

## **Lab 17: Molecular Clock**

Today we are going to use several different methods of testing the molecular clock and estimating node times. We will use a couple of likelihood ratio tests to test the molecular clock against a totally unconstrained tree and a tree with a few branches allowed to vary independently. We will also use several rate smoothing methods to infer divergence times. We will not deal with several commonly used methods. In particular we will not use any relative rate tests to test the molecular clock. This is a very active field and there are constantly new methods and new programs being developed.

### **Testing for Global Molecular Clock**

Under the null hypothesis, the phylogeny is rooted and the branch lengths are constrained such that all of the tips can be drawn at a single time plane. Under the alternative hypothesis, each branch is allowed to vary independently. The alternative hypothesis invokes  $s - 2$  additional parameters, where  $s$  is the number of sequences. The likelihood ratio test statistic is  $-2\log L = 2(\log L_0 - \log L_1)$ , where  $L_0$  and  $L_1$  are the likelihoods under the null and alternative hypotheses, respectively. The significance of the likelihood ratio test statistic can be approximated using a chi-square distribution (with  $s - 2$  degrees of freedom).

The following example shows how to perform the likelihood ratio test of the molecular clock using PAUP\*. Use the DNA sequence data and tree from the file *Cephalopod.nex*. For speed the Hasegawa, Kishino, and Yano (1985) model of DNA substitution with among site rate variation described using a gamma distribution. Go to the following window: Trees→treescor→likelihood→likelihood settings and estimate model parameters for the Ts:tv ratio and the gamma distribution shape parameter. Record the  $-\ln L$  score. This is the likelihood score for the alternative hypothesis, which allows branches to vary independently.

Root the phylogeny (Trees→root trees), and return to the likelihood settings window and select "miscellaneous". Choose to enforce a molecular clock. Recalculate the likelihood score under this null model

Conduct the likelihood ratio test and determine if you can reject the null model. Use the **Chidist** function in *Excel* to get a p-value. Why is the likelihood score of the alternative model higher than the null model?

### **Testing for Local Molecular Clock**

In the previous example we tested whether the entire tree fit a clock as opposed to every branch on the tree having an independent rate. We could also test whether a clade has a different rate from the rest of the tree. We can not do this in PAUP\*, because PAUP\* does not allow us to specify different rates on different branches. Instead we will use *BASEML* a program from the *PAML* package of phylogeny programs by Ziheng Yang. This program does ML analysis of DNA sequences, and allows us to specify a tree and different distributions of rates on the tree. All these programs can be found at <http://abacus.gene.ucl.ac.uk/software/paml.html>.

This program is entirely controlled by the input files. Open the tree file *CephTree* in the *BaseML* folder with a text editor. As you can see this tree contains the same tree in Newick

format as we used in the previous example. You will also see a '\$1' after the clade containing Joubiniteuthis and Moroteuthis. This specifies that all the branches in this clade will have a different rate than the other branches in the tree. Open the file **BaseML** with a text editor. When *BASEML* is run, it automatically opens the control file, named **BaseML**, in the same folder as it. The first line of the file specifies the file with the DNA sequences. The second line specifies the tree file. There are many other options in this file, but the only one that we are concerned with here is the **clock** option at the bottom of the tree. Here you can specify how the rates on the branches are grouped. 0 allows the rates on all the branches to vary independently; 1 enforces a molecular clock; and 2 enforces separate molecular clocks on each set of specified branches. We'll start with 0.

Open up the terminal Applications>Utilities>Terminal. Type `cd`, space, and drag the `base ml` folder into the terminal and hit enter. Now type `./baseml` to run *BASEML*. It will take second. When it is done, record the likelihood score. Open up the **BaseML** file in a text editor. Change the **clock** setting to 1. Run *BASEML* and record the likelihood score. Repeat the process for a **clock** setting of 2.

Use the likelihood ratio test to compare these models. Which model has the highest likelihoods? Why? Which model is the best?

### Estimating Divergence Times Using r8s

Now we will use r8s (that's a pun pronounced "rates") to estimate divergence times. R8s uses a tree with branch lengths derived from another program, and then tries to estimate the node times by some measure of the rate differences between these branches. It is by Mike Sanderson and is freely available at <http://loco.biosci.arizona.edu/r8s/>. r8s uses normal nexus files as input files but you need to make a few additional commands.. In particular you need to specify the timing of the nodes which we can locate in time. As we learned in lecture, it is in fact very difficult to locate a node in time.

Open the **CEPHr8s** file in a text editor. The tree block is the same as you would find in any nexus file. Branch lengths are included after the colons. Some of the internal nodes are also named, after the closed parentheses, but before the columns. It is necessary to name nodes so that dates can be assigned to them. The r8s block has several commands. **lengths=persite** means that the branch length is in changes per site not total number of changes, and **ultrametric=no** means that the input tree is not ultrametric. **fixage taxon=Clade1 age=150** sets the age of node Clade1 to 150, and **constrain taxon=node2 min\_age=200 max\_age=300** forces node two to be between . These times are measured backwards from the present, so that **min\_age=200** means that this divergence happened at least 200 (million?) years ago. Finally **divtime method=LF** starts the fitting algorithm using the Langley-Fitch method which deduces node times using maximum likelihood of the branch lengths assuming a constant rate of substitution. If **method = NPRS**, the program minimizes the squares of the differences between adjacent branch lengths. Algorithm specifies the numerical search method.

Open the command terminal, type `cd` space, drag the folder labeled r8s into the terminal and hit return. Type `./r8s -v`, hit return and then type **execute CEPHr8s.nex**. This will execute the file and the commands that we inputted in the r8s block. According to the instructions this should be easier to do but I couldn't get it to work the other way. When the program is done running type **showage** to output a table of node ages. Repeat the program using the NPRS method. How do the node estimates for the two methods compare?