Integrative Biology 200, Spring 2014
Principles of Phylogenetics: Systematics

University of California, Berkeley
by Traci L. Grzymala

Lab 23: Coevolution

Introduction

Today we will explore the program Jane4 (Conow *et al*, 2010) used for cophylogenetic analyses in which two phylogenetic trees for groups are constructed and then the relationships between these trees are explored. Examples of coevolutionary systems are hosts and their parasites, insect-plant relations, or symbiotic relationships. You can also use these methods to explore gene tree-species tree questions or biogeography questions. One common approach to cophylogenetic problems is to use an evolutionary model that describes the set of possible types of events that happened during coevolution and assign costs for the different types of events. The problem is then to find a reconstruction of the common history with the minimal sum of event costs. Algorithms that employ this idea are called event-based methods. With these methods, there are different types of events:

1. Codivergence
2. Duplication
3. Host Switch
4. Loss (extinction, sampling failure, or missing the boat)
5. Failure to Diverge

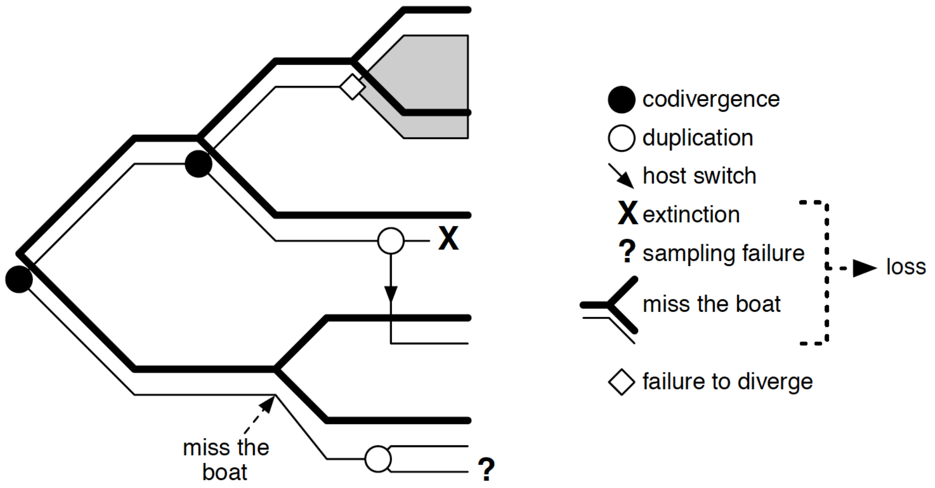


Figure from Charleston & Perkins, 2006

Jane4 uses event-based methods and the lab today is based primarily from the tutorial available here: http://www.cs.hmc.edu/~hadas/jane/tutorial.html

The goals of the lab today are to:

I. Load phylogenetic data for gophers and their parasites into Jane4
II. Perform a reconstruction minimizing costs for events between these two phylogenies
III. Run randomizations for statistical analyses

Exercise 1: Coevolution with Jane

Let's open the program Jane by double clicking on the program icon. We'll be using the GUI version today, but you can also use command line prompts if you wish. You will first see the Main window comprised of three panels. The upper left panel is the "Information" panel and displays basic information about the file that you load into the program. The upper right is the "Actions" panel and contains the buttons for running reconstructions. The bottom is the "Solution" panel.

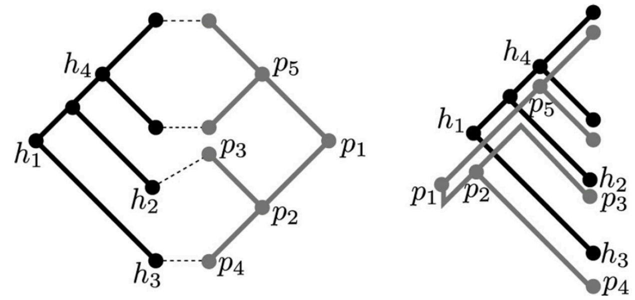


Figure depicting two phylogenies on the left and their reconstruction on the right
(from Merkle et al. 2010)

What type of input data does Jane need?

1. A pair of phylogenetic trees ("host" and "parasite")
2. A mapping between the tips of these trees (*Note: the mapping does NOT need to be 1-to-1. i.e. several parasites can be on one host)

Download the example data set "gopher_louse.tree". This is a now classic data set consisting of a pocket gopher phylogeny and a phylogeny of their parasitic lice (Hafner & Nadler, 1988). Place this file into a folder for today and load this file into Jane

File → Launch Tree Editor.

Select the Folder in the top left. Find the file named "gopher_louse.tree". You will now see the phylogeny for the gophers on the left and the phylogeny for the lice on the right. Also open this file in a text editor to explore the formatting. There are several blocks with information following (similar to what we've seen in Nexus format, PAUP blocks, or MrBayes blocks).

***Question #1:* What does the 'PHI' block indicate? This may not be intuitive at first, so take a minute to look at the file. Ask if this doesn't make sense.**

Explore the buttons at the top of the page to see what your options are in this window. Feel free to manipulate the phylogeny as you'd like – don't worry, you won't break anything! When in the Time Zone mode, click and drag the dotted red lines to move them around. See what options are in the Region mode.

***Question #2:* What are the Time Zone modes and the Region mode buttons? Why do you think each of these would be helpful to include in your analysis?**

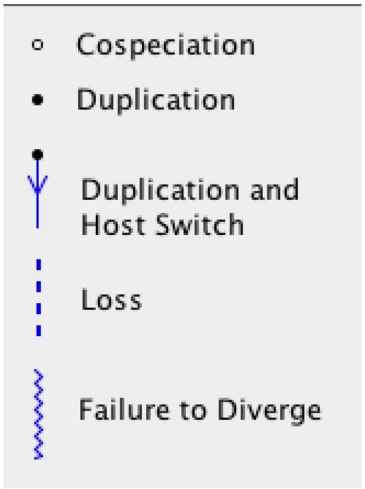
You can either change some of the options or leave them as they are for our example today. Select the “Load to Solver” button.   Settings → Set Costs

You can set the costs of each type of event: Cospeciation, duplication, duplication with host switching, loss, and failure to diverge. Jane requires event costs to be integers, so if you want higher precision of your costs, you simulate this by scaling up the costs (e.g. costs of “100” and “101” are analogous to “1.00” and “1.01”). At this point you might be asking ‘how does one determine the cost values’. This is, of course, the million-dollar question. Often cospeciation costs are considered to be small, and duplication and host switch costs are usually assumed to be high. However, from a biological point of view, the exact values for these costs are basically unknown. Because of this Jane also supports solving for ranges of costs. For example, you can specify that cospeciations cost 0, duplications cost between 1 and 3, duplications with host switch cost between 3 and 5, etc. There is also a relatively new program CoRe-Pa (Merkle *et al*, 2010) that uses a parameter-adaptive approach where you do not have to assign costs prior to the reconstruction. We will not cover this program today, but the reference is given at the end of the lab in case you want to explore this option later.

For now let’s just keep the default settings. Select “OK”. Then in the Actions panel, select “Go”.

A Solutions table will be displayed. You will notice that many of the solutions Jane finds will appear to be identical. While they look identical, they are based on different “timings” or relative orderings of the speciation events in the host tree. To compress these seemingly identical solutions so that only truly distinct numbers of events are presented, check the “Compress Isomorphic Solutions” button. Okay, go back to the list and double click on the first solution. A visualization of the solution should pop up in a new window.

Each type of event is indicated by a specific symbol as shown below:



To access this menu for yourself, select “Options” → “Show Key”

The colors indicate the existence of other possible locations for the association. A green node means there is a location of lower cost where the parasite node and its descendants may be mapped. A yellow node indicates that there is another location of equal cost, and a red node means that all other locations it may be mapped to are of higher cost.

Question #3: **How many of each of the different events are shown to have occurred? What is the total cost of the reconstruction?**

Exercise 2: Statistical Analyses

Now that we have generated a reconstruction with a specific cost associated with it, we want to generate samples of random parasite trees or tip mappings to obtain their cost. These costs can then be used to perform statistical analyses. To get started, select “Stats Mode” in the middle of the main screen.

In the “Statistical Parameters” box on the right, increase the sample size to 10,000. We will include the original problem instance and use Random Tip Mapping so leave the other settings as they are. Select Go from the “Actions” panel. This will take a minute or so to run. [Note: If this takes more than 5 minutes, stop the run and reset the sample size to be lower 5,000 or less]

When the run is finished, you should have a histogram on the bottom left. The horizontal axis represents the cost of the sample and the vertical axis represents the number of samples with the corresponding cost. The original problem instance will appear as a red dashed line. You can save this image with the “Save as Picture” option. You can also use this information to calculate a p-value for your original reconstruction score.

Question #4: **Save a .png picture of your histogram and send that to me. What is your computed p-value? How would you interpret this value?**

References

Charleston MA, Perkins SL (2006) Traversing the tangle: Algorithms and applications for cophylogenetic studies. *Journal of Biomedical Informatics*, 39: 62–71.

Conow C, Fielder D, Ovadia Y, and Libeskind-Hadas R. (2010) Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, 5:16.

Hafner MS, Nadler SA (1988) Phylogenetic trees support the coevolution of parasites and their hosts. *Nature*, 332: 258–259.

Merkle D, Middendorf M, Wieseke N (2010) A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinformatics*, 11 (Suppl 1): S60.