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 clade distribution, we sample one topology. still don't know how? which output file to use?

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 JC (or TN) model, and $\eta = 0.5$ (see JCvsTN.pdf for details on choosing 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...`

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

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

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

Q: how to estimate this Q? Also, L_j^x is the likelihood, we want $P(x|1,2)$
So that the sequence matrix for x is given by S_x:

$$S_x[i,j] = L_j^x(i)$$

$$S_x = \begin{matrix} \text{A} & L_1^x(A) & L_2^x(A)... \\ \text{C} & L_1^x(C) & L_2^x(C)... \\ \text{G} & L_1^x(G) & L_2^x(G)... \\ \text{T} & L_1^x(T) & L_2^x(T)... \end{matrix}$$

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

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

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

Compute importance weight

$$w(tree) = posterior(tree)/g(tree)$$
$$g(tree) = p(topology)density(branchlengths|topology)$$

where $p(topology)$ comes from the clade distribution.