

# Minimum evolution using ordinary least-squares is less robust than neighbor-joining

Stephen J. Willson  
Department of Mathematics  
Iowa State University  
Ames, IA 50011 USA  
email: swillson@iastate.edu

November 30, 2004

Proposed Running Head: Minimum evolution is less robust  
FAX: (515) 294-5454

Abstract: The method of minimum evolution reconstructs a phylogenetic tree  $T$  for  $n$  taxa given dissimilarity data  $d$ . In principle for every tree  $W$  with these  $n$  leaves an estimate for the total length of  $W$  is made, and  $T$  is selected as the  $W$  that yields the minimum total length. Suppose that the ordinary least square formula  $S_W(d)$  is used to estimate the total length of  $W$ . A theorem of Rzhetsky and Nei shows that when  $d$  is positively additive on a completely resolved tree  $T$ , then for all  $W \neq T$  it will be true that  $S_W(d) > S_T(d)$ . The same will be true if  $d$  is merely sufficiently close to an additive dissimilarity function. This paper proves that as  $n$  grows large, even if the shortest branch length in the true tree  $T$  remains constant and  $d$  is additive on  $T$ , then the difference  $S_W(d) - S_T(d)$  can go to zero. It is also proved that, as  $n$  grows large, there is a tree  $T$  with  $n$  leaves, an additive distance function  $d_T$  on  $T$  with shortest edge  $\epsilon$ , a distance function  $d$ , and a tree  $W$  with the same  $n$  leaves such that  $d$  differs from  $d_T$  by only approximately  $\epsilon/4$ , yet minimum evolution incorrectly selects the tree  $W$  over the tree  $T$ . This result contrasts with the method of neighbor-joining, for which Atteson showed that incorrect selection of  $W$  required a deviation at least  $\epsilon/2$ . It follows that, for large  $n$ , minimum evolution with ordinary least-squares can be only half as robust as neighbor-joining.

## 1 Introduction

Minimum evolution methods for identifying phylogenetic trees were first proposed by Kidd and Sgaramella-Zonta (1971) and independently by Rzhetsky

and Nei (1992a, 1992b). A recent review of the varieties, strengths, and limitations of such methods is given in Gascuel *et al.* (2001). Simulation studies detailing some shortcomings of the use of minimum evolution in reconstructing trees include Saitou and Imanishi (1989), Gascuel (2000), and Desper and Gascuel (2002).

Briefly, suppose that a dissimilarity function  $d(i, j)$  is given between each pair  $i$  and  $j$  in a set  $S$  of taxa, and assume that  $d$  is approximately additive. Some dissimilarity functions that are commonly used include the Jukes-Cantor formula (Jukes and Cantor 1969), the Kimura 2-parameter formula (Kimura 1980), the HKY formula (Hasegawa *et al.* 1985) and the log determinant formula (Lake 1994; Steel 1994). An excellent overview of dissimilarity formulas is found in Swofford *et al.* (1996).

In principle, given such a dissimilarity function  $d$ , for each tree  $W$  with the set  $S$  of leaves, one computes an estimate  $L(W)$  for the sum of the branch lengths of all edges in  $W$ , assuming that  $d$  arose from the tree  $W$ . The minimum evolution criterion selects the inferred phylogenetic tree  $T$  to be the tree  $W$  for which  $L(W)$  is minimal.

There are many formulas that can be used to estimate the sum  $L(W)$  of the branch lengths in a tree  $W$  from the dissimilarities  $d(i, j)$ . Rzhetsky and Nei (1993) focus on the natural choice in which the ordinary least-squares (OLS) estimate  $S_W(d)$  is utilized. Since the variance of a large distance  $d(i, j)$  is probably larger than the variance of a small distance  $d(i, j)$ , another natural approach is to estimate the variances and utilize weighted least-squares approximations to estimate  $L(W)$ . Recent algorithms for such weighted or generalized least-square estimates have been given by Felsenstein (1997), Makarenkov and Leclerc (1999), and Bryant and Waddell (1998).

Suppose that  $T$  is a completely resolved tree with the set  $S$  of leaves and  $W$  is a completely resolved tree with the same set  $S$  of leaves but  $W \neq T$ . Suppose that  $d_T$  is additive on  $T$  in such a manner that each edge  $e$  of  $T$  has positive branch length  $w(e) > 0$ . In this situation Rzhetsky and Nei (1993) proved that the ordinary least-square estimates  $S_T(d_T)$  for the length of  $T$  and  $S_W(d_T)$  for the length of  $W$  necessarily satisfy  $S_W(d_T) > S_T(d_T)$ . This result guarantees the correctness of minimum evolution when  $d_T$  is in fact additive on  $T$ . A complete proof of the consistency of minimum evolution in a more generalized context appears in Denis and Gascuel (2003).

In practice, a measured dissimilarity function  $d$  is rarely additive on any tree. If  $d$  is “sufficiently close” to such an additive function  $d_T$  on  $T$ , then continuity shows that  $S_W(d) > S_T(d)$  will also hold. An issue is what constitutes being “sufficiently close”. This paper contains theorems concerning how close  $d$  must be to an additive function  $d_T$  so that it is guaranteed that  $S_W(d) > S_T(d)$ .

This paper presents examples showing that  $S_W(d) - S_T(d)$  can become arbitrarily small as the number  $n$  of leaves grows, even if  $d$  is additive on  $T$  and the shortest branch length is bounded below. More precisely, we obtain the following:

**Theorem 1.1.** *For every  $n \geq 6$ , there is a tree  $T$  with  $n$  leaves, an additive*

dissimilarity function  $d$  on  $T$  with minimum branch length  $\epsilon$  and a tree  $W$  different from  $T$  but with the same  $n$  leaves such that

$$S_W(d) - S_T(d) \leq \frac{4(n-2)\epsilon}{n^2-1}.$$

The resolution of the correct tree in the presence of errors is often hard. By Theorem 1.1, as the number  $n$  of taxa increases, one is forced to distinguish not by  $\epsilon$  but by  $\epsilon O(1/n)$ . Hence the required distinction grows small purely as a matter of geometry, making the problem even harder.

Suppose that  $d_T$  is additive on  $T$ . Consider the problem of using minimum evolution with OLS approximations to determine  $T$  from a dissimilarity function  $d$  which is an empirical estimate of  $d_T$ . A basic concern is when the use of OLS approximations can lead to the selection of an incorrect tree  $W$  because  $S_W(d) < S_T(d)$ . Let  $\|d - d_T\|_\infty$  denote the largest value of  $|d(i, j) - d_T(i, j)|$  as  $i$  and  $j$  range over the taxa. The following result shows that  $\|d - d_T\|_\infty$  can be disappointingly small and lead to the selection of an incorrect tree:

**Theorem 1.2.** *For every  $n \geq 6$ , there is a tree  $T$  with  $n$  leaves, an additive dissimilarity function  $d_T$  on  $T$  with minimum branch length  $\epsilon$ , a dissimilarity function  $d$ , and a tree  $W$  with the same  $n$  leaves such that*

$$\|d - d_T\|_\infty \leq \epsilon \frac{n^2 - 1}{4(n^2 - 4n + 8)}$$

but  $S_W(d) < S_T(d)$ .

Atteson (1999) defines the *edge  $l_\infty$  radius* for a method of tree reconstruction to be the largest number  $\alpha$  such that, whenever  $d_T$  is additive on a tree  $T$  with shortest internal branch length  $s(d_T)$  and  $d$  is a dissimilarity function such that, for all  $i$  and  $j$ ,  $|d(i, j) - d_T(i, j)| < \alpha s(d_T)$ , then the method selects  $T$  as the correct tree. Thus the edge  $l_\infty$  radius is a measure of robustness of the method. Atteson shows that the largest possible edge  $l_\infty$  radius is  $1/2$ . He also shows that the commonly used method of neighbor-joining due to Saitou and Nei (1987) has edge  $l_\infty$  radius  $1/2$ .

Let  $l_\infty(n)$  denote the edge  $l_\infty$  radius of the method in which minimum evolution is utilized to select a tree given a distance function on  $n$  taxa, where lengths are estimated using OLS formulas. Theorem 1.2 implies the following result:

**Theorem 1.3.**  $\limsup_{n \rightarrow \infty} l_\infty(n) \leq 1/4$ .

Consequently, one may say that for large numbers of taxa, minimum evolution using OLS estimates can be only half as robust as neighbor-joining. Briefly, Theorem 1.2 says that, for large  $n$ , when the correct tree  $T$  has shortest branch length  $\epsilon$ , then use of minimum evolution can lead to the incorrect selection of the tree  $W$  over the tree  $T$  when  $\|d - d_T\|_\infty$  is approximately  $\epsilon/4$ . In contrast, for neighbor-joining, Atteson shows that incorrect selection of  $W$  requires that this deviation be at least  $\epsilon/2$ .

The results of this paper apply only to the use of ordinary least-square (OLS) estimates. In fact, Gascuel *et al.* (2001) give an example where weighted least-squares estimates  $f_W(d)$  for trees  $W$  are utilized, where the dissimilarity function  $d_T$  is additive on the tree  $T$ , but for which there exists a tree  $W$  satisfying  $f_W(d_T) < f_T(d_T)$ . Hence the method of minimum evolution is not guaranteed to work in principle using weighted least-square estimates even for additive dissimilarity functions.

## 2 Preliminaries

A good introduction to tree inference using distance methods may be found in Semple and Steel (2003). Let  $S = \{1, 2, \dots, n\}$  denote a set of distinct taxa. A *dissimilarity function* on  $S$  is a function  $d : S \times S \rightarrow R$  such that  $d(i, i) = 0$  for all  $i$  in  $S$  and  $d(i, j) = d(j, i)$  for all  $i$  and  $j$  in  $S$ . The collection of dissimilarity functions on  $S$  will be denoted  $D(S)$ .

A *split*  $A|B$  of  $S$  is a partition of  $S$  into two disjoint nonempty subsets  $A$  and  $B$ ; thus  $A|B = \{A, B\}$  where  $A \cup B = S$  and  $A \cap B = \emptyset$ . Given any edge  $e$  of a tree  $T$ , removal of the edge  $e$  (but not its endpoints) disconnects  $T$  into two components. Let  $A$  and  $B$  denote the sets of leaves from  $S$  in the two different components, and call  $A|B = B|A$  the *split corresponding* to  $e$ . We say that the split *separates*  $i$  and  $j$  if either (1)  $i \in A$  and  $j \in B$ , or (2)  $i \in B$  and  $j \in A$ .

A *branch length function*  $w$  for the tree  $T$  is a function which assigns to each edge  $e$  of  $T$  a real number  $w(e)$ , called its *length* or *branch length*. The branch length function  $w$  is *positive* if for every edge  $e$  it is true that  $w(e) > 0$ . Any branch length function  $w$  for a tree  $T$  with leaves labelled by  $S$  leads to a dissimilarity function  $d$  on  $S$  such that if  $i$  and  $j$  are in  $S$  and  $P_{ij}$  is the unique path in  $T$  between nodes  $i$  and  $j$ , then  $d(i, j)$  is the sum of the lengths of edges on  $P_{ij}$ :  $d(i, j) = \Sigma\{w(e) : e \text{ is on } P_{ij}\}$ , where for a multiset  $U$  of real numbers,  $\Sigma U$  denotes the sum of the elements in  $U$ . Say that a dissimilarity function  $d$  is *additive* on  $T$  if there exists a branch length function  $w$  on  $T$  such that  $d$  is obtained in this manner. Say that  $d$  is *positively additive on*  $T$  if in addition  $w$  is positive. The *total length* or *treelength*  $L$  of  $T$  is

$$L(T; d) = L(T) = \Sigma\{w(e) : e \text{ is an edge of } T\}.$$

Given a split  $A|B$  of the set  $S$  of leaves, define a dissimilarity function  $d_{A|B}$  by

$$d_{A|B}(i, j) = \begin{cases} 1 & \text{if } A|B \text{ separates } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}$$

Given a labelled tree  $T$  and an edge  $e$  of  $T$ , form a branch length function  $w$  by  $w(e) = 1$  and  $w(e') = 0$  for all edges  $e'$  of  $T$  distinct from  $e$ . Let  $d_e$  denote the corresponding additive dissimilarity function, so

$$d_e(i, j) = \begin{cases} 1 & \text{if } e \text{ lies on the path } P_{ij} \text{ from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

Note that  $L(T; d_e) = 1$ . It is clear that the functions  $d_e$  as  $e$  ranges over all edges of  $T$  form a basis for the vector space of additive dissimilarity functions on  $T$ . For any additive dissimilarity function  $d$  with the branch length function  $w(e)$  it follows that

$$d = \Sigma\{w(e)d_e : e \text{ is an edge of } T\}.$$

**Lemma 2.1.** *Suppose  $A|B$  denotes the split of  $S$  that corresponds to the edge  $e$  of  $T$ . Then  $d_e = d_{A|B}$ .*

*Proof.* For each  $i$  and  $j$ , the path  $P_{ij}$  includes the edge  $e$  if and only if  $A|B$  separates  $i$  and  $j$ .  $\square$

Every linear functional  $f : D(S) \rightarrow R$  has the form  $f(d) = \Sigma c(i, j)d(i, j)$  where  $d$  is a dissimilarity function, the summation is over all pairs  $(i, j)$  with  $i$  and  $j$  in  $S$ , and for each  $i$  and  $j$ ,  $c(i, j)$  is a real constant. Since  $d(i, i) = 0$  we will assume  $c(i, i) = 0$  for each  $i$  in  $S$ ; since  $d(i, j) = d(j, i)$  we will assume that for each pair  $\{i, j\}$  of distinct elements of  $S$  at least one of  $c(i, j)$  or  $c(j, i)$  is 0. For example, we may write  $f(d) = \Sigma\{c(i, j)d(i, j) : i \in S, j \in S, i < j\}$ .

Given a labelled tree  $T$  with the set  $S$  of leaves and a dissimilarity function  $d$ , the ordinary least-squares (OLS) formula  $S_T(d)$  is a linear expression that estimates the total length of  $T$  given  $d$ . Briefly, for each edge  $e$  of  $T$ , an estimate  $\hat{w}(e) = S_T(d, e)$  is made for the branch length  $w(e)$  of  $e$  using OLS; then  $S_T(d) = \Sigma\{S_T(d, e) : e \text{ is an edge of } T\}$ . See, for example, Rzhetsky and Nei (1992ab, 1993) or Gascuel *et al.* (2001).

It is well known that whenever  $d$  is additive on  $T$ , then  $S_T(d)$  tells the correct value of the total length of  $T$ . More specifically, if

$$d = \Sigma\{w(e)d_e : e \text{ is an edge of } T\},$$

then  $S_T(d) = \Sigma\{w(e) : e \text{ is an edge of } T\}$ . Indeed, in this situation,  $S_T(d, e) = w(e)$  for each edge  $e$ .

### 3 Estimates for errors in measured lengths

The minimum evolution criterion is based on comparison of the estimated lengths of trees; hence it is desirable to study the errors in these measurements. We shall typically study the situation where  $T$  denotes the “true tree” and  $W$  denotes the “wrong tree.”

Let  $T$  and  $W$  denote two different completely resolved trees for the same set  $S$  of  $n$  taxa. Let the OLS formulas be written  $S_T(d) = \Sigma c(i, j; T)d(i, j)$  and  $S_W(d) = \Sigma c(i, j; W)d(i, j)$ . Of special interest is the difference between the estimates:  $S_W(d) - S_T(d)$ .

**Lemma 3.1.** *Suppose that  $d$  is additive on  $T$ . For each edge  $e$  of  $T$ , let  $w(e)$  be the branch length and let  $e$  correspond to the split  $s(e) = A_e|B_e$  of  $S$ . Then (1)  $S_W(d) - S_T(d) = \Sigma w(e)(S_W(d_{A_e|B_e}) - 1)$ .*

- (2) For each  $e$ ,  $S_W(d_{A_e|B_e}) - 1 \geq 0$ .  
(3) If  $A|B$  is a split of  $W$ , then  $S_W(d_{A|B}) - 1 = 0$ .  
(4) If the edge  $e$  of  $T$  is the edge to a leaf  $j$ , then  $S_W(d_{A_e|B_e}) - 1 = 0$ .

*Proof.* Since  $d$  is additive,  $d = \sum w(e)d_e$ , whence  $S_W(d) = \sum\{w(e)S_W(d_e)\} = \sum w(e)S_W(d_{A_e|B_e})$  by Lemma 2.1. Since  $S_T$  gives the correct total length of the tree when  $d$  is additive on  $T$ , it follows that  $S_T(d) = \sum w(e)$ . Hence

$$S_W(d) - S_T(d) = \sum w(e)(S_W(d_{A_e|B_e}) - 1).$$

By Rzhetsky and Nei (1993), when each  $w(e) > 0$ , it follows that  $S_W(d) > S_T(d)$ . Hence  $S_W(d) - S_T(d) > 0$  for all choices of positive branch lengths, and so for each  $e$ ,  $S_W(d_{A_e|B_e}) - 1 \geq 0$ . If  $A_e|B_e$  happens to be a split of  $W$  as well as  $T$  corresponding to some edge of  $W$ , then  $S_W(d_{A_e|B_e}) = 1$  since  $d_{A_e|B_e}$  is then additive on  $W$  with total length 1 and  $S_W$  tells the correct length for dissimilarity functions additive on  $W$ . Hence in this case  $S_W(d_{A_e|B_e}) - 1 = 0$ . If  $e$  is the edge in  $T$  to a leaf  $j$ , then the split  $A_e|B_e$  is  $\{j\}|(S - \{j\})$  which is also a split of  $W$ , so (4) follows from (3).  $\square$

Suppose that  $d$  is additive on  $T$  with nonnegative branch length function  $w(e)$ . Suppose  $W$  is any tree with the same set of leaves. From Lemma 3.1 it follows that

$$S_W(d) - S_T(d) = \sum\{u(e)w(e) : e \text{ is an edge of } T\}$$

where

$$u(e) = S_W(d_{A_e|B_e}) - 1 \geq 0$$

and the summation may be taken over all internal edges  $e$  of  $T$  since all other edges  $e$  satisfy  $u(e) = 0$  by 3.1(4). Define

$$M(T, W) = \sum\{u(e) : e \text{ is an edge of } T\}.$$

Related quantities are

$$M(T) = \min\{M(T, W) : W \neq T\}$$

and

$$M(n) = \min\{M(T) : T \text{ has } n \text{ leaves}\}$$

There is an easier computation of  $M(T, W)$  as follows: Given the tree  $T$ , let  $d'_T$  denote the additive dissimilarity function on  $S$  such that  $w(e) = 1$  for each edge of  $T$ . Thus  $d'_T(i, j)$  is the number of edges on the path  $P_{ij}$  in  $T$  connecting vertices  $i$  and  $j$ .

**Lemma 3.2.** *Assume that the trees  $T$  and  $W$ ,  $T \neq W$ , have the same  $n$  leaves.*

- (1)  $M(T, W) = S_W(d'_T) - S_T(d'_T)$ .  
(2)  $M(T, W) > 0$ .

*Proof.* If  $w(e_i) = 1$  for all  $i$ , then  $S_W(d'_T) - S_T(d'_T) = \sum u(e_i)$ . Since  $d'_T$  is additive on  $T$  and each branch length is positive, Rzhetsky and Nei (1993) showed  $S_W(d'_T) > S_T(d'_T)$ , whence  $M(T, W) > 0$ .  $\square$

If  $d$  is additive on  $T$  with branch lengths  $w(e)$ , let the *shortest internal edge length* be  $s = s(d) = \min\{w(e) : e \text{ is an internal edge of } T\}$ . The following result shows that  $S_W(d) - S_T(d) \geq M(T, W)s(d)$  and equality is possible. Hence  $M(T, W)$  measures the minimum separation between  $S_W(d)$  and  $S_T(d)$  when the shortest branch length in  $T$  is controlled.

**Theorem 3.3.** *Suppose that  $d$  is additive on the completely resolved tree  $T$  via the positive branch length function  $w(e)$ . Let  $s(d)$  be the shortest internal edge length for  $T$ . Suppose  $W \neq T$  is a tree with the same set of leaves as  $T$ , and  $S_T$  and  $S_W$  are the OLS functionals for the trees  $T$  and  $W$  respectively. Then*

- (1)  $S_W(d) - S_T(d) \geq M(T, W)s(d)$ .
- (2) *There exists an additive dissimilarity function  $d$  on  $T$  such that  $S_W(d) - S_T(d) = M(T, W)s(d)$ .*

*Proof.*  $S_W(d) - S_T(d) = \sum u(e)w(e) \geq \sum u(e)s(d)$  [since  $u(e) \geq 0$  by Lemma 3.1]  $= M(T, W)s(d)$ . Note that equality is attained if  $w(e) = s(d)$  for each edge  $e$  with  $u(e) > 0$ .  $\square$

Suppose  $S_T(d) = \sum c(i, j; T)d(i, j)$  and  $S_W(d) = \sum c(i, j; W)d(i, j)$ , so  $S_W(d) - S_T(d) = \sum [c(i, j; W) - c(i, j; T)]d(i, j)$ . Define

$$K(T, W) = \sum \{|c(i, j; W) - c(i, j; T)| : i < j\}.$$

The next lemma shows that a perturbation of an additive dissimilarity function by an amount  $\epsilon$  can change  $|S_W(d) - S_T(d)|$  by at most  $K(T, W)\epsilon$ , and equality is possible.

**Lemma 3.4.** *Suppose that  $d_T$  is positively additive on the completely resolved tree  $T$  with shortest internal edge length  $s(d_T)$ , and  $\epsilon > 0$ . Suppose  $W$  is a tree with the same set  $S$  of leaves,  $W \neq T$ .*

- (1) *Suppose that  $d$  is a dissimilarity function (not necessarily additive on  $T$ ) such that for all  $i$  and  $j$ ,  $|d(i, j) - d_T(i, j)| < \epsilon$ . Then*

$$S_W(d) - S_T(d) > S_W(d_T) - S_T(d_T) - K(T, W)\epsilon.$$

- (2) *There exists a dissimilarity function  $d$  such that for all  $i$  and  $j$ ,  $|d(i, j) - d_T(i, j)| \leq \epsilon$  and*

$$S_W(d) - S_T(d) = S_W(d_T) - S_T(d_T) - K(T, W)\epsilon$$

*Proof.* For (1),  $S_W(d) - S_T(d) - (S_W(d_T) - S_T(d_T))$   
 $= S_W(d - d_T) - S_T(d - d_T)$   
 (since both  $S_W$  and  $S_T$  are linear)  
 $= \sum [c(i, j; W) - c(i, j; T)][d(i, j) - d_T(i, j)]$ .  
 Hence

$|S_W(d - d_T) - S_T(d - d_T)| \leq \Sigma |c(i, j; W) - c(i, j; T)| |d(i, j) - d_T(i, j)|$   
 $< \Sigma |c(i, j; W) - c(i, j; T)| \epsilon = K(T, W) \epsilon.$   
 Now  $S_W(d) - S_T(d) = (S_W(d_T) - S_T(d_T)) + [S_W(d - d_T) - S_T(d - d_T)]$   
 $\geq (S_W(d_T) - S_T(d_T)) - |S_W(d - d_T) - S_T(d - d_T)|$   
 $> (S_W(d_T) - S_T(d_T)) - K(T, W) \epsilon \geq M(T, W) s(d_T) - K(T, W) \epsilon$   
 by Theorem 3.3. This proves (1).

For (2), define

$$d(i, j) = \begin{cases} d_T(i, j) - \epsilon & \text{if } c(i, j; W) - c(i, j; T) \geq 0 \\ d_T(i, j) + \epsilon & \text{if } c(i, j; W) - c(i, j; T) < 0. \end{cases}$$

Then for all  $i$  and  $j$  it follows

$$[c(i, j; W) - c(i, j; T)] [d(i, j) - d_T(i, j)] = -|c(i, j; W) - c(i, j; T)| \epsilon.$$

Hence

$$S_W(d) - S_T(d) - (S_W(d_T) - S_T(d_T)) = -\Sigma |c(i, j; W) - c(i, j; T)| \epsilon = -K(T, W) \epsilon$$

so

$$S_W(d) - S_T(d) = (S_W(d_T) - S_T(d_T)) - K(T, W) \epsilon.$$

□

**Corollary 3.5.** *Suppose that  $d_T$  is additive on  $T$  such that for all edges  $e$  of  $T$  the branch length satisfies  $w(e) = s(d_T)$ . Let  $\epsilon > 0$ . Then there exists a dissimilarity function  $d$  such that for all  $i$  and  $j$ ,  $|d(i, j) - d_T(i, j)| \leq \epsilon$  and*

$$S_W(d) - S_T(d) = M(T, W) s(d_T) - K(T, W) \epsilon.$$

*Proof.* For the hypothesized  $d_T$ , let  $d$  be as in 3.4(2). Then  $(S_W(d_T) - S_T(d_T)) = M(T, W) s(d_T)$  by Lemma 3.2, and the result follows. □

From Corollary 3.5, the balance between the quantities  $M(T, W)$  and  $K(T, W)$  helps to control the possible sign of  $S_W(d) - S_T(d)$ . Define

$$R(T, W) = M(T, W) / K(T, W).$$

Related quantities are

$R(T) = \min\{R(T, W) : W \text{ is a tree with the same } n \text{ leaves as } T, \text{ but } W \neq T\}$   
 and  $R(n) = \min\{R(T, W) : T \text{ and } W \text{ have the same } n \text{ leaves, } T \neq W\}.$

Suppose that the additive dissimilarity function  $d_T$  on the tree  $T$  is perturbed to a new dissimilarity function  $d$ . The following theorem shows that if the perturbation is less than  $R(T, W) s(d_T)$ , then it will still be guaranteed that  $S_W(d) - S_T(d) > 0$ ; while if the perturbation equals  $R(T, W) s(d_T)$  then it is possible that instead  $S_W(d) - S_T(d) \leq 0$ . Thus  $R(T, W)$  gives information about how much an additive dissimilarity function  $d_T$  on the tree  $T$  can be perturbed to a new dissimilarity function  $d$  without losing the basic inequality  $S_W(d) - S_T(d) > 0$ . It thus measures the robustness of the method, while controlling for short branch lengths.

**Theorem 3.6.** *Let  $T$  and  $W$  be distinct completely resolved trees with the same set of leaves.*

(1) *Let  $d_T$  be positively additive on  $T$  with shortest internal branch length  $s(d_T)$ . Suppose that  $d$  is any dissimilarity function satisfying that for all  $i$  and  $j$ ,  $|d(i, j) - d_T(i, j)| < R(T, W)s(d_T)$ . Then  $S_W(d) - S_T(d) > 0$ .*

(2) *There exists an additive dissimilarity function  $d_T$  on  $T$  with shortest internal branch length  $s(d_T)$  and a dissimilarity function  $d$  satisfying that for all  $i$  and  $j$ ,  $|d(i, j) - d_T(i, j)| \leq R(T, W)s(d_T)$  such that  $S_W(d) - S_T(d) = 0$ .*

(3) *For any  $\epsilon > 0$ , there exists an additive dissimilarity function  $d_T$  on  $T$  with shortest internal branch length  $s(d_T)$  and there exists a dissimilarity function  $d$  such that for all  $i$  and  $j$ ,  $|d(i, j) - d_T(i, j)| \leq (R(T, W) + \epsilon)s(d_T)$  and such that  $S_W(d) - S_T(d) < 0$ .*

*Proof.* For (1), by Lemma 3.4(1)

$$\begin{aligned} S_W(d) - S_T(d) &> S_W(d_T) - S_T(d_T) - K(T, W)R(T, W)s(d_T) \\ &\geq M(T, W)s(d_T) - K(T, W)R(T, W)s(d_T) \text{ (by Theorem 3.3)} \\ &= M(T, W)s(d_T) - M(T, W)s(d_T) = 0. \end{aligned}$$

For (2), by Corollary 3.5 there exists an additive  $d_T$  and a dissimilarity function  $d$  such that for all  $i$  and  $j$ ,  $|d(i, j) - d_T(i, j)| \leq R(T, W)s(d_T)$  and  $S_W(d) - S_T(d) = M(T, W)s(d_T) - K(T, W)R(T, W)s(d_T) = M(T, W)s(d_T) - M(T, W)s(d_T) = 0$ .

For (3), let  $d_T$  be the additive dissimilarity function on  $T$  for which each edge has branch length  $s = s(d_T) > 0$ . By 3.5, replacing  $\epsilon$  by  $(R(T, W) + \epsilon)s$ , we see that there exists  $d$  such that for all  $i$  and  $j$ ,

$$\begin{aligned} |d(i, j) - d_T(i, j)| &\leq (R(T, W) + \epsilon)s, \text{ and} \\ S_W(d) - S_T(d) &= M(T, W)s - K(T, W)(R(T, W) + \epsilon)s \\ &= M(T, W)s - K(T, W)R(T, W)s - K(T, W)\epsilon s \\ &= M(T, W)s - M(T, W)s - K(T, W)\epsilon s \\ &= -K(T, W)\epsilon s < 0. \end{aligned} \quad \square$$

Given two dissimilarity functions  $d$  and  $d'$ , define the *l-infinity norm* or  $l_\infty$  norm by  $\|d - d'\|_\infty = \min\{|d(i, j) - d'(i, j)| : i \text{ and } j \text{ are taxa}\}$

**Corollary 3.7.** (1) *Suppose  $T$  is a completely resolved tree having  $n$  leaves and  $d_T$  is an additive dissimilarity function on  $T$  for which the shortest internal branch length is  $s(d_T)$ . If  $\|d - d_T\|_\infty < R(n)s(d_T)$ , then for all trees  $W \neq T$  with the same  $n$  leaves,  $S_W(d) > S_T(d)$ .*

(2) *There exists a completely resolved tree  $T$  with  $n$  leaves, a tree  $W$  with the same set of leaves, an additive dissimilarity function  $d_T$  on  $T$ , and a dissimilarity function  $d$  with  $\|d - d_T\|_\infty > R(n)s(d_T)$  for which  $S_W(d) < S_T(d)$ .*

Atteson (1999) defines the *edge  $l_\infty$  radius* for a method of tree reconstruction to be the largest number  $\alpha$  such that, whenever  $d_T$  is additive on a tree  $T$  with shortest internal branch length  $s$  and  $d$  is a dissimilarity function such that, for all  $i$  and  $j$ ,  $|d(i, j) - d_T(i, j)| < \alpha s$ , then the method selects  $T$  as the correct tree. Let  $l_\infty(n)$  denote the edge  $l_\infty$  radius for the reconstruction of phylogenetic trees with  $n$  leaves by the use of minimum evolution with OLS formulas to estimate the lengths of trees.

**Theorem 3.8.**  $l_\infty(n) = R(n)$ .

*Proof.* Suppose  $d_T$  is additive on the tree  $T$  with shortest internal branch length  $s$ . If  $|d(i, j) - d_T(i, j)| < R(T, W)s$ , then by Theorem 3.6(1),  $S_W(d) > S_T(d)$  so  $T$  will be preferred by minimum evolution over  $W$ . Hence for any  $T$ , if  $d_T$  is additive on  $T$  with minimal internal branch length  $s$ , and if  $|d(i, j) - d_T(i, j)| < R(n)s$  then  $T$  will be preferred over any  $W$  and the method will select the correct tree  $T$ . Consequently,  $l_\infty(n) \geq R(n)$ .

Conversely, suppose  $\alpha > R(n)$ . Choose  $T$  and  $W$  so  $R(T, W) = R(n)$ . By Theorem 3.6(3) there exists an additive dissimilarity function  $d_T$  on  $T$  with shortest internal edge length  $s$  and a dissimilarity function  $d$  satisfying that  $|d(i, j) - d_T(i, j)| \leq \alpha s$  but  $S_W(d) < S_T(d)$ . It follows that even though  $|d(i, j) - d_T(i, j)| < \alpha s$ , the method may not select  $T$  as the correct tree. Hence  $l_\infty(n) \leq \alpha$  whenever  $\alpha > R(n)$ . It follows that the edge  $l_\infty$  radius is exactly  $R(n)$ .  $\square$

## 4 Estimates for $M(n)$ and $R(n)$

Let  $t(n)$  denote the number of labelled trees with  $n$  leaves. It is well-known that  $t(n) = (3)(5)(7)\dots(2n-5)$ , so that  $t(n)$  grows rapidly as  $n$  increases. Note  $t(4) = 3, t(5) = 15, t(6) = 105, t(7) = 945, t(8) = 10395$ . For these values of  $n$ , for each topological type of tree  $T$  with  $n$  leaves, and for each tree  $W \neq T$  with  $n$  leaves, a computer was used to find  $M(T, W), K(T, W)$ , and  $R(T, W)$ . It was then easy to compute  $M(T)$  and  $R(T)$ . The results are shown in Tables 1 and 2.

$n$	$T$	$M(T)$	$W$
4*	((12)(34))	0.5	((13)(24))
5*	((12)(3(45)))	0.5	((13)(2(45)))
6*	((12)(3(4(56))))	0.444444	((12)(4(3(56))))
6	((12)((34)(56)))	0.555556	((12)(3(4(56))))
7*	((12)(3(4(5(67)))))	0.416667	((12)(3(5(4(67)))))
7	((12)(3((45)(67))))	0.5	((13)(2((45)(67))))
8*	((12)(3(4(5(6(78))))))	0.375	((12)(3(5(4(6(78))))))
8	((12)3((4(56))(78)))	0.5	((23)(1((4(56))(78))))
8	((12)(3(4((56)(78)))))	0.4	((12)(4(3((56)(78)))))
8	((12)((34)((56)(78))))	0.5	((12)((56)((34)(78))))

Table 1: The different topological types of tree  $T$  with  $n$  leaves,  $4 \leq n \leq 8$ . For each  $n$ , the table shows  $M(T)$  and a tree  $W$  which is closest to  $T$  in the sense that  $M(T, W) = M(T)$ . For each  $n$ , an asterisk marks the case where  $M(n) = M(T, W)$ .

For each topological type of tree  $T$  with  $n$  leaves,  $4 \leq n \leq 8$ , Table 1 shows  $M(T)$  and one choice of  $W$  such that  $M(T) = M(T, W)$ . Usually there are several choices of  $W$  that yield the same  $M(T)$ . Note that  $M(n)$  appears from

the table to decrease, at least for small  $n$ . It is noteworthy that different trees  $T$  with  $n$  leaves may have different values of  $M(T)$ .

Similarly, for each topological type of tree  $T$  with  $n$  leaves,  $4 \leq n \leq 8$ , Table 2 shows the value  $R(T)$  as well as a choice of  $W$  such that  $R(T, W) = R(T)$ . Note that  $R(n)$  appears to decrease as  $n$  increases.

$n$	$T$	Count	$R(T)$	$W$
4*	$((12)(34))$	3	0.5	$((13)(24))$
5*	$((12)(3(45)))$	15	0.5	$((12)(4(35)))$
6*	$((12)(3(4(56))))$	90	0.45	$((12)((34)(56)))$
6	$((12)((34)(56)))$	15	0.5	$(12)(3(4(56)))$
7*	$((12)(3(4(5(67))))$	630	0.428571	$((12)((34)(5(67))))$
7	$((12)(3((45)(67)))$	315	0.5	$((12)(3(4(5(67))))$
8*	$((12)(3(4(5(6(78))))$	5040	0.4	$((12)(3((45)(6(78))))$
8	$((12)(3(4((56)(78))))$	2520	0.416667	$((12)((34)((56)(78)))$
8	$((12)(3((4(56))(78)))$	2520	0.5	$((12)((3(4(56)))(78)))$
8	$((12)((34)((56)(78)))$	315	0.5	$((12)(3(4((56)(78))))$

Table 2: For each topological type of tree  $T$  with  $n$  leaves ( $4 \leq n \leq 8$ )  $R(T)$  is given, together with a tree  $W$  so  $R(T) = R(T, W)$ . An asterisk marks the pair yielding  $R(n)$ . Count tells the number of distinct trees of the same topological type as  $T$ .

It is noteworthy that not all trees  $T$  with  $n$  leaves have the same  $R(T)$ . Thus for  $n = 8$  there is a tree  $T_1$  with  $R(T_1) = 0.4$  and a tree  $T_2$  with  $R(T_2) = 0.5$ . As a consequence if it happens in a biological situation that the correct tree is  $T_1$  with shortest internal branch length  $s$  and additive dissimilarity  $d_{T_1}$ , then minimum evolution is guaranteed to work only when  $|d(i, j) - d_{T_1}(i, j)| < 0.4s$  for all  $i$  and  $j$ . By contrast, if the correct tree is  $T_2$  then minimum evolution is guaranteed to work when  $|d(i, j) - d_{T_2}(i, j)| < 0.5s$  for all  $i$  and  $j$ . Hence we expect greater tolerance of errors if  $T_2$  is the correct tree than if  $T_1$  is the correct tree.

For general  $n$ , this paper shows the following results:

**Theorem 4.1.** *For  $n \geq 6$  there exist labelled trees  $T^n$ ,  $W^n$ , and  $V^n$  with  $n$  leaves such that*

- (1)  $M(T^n, W^n) = 4(n-2)/n^2$  if  $n$  is even and  $M(T^n, W^n) = 4(n-2)/(n^2-1)$  if  $n$  is odd.
- (2)  $R(T^n, W^n) = 1/2$ .
- (3)  $M(T^n, V^n) = 1/2$ .
- (4)  $R(T^n, V^n) = (n^2)/[16 + 4(n-2)^2]$  if  $n$  is even and  $R(T^n, V^n) = (n^2-1)/[16 + 4(n-3)(n-1)]$  if  $n$  is odd.

The proof is deferred until section 5. The trees have the form given in Figures 1, 2, and 3, in which  $H$  and  $J$  represent the same graphs in all three figures

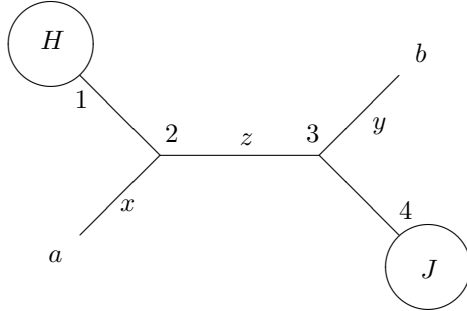


Figure 1: The tree  $T$ . In the calculations,  $d$  will be additive on  $T$ . The length of edge  $(2,3)$  is  $z$ .

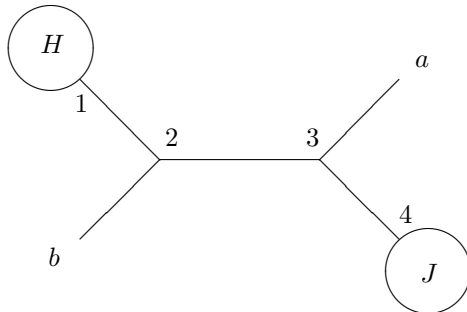


Figure 2: The tree  $W$ . Here  $H$  and  $J$  are the same as in tree  $T$  shown in Figure 1.

and there are some constraints on the number of leaves in each section. Theorem 1.1 follows immediately from (1) and Theorem 3.3. Theorem 1.2 follows immediately from (4) and Theorem 3.6 since  $\frac{n^2-1}{4(n^2-4n+8)}$  is less than or equal to both estimates in (4). The formulas in Theorem 4.1 have several interesting consequences.

**Theorem 4.2.**  $\lim_{n \rightarrow \infty} M(n) = 0$ .

*Proof.* From the definition it is clear that  $0 \leq M(n) \leq M(T^n, W^n)$ . As  $n$  goes to infinity, both expressions for  $M(T^n, W^n)$  in Theorem 4.1(1) go to 0. In fact,

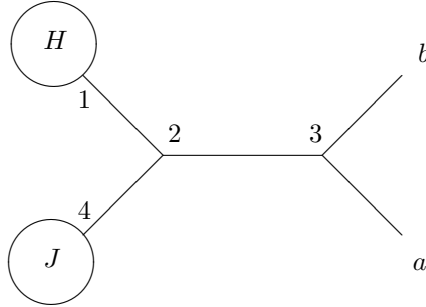


Figure 3: The tree  $V$ . Here  $H$  and  $J$  are the same as in tree  $T$  shown in Figure 1.

we see that  $M(T^n, W^n) = O(1/n)$ . □

**Theorem 4.3.**  $\limsup_{n \rightarrow \infty} R(n) \leq 1/4$ .

*Proof.* By the definition of  $R(n)$ ,  $0 \leq R(n) \leq R(T^n, V^n)$ . As  $n$  goes to infinity, both expressions for  $R(T^n, V^n)$  in Theorem 4.1(4) have limit  $1/4$ . □

Theorems 3.8 and 4.3 combine to yield Theorem 1.3.

Of the trees in the theorem, the pairs  $T^n$  and  $V^n$  appear hardest to distinguish because  $R(T^n, V^n)$  is small. Nevertheless  $M(T^n, V^n)$  remains constantly  $1/2$  while  $M(T^n, W^n)$  gets arbitrarily small. It is surprising that the trees which are closest together in terms of their dissimilarity estimates are not the trees that are hardest to distinguish in the presence of errors. Note also that for distinguishing particular families of trees (such as  $T^n$  and  $W^n$ ) the use of OLS minimum evolution is just as good as neighbor-joining since the robustness is 0.5. Similarly, Table 2 exhibits some trees with 6, 7, or 8 leaves for which the robustness is 0.5, so that again OLS minimum evolution would be as good as neighbor-joining if the correct tree is one of these.

It is noteworthy that the formulas for  $R(T^n, V^n)$  in Theorem 4.1 in fact yield the exact value of  $R(n)$  for  $4 \leq n \leq 8$ . The author does not know whether this is true for all  $n \geq 4$ .

## 5 Proof of Theorem 4.1

Figure 1 shows, for  $n \geq 4$  a tree  $T$  with  $n$  leaves. Here  $a$  and  $b$  are leaves; the edge joining  $a$  to  $T$  is  $(2, a)$  and the edge joining  $b$  to  $T$  is  $(3, b)$ .  $H$  represents a rooted tree containing  $h$  leaves, whose root is at the vertex 1. Similarly  $J$

represents a rooted tree containing  $j$  leaves with root at 4. Assume  $h$  and  $j$  are at least one and  $h + j = n - 2$  so that there are exactly  $n$  leaves. If, for example,  $h = 1$ , then  $H$  is identically the same as a single leaf 1. If  $h = 2$  with leaves  $p$  and  $q$ , then there are edges  $(1, p)$  and  $(1, q)$ . There will be an additive distance function on  $T$ , and the lengths of three edges are indicated as  $x$ ,  $y$ , and  $z$ . Figure 2 shows a corresponding tree  $W$ , also with  $n$  leaves. The trees  $H$  and  $J$  are identical with those in  $T$ . The only difference between  $T$  and  $W$  is that the leaves  $a$  and  $b$  have been interchanged. Similarly, Figure 3 shows a tree  $V$  in which  $H$  and  $J$  are identical with those in  $T$  and  $W$ . The trees in Theorem 4.1 are these trees with some additional constraints that make  $h$  and  $j$  approximately  $n/2$ .

The heart of the calculations will be a formula of Desper and Gascuel (2002). Suppose an arbitrary tree  $T$  has the clusters  $A$ ,  $B$ ,  $C$ , and  $D$  arranged as  $T = ((AB)(CD))$ , as shown in Figure 4. Let  $|X|$  denote the number of leaves in  $X$ , and write  $d_{ij}$  for  $d(i, j)$ . If  $U$  and  $V$  are disjoint subsets of leaves from the tree, the average distance between  $U$  and  $V$  is defined as

$$\Delta_{U|V} = \frac{1}{|U||V|} \sum_{i \in U, j \in V} d_{ij}.$$

**Lemma 5.1.** *Suppose that  $T$  is the tree in Figure 4 and  $W$  is the tree from Figure 4 in which  $B$  and  $C$  have been swapped. Then*

$$S_T(d) - S_W(d) = (1/2)[(\lambda - 1)(\Delta_{A|C} + \Delta_{B|D}) - (\lambda' - 1)(\Delta_{A|B} + \Delta_{C|D}) - (\lambda - \lambda')(\Delta_{A|D} + \Delta_{B|C})]$$

where

$$\lambda = \frac{|A||D| + |B||C|}{(|A| + |B|)(|C| + |D|)}$$

and

$$\lambda' = \frac{|A||D| + |B||C|}{(|A| + |C|)(|B| + |D|)}.$$

*Proof.* Desper and Gascuel (2002) prove this result as their formula (9).  $\square$

**Lemma 5.2.** *Suppose that  $T$  and  $W$  are the trees of Figures 1 and 2 respectively. Assume  $d$  is additive on  $T$  with branch length function  $w$ , and  $z = w((2, 3))$ . Then*

$$S_W(d) - S_T(d) = z \frac{h + j}{(h + 1)(j + 1)}$$

and

$$M(T, W) = \frac{h + j}{(h + 1)(j + 1)}$$

where  $h = |H|$  and  $j = |J|$ .

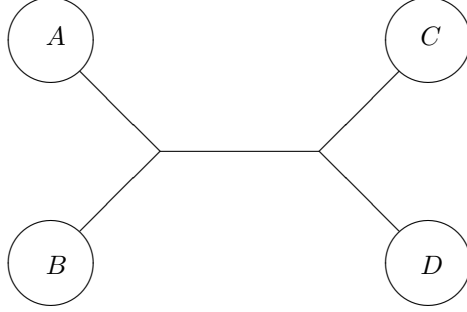


Figure 4: The arrangement of a tree  $T$  for the formula of Desper and Gascuel 2002.

*Proof.*  $W$  is obtained from  $T$  by interchanging  $a$  and  $b$ . Hence by 5.1

$$S_T(d) - S_W(d) = (1/2)[(\lambda-1)(\Delta_{H|b} + \Delta_{a|J}) - (\lambda'-1)(\Delta_{H|a} + \Delta_{b|J}) - (\lambda-\lambda')(\Delta_{H|J} + \Delta_{a|b})]$$

where

$$\lambda = \frac{hj+1}{(h+1)(j+1)} = \lambda'.$$

$$\begin{aligned} \text{Now } S_T(d) - S_W(d) &= (1/2)[(\lambda-1)(\Delta_{H|b} + \Delta_{a|J}) - (\lambda-1)(\Delta_{H|a} + \Delta_{b|J})] \\ &= (1/2)(\lambda-1)(\Delta_{H|b} - \Delta_{H|a} + \Delta_{a|J} - \Delta_{b|J}). \end{aligned}$$

Moreover, since the distance function is additive on  $T$  it is clear from Figure 1 that  $\Delta_{H|b} = \Delta_{H|a} + y + z - x$  while  $\Delta_{J|a} = \Delta_{J|b} + z + x - y$ . Hence

$$\begin{aligned} S_T(d) - S_W(d) &= (1/2)(\lambda-1)((y+z-x) + (z+x-y)) \\ &= z(\lambda-1) = z \frac{-h-j}{(h+1)(j+1)}. \end{aligned}$$

The formula for  $M(T, W)$  then follows from Lemma 3.2.  $\square$

**Lemma 5.3.** *Suppose that  $T$  and  $V$  are the trees of Figures 1 and 3 respectively. Assume  $d$  is additive on  $T$  with branch length function  $w$ , and  $z = w((2, 3))$ . Then  $S_V(d) - S_T(d) = z/2$  and  $M(T, V) = 1/2$ .*

*Proof.*  $V$  is obtained from  $T$  by interchanging  $a$  and  $J$ . Hence by 5.1

$$S_T(d) - S_V(d) = (1/2)[(\lambda-1)(\Delta_{H|J} + \Delta_{a|b}) - (\lambda'-1)(\Delta_{H|a} + \Delta_{J|b}) - (\lambda-\lambda')(\Delta_{H|b} + \Delta_{a|J})]$$

where

$$\lambda = \frac{h+j}{(h+1)(j+1)}$$

and

$$\lambda' = \frac{h+j}{(h+j)(1+1)} = 1/2.$$

Since  $d$  is additive on  $T$ , write  $\Delta_{H|2}$  for the average distance in  $T$  from a member of  $H$  to vertex 2 in  $T$ , while similarly write  $\Delta_{J|3}$  for the average distance from a member of  $J$  to vertex 3 in  $T$ . From Figure 1 it is clear that

$$\Delta_{H|a} = \Delta_{H|2} + x,$$

$$\Delta_{H|b} = \Delta_{H|2} + y + z,$$

$$\Delta_{J|a} = \Delta_{J|3} + x + z,$$

$$\Delta_{J|b} = \Delta_{J|3} + y,$$

$$\Delta_{H|J} = \Delta_{H|2} + \Delta_{J|3} + z,$$

and  $\Delta_{a|b} = x + y + z$ . Hence

$$\begin{aligned} S_T(d) - S_V(d) &= (1/2)[(\lambda - 1)(\Delta_{H|2} + \Delta_{J|3} + z + (x + y + z)) \\ &\quad - (\lambda' - 1)(\Delta_{H|2} + x + \Delta_{J|3} + y) - (\lambda - \lambda')(\Delta_{H|2} + y + z + \Delta_{J|3} + x + z)] \\ &= z(\lambda' - 1) = -z/2. \end{aligned}$$

That  $M(T, V) = 1/2$  follows from Lemma 3.2.  $\square$

**Lemma 5.4.** *Let  $T$  and  $W$  be as in Figures 1 and 2 respectively. Then*

$$K(T, W) = \frac{2(h+j)}{(h+1)(j+1)}$$

and  $R(T, W) = 1/2$ .

*Proof.* From the proof of 5.2 we have

$$S_T(d) - S_W(d) = (1/2)[(\lambda - 1)(\Delta_{H|b} + \Delta_{a|J}) - (\lambda' - 1)(\Delta_{H|a} + \Delta_{b|J}) - (\lambda - \lambda')(\Delta_{H|J} + \Delta_{a|b})]$$

where

$$\lambda = \lambda' = \frac{hj+1}{(h+1)(j+1)}.$$

Clearly  $\lambda < 1$ . Each of the  $|H|$  terms  $d_{hb}$  for  $h \in H$  appears only in  $\Delta_{H|b}$  and hence has coefficient  $(1/2)(\lambda - 1)/|H|$ . Each such term then contributes  $(1/2)(1 - \lambda)/|H|$  to  $K(T, W)$ . Since there are  $|H|$  such terms, all the terms in  $\Delta_{H|b}$  contribute  $(1/2)(1 - \lambda)$  to  $K(T, W)$ . Similarly each term  $d_{ha}$  for  $h \in H$  appears only in  $\Delta_{H|a}$  and with coefficient  $(1/2)(-\lambda' - 1)/|H|$ . Since there are  $|H|$  such terms, all of them contribute  $(1/2)(-\lambda' - 1)$  to  $K(T, W)$ . In this manner we see that

$$\begin{aligned} K(T, W) &= (1/2)[(1 - \lambda)(1 + 1) - (\lambda' - 1)(1 + 1)] \\ &= 2 - 2\lambda = 2 \frac{h+j}{(h+1)(j+1)} \end{aligned}$$

From 5.2,  $R(T, W) = M(T, W)/K(T, W) = 1/2$ .  $\square$

**Lemma 5.5.** *Suppose that  $T$  and  $V$  are the trees of Figures 1 and 3 respectively. Assume  $d$  is additive on  $T$  with branch length function  $w$ , and  $z = w((2, 3))$ . Assume  $h \geq 2$  and  $j \geq 2$ . Then*

$$K(T, V) = \frac{2(hj + 1)}{(h + 1)(j + 1)}$$

and

$$R(T, V) = \frac{(h + 1)(j + 1)}{4(hj + 1)}.$$

*Proof.* From the proof of 5.3,

$$S_T(d) - S_V(d) = (1/2)[(\lambda - 1)(\Delta_{H|J} + \Delta_{a|b}) - (\lambda' - 1)(\Delta_{H|a} + \Delta_{J|b}) - (\lambda - \lambda')(\Delta_{H|b} + \Delta_{a|J})]$$

where

$$\lambda = \frac{h + j}{(h + 1)(j + 1)}$$

and

$$\lambda' = \frac{h + j}{(h + j)(1 + 1)} = 1/2.$$

Clearly  $\lambda < 1$ , so  $\lambda - 1 < 0$ . Moreover, if  $h \geq 2$  and  $j \geq 2$  it is easy to see that  $hj + 1 > h + j$ , so that  $2(h + j) = 2h + 2j < h + j + hj + 1 = (h + 1)(j + 1)$  and  $\lambda < 1/2$ . Hence  $-(\lambda - \lambda') > 0$ . Trivially,  $-(\lambda - 1) > 0$ .

Each term  $d_{ij}$  for  $i \in H$  and  $j \in J$  appears in this formula for  $S_T(d) - S_V(d)$  only in the term  $\Delta_{H|J}$  hence with coefficient  $(1/2)(\lambda - 1)/(|H||J|)$ . The absolute value of this coefficient is therefore  $(1/2)(-\lambda - 1)/(|H||J|)$ . Since there are  $|H||J|$  such choices for  $i$  and  $j$ , it follows that the contribution of all such terms to  $K(T, V)$  is  $(1/2)[-(\lambda - 1)(1)]$ . In like manner the contribution of all terms  $d_{ia}$  with  $i \in H$  arise from  $\Delta_{H|a}$  and each such term has coefficient  $(1/2)[-(\lambda' - 1)/|H|]$ ; since  $-(\lambda' - 1) > 0$  and there are  $|H|$  such terms, the contribution of all such terms to  $K(T, V)$  is  $(1/2)[-(\lambda' - 1)(1)]$ .

In this manner we see that

$$\begin{aligned} K(T, V) &= (1/2)[-(\lambda - 1)(1 + 1) - (\lambda' - 1)(1 + 1) - (\lambda - \lambda')(1 + 1)] \\ &= -(\lambda - 1) - (\lambda' - 1) - (\lambda - \lambda') = 2 - 2\lambda = \frac{2(hj + 1)}{(h + 1)(j + 1)}. \end{aligned}$$

From 5.3,  $M(T, V) = 1/2$ , so

$$R(T, V) = M(T, V)/K(T, V) = \frac{(h + 1)(j + 1)}{4(hj + 1)}.$$

□

*Proof. (Completion of the proof of Theorem 4.1.)*

If  $n \geq 6$ , we choose trees  $T^n$ ,  $W^n$ , and  $V^n$  as in Figures 1, 2, and 3 respectively. Then (2) and (3) from Theorem 4.1 follow immediately from Lemmas 5.4 and 5.3. If  $n$  is even, select  $h = j = (n - 2)/2$ . Then the total number of

leaves is  $n$ . Moreover, in this case, by Lemma 5.2,

$$M(T^n, W^n) = \frac{h+j}{(h+1)(j+1)} = 4 \frac{n-2}{n^2}$$

proving (1) when  $n$  is even. For (4), by Lemma 5.5,

$$R(T^n, V^n) = \frac{(h+1)(j+1)}{4(1+hj)} = \frac{n^2}{16+4(n-2)^2}.$$

If  $n$  is odd, select  $h = (n-3)/2$ , and  $j = (n-1)/2$ . Then the total number of leaves is  $n$ . By Lemma 5.2,

$$\begin{aligned} M(T^n, W^n) &= \frac{h+j}{(h+1)(j+1)} \\ &= \frac{n-2}{((n-1)/2)((n+1)/2)} = 4 \frac{n-2}{n^2-1} \end{aligned}$$

proving (1) when  $n$  is odd. For (4), by Lemma 5.5,

$$R(T^n, V^n) = \frac{(h+1)(j+1)}{4(hj+1)} = \frac{n^2-1}{16+4(n-3)(n-1)}.$$

□

## 6 Discussion

In summary, this paper shows that as the number  $n$  of leaves increases without bound, the edge  $l_\infty$  radius becomes at most  $1/4$ . This contrasts with neighbor-joining, for which the edge  $l_\infty$  radius remains at the maximal possible value  $1/2$ .

Theorem 4.1 gives insight, however, into a more refined analysis. For some pairs  $T$  and  $W$  it is nevertheless true that  $R(T, W)$  remains at the maximal possible value  $1/2$ . Thus the trees  $T$  and  $W$  of Figures 1 and 2 tend to be distinguishable by minimum evolution for large  $n$  since, in the cases considered,  $R(T, W)$  remains  $1/2$  for all  $n$ . If  $T$  is the true tree, it is unlikely to be confused with  $W$ . On the other hand, the trees  $T$  and  $V$  of Figures 1 and 3 become less distinguishable by minimum evolution as  $n$  gets large since  $R(T, V)$  goes  $1/4$ . This suggests that special attention is needed in distinguishing  $T$  from  $V$ , more than in distinguishing  $T$  from  $W$ .

The theorems potentially have consequences for the method of taxon sampling applied to minimum evolution. See, for example, Hillis (1996), Graybeal (1998), Poe (1998), or Ackerly (2000). Suppose that  $T$  is the correct tree and we wish to distinguish  $T$  from  $V$ , as in Figures 1 and 3. Then adding more taxa to  $H$  or  $J$  will make the problem harder, not easier. Indeed, using the formula from Lemma 5.5 we see that  $R(T, V) = \frac{(h+1)(j+1)}{4(1+hj)}$  gets smaller when  $h$  or  $j$  is increased.

Further research in these subjects is desirable. The author is unable to prove, for example, that  $M(n + 1) \leq M(n)$  and  $R(n + 1) \leq R(n)$  nor that  $\lim_{n \rightarrow \infty} R(n) = 1/4$ .

## Acknowledgments

Thanks go to Olivier Gascuel for suggesting dramatic simplifications in the proof of Theorem 4.1. Thanks go to the anonymous referees for many helpful suggestions to improve the readability and pertinence of the manuscript.

## References

- Ackerly, D.D. (2000). Taxon sampling, correlated evolution, and independent contrasts. *Evolution* 54(5), 1480-1492.
- Atteson, K. (1999). The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica* 25, 251-278.
- Bryant, D. and P. Waddell (1998). Rapid Evaluation of Least-Squares and Minimum-Evolution Criteria on Phylogenetic Trees. *Mol. Biol. Evol.* 7,1346-1359.
- Denis, F. and O. Gascuel (2003). On the consistency of the minimum evolution principle. *Discrete Applied Mathematics* 127(1), 63-77.
- Desper, R. and O. Gascuel (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology* 9(5), 687-705.
- Farris, J.S., A.G. Kluge, and M.J. Eckardt (1970). A numerical approach to phylogenetic systematics. *Syst. Zool.* 19, 172-191.
- Felsenstein, J. (1997). An alternating least-squares approach to inferring phylogenies from pairwise distances. *Syst. Biol.* 46, 101-111.
- Gascuel, O. (2000). On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. *Mol. Biol. Evol.* 17(3), 401-405.
- Gascuel, O., D. Bryant, and F. Denis (2001). Strengths and limitations of the minimum evolution principle. *Syst. Biol.* 50(5), 621-627.
- Graybeal, A. (1998). Is it better to add taxa or characters to a difficult phylogenetic problem. *Syst. Biol.* 47, 9-17.
- Hasegawa, M., H. Kishino, and K. Yano (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160-174.
- Hillis, D.M. (1996). Inferring complex phylogenies. *Nature* 383, 130.
- Jukes, T.H. and C.R. Cantor (1969). Evolution of protein molecules, in S. Osawa and T. Honjo (Eds), *Evolution of Life: Fossils, Molecules, and Culture*, Springer-Verlag, Tokyo, pp. 79-95.
- Kidd, K.K. and L.A. Sgaramella-Zonta (1971). Phylogenetic analysis: concepts and methods. *Am. J. Human Genet.* 23, 235-252.
- Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol.*

Evol. 16, 111-120.

Lake, J.A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Natl. Acad. Sci. USA* 91, 1455-1459.

Makarenkov, V. and B. Leclerc (1999). An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. *J. Classif.* 16, 3-26.

Poe, S. (1998). Sensitivity of phylogeny estimation to taxonomic sampling. *Syst. Biol.* 47, 18-31.

Rzhetsky, A. and M. Nei (1992a). A simple method for estimating and testing minimum- evolution trees. *Mol. Biol. Evol.* 9, 945-967.

Rzhetsky, A. and M. Nei (1992b). Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *J. Mol. Evol.* 35, 367-375.

Rzhetsky, A. and M. Nei (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.* 10, 1073-1095.

Saitou, N. and M. Nei (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406-425.

Saitou, N. and T. Imanishi (1989). Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Mol. Biol. Evol.* 6(5), 514-525.

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

Steel, M.A. (1994). Recovering a tree from the leaf colourations it generates under a Markov model. *Appl. Math. Lett.* 7(2), 19-23.

Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis (1996). Phylogenetic inference, in D. Hillis, C. Moritz, and B. Mable, (Eds), *Molecular Systematics*, second edition, Sinauer, Sunderland, MA, pp. 407-514.