Algorithmica (1999) 25: 251–278

Algorithmica

The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction[1]

K. Atteson[2]

Abstract. We analyze the performance of the popular class of neighbor-joining methods of phylogeny reconstruction. In particular, we find conditions under which these methods will determine the correct tree topology and show that these perform as well as possible in a certain sense. We also give indications of the performance of these methods when the conditions necessary to show that they determine the entire tree topology correctly, do not hold. We use these results to demonstrate an upper bound on the amount of data necessary to reconstruct the topology with high confidence.

Key Words. Phylogenetic reconstruction, Neighbor-joining, Evolutionary trees.

1. Introduction. The phylogenetic reconstruction problem is to determine the evolutionary relationships between a set of species typically from information contained in biomolecular sequence data. These evolutionary relationships may be represented by a phylogenetic tree, that is, a tree in which the leaves represent extant species and the internal nodes represent possibly extinct common ancestors of the extant species. Besides being of scientific interest, methods of phylogenetic reconstruction can have important applications to human health as, for instance, in the choice of drugs for targeting particular parasites [KDD$^+$]. Particular evolutionary relationships are often debated among biologists and different relationships can be obtained by the multitude of different phylogenetic reconstruction methods. In recent years, the growth of large-scale DNA sequencing has begun to provide a wealth of data for phylogenetic reconstruction.

While over the last several decades, many methods for phylogenetic reconstruction have been proposed, there have been few proven performance guarantees for these methods until recently. One such performance guarantee is given in [ABF$^+$], which demonstrates a method which outputs an additive distance matrix (tree distance) which is within a factor of 3 of the additive distance matrix which is closest under the l_∞ norm on distance matrices (see Section 3.1 for definitions of these terms). Assuming the Cavender–Farris stochastic model of evolution, Farach and Kannan [FK] demonstrate sample-size bounds for obtaining a tree which is nearby the true model tree with respect to the variational distance between distributions defined by these trees. However, the performance guarantees of these works are difficult to interpret in terms of finding the tree which represents the actual evolutionary relationship between the species. Here, as in [ESSW], we take the view, prevalent among biologists, that the primary goal of phylogenetic reconstruction

[1] This work was supported by NSF Grant Number BIR 9413215 while the author was at the University of Pennsylvania.
[2] Yale University, Ecology and Evolutionary Biology, New Haven, CT 06520, USA. atteson@peaplant.biology.yale.edu.

Received June 1, 1997; revised March 16, 1998. Communicated by D. Gusfield and M.-Y. Kao.

is to reconstruct all or some of the edges of the true tree. We give conditions under which the neighbor-joining methods, some of the most popular of computationally efficient methods, will do so. In particular, we find the radius around the true tree, for a certain metric, in which the observed distances must be in order to guarantee that these methods reconstruct all or some of the edges of the tree. These conditions yield upper bounds on the sequence length needed for these methods to reconstruct all or some of the edges of the tree. In fact, these methods do the best possible at reconstructing the entire tree, that is, no method can be guaranteed to reconstruct the tree for observed distance matrices in a larger radius around the true tree. When the observed distance matrix is not within the radius mentioned above, so that we cannot show that the topology can be determined completely, a slight modification of one of the neighbor-joining methods can be shown to do the best possible at reconstructing some of the edges of the true tree.

In the next section we introduce some notation. In Section 3 we discuss the details of the results of the paper and their significance. The subsequent sections of the paper present the proofs of these results.

2. Some Notation.

As mentioned previously, we represent evolutionary relationships by trees, which we now define. We assume the reader is familiar with the basic concepts of graph theory, see, e.g., [Bo]. Since we are trying to determine the topology of the tree relative to the extant species which are represented as leaves, evolutionary trees are leaf-labeled trees, that is, two evolutionary trees are the same if they have the same topology relative to the leaves.

DEFINITION 1. A *tree* is a connected acyclic graph. We write $V(T)$ and $E(T)$ for the vertex set and edge set, respectively, of a tree T. A *leaf* of a tree is a node of degree 1. We write $L(T)$ for the set of leaves of tree T. When the tree T is implicitly understood, we write V, E, and L for the vertex, edge, and leaf sets of T, respectively. Two trees T and T' are (*leaf-labeled*) *isomorphic*, written $T \sim T'$ if there is a bijection $f\colon V(T) \to V(T')$ which preserves adjacency, that is, $E(T') = \{(f(v), f(v'))\colon (v, v') \in E(T)\}$, and which preserves leaves, that is, $f(v) = v$ for all leaves $v \in L(T)$. Isomorphism is an equivalence relation and we define the *topology* of a tree as the equivalence class of trees isomorphic to it. Isomorphic trees are trees which are the same for our purposes and so we sometimes blur the distinction between isomorphic trees and trees which are equal. A *rooted tree* is a tree along with a special node called the root. A *binary tree* is a tree in which every internal node has degree 3. A *rooted binary tree* is a tree having a single node of degree 2, called the *root*, and such that every other internal node has degree 3.

For a tree T and an edge $e \in E(T)$, the graph $T - e$ is the graph obtained by removing e from T, that is, if $T - e = (V, E - \{e\})$. Note that $T - e$ has exactly two components and so partitions the set of leaves into two components. For $k \in V$, we use the notation $L_k(T - e)$, or just $L_k(e)$ if the tree T is implicitly understood, for the set of leaves in the component of $T - e$ containing k (see Figure 1). Let $s(T - e) = \{L_k(T - e), L - L_k(T - e)\}$ which we refer to as the *split* of T generated by e. Let $S(T)$ denote the set of all splits of T, that is, $S(T) = \{s(T - e)\colon e \in E(T)\}$. Note that $S(T) = S(T')$ if and only if $T \sim T'$ (see, e.g., [BD]).

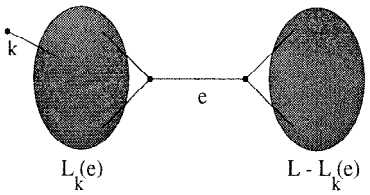

Fig. 1. An illustration of the notation "$L_k(e)$." "$L_k(e)$" denotes the set of leaves in the component of $T - e$ containing the leaf k.

For i, j, k, and l in $L(T)$, the tree T *induces the quartet* $\{\{i, j\}, \{k, l\}\}$ if there is an edge $e \in E(T)$ such that i and j are separated from k and l by e. We denote the set of quartets induced by the tree T by $Q(T)$. Note that the topology of a tree is determined by the quartets that it induces, that is, the $T \sim T'$ if and only if $Q(T) = Q(T')$ (see, e.g., [BD]).

We now introduce some definitions for distance matrices which are the inputs for the methods which we will discuss.

DEFINITION 2. A *distance matrix*, $\hat{D}$, is a symmetric nonnegative matrix, indexed by a set of taxa L, and having 0 diagonal. Note that the terminology *distance matrix* is standard in the systematic biology literature even though we do not assume that a distance matrix defines a metric or distance (nor a pseudometric) on the set of species since we do not require the triangle inequality. If $\hat{\mathcal{D}}$ is the set of distance matrices and $\mathcal{T}$ is the set of trees, a *distance-based method* for phylogenetic reconstruction[3] is a function $f\colon \hat{\mathcal{D}} \to \mathcal{T}$. A *weighted tree* τ is a tree T_τ along with a function $l\colon E(T_\tau) \to [0, \infty)$, from the edges of the tree into the positive real numbers. Similarly, a *weighted binary tree* is a weighted tree τ where T_τ is binary. For any nodes x and y of a weighted tree τ, we define the *distance* between x and y as

$$D^\tau_{xy} = \sum_{e \in P_{x,y}} l(e),$$

where $P_{i,j}$ denotes the set of edges on the unique path between i and j in T_τ. We use the symbol D^τ for the distance matrix on the leaves of T_τ. An *additive distance matrix* is a distance matrix $\hat{D}$ for which there is a weighted tree τ such that $\hat{D} = D^\tau$. Note that the weighted tree corresponding to an additive distance matrix is unique.

The following result is a local characterization of additive distance matrices which will be useful.

[3] While most methods output a weighted tree, we do not consider the weights of the output here. Also, we are avoiding reference to a computational model since it is not required here. Finally, note that there are many methods, known as *sequence-based method*, which use sequences as input rather than distances.

LEMMA 1 (Four Point Condition). *Let D be a distance matrix. Any four taxa can be labeled as i, j, k, and l in a way such that*

$$D_{ij} + D_{kl} \leq D_{ik} + D_{jl} = D_{il} + D_{jk} \tag{1}$$

if and only if D is an additive distance matrix. If D corresponds with a weighted binary tree τ, then there is an edge e which separates i and j from k and l, that is, such that i and j are in a different component of $T_\tau - e$ than k and l, and the difference between the right-hand side and left-hand side of the above inequality is at least $2l(e)$.

For the history and proof of this important result, see, e.g., [BG].

3. Discussion

3.1. *Finding the True Tree.* Intuitively, the distance matrix $\hat{D}$, which is given as input into a distance-based method for phylogenetic reconstruction, represents an estimate of the amount of evolutionary divergence between species i and j. We imagine that there is a "true" additive distance matrix $D = D^\tau$ of which the observed distance matrix $\hat{D}$ is a noisy or corrupted version. A reasonable distance-based method should return the tree T_τ when given D^τ or a distance matrix sufficiently close to D^τ as input, that is, if the noise is sufficiently small. We now demonstrate that if the observed distance matrix is too far from the actual distance matrix, then no method can be guaranteed to reconstruct the true tree correctly. First we define our notion of closeness:

DEFINITION 3. The l_∞ *norm or error* between distance matrices $\hat{D}$ and $\hat{D}'$, written $\|\hat{D} - \hat{D}'\|_\infty$, is defined as

$$\|\hat{D} - \hat{D}'\|_\infty = \max_{i,j} |\hat{D}_{ij} - \hat{D}'_{ij}|.$$

We say that a method f has l_∞ *radius* α if, for every weighted binary tree τ and every distance matrix $\hat{D}$ such that

$$\|D^\tau - \hat{D}\|_\infty < \alpha \min_{e \in E(T)} l(e),$$

the method reconstructs T_τ, that is, $f(\hat{D}) = T_\tau$.

The l_∞ radius of a method is the radius of the largest ball (in the l_∞ metric space on distance matrices), in multiples of the length of the shortest edge, around a true weighted binary tree, within which the method is guaranteed to reconstruct the true tree. In fact, we can show that no method has l_∞ radius more than $\frac{1}{2}$ using the following fact:

LEMMA 2. *For every additive distance matrix $D = D^\tau$, there is an additive distance matrix $D' = D^{\tau'}$ and a distance matrix $\hat{D}$ such that $S(T_{\tau'}) \neq S(T_\tau)$ and*

$$\|D - \hat{D}\|_\infty = \min_{e \in E(T_\tau)} \frac{l(e)}{2}, \tag{2}$$

$$\|D' - \hat{D}\|_\infty = \min_{e \in E(T_{\tau'})} \frac{l(e)}{2}. \tag{3}$$

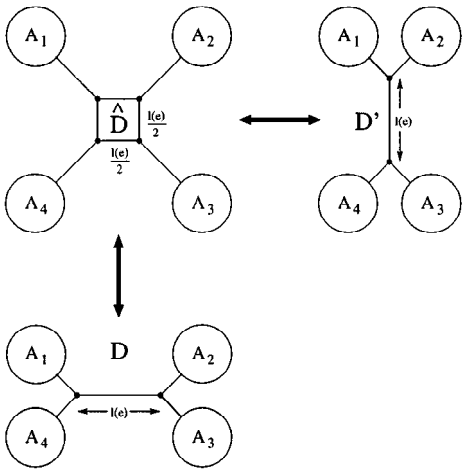

Fig. 2. A graphical representation of two "nearest" distinct tree topologies.

PROOF. In fact, D' is formed from D by rearranging the subtrees around an edge e of smallest length in T_τ as demonstrated in Figure 2. Similarly, $\hat{D}$ is constructed as shown by the weighted graph in the figure. In particular,

$$\hat{D}_{ij} = \begin{cases} D_{ij} & \text{if } i, j, \in A_k \text{ for some } k, \\ \frac{1}{2}(D_{ij} + D'_{ij}) & \text{if } i \in A_k \text{ and } j \in A_l \text{ and } |k - l| \text{ is odd}, \\ D_{ij} - \dfrac{l(e)}{2} & \text{if } i \in A_k \text{ and } j \in A_l \text{ and } |k - l| = 2. \end{cases}$$

We leave the details of this construction to the reader. See [ESSW] for a similar construction. □

This implies that no method can have l_∞ radius larger than $\frac{1}{2}$:

LEMMA 3. *No method has l_∞ radius larger than $\frac{1}{2}$.*

PROOF. Suppose a method f has l_∞ radius larger than $\frac{1}{2}$. Choose any additive distance matrix D_τ and let $D' = D^{\tau'}$ and $\hat{D}$ be, respectively, an additive distance matrix and distance matrix satisfying the properties of Lemma 2. Since the method has l_∞ radius larger than $\frac{1}{2}$ and (2) holds, f reconstructs T_τ on $\hat{D}$, that is $f(\hat{D}) = T_\tau$. However, it must also reconstruct $T_{\tau'}$ on $\hat{D}$ because of (3) and so $f(\hat{D}) = T_{\tau'}$. However, T_τ and $T_{\tau'}$ are not isomorphic since $S(T_\tau) \neq S(T_{\tau'})$ and so, in particular, $T_\tau \neq T_{\tau'}$ which is a contradiction.[4] □

[4] A subtlety here is that for sequence-based methods, that is, methods which use sequences as input, we must also show the existence of sequences which correspond to the distances $\hat{D}$. However, this can be done in many cases.

The larger the l_∞ radius of a given method, the larger the set on which we can guarantee that method will correctly reconstruct the true tree. In this paper we show that the neighbor-joining methods have the optimal l_∞ radius of $\frac{1}{2}$. In particular, we demonstrate this for the ADDTREE method of Sattah and Tversky [ST]:

THEOREM 2. *ADDTREE has l_∞ radius $\frac{1}{2}$.*

PROOF. See Section 5.1. □

Also, we demonstrate this for the popular NJ method of Saitou and Nei [SN]:

THEOREM 4. *NJ has l_∞ radius $\frac{1}{2}$.*

PROOF. See Section 6.1. □

The later result also holds for the UNJ and BIONJ methods of Gascuel [G2], [G3] which are modifications of NJ. Note that methods described in [ESSW] and the Buneman tree method [Bu] are also known to have this property but known algorithms implementing these methods have higher computational complexity than some of the neighbor-joining methods. A method which finds the closest additive distance matrix to the input distance matrix under the l_∞ norm would have l_∞ at least $\frac{1}{4}$ (see [ESSW]). However, this problem is NP-hard to approximate within a factor of $\frac{9}{8}$ [ABF$^+$]. A 3-approximation to this problem is known [ABF$^+$] which has l_∞ radius between $\frac{1}{8}$ and $\frac{1}{6}$ (see [ESSW]).

Motivated by Lemma 3, we now give a name to a distance matrix which is near enough to a weighted binary tree so that it can be guaranteed to be correctly reconstructed by a method with optimal l_∞ radius:

DEFINITION 4. A distance matrix $\hat{D}$ is *nearly additive* with respect to a weighted binary tree τ if

$$\|\hat{D} - D^\tau\|_\infty < \min_{e \in E(T_\tau)} \frac{l(e)}{2}. \tag{4}$$

A distance matrix D is *nearly additive* if there is a weighted binary tree τ such that D is nearly additive with repect to τ.

In fact, the binary tree (but not the edge weights) corresponding to a nearly additive distance matrix is unique:

LEMMA 4. *The binary tree corresponding to a nearly additive distance matrix is unique.*

PROOF. Let $\hat{D}$ be a nearly additive distance matrix. Suppose there are weighted binary trees τ and τ' such that formula (4) holds for both τ and τ'. Let e and e' be edges of minimal length in τ and τ', respectively. For any $\{\{i, j\}, \{k, l\}\} \in Q(T_\tau)$, we have

$$D_{ij}^{\tau'} + D_{kl}^{\tau'} < \hat{D}_{ij} + \hat{D}_{kl} + l(e') < D_{ij}^{\tau} + D_{kl}^{\tau} + l(e') + l(e) \tag{5}$$

$$\begin{aligned} &\leq D^{\tau}_{ik} + D^{\tau}_{jl} + l(e') - l(e) < \hat{D}_{ik} + \hat{D}_{jl} + l(e') \\ &< D^{\tau'}_{ik} + D^{\tau'}_{jl} + 2l(e'), \end{aligned}$$

where the third inequality is from the four point condition. Lemma 1, and the others from formula (4). Now suppose that $\{\{i,k\},\{j,l\}\} \in Q(T_{\tau'})$. By the four point condition, Lemma 1, there is an edge e'' such that

$$\begin{aligned} D^{\tau'}_{ik} + D^{\tau'}_{jl} &\leq D^{\tau'}_{ij} + D^{\tau'}_{jl} - 2l(e'') \\ &< D^{\tau'}_{ik} + D^{\tau'}_{jl} + 2l(e') - 2l(e'') \leq D^{\tau'}_{ik} + D^{\tau'}_{jl}, \end{aligned}$$

where the second inequality is formula (5) and the last from the fact that $l(e') \leq l(e'')$. This is a contradiction and so $\{\{i,k\},\{j,l\}\} \notin Q(T_{\tau'})$. Similarly, $\{\{i,l\},\{j,k\}\} \notin Q(T_{\tau'})$ and so, by the four point condition, $\{\{i,j\},\{k,l\}\} \in Q(T_{\tau'})$. Hence, $Q(T_{\tau}) \subseteq Q(T_{\tau'})$ and so, by symmetry, $Q(T_{\tau}) = Q(T_{\tau'})$ and $T_{\tau} \sim T_{\tau'}$. □

Because of this uniqueness, we sometimes say that $\hat{D}$ is nearly additive with respect to a tree T. The concept of nearly additive distance matrices was introduced in [ESSW] (without the name).

Finally, we would like to make some comments about the tightness of the results in this paper. Letting $f^{-1}(T)$ denote the set of distance matrices for which method f yields tree T and $N(T)$ the set of distance matrices which are nearly additive with respect to T, we will show here that $N(T) \subset f^{-1}(T)$ for certain methods. In fact, there are many distance matrices which are not in $N(T)$ for any tree T and so, in fact, $f^{-1}(T)$ is generally substantially larger. All we can say is that $f^{-1}(T)$ contains no other l_∞ balls centered at an additive distance matrix as previously noted.

3.2. *Finding Long Edges.* Let $\varepsilon = \|D - \hat{D}\|_\infty$ where $D = D^{\tau}$ and $\hat{D}$ are the true and observed distances, respectively. As discussed in the preceding sections, we can guarantee that the neighbor-joining methods will correctly reconstruct the topology of the true tree if ε is less than $\min_{e \in E(T_\tau)}(l(e)/2)$, that is, if all edges are longer than 2ε. In many cases of interest, only some of the edges of the actual tree will be of this length. In such cases, although the methods would not be guaranteed to reconstruct the entire tree correctly, they might correctly reconstruct the edges which are longer than 2ε. In particular, we associate an edge with the split that it generates and define the edge l_∞ radius analogously to the l_∞ radius:

DEFINITION 5. For a weighted tree τ, we say that the distance-based method f correctly reconstructs edge $e \in E(T_\tau)$ on input distance matrix $\hat{D}$ if there is some edge $e' \in f(\hat{D})$ such that the split generated by e in T_τ is the same as the split generated by e' in $f(\hat{D})$, that is, if $s(T_\tau - e) \in S(f(\hat{D}))$. We say that a method f has *edge l_∞ radius α* if, for every weighted binary tree τ, every edge $e \in T_\tau$, and every distance matrix $\hat{D}$ such that

$$\|D^{\tau} - \hat{D}\|_\infty < \alpha l(e),$$

the method correctly reconstructs edge e on input $\hat{D}$. Note that if a method has edge l_∞ radius larger than α, then it has l_∞ radius at least α.

Just as the l_∞ radius gives us conditions under which a method will correctly reconstruct the true tree, the edge l_∞ radius gives us conditions under which a method will correctly reconstruct sufficiently large edges of the true tree. Similarly, since no method has l_∞ radius larger than $\frac{1}{2}$, no method has edge l_∞ radius larger than $\frac{1}{2}$.

LEMMA 5. *No method has edge l_∞ radius larger than $\frac{1}{2}$.*

PROOF. If f has edge l_∞ radius larger than $\frac{1}{2}$, then it has l_∞ radius larger than $\frac{1}{2}$ which contradicts Lemma 3. □

In this paper we will show that Sattah and Tversky's method has edge l_∞ radius 0.

LEMMA 9. *ADDTREE has edge l_∞ radius* 0.

PROOF. See Section 5.2. □

However, this negative result only occurs if a pathological condition which we refer to as a tie in the four point condition (see Section 5.2) occurs. If ties are excluded, either because they occur with negligible probability or by minor modifications to the method, the method obtains edge l_∞ radius $\frac{1}{2}$ and so again performs as well as possible. See Theorem 3 of Section 5.2 for a precise statement and proof of this result. Note that the Buneman tree method [Bu] also has edge l_∞ radius $\frac{1}{2}$ (Tandy Warnow, personal communication). However, Saitou and Nei's NJ method does not do best possible:

LEMMA 13. *NJ has edge l_∞ radius at most $\frac{1}{4}$.*

PROOF. See Section 6.2. □

We conjecture that NJ has edge l_∞ equal to $\frac{1}{4}$. By comparison, the 3-approximation of the closest l_∞ additive distance matrix has edge l_∞ radius between $\frac{1}{8}$ and $\frac{1}{6}$ (see [ESSW]).

3.3. *A Probabilistic Interpretation.* By introducing a specific model of the evolution of biomolecular sequences, we can determine an upper bound on the sample-size complexity, that is, the number of samples required so that the neighbor-joining methods will correctly reconstruct either the entire topology (or edges above a certain length) with high probability. This is done by choosing the number of samples so that the error will be within the l_∞ radius (edge l_∞ radius) with high probability. The model we introduce is the Cavender–Farris model [F], [C]. Under this model, there is a true rooted tree T with n leaves corresponding to extant observed species and internal nodes corresponding to ancestral species. For any species i (extant or ancestral), there is a sequence of k binary random variables (with only the random variables corresponding to the leaf sequences being observed). The sequence at the root is generated by fair coin flips (i.i.d. uniform). With each edge $e \in E$, we associate a probability $p(e)$, the probability that a given site of the sequence will change along that edge. Each site of the sequence is assumed to change in a Markovian fashion with respect to other species, that is, any species is dependent

upon any ancestral species (more precisely, any nondescendant species) only through its most recent ancestral species. Each site is assumed to be i.i.d. (this is perhaps the biologically most unrealistic assumption of the model). The model generates a sequence of k i.i.d. binary vectors, each having one element for each extant species.

Given the Cavender–Farris model, let $p_{i,j}$ be the probability that random variables corresponding to extant species i and j at a given site will differ. Assuming $\max_{i,j} p_{i,j} < \frac{1}{2}$, the distance matrix given by

$$D_{ij} = -\tfrac{1}{2}\ln(1-2p_{i,j})$$

is additive. Note that D_{ij} is the expected number of mutations between species i and j. Letting $\hat{p}_{i,j}$ denote the number of observed mutations per site which occur between species i and j, we can estimate D_{ij} by

$$\hat{D}_{ij} = -\tfrac{1}{2}\ln(1-2\hat{p}_{i,j}).$$

In fact, using the Azuma–Hoeffding inequality [GS], we can guarantee that $\|D-\hat{D}\|_\infty < \varepsilon$ with probability at least $1-\delta$ if (see [Be] for a proof)

$$k \geq \frac{8\ln(n^2/\delta)}{(1-\exp(-\varepsilon))^2}\exp\left(\max_{i,j} 4D_{ij}\right),$$

where n is the number of species and k is the required number of observed potential mutation sites. Hence, if we allow ε to equal $\min_{e\in E}(l(e)/2)$, we can guarantee that the neighbor-joining method will find the true tree with probability at least $1-\delta$ if at least k sites are observed. Note, however, that the number of samples needed for these guaranteed performance rates would often not be practical in many situations of interest. Similarly, if we let $\varepsilon = \frac{1}{2}$, then the modification of ADDTREE discussed in Section 5.2 can be guaranteed to reconstruct every edge of length at least l correctly with probability $1-\delta$ if at least k sites are observed.

4. Neighbor-Joining Methods

4.1. *The Methods*. The neighbor-joining methods are agglomerative clustering algorithms, that is, they produce a tree in a bottom-up fashion, by iteratively combining taxa. For the purpose of this paper, we say that two taxa $i, j \in L$ are *neighbors* in a tree T if and only if $|P_{i,j}| = 2$, that is, if there are exactly two edges on the path in the tree between the taxa i and j. Every tree with at least three vertices has a pair of neighbors. The basic idea behind the neighbor-joining methods is to attempt to find a pair of species i and j which are neighbors in the tree, modify the distances so as to combine i and j into a new species u, and repeat. The pair of taxa to be combined is chosen to optimize a criterion which we refer to as the *neighbor selection criterion* (or sometimes simply as the *score*) which is a function of the observed distances $\hat{D}$ and the pair of taxa, i and j, under consideration. We denote the neighbor selection criterion for taxa i and j by $X_{i,j}(D)$. The particular neighbor selection criterion used differs between different neighbor-joining methods. For the specific neighbor selection criteria for the methods

which are analyzed in this paper, see Sections 5.1 and 6.1. After finding a pair i and j to combine into a new species u, the distances are updated in the following manner:

$$D_{uk} = \lambda_u D_{ik} + (1 - \lambda_u) D_{jk} \tag{6}$$

with distances between all other taxa remaining unchanged. The most popular methods use $\lambda_u = \frac{1}{2}$ for all u but we consider the general case here where $0 \leq \lambda_u \leq 1$ in order to be able to apply the results more generally. We summarize the workings of the neighbor-joining methods and introduce some notation in the following:

Let $L^1 = L$, $\hat{D}^1 = \hat{D}$, and $L_i = \{i\}$ for all $i \in L$.
For $m = 1, \ldots, n-2$:

1. Choose i^m and j^m which optimize $X_{i^m, j^m}(\hat{D}^m)$.
2. Fixing some new taxon u^m (e.g., $u^m = \{i^m, j^m\}$), let $L^{m+1} = L^m - \{i^m, j^m\} \cup \{u^m\}$ and $L_{u^m} = L_{i^m} \cup L_{j^m}$ and

$$\hat{D}^{m+1}_{kl} = \begin{cases} \hat{D}^m_{kl} & \text{if } k, l \neq u^m, \\ \lambda_{u^m} \hat{D}^m_{il} + (1 - \lambda_{u^m}) \hat{D}^m_{jl} & \text{if } k = u^m, \end{cases}$$

Output a tree T such that $S(T) = \{\{L_u, L - L_u\}\colon u \in \bigcup_{m=1}^{n-1} L^m\}$.

Here, the set L^m denotes the set of species and $\hat{D}^m$ denotes the distances which are input into the *mth* iteration of the method. For $u \in \bigcup_{m=1}^{n-1} L^m$, the set L_u is the subset of L which has been combined to form u which we refer to as the *representatives* of u.

4.2. *Finding the True Tree.* In this section we derive a result about conditions under which a general neighbor-joining method can be guaranteed to find the true tree. In particular, we show that any neighbor-joining algorithm which correctly chooses a pair of neighbors in the first iteration, has the optimal l_∞ radius $\frac{1}{2}$. Our first lemma says that if, when given additive distances as input, we combine a pair which are neighbors in the corresponding tree during the updating step of the method, then the result is the distances of the original tree with the pair of neighbors replaced by a single leaf hanging off of the node adjacent to the pair of neighbors. Note that similar results were proved by Bandelt and Dress [BD] but we present them here in the form we require.

LEMMA 6. *Let $D = D^\tau$ be an additive distance matrix with corresponding weighted binary tree τ. Fix neighbors i and j in T_τ. Let u' denote the internal vertex adjacent to i and j. The distance matrix on the set of taxa $(L - \{i, j\}) \cup \{u'\}$ with distances given by formula* (6) *is additive and corresponds to a weighted binary tree τ'. In particular, we can choose $T_{\tau'}$ to be the tree formed from T_τ by removing i and j. Furthermore, the edge lengths, $l'(e)$ for edges $e \in E(\tau')$, are given by*

$$l'(x, y) = \begin{cases} l(x, y) & \text{if } u' \notin \{x, y\}, \\ l(x, y) + \lambda_u l(i, u') + (1 - \lambda_u) l(\{j, u'\}) & \text{otherwise.} \end{cases} \tag{7}$$

PROOF. Let D denote the distances given by formula (6). We must show that $D^{\tau'} = D$. Clearly, if neither k nor l is u', we have that $D^{\tau'}_{kl} = D_{kl}$ since the construction of τ' does not affect the path between k and l. Otherwise, we have

$$\begin{aligned} D^{\tau'}_{ku'} &= D^{\tau}_{ku'} + \lambda_u l(i,u') + (1-\lambda_u) l(j,u') \\ &= \lambda_u (D_{ku'} + l(i,u')) + (1-\lambda_u)(D_{ku'} + l(j,u')) \\ &= \lambda_u D_{ik} + (1-\lambda_u) D_{jk}. \end{aligned}$$

Hence, $D^{\tau}_{ku} = D_{ku}$ and so the lemma holds. □

In practice, the method is not given the "actual" distances D but the approximate distances $\hat{D}$ which we will eventually assume are sufficiently close to D. It is important that, in applying the update formula (6), the "actual" distances of the tree with taxa combined and the distances used by the method do not grow further apart. We demonstrate that this is so after introducing some notation. In analogy to the observed distances used as input for the mth iteration, $\hat{D}^m$, we let D^m denote the "actual" distances at the mth iteration, when the pair which is chosen using $\hat{D}$ are combined. In other words, $D^1 = D$ and

$$D^{m+1}_{kl} = \begin{cases} D^m_{kl} & \text{if } k, l \neq u^m, \\ \lambda_{u^m} D^m_{i^m l} + (1-\lambda_{u^m}) D^m_{j^m l} & \text{if } k = u^m. \end{cases}$$

It is important to keep in mind that it is the pair, i^m and j^m, chosen by the method using $\hat{D}^m$ (and not D^m) which is used in calculating D^{m+1}. The following lemma demonstrates that the approximate and actual distances do not grow further apart.

LEMMA 7. *For any m, we have*

$$\|\hat{D}^m - D^m\|_\infty \leq \|\hat{D} - D\|_\infty.$$

PROOF. The proof is by induction. The result holds for $m = 1$ by definition. Now suppose that the result holds for m. If $k, l \neq u^m$, then the distances are unchanged and so $|\hat{D}^{m+1}_{kl} - D^{m+1}_{kl}| = |\hat{D}^m_{kl} - D^m_{kl}|$. Otherwise

$$\begin{aligned} \left|\hat{D}^{m+1}_{u^m k} - D^{m+1}_{u^m k}\right| &= \left|\left(\lambda_{u^m}\hat{D}^m_{i^m k} + (1-\lambda_{u^m})\hat{D}^m_{j^m k}\right) - \left(\lambda_{u^m}\hat{D}^m_{i^m k} + (1-\lambda_{u^m}) D^m_{j^m k}\right)\right| \\ &= \left|\lambda_{u^m}\left(\hat{D}^m_{i^m k} - D^m_{i^m k}\right) + (1-\lambda_{u^m})\left(\hat{D}^m_{j^m k} - D^m_{j^m k}\right)\right| \\ &\leq \lambda_{u^m}\left|\hat{D}^m_{i^m k} - D^m_{i^m k}\right| + (1-\lambda_{u^m})\left|\hat{D}^m_{j^m k} - D^m_{j^m k}\right| \\ &\leq \lambda_{u^m}\|\hat{D}^m - D^m\|_\infty + (1-\lambda_{u^m})\|\hat{D}^m - D^m\|_\infty \\ &= \|\hat{D}^m - D^m\|_\infty \end{aligned}$$

Hence, the result holds. □

The previous two lemmas allow us to characterize the performance of an arbitrary neighbor-joining method in terms of its performance on the first iteration:

THEOREM 1. *Fix a neighbor-joining method such that*:

1. *Given a nearly additive distance matrix $\hat{D}$ with respect to a tree T, any pair which optimizes the neighbor selection criterion are neighbors in T.*
2. *The update formula is given by* (6).

For any nearly additive distance matrix $\hat{D}$ with respect to T, the neighbor-joining method outputs T, that is, the method has the optimal l_∞ radius $\frac{1}{2}$.

PROOF. We assume that $\hat{D}$ is nearly additive and let $D = D^{\tau^1}$ be an additive distance matrix which is nearby (any one of which has the same topology by Lemma 4). Let $E^1 = E(T_{\tau^1})$ and $l^1\colon E^1 \to [0,\infty)$ denote the edge weights of τ^1. First we show that D^m is additive and i^m and j^m are neighbors. We prove this by using the induction hypothesis that the following three conditions hold simultaneously:

(a) $D^m = D^{\tau^m}$ for some weighted binary tree τ^m with edge set E^m and edge weights $l^m\colon E^m \to [0,\infty)$.
(b) $\min_{e\in E^m} l^m(e) \geq \min_{e\in E^1} l^1(e)$.
(c) i^m and j^m are neighbors in T_{τ^m}.

The base case, $m = 1$, follows directly from the assumptions of the lemma. Now suppose by induction that D^m is additive and that item (b) holds. We use the notation $T^m = T_{\tau^m}$. We have

$$\|\hat{D}^m - D^m\|_\infty \leq \|\hat{D} - D\|_\infty < \min_{e\in E^1} \frac{l(e)}{2} \leq \min_{e\in E^m} \frac{l(e)}{2},$$

where the first inequality is from Lemma 7, the second is assumed for proof of the lemma, and the third is by the induction hypothesis. Hence, since D^m is additive for the weighted binary tree τ^m, we have that $\hat{D}^m$ is nearly additive and so, by assumption of the lemma, any pair, i^m and j^m, optimizing the neighbor selection criterion is a pair of neighbors in T^m which verifies item (c). Hence, by Lemma 6, D^{m+1} is additive and T^{m+1} is binary which verifies item (a). Also from Lemma 6, $\min_{e\in E^{m+1}} l^{m+1}(e) \geq \min_{e\in E^m} l^m(e)$ since, if $e = (x,y)$, then either $e \in E^m$ or $u' \in \{x,y\}$, in which case $l^{m+1}(e) \geq l^m(e)$. Hence, by the induction hypothesis, $\min_{e\in E^{m+1}} l^{m+1}(e) \geq \min_{e\in E^1} l^1(e)$ verifying item (b) and completing the induction. Hence, the neighbor-joining method chooses a pair of neighbors at every iteration.

For $e \in E^m$ and an arbitrary $k \in L^m$, let $S = L_k(T^m - e)$. Define $s'(T^m - e) = \{\bigcup_{u\in S} L_u, L - \bigcup_{u\in S} L_u\}$, that is, $s'(T^m - e)$ is the split generated by e considered as sets of L by using the representatives of vertices $u \in L^m$. We will now show that

$$\left\{\{L_u, L - L_u\}\colon u \in \bigcup_{l=1}^{m} L^l\right\} \cup \{s'(T^m - e)\colon e \in E^m\} = S(T^1) \tag{8}$$

holds for every m. The proof is by induction on m. For the base case, we have that $\bigcup_{u\in S} L_u = \bigcup_{u\in S}\{u\} = S$ for any $S \subseteq L^1$ and so $s'(T^1 - e) = s(T^1 - e)$. This means that $\{s'(T^m - e)\colon e \in E^m\} = S(T)$ and so (8) holds in the base case. We now prove the induction step. As in Lemma 6, let $u^{m\prime}$ be the vertex adjacent to i^m and j^m. For any $e \in E^m - \{(u^{m\prime}, i^m), (u^{m\prime}, j^m)\}$, we have that $L_{i^m}(T^m - e) = L_{u^m}(T^{m+1} - e) -$

$\{i^m, j^m\} \cup \{u^m\}$. Letting $S = L_{u^m}(T^{m+1} - e)$, we have that $\bigcup_{u \in S} L_u = \bigcup_{u \in L_{i^m}(T^m - e)} L_u - (L_{i^m} \cup L_{j^m}) \cup L_{u^m} = \bigcup_{u \in L_{i^m}(T^m - e)} L_u$. Hence, $s'(T^m - e) = s'(T^{m+1} - e)$ for $e \in E^m - \{(u^{m'}, i^m), (u^{m'}, i^m)\}$. Note that the first term of the union on the left-hand side of (8) is strictly increasing. Hence, we need only show that $s'(T^m - (i^m, u^{m'}))$ is contained in the first term of the union and similarly with i^m replaced by j^m. However, we have that $s'(T^m - (i^m, u^{m'})) = \{L_{i^m}, L - L_{i^m}\}$ and similarly for j^m and so (8) holds in the induction case. Hence, we have that $\{\{L_u, L - L_u\}\colon u \in \bigcup_{l=1}^{n-1} L^l\} \cup \{s'(T^{n-1} - e)\colon e \in E^{n-1}\} = S(T^1)$. However, since we decrease the number of leaves by one in each iteration, we have that $E^{n-1} = \{(u, u')\}$ for some $u, u' \in L^{n-1}$. Hence, $\{L_u, L - L_u\} \in \{\{L_u, L - L_u\}\colon u \in \bigcup_{l=1}^{n-1} L^l\}$ and so $\{\{L_u, L - L_u\}\colon u \in \bigcup_{l=1}^{n-1} L^l\} = S(T^1)$. This means that the neighbor-joining method correctly reconstructs the topology of the tree. □

5. Sattah and Tversky's Method

5.1. *Finding the True Tree.* The first neighbor-joining method that we study was introduced by Sattah and Tversky [ST] and is often called ADDTREE. Before defining the neighbor selection criterion of this method, we first introduce a useful terminology. We say that a pair of taxa i and j *win the four point condition* for a quartet $\{i, j, k, l\}$ (under $\hat{D}$) if

$$\hat{D}_{ij} + \hat{D}_{kl} < \min(\hat{D}_{ik} + \hat{D}_{jl}, \hat{D}_{il} + \hat{D}_{jk}).$$

In other words, i and j win the four point condition for the quartet $\{i, j, k, l\}$ if the inequality of the four point condition holds strictly. Note that of the six (unordered) pairs of any quartet, either none or two pairs win the four point condition for the quartet.

Sattah and Tversky's method maximizes the neighbor selection criterion $X_{i,j}(D) = C_{i,j}$ defined as follows:

$$C_{i,j} = |\{(k, l)\colon D_{ij} + D_{kl} < \min(D_{ik} + D_{jl}, D_{il} + D_{jk})\}|, \tag{9}$$

where we assume i, j, k, l are all distinct. This is the number of pairs $\{k, l\}$ such that i and j win the four point condition for the quartet $\{i, j, k, l\}$. Furthermore, ADDTREE uses $\lambda_u = \frac{1}{2}$ for all u in the updating formula (6).

In this section we prove that Sattah and Tversky's method has an optimal l_∞ radius for determining the true topology. In order to show that ADDTREE outputs the tree corresponding to a nearly additive distance matrix given it as input, it only remains to show that (1) of Lemma 1 holds. While this is a simple matter, we present it here in the form of a more general lemma for later use.

LEMMA 8. *Let $\varepsilon = \|\hat{D} - D^\tau\|_\infty$ for actual distance matrix $D = D^\tau$ and observed distance matrix $\hat{D}$. Suppose $i, j, k, l \in L^m$ are such that there is an edge e in the binary tree T_τ such that $l(e) > 2\varepsilon$ and such that L_i and L_j are contained in the same component of $T - e$ and L_k and L_l are contained in the other component, then i and j win the four point condition for the quartet $\{i, j, k, l\}$ under $\hat{D}^m$ for any m.*

PROOF. We prove by induction that $\hat{D}^m_{ij} + \hat{D}^m_{kl} < \hat{D}^m_{ik} + \hat{D}^m_{jl}$. A symmetric argument shows that $\hat{D}^m_{ij} + \hat{D}^m_{kl} < \hat{D}^m_{il} + \hat{D}^m_{jk}$. For the base case, fix $i, j, k, l \in L$ such that e with $l(e) > 2\varepsilon$ separates i and j from k and l (note that $L_i = \{i\}$, etc.). By the four point condition, Lemma 1, we have

$$D_{ij} + D_{kl} < D_{ik} + D_{jl} - 4\varepsilon$$

Hence,

$$\begin{aligned} \hat{D}_{ij} + \hat{D}_{kl} &\leq D_{ij} + D_{kl} + 2\varepsilon \\ &< D_{ik} + D_{jl} - 2\varepsilon \leq \hat{D}_{ik} + \hat{D}_{jl}, \end{aligned}$$

where the first and last inequalities are by assumption of the lemma and the middle from above. Hence, the result holds for the base case. Now suppose that the result holds for m. Since $L^{m+1} = (L^m - \{i^m, j^m\}) \cup \{u^m\}$, the result will hold by induction if $u^m \notin \{i, j, k, l\}$. Hence, we may assume without loss of generality that $i = u^m$. Note that $L_{u^m} = L_{i^m} \cup L_{j^m}$ and so we have that e separates $i' \in L_{i^m}$ and $j' \in L_{j^m}$ from $k' \in L_k$ and $l' \in L_l$. By the induction hypothesis $\hat{D}^m_{i^m j} + \hat{D}^m_{kl} < \hat{D}^m_{i^m k} + \hat{D}^m_{jl}$ and $\hat{D}^m_{j^m j} + \hat{D}^m_{kl} < \hat{D}^m_{j^m k} + \hat{D}^m_{jl}$. Hence,

$$\begin{aligned} \hat{D}^m_{ij} + \hat{D}^m_{kl} &= \tfrac{1}{2}\left(\hat{D}^m_{i^m j} + \hat{D}^m_{kl} + \hat{D}^m_{j^m j} + \hat{D}^m_{kl}\right) \\ &< \tfrac{1}{2}\left(\hat{D}^m_{i^m k} + \hat{D}^m_{jl} + \hat{D}^m_{j^m k} + \hat{D}^m_{jl}\right) \\ &= \hat{D}^m_{ik} + \hat{D}^m_{jl}, \end{aligned}$$

which demonstrates the lemma. □

Now we can demonstrate that the method of Sattah and Tversky has an optimal l_∞ radius.

THEOREM 2. *ADDTREE has l_∞ radius $\frac{1}{2}$.*

PROOF. From Lemma 1, we need only show that, for any nearly additive distance matrix $\hat{D}$ with respect to τ, every maximizing pair of the neighbor selection criterion given by formula (9) is a pair of neighbors of T_τ. Fix a pair of neighbors i and j. The pair of taxa i and j are separated from any other pair of taxa k and l by some internal edge e. Since all edges are length at least 2ε by assumption, we have by Lemma 8 that i and j win the four point condition for the quartet $\{i, j, k, l\}$. Hence, $C_{i,j} = ((n-2)(n-3))/2$, which is its maximal value. We must now demonstrate that no nonneighbors achieve this value. Fix a pair of nonneighbors k and l. Since k and l are nonneighbors, they are separated by an internal edge e. Let i be a member of the component of $T - e$ containing k and let j be a member of the component containing l (these must exist since otherwise e is not an internal edge). Since e separates i and k from j and l, we see, again by Lemma 8, that k and l do not win the four point condition for the quartet $\{i, j, k, l\}$ (since i and k as well as j and l do). Hence, $C_{k,l}$ cannot achieve the maximal value of $C_{i,j} = ((n-2)(n-3))/2$. Hence, the neighbor selection criterion is only maximized at pairs of neighbors and so the method outputs the topology of T_τ. □

5.2. *Finding Long Edges*. In the previous section we have shown that Sattah and Tversky's neighbor-joining method performs well when all edges have length at least 2ε where $\varepsilon = \|D - \hat{D}\|_\infty$ for actual and observed distances D and $\hat{D}$. In this section we examine what happens when this assumption is violated. First we introduce some terminology. We say that a pair of taxa i and j *tie* the four point condition for a quartet $\{i, j, k', l'\}$ if there is a labeling of k' and l' as k and l such that

$$D_{ij} + D_{kl} = D_{ik} + D_{jl} \leq D_{il} + D_{jk}.$$

If the inequality in the above expression is strict, then four pairs tie the four point condition for that quartet and we call this a *two-way tie*. If equality holds, then all six pairs tie the four point condition for that quartet and we call this a *three-way tie*. For any quartet, there is a pair which either wins or ties the four point condition for that quartet and there are never pairs which win and tie simultaneously. Note, however, that no "points" are assigned for ties in Sattah and Tversky's method (that is, the neighbor selection criterion is not directly dependent of how many quartets for which a pair of taxa ties the four point condition). This fact allows us to construct a counterexample demonstrating arbitrarily bad performance of the method when there are short edges:

LEMMA 9. *ADDTREE has edge l_∞ radius* 0.

PROOF. We must show that, for any sufficiently large number M and any $\varepsilon > 0$, there is an additive distance matrix D with tree τ which has an edge of length $M\varepsilon$ and there is a distance matrix $\hat{D}$ such that $\|D - \hat{D}\|_\infty \leq \varepsilon$ and such that Sattah and Tversky's method will not correctly reconstruct the edge of length $M\varepsilon$. We present a counterexample which works for all such M below. First we provide some of the intuition behind the counterexample and subsequently the details, many of which are tedious and left to the reader to verify. Note that the amount by which the inequality of the four point condition, Lemma 1, is satisfied is at least twice the length of any edge which separates the two pairs of neighbors. The intuition is that we will choose a tree with a single long edge and make all remaining edges short so that we can force pairs which do not span the long edge to have ties. Fix two taxa on opposite sides of the long edge, k^* and l^*. We wish to force the method to choose k^* and l^* in the first iteration. There are three types of quartets:

1. Quartets with pairs on both sides of the long edge.
2. Quartets with a triplet on one side of the long edge and a single taxa on the other.
3. Quartets with all taxa on one side of the long edge.

We (judiciously) choose an equal number of taxa on either side of the long edge, namely, $m = n/2$ taxa on each side. The outcome (win or tie and for which pairs) for quartets of type 1 cannot be changed because we assume that the long edge is at least length 2ε. Each pair on one side of the long edge wins the four point condition for $\binom{m}{2}$ of these pairs (once for each pair on the other side). We can, however, make it so that k^* and l^* win all quartets of type 2 in which they are involved. There are $2\binom{m-1}{2}$ such quartets. For each of these, a pair on one side or the other will win as well and so, assuming that we can split these up evenly, there will be approximately $\binom{m-1}{2}/\binom{m-1}{2}$ such quartets for which

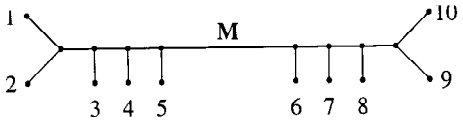

Fig. 3. The actual weighted tree for the counterexample for Lemma 9. All edges have length $\varepsilon/8$ except for the single long edge which has length $M\varepsilon$.

each pair on a single side of the long edge will win. We can arrange all other quartets (namely, those of type 3) to have ties. The resulting score for pairs $\{i, j\}$ on the same side of the long edge will be at most $C_{i,j} \leq \binom{m}{2} + 1$ and $C_{k^*,l^*} = 2\binom{m-1}{2}$. For $m = 5$, i.e., 10 leaves total, we can see that k^* and l^* will be chosen on the first iteration, thereby incorrectly reconstructing the long edge. Now we present the details of the construction.

The actual weighted binary tree used in the construction of the counterexample is shown in Figure 3. We choose all edges to be length $\varepsilon/8$ except for the long edge, which we choose to be length $M\varepsilon$. Table 1 presents the observed distances $\hat{D}$ of the counterexample in multiples of ε. It can be verified that the l_∞ distance between this distance matrix and the weighted binary tree given in Figure 3 is ε. Finally, Table 2 shows the neighbor selection criterion, $C_{i,j}$ for each pair of leaves i and j. This table can be validated from the previous table assuming that M is sufficiently large so that a sum of distances will always be larger than another if it contains a higher multiple of M. It can be seen that 5 and 6 maximize the neighbor selection criterion and so will be chosen as i^1 and j^1. Also note that $\{5, 6\} = L_{u^1}$ and $L_{u^1} \subseteq L_{u^m}$ or $L_{u^1} \subseteq L - L_{u^m}$ for all m. Hence, it is not possible that the method finds the split $s(T - e)$ where e is the long edge, since 5 and 6 are in different components of $T - e$. □

The above demonstration is unsatisfying in that it is dependent upon the occurrence of ties, which should not occur often and, furthermore, can be handled correctly by modifying the method slightly. For instance, if we modify the neighbor selection criterion so as to assign points for ties, the above counterexample no longer holds. In fact, we demonstrate below that if ties do not occur, or if they are assigned large enough scores,

Table 1. The approximate distances of the counterexample for Lemma 9 in multiples of ε. The diagonal entries are 0 and other blank entries can be filled in by symmetry.

Leaf no.	2	3	4	5	6	7	8	9	10
1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	1	$M + \frac{1}{2}$	M	M	M	M
2		$\frac{1}{2}$	$\frac{1}{2}$	1	$M + \frac{1}{2}$	M	M	M	M
3			$\frac{1}{2}$	1	$M + \frac{1}{2}$	M	M	M	M
4				1	$M + \frac{1}{2}$	M	M	M	M
5					$M + \frac{1}{2}$	$M + \frac{1}{2}$	$M + \frac{1}{2}$	$M + \frac{1}{2}$	$M + \frac{1}{2}$
6						1	1	1	1
7							$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
8								$\frac{1}{2}$	$\frac{1}{2}$
9									$\frac{1}{2}$

Table 2. The neighbor selection criterion, $C_{i,j}$, of the counterexample for Lemma 9 in multiples of ε. The diagonal entries are irrelevant and other blank entries can be filled in by symmetry.

Leaf no.	2	3	4	5	6	7	8	9	10
1	11	11	11	10	0	0	0	0	0
2		11	11	10	0	0	0	0	0
3			11	10	0	0	0	0	0
4				10	0	0	0	0	0
5					12	0	0	0	0
6						10	10	10	10
7							11	11	11
8								11	11
9									11

then Sattah and Tversky's method will work optimally and correctly reconstruct every edge of length at least 2ε.

THEOREM 3. *Suppose one of the following conditions hold*:

1. *There are no ties*.
2. *Ties are scored so that the total contribution of a quartet in which there is a tie at least equals the total contribution of a quartet in which there is a win, i.e., each pair in a two-way tie scores at least* $\frac{1}{2}$ *point and each pair in a three-way tie scores at least* $\frac{1}{3}$ *point*.

Then the resulting neighbor-joining method has edge l_∞ *radius* $\frac{1}{2}$.

PROOF. Fix an input distance matrix $\hat{D}$ which is nearly additive with respect to some weighted binary tree τ and assume that there is an edge $e \in E(T_\tau)$ of length at least 2ε that the method will not correctly reconstruct. Let $s(T_\tau - e) = \{L', L - L'\}$. The method does not correctly reconstruct edge e if and only if $\{L_u, L - L_u\} \neq \{L', L - L'\}$ for all $u \in \bigcup_{m=1}^{n-1} L^m$. Hence, for every $u \in \bigcup_n L^m$, one of the following must hold:

(a) L_u is strictly contained in L' or $L - L'$.
(b) L_u strictly contains L' or $L - L'$.
(c) Each of the four intersections, $L_u \cap L'$, $L_u \cap (L - L')$, $(L - L_u) \cap L'$ and $(L - L_u) \cap (L - L')$, are nonempty.

Note that for all $u \in L$, item (a) holds. Let $L^{n-1} = \{u, u'\}$ (by this iteration, there are only two remaining taxa). Item (a) cannot hold for both u and u' because, since $L_u = L - L_{u'}$, this would imply either $L_u = L'$ or $L_{u'} = L'$. Hence, item (b) or item (c) holds for either u or u'. Let $m + 1$ be the smallest number such that there is a $u \in L^{m+1}$ for which either item (b) or item (c) holds. Since $L^{m+1} = (L^m - \{i^m, j^m\}) \cup \{u^m\}$, it must be that item (b) or item (c) holds for $u = u^m$. However, since item (a) holds for i^m and j^m and $L_{u^m} = L_{i^m} \cup L_{j^m}$, it cannot be that item (b) holds for $u = u^m$. Hence, item (c) holds for $u = u^m$. We summarize this situation by saying that the edge e is broken at iteration m. We have just shown that every edge which is not correctly reconstructed is broken at some iteration.

Now assume, without loss of generality, that m is the earliest iteration at which an edge of length at least 2ε is broken. Let $e_1, e_2, \ldots, e_p$ denote all edges of length at least 2ε which are broken at iteration m. It follows from the previous paragraph that, for every $u \in L^m$, the set L_u will be contained in a single component of $T - \{e_1, \ldots, e_p\}$. Let $L^{i^m} \subseteq L^m$ denote set of all u such that L_u is contained in the same component of $T_\tau - \{e_1, \ldots, e_p\}$ as L_{i^m}. We proceed to bound the values of $C_{i,j}$ for $i, j \in L^m$. In order to do this, we determine, when possible, which pairs win the four point condition for various quartets. From the previous paragraph, it can be seen that L^{i^m} and L^{j^m} are disjoint. Hence, i^m and j^m can only win the four point condition for quartets of the form $\{i^m, j^m, k, l\}$ such that k and l are both in either L^{i^m}, L^{j^m}, or $L^m - L^{i^m} - L^{j^m}$. There are at most the following number of such quartets, letting $x_1 = |L^{i^m}|$ and $x_2 = |L^{j^m}|$ and $x_3 = |L - L^{i^m} - L^{j^m}|$:

$$\binom{x_1 - 1}{2} + \binom{x_2 - 1}{2} + \binom{x_3}{2}. \tag{10}$$

Now consider a pair $i, j \in L^{i^m}$ and assume without loss of generality that $x_1 \leq x_2$. The pair i and j wins the four point condition for any quartet of the form $\{i, j, k, l\}$, where $k, l \in L^m - L^{i^m}$ by Lemma 8 since some edge from $\{e_1, \ldots, e_p\}$ separates L_i and L_j from L_k and L_l. There are $\binom{x_2+x_3}{2}$ pairs of this form. Also, if there are no ties as assumed in one of the alternatives in the statement of the theorem, then, for any quartets $i, j, k, l \in L^{i^m}$, there are two pairs in L^{i^m} which win the four point condition for that quartet. Alternatively, if ties are scored as mentioned in the lemma, then the total score for each pair in L^{i^m} from a quartet of the kind mentioned is 2. Hence, in either of these cases, the total contribution to the sum of scores of pairs in L^{i^m} from quartets of this form is at least twice the number of such quartets, $\binom{x_1}{4}$. For quartets of the form $i, j, k \in L^{i^m}$ and $l \in L^m - L^{i^m}$, at least one pair in L^{i^m} wins the four point condition for that quartet (e.g., if i and l win the four point condition for the quartet, then so do j and k). Similarly, if ties are scored as mentioned, then the total score from ties for pairs in L^{i^m} will be 1 since each tying pair is either in L^{i^m} or its complement in the quartet is. The total contribution to the sum of scores of pairs in L^{i^m} from quartets of this form is at least the number of such quartets, $\binom{x_1}{3}(x_2 + x_3)$. Summing the contributions from various types of quartets, there must be a pair, $i^*, j^* \in L^m$, whose score exceeds the average:

$$\binom{x_2 + x_3}{2} + \frac{2\binom{x_1}{4} + \binom{x_1}{3}(x_2 + x_3)}{\binom{x_1}{2}}. \tag{11}$$

We will demonstrate that (11) exceeds (10). Subtracting (10) from (11), expanding and simplifying yields

$$\begin{aligned}
&\tfrac{1}{6}(-2x_1^2 + 2x_1x_2 + 2x_1x_3 + 6x_2x_3 + 4x_1 + 2x_2 - 4x_3 - 6) \\
&\quad = \tfrac{1}{3}(x_1(x_2 - x_1) + (x_1 + 3x_2 - 2)x_3 + (2x_1 + x_2 - 3)) \\
&\quad \geq \tfrac{1}{3}((x_1 + 3x_2 - 2)x_3 + (2x_1 + x_1 - 3)) > 0,
\end{aligned}$$

where we have used the fact that $x_2 \geq x_1$ for the second inequality and the facts that $x_1 \geq 1$ and $x_2 \geq 1$ and the fact that either $x_1 > 1$, $x_2 > 1$, or $x_3 > 0$ (since we can assume that there are at least three leaves) for the third. Hence, $C_{i^*,j^*} > C_{i^m,j^m}$ which is a contradiction, since the method chooses $i^m, j^m \in L^m$ to maximize C_{i^m,j^m}. □

The above result is somewhat counterintuitive in light of the previous result. Since a tie would seem to provide no evidence, it is more intuitive to throw away ties than to score them. However, throwing them away gives an advantage to pairs which are not involved in quartets with ties.

6. Saitou and Nei's Method

6.1. *Finding the True Tree.* In this section we demonstrate that the neighbor-joining method of Saitou and Nei [SN] has the optimal l_∞ radius of $\frac{1}{2}$. In Saitou and Nei's method, the neighbor selection criterion [SK] is given by $X_{i,j}(D) = S_{i,j}$:

$$S_{i,j} = (n-2)D_{ij} - \sum_k D_{ik} - \sum_k D_{jk}, \tag{12}$$

which is minimized to choose a pair to combine at each iteration. Another version of the method is given in [SK] but these versions are proved equivalent in [G1]. Saitou and Nei [SN] suggest using update formula (6) with $\lambda_u = \frac{1}{2}$ but the results presented in this section also apply to variants of the method which use other choices for λ_u, such as the methods presented in [G2], [G3]. Note that the neighbor selection criterion, formula (12), is linear in the distances which are linear in the edge weights when D is additive. We now present a lemma which determines the weights of this formula for each edge.

LEMMA 10. *Suppose that the additive distances D correspond with the weighted tree τ, with edge set E. We have*

$$S_{i,j} = \sum_{e \in E} w_e(i,j)l(e), \tag{13}$$

where

$$w_e(i,j) = \begin{cases} -2 & \text{if } e \in P_{i,j}, \\ -2|L - L_i(e)| & \text{otherwise,} \end{cases} \tag{14}$$

where $P_{i,j}$ denotes the set of edges on the path between i and j and L denotes the set of taxa.

Note that $w_e(i,j)$ is symmetric in i and j since, for $e \in E - P_{i,j}$, we have that $|L - L_i(e)| = |L - L_j(e)|$.

PROOF. First note that for taxa k and l, the distance D_{kl} is the sum of the branch lengths of the edges in $P_{k,l}$:

$$D_{kl} = \sum_{e \in P_{k,l}} l(e).$$

Hence,

$$\begin{aligned} S_{i,j} &= (n-2)D_{ij} - \sum_k D_{ik} - \sum_k D_{jk} \\ &= \sum_{e \in P_{i,j}} (n-2)l(e) - \sum_k \sum_{e \in P_{i,k}} l(e) - \sum_k \sum_{e \in P_{j,k}} l(e). \end{aligned} \tag{15}$$

Note that $e \in P_{i,k}$ if and only if $k \in L - L_i(e)$. Hence,

$$\sum_k \sum_{e \in P_{i,k}} l(e) = \sum_{e \in E} \sum_{k \in L - L_i(e)} l(e) = \sum_{e \in E} |L - L_i(e)| l(e).$$

Incorporating this into (15) yields

$$\begin{aligned} S_{i,j} &= \sum_{e \in P_{i,j}} (n-2) l(e) - \sum_k \sum_{e \in P_{i,k}} l(e) - \sum_k \sum_{e \in P_{j,k}} l(e) \\ &= \sum_{e \in P_{i,j}} (n-2) l(e) - \sum_{e \in E} (|L - L_i(e)| + |L - L_j(e)|) l(e) \\ &= \sum_{e \in P_{i,j}} ((n-2) - (|L - L_i(e)| + |L - L_j(e)|)) l(e) \\ &\quad - \sum_{e \in E - P_{i,j}} (|L - L_i(e)| + |L - L_j(e)|) l(e). \end{aligned}$$

However, for $e \in P_{i,j}$, we have that $|L - L_i(e)| + |L - L_j(e)| = n$. Also, for $e \in E - P_{i,j}$, we have that $|L - L_i(e)| = |L - L_j(e)|$. Hence,

$$S_{i,j} = -2 \sum_{e \in P_{i,j}} l(e) - 2 \sum_{e \in E - P_{i,j}} |L - L_i(e)| l(e),$$

which was to be shown. □

We now find the difference in the neighbor selection criterion, $S_{i,j}$, for nonneighbors and neighbors when calculated from the actual distances. Let N be the set of pairs of L which are neighbors. It would be desirable to determine a lower bound on

$$\min_{\{k,l\} \subseteq \overline{N}} S_{k,l} - \min_{\{i,j\} \subseteq N} S_{i,j}$$

so that when we consider $S_{k,l}$ calculated using approximate distances, we will know the tolerance within which the distances can vary. However, it turns out to be easier to bound

$$\min_{\{k,l\} \subseteq \overline{N}} (S_{k,l} - S_{i,j}),$$

where i and j are neighbors chosen to depend upon k and l. In fact, there is no loss in the tightness of the overall results using this less strict bound.

LEMMA 11. *Let $D = D^\tau$ for a weighted binary tree τ. Let S denote the results of formula* (12) *applied to the distance D. If $k, l \in L$ are not neighbors, then there is a pair of neighbors $i, j \in L$ such that either*

$$S_{k,l} - S_{i,j} \geq 3(n-4) \min_e l(e)$$

and $\{i, j\} \cap \{k, l\} = \varnothing$ *or*

$$S_{k,l} - S_{i,j} \geq 2(n-3) \min_e l(e)$$

and $|\{i, j\} \cap \{k, l\}| = 1$.

PROOF. We first summarize the proof. We consider the subtrees hanging off of the path from k to l and choose i and j to be any pair of neighbors in one of these subtrees which does not uniquely contain a maximal number of leaves. The proof then follows by case analysis. Using Lemma 10, for each $e \in E$, the weight on $l(e)$ in $S_{k,l}$ can be shown to be at least that in $S_{i,j}$. In particular, for edges which separate i and j from k and l, the weight in $S_{i,j}$ will be substantially less than the weight in $S_{k,l}$ since the component containing i and j will be smaller than the component containing k and l. Summing bounds on the differences in the weights on various edges leads to the desired result.

We now proceed with the details. Let $T = T_\tau$. Fix any taxa, $k, l \in L$ which are not neighbors as in the statement of the lemma. Choose an edge $e^* \in E - P_{k,l}$ which minimizes $|L_k(e^*)|$. Note that the component of $T - e^*$ which contains k and l has at least three leaves (k, l, and at least one other which must come off of $P_{k,l}$ since k and l are not neighbors and only one edge e^* has been deleted). Hence, there are a pair of neighbors, say i and j in this component, which must also be neighbors in T (see Figure 4). We show that $w_e(i, j) \leq w_e(k, l)$ and by how much via case analysis on e:

1. Suppose that $e \in P_{k,l}$. We have that $w_e(i, j) = -2|L - L_i(e)| \leq -2 = w_e(k, l)$ since $|L - L_i(e)| \geq 1$ for all e. We now determine a lower bound on the quantity $\sum_{e \in P_{k,l} - P_{i,j}} (w_e(k, l) - w_e(i, j))$ for the case in which $\{k, l\} \cap \{i, j\} = \varnothing$. From what we have just shown, this quantity is nonnegative. Let $P_{k,l} = \{k, v_1, \ldots, v_m, l\}$ with $m \geq 2$ since k and l are not neighbors. Note that each v_p, for $1 \leq p \leq m$, has an edge e_p incident on it which is not in $P_{k,l}$ since the tree is binary and every internal node is incident with three edges. For some e_{p^*} (see Figure 5), we have $\{i, j\} \in L - L_k(e_{p^*})$ since i and j are neighbors and neither equals k or l. Note that it must be that $L_i(e_{p^*}) \subseteq L_k(e^*)$ since $e_{p^*} \in P_{i,k} \subseteq L_k(e^*)$. Also, by the minimality of $|L_k(e^*)|$, we have $|L_k(e^*)| \leq |L_k(e_{p^*})|$ and so $|L - L_k(e_{p^*})| = |L_i(e_{p^*})| \leq |L_k(e^*)| \leq |L_k(e_{p^*})|$ or $|L_k(e_{p^*})| \geq n/2$. Now letting $e' = (v_{p^*-1}, v_{p^*})$ and $e'' = (v_{p^*}, v_{p^*+1})$ (see Figure 5), we have $w_{e'}(i, j) + w_{e''}(i, j) = -2(|L - L_i(e')| + |L - L_i(e'')|) = -2|L_k(e_{p^*})| \leq -n$. Hence, for this pair of edges, $w_{e'}(k, l) + w_{e''}(k, l) - w_{e'}(i, j) - w_{e''}(i, j) \geq n - 4$.
2. Suppose $e \in P_{i,k} - P_{k,l}$. First we show that e^* is the component of $T - e$ containing k. Otherwise $L_k(e) \subseteq L_K(e^*)$ strictly (since $i \in L_k(e^*) - L_k(e)$). Since $|L_k(e^*)|$ is minimal among $|L_k(e)|$ for all $e \notin P_{k,l}$, it must be that e^* is in the component of $T - e$ containing k.

   Hence, e^* is in the component of $T - e$ containing k. In this case we have that $L - L_k(e^*) \subset L_k(e)$, strictly, since $k \in L_k(e)$ and $k \notin L - L_k(e^*)$. By the minimality of $|L_k(e^*)|$ among $\{|L_k(e)|\colon e \in E - P_{k,l}\}$, we have that $n - |L_k(e^*)| \geq n - |L_k(e)|$ and so $w_e(k, l) = -2(n - |L_k(e)|) \geq -2(n - |L_k(e^*)|) > -2|L_k(e)| = -2(n - |L_i(e)|) = w_e(i, j)$. We now determine a lower bound on the sum of $w_e(k, l) - w_e(i, j)$ over edges for which this case occurs, which we have already shown is positive (assuming there is at least one such edge). Note that since i and j are neighbors and k and

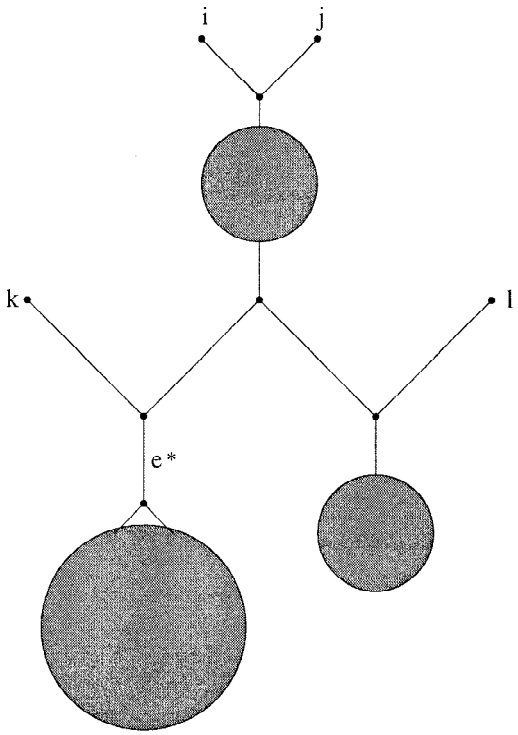

Fig. 4. An illustration of the choice of neighbors in Lemma 11. The neighbors are chosen in any nonmaximal component coming off of $P_{k,l}$, the path between k and l.

l are not, we can assume without loss of generality that $k \notin \{i, j\}$ (by choosing $l \in \{i, j\}$ if necessary). Hence, there is an edge $e''' \in P_{i,k}$ which separates i and j from the remaining taxa. For this edge, we have that $w_{e'''}(i, j) = -2(n-2)$ and $w_{e'''}(k, l) = -2$ if $l \in \{i, j\}$ and $w_{e'''}(k, l) = -2|L - L_k(e''')| = -4$ if $l \notin \{i, j\}$.

3. Suppose $e \in P_{j,l} - P_{k,l}$. If $l \notin \{i, j\}$, this case is subsumed by case 2 since i and j are symmetric. Otherwise, we assume without loss of generality that $l = j$ and so there are no edges in $P_{j,k}$.
4. Suppose $e \in E - P_{k,l} - P_{i,k} - P_{j,l}$. In this case, $\{i, j, k, l\} \subseteq L_k(e)$. Hence, $w_e(i, j) = -2(n - |L_i(e)|) = -2(n - |L_k(e)|) = w_e(k, l)$.

Hence, assuming $\{i, j\} \cap \{k, l\} = \varnothing$ and using the differences in weights found in cases 1 and 2:

$$\begin{aligned} S_{k,l} - S_{i,j} &= \sum_{e \in E} (w_e(k, l) - w_e(i, j)) l(e) \\ &\geq \left(\sum_{e \in E} w_e(k, l) - w_e(i, j) \right) \min_e l(e) \end{aligned}$$

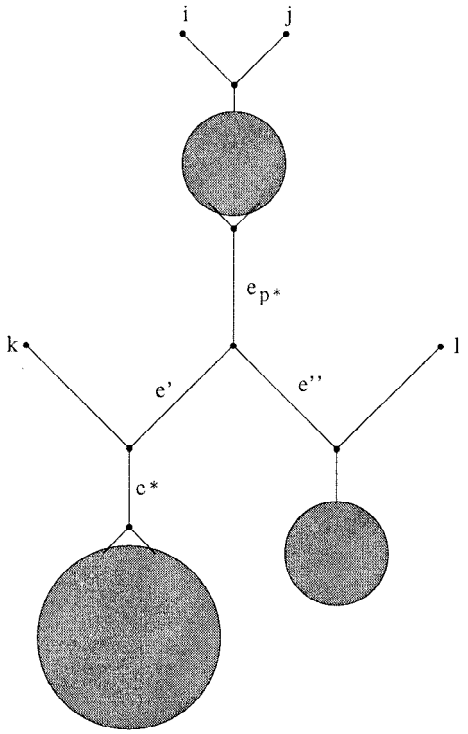

Fig. 5. Illustration of case 1 of the proof of Lemma 11.

$$
\begin{aligned}
&\geq \left(\sum_{e \in \{e', e'', e'''\}} w_e(k, l) - w_e(i, j) \right) \min_e l(e) \\
&= (w_{e'}(k, l) + w_{e''}(k, l) - w_{e'}(i, j) - w_{e''}(i, j) \\
&\quad + w_{e'''}(k, l) - w_{e'''}(i, j)) \min_e l(e) \\
&\geq (n - 4 + -4 - -2(n - 2)) \min_e l(e) = 3(n - 4) \min_e l(e).
\end{aligned}
$$

Similarly, assuming $|\{i, j\} \cap \{k, l\}| = 1$ and using the difference in weights found in case 2,

$$
\begin{aligned}
S_{k,l} - S_{i,j} &= \sum_{e \in E} (w_e(k, l) - w_e(i, j)) l(e) \\
&\geq (w_{e'''}(k, l) - w_{e'''}(i, j)) l(e''') \\
&\geq (-2 - -2(n - 2)) \min_e l(e) = 2(n - 3) \min_e l(e)
\end{aligned}
$$

Hence, the theorem is proved. □

Now let $\hat{S}$ denote the results of formula (12) applied to distances $\hat{D}$. We wish to show that, for any nonneighbors k, l, there are neighbors i, j such that $\hat{S}_{k,l} > \hat{S}_{i,j}$. First we

decompose

$$\hat{S}_{k,l} - \hat{S}_{i,j} = \hat{S}_{k,l} - S_{k,l} + S_{k,l} - S_{i,j} + S_{i,j} - \hat{S}_{i,j}.$$

Lemma 11 bounds the middle pair of terms of the above and so it would be natural to seek a bound for $|\hat{S}_{k,l} - S_{k,l}|$ for any pair k, l in order to bound the outer terms. However, this does not lead to the tightest results and so we instead bound $\hat{S}_{k,l} - S_{k,l} + S_{i,j} - \hat{S}_{i,j}$ directly which, due to a cancelation of terms, yields a tighter bound:

LEMMA 12. *Let D and $\hat{D}$ denote two distance matrices. We have*

$$\hat{S}_{k,l} - S_{k,l} + S_{i,j} - \hat{S}_{i,j} \geq -6(n-4)\|D - \hat{D}\|_\infty$$

when $\{i, j\} \cap \{k, l\} = \varnothing$ and

$$\hat{S}_{k,l} - S_{k,l} + S_{i,j} - \hat{S}_{i,j} \geq -4(n-3)\|D - \hat{D}\|_\infty$$

when $|\{i, j\} \cap \{k, l\}| = 1$.

PROOF. Let $\varepsilon_{i,j} = \hat{D}_{ij} - D_{ij}$. We have, from (12),

$$\begin{aligned}(16)\qquad &\hat{S}_{k,l} - S_{k,l} + S_{i,j} - \hat{S}_{i,j} \\ &\quad = (n-2)(\varepsilon_{k,l} - \varepsilon_{i,j}) + \sum_m (\varepsilon_{i,m} + \varepsilon_{j,m} - \varepsilon_{k,m} - \varepsilon_{l,m}).\end{aligned}$$

Considering the two cases of the lemma separately:

1. Suppose $\{i, j\} \cap \{k, l\} = \varnothing$ and so

$$\begin{aligned}&\hat{S}_{k,l} - S_{k,l} + S_{i,j} - \hat{S}_{i,j} \\ &\quad = (n-4)(\varepsilon_{k,l} - \varepsilon_{i,j}) + \sum_{m \notin \{i,j,k,l\}} (\varepsilon_{i,m} + \varepsilon_{j,m} - \varepsilon_{k,m} - \varepsilon_{l,m}) \\ &\quad \geq -6(n-4)\|D - \hat{D}\|_\infty\end{aligned}$$

   and so the result holds in this case.
2. Suppose $|\{i, j\} \cap \{k, l\}| = 1$. Assume without loss of generality that $i = k$. In this case, (16) reduces to

$$\begin{aligned}&\hat{S}_{k,l} - S_{k,l} + S_{i,j} - \hat{S}_{i,j} \\ &\quad = (n-2)(\varepsilon_{k,l} - \varepsilon_{i,j}) + \sum_m (\varepsilon_{j,m} - \varepsilon_{l,m}) \\ &\quad = (n-3)(\varepsilon_{k,l} - \varepsilon_{i,j}) + \sum_{m \notin \{j,k,l\}} \varepsilon_{j,m} - \sum_{m \notin \{i,j,l\}} \varepsilon_{l,m} \\ &\quad \geq -4(n-3)\|D - \hat{D}\|_\infty\end{aligned}$$

   Hence, the result holds in this case. □

Finally, we are in the position to show our main result for Saitou and Nei's neighbor-joining method, that it has the optimal l_∞ radius of $\frac{1}{2}$:

THEOREM 4. *NJ has l_∞ radius $\frac{1}{2}$.*

PROOF. Let D be an additive distance matrix corresponding to the weighted binary tree τ. Let $\hat{D}$ be an observed distance matrix and suppose that $\|D - \hat{D}\|_\infty < \min_{e \in E}(l(e)/2)$, that is, $\hat{D}$ is nearly additive. We must show that NJ yields T_τ on input $\hat{D}$. From Lemma 1, we need only show that the method chooses a pair which are neighbors in the first iteration when given a nearly additive distance matrix as input. Fix nonneighbors k and l and let i and j be the neighbors whose existence is demonstrated in Lemma 11. Suppose first that $\{k, l\} \cap \{i, j\} = \varnothing$:

$$\begin{aligned}
\hat{S}_{k,l} - \hat{S}_{i,j} &= \hat{S}_{k,l} - S_{k,l} + S_{i,j} - \hat{S}_{i,j} + S_{k,l} - S_{i,j} \\
&\geq \hat{S}_{k,l} - S_{k,l} + S_{i,j} - \hat{S}_{i,j} + 3(n-4) \min_{e \in E} l(e) \\
&\geq -6(n-4)\|D - \hat{D}\|_\infty + 3(n-4) \min_{e \in E} l(e) \\
&> 0,
\end{aligned}$$

where we have used Lemma 11 for the first inequality, Lemma 12 for the second, and the assumption that $\hat{D}$ is nearly additive for the third. Similarly, if $|\{k, l\} \cap \{i, j\}| = 1$, we have

$$\begin{aligned}
\hat{S}_{k,l} - \hat{S}_{i,j} &= \hat{S}_{k,l} - S_{k,l} + S_{i,j} - \hat{S}_{i,j} + S_{k,l} - S_{i,j} \\
&\geq \hat{S}_{k,l} - S_{k,l} + S_{i,j} - \hat{S}_{i,j} + 2(n-3) \min_{e \in E} l(e) \\
&\geq -4(n-3)\|D - \hat{D}\|_\infty + 2(n-3) \min_{e \in E} l(e) \\
&> 0.
\end{aligned}$$

Hence, the method must choose a pair of neighbors at the first iteration and so, by Lemma 1, outputs a tree with topology T. □

6.2. *Finding Long Edges*. In Section 5.2 we showed that a variant of Sattah and Tversky's ADDTREE method has optimal edge l_∞ radius $\frac{1}{2}$. Here we discuss this problem for the method of Saitou and Nei's NJ method. We show that the edge l_∞ radius of NJ is at most $\frac{1}{4}$. To demonstrate an upper bound of $1/M$, we must provide a distance matrix which is within ε of the distance matrix for a weighted binary tree with an edge e of length at least $M\varepsilon$ such that NJ does not correctly reconstruct e given the distance matrix as input. In order for this to happen, NJ must combine a pair on opposite sides of e before combining a pair on each side of e (for further details on this, see the proof of Theorem 3). In fact, we can use linear programming to help the search for counterexamples by allowing the edge lengths and errors to become variables. Given a particular topology, a particular edge to be broken, and a particular sequence of pairs to combine, the longest that the edge can be such that it can be broken can be found by a linear program. We have experimented with numerous such possibilities, including many which seem likely to be

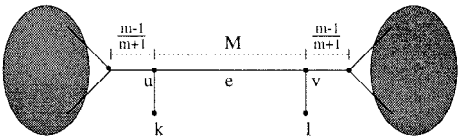

Fig. 6. The actual weighted tree for the counterexample for Lemma 13. All edges are chosen to be short except for the long edge of length $M\varepsilon$ and the two edges of length $((m-1)/(m+1))\varepsilon$ where $m = (n-2)/2$.

counterexamples and we were unable to come up with a counterexample containing an edge of length 4ε or more which could be broken. Hence, we conjecture that Saitou and Nei's neighbor-joining method will correctly reconstruct any edge of length at least 4ε. Here, we show that NJ cannot be guaranteed to do so for any lower number.

LEMMA 13. *NJ has edge l_∞ radius at most $\frac{1}{4}$.*

PROOF. Unlike in Lemma 9, the number of leaves of the counterexample must grow in order to approach 4ε. We fix a number m, which we will eventually constrain to be sufficiently large, and construct a weighted binary tree τ containing $n = 2m + 2$ taxa. The counterexample is illustrated in Figure 6. The tree will contain an internal edge, e, with endpoints u and v, and have length $M\varepsilon$ which separates the set of all taxa into two groups of equal size. There will be two special leaves, which we label k and l, hanging directly off of u and v, respectively. Besides e and the branch leading to k, there will be an internal branch of length $((m-1)/(m+1))\varepsilon$ hanging off of u which separates k from the m remaining species on that side of e and similarly for v. Let L_u denote the set of leaves on the same side of e as k but not including k, i.e., $L_u = L_k(e) - \{k\}$ and similarly for L_v. We leave the topology and branch lengths on the remaining species open, subject to the constraint that the longest edge in these subtrees is at most length $\gamma\varepsilon$ where γ will eventually be chosen sufficiently small. We now choose $\hat{D}$ by choosing $\varepsilon_{i,j} = \hat{D}_{ij} - D^{\tau}_{ij}$ as follows:

$$\varepsilon_{i,j} = \begin{cases} -\varepsilon & \text{if } i \in L_u \text{ and } j \in L_v \text{ or if } i \in L_v \text{ and } j \in L_u \text{ or if } \{i, j\} = \{k, l\}, \\ \varepsilon & \text{otherwise.} \end{cases}$$

We demonstrate that $\hat{S}_{k,l} < \hat{S}_{i,j}$ for every pair $i, j \in L_u \cup \{k\}$ and every pair $i, j \in L_v \cup \{l\}$. These two cases are symmetric and so we assume $i, j \in L_u \cup \{k\}$. First consider $i, j \in L_u$. Using Lemma 10, it can be seen that

$$S_{k,l} - S_{i,j} \le 2mM\varepsilon + 4\frac{m-1}{m+1}\varepsilon + c(m)\gamma\varepsilon$$

for some function $c(n)$. Now using case 1 from the proof of Lemma 12 and simplifying, it can be seen that

$$\hat{S}_{k,l} - \hat{S}_{i,j} \le 2mM\varepsilon + 4\frac{m-1}{m+1}\varepsilon + c(m)\gamma\varepsilon - 8m\varepsilon + 4\varepsilon.$$

Since $M < 4$, it can be seen that if n is sufficiently large (so that $2mM - 8m + 8 < 0$), we can choose γ sufficiently small so that the above expression will be negative. Now

suppose that $i = k$ and $j \in L_u$. Again, using Lemma 10, it can be seen that

$$S_{k,l} - S_{i,j} \leq 2mM\varepsilon - (2m-2)\frac{m-1}{m+1}\varepsilon + c(m)\gamma\varepsilon$$

for some function $c(n)$. Using case 2 from the proof of Lemma 12 and simplifying, it can be seen that

$$\hat{S}_{k,l} - \hat{S}_{i,j} \leq 2mM\varepsilon - (2m-2)\frac{m-1}{m+1}\varepsilon + c(m)\gamma\varepsilon - 6m\varepsilon + 2\varepsilon.$$

Again, since $M < 4$, it can be seen that if m is sufficiently large, we can choose γ sufficiently small so that the above expression will be negative. Putting these results all together, Saitou and Nei's method will choose k and l as neighbors on the first iteration, thereby breaking the edge e. □

Acknowledgments. I would like to thank Tandy Warnow for suggesting this problem and for many useful discussions related to it, Olivier Gascuel for many suggested corrections to an earlier manuscript, Shibu Yooseph for working with me on finding counterexamples for trees with short edges, and Junhyong Kim for many useful suggestions and discussions. Thanks also to the other reviewers for many useful comments.

References

[ABF+] R. Agarwala, V. Bafna, M. Farach, B. O. Narayanan, M. Paterson, and M. Thorup. On the approximability of numerical taxonomy. In *Proceedings of the Seventh Annual ACM–SIAM Symposium on Discrete Algorithms*, pages 365–372, 1996.

[BD] H.-J. Bandelt and A. Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics*, 7:309–343, 1986.

[Be] V. Berry. Méthodes et Algorithmes pour reconstruire les arbres de l'Évolution. Ph.D. thesis, Université de Montpellier, 1997.

[BG] J.-P. Barthélemy and A. Guénoche. *Trees and Proximity Representations*. Wiley, New York, 1991.

[Bo] B. Bollobás. *Graph Theory*. Springer-Verlag, New York, 1979.

[Bu] P. Buneman. The recovery of trees from measures of dissimilarity. In *Mathematics in the Archeological and Historical Sciences* (F. R. Hodson, D. G. Kendall, and P. Tautu, eds.). Edinburgh University Press, Edinburgh, 1971.

[C] J. A. Cavender. Taxonomy with confidence. *Mathematical Biosciences*, 40:271–280, 1978.

[ESSW] P. L. Erdős, M. A. Steel, L. A. Székely, and T. J. Warnow. A few logs suffice to build (almost) all trees (ii). Technical Report 97-72, DIMACS, 1997.

[F] J. S. Farris. A probability model for inferring evolutionary trees. *Systematic Zoology*, 22:250–256, 1973.

[FK] M. Farach and S. Kannan. Efficient algorithms for inverting evolution. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on the Theory of Computing*, pages 230–235, 1996.

[G1] O. Gascuel. A note on Sattah and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Molecular Biology and Evolution*, 11(6):961–963, 1994.

[G2] 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, 1997.

[G3] O. Gascuel. Concerning the NJ algorithm and its unweighted version, UNJ. In *Mathematical Hierarchies and Biology*, pages 149–170, American Mathematical Society, Providence, RI, 1997.

[GS] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford University Press, Oxford, 1992.

[KDD+] S. Köhler, C. F. Delwiche, P. W. Denny, L. G. Tilney, P. Webster, R. J. M. Wilson, J. D. Palmer, and D. S. Roos. A plastid of probable green algal origin in apicomplexan parasites. *Science*, 275:1485–1489, 1997.

[SK] J. A. Studier and K. J. Keppler. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5(6):729–731, 1988.

[SN] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[ST] S. Sattath and A. Tversky. Additive similarity trees. *Psychometrika*, 42(3):319–345, 1977.