Importance sampling: 4 taxa case

Data. 4 sequences for cats: cat, tiger, leopard, clouded leopard in phylip file `4taxa-cats.phy`:

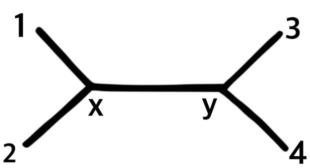
```
Cat ATGTTCATAAACCGGTGACTATTTTCAACTAATCACAAACTGAGCTGGCATGGTGGGGACTGC...
CloudedLeopard ATGTTCATAAACCGCTGACTATTTTCAACTAACCATCGCTTGGGCCGGTATAGTA...
Leopard ATGTTCATAAACCGCTGACTATTTTCAACCAATCACAAAGATAGCTGGCATGGTGGGGACTGC...
Tiger ATGTTCATAAACCGCTGACTATTTTCAACCAATCACAAGGATATTTGGTATAGTGGGGACTGC...
```

Conditional clade distribution. From the phylip input file, we obtain the conditional clade distribution from a sample bootstrapped NJ trees with the perl script `seq2ccdprobs.pl`.

Sample topology. From the conditional clade distribution, we sample one topology. Denote by $p_{ccd}(T)$ the probability of sampling the particular topology T from the conditional clade distribution.

Sample branch lengths given a topology. Let T be the 4-taxon topology sampled from the conditional clade distribution.

1. Choose one tip at random to exclude. Denote the other three sequences by seq_1, seq_2, seq_3, where seq_1, seq_2 are sisters.



2. Compute the matrices of counts between all pairs of the three sequences: x_{12}, x_{13}, x_{23}, and simulate the branch length between each pair with Tamura-Nei (TN) model, and $\eta = 0.5$ (see JCvsTN.pdf for details on choosing TN and η): d_{12}, d_{13}, d_{23}

3. Compute the distances between seq_1, seq_2 and its parent x:

$$d_{1x} = \frac{(d_{12} + d_{13} - d_{23})}{2}$$
$$d_{2x} = \frac{(d_{12} + d_{23} - d_{13})}{2}$$

4. Convert $seq_1, seq_2, seq_3, seq_4$ to matrices like

```
Cat ATGTTCAT...
```

$$S_{cat} = \begin{array}{cccccccc} A & 1 & 0 & 0 & 0 & 0 & 0 & 1 \ldots \\ C & 0 & 0 & 0 & 0 & 0 & 1 & 0 \ldots \\ G & 0 & 0 & 1 & 0 & 0 & 0 & 0 \ldots \\ T & 0 & 1 & 0 & 1 & 1 & 0 & 0 \ldots \end{array}$$

5. Estimate the sequence distribution at x from the sequences seq_1, seq_2. The formula for the likelihood at site j for node x, parent of $1, 2$ is:

$$L_j^x(s) = \left[\sum_{i \in \{A,C,G,T\}} P_{si}(d_{1x}) L_j^1(i)\right] * \left[\sum_{i \in \{A,C,G,T\}} P_{si}(d_{2x}) L_j^2(i)\right]$$
$$s \in \{A, C, G, T\}$$

where

$$L_j^k(i) = S_k[i, j], k = 1, 2$$
$$P(t) = \exp(\hat{Q}t)$$

So that the sequence matrix for x is given by S_x:

$$S_x[i, j] = \frac{\pi_i L_j^x(i)}{\sum_{i=1}^4 \pi_i L_j^x(i)}$$

6. Compute the count matrix between seq_3, seq_4: x_{34} and simulate the branch length d_{34} with TN model and $\eta = 0.5$

7. Compute the equivalent to the count matrix between x and seq_3, seq_4:

$$x_{x3} = \sum_{j=1}^{nsites} S_3[, j] * S_x[, j]^T$$

8. Simulate branch lengths d_{3x}, d_{4x} with TN model and $\eta = 0.5$

9. Compute the distances between seq_3, seq_4 and its parent y, and between x, y:

$$d_{3y} = \frac{(d_{34} + d_{3x} - d_{4x})}{2}$$
$$d_{4y} = \frac{(d_{34} + d_{4x} - d_{3x})}{2}$$
$$d_{xy} = \frac{(d_{3x} + d_{4x} - d_{34})}{2}$$

10. Compute the density of the branch lengths given the topology, denoted as $f_{TN}(d|T)$ for $d = (d_{1x}, d_{2x}, d_{xy}, d_{3y}, d_{4y})$. We simulate the branch lengths: $d_{12}, d_{13}, d_{23}, d_{3x}, d_{4x}, d_{34}$ with the TN model as gamma random variables. These 6 branch lengths were simulated independently, so the joint density is given by

$$f(d_{12}, d_{13}, d_{23}, d_{3x}, d_{4x}, d_{34}) = \prod_i \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} d_i^{\alpha_i - 1} \exp(-\beta_i d_i)$$

We then transform those branch lengths into the desired parameters (keeping the variable d_{13} for the transformation to be bijective):

$$\begin{aligned}
d_{1x} &= \frac{(d_{12} + d_{13} - d_{23})}{2} \\
d_{2x} &= \frac{(d_{12} + d_{23} - d_{13})}{2} \\
d_{13} &= d_{13} \\
d_{3y} &= \frac{(d_{34} + d_{3x} - d_{4x})}{2} \\
d_{4y} &= \frac{(d_{34} + d_{4x} - d_{3x})}{2} \\
d_{xy} &= \frac{(d_{3x} + d_{4x} - d_{34})}{2}
\end{aligned}$$

The determinant of the Jacobian in absolute value is 4. Thus, the joint density for the transformed variables is

$$\begin{aligned}
f(d_{1x}, d_{2x}, d_{3y}, d_{4y}, d_{xy}, d_{13}) = &\left[\prod_i \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}\right] (d_{1x} + d_{2x})^{\alpha_{12}-1} \exp\left(-\beta_{12}(d_{1x} + d_{2x})\right) \\
&* d_{13}^{\alpha_{13}-1} \exp\left(-\beta_{13} d_{13}\right) \\
&* (d_{13} - d_{1x} + d_{2x})^{\alpha_{23}-1} \exp\left(-\beta_{23}(d_{13} - d_{1x} + d_{2x})\right) \\
&* (d_{3y} + d_{xy})^{\alpha_{3x}-1} \exp\left(-\beta_{3x}(d_{3y} + d_{xy})\right) \\
&* (d_{4y} + d_{xy})^{\alpha_{4x}-1} \exp\left(-\beta_{4x}(d_{4y} + d_{xy})\right) \\
&* (d_{3y} + d_{4y})^{\alpha_{34}-1} \exp\left(-\beta_{34}(d_{3y} + d_{4y})\right) * 4
\end{aligned}$$

We want to integrate out d_{13},

$$\begin{aligned}
f_{TN}(d_{1x}, d_{2x}, d_{3y}, d_{4y}, d_{xy}) &= \int_0^\infty f(d_{1x}, d_{2x}, d_{3y}, d_{4y}, d_{xy}, d_{13}) dd_{13} \\
&= \left[\prod_i \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}\right] (d_{1x} + d_{2x})^{\alpha_{12}-1} \exp\left(-\beta_{12}(d_{1x} + d_{2x})\right) \\
&* 4 \exp\left(-\beta_{23}(-d_{1x} + d_{2x})\right) \\
&* (d_{3y} + d_{xy})^{\alpha_{3x}-1} \exp\left(-\beta_{3x}(d_{3y} + d_{xy})\right) \\
&* (d_{4y} + d_{xy})^{\alpha_{4x}-1} \exp\left(-\beta_{4x}(d_{4y} + d_{xy})\right) \\
&* (d_{3y} + d_{4y})^{\alpha_{34}-1} \exp\left(-\beta_{34}(d_{3y} + d_{4y})\right) \\
&* \int_0^\infty d_{13}^{\alpha_{13}-1} (d_{13} - d_{1x} + d_{2x})^{\alpha_{23}-1} \exp\left(-(\beta_{13} + \beta_{23}) d_{13}\right) dd_{13}
\end{aligned}$$

The integral is finite, but it does not have an analytic expression. The integral is a function of d_{1x}, d_{2x}. We will ignore this integral for the moment, and assume it is a constant.

Importance weight. Let $p(T)$ denote the prior distribution of tree T, and let $L(T, d)$ denote the likelihood of T under the GTR model.

$$L(T,d) = \prod_k L_k(T,d)$$

$$L_k(T,d) = \sum_{i=1}^{4} \sum_{j=1}^{4} \pi_i P_{ij}(d_{xy}) P_{i1}(d_{1x}) P_{i2}(d_{2x}) P_{j3}(d_{3y}) P_{j4}(d_{4y})$$

Then, the importance weight is

$$w(T) = \frac{p(T)L(T,d)}{g(T)}$$

$$g(T) = p_{ccd}(T) f_{TN}(d|T)$$

Algorithm 1: Importance sampling

Input: PHYLIP or NEXUS file with DNA sequences for 4 taxa; likelihood model (e.g. GTR) and prior

- Compute the conditional clade probabilities by bootstrapping and NJ: p_{ccd}
- **for** $i = 1$ **to** N **do**
  - Sample a topology $T \sim p_{ccd}$
  - Sample branch lengths from TN model $d|T \sim f_{TN}(d|T)$
  - Compute importance weight $w(T)$ with likelihood and prior
- Normalize importance weights
