On the Uniqueness of the Selection Criterion in Neighbor-Joining

David Bryant

February 1, 2005

Abstract

The Neighbor-Joining (NJ) method of Saitou and Nei is the most widely used distance based method in phylogenetic analysis. Central to the method is the selection criterion, the formula used to choose which pair of objects to amalgamate next. Here we analyze the NJ selection criterion using an axiomatic approach. We show that any selection criterion that is linear, permutation equivariant, statistically consistent and based solely on distance data will give the same trees as those created by NJ.

1 Introduction

The Neighbor-joining (NJ) method of Saitou and Nei (1987) is a phylogenetics classic. It is, arguably, still the most widely used distance based phylogenetic method. Saitou and Nei's original article has been cited at least 10,000 times[1], and continues to be cited extensively, a full 15 years after it first appeared. Numerous simulation studies have established the effectiveness of NJ as a statistical estimator (Nei, 1991; Charleston et al., 1994; Kuhner and Felsenstein, 1994).

NJ is an agglomerative method, putting it in the same class of tree construction and clustering methods as the single linkage method (Gower and Ross, 1969), UPGMA (Sokal and Michener, 1958), and AddTree (Sattath and Tversky, 1977), to name a few. Initially, each element is placed in a separate cluster. Two clusters or nodes are selected. These are combined into one cluster, the dissimilarity matrix is reduced, and the method proceeds recursively. Within this agglomerative framework, NJ is defined by three components: the criterion used to select pairs of nodes; the formulae used to reduce the dissimilarity matrix; and the branch length estimation formulae. These components are discussed formally in Section 2.2.

Of all the components of the NJ algorithm, the least intuitive is the selection criterion. The original paper of Saitou and Nei (1987) motivates the selection criterion indirectly, using

[1]Science Citation Index: http://www.isinet.com/products/citation/sci/

the principle of *minimum evolution*. Each pair x, y of objects or nodes is assigned a score as follows:

- First, branch lengths are estimated for the tree in Figure 1 using ordinary least squares, so that the sum of squared differences between the input distances and the induced path length distances in the tree is minimized (following Cavalli-Sforza and Edwards (1967)).
- Second, the sum of these branch lengths is determined and used as a score for x, y.

The pair x, y giving the minimum score is selected for amalgamation. Saitou and Nei (1987) provide a direct formula for computing this score. Studier and Keppler (1988) derived the simpler but equivalent formula

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

that is in use today. Here $\delta(x, y)$ is the input distance between x, y and n is the number of objects (taxa). The equivalence between the formula of Saitou and Nei and the formula of Studier and Keppler was first proven formally by Gascuel (1994).

The problem with Saitou and Nei's formulation of the selection criterion is that it is not immediately apparent why this selection criterion would be effective. Indeed it was many years before a correct consistency result was finally published (see Atteson (1999) and the discussion in Gascuel (1997b)). While the minimum evolution *is* a consistent criterion for selecting trees (Rzhetsky and Nei, 1993), the tree in Figure 1 will, in general, be different from the *true* tree. The formula itself does not help much. As noted by Felsenstein (2003) the coefficient $\frac{1}{n-2}$ looks (at first glance) like a mistake that should be replaced by $\frac{1}{n-1}$ in order to give a correct average. The fact that this replacement gives an inconsistent criterion is a direct consequence of our main result.

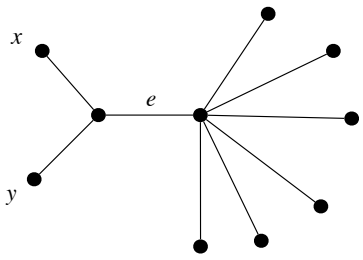

Figure 1: Tree used to evaluate the NJ selection criterion for combining x and y. Vach and Degens Vach and Degens (1991) observed that the minimizing NJ criterion is equivalent to maximizing the length of the branch e.

A small number of authors have proposed theoretical justifications for the NJ selection formula. Vach and Degens (1991) observed that the criterion can also be interpreted as a branch length estimate in this incorrect tree (see Figure 1). Charleston et al. (1993) demonstrated that the NJ criterion was unique in a parameterized family of *net divergence builders* (see Section 3). Atteson (1999) made a significant advance by showing that NJ will correctly construct trees when the observed distances are 'sufficiently close' to the true distances. Gascuel (1997b), following Mirkin (1996), suggested that the NJ criterion can be considered an estimator of the *a-centrality* of a pair of taxa, though this explanation would be more satisfying if the a-centrality definition didn't have to be tweaked in order to equal the NJ criterion. In any case, it is not clear why an estimator of these variables would provide an effective criterion for selecting pairs to agglomerate.

Swofford et al. (1996) take a different angle, and describe the NJ criterion as a rate-corrected distance. Unfortunately, the criterion will also correct ultra-metric distances, for which no rate correction is necessary.

The two other components of the NJ method, the reduction formula and the branch length formula, are far better understood. For both of these formulae it is clear what is being estimated. An extensive investigation of the statistical properties of the reduction and branch length estimation formulae appeared in a series of papers by Gascuel (1997a,b, 2000). Gascuel described the family of possible linear reduction formulae, showed that the formula used in NJ was optimal in this family when variances are unknown, and proposed new reduction formulae that incorporated variance information. Experimental studies suggested that the resulting method (BioNJ) recovers the model tree more often than NJ, under standard models of character evolution (Gascuel, 1997a).

In this paper we continue along lines similar to Gascuel, except that we focus exclusively on the selection criterion. We run immediately into a problem: if we are to justify the choice of selection formula statistically we need to know what is being estimated. This is unclear, so we attack the problem sideways. We show, under appropriate assumptions, that whatever we estimate we will always end up with a criterion equivalent to that used in NJ. Saitou and Nei did not happen to pick one particularly effective criterion out of many—they chose the only criterion possible. Put another way, the NJ selection criterion is the best in a class of one.

Of course, this result is only as significant as our assumptions are reasonable. These are:

- the selection criterion is linear, and is a deterministic function only of the distances;
- there is no a priori significance given to any taxa (or objects under study), so varying the input order does not affect the relative ordering of pairs;
- the criterion is *consistent*: given tree-like data the criterion will pick a pair of neighboring taxa.

These assumptions are formalized in Section 3. Any selection criterion must either violate one of these assumptions or be equivalent to NJ. We discuss a few of these alternative criteria in Section 4.

Note that we have assumed that the selection criterion uses *only* the distance data. If we drop this assumption then the NJ selection criterion will be far from unique. Future methods (and some existing methods, see Section 4) could and should utilize far more information when selecting pairs. What we demonstrate here is that this extra information *needs* to be incorporated if we are to progress beyond the NJ method, at least when it comes to selecting pairs to amalgamate.

The outline of this paper is as follows. In Section 2 we present definitions and notation, provide a formal description of the NJ method as well as a new, simple, proof of consistency. Our main result appears in Section 3, where we present several properties satisfied by the NJ selection criterion then prove that these properties characterize the criterion. We conclude with a brief discussion of criteria used in alternative methods.

2 Definitions and notation

For the most part, we adopt the notation of Semple and Steel (2003).

2.1 X-trees and dissimilarities

In phylogenetics, the objects, species, or sequences under study are called *taxa* (sing. *taxon*), which is an abbreviation for *operational taxonomic units (OTUs)*. 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 bijection from X to the leaves of T. Two taxa x, y in a phylogenetic X-tree are *neighboring* if the path from $\phi(x)$ to $\phi(y)$ contains at most one internal vertex.

A map $\delta : X \times X \to \mathbb{R}$ is called a *dissimilarity map* if it satisfies $\delta(x, x) = 0$ and $\delta(x, y) = \delta(y, x)$ for all $x, y \in X$.

Let $w : E(T) \to \mathbb{R}$ be an assignment of a real-valued lengths to each edge of $\mathcal{T}$. We will assume that all weights are non-zero, as zero weight edges can be contracted. Edge weights induce a dissimilarity measure on X, denoted $d_{(\mathcal{T},w)}$, where for each $x, y \in X$, we define $d_{(\mathcal{T},w)}(x, y)$ as the sum of the weights on the unique path between $\phi(x)$ and $\phi(y)$. If the weights are all positive then the resulting dissimilarity map is an *additive distance*[2]. A method for constructing trees from dissimilarity functions is said to be *consistent* if it returns $\mathcal{T}$ (and w) when applied to the additive distance $d_{(\mathcal{T},w)}$ corresponding to a weighted X-tree.

[2]Our use of the term *additivity* differs from that in measure theory

2.2 The Neighbor-joining method

Let δ be a dissimilarity map on X. For each $x \in X$ we define

$$r_\delta(x) = \sum_{y \in X} \delta(x, y). \tag{1}$$

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

$$Q_\delta(x, y) = \delta(x, y) - \frac{1}{(n-2)}(r_\delta(x) + r_\delta(y)), \tag{2}$$

where $n = |X|$. This formula was proposed by Studier and Keppler (1988) and is equivalent to the minimum evolution based criterion of Saitou and Nei (1987).

The Neighbor-joining algorithm is recursive. If $|X| = 3$, say $X = \{x_1, x_2, x_3\}$, then we return the X-tree with three leaves x_1, x_2, x_3 joined to a central vertex and weights as given in Figure 2.

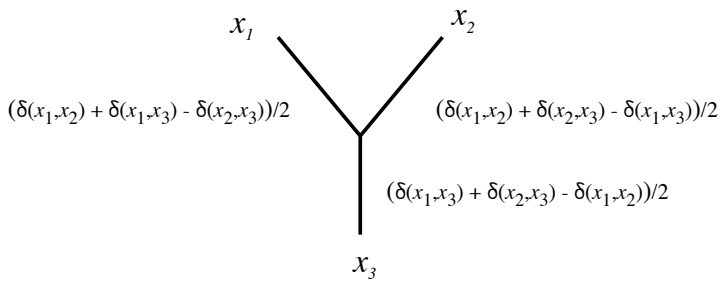

Figure 2: Tree constructed for three taxa

If $|X| > 3$ then we choose the pair $x, y \in X$ that minimizes $Q_\delta(x, y)$. We create a new element v_{xy} and set $X' = (X - \{x, y\}) \cup \{v_{xy}\}$. We construct a new dissimilarity δ' on X' by setting $\delta'(u, v) = \delta(u, v)$ and

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

for all $u, v \in X' - \{v_{xy}\}$. This is called the **reduction step**.

We now have a dissimilarity map on a smaller set. Applying the algorithm recursively we obtain 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 incident with x is assigned weight

$$b_x = \frac{1}{n-2} \sum_{z \neq x,y} (\delta(x, z) + \delta(x, y) - \delta(y, z)) \tag{3}$$

and the edge adjacent to y is assigned weight

$$b_y = \frac{1}{n-2} \sum_{z \neq x,y} (\delta(y,z) + \delta(x,y) - \delta(x,z)) \tag{4}$$

where $n = |X|$.

We make two preliminary observations. First, the algorithm can be easily implemented to run in $O(n^3)$ time. Second, it is not clear what to do when there are two pairs x, y and x', y' that both minimize $Q_\delta(x,y)$. In practice, the dissimilarity maps used are derived from discrete data, so these ties can and do arise. Ties are usually broken at random, with multiple runs used to assess the consequences of different choices.

2.3 Consistency of the NJ criterion

For NJ to be consistent, the selection criterion Q_δ must select neighboring taxa in $\mathcal{T}$ when applied to an additive distance $d_{(\mathcal{T},w)}$. Several proofs of consistency, with varying degrees of accuracy, have been proposed—see the discussion in Gascuel (1997b). Here we give a new, and particularly simple, proof.

Let $d = d_{(\mathcal{T},w)}$ be the additive distance corresponding to a phylogenetic X-tree $\mathcal{T}$ with positive weights w. We can extend d to all the vertices of $\mathcal{T}$ by letting the distance between any two vertices equal the sum of weights on the path connecting them. Likewise, for each vertex v we let $r_d(v)$ denote the sum

$$r_d(v) = \sum_{x \in X} d(v,x),$$

thereby extending (1) to all the vertices of $\mathcal{T}$.

Theorem 1 *If $d = d_{(\mathcal{T},w)}$ and x, y minimize $Q_d(x,y)$ then x and y are neighboring taxa in $\mathcal{T}$.*

Proof
First observe that for each $x \in X$, if we add a constant amount K_x to each distance $d(x,z)$ then the only affect on Q is that $\frac{2K_x}{n-2}$ is subtracted from the value of $Q_d(u,v)$ for all $u, v \in X$. This does not change the relative ordering of pairs with respect to Q. Consequently, we can assume that every external edge in $\mathcal{T}$ has weight zero.

For each non-zero weight edge $e = \{u, v\}$ we have that $r_d(u) > r_d(v)$ exactly when there are more taxa closer to v than u. Thus the maximum of $r(v)$ over all vertices is obtained for some vertex v^* adjacent to exactly one internal edge. Clearly $r_d(x) = r_d(v^*)$ for all leaves x adjacent to v^*, as all external edges have zero weight.

Suppose that x and y are not neighbors. Let z, z' be any two taxa adjacent to v^*. Since there is an internal edge separating x and y but none separating z, z' we have

$$d(x,y) > 0 = d(z,z').$$

As well, by the choice of v^* we have $r_d(x) \leq r_d(v^*) = r_d(z)$ and $r_d(y) \leq r_d(v^*) = r_d(z')$. Hence $Q_d(x,y) > Q_d(z,z')$ and x, y do not minimize the Q_d criterion. □

3 Uniqueness of the Q criterion

Gascuel (1997a,b, 2000) showed that there exists a family of admissible reduction formulae that can replace (3) and (4). This study led to two new versions of NJ. The first, UNJ, incorporates cluster sizes in order to make branch length estimates identical to those given by ordinary least squares. The second, BioNJ, incorporates variances and covariances into the reduction step. Simulation experiments indicate that this second modification improves the accuracy of tree estimation from sequence data (Gascuel, 1997a).

However, no such investigation was made into the space of admissible selection criteria. NJ, BioNJ and UNJ all use the Q criterion. Earlier work of Charleston et al. (1993) indicates that there may not be appropriate alternatives. They showed that the Q criterion is the only consistent formula in the family of *net divergence builders*

$$\delta(x,y) - \omega(r_\delta(x) + r_\delta(y)) \qquad \omega \in \mathbb{R}.$$

Our results strengthen this observation. We show that the Q criterion is unique among all *linear* selection criteria, subject to conditions on consistency and independence of input order. We now formalize the list of proposed conditions that a distance based selection criterion should satisfy.

(Q1) Q is *consistent*. If $\delta = d_{(\mathcal{T},w)}$ for some positively edge-weighted 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 equivariant*. 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 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) would need to be satisfied by any alternative to the Q criterion. Condition (Q2) ensures that the order in which the data are input does not affect the ranking of pairs - in phylogenetics, the order of sequences should be irrelevant to tree reconstruction (see Moulton and Steel (1999)). Note that, because of the way ties are broken, the same does not apply for the NJ method as a whole. The final condition (Q3) is quite restrictive, but given the centrality of linear estimators in statistics, characterizing the space of linear selection criterion is a natural starting point. Later (Section 4) we will discuss selection criteria based on non-linear functions.

The main result for this section is then

Theorem 2 *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 $Q_\delta(x, y)$.*

The theorem is proved in several steps. By (Q3) we can find coefficients $\lambda_{xy;wz}$ such that

$$\hat{Q}_\delta(x, y) = \sum_{w=1}^{n} \sum_{z=w+1}^{n} \lambda_{xy;wz}\delta(w, z)$$

for all $\delta \geq 0$, where $n = |X|$ and the coefficients $\lambda_{xy;wz}$ are functions of n. Next we will use (Q2) to re-express $\hat{Q}_\delta(x, y)$ as a linear combination involving only three different coefficients. Finally, we will use (Q1) to show that $\hat{Q}_\delta$ is equivalent to a net divergence builder, allowing us to apply Lemma 1 of (Charleston et al., 1993) to finish the proof.

Claim 1 *There exist coefficients α, β and γ such that for all $x, y \in X$,*

$$\hat{Q}_\delta(x, y) = \alpha\delta(x, y) + \beta\left(\sum_{z\neq x} \delta(x, z) + \sum_{z\neq y} \delta(y, z)\right) + \gamma\left(\sum_{w,z} \delta(w, z)\right) \tag{5}$$

The coefficients α, β, γ are all functions of n.

For any u, v, let e_{uv} denote the dissimilarity map for which $e_{uv}(u, v) = 1$ and all other values are zero. We repeatedly apply two observations. First, for any $u, v \in X$ we have $\hat{Q}_{e_{uv}}(x, y) = \lambda_{xy;uv}$. Second, if σ is any permutation on X with $s = \sigma(u)$ and $t = \sigma(v)$ then $\sigma^{-1}(e_{uv}) = e_{st}$.

Let $\{w, x, y, z\}$ and $\{i, j, k, l\}$ be subsets of X, each containing four distinct elements. Let σ be any permutation on X taking w, x, y, z to i, j, k, l respectively. From (Q2) we have $\hat{Q}_\delta(u, v) = \hat{Q}_{\sigma^{-1}(\delta)}(\sigma(u), \sigma(v))$, giving

$$\lambda_{wx;yz} = \hat{Q}_{e_{yz}}(w, x) = \hat{Q}_{e_{kl}}(i, j) = \lambda_{ij;kl}$$

$$\lambda_{wz;yz} = \hat{Q}_{e_{yz}}(w, z) = \hat{Q}_{e_{kl}}(i, l) = \lambda_{il;kl}$$

$$\lambda_{yz;yz} = \hat{Q}_{e_{yz}}(y, z) = \hat{Q}_{e_{kl}}(k, l) = \lambda_{kl;kl}.$$

Thus all coefficients indexed by four distinct elements are equal, as are all coefficients indexed by three or two distinct elements. Let λ_2, λ_3 and λ_4 be the common values for coefficients with two, three, and four distinct indices. Set $\alpha = \lambda_2 - 2\lambda_3 + \lambda_4$, $\beta = \lambda_3 - \lambda_4$ and $\gamma = \lambda_4$, giving (5). ∘

Let $S(\delta)$ denote the sum

$$S(\delta) = \gamma \sum_{u<v} \delta(u, v).$$

Since $r_\delta(x) = \sum_y \delta(x, y)$ we can rewrite (5) as

$$\hat{Q}_\delta(x, y) = \alpha\delta(x, y) + \beta(r_\delta(x) + r_\delta(y)) + S(\delta). \tag{6}$$

Claim 2 *The coefficient α in (6) is positive.*

Let T be the tree (i) in Figure 3 if $n = |X|$ is even and tree (ii) if n is odd. Assign all edges (branches) weight one, and let d be the resulting additive distance on X.

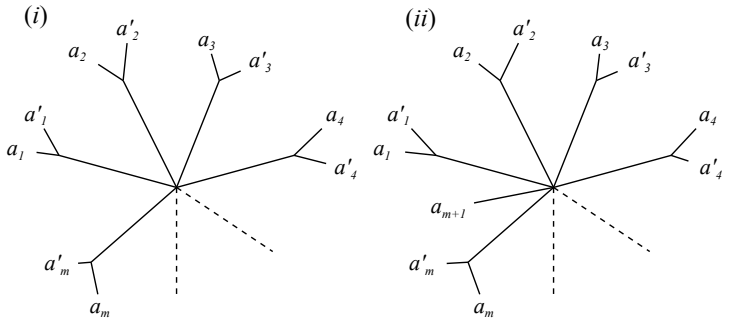

Figure 3: Two trees used to prove Claim 2. Let $m = \lfloor n/2 \rfloor$. All edges have length 1.

First note that there is some constant K such that $r_d(a_i) = r_d(a'_i) = K$ for all $i = 1, 2, \ldots, m$. For each $i \neq j$ we have

$$\hat{Q}_d(a_i, a'_i) = \hat{Q}_d(a_j, a'_j) = \alpha \cdot 2 + \beta \cdot 2K + S(d)$$

while

$$\hat{Q}_d(a_i, a_j) = \hat{Q}_d(a_i, a'_j) = \hat{Q}_d(a'_i, a_j) = \hat{Q}_d(a'_i, a'_j) = \alpha \cdot 4 + \beta \cdot 2K + S(d).$$

For $\hat{Q}$ to satisfy (Q1) we must have $\hat{Q}_d(a_i, a'_i) < \hat{Q}_d(a_i, a_j)$, giving $\alpha > 0$. ∘

Rearranging (6) we see that

$$\frac{1}{\alpha}(\hat{Q}_\delta(x, y) - S(d)) = \delta(x, y) + \frac{\beta}{\alpha}(r_\delta(x) + r_\delta(y))$$

is a *net divergence builder* (see Charleston et al. (1993)) with $\omega = \frac{\beta}{\alpha}$. Since α is positive and $S(d)$ does not depend on x, y we have that $(\hat{Q}_\delta - S(d))$ puts pairs in the same order as $\hat{Q}_\delta$ and therefore also satisfies (Q2). Lemma 1 of Charleston et al. (1993) states that that the only value of ω for which

$$\delta(x, y) + \omega(r_\delta(x) + r_\delta(y))$$

gives a consistent selection criterion is $\omega = \frac{1}{n-2}$. Hence $\frac{\alpha}{\beta} = \frac{1}{n-2}$, giving $(\hat{Q}_\delta - S(d)) = Q_\delta$ and

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

for all x, y. As $\alpha > 0$ and $S(d)$ is independent of x, y, $\hat{Q}$ ranks pairs in the same order as Q_d. □

The proof of Theorem 2 provides the following result.

Corollary 1 *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*

$$\hat{Q}_\delta(x, y) = \alpha Q_\delta(x, y) + \beta \tag{7}$$

for all $x, y \in X$.

4 Discussion

We have shown that under the appropriate, perhaps restrictive, assumptions, the selection criterion used in NJ, UNJ and BioNJ is unique. Consequently, agglomerative method differing from NJ in the selection step must violate one of the properties (Q1)—(Q3). We discuss here three well known methods, each of which satisfy two out of the three conditions. The discussion also provides a demonstration that the conditions (Q1)—(Q3) are independent: no two conditions imply the third.

The first method, UPGMA (Sokal and Michener, 1958), (or any of the linkage tree methods for that matter) chooses a pair of taxa x, y for which $\delta(x, y)$ is minimized. This criterion satisfies (Q2) and (Q3), but not (Q1), since UPGMA is not always consistent on additive dissimilarities that are not ultrametric (see, for example, the discussion in Felsenstein (2003)).

The lack of consistency in UPGMA can be corrected by applying an appropriate transformation to the dissimilarity map, such as that introduced by Farris et al. (1970). An *outgroup taxon* v_0 is fixed *a priori*. The selection process is then equivalent to choosing a pair x, y that minimizes

$$\delta(x, y) - \delta(v_0, x) - \delta(v_0, y).$$

This criterion satisfies (Q1) and (Q3), but not (Q2) since v_0 was fixed beforehand. The criterion therefore gives greater significance to the dissimilarities involving v_0, which could be problematic if v_0 is a distantly related taxon for which our distance estimates have high variance.

Bruno et al. (2000) describe an alternative to the NJ criterion based on weighted least squares. The variances in the distance estimates are incorporated directly into the selection criterion, thereby improving the statistical efficiency of the method. The variances are estimated directly from the distance matrix, so the method is still distance based. The selection criterion used in their method satisfies (Q1) and (Q2), but not (Q3).

Felsenstein concluded a recent discussion on neighbor-joining type methods (Felsenstein, 2003) by stating that there was much left to do in the development of weighted versions of neighbor-joining that properly reflect the kinds of noise that occur in biological sequence data. While Bruno et al. (2000) describe their method as likelihood based, their model of sequence evolution is not that used in the full character likelihood method of Felsenstein (1981). The closest NJ type algorithm based on character likelihood is the star decomposition heuristic implemented in PAML (Yang, 1995). However it has not been verified that this

heuristic provides a consistent phylogenetic estimator. We predict that the appropriate method might be derived from a character based version of the NJ selection criterion, though it is not clear exactly how this might be done. A proper understanding of this criterion is just a beginning.

Acknowledgements

Thanks to Olivier Gascuel, Israel Malka, Eldar Ronen and Mike Steel for comments and suggestions on earlier versions of this manuscript. This work was supported by NSERC grant 238975-01 and FQRNT grant number 2003-NC-81840.

References

K. Atteson. The performance of the neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25:251–278, 1999.

W.J. Bruno, N.D. Socci, and A.L. Halpern. Weighted neighbor joining: A likelihood-based approach to distance based phylogeny. *Molecular Biology and Evolution*, 17(1):189–197, 2000.

L. Cavalli-Sforza and A. Edwards. Phylogenetic analysis models and estimation procedures. *Evolution*, 32:550–570, 1967.

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

M. Charleston, M. Hendy, and D. Penny. The effect of sequence length, tree topology, and number of taxa on the performance of phylogenetic methods. *Journal of Computational Biology*, 1:133–151, 1994.

J.S. Farris, A.G. Kluge, and M.J. Eckart. A numerical approach to phylogenetic systematics. *Systematic Zool.*, 19:172–189, 1970.

J. Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates Inc, 2003.

O. Gascuel. A note on sattath and tversky's, saitou and nei's, and studier and keppler's algorithms for inferring phylogenies from evolutionary distances. *Mol. Biol. Evol.*, 11(6): 961–963, 1994.

O. Gascuel. Bionj: an improved version of the nj algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–695, 1997a.

O. Gascuel. Concerning the nj algorithm and its unweighted version, unj. In B. Mirkin, F.R. McMorris, F.S. Roberts, and A. Rhetsky, editors, *Mathematical Hierarchies and Biology*, pages 149–170. AMS, Providence, 1997b.

Olivier Gascuel. Data model and classification by trees: the minimum variance reduction (MVR) method. *J. Classification*, 17(1):67–99, 2000. ISSN 0176-4268.

J. C. Gower and G. J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Appl. Statist.*, 18:54–61, 1969.

M.K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11:459–468, 1994.

Boris Mirkin. *Mathematical classification and clustering*, volume 11 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 1996. ISBN 0-7923-4159-7.

V. Moulton and M. Steel. Retractions of finite distance functions onto tree metrics. *Discrete Appl. Math.*, 91(1-3):215–233, 1999. ISSN 0166-218X.

M. Nei. Relative efficiencies of different tree-making methods for molecular data. In M. M. Miyamoto and J. Cracraft, editors, *Phylogenetic analysis of DNA sequences*, pages 90–128. Oxford University Press, 1991.

A. Rzhetsky and M. Nei. Theoretical foundation of the minimum evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10:1073–1095, 1993.

N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstruction of phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

S Sattath and A. Tversky. Additive similarity trees. *Psychometrika*, 42:319–345, 1977.

C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, 2003.

R.R. Sokal and C.D. Michener. A statistical method for evaluating systematic relationships. *Univ. Kansas Science Bull.*, 38:1409–1438, 1958.

J.A. Studier and K.J. Keppler. A note on the neighbor-joining method of Saitou and Nei. *Molecular Biology And Evolution*, 5:729–731, 1988.

D. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, C. Moritz, and B.K. Mable, editors, *Molecular Systematics*, pages 407–514. Sinauer, 2nd edition, 1996.

W. Vach and P. O. Degens. Least squares approximation of additive trees to dissimilarities—characterizations and algorithms. *CSQ—Comput. Statist. Quart.*, 6(3):203–218, 1991. ISSN 0723-712X.

Z. Yang. *PAML: Phylogenetic analysis by maximum likelihood*. Institute of Molecular Genetics, Pennsylvania State University, 1995.