On the Uniqueness of the Selection Criterion in NJ

MATH285K - Spring 2010 *Presenter: Selim Bahadir*

1 Introduction

The Neighbor-Joining (NJ) algorithm of Saitou and Nei (1987) is a recursive procedure for reconstructing trees that is based on the pairwise distances between leaves. It is the most widely used distance based phylogenetic method. Center of the method is the the selection criterion, the formula used to determine which pair of objects to be paired next. This paper shows that any selection criterion based on distance data that is linear, permutation equivalent and statistically consistent will give the same trees as those created by NJ.

2 Definitions and Background

2.1 X-trees and Dissimilarities

Let X be a finite set of taxa. A $phylogenetic$ X-tree is a pair $\mathcal{T} = (T, \phi)$ where $T = (V(T), E(T)$ is a tree without vertices of degree two and $\phi : X \to V$ is a bijection from X to the leaves of $\mathcal{T}$. Two taxa x, y in a phylogenetic Xtree are $neighboring$ if the path from $\phi(x)$ to $\phi(y)$ contains only one internal vertex.

A map $\delta : X \times X \ \to \mathbb{R}$ is called a $dissimilarity\ map$ if $\delta(x, y) = \delta(y, x)$ and $\delta(x, x) = 0$ for all $x, y \in X$. Let $w : E(T) \to \mathbb{R}^+$ be an assignment of positive real-valued lengths to each edge of $\mathcal{T}$. Such an assignment induces a dissimilarity map on X, denoted $d_{(\mathcal{T},w)}$ and called $additive\ distance$, where for each $x, y \in X$, we define $d_{(\mathcal{T},w)}(x, y)$ as the sum of the weights on the path between $\phi(x)$ to $\phi(y)$.

2.2 The Neighbor-Joining Method

The Q-$criterion$ for δ is the function $Q_\delta : X \times X \to \mathbb{R}$ given by

$$Q_\delta(x, y) = \delta(x, y) - \frac{1}{n-2} \sum_{z \in X} \delta(x, z) - \frac{1}{n-2} \sum_{z \in X} \delta(y, z)$$

where $n = |X|$.
The Neighbor-joining algorithm is recursive. If $n = 3$, let $X = \{x_1, x_2, x_3\}$, then the output is the x-tree with leaves x_1, x_2, x_3 joined to central vertex v and the edge weight of (x_i, v) is equal to $\frac{1}{2}(\delta(x_i, x_j) + \delta(x_i, x_k) - \delta(x_j, x_k))$ where $\{i, j, k\} = \{1, 2, 3\}$.
If $n > 3$ then we choose the pair $x, y \in X$ that minimizes $Q_\delta(x, y)$. We create a new element v_{xy} and let $X' = (X - \{x, y\}) \cup \{v_{xy}\}$. We construct a new dissimilarity δ'on X' by setting $\delta'(u, v) = \delta(u, v)$ and

$$\delta'(u, v_{xy}) = \frac{1}{2}(\delta(u, x) + \delta(u, y) - \delta(x, y))$$

for all $u, v \in X' - \{v_{xy}\}$.
Applying the algorithm recursively we get a phylogenetic X'-tree $\mathcal{T}'$ with edge weights. We attach vertices labeled x and y adjacent to v_{xy} to obtain an X'-tree. The edge (x, v_{xy}) is assigned weight

$$\frac{1}{n-2} \sum_{z \neq x,y} (\delta(x, z) + \delta(x, y) - \delta(y, z)),$$

and the edge (x, v_{xy}) is assigned weight

$$\frac{1}{n-2} \sum_{z \neq x,y} (\delta(y, z) + \delta(x, y) - \delta(x, z)).$$

2.3 Selection Criterion Conditions

(Q1) Q is $consistent.$ If $\delta = d_{(\mathcal{T},w)}$ for some X-tree $\mathcal{T}$ and x, y minimize $Q_\delta(x, y)$ then x and y are neighboring in $\mathcal{T}$.
(Q2) Q is $permutation\ equivalent.$ For any permutation σ of X we have $Q_{\sigma(\delta)}(x, y) = Q_\delta(\sigma(x), \sigma(y))$ for all $x, y \in X$. Here $\sigma(\delta)$ is the dissimilarity map defined by $\sigma(\delta)(x, y) = \delta(\sigma(x), \sigma(y))$ for all $x, y \in X$.
(Q3) Q is $linear\ and\ continuous$ in δ. Given any two dissimilarities δ and δ' and constants λ,λ' we have

$$Q_{\lambda\delta + \lambda'\delta}(x, y) = \lambda Q_\delta(x, y) + \lambda' Q_{\delta'}(x, y).$$

Condition Q1 is necessary for any alternative to the Q criterion. Condition Q2 ensures that the ranking of pairs are independent of the order of data input. Condition Q3 is quite restrictive, but given the centrality of linear estimators in statistics, linear selection criterion is a natural starting point.

3 Main Result

Earlier work of Charleston et al. (1993) indicates that there may not be convenient alternatives of the Q criterion. They showed that the Q criterion is the only consistent formula in the family of

$$\delta(x,y) - \omega(\sum_{z \in X} \delta(x,z) + \sum_{z \in X} \delta(y,z)) \quad \omega \in \mathbb{R}.$$

Work of Bryant (2005) strengthens this observation; the Q criterion is unique among all linear selection criteria, subject to conditions on consistency and independence of input order.

Theorem Let $\hat{Q}_\delta : X \times X \to \mathbb{R}$ be any function that satisfies Q1, Q2 and Q3.Then $\hat{Q}_\delta$ and Q_δ order pairs of taxa in the same way. Hence a pair $x, y \in X$ minimizes $\hat{Q}_\delta(x,y)$ if and only if it minimizes $Q_\delta(x,y)$.

Corollary Let $\hat{Q}_\delta : X \times X \to \mathbb{R}$ be any function that satisfies Q1, Q2,and Q3. Then there are $\alpha > 0$ and β, dependent only on n, such that

$$\hat{Q}_\delta(x,y) = \alpha Q_\delta(x,y) + \beta$$

for all $x, y \in X$.

References

Bryant, D.: On the Uniqueness of the Selection Criterion in Neighbor-Joining. Journal of Classification 22 (2005) :3-15

Charleston, M.,Hendy, M., and Penny, D.: Neighbor-joining uses the optimal weight for net divergence. Molecular Phylogenetics and Evolution 222 (1993):6-12.