Integrative Biology 200A *University of California, Berkeley*
Principals of Phylogenetics *Spring 2008*

Biogeographic Reconstruction on Phylogenies

Today we're going to be looking at programs that compare trees between two associated groups of objects to deduce their common history. This could be a comparison of host & parasite, organism & gene or area & organism trees. The different relationships can be analogized like this (read down):

Host	Organism	Area
Parasite	Gene	Organism
Host switch	Horizontal transfer	Dispersal
Cospeciation	Orthology	Vicariance
Parasite speciation on one host	Gene duplication or allelic divergence	Sympatric speciation (kind of)
Parasite extinction	Gene loss or fixation	Extinction

In all three cases comparisons can be made between the two trees to see how often dispersal or vicariance (or their analogous events) best explains the situation. We are going to try two different programs that use different criteria to determine the relationship between areas and associated organisms.

Treemap

COMPONENT by Rod Page is a good program for analyzing and comparing trees and can do some of these comparisons. However, it can only reconcile the trees. To reconcile trees is to add hypothesized extinct taxa to the organism/parasite tree based on the assumption that there is no dispersal/ host switching. Thus all differences between the trees are a consequence of an ancestral area/host having two organisms/parasites one of which has since gone extinct. The reconciled tree adds hypothesized extinct organisms/parasites.

There is another program *Treemap* from Rod Page's lab. It is only "experimental", and is not widely used, but it looks cool and allows you to do diverse comparisons of the trees. Both programs are available for free. We are going to use *Treemap* to explore comparisons between the host-parasite trees that they provide as examples. We will look at host parasite-data, but the same principles apply for biogeography.

1. Go to http://taxonomy.zoology.gla.ac.uk/rod/treemap.html, and download the program for "A PC running Windows 3.1 or later." Save treemap.zip on the desktop.
2. Extract Treemap by right clicking treemap.zip and choosing "extract all" to make an unzipped folder, also on the desktop.
3. Open Treemap by double-clicking it, then choose open from the File menu. Open the **HAFFNER88.NEX** file.

This is a file with a phylogeny of gophers and their associated lice. This file contains two trees, one for the host and one for the parasite, and a description of which hosts are associated with which parasites. You will see four windows. The first window is the **Tanglegram**, which shows the parasite tree on the right, the host tree on the left and arrows connecting the associated hosts and parasites. This is basically a graphical representation of the data in the input file.

4. Click on the nodes to switch the branches around and try to untangle the intersecting lines. This will not change the topology of the trees or the information that you are looking at, only the appearance. Pull down the **view** menu and select **phylogram**. Then pull down the **view** menu again and select **Internal labels**.

The second window is called the **Reconstruction Window**. This is where the program does its real work. This window can be difficult to read. It shows the two trees overlaid. The parasite tree is black and the host tree is grey with parasites below their associated hosts. Circles at nodes of the parasite tree represent cospeciation and squares represent speciations of a parasite on a single host. Initially it will show the reconciled parasite tree, which assumes no host switching. Thus you will see that for some of the branches, there are two parasites on the branch. This indicates that for this reconstruction there were two species of parasites living on the host at that time.

The third window is the **Branch lengths** window. It shows a graph comparing the distances of branches shared by the host and the parasite. If the molecular clock holds and your reconstruction is correct, then these points should fall on a straight line. Why? As you can see the only two points plotted seam to fit a straight line.

The fourth window is the **Histogram Window**, but you will have to run an analysis for this to show anything. This file does not actually have branch length data, so it is better to look at coalescence times.

5. In the **Branch lengths** window pull down the **View** menu and select **Plot coalescence times**.

The plot will now change to show a comparison of the "age" of the nodes shared by the parasite and the host. These are the nodes where they cospeciated. As you can see, you now have many more points, and, although three of them distinctly fall on a line, the other two do not.

Reconstructions

6. In the **Reconstruction** Window click the square next to the node labeled **13**. This will make the parasite tree change, so that the clade with *cheriei* and *costaricensis* has undergone a host switch. (Although this may be hard to tell, because of crappy graphics. It is experimental.)
7. Look at the coalescence time graph in the **Branch lengths** window again. It has changed to reflect the fact that you removed a cospeciation node, but this does not lead to any improvement. Click the square in the **Reconstruction** window again to return the tree to its default state.

You can try clicking different combinations of parasites and nodes to get a reasonable parasite tree. This sometimes works well, but not always. (Once again, it's experimental).

There is also a reconstruction that was saved with this file.

8. Pull down the menu at the top of the **Reconstruction** window that says **None** and select **Pagel_1990**. How does this one look? What about its coalescence times?

You can also search for the "best" reconstructions. This program defines "best" as the tree with the greatest number of cospeciation events.

9. While in the **Reconstruction** window, go to the **Reconstruction** menu and select **Heuristic Search**. The program will search for the best tree. How do the coalescence times look now?
10. This time do an **Exact Search** for the "best" tree. When you're done, pull down the reconstruction menu labeled **None**.

You will find six trees labeled best. Look through these trees. How do they look? Do any of them have completely consistent coalescence times? What assumptions may be violated that could explain this? Which node seams to be particularly problematic? What might be going on here if none of the assumptions are violated?

Randomization

One way to test if your pattern is significant is to randomize your data, and see how often you get results with as many cospeciations as you got from your actual data. If you rarely get that many cospeciations in the best reconstruction, then your results are probably significant.

11. Pull down the **Randomisation** (they're Scottish) menu and select **Parasite tree.** Type **100** for the number of trees and hit **OK**. This will generate 100 randomizations of the parasite tree and count the maximum number of cospeciations on each one.
12. Go to the **Histogram** window. You will see a distribution of the results from your randomization. How many times did you get as many or more cospeciations than you found in the real data? Is this a significant result? What if you randomize the host tree or both trees?

Adding host/parasite associations

13. Go to the Edit menu. Choose "Associations…" To edit the parasites on a host, click that host, then double-click the parasites to add or subtract them. You can play with this and see how it affects your reconstructions.

Just for fun

14. Close this file and open **HAFFNER94.NEX**
15. Try to rearrange the **Tanglegram** so that it makes since. You can't get it perfect, but you can improve it.
16. Look at the **Branch Lengths** window. This data set has actual branch lengths, and as you can see the graph is a lot messier. Can you improve it?

DIVA

DIVA is a program by Fredrick Ronquist, which is freely available on the web at http://www.ebc.uu.se/systzoo/research/diva/diva.html. (However, I have already downloaded and installed it and you should be able to find it in the 200A folder.) Unlike *Treemap* it is made specifically for biogeography. Furthermore, it does not just maximize the number of cospeciation events, but instead has a cost matrix that describes the cost of all possible events. Should these assumptions be different for a host-parasite as opposed to an area-organism reconstruction? *Diva* does not require a cladogram for the relationships among different areas. It only requires a tree describing the relationship between the different taxa and a description of which areas those taxa are associated with.

17. Go to the folder C:\Program Files\IB 200A\Divawin
18. Open the file **transp.txt** in a text editor. This file describes the relationships and distributions of several species of domestic fruit. The matrix is a description of where the taxa are found. A 0 represents absence from that area and a 1 indicates presence. Following that is a tree describing the relationship of the taxa. Can you read the tree in this format to understand the relationships?
19. Make a list of the each of the geographical areas, starting with South America:

    A. South America

    B.

    C.

    D.

    E.

    F.

    G.

20. Open *DIVA*, either from the folder or the start menu.
21. Type
    ```
    proc transp.txt;
    ```
    and hit **enter**. This will execute the file we were just looking at. Several lines will follow saying that it has opened the file successfully.
22. Now type
    ```
    optimize;
    ```
    and hit **enter**. It will quickly optimize the data to fit the tree, although it would take much longer if you had more taxa or more areas.

To understand the output you must recognize two things. First what is the tree that describes the relationship among the taxa. Each line of output gives the name of only two taxa to describe a node so you must know the tree to understand what other taxa are descended from that node. The second thing you need to know is the order that the areas were listed in, as *DIVA* refers to them only with letters. Thus **A** refers to the first area listed, South America, **B** to Africa, etc.

This should be enough for you to interpret the output. For example the second line of output:

Node 10 (anc. of terminals orange-kiwi):E

means that the common ancestor of oranges, bananas, papayas and kiwis lived in Asia.

Trying Again with BaseML: Testing for a Local Molecular Clock

Last week, we tried to use BaseML, from the PAML package by Ziheng Yang, to test whether a clade has a different rate from the rest of the tree. It didn't go so well. Now, if you would like to try again, we can see if we can get it work. let's try it again and see if we

This program is entirely controlled by the input files. You will need to download these from the web.

1. Go to the syllabus page of the IB 200A website. Download four files:

   CephTree.trees, BaseML.ctl, CephSeq.nuc, BaseML.exe

   Put all the files in the same folder on your desktop.

The first file is *CephTree.trees* – open it 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.ctl* with a text editor. This is the control file for the BaseML program. When BASEML.exe is run, it automatically opens the control file, which must be 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.

2. To run BaseML, double-click on the BaseML
3. When it is done, record the likelihood score. (It will be at the bottom of the screen, after `–lnL =`)
4. Open up the *BaseML.ctl* 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.
5. Use the likelihood ratio test to compare these models. Which model has the highest likelihoods? Why? Which model is the best? *Remember, the likelihood ratio test statistic is -2logL = 2(logL0 - logL1)*