Counting Mutations on a Tree

Alan R. Rogers

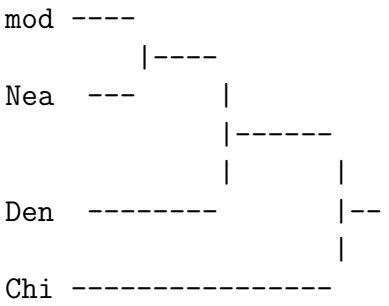
October 24, 2011

This is supplementary info, intended to clarify the homework problem about allocating mutations on a tree. We expected students to attack this problem with arithmetic rather than algebra. Nonetheless, some of you may find an algebraic approach helpful.

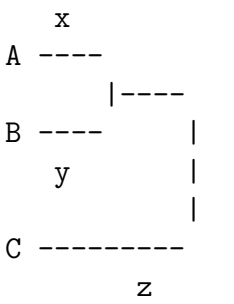
Here is a fragment of the table of pairwise differences from the homework:

Species	mod	Nea	Den	chi
mod	—	168	330	1304
Nea		—	318	1289
Den			—	1309

Your task is to allocate mutations from this table onto the following tree:



Let us first consider a smaller and more abstract problem involving three arbitrary species, A, B, and C, which have the following phylogenetic tree:



Here, x counts mutations along the branch leading to A, y counts those along the branch to B, and z counts the number along the *entire* branch leading to C from the common ancestor of A and B. In some alignment, the numbers of nucleotide differences between the three pairs of species are D_{AB}, D_{AC}, and D_{BC}.

Under the infinite sites model, the number of nucleotide site differences between two species must equal to the number of mutations along the two branches that separate them. This implies that

$$\begin{aligned} D_{AB} &= x + y \\ D_{AC} &= x + z \\ D_{BC} &= y + z \end{aligned}$$

These equations holds exactly under the model of infinite sites but are approximations under other models of mutation.

This is a system of three equations in three unknowns. Solving it gives

$$x = \Big(D_{AB} + D_{AC} - D_{BC}\Big)\Big/2 \quad (1)$$

$$y = \Big(D_{AB} - D_{AC} + D_{BC}\Big)\Big/2 \quad (2)$$

For example, let us take A, B, and C as "mod," "Nea," and "Den." Then (using the table above) these formulas become

$$\begin{aligned} x &= \Big(168 + 330 - 318\Big)\Big/2 = 90 \\ y &= \Big(168 - 330 + 318\Big)\Big/2 = 78 \end{aligned}$$

In other words, the model allocates 90 mutations to the branch leading to modern humans and 78 to the one going to Neanderthals.

We can do exactly the same exercise with any group of three species. For example, moderns, Denisovans and chimpanzees:

$$\begin{aligned} x &= \Big(330 + 1304 - 1309\Big)\Big/2 = 162.5 \\ y &= \Big(330 - 1304 + 1309\Big)\Big/2 = 167.5 \end{aligned}$$

Had the data been in perfect agreement with the infinite sites model, these numbers should be integers. As you can see, they are not. The result allocates 167.5 mutations to the branch leading to Denisovans and 162.5 to the branch leading to moderns. This latter branch has two parts, separated by the common ancestor of humans and Neanderthals. We already know that 90 mutations occurred on the more recent portion of this branch. The number on the more ancient portion must therefore be $167.5 - 90 = 72.5$.

You can use apply this method repeatedly to allocate mutations throughout the larger tree in the homework assignment. The only remaining ambiguity will involve the branch separating gorillas from the other species. The method above cannot help with that problem, because there is no outgroup.