02-251: Great Ideas in Computational Biology

Mathematical Notes on Evolutionary Trees

Spring 2019

A **tree** is a connected acyclic graph.

Theorem 1. *In any tree, for any two nodes there is exactly one path connecting them.*

Proof. We proceed by contradiction. Consider arbitrary nodes v and w in a tree T. If there were two paths P_1 and P_2 connecting v and w, then we could form a cycle by traversing v to w via P_1 and returning to v via P_2. □

A corollary of Theorem 1 is that the removal of any edge $e = \{x, y\}$ from a tree disconnects the tree into two smaller trees, T_x (containing x) and T_y (containing y). Each subtree contains a collection of nodes such that for every pair of nodes $\{v, w\}$, the unique path in T connecting v and w does not pass through e.

Connected graphs tend to have more edges, and acyclic graphs tend to have fewer edges. The battle between these conflicting forces mean that trees, which are both connected and acyclic, have a fixed number of edges with respect to the number of nodes.

Theorem 2. *Every tree with n nodes has $n - 1$ edges.*

Proof. We proceed inductively. Clearly the statement holds if n is equal to 1. Assume that it holds for all values of $k < n$, and consider an arbitrary tree T with n nodes. Remove an arbitrary edge from this tree. By the corollary to Theorem 1, we have two trees containing j and k nodes, respectively, where $j + k = n$. By the inductive step, these two trees contain $j - 1$ edges and $k - 1$ edges, and so T must contain $(j - 1) + (k - 1) + 1 = n - 1$ edges as desired. □

Recall that a **distance matrix** is a matrix satisfying the three properties of nonnegativity, symmetry, and the triangle inequality. The **Hamming distance** of two strings is the number of mismatched symbols between the strings.[1]

Theorem 3. *Given a multiple sequence alignment M for n species, if $D_{i,\,j}$ is defined as the Hamming distance between rows i and j of the alignment, then D is a distance matrix.*

Proof. Clearly D is symmetric and nonnegative by how it is defined. For the triangle inequality, assume that i, j, and k are arbitrary rows of the alignment. We will show that $D_{i,\,k} \leq D_{i,\,j} + D_{j,\,k}$.

Consider a column x of M, and let $\delta_x[i,\,k]$ equal 1 if $M[i, x]$ matches $M[k, x]$ and 0 otherwise. Note that summing $\delta_x[i,\,k]$ over all columns of M yields the Hamming distance $D_{i,\,k}$. If value $M[i, x]$ is the same as $M[k, x]$, then $\delta_x[i,\,k]$ is zero and certainly $\delta_x[i,\,k] \leq \delta_x[i,\,j] + \delta_x[j,\,k]$. If value $M[i, x]$ differs from $M[k, x]$, then it is impossible for $M[j, x]$ to match both $M[i, x]$ and $M[k, x]$, so again $\delta_x[i,\,k] \leq \delta_x[i,\,j] + \delta_x[j,\,k]$. Summing this inequality over all columns x yields the desired result. □

[1] It is rare to find a simpler concept that was named after someone.

A distance matrix is **additive** if there is an edge-weighted tree "fitting" the matrix, meaning that the distance between any two leaves $d_{i,j}$ in the tree is equal to the corresponding distance matrix value $D_{i,j}$. We stated without proof that if a distance matrix is additive, then there is a unique **simple tree** fitting the distance matrix, which is a tree with no (internal) nodes of degree 2. We denote this simple tree *Tree*(D).

Theorem 4. *Every simple tree with at least three nodes has a pair of neighboring leaves.*

Proof. Consider a path $P = (v_1, v_2, \ldots, v_k)$ of maximum length k in an arbitrary simple tree T connecting nodes v_1 and v_k. Clearly v_1 and v_k are leaves (otherwise P could be extended). Because T is simple, v_2 must be adjacent to not only v_1 and v_3 but also some other node w. Note that w must be a leaf, or else we would have a longer path formed by traveling from v_k to v_2, then to w, and then to the other node w is connected to. So v_1 and w must be neighbors. □

We then saw the **Limb Length Theorem**, stated below, which powered our recursive algorithm for producing the simple tree fitting a given additive matrix.

Theorem 5. *If D is an additive matrix and j is a leaf in Tree(D), then LimbLength(j) is equal to the minimum value of $(D_{i,j} + D_{j,k} - D_{i,k})/2$ over all leaves i and k of Tree(D),*

Proof. Let m denote the parent node of j. There are two possibilities for leaves i and k in *Tree*(D). The paths $P_{i,m}$ and $P_{k,m}$ connecting i to m and k to m can either share an edge or not. If these two paths do not share an edge, then we have that

$$\begin{aligned} d_{i,j} + d_{j,k} &= (d_{i,m} + d_{m,j}) + (d_{k,m} + d_{m,j}) \\ &= d_{i,k} + 2 \cdot \mathit{LimbLength}(j) \end{aligned}$$

Rearranging terms gives that $\mathit{LimbLength}(j) = \dfrac{d_{i,j} + d_{j,k} - d_{i,k}}{2}$. All the lower case d on the right side are present in the original distance matrix, and we can replace them with capital D.

If, on the other hand, $P_{i,m}$ and $P_{k,m}$ share an edge, then a similar argument will produce that

$$\begin{aligned} d_{i,j} + d_{j,k} &= (d_{i,m} + d_{m,j}) + (d_{k,m} + d_{m,j}) \\ &\geq d_{i,k} + 2 \cdot \mathit{LimbLength}(j) \end{aligned}$$

In this case, rearranging terms yields that $\mathit{LimbLength}(j) \leq \dfrac{D_{i,j} + D_{j,k} - D_{i,k}}{2}$.

Because *Tree*(D) is simple, m must have degree at least equal to 3, which means that we can always find i and k such that $\mathit{LimbLength}(j) = \dfrac{D_{i,j} + D_{j,k} - D_{i,k}}{2}$. Taking the minimum of the right expression over all leaves i and k not equal to j therefore yields the result. □

We then saw the neighbor-joining algorithm, which we claimed constructs *Tree*(D) when D is additive but also provides a reasonable heuristic when D is not additive. But why?

Our claim was that once we inferred neighbors i and j, we set the limb length of a leaf i equal to

$$\mathit{LimbLength}^*(i) = \tfrac{1}{2}(D_{i,j} + \Delta_{i,j}) \,,$$

where

$$\Delta_{i,j} = \frac{\textit{TotalDistance}_D(i) - \textit{TotalDistance}_D(j)}{n-2}\,.$$

Here, the function $\textit{TotalDistance}_D(i)$ sums the distances between a node i and all other leaves. The * in the first equation indicates the acknowledgement that if D is non-additive, then we are not inferring the limb length of i, but rather setting it equal to an expression as part of a heuristic.

To see where this equation comes from, first suppose that D is additive, and select some leaf k other than i and j. Letting m denote the parent of i and j, we can see from the proof of the Limb Length Theorem that because the paths connecting j to m and k to m must be disjoint, then the limb length of i is equal to $\frac{D_{i,j} + D_{i,k} - D_{j,k}}{2}$.

We could just use this formula to inspire setting $\textit{LimbLength}^*(i)$ when D is non-additive. However, this formula would be biased based on our selection of the third leaf k. To reduce this bias, we could use the *average* of this formula over all k:

$$\textit{LimbLength}^*(i) = \frac{1}{n-2} \cdot \sum\nolimits_{\text{all leaves } k \text{ other than } i \text{ or } j} \frac{D_{i,j} + D_{i,k} - D_{j,k}}{2}$$

Note that each term in the sum has a $D_{i,j}$ term. So we can pull these $n-2$ terms out of the sum and obtain

$$\begin{aligned}
\textit{LimbLength}^*(i) &= \frac{D_{i,j}}{2} + \frac{1}{n-2} \cdot \sum_{k \neq i,j} \frac{D_{i,k} - D_{j,k}}{2} \\
&= \frac{D_{i,j}}{2} + \frac{1}{n-2} \cdot \left(\sum_{k \neq i,j} \frac{D_{i,k}}{2} - \sum_{k \neq i,j} \frac{D_{j,k}}{2} \right) \\
&= \frac{1}{2} \left[D_{i,j} + \frac{1}{n-2} \left(\sum_{k \neq i,j} D_{i,k} - \sum_{k \neq i,j} D_{j,k} \right) \right] \\
&= \frac{1}{2} \left[D_{i,j} + \frac{1}{n-2} \left(\sum_{k \neq i,j} D_{i,k} + D_{i,j} - \sum_{k \neq i,j} D_{j,k} - D_{i,j} \right) \right] \\
&= \frac{1}{2} \left[D_{i,j} + \frac{1}{n-2} \left(\sum_{k \neq i} D_{i,k} - \sum_{k \neq j} D_{j,k} \right) \right] \\
&= \frac{1}{2} \left(D_{i,j} + \frac{\textit{TotalDistance}_D(i) - \textit{TotalDistance}_D(j)}{n-2} \right) \\
&= \frac{1}{2} \left(D_{i,j} + \Delta_{i,j} \right)
\end{aligned}$$

We can now see that this derivation yields the previous formula.