Note

On the edge $l_\infty$ radius of Saitou and Nei’s method for phylogenetic reconstruction

Wenqiang Dai\textsuperscript{a,}\textsuperscript{*}, Yinfeng Xu\textsuperscript{a,b}, Binhai Zhu\textsuperscript{c}

\textsuperscript{a}School of Management, Xi’an Jiaotong University, Xi’an, 710049, PR China
\textsuperscript{b}The State Key Lab for Manufacturing Systems Engineering, PR China
\textsuperscript{c}Department of Computer Science, Montana State University, Bozeman, MT 59717-3880, USA

Received 25 May 2006; received in revised form 28 August 2006; accepted 30 August 2006

Communicated by Ding-Zhu Du

Abstract

In this paper, we study the performance of Saitou and Nei’s neighbor-joining method for phylogenetic reconstruction. We show that the edge $l_\infty$ radius of the method is $\frac{1}{4}$. This improves an result by Atteson [The performance of neighbor-joining methods of phylogenetic reconstruction, Algorithmica 25 (1999) 251–278] and Xu et al. [A lower bound on the edge $l_\infty$ radius of Saitou and Nei’s method for phylogenetic reconstruction, Inform. Process. Lett. 94(5) (2005) 225–230]. Previously, only an upper bound $\frac{1}{4}$ and a lower bound $\frac{1}{6}$ were known.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Phylogenetic reconstruction; Neighbor-joining; Evolutionary trees

1. Introduction

Inferring evolutionary relationships is a fundamental problem in biology. Phylogenies, or leaf-labeled trees, are useful tools to represent evolutionary relationships between species [3,4,11]. In recent years, with the development of bioinformatics and computational biology, these methods have been used in other areas, such as historical linguistics.

In general, the phylogenetic reconstruction problem is to reconstruct a leaf-labeled tree from a distance matrix. The problem is known to be NP-complete [7]. A lot of heuristic algorithms have been proposed in the past decades [12]. The neighbor-joining algorithm of Saitou and Nei [9] is one of the most popular methods for reconstructing phylogenetic trees from a matrix of pairwise evolutionary distances. This is not only because of its speed, but also because of its accuracy: when exact distances are given, it is guaranteed to reconstruct the correct tree. (Notice that in practice the matrix is usually not perfect, its entries could be incomplete or contain errors.)

The theoretical performance of phylogenetic reconstruction methods have been studied a few years ago. Atteson [2] defined the $l_\infty$ radius and edge $l_\infty$ radius to guarantee that the methods should conform with the demand of actual biological meaning, that is, the primary goal of phylogenetic reconstruction method is to correctly reconstruct all edges.

\textsuperscript{*} This research is supported by NSF of China under Grants Nos. 10371094, 70471035 and 70525004, and NSF.

\textsuperscript{*} Corresponding author.

E-mail addresses: wq dai@mail.xjtu.edu.cn (W. Dai), yf xu@mail.xjtu.edu.cn (Y. Xu), bhz@cs.montana.edu (B. Zhu).

0304-3975/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2006.08.044
(using the $l_\infty$ radius) or some of the edges (using the edge $l_\infty$ radius) of the tree for observed distance matrices. Among several interesting bounds, Atteson proved that Saitou and Nei’s neighbor-joining method has an edge $l_\infty$ radius at most $\frac{1}{4}$. Recently, Xu et al. [13] improved the lower bound from 0 to $\frac{1}{6}$. In this paper, we show that Saitou and Nei’s neighbor-joining method has an edge $l_\infty$ radius at least $\frac{1}{3}$, which is tight.

The paper is organized as follows. In Section 2, we introduce some basic notations and definitions regarding phylogenetic reconstruction and the neighbor-joining method. In Section 3, we present the theoretical proof.

2. Preliminaries

In this section we first present some necessary notations and definitions. We assume that the reader is familiar with the basic concepts of graph theory, see, e.g., Bollobás [5]. We will also present the details of the neighbor-joining method and the edge $l_\infty$ radius.

2.1. Definitions

**Definition 1.** Given a tree $T = (V(T), E(T))$, we write $L(T)$ for the set of leaves of $T$. When the tree $T$ is implicitly understood, we simply write $V, E$ and $L$ for the vertex, edge and leaf set of $T$, respectively. A binary tree is a tree in which every internal node has degree three.

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, $T - e = (V, E - \{e\})$. Note that $T - e$ has two disjoint components and partitions the set of leaves into two disjoint ones. For $k \in L$, we use the notation $L_k(T - e)$ for the set of leaves in the component of $T - e$ containing $k$, thus $L - L_k(T - e)$ will be used to denote the other component. Let $s(T - e) = \{L_k(T - e), L - L_k(T - e)\}$, which we refer to as the split of $T$ generated by $e$ and let $S(T)$ denote the set of all splits of $T$, that is $S(T) = \{s(T - e) : e \in E(T)\}$.

The evolutionary trees are leaf-labeled trees, whose leaves are represented as the extant species in evolutionary biology. Two evolutionary trees are the same if one is isomorphic to the other.

**Definition 2.** A weighted tree $T$ is a tree $T$ along with a function $l : E(T) \to (0, +\infty)$. A distance matrix, $D$, is a symmetric matrix with zero diagonal elements indexed by a set of species $L$. For any nodes $x$ and $y$ of a weighted tree $T$, we define the distance between $x$ and $y$ as

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

where $P_{i,j}$ denotes the set of edges on the unique path between $i$ and $j$ in $T$. A distance matrix $\hat{D}$ is called an additive distance matrix if there is a weighted tree $T$ such that $\hat{D} = D^T$. Obviously the weighted tree corresponding to an additive distance matrix is unique.

**Definition 3.** A distance-based method is a function $f : \mathcal{D} \to \mathcal{T}$, where $\mathcal{D}$ is the set of distance matrices and $\mathcal{T}$ is the set of trees. Intuitively, given a distance matrix the method tries to reconstruct a phylogeny.

2.2. Saitou and Nei’s neighbor-joining method

The neighbor-joining method is a special distance-based method and its output is a binary tree. There are several versions for this method which differ in choosing different objective functions at each step. We will focus on Saitou and Nei’s method [9]. Given a distance matrix $D$, the algorithm computes $S_{i,j} = (n - 2)D_{i,j} - \sum_k D_{k,i} - \sum_k D_{k,j}$ for each species $i, j$ and chooses a pair of species which minimize $S_{i,j}$, creates a new node that represents the cluster of these species, and then computes a new distance matrix with reduced size where both species are replaced by this new node. The process is repeated until the number of species become three (or two for rooted tree).

We summarize the algorithm of Saitou and Nei’s neighbor-joining method following [2]. Let $L^1 = L = \{1, 2, \ldots, n\}$, $\hat{D}^1 = \hat{D}$ and $L_i = \{i\}$ for all $i \in L$. 

For \( m = 1, 2, \ldots, n - 2 \):
1. Choose \( i^m \) and \( j^m \) which minimize the \( S_{i^m, j^m} \).
2. Fixing the new specie \( u^m = \{i^m, j^m\} \), let \( L^{m+1} = (L^m - \{i^m, j^m\}) \cup \{u^m\} \) and \( L_{u^m} = L_{i^m} \cup L_{j^m} \).

Then compute
\[
\hat{D}^{m+1}_{k,l} = \begin{cases} 
\hat{D}^m_{k,l} & k, l \neq u^m, \\
\frac{1}{2} \hat{D}^m_{i,l} + \frac{1}{2} \hat{D}^m_{j,l} & k = u^m.
\end{cases}
\]

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

Here, the set \( L^m \) denotes the set of species and \( \hat{D}^m \) denotes the distances matrix which are fed into the \( m \)th iteration of the method.

### 2.3. Reconstruction performance

Given an exact distance matrix we can reconstruct a correct tree using the neighbor-joining method [1,2]. In practice, due to the noise or incomplete data, we can only obtain an observed distance matrix, possibly with errors. Obviously, one way to measure the performance of a method is to check whether it can reconstruct all or some of the tree edges when the noise is sufficiently small. The following definition is presented in [2].

**Definition 4.** For any two distance matrices \( \hat{D} \) and \( \hat{D}' \), the \( l_\infty \) norm error between them, written \( \| \hat{D} - \hat{D}' \|_\infty \) is defined by
\[
\| \hat{D} - \hat{D}' \|_\infty = \max_{i,j} |\hat{D}_{i,j} - \hat{D}'_{i,j}|.
\]

For a weighted tree \( T \), the distance-based method \( f \) correctly reconstructs \( e \in E(T) \) on input distance matrix \( \hat{D} \), if, there is an edge \( e' \in E(f(\hat{D})) \) such that \( s(T - e) = s(f(\hat{D}) - e') \) (that is, \( s(T - e) \in S(f(\hat{D})) \)). A method \( f \) has edge \( l_\infty \) radius \( \alpha \) if for every weighted binary tree \( T \), every edge \( e \in E(T) \) and every distance matrix \( \hat{D} \) whenever
\[
\| \hat{D} - D^T \|_\infty < 2\alpha l(e),
\]
the method correctly reconstructs edge \( e \) on input \( \hat{D} \).

Atteson [2] presented a counterexample to show that the edge \( l_\infty \) radius of Saitou and Nei’s method is at most \( \frac{1}{4} \). Xu et al. [13] proved that this edge \( l_\infty \) radius is at least \( \frac{1}{6} \). In this paper, we show that this edge \( l_\infty \) radius is at least \( \frac{1}{4} \), thus obtaining a tight bound on the edge \( l_\infty \) radius of Saitou and Nei’s method.

### 3. The edge \( l_\infty \) radius of Saitou and Nei’s method

We first review the following result [2] which will be useful in our proof. For the history and proof of this important result, see, e.g., [4,6].

**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
\[
D_{i,j} + D_{k,l} \leq D_{i,k} + D_{j,l} = D_{i,l} + D_{j,k},
\]
if and only if \( D \) is an additive distance matrix. If \( D \) corresponds with a weighted binary tree \( T \), then there is an edge \( e \) which separates \( i \) and \( j \) from \( k \) and \( l \), that is, \( i \) and \( j \) are in a different component of \( T - e \) from \( k \) and \( l \), and the difference between the right-hand side and left-hand side of the above inequality is at least \( 2l(e) \).

The main result of this paper is as follows.

**Theorem 2.** Saitou and Nei’s neighbor-joining method has edge \( l_\infty \) radius at least \( \frac{1}{4} \).
Proof. We need to prove that, for every weighted binary tree $T$, every input distance matrix $\hat{D}$, the method correctly reconstructs edge $e \in E(T)$ given that the length of edge $e$ is no less than $4\epsilon$, where $e = \|D^T - \hat{D}\|_\infty$. In the light of Definition 4, we must show that at any iteration, for any pair of leaves $k, l$ on the opposite sides of $e$, there exist a pair of leaves on the same sides of $e$—denoted by $i$ and $j$, such that $\hat{S}_{k,l} - \hat{S}_{i,j} > 0$. That is, the theorem is proved if we can prove that there exist a pair of leaves $i, j$ in $L_k(T - e)$ such that $\hat{S}_{k,l} - \hat{S}_{i,j} > 0$ or, a pair of leaves in $L - L_k(T - e)$—denoted by $p$ and $q$, such that $\hat{S}_{k,l} - \hat{S}_{p,q} > 0$.

Without loss of generality, let $|L_k(T - e)| \leq |L_l(T - e)|$, we now prove that there exist a pair of leaves $i, j$ in $L_k(T - e) - \{k\}$ such that $\hat{S}_{k,l} - \hat{S}_{i,j} > 0$.

Firstly, for any pair of leaves $k, l$ on the opposite sides of $e$ and any pair of leaves $i, j$ on the same sides of $e$ with $k$, we have

$$
\hat{S}_{k,l} - \hat{S}_{i,j} = (n - 2)(\hat{D}_{k,l} - \hat{D}_{i,j}) + \sum_{t} (\hat{D}_{t,i} + \hat{D}_{t,j} - \hat{D}_{t,k} - \hat{D}_{t,l})
$$

$$
= (n - 2)(\hat{D}_{k,l} - \hat{D}_{i,j}) + 2(\hat{D}_{i,j} - \hat{D}_{k,l}) + \sum_{t \neq \{k, l, i, j\}} (\hat{D}_{t,i} + \hat{D}_{t,j} - \hat{D}_{t,k} - \hat{D}_{t,l})
$$

$$
= (n - 4)(\hat{D}_{k,l} - \hat{D}_{i,j}) + \sum_{t' \in L_k(T - e) - \{l\}} (\hat{D}_{t',i} + \hat{D}_{t',j} - \hat{D}_{t',k} - \hat{D}_{t',l})
$$

$$
+ \sum_{t' \in \{k, i, j\}} (\hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{t',i} + \hat{D}_{t',j} - \hat{D}_{t',k} - \hat{D}_{t',l}).
$$

As $|L_k(T - e)| \leq |L_l(T - e)|$, we can (arbitrary) choose a subset $A \subseteq L_l(T - e) - \{l\}$ such that $|A| = |L_k(T - e) - \{k, i, j\}| = m$. Let $B = L_l(T - e) - \{l\} - A$. Then

$$
\hat{S}_{k,l} - \hat{S}_{i,j} = \frac{1}{m} \sum_{t' \in A} \sum_{t' \in B} (\hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{t',i} + \hat{D}_{t',j} - \hat{D}_{t',k} - \hat{D}_{t',l})
$$

$$
+ (\hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{t',i} + \hat{D}_{t',j} - \hat{D}_{t',k} - \hat{D}_{t',l})
$$

$$
+ \sum_{t' \in B} (\hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{t',i} + \hat{D}_{t',j} - \hat{D}_{t',k} - \hat{D}_{t',l})
$$

$$
= \frac{1}{m} \Omega_1 + \Omega_2,
$$

where

$$
\Omega_1 = \sum_{t' \in A} \sum_{t' \in B} f(k, l, i, j, t, t'),
$$

$$
\Omega_2 = \sum_{t' \in B} \sum_{t' \in A} g(k, l, i, j, t, t').
$$

By the following lemmas, given $l(e) > 4\epsilon$, there exist a pair of leaves $i, j$ on the same sides of $e$ with $k$ such that $\Omega_1 > 0$ (Lemma 3) and $\Omega_2 > 0$ (Lemma 6), thus $\hat{S}_{k,l} - \hat{S}_{i,j} > 0$. So the theorem is proven. \qed
Lemma 3. For any pair of leaves $k, l$ on the opposite sides of $e$, given $l(e) > 4\epsilon$ and $|L_k(T - e)| \leq |L_l(T - e)|$, there exist a pair of leaves, $i$ and $j$, on the same sides of $e$ with $k$ such that $\Omega_1 = \sum_{t \in L_k(T - e) - \{k, i, j\}} f(k, l, i, j, t, t') > 0$.

Proof. For any $i \in L_k(T - e) - \{k\}$, $j \in L_l(T - e) - \{k, i\}$ and $t \in L_k(T - e) - \{k, i, j\}$, given $l(e) > 4\epsilon$ and $|L_k(T - e)| \leq |L_l(T - e)|$, by the following Lemma 4, the function $f(k, l, i, j, t, t')$ satisfies

$$f(k, l, i, j, t, t') + f(k, l, i, t, j, t') + f(k, l, j, i, t, t') > 0.$$ 

Then by Lemma 5, we obtain

$$\Omega_1 = \sum_{t \in L_k(T - e) - \{k, i, j\}} \sum_{t' \in A} f(k, l, i, j, t, t') > 0.$$ 

Therefore, there must exist a leaf $i \in L_k(T - e) - \{k\}$, and $j \in L_l(T - e) - \{k, i\}$, such that $\sum_{t \in L_k(T - e) - \{k, i, j\}} f(k, l, i, j, t, t') > 0$, which implies

$$\Omega_1 = \sum_{t \in L_k(T - e) - \{k, i, j\}} \sum_{t' \in A} f(k, l, i, j, t, t') > 0. \quad \Box$$

Lemma 4. For any pair of leaves $k, l$ on the opposite sides of $e$, and any $i, j, t \in L_k(T - e) - \{k\}, t' \in L_l(T - e) - \{l\}$, if $l(e) > 4\epsilon$, then

$$f(k, l, i, j, t, t') + f(k, l, i, t, j, t') + f(k, l, j, i, t, t') > 0.$$ 

Proof. Note that $\hat{D}_{a,b} = \hat{D}_{b,a}$ and we have

$$f(k, l, i, j, t, t') + f(k, l, i, t, j, t') + f(k, l, j, i, t, t')$$

$$= (\hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{l,i} + \hat{D}_{l,j} - \hat{D}_{i,k} - \hat{D}_{i,l}) + (\hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{l,i} + \hat{D}_{l,j} - \hat{D}_{i,k} - \hat{D}_{l,t})$$

$$+ (\hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{l,i} + \hat{D}_{l,j} - \hat{D}_{i,k} - \hat{D}_{l,t})$$

$$+ (\hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{l,i} + \hat{D}_{l,j} - \hat{D}_{i,k} - \hat{D}_{l,t})$$

$$= 6\hat{D}_{k,l} + 2(\hat{D}_{i,j} + \hat{D}_{l,i} - \hat{D}_{l,t}) - \hat{D}_{k,l} + \hat{D}_{i,j} + \hat{D}_{l,i} + \hat{D}_{l,j} + \hat{D}_{l,t} - 3(\hat{D}_{k,t'} + \hat{D}_{l,t'})$$

$$= \left[ (\hat{D}_{k,l} + \hat{D}_{i,j} - \hat{D}_{l,t} - \hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{l,t}) (\hat{D}_{k,l} + \hat{D}_{i,j} - \hat{D}_{l,t} - \hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{l,t}) \right]$$

$$\equiv A_1 + A_2,$$ 

where

$$A_1 = (\hat{D}_{k,l} + \hat{D}_{i,j} - \hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{l,t} - \hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{l,t})$$

$$A_2 = (\hat{D}_{k,l} + \hat{D}_{i,j} - \hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{l,t} - \hat{D}_{k,l} - \hat{D}_{i,j} + \hat{D}_{l,t}).$$

Let $D = D^T$. By Four Point Condition (Lemma 1), we have

$$A_1 = (D_{k,l} + D_{i,j} - D_{k,l} - D_{i,j}) + (D_{k,l} + D_{i,j} - D_{k,j} - D_{l,t}) + (D_{k,l} + D_{i,j} - D_{k,j} - D_{l,t})$$

$$+ (e_{k,l} + e_{l,t} - e_{k,l} - e_{l,t}) + (e_{k,l} + e_{l,t} - e_{k,j} - e_{l,j}) + (e_{k,l} + e_{l,t} - e_{k,t} - e_{l,j})$$

$$\geq (D_{k,l} + D_{l,t} - D_{k,l} - D_{l,t}) + (D_{k,l} + D_{l,t} - D_{k,j} - D_{l,t}) + (D_{k,l} + D_{l,t} - D_{k,t} - D_{l,t}) - 12\epsilon$$

$$\geq 6l(e) - 12\epsilon.$$
Now we compute $A_2$. Firstly we have

\[
A_2 = (D_{k,l} + D_{l,t'} - D_{k,t'} - D_{l,i}) + (D_{k,l} + D_{j,t'} - D_{k,t'} - D_{l,j}) + (D_{k,l} + D_{l,t'} - D_{k,t'} - D_{l,i})
\]

\[
+ (e_{k,l} + e_{l,t'} - e_{k,t'} - e_{l,i}) + (e_{k,l} + e_{l,t'} - e_{k,t'} - e_{l,j}) + (e_{k,l} + e_{l,t'} - e_{k,t'} - e_{l,t})
\]

\[
\geq (D_{k,l} + D_{l,t'} - D_{k,t'} - D_{l,i}) + (D_{k,l} + D_{j,t'} - D_{k,t'} - D_{l,j}) + (D_{k,l} + D_{l,t'} - D_{k,t'} - D_{l,i}) - 12\varepsilon.
\]

Let $e = (x, y)$, without loss of generality, we assume that $x$ is the left vertex of $e$ and $y$ is the right vertex of $e$, at the same time assume that the leaf $k$ is on the same side with endpoint $x$ and leaf $l$ is on the same side with endpoint $y$. (See Fig. 1.)

Let $l(t, x) = b_t$ for $t \in L_k(T - e)$ and $l(t', y) = b_t'$ for $t' \in L_k(T - e)$, then

\[
D_{k,l} = b_k + b_l + l(e), \quad D_{l,t'} = b_l + b_t' + l(e), \quad D_{k,l} = b_k + b_l + l(e), \quad D_{l,i} = b_l + b_i + l(e),
\]

\[
D_{j,t'} = b_j + b_t' + l(e), \quad D_{l,j} = b_l + b_j + l(e), \quad D_{l,t} = b_l + b_t + l(e), \quad D_{l,t} = b_l + b_t + l(e).
\]

Thus,

\[
D_{k,l} + D_{l,t'} - D_{k,t'} - D_{l,i} = (b_k + b_l + l(e) + b_l + b_t' + l(e)) - (b_k + b_l + l(e) + b_l + b_i + l(e)) = 0,
\]

\[
D_{k,l} + D_{j,t'} - D_{k,t'} - D_{l,j} = (b_k + b_l + l(e) + b_j + b_t' + l(e)) - (b_k + b_l + l(e) + b_j + b_l + l(e)) = 0,
\]

\[
D_{k,l} + D_{l,t'} - D_{k,t'} - D_{l,i} = (b_k + b_l + l(e) + b_l + b_t' + l(e)) - (b_k + b_l + l(e) + b_l + b_t + l(e)) = 0.
\]

Therefore, we have

\[
A_2 \geq - 12\varepsilon.
\]

Now we obtain that

\[
f(k, l, i, j, t, t') + f(k, l, i, t, j, t') + f(k, l, j, t, i, t') = A_1 + A_2 \geq 6l(e) - 24\varepsilon.
\]

As $l(e) > 4\varepsilon$, finally we have

\[
f(k, l, i, j, t, t') + f(k, l, i, t, j, t') + f(k, l, j, t, i, t') > 0.
\]

This lemma is proven. □

**Lemma 5.** If $f(k, l, i, j, t, t') + f(k, l, i, t, j, t') + f(k, l, j, t, i, t') > 0$, then

\[
\sum_{i \in L_k(T - e) - \{k\}} \sum_{j \in L_k(T - e) - \{k,j\}} \sum_{t \in L_k(T - e) - \{k,i,j\}} f(k, l, i, j, t, t') > 0
\]

**Proof.** As the sum is independent of $l, t'$, for simplicity, we let $f(k, l, a, b, c, t') = f(k, a, b, c)$. We also let $L_k(T - e) - \{k\} = X$. Then this lemma is to prove

\[
\sum_{i \in X} \sum_{j \in X - \{i\}} \sum_{t \in X - \{i,j\}} f(k, i, j, t) \geq 0.
\]
The proof is by induction. Firstly, if \(|X| = 3\), denote \(X = \{x_1, x_2, x_3\}\). Then
\[
\sum_{i \in X} \sum_{j \in X - \{i\}} \sum_{t \in X - \{i, j\}} f(k, i, j, t)
\]
\[
= f(k, x_1, x_2, x_3) + f(k, x_1, x_3, x_2) + f(k, x_2, x_1, x_3)
\]
\[
+ f(k, x_2, x_3, x_1) + f(k, x_3, x_1, x_2) + f(k, x_3, x_2, x_1)
\]
\[
= 2[f(k, x_1, x_2, x_3) + f(k, x_1, x_3, x_2) + f(k, x_2, x_3, x_1)] > 0,
\]
where the second equality comes from \(\forall a, b, f(k, a, b, t) = f(k, b, a, t)\), and the third inequality comes from the input constraints \(f(k, x_1, x_2, x_3) + f(k, x_1, x_3, x_2) + f(k, x_2, x_3, x_1) > 0\).

Now, assume that this lemma is correct for any set \(A\), we prove that it also holds for set \(A + \{x\}\), where \(x \notin A\). We have
\[
\sum_{i \in A + \{x\}} \sum_{j \in A + \{x\} - \{i\}} \sum_{t \in A + \{x\} - \{i, j\}} f(k, i, j, t)
\]
\[
= \sum_{i \in A} \sum_{j \in A - \{i\}} \sum_{t \in A - \{i, j\}} f(k, i, j, t) + \sum_{i \in A} \sum_{j \in A - \{i\}} f(k, i, j, x)
\]
\[
+ \sum_{i \in A} \sum_{t \in A - \{i\}} f(k, i, x, t) + \sum_{j \in A} \sum_{t \in A - \{j\}} f(k, x, j, t)
\]
\[
= \sum_{i \in A} \sum_{j \in A - \{i\}} \sum_{t \in A - \{i, j\}} f(k, i, j, t) + \sum_{i \in A} \sum_{j \in A - \{i\}} [f(k, i, j, x) + f(k, i, x, j) + f(k, x, i, j)] > 0.
\]
Hence, the lemma is proved. \(\square\)

**Lemma 6.** For any pair of leaves \(k, l\) on the opposite sides of \(e\), if \(|L_k(T - e)| \leq |L_l(T - e)|\), then for any pair of leaves, \(i, j \in L_k(T - e) - \{k\}\), given \(l(e) > 4\varepsilon\), we have \(\Omega_2 = \sum_{t' \in B} g(k, l, i, j, t') > 0\).

**Proof.** We will prove a stronger result: for any pair of leaves, \(i, j \in L_k(T - e) - \{k\}\) and any \(t' \in L_l(T - e) - \{l\}\), given \(l(e) > 4\varepsilon\), we have \(g(k, l, i, j, t') > 0\).

We also let \(D = D^T\), then
\[
g(k, l, i, j, t') = \hat{D}_{k, l} - \hat{D}_{l, i} + \hat{D}_{t', i} + \hat{D}_{t', j} - \hat{D}_{t', k} - \hat{D}_{t', l}
\]
\[
= D_{k, l} - D_{l, i} + D_{t', i} + D_{t', j} - D_{t', k} - D_{t', l} + \varepsilon_{k, l} - \varepsilon_{l, i} + \varepsilon_{t', i} + \varepsilon_{t', j} - \varepsilon_{t', k} - \varepsilon_{t', l}
\]
\[
\geq (D_{k, l} + D_{t', j} - D_{t', k} - D_{t', i} + (D_{l, i} + D_{t', i} - D_{t', l} - D_{l, j}) - 6\varepsilon.
\]

By Four Point Condition (Lemma 1), we have
\[
D_{t', i} - D_{t', l} - D_{t', j} \geq 2l(e).
\]
Similar to the proof of Lemma 4, let \(l(t, x) = b_t\) for \(t \in L_k(T - e)\) and \(l(t', y) = b_{t'}\) for \(t' \in L - L_k(T - e)\) (see also Fig. 1), we have
\[
D_{k, l} = b_k + b_l + l(e), D_{t', j} = b_{t'} + b_j + l(e), D_{t', k} = b_k + b_{t'} + l(e), D_{l, j} = b_l + b_j + l(e)
\]
and this implies
\[
D_{k, l} + D_{t', j} - D_{t', k} - D_{l, j} = 0.
\]
Thus, we obtain
\[
g(k, l, i, j, t') \geq 2l(e) - 6\varepsilon \geq 2 \times 4\varepsilon - 6\varepsilon = 2\varepsilon > 0. \square
4. Closing remarks

In this paper we show that $\frac{1}{4}$ is a tight bound for the edge $l_\infty$ radius of Saitou and Nei’s neighbor-joining method for phylogenetic reconstruction. An interesting problem is to appropriately modify the method to obtain the largest edge $l_\infty$ radius $\frac{1}{2}$. Another interesting problem is to compare empirically the performance of different neighbor-joining methods, e.g., the one proposed by Sattath and Tversky [10]. We remark that Moret et al. recently started some related empirical investigations [8].

References