Notes on a mean phylogenetic tree

Background

We desire a method to define and calculate the mean of a sample of phylogenetic trees. Many others have done recent work on the Fréchet mean of phylogenetic trees using a distance defined as the minimal path distance between two phylogenetic trees in tree space (xxx references). The Fréchet mean tree a is defined as

$$a = \arg\min_{t \in \mathcal{T}} \sum_{i=1}^{n} d^2(t_i, t)$$

where $t_1, \ldots, t_n$ is a sample of trees and $\mathcal{T}$ is the set of all phylogenetic trees for some distance d.

Each phylogenetic tree t is defined by its set of splits $t_S = \{s \in \mathcal{S} : s \in t\}$ and the edge lengths associated with each split, $t(s)$ where $\mathcal{S}$ is the set of all possible splits. Extend the definition so that $t(s) = 0$ if $s \notin t$. Then define the distance between two trees t_1 and t_2 as

$$d(t_1, t_2) = \sqrt{\sum_{s \in \mathcal{S}} (t_1(s) - t_2(s))^2}$$

I do not know if this is equivalent to the path distance between trees or not.

With this definition of pairwise distance, let

$$D^2(a) = \sum_{i=1}^{n} d^2(t_i, a)$$

be the sum of squared distances between a phylogenetic tree a and sample trees $t_1, \ldots, t_n$. A tree which minimizes D^2 is a Fréchet mean tree.

Note the following straightforward derivation. For simplicity, let $\sum_i$ represent $\sum_{i=1}^{n}$ and $\sum_s$ represent $\sum_{s \in \mathcal{S}}$.

$$\begin{aligned}
D^2(a) &= \sum_i \sum_s (t_i(s) - a(s))^2 \\
&= \sum_i \sum_s (t_i(s))^2 - 2 \sum_s a(s) \sum_i t_i(s) + n \sum_s (a(s))^2 \\
&= n \Big(\frac{\sum_i \sum_s (t_i(s))^2}{n} - 2 \sum_s a(s) \frac{\sum_i t_i(s)}{n} + \sum_s (a(s))^2 \Big)
\end{aligned}$$

The first term is constant in a, and so we can eliminate it when seeking a minimizer. Define $\bar{s} = \sum_i t_i(s)/n$ to be the mean length of edges corresponding to the split s over the sample, treating trees without the split as zero values when computing the mean. With this new notation, we seek to minimize

$$\sum_s (a(s))^2 - 2 \sum_s a(s) \bar{s}$$

As $\sum_s \bar{s}^2$ does not vary with a, we can add it to the previous expression and we see that the Fréchet mean minimizes

$$\sum_s (a(s))^2 - 2\sum_s a(s)\bar{s} + \sum_s \bar{s}^2 = \sum_s (a(s) - \bar{s})^2$$

Decompose this sum over the splits in a and those not in a.

$$\sum_{s \in a} (a(s) - \bar{s})^2 + \sum_{s \notin a} (a(s) - \bar{s})^2$$

The first term will be zero if the edges corresponding to splits $s \in a$ are given lengths $a(s) = \bar{s}$ and the second sum simplifies to $\sum_{s \notin a} \bar{s}^2$ as $a(s) = 0$ for all terms. To minimize this sum, we seek the tree a with the set of compatible splits that maximizes the sum of squared mean samples splits. Thus, the Fréchet mean tree is

$$a = \underset{\substack{t \in \mathcal{T} \\ t(s) = \bar{s} \text{ for all } s \in t}}{\arg\max} \sum_{s \in t} \bar{s}^2$$