Inference of ancestral locations

Richard Neher

(Dated: August 19, 2023)

A common phylogeographic model is diffusion. Given the sampling locations r_i of the tips i of the tree, the ancestral locations r_n of internal nodes of the tree have a likelihood function

$$\log L = -\sum_{n\neq root} \frac{(r_n - r_p)^2}{4Dt_n} \tag{1}$$

where r_p is the location of the parent of n and t_n is the length of the branch leading to node n. The ancestral locations are not observed and to calculate the overall likelihood need to be integrated over.

This can be done recursively, where the likelihood of a subtree n given the position of node r_n is given by

$$\begin{aligned}
L_n(r_n) &= \prod_{c\in n} \frac{1}{4\pi D t_c} \int d\,r_c e^{-\frac{(r_c-r_n)^2}{4Dt_c}} P(r_c) \\
&= \prod_{c\in n} \frac{\sqrt{d_c}}{\sqrt{\pi}} \int d\,r_c e^{-d_c r_n^2 + 2d_c r_n r_c - d_c r_c^2 - a_c r_c^2 + 2b_c r_c - c_c} \\
&= \prod_{c\in n} \frac{\sqrt{d_c}}{\sqrt{\pi}} \int d\,r_c e^{-d_c r_n^2 + 2r_c(b_c + r_n d_c) - r_c^2(d_c + a_c) - c_c} \\
&= \prod_{c\in n} \frac{\sqrt{d_c}}{\sqrt{\pi}} \int d\,r_c e^{-d_c r_n^2 - (d_c+a_c)(r_c^2 - 2r_c \frac{b_c + d_c r_n}{d_c + a_c} + \frac{(b_c+d_c r_n)^2}{(d_c+a_c)^2}) + \frac{(b_c+d_c r_n)^2}{d_c+a_c} - c_c} \\
&= \prod_{c\in n} \frac{\sqrt{d_c}}{\sqrt{\pi}} \int d\,r_c e^{-d_c r_n^2 + (d_c+a_c)(r_c - \frac{b_c+d_c r_n}{d_c+a_c})^2 + \frac{(b_c+d_c r_n)^2}{d_c+a_c} - c_c} \\
&= \frac{\sqrt{d_c}}{\sqrt{a_c + d_c}} e^{-d_c r_n^2 + (b_c^2 + 2r_n b_c d_c + d_c^2 r_n^2)/(d_c+a_c) - c_c} \\
&= \frac{\sqrt{d_c}}{\sqrt{a_c + d_c}} e^{-d_c\left(1 - \frac{d_c}{a_c+d_c}\right) r_n^2 + 2\frac{b_c d_c}{d_c+a_c} r_n - c_c + \frac{b_c^2}{d_c+a_c}}
\end{aligned} \tag{2}$$

This allows calculation of the parameters a_n, b_n, and c_n of node n from the children that are not terminal nodes as.

$$a_n = \sum_{c\in n} d_c \left(1 - \frac{d_c}{a_c + d_c}\right) = \sum_{c\in n} \frac{d_c a_c}{a_c + d_c} \tag{3}$$

$$b_n = \sum_{c\in n} \frac{b_c d_c}{a_c + d_c} \tag{4}$$

$$c_n = \sum_{c\in n} c_c + \frac{b_c^2}{d_c + a_c} + \frac{\log(d_c) - \log(a_c + d_c)}{2} \tag{5}$$

If a child is a terminal node, the terms in the sum need to be replaced by

$$a_n = \sum_{c\in n} d_c \tag{6}$$

$$b_n = \sum_{c\in n} d_c r_c \tag{7}$$

$$c_n = \sum_{c \in n} d_c r_c^2 - \log(2\pi/d_c)/2 \tag{8}$$

Note that for a single child, the variances add ($a_n^{-1} = a_c^{-1} + d_c^{-1}$) and the most likely positions don't change ($b_n/a_n = b_c/a_c$).

The same propagation can be used up the tree

$$a'_n = \frac{d_p a'_p}{a'_p + d_c} + \sum_{c \in p, c \neq n} \frac{d_c a_c}{a_c + d_c} \tag{9}$$

$$b'_n = \frac{b'_p d_p}{a'_p + d_p} + \sum_{c \in p, c \neq n} \frac{b_c d_c}{a_c + d_c} \tag{10}$$

$$c'_n = c'_p + + \frac{{b'_p}^2}{d_p + a'_p} + \frac{\log(d_p) - \log(a'_p + d_p)}{2} + \sum_{c \in p, c \neq n} c_c + \frac{b_c^2}{d_c + a_c} + \frac{\log(d_c) - \log(a_c + d_c)}{2} \tag{11}$$