes all genes. If at the end of the lab you find that your data reject the molecular clock hypotheses, you may want to repeat the analysis on individual genes, using *Modeltest* to choose the best model for individual genes. (For your final ML phylogenetic reconstruction, hopefully you will infer one tree based on all available data. Be aware, however, that in the current version of *PAUP** you can only apply one model to all the data.)

You can define within your NEXUS data file a `taxset` called `outgroup`, that contains only your outgroup sequences. (You may have any number of `taxset` commands.) For example:

```
begin SETS;
       taxset outgroup = bicho_X bicho_Y mata_Z ;
END;
```

However, for today's lab, we have no outgroups. If you have an outgroup, you could now find (or add) a `PAUP` block, and add the command "`exclude outgroup /only;`" e.g.:

```
begin PAUP;
       exclude outgroup /only ;
END;
```

This way, we would never forget to remove the outgroup before running the *jModeltest* analysis.

To make today's lab analysis run faster, you should only analyze only 12-20 DNA sequences. If you need to prune your data, delete either a complete clade, or a wide sample of sequences from across the phylogeny. If many students are using the marine gastropod data (ask me), each group should select different subsets of the data so we can compare results at the end of the lab session today. To analyze only a sub-sample of the data, you may either add an `include` or an `exclude` command to a `PAUP` block, or simply toss out sequences and then export the remaining sequences to a new file.

III. Running *jModeltest 0.1.1*.

With the old version of *Modeltest*, one had to first run *PAUP**, then use *Modeltest* to analyze the output. The newer *jModeltest* now does both for us. It uses the free (open source) phylogenetics program, *PhyML* (Guindon & Gascuel 2003) to calculate likelihoods, and if the user chooses, calculate BioNJ (Gascuel 1997) trees or do ML tree searches.

To launch the program, click on the java application called "jModelTest.jar" in the folder, "jModelTest 0.1 package". Your data can be in a different folder from the application (unless you run into problems, then you might consider moving your data file). Select File > Load DNA alignment and navigate to your data file and click on it. Select Analysis > Compute likelihood scores. Now decide how many models of which type you would like to test. Number of substitution schemes: the number of different DNA subst. models to evaluate. If you plan on running *MrBayes*, you should select 3, since MrBayes only implements 3 substitution models: JC, K2P/HKY, and GTR. Today we will evaluate 7, because that's how many *PAUP** has pre-set (try to enumerate them!), though any reversible model of nucleotide substitution can be implemented.

Base frequencies: whether nucleotide frequencies are free to vary, or are set at 0.25 each. Select this.
Rate variation: Add "I" and/or "G" rate heterogeneity parameters. Select both. (nCat 4 is fine).
Base tree for likelihood calculations: Two options allow the tree to vary with the model (bottom 2). ML will take too long, perhaps, so we'll try BioNJ (**not** "Fixed BIONJ-JC," which is was the old *Modeltest* used).

jModeltest needs to now be calculate log-likelihood scores for a BioNJ tree under each model. *jModeltest* passes the work to the program, *PhyML*. This may take 5-10 minutes for the "DirtFrog" data (14 samples). The output files should appear in the folder exe > phyml, as well as on the screen. We can save the screen output by selecting Edit > Save console, which is never a bad idea. You might look at this output and consider these questions:

¿Do more complex models have higher or lower log likelihood scores, or either one?

¿Are more complex models always better?

Once we have the likelihood scores calculated, we can apply our model selection criterion of choice on these data. First, let's apply the AIC. Choose the menu Analysis > Do AIC calculations... Select "Use AICc correction" (Burnham & Anderson 2004). Also click on "Write PAUP* block". Confidence interval can be left at 100% or lowered to 95% or even 50%. It shouldn't matter too much. (Be sure that "Do model averaging" is selected. I also recommend selecting "Calculate parameter importance".)

The best model according to the AICc will be found under this heading:
CORRECTED AKAIKE INFORMATION CRITERION (AICc)
along with associated parameter values. Below that, you will find text much like this:

```
[!
Likelihood settings from best-fit model (TIM2+G) selected by AICc
with jModeltest 0.1.1 on Tue Jun 23 03:18:54 GMT-00:05 2009]

BEGIN PAUP;
Lset base=(0.2890 0.3056 0.1145 0.2909) nst=6  rmat=(2.8068 28.0066 2.8068 1.0000 17.7482
       1.0000) rates=gamma shape=0.2830 ncat=4 pinvar=0;
END;
```

Copy this information and save it in a new file called, e.g., "PAUP_cmds_AICc_modelaverage.txt" BUT now edit this file by replacing each parameter value above with the values displayed under this heading:
* AICc MODEL SELECTION : Model averaged estimates
Be sure to get the parameter values in the right places! The PAUP command block is read as follows:

```
Lset                                  = Settings for the likelihood analysis.
Base=(0.2810 0.3028 0.1184)           = Frequency of bases (A C G), you get T by subtraction.
Nst=6                                 = Assume a 6-parameter rate matrix.
Rmat=(3.8047 30.0543 3.1745 1.0656 18.3693)  = rates = rAC rAG rAT rCG rCT
                                      ¿Which rates are higher, the transversions or the transitions?
Rates=gamma  Shape=0.3704             = Assume Γ distribution of rates among sites with α = 0.3704
Pinvar=0;                             = No "I" parameter (proportion of invariant sites = 0)
```

So which model selection criterion should you use? According to my reading of Burnham & Anderson (2004) I suggest the following. AICc is preferable to regular AIC. For phylogenetic model selection problems, AICc is likely to perform better than BIC. Note that the choice of criterion is *not* a matter of frequentist vs. Bayesian statistics (Burnham & Anderson 2004). You can use your output model and the attending parameter values for our likelihood tree search, but I suggest using the MODEL-AVERAGED parameter values, as stated above.

Compare the "best" models selected under the AICc: how many models were contained in the set that included at least 95% of the cumulative Akaike weights? Look at the table entitled:
* AICc MODEL SELECTION : Selection uncertainty
How do these best models differ, i.e., what parameters do they have in common or not in common?

If you have time, try some more options, such as the hLRT or dLRT (only work when the topology is **fixed**), the BIC or the DT criterion. For more information on each criterion, see the manual, "jModelTest.0.1.1.pdf".

Questions for consideration/discussion:

(1) Why do we use *jModeltest* or similar program? Why not just *a priori* adopt the simplest (JC) or the most complex model usable in *PAUP** (*i.e.,* GTR+Γ+I)?

(2) Looking at your output of likelihood scores: you may have noticed that it is possible for a model with fewer parameters to have *lower* log-likelihood support than a model with more parameters. How could this be if (in lecture we said that?) more complex models "always" give higher support? Give an example (real or hypothetical).

(3) Why do you suppose that Burnham & Anderson (2004) recommend the model-averaged parameter values?

A few additional references:

Alfaro ME, Huelsenbeck JP (2006) Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Systematic Biology* **55**, 89–96. *(Compares AIC to a Bayes method not in* Modeltest*.)*

Minin V, Abdo Z, Joyce P, Sullivan J (2003) Performance-based selection of likelihood models for phylogeny estimation. *Systematic Biology* **52**, 674–683. *(An alternative to* Modeltest, *but still uses* PAUP**.)*

Pol D (2004) Empirical problems of the hierarchical likelihood ratio test for model selection. *Systematic Biology* **53**, 949-962.

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**, 406-425. *(Citation for the neighbor-joining method.)*

Sullivan J, Holsinger KE, Simon C (1996) The effect of topology on estimates of among-site rate variation. *Journal of Molecular Evolution* **42**, 308–312. *(Regarding the effect of topology on parameter estimation.)*

Additional references for models, parameters, and methods can be found in the *Modeltest* documentation.