**Integrative Biology 200A** _**University of California, Berkeley**_
**"PRINCIPLES OF PHYLOGENETICS"** _**Spring 2006**_

Lab 2: Introduction to PAUP*

Today we will be learning about some of the basic features of **PAUP* (P**hylogenetic **A**nalysis **U**sing **P**arsimony [*and other methods]), a phylogenetics program developed by David Swofford. PAUP* can infer phylogenies using distance, parsimony and likelihood. Today, we will run these types of analyses using a sample mtDNA data set. We will learn how to use more features in PAUP* in later labs.

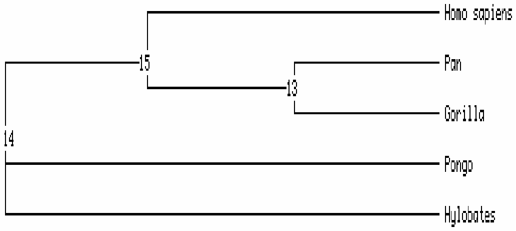
EXERCISE I: Basic PAUP.

Open PAUP*. From the **Sample NEXUS files** folder select the **Homonoid mtDNA** file. Select **Edit** and hit the **Edit** button. This shows you the NEXUS file with the data we will be using for our analysis. You can make changes to the file in this window. Now pull down the **File** menu and select **Execute Hominoid mtDNA.** Now PAUP will be able to analyze your file.

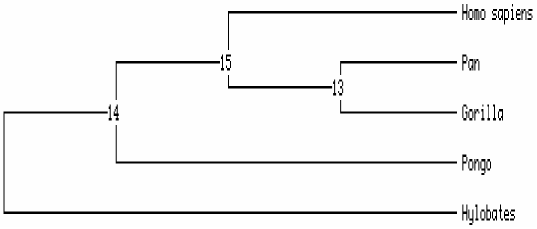
Under the **DATA** menu select **Show character status (Brief summary)**. The output tells you the current optimality criteria (parsimony, likelihood or distance), the number of characters, the character status, coding, number of parsimony informative characters, etc. These are some of the basic summary statistics of your data. Next, select **Show pairwise distances** under the same menu. Continue to run through the various options to see what types of output PAUP will generate with the click of a button.

EXERCISE II: Defining an outgroup and rooting.

Go to the **Data** menu and select **Define Outgroup**. From the list of taxa, select the appropriate outgroup. This will tell PAUP to "root" the tree with this taxon, but the tree can still be drawn in different ways. Go to the **Options** menu and select **Rooting**. This is where you can tell PAUP how to draw the root of your tree, either as a basal polytomy or a separate monophyletic group. You can pick whichever one you prefer.



Basal Polytomy



Ingroup Monophyletic

EXERCISE III: Distance analyses.

PAUP can run multiple types of distance analyses. Go to the **Analysis** menu and select **Distance** to change the current optimality criteria. Next, go to the **Analysis** menu and select **Distance settings.** Select whichever type of distance settings you like and select **OK.** There are lots of options in this menu. You have to read the literature to know what many of them are, but I'll be happy to answer any of your questions. Now that you have changed your distance settings, return to the **Show pairwise distance** command and notice that the output is different from before. Return again to the **Analysis** menu and select **Neighbor Joining/UPGMA.** Selecting neighbor-joining or UPGMA and selecting OK from this screen will run an analysis.

Run both NJ and UPGMA analyses and compare the results. If you're interested, try changing the distance settings and rerun your Neighbor joining analysis. Remember, every time you change the distance settings the pairwise distances will change.

EXERCISE IV: Parsimony analyses.

PAUP uses a variety optimality criteria to determine which tree is best, such as Parsimony or Likelihood. Obviously these different optimality criteria include very different assumptions. When using PAUP* to analyze your data, make sure you are using the optimality criterion that includes the assumptions you want. We'll discuss these assumptions in class.*

First we'll focus on various search strategies (e.g., exhaustive, branch and bound and heuristic). These are methods of finding the best tree under whatever optimality criterion you are using. As we noted in lecture, there is not an exact algorithm to calculate the best tree for a given data set. Thus we need to compare the scores of many trees in order to find the best tree.

***Exhaustive search** – Looks at every possible tree. Too many trees for data sets of more than 12 taxa.*

***Branch and Bound** – Guaranteed to find the shortest tree, but doesn't look at all possible trees. Prohibitively slow with large data sets. Branch and Bound searches save time by ignoring trees that are longer than those that have been previously evaluated.*

***Heuristic Searches** – Not guaranteed to find the shortest tree, but are far less time consuming and our only option for large data sets. Heuristic search methods save time by trying to sample different portions of the possible 'tree space' in hopes of finding something close to the best trees.*

As implemented in PAUP, a heuristic search builds a starting cladogram using **Stepwise Addition**. Stepwise addition starts by building a tree with three taxa, and then adds each additional taxon on the branch that is most parsimonious. Because Stepwise Addition can be mislead by local tree optima, PAUP* then undertakes branch swapping. Branch swapping essentially rearranges the branches on the cladogram to see if it can find any cladograms that are shorter. Different branch swapping options include nearest-neighbor interchange (NNI), subtree pruning-regrafting (SPR), and tree bisection-reconnection (TBR). These methods differ in the details of how they remove branches from the cladogram and reconnect them.*

Change the optimality criteria back to **Parsimony.** Next, we want to conduct an **Exhaustive Search** (remember, this isn't possible with 12 taxa or more). All the settings will be the same as your previous run, if you don't change them. Return to the **Analysis** menu and select **Exhaustive Search**. Click **OK** and when the analysis is done click **Close**. The Exhaustive search output is a frequency distribution of every possible tree. The best tree(s) is the one with the fewest number of steps. To view the shortest tree(s), simply go to the **Trees** menu and select **Describe trees.** Notice all the various summary statistics you can view in addition to just the tree. Pick some output that sound interesting to you. Also, select **Cladogram** and **Phylogram** as output types. If you ever get multiple trees, you can view a consensus tree by selecting **Compute consensus** from the **Trees** menu. Run the same analysis using the Branch and Bound option. Does this speed up the analysis? Why? Next, run a heuristic search. Does this speed up the analysis? Why?

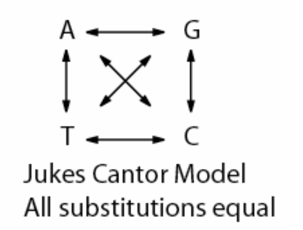
EXERCISE V: Estimating support by bootstrapping.

*You might be interested to know the support for your tree. One measure of support is called the **Bootstrap.** Bootstrapping is a statistical method of resampling the data with replacement. We'll go over this more in class. Bootstrapping provides number on the nodes (0-100%) that correspond to the support. The highest support value is 100, while values below between 50 -70 are usually considered weak. It's important to know that values below 50 aren't shown. In fact, branches below 50 are collapsed and shown as a polytomy.*

Go to the **analysis** menu and select **Bootstrap/jackknife.** Set the **number of replicates** to 100 and the **Type of search** to **Full heuristic**, select **Continue,** then **Search.**

EXERCISE VI: Maximum likelihood analyses.

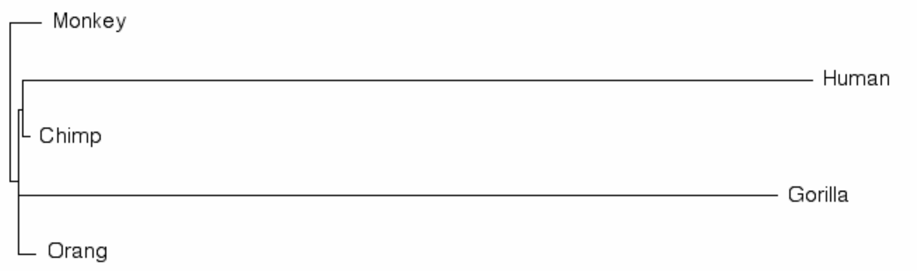
Maximum likelihood (ML) is a statistical method for reconstructing trees. We'll discuss ML in lecture. Basically, ML operates by calculating the following conditional equation: What is the likelihood of observing a data set given a phylogeny and a model of DNA sequence evolution? The tree with the highest likelihood score is considered the best tree. When using maximum likelihood to build trees we have to select a model of DNA sequence evolution. Today we'll use the Jukes Cantor model, which assumes all substitution types are equal.



Set the optimality criteria to **Likelihood.** Next, we have to specify the DNA substitution model. First, go to **likelihood settings** and verify that the **substitution model** is set to **All rates equal ("1 st")** Second, go to **Base frequency** and select **Assume equal frequencies.** After you have made these changes, select **OK.** This is the jukes cantor model. ML analyses are notorious for their slow computational speed. Make sure you run a heuristic search!

EXERCISE VIII: Different methods, different trees.

Are the trees that we have made so far correct? This exercise will demonstrate a situation where different methods get different answers. When two taxa in a data set have long branches (i.e., lots of autapomorphies), and are separated by short internodes, parsimony will tend to group the long branches together with strong support, and the problem gets worse as you ad more data. Maximum likelihood overcomes this problem, but can still fail to get the right tree, if we use an inappropriate model of sequence evolution.



For this exercise, this phylogeny represents the "true" evolutionary tree. Notice the branches for Human and Gorilla are much longer than the other branches, and that the "true" tree groups Humans with Chimps by a very short internode.

Under the **Likelihood settings** set the substitution model to 2ST. This allows for different transition and transversion rates, and thus is more realistic than the model we used before. Run a **Heuristic** search. Go to **Describe trees.** How does the tree look now? Now, conduct a **Bootstrap** analysis under ML – make sure you select **Full heuristic.** Does the tree strongly support any relationships? Why? Do you consider this outcome good or bad? What would happen if you changed the likelihood settings to an even less restrictive model?