Contents

1 Background

Lewis (2001) discussed the need to correct likelihood calculations of discrete morphological traits for ascertainment biases that exclude the possibility of a character being constant. The model that corrects for this ascertainment bias is referred to as the Mk_v model. The correction entails calculating: $\Pr(X \mid T, \text{each character is variable})$ rather than $\Pr(X \mid T)$, where X is the data matrix and T is the tree.

$$\Pr(X \mid T, \text{each character is variable}) = \prod_{i=1}^{M} \Pr(x_i \mid T, x_i \text{ is variable})$$

where M is the number of characters (columns) in the data matrix. Fortunately,

$$\Pr(x_i \mid T, x_i \text{ is variable}) \Pr(x_i \text{ is variable} \mid T) = \Pr(x_i \mid T)$$

so:

$$\Pr(x_i \mid T, x_i \text{ is variable}) = \frac{\Pr(x_i \mid T)}{\Pr(x_i \text{ is variable} \mid T)}$$

$\Pr(x_i \text{ is variable} \mid T)$ is typically calculated by augmenting the real data matrix with a set of "fake" columns of constant characters:

$$\begin{aligned} \Pr(x_i \text{ is variable} \mid T) &= 1 - \Pr(x_i \text{ is constant} \mid T) \\ &= 1 - \sum_{j=1}^{K} \Pr(C(j) \mid T) \end{aligned}$$

where K is the number of character states and $C(j)$ denotes a character that is made up of a column of taxa all displaying state j. Because there are only K fake columns for every character, the calculation is only $(1 + K)$ times slower than a likelihood calculation that does not correct for ascertainment bias.

As noted by Matzke and Irmis (2018), if a researcher only collects parsimony informative sites, then a different form of ascertainment bias correction is needed. One could sum up all of the parsimony uninformative sites:

$$\begin{aligned} \Pr(x_i \text{ is informative} \mid T) &= 1 - \Pr(x_i \text{ is uninformative} \mid T) \\ &= 1 - \sum_{j \in \mathcal{U}} \Pr(j \mid T) \end{aligned}$$

where $\mathcal{U}$ denotes the set of all columns that are parsimony-uninformative. However the size of $\mathcal{U}$ grows as a function of the number of tips (N). Indeed it grows quickly for $K > 2$. Here we describe a dynamic programming approach that avoids the need to enumerate all uninformative patterns. This let's us calculate $\Pr(x_i \text{ is uninformative} \mid T)$ more efficiently.

Algorithm

For the typical case $N > K$, and a parsimony informative character is one that has one repeatedly observed state and a set of singleton states. The set of singleton states is the empty set in the case of constant character. When $N \leq K$, one must consider the possibility that no state is repeated.

Let $\mathcal{A}$ denote the alphabet of states (so $|\mathcal{A}| = K$). Let, $\mathcal{B}(y)$ denote $\mathcal{A} - y$; this is the set of states other than some state y; and $\mathcal{B}(y, w)$ denote $\mathcal{A} - w - y$; this is the set of states excluding state y and state w. Let $\mathcal{S}(y)$ be the power set of $\mathcal{B}(y)$. Similarly $\mathcal{S}(y, w)$ is the power set of $\mathcal{B}(y, w)$.

Below we will suppress the index (i above) that indicates which character we are correcting. The ascertainment bias correction depends only on the tree and branch length induced by considering leaves that are scored (not missing data). Thus the calculations do not depend on the specific states of column x_i; so using x is sufficient.

We want to calculate $\Pr(x \text{ is uninformative} \mid T) = \Pr(x \in \mathcal{U} \mid T)$

$$\Pr(x \in \mathcal{U} \mid T) \quad = \quad \mathcal{D}(T) + \sum_{r \in \mathcal{A}} \sum_{s \in \mathcal{S}(r)} \Pr(\mathrm{r} = r, \mathrm{s} = s \mid T)$$

where "r" is the random variable representing the repeated state; "s" is the random variable representing the singelton state set; and $\mathcal{D}(T)$ is the summation (only needed for small trees) over all patterns with no repeated states:

$$\mathcal{D}(T) = \sum_{s \in \mathcal{S}(\emptyset)} \Pr(\mathrm{r} = \emptyset, \mathrm{s} = s \mid T)$$

We can calculate $\Pr(\mathrm{r} = r, \mathrm{s} = s \mid T)$ via postorder traversal similar to Felsenstein's pruning algorithm. Let z denote an internal node, and let the lookup table element $Y[z][r][s][c]$ hold the probability of observing repeated state r and singleton state set s among the descendants of node z if node z had ancestral character state c.

If ρ denotes the node indicator for the root of the tree, then

$$\Pr(\mathrm{r} = r, \mathrm{s} = s \mid T) \quad = \quad \sum_{c \in \mathcal{A}} \pi_c Y[\rho][r][s][c]$$

where π_c is the root prior probability of state c (which is usually the equilibrium state frequency of c).

Let $l(z)$ and $r(z)$ denote the the left and right children of node(z). Let $\nu_{l(z)}$ and $\nu_{r(z)}$ denote the branch lengths leading to the left and right children of node(z).

1.1 Parent of a cherry

If z is the parent of two tips then initialize $Y[z][][][]$ using the following rules

For every $c \in \mathcal{A}$:

$$Y[z][c][\emptyset][c] = \Pr(c \to c \mid \nu_{l(z)}) \Pr(c \to c \mid \nu_{r(z)})$$

where $\Pr(i \to j \mid \nu_k)$ denotes the probability of an seeing the descendant state j if the ancestral state is i and the edge length is ν_k.

For every $c \in \mathcal{A}$ and every single state $d \in \mathcal{B}(c)$:

$$\begin{aligned} Y[z][\emptyset][\{c,d\}][c] &= \Pr(c \to c \mid \nu_{l(z)}) \Pr(c \to d \mid \nu_{r(z)}) + \Pr(c \to d \mid \nu_{l(z)}) \Pr(c \to c \mid \nu_{r(z)}) \\ Y[z][d][\emptyset][c] &= \Pr(c \to d \mid \nu_{l(z)}) \Pr(c \to d \mid \nu_{r(z)}) \end{aligned}$$

For every $c \in \mathcal{A}$ and every single state $d \in \mathcal{B}(c)$ and every single state $f \in \mathcal{B}(c,d)$:

$$Y[z][\emptyset][\{d,f\}][c] = \Pr(c \to d \mid \nu_{l(z)}) \Pr(c \to f \mid \nu_{r(z)}) + \Pr(c \to f \mid \nu_{l(z)}) \Pr(c \to d \mid \nu_{r(z)})$$

All other elements of $Y[z][][][]$ are set to 0.

1.2 Parent of two internals

If z is the parent of two other internal nodes do the following.

For every $r \in \mathcal{A}$, $c \in \mathcal{A}$ and $s \in \mathcal{S}(c)$:

$$\begin{aligned} Y[z][r][s][c] &= \sum_{d \in \mathcal{A}} Pr(c \to d \mid \nu_{l(z)}) \sum_{f \in \mathcal{A}} Pr(c \to f \mid \nu_{r(z)}) V(z,r,s,d,f) \\ V(z,r,s,d,f) &:= \sum_{t \in \mathcal{S}(r)} \Big(Y[l(z)][r][t][d] Y[r(z)][r][s-t][f] \\ &\quad \ldots + Y[l(z)][r][t][d] Y[r(z)][\emptyset][s-t][f] \\ &\quad \ldots + Y[l(z)][r][t][d] Y[r(z)][\emptyset][r+s-t][f] \\ &\quad \ldots + Y[l(z)][\emptyset][t][d] Y[r(z)][r][s-t][f] \\ &\quad \ldots + Y[l(z)][\emptyset][r+t][d] Y[r(z)][r][s-t][f] \\ &\quad \ldots + Y[l(z)][\emptyset][r+t][d] Y[r(z)][\emptyset][r+s-t][f] \Big) \end{aligned}$$

1.3 Parent of one internal and one terminal

Without loss of generality, consider rotating the node such that the left child is the terminal: For every $r \in \mathcal{A}$, $c \in \mathcal{A}$ and $s \in \mathcal{S}(c)$:

$$Y[z][r][s][c] = \sum_{d \in \mathcal{A}} Pr(c \to d \mid \nu_{l(z)}) \sum_{f \in \mathcal{A}} Pr(c \to f \mid \nu_{r(z)}) W(z,r,s,d,f)$$

where if $d = r$:

$$W(z, r, s, d, f) \quad := \quad Y[r(z)][r][s][f] + Y[r(z)][\emptyset][r + s][f]$$

if $d \neq r$:

$$W(z, r, s, d, f) \quad := \quad Y[r(z)][r][s - d][f]$$

References

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology*, 50(6):913–925.

Matzke, N. J. and Irmis, R. B. (2018). Including autapomorphies is important for paleontological tip-dating with clocklike data, but not with non-clock data. *PeerJ*, 6:e4553.