Filogeografía

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

Lab 5
*PAUP** 4beta10 and a test of the molecular clock hypothesis.
24 June 2009

The goal for today's lab is to infer a phylogenetic tree using the criterion of maximum likelihood (ML), then infer the same tree again, with a molecular clock enforced. Finally, we will compare the results using a χ^2 test. In this test, we will assume that the clock hypothesis is the null hypothesis and try to reject it. The "clock" hypothesis means that rates of evolution are the same on all branches, thus the tree is "ultrametric." Note that an ultrametric tree is **not** the same as a cladogram.

Software

*PAUP** used to be freely available, but unfortunately we now are asked to pay ~US$100. PC, Mac OS9, and unix versions are sold by Sinauer Assoc. A new version of *PAUP** is currently in pre-beta stage.

Citations

Swofford DL (1998) *PAUP**. *Phylogenetic Analysis Using Parsimony *and Other Methods*, Version 4b10. Sinauer Associates, Sunderland, Massachusetts.

Maddison DR, Swofford DL, Maddison WP (1997) NEXUS: An extensible file format for systematic information. *Systematic Biology* **46**, 590-621.

Additional references:

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.)*

O'Meara, Brian (2006) guide to preparing NEXUS batch files and running the basic analyses using *PAUP**. I highly recommend this webpage: `http://www.brianomeara.info/phylogenetics.html`

Overview of today's lab:

1) Use your preferred model to infer the maximum likelihood tree for the data, *with* outgroup.
2) Calculate the likelihood (-Ln) of **the same tree** with a molecular clock enforced, using *PAUP**.
3) Calculate a LRT of the molecular clock hypothesis (using a calculator and χ^2 table).

I. Keep the documentation handy

For every analysis, I recommend first locating and opening the documentation. Inside the PAUP folder look for another folder called "PAUP Docs folder," or something like that. Inside that folder, locate and open a file called "PAUP_Cmd_ref_v2.pdf" and keep this document handy at all times. I still open this file **every time** I need to do anything in *PAUP**. If you are analyzing your own data, you may also want to run your data through the recommended instructions of Brian O'Meara [URL above].)

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. For running the molecular clock test, you have the option to include or exclude the outgroup. If you are reasonably confident of the outgroup, and it is not so genetically distant that we have problems in estimated the branch lengths between it and the ingroup, then I would include the outgroups in the test. On the other

hand, if the outgroup seems somehow "problematic" perhaps you can justify excluding it from the clock test. If you are using the "DirtFrog" example data, first try this lab **with** the outgroup.

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.) To edit your PAUP file, you can use *WORDPAD*, or even *PAUP** itself. Even better would be *BBEdit* (costs money) or *TextWrangler* (Mac only, I believe).

For example:

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

If you have an outgroup and you want to exclude it, 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 analysis.

To make today's lab analysis run faster, you should only analyze only 12-20 DNA sequences, such as are contained the "DirtFrog" example file. If you need to prune your data, delete either a complete clade, or a wide sample of sequences from across the phylogeny. 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.

Find the command line from the *jModeltest* lab. Be sure the parameter values reflected the MODEL-AVERAGED values that you edited by hand in the previous lab. Copy the command block into your NEXUS data file. It will be similar to this:

```
BEGIN PAUP;
Lset  Base=(0.2810 0.3028 0.1184)  Nst=6  Rmat=(3.8047 30.0543 3.1745 1.0656
18.3693)  Rates=gamma  Shape=0.3704  Pinvar=0;
END;
```

Paste the text at the bottom of your NEXUS file. You can do this directly in *PAUP** or in a text editor such as *WordPad* or *TextWrangler*.

A. Type into your NEXUS (data) file another block of text, but change the names of the log file and output trees as you see fit:

```
begin PAUP;
log start file=YOUR_dataNAME_MLsearch_log.txt ;
undelete / ClearTrees=Yes;
Outgroup outgroup /only ;
Set criterion=likelihood ;
Set MaxTrees=100 increase=no ;
HSEARCH  start=NJ  swap=TBR  Dstatus=240 ;
END;

begin PAUP;
     set TOrder=Right;
     savetrees file=YOUR_dataNAME_ML.tre brlens=yes;
     DescribeTrees;
savetrees file=YOUR_dataNAME_no_brlens.tre brlens=no;
log stop ;
END;
```

II. Running *PAUP**

SAVE the file under the FILE menu in *PAUP** (or other text editor). Hit EXECUTE. Now your ML search should be running. Hopefully this will be done some time today!

Alternatively, you could limit the **length of time** of the search. Normally, this is a bad idea, but for today's lab you might want to, just so you can see some output before class is over. Also, for today's lab we really just need the log-likelihood score, therefore a complete search may not be necessary for our purposes. If you choose to impose a

time limit on your search, you can do this by placing the following text after the `HSEARCH` command (but before the "`;`"): `TimeLimit=3600` Time is in seconds, so this option would limit your search to one hour.

III. Calculating the likelihood of the same ML tree but with clock enforced!

A. Save your NEXUS file again, with a new name such as "`YOUR_dataNAME_CLOCK_search.nex`". In your new NEXUS file, in the 2nd `PAUP` block, add a line that says "`constraints fixMLtopol = `"

Open your file "`YOUR_dataNAME_no_brlens.tre`" in a text editor and copy the tree that is there in NEWICK format, like this: `= tree PAUP_1 = [&U] (1,(((2,(3,4)),(5,6)),(7,(((((8,9),10),11),12),(13,14)))));`

Copy this "tree" and paste it into your NEXUS file after you new `constraints` command, like this:

```
constraints fixMLtopol = [&U] (1,(((2,(3,4)),(5,6)),(7,(((((8,9),10),11),
        12),(13,14)))));
```

Be sure this line ends with a semi-colon "`;`". See below for a completed example.

B. More edits to make to your file:

At the end of your `LSET` command add the text: " `clock=YES` "

In your `LSET` command change all fixed parameter values to "`estimate`", but do not change the model (you may *not* need all parameters shown in the example below. For example, I added the "`RClass`" option to enforce the Tamura-Nei model, but you may not want this (see PAUP manual for more info).

At the end of your `HSEARCH` command add the text: "`constraints=fixMLtopol enforce=YES;`"

You *may* have to add a line that says "`RootTrees OutRoot=Monophyl UserBrLens=Yes ;`" But this sometimes doesn't work! We'll see what happens. (Clock enforced trees need to be rooted!)

Modify the name of your `log` file like this: "`log start file=dataNAME_CLOCK_search.out ;`"

Modify the name of your `.tre` file to be like this: "`savetrees file=dataNAME_CLOCK.tre`"

So, your final new `PAUP` block might look something like this:

```
begin PAUP;
        log start file=dataNAME_CLOCK_search_log.txt ;
        Outgroup outgroup /only ;
        constraints fixMLtopol = [&U]
        (1,(((2,(3,4)),(5,6)),(7,(((((8,9),10),11),12),(13,14)))));
        Set criterion=likelihood ;
        Lset  Base=estimate  Nst=6  Rmat= estimate  RClass=(a b a a c a)
              Rates=gamma  Shape=estimate Pinvar=estimate clock=YES;
        Set MaxTrees=100 increase=no ;
        HSEARCH  start=NJ  swap=TBR  Dstatus=240 constraints=fixMLtopol
             enforce=YES ;
        END;
begin PAUP;
     set TOrder=Right;
     savetrees file=dataNAME_CLOCK.tre brlens=yes;
     DescribeTrees;
log stop ;
END;
```

And cross your fingers that this works! It should take only minutes, because we are **not** searching for topologies, only optimizing model parameter values (including branch lengths). If it works, then calculate the likelihood ratio test (as discussed in class?). Log-likelihood scores (ln*L*) can be found in your "`.tre`" files or in your "`log.txt`" files. The formula is:

$-2 \times \ln (L_0 / L_A) = -2 \times (\ln L_0 - \ln L_A)$ which is approximately distributed as a χ^2 with *n*-2 degrees of freedom, where ***n* = number of samples** (i.e., external branches or DNA sequences)**, L_0 = likelihood (null model | data)**, L_A **= likelihood (alternative model | data)**, and of course **$\ln L_0$ & $\ln L_A$ are their respective log-likelihoods.** *PAUP** calculates and reports ln*L*. Recall that our null model assumes constant rates of evolution among branches (molecular clock), and our alternative model allows variable rates.

Consult an on-line table of χ^2 to determine if your data significantly reject the null hypothesis of a molecular clock or use **R** to calculate the probability exactly. Report the negative log-likelihood (–Ln) for your unconstrained and constrained (clock enforced) trees. Which values is higher, *i.e.*, which score represents the greater level of support? Report your χ^2 value, degrees of freedom and resulting *p* value. Did you reject the molecular clock hypothesis?