Laboratory excercise written by Andrew J. Crawford <andrew@dna.ac>
with the support of CIES Fulbright Program and Fulbright Colombia.
`http://filogeografia.dna.ac`
Enjoy!

Filogeografía: genética evolutiva espacial.

BIOL 4211, Universidad de los Andes, Bogotá. Coord.: -074.0657, 04.6012
25 de enero a 01 de abril 2006

Lab 9a
The SH test of tree topology using *PAUP** 4b10

25 marzo 2006

NB: Always download the latest version of all software
and read carefully the latest accompanying documentation.

The goal of this first lab today is to perform a conservative but easily implemented test of topology to your data set. For this task, you must have one or more *a priori* hypotheses of relationships, hopefully derived from questions about the history of your species. You must also decide on the constraint topology predicted by each hypothesis.

You will use *PAUP** to perform a likelihood-based non-parametric test of tree topology, the paired-sites test (aka, SH test) of Shimodaira and Hasegawa (1999). This test uses bootstrap resampling and corrects critical values for multiple comparisons. This test is known to be conservative (Buckley 2002; Shimodaira 2002). For a given topological test, you will constrain only the node/s in question and then perform a new ML search, using the same non-partitioned ML model of sequence evolution you selected based on the *Modeltest* exercise. The significance of the difference in the sum of site-wise log-likelihoods for all trees is evaluated by bootstrap sampling of site scores (RELL sampling) with 1000 replicates (Kishino and Hasegawa, 1989) and then calculating how far the observed differences are from the mean of the bootstrap replicates. The accuracy of this test is increased with the inclusion of all reasonable trees, in addition to the ML tree (H_1) and the constrained tree (H_0), while the power is reduced with the inclusion of improbable trees (Shimodaira and Hasegawa 1999). Therefore, you could included, for example, a set of maximum parsimony trees as well, calculated their likelihoods under the same model of evolution, and then obtain the one-tailed *p*-values for all trees including your null hypotheses of relationships. However, to save time, we will just compare the unconstrained (H_1) and one or more constrained (H_0) ML trees, as is often done in published studies.

Students should be aware that one can also use *PAUP** to perform other topological tests, such as the parsimony-based *z*- or *t*-test. For very detailed instructions on performing a parameter bootstrap test (*e.g.*, SOWH test) see the webpage: `http://dna.ac/genetics.html`

References for Topology Tests

Buckley TR (2002) Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Systematic Biology* **51**, 509-523.

Felsenstein J (2004) *Inferring Phylogenies*. Sinauer Associates Inc., Sunderland, Massachusetts, USA.

Goldman N, Anderson JP, Rodrigo AG (2000) Likelihood-based tests of topologies in phylogenetics. *Systematic Biology* **49**, 652-670. *(Compares SH, SOWH, KH, and other ML-based tests of topology.)*

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution* **29**, 170-179.

Shi X, Gu H, Susko E, Field C (2005) The comparison of the confidence regions in phylogeny. *Molecular Biology and Evolution* **22**, 2285–2296. *(Compares various tests and recommends two new tests.)*

Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection *Systematic Biology* **51**, 492-508.

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* **16**, 1114-1116.

Software: availability and assistance with analyses

Download the latest updaters for *MacClade* here (the latest version is 4.08):
`http://macclade.org/index.html`
An excellent forum for keeping up with the latest developments in phylogenetics:
`http://www.yphy.org/phycom/`
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`

Citations for software.

WP Maddison, DR Maddison (1992) MacClade 4: Analysis of Phylogeny and Character Evolution, Version 4.0. Sinauer Associates Inc., Sunderland, Massachusetts, USA.
Swofford DL (1998) *PAUP*. Phylogenetic Analysis Using Parsimony *and Other Methods*, Version 4b10. Sinauer Associates, Sunderland, Massachusetts.

Overview of today's lab:

1) Decide on one or more *a priori* hypotheses of relationships.
2) Create a constraint tree by writing a Newick-formatted tree by hand, or using *MacClade*.
3) Run a ML search under each constraint (each tree is a H_0 tree).
4) Import your ML tree from Lab 7 (this is your H_1 tree), plus additional (optional) MP trees.
5) Run the SH test to see if H_0 (as well as H_1) adequately explains the data, or whether H_0 is rejected.

I. *A priori* hypotheses

First, you need one or more *a priori* hypotheses of topological relationships. I recommend you think carefully about this. Also, present your hypotheses to a critical friend first. Discuss them. I recommend constraining the minimum number of nodes possible under any given *a priori* hypothesis. If you constrain more nodes than absolutely necessary, you will (a) more likely reject your hypothesis, but (b) possibly reject it for the wrong (unintended or uninteresting) reasons. Also, consider alternative outcomes and how you might interpret them. Consider the undesirable but possibly necessary option of removing taxa if they're inclusion precludes clear alternative *a priori* topological hypotheses (constraints) AND they have no bearing the hypotheses under consideration. However, if you delete any taxa you will have to infer a new ML tree (repeating Lab 7). Any test of topology (*e.g.*, SH test) can only compare trees inferred from the exact same set of taxa/samples.

Make a new folder for this lab. If you decide that you cannot meaningfully include all taxa (samples) in your hypothesis testing, create a new trimmed NEXUS file now, before proceeding. Otherwise, directly load your complete NEXUS file into *MacClade*. (You can also manipulate and save trees using *TreeView* for PC but not in *TreeView* for Mac. You can move or delete branches in *TreeEdit* but you cannot collapse nodes.) Go to Windows menu, select "Tree Window". Move branches around with the arrow tool until you have your desired tree topology. Preferably, this tree will represent just a single monophyletic group with a big polytomy outside of this one designated clade. Go to Trees menu and select "Store Tree..." and give it an informative name. For multiple predicted topologies (one per *a priori* hypothesis), make new constraints and store them. Then save the file AND under Trees menu also select "Save Tree File as..." just in case. These saved trees will contain your constrain written in Newick format (see Lab 7).

II. Get ML tree without constraints (the ML tree)

Hopefully, you did not have to remove any taxa to conduct your hypothesis testing, in which case you can just find your ML tree from Lab 7. If you decide to test your hypotheses on a trimmed data set, you will have to re-run *Modeltest* and infer a new ML tree using, *e.g.*, *PAUP** (see Lab 7). Be sure to save the log and .tre files. If you had two or more ML trees, include all of them in the SH test.

III. Get ML tree with constraints (the H_0 tree)

Now you need to find best tree under constraint hypothesis. This search will be at least a little faster than the ML search because the constraint restricts the number of possible topologies considered. We impose constraints the same way we did in Lab 7. To your NEXUS file, you might add a PAUP block somewhat like this:

```
begin PAUP;
CONSTRAINTS Pleist_fragm =
((27,14,6,20,15,17,13,29,1,25,18,28,24,19,12,16),2,3,26,21,9,10,22,11,8,7,23,5,4);
set criterion=likelihood ;
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 ;
HSEARCH  Start=NJ  swap=TBR  enforce=YES Constraints=Pleist_fragm ;
savetrees file=Pleist_fragm_ML.tre brlens=yes ;
END;
```

Actually, it would probably make more sense to run *Modeltest* on your constrained topology, estimate the necessary parameter values, and then apply this model to the ML search. However, it's not clear which method would be correct, but certainly any difference would be miniscule. Such considerations may be more important for the SOWH test than the SH test. And for smaller data sets remember that your best option is to estimate parameter values while searching for the best tree, using an `lset` command such as this:

```
Lset  BaseFreq=estimate  Nst=6  Rmat=estimate  Rates=gamma
     Shape=estimate  Pinvar=estimate ;
```

Obtain the H_0 (constraint) tree for all hypotheses (predictions) of interest. Then proceed to the next step. While this analysis is running, you may want to start Lab 9b.

IV. Conducting the SH test in *PAUP**

You need to have your data executed in *PAUP** and have in *PAUP**'s memory all tress you want to analyze. These trees will include the ML tree (H_1), all trees obtained under constraints, plus any extra optimal trees, e.g., the set of maximum parsimony trees. You will be comparing all trees at once. Remember that the Get Trees from File command can be executed only after the data file is executed. The default setting in *PAUP** is to overwrite (erase) the trees already in memory when importing new trees. Therefore, if *PAUP** has trees already in memory and you want to import more trees using the command Trees > Get Trees from File..., then be sure to click on Options... and click on the overlapping circles such that all parts of both circles are darkened. Also, you will need to know which tree is which. *PAUP** numbers the tree in the order in which they are loaded. If you have only two trees to be compared, do not worry, the ML tree will be labeled "best."

The command line argument would be:

```
GetTrees File=YourTreeFilenName.tre StoreBrLens=Yes WarnTree=Yes Mode=7 ;
```

"Mode=7" should correspond to combining all trees, but this command seems buggy to me. After importing all your trees, you should confirm that *PAUP** contains all the trees you intended to import. If the `GetTrees` command keeps erasing your trees, you will have to cut-and-paste all the Newick trees into one single text file (ending with .tre) and then run `GetTrees` on this new file.

Find the SH window in *PAUP** OS9 under the menu, Trees > Tree Scores > likelihood...

Click on the button labeled Likelihood Settings... and confirm that your desired model and parameter settings are represented there.

Click on the button labeled Topology Tests (KH,SH)...

Check the box labeled Shimodaira-Hasegawa test

Under Test distribution click the radio button next to RELL.

Under Number of bootstrap replicates for RELL and full optimization, you can leave the default values of 1000.

Hit OK. Now you are back at the previous window. Check any of the boxes you want to.

Hit OK, and the analysis should run for a few seconds or minutes. The output provides a table showing the –lnL score of all trees (according to your model), the difference in –lnL between each tree and the best tree, and the associated *P*-value for probability that a given tree can explain the data. All trees with $P < 0.05$ are rejected.

The *PAUP** command line argument for the above SH analysis starts with:

```
lscores all/ SHtest=RELL
```
... and ends with the settings for your evolutionary model that you obtained from *Modeltest* in Lab 7, *e.g.*,
```
lscores all/ SHtest=RELL Nst=6  Rmat=(3.8047 30.0543 3.1745 1.0656 18.3693)  Rates=gamma
       Shape=0.3704  Pinvar=0 ;
```
Copy and save the output to a text file.

REPORT Lab 9a

Instead of a report, you can simply include your finding in your final paper. You should include a paragraph explaining your a *priori* hypotheses and the topological predictions made by each. If your predictions are not immediately clear and obvious to the reader, you should consider illustrating your predictions with cladograms. Report which predictions were (or were not) rejected. Recall that the SH test is very conservative. Any predictions rejected by the SH test would very likely be rejected by all other topological tests as well.