HAL
open science

Consistency of Topological Moves Based on the Balanced Minimum Evolution Principle of Phylogenetic Inference

Magnus Bordewich, Olivier Gascuel, Katharina Huber, Vincent Moulton

To cite this version:

Magnus Bordewich, Olivier Gascuel, Katharina Huber, Vincent Moulton. Consistency of Topological Moves Based on the Balanced Minimum Evolution Principle of Phylogenetic Inference. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2009, 6 (1), pp.110-117. 10.1109/TCBB.2008.37 . lirmm-00324146

HAL Id: lirmm-00324146

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00324146v1

Submitted on 24 Sep 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Consistency of Topological Moves Based on the Balanced Minimum Evolution Principle of Phylogenetic Inference

M. Bordewich, O. Gascuel, K. T. Huber,V. Moulton

M. Bordewich is with the Department of Computer Science, Durham University, U. K.; O. Gascuel is with LIRMM, CNRS-Université Montpellier II, France; K. T. Huber and V. Moulton are with the Department of Computer Science, University of East Anglia, U. K.

Abstract

Many phylogenetic algorithms search the space of possible trees using topological rearrangements and some optimality criterion. FastME is such an approach that uses the *balanced minimum evolution (BME)* principle, which computer studies have demonstrated to have high accuracy. FastME includes two variants: *balanced subtree prune and regraft (BSPR)* and *balanced nearest neighbor interchange (BNNI)*. These algorithms take as input a distance matrix and a putative phylogenetic tree. The tree is modified using SPR or NNI operations, respectively, to reduce the BME length relative to the distance matrix, until a tree with (locally) shortest BME length is found.

Following computer simulations, it has been conjectured that BSPR and BNNI are consistent, i.e. for an input distance that is a tree-metric, they converge to the corresponding tree. We prove that the BSPR algorithm is consistent. Moreover, even if the input contains small errors relative to a tree-metric, we show that the BSPR algorithm still returns the corresponding tree. Whether BNNI is consistent remains open.

Index Terms

phylogenetic tree, topological move, subtree prune and regraft (SPR), BSPR algorithm, Nearest Neighbor Interchange (NNI), BNNI algorithm, balanced minimum evolution principle (BME), tree-length, quartet-distance, Robinson Foulds distance, consistency, safety radius.

I. Introduction

Many practical methods for phylogenetic tree inference proceed by repeatedly updating a proposed tree using topological rearrangements, until a locally optimal tree is found according to some optimality criterion. Such methods include those implemented in the widely used PAUP* [29] and PHYLIP packages [12], and optimality criteria include likelihood and parsimony scores. The most commonly used topological rearrangements are Subtree Prune and Regraft (SPR), Nearest Neighbor Interchange (NNI), and Tree Bisection and Reconnection (TBR); see [25] for definitions and properties, and the next section for a brief description of SPR and NNI moves.

Recently, such a local topology search approach was introduced for inferring phylogenetic trees from distance matrices, based on the *balanced minimum evolution (BME)* principle [6]. The optimality criterion used is to minimize Pauplin's [20] tree-length estimate relative to the given distance matrix. This approach is implemented in a software called FastME [6]. Two topological rearrangement possibilities are available in the latest release of FastME: the *balanced*

subtree prune and regraft (BSPR) algorithm [17] and the *balanced nearest neighbor interchange (BNNI) algorithm* [6]. FastME has been shown [6], [7] to be a fast and accurate method for tree inference, compared to other popular distance-based methods such as NJ [23], BIONJ [15], FITCH [13] or WEIGHBOR [3]. Vinh et al. [30] even concluded "We found that BNNI boosts the topological accuracy of all [distance-based] methods." Note that the local search range under NNI operations is a subset of that under SPR operations, so BSPR is expected to be at least as accurate as BNNI.

A number of studies have been dedicated to the greedy algorithms used to infer an initial tree for use in a topological search. For example, Atteson's study of NJ [2]. However, to the best of our knowledge, no one has explored theoretical properties of topological moves in the context of tree inference. Here we will make a first step towards filling this gap in relation to the BME framework, and in this way, shed some light on why BSPR and BNNI work so well in practise. In particular, we consider the following question. Suppose the matrix of pairwise distances given as input is in fact a *tree-metric* δ^*, *i.e.* there is a unique phylogenetic tree T^* with positive edge lengths for T^* so that, for each $x, y \in X$, the distance δ^*_{xy} is the length of the path between x and y in T^*. If we apply the BSPR (BNNI) algorithm starting with distance δ^* and any initial phylogenetic tree T, is the algorithm guaranteed to output T^*? That is to say, is the BSPR (BNNI) algorithm *consistent*?

Numerous computer simulations have suggested that both the BSPR and BNNI algorithms are consistent [7]. Here we prove that the BSPR algorithm is indeed consistent. In fact, we show that even if the input δ contains some errors, but remains sufficiently close to δ^*, then the BSPR algorithm will still output T^* (Theorem 5.2). Here, sufficiently close means $|\delta_{xy} - \delta^*_{xy}|$ is less than $1/3$ of the smallest edge weight of T^*, for all $x, y \in X$, i. e. the BSPR algorithm has a *safety radius* of at least 1/3. As a corollary, we show that the BME principle itself has a safety radius of at least 1/3, which solves an open question [8]. Safety radius analysis was introduced by Atteson [2], and has become a standard approach to characterize the performance of distance-based, tree building algorithms (see *e.g.* [9] for a review). In particular, Atteson showed that no distance method can have a safety radius larger than 1/2, and that NJ and related greedy algorithms have optimal 1/2 safety radius.

The rest of the paper is organised as follows. In the following section, we review some basic definitions concerning phylogenetic trees and balanced minimum evolution, and prove a

key lemma concerning the structure of pairs of trees. In Sections III and IV, we prove some results analogous to consistency of the BSPR algorithm for the Robinson-Foulds [22] and the quartet [11] tree comparison metrics. In particular, in Section III we show that for two distinct phylogenetic trees T and T^* there is a sequence of SPR operations which transforms T into T^* and decreases the Robinson-Foulds distance to T^* at every step. In Section IV, we prove a similar result for the quartet distance. In Section V, we show that the BSPR algorithm is consistent and has safety radius at least $1/3$. However, the question remains open for BNNI. This is discussed along with other open questions in Section VI.

II. Basics, definitions and notation

A *phylogenetic tree* is a binary tree T whose leaves are bijectively labelled by the elements of some finite set X. The set X usually denotes a set of species or taxa, and the tree T represents the evolutionary relationships between them. Unless stated otherwise, from now on X will denote a finite set and all trees considered will be phylogenetic trees on X. Throughout we consider phylogenetic trees as unweighted, i. e. they do not have intrinsic edge lengths, with the exception of the true tree T^* which does have edge lengths (or weights). Furthermore, capital letters will be used in all figures to represent subtrees.

The NNI and SPR tree rearrangement operations can be described as follows [25]. Suppose that T is the tree depicted in Fig. 1 that A, B, C, $C_0, \ldots, C_k$ and D are subtrees of T as indicated in that figure, and that T' is a tree resulting from one NNI or SPR operation applied to T. Regarding NNI, T' is obtained from T by deleting some edge $e = \{u, r_B\}$ of T where r_B is the root of B, suppressing vertex u, and adding an edge e' between r_B and a vertex that subdivides the edge between v and D or between v to C where v is the neighbor of u in $T - e$ (cf. Fig. 1(a)). Regarding SPR, T' is obtained from T by deleting some edge $e = \{u, r_B\}$ in T where again r_B is the root of B, suppressing u, and adding an edge e' between r_B and a vertex that subdivides an edge in the component of $T - e$ that does not contain B (cf. Fig. 1(b)). Note that in both operations the root of B is unchanged, i.e. the edges e and e' share the same vertex of B.

The BSPR (BNNI, respectively) algorithm works as follows. For an input distance matrix δ, with entries δ_{xy}, $x, y \in X$ and some phylogenetic tree T on X, the *total tree length* $\hat{l}(T)$ of T

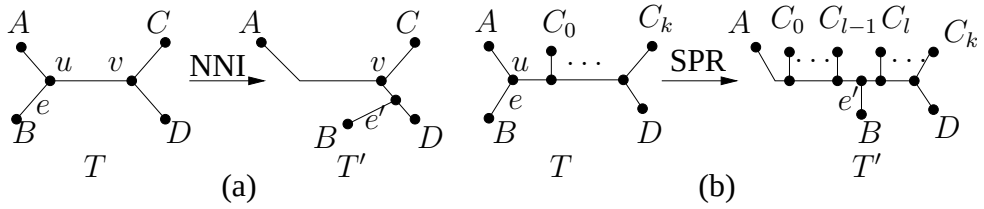

Fig. 1. A schematic description of an NNI and an SPR operation. See text for details.

(relative to δ) is defined according to the following formula due to Pauplin [20]

$$\hat{l}(T) = \sum_{x,y \in X} 2^{1-p_{xy}} \delta_{xy}, \tag{1}$$

where p_{xy} denotes the number of edges in the path from x to y. Semple and Steel [26] provided an elegant interpretation of Equation (1) which we present in Fig. 2 for the convenience of the reader. Then, for all trees T' that can be obtained from T by performing a single SPR (NNI,

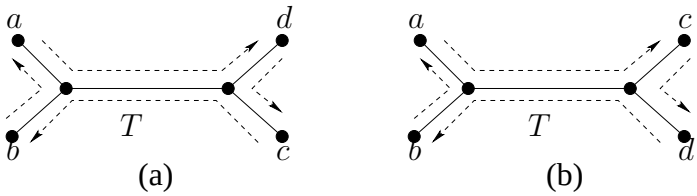

Fig. 2. The figure depicts two drawings of the same tree T on the set $X = \{a, b, c, d\}$. By crossing each edge twice as indicated, the tree length $\hat{l}(T)$ of the tree T depicted in (a) equates to $\frac{1}{2}(\delta_{ad} + \delta_{dc} + \delta_{cb} + \delta_{ba})$ and to $\frac{1}{2}(\delta_{ac} + \delta_{cd} + \delta_{db} + \delta_{ba})$ in (b) where δ_{xy} denotes the distance between any two elements in X. Pauplin's formula for $\hat{l}(T)$ is the average of these two alternative ways to compute $\hat{l}(T)$ i.e. $\hat{l}(T) = \frac{1}{2}(\frac{1}{2}(\delta_{ad} + \delta_{dc} + \delta_{cb} + \delta_{ba}) + \frac{1}{2}(\delta_{ac} + \delta_{cd} + \delta_{db} + \delta_{ba}))$. This interpretation can be extended to larger trees using circular orderings of X, see [26].

respectively) operation on T (see Fig. 1), it is checked whether $\hat{l}(T) - \hat{l}(T') > 0$. If this holds, *i.e.* the total tree length of T' is less than that of T, the tree T' is taken in preference to T and the process is iterated. This process is repeated until a tree T'' is found with the property that no SPR operation (NNI, respectively) on T'' yields a tree having shorter total tree length. Note that (i) if δ is a tree metric and T an edge weighted phylogenetic tree that realizes δ then $\hat{l}(T)$ is the sum of the branch lengths of T [26], (ii) the local search range under NNI operations is a subset of that under SPR, and (iii) the check $\hat{l}(T) - \hat{l}(T') > 0$ can be performed efficiently.

Indeed in both BSPR and BNNI it takes time $O(|X|^2)$ to evaluate all moves and update all data structures corresponding to the new current tree, see [6], [17] for details.

A *split* $S = \{A, B\}$ on a taxa set X is a bipartition of X into two non-empty disjoint subsets $A, B \subseteq X$ whose union is X. For ease of notation, we will write $A|B$ or, equivalently $B|A$ for the split $\{A, B\}$. In general, a collection of splits of X is called a *split system* of X.

Suppose that T is a tree on X. Then a split system $\mathcal{S}(T)$ can be associated to T in the following way. Consider some edge $e \in E(T)$. Then deleting e induces a split $S_e = A|B$ of the leaf set $\mathcal{L}(T) = X$ where A is the leaf-label set of one of the resulting connected components and B is the leaf-label set of the other. The collection of splits of X obtained by deleting, in turn, every edge in T is the split system $\mathcal{S}(T)$.

A *subtree* T' of T is any tree that can be obtained from T by removing an edge of T and picking of the connected components in the resulting graph[1]. Note that T' can always be thought of as a tree rooted at the unique vertex in $e \cap V(T')$, or as unrooted by suppressing this degree 2 vertex. For convenience, we will always denote the root of a subtree T' of T by $r_{T'}$. Note also that every leaf of T is a subtree of T.

Given two subtrees A and B of T, we call A and B *disjoint* if $V(A) \cap V(B) = \emptyset$. If A and B are disjoint and there exist some vertex $x \in V(T)$ such that $e_{r_A} = \{x, r_A\}, e_{r_B} = \{x, r_B\} \in E(T)$, then we denote the subtree of T with vertex set $V(A) \cup V(B) \cup \{x\}$ and edge set $E(A) \cup E(B) \cup \{e_{r_A}, e_{r_B}\}$ by $A \cup B$.

We conclude this section with a lemma concerning trees that will be helpful throughout the paper. Given a tree T, we call a pair of leaves a, b in T which are incident with the same vertex a *cherry* of T, and denote the set of cherries of T by $\mathcal{C}(T)$.

Lemma 2.1: Suppose T and T^* are two trees with distinct topologies. Then there exist disjoint subtrees B, D in T such that B, D, and $B \cup D$ are subtrees of T^* but $B \cup D$ is not a subtree of T.

Proof: Suppose T and T^* are two trees with distinct topologies. To prove the lemma, we distinguish between the cases that (a) there exist elements $x, y \in X$ such that x and y form a cherry in T^* but not in T, and (b) $\mathcal{C}(T^*) \subseteq \mathcal{C}(T)$.

Suppose that (a) holds, i.e., there exist $x, y \in X$ such that x and y form a cherry in T^* but

[1]Note that this definition of a subtree is more restrictive than the one that is commonly used, as described in e.g. [25].

not in T. Then taking B to be the subtree x and D to be the subtree y, the statement holds.

Now suppose (b) holds, i.e., $\mathcal{C}(T^*) \subseteq \mathcal{C}(T)$. Associate to T and T^* new trees $\overline{T}$ and $\overline{T^*}$, respectively, by contracting every cherry, with labels a and b say, of $\mathcal{C}(T^*)$ in both T and T^*, into a leaf which we label $\{a, b\}$. Clearly, since T and T^* have distinct topologies, $\overline{T}$ and $\overline{T^*}$ have distinct topologies.

Now, define $\overline{X}$ to be the leaf-label set of $\overline{T}$. If there exist $x, y \in \overline{X}$ such that x and y form a cherry in $\overline{T^*}$ but not in $\overline{T}$, then we define the trees $\overline{B}$ and $\overline{D}$ as described in case (a) (with X, T, and T^* replaced by $\overline{X}$, $\overline{T}$ and $\overline{T^*}$, respectively). The required subtrees B and D of T and T^* can then be obtained from $\overline{B}$ and $\overline{D}$ by expanding every leaf labelled by a subset A of X of size 2, to a cherry with label set A. If, on the other hand, $\mathcal{C}(\overline{T^*}) \subseteq \mathcal{C}(\overline{T})$, then we iterate the contraction process until we have found two binary leaf labelled trees $\overline{\overline{T}}$ and $\overline{\overline{T^*}}$ for which there is a cherry in $\mathcal{C}(\overline{\overline{T^*}})$ which is not in $\mathcal{C}(\overline{\overline{T}})$. From this cherry we obtain $\overline{\overline{B}}$ and $\overline{\overline{D}}$, and the required subtrees B and D of T and T^* can then be obtained by repeatedly applying the above described expansion process. ■

III. Robinson-Foulds distance

The *Robinson-Foulds distance* [22] is tree comparison metric that is commonly used to measure dissimilarity between phylogenetic trees on the same leaf set. For two trees T_1 and T_2 on X, it is defined by

$$d_{RF}(T_1, T_2) = |\mathcal{S}(T_1) - \mathcal{S}(T_2)| + |\mathcal{S}(T_2) - \mathcal{S}(T_1)|.$$

Note that T_1 and T_2 have the same *topology* if and only if $d_{RF}(T_1, T_2) = 0$.

In this section, we prove the following result.

Theorem 3.1: If T^* is a fixed tree and T is any other tree, then there is a sequence of trees $T_0 = T, T_1, \ldots, T_k = T^*$, such that

1) tree T_{i+1} is obtained from T_i by a single SPR-operation, and

2) $d_{RF}(T_i, T^*) - d_{RF}(T_{i+1}, T^*) > 0$,

for all $0 \leq i \leq k-1$.

This result is a direct consequence of the following lemma. For two trees T_1 and T_2 the *SPR-distance* $d_{SPR}(T_1, T_2)$ between T_1 and T_2 is the minimal number of SPR-operations needed to transform the topology of T_1 into that one of T_2.

Lemma 3.2: Suppose T and T^* are two trees with distinct topologies. Then there exists a tree T' such that $d_{SPR}(T, T') = 1$ and $d_{RF}(T^*, T') < d_{RF}(T^*, T)$.

Proof: Suppose T and T^* are two trees with distinct topology. Then, by Lemma 2.1, there exist disjoint subtrees B, D in T such that B and D are subtrees of T^* and the subtree $B \cup D$ is also a subtree of T^* but not of T. Consider the tree T' obtained from T by pruning the subtree B and regrafting it adjacent to D (see Fig. 3) giving rise to a new vertex p. Clearly, $d_{SPR}(T, T') = 1$.

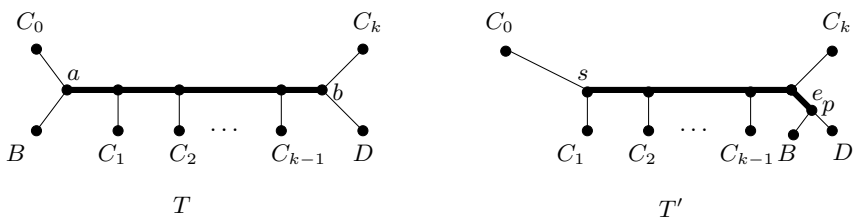

Fig. 3. The trees T and T' considered in the proof of Lemma 3.2.

To see that the inequality stated in the lemma holds, we distinguish between two types of splits displayed by T. For R denoting either T or T', let $\mathcal{S}_b(R)$ denote the set of splits in $\mathcal{S}(R)$ which correspond to the edges in the path from a to b in case $R = T$ and the edges in the path from s to p in case $R = T'$. For the convenience of the reader we indicate these edges in bold (see Fig. 3). Put $\mathcal{S}_{nb}(R) = \mathcal{S}(R) - \mathcal{S}_b(R)$. Note that the latter set also contains those splits that correspond to an edge in the subtrees B, D or in one of the subtrees of R indicated by $C_0, \ldots, C_k$, $k \geq 0$, in Fig. 3.

Now suppose that S is a split on X. Then, by construction, $S \in \mathcal{S}_{nb}(T)$ if and only if $S \in \mathcal{S}_{nb}(T')$. Let $S_1 = \mathcal{L}(B)|X - \mathcal{L}(B)$ and $S_2 = \mathcal{L}(D)|X - \mathcal{L}(D)$. Note that $S_1, S_2 \in \mathcal{S}_{nb}(T) \cap \mathcal{S}_{nb}(T') \cap \mathcal{S}(T^*)$. Let S_e denote the split in $\mathcal{S}(T')$ that corresponds to the edge $e \in E(T')$ as specified in Fig. 3. Observe that

1) $\mathcal{S}_{nb}(T) = \mathcal{S}_{nb}(T')$,

2) $\mathcal{S}_b(T) \cap \mathcal{S}(T^*) = \emptyset$, since the only splits of T^* which separate B and D are S_1 and S_2,

3) $\mathcal{S}_b(T') \cap \mathcal{S}(T^*) \neq \emptyset$ since S_e is a split of T' and T^*.

Hence it follows that

$$\begin{aligned} |\mathcal{S}(T^*) - \mathcal{S}(T)| &= |\mathcal{S}(T^*) - \mathcal{S}_{nb}(T) - \mathcal{S}_b(T)| \\ &= |\mathcal{S}(T^*) - \mathcal{S}_{nb}(T') - \mathcal{S}_b(T)| \\ &> |\mathcal{S}(T^*) - \mathcal{S}_{nb}(T') - \mathcal{S}_b(T')| \\ &= |\mathcal{S}(T^*) - \mathcal{S}(T')|. \end{aligned}$$

Since the trees are binary, they all have the same number of internal edges and hence splits. Thus

$$|\mathcal{S}(T) - \mathcal{S}(T^*)| = |\mathcal{S}(T^*) - \mathcal{S}(T)| > |\mathcal{S}(T^*) - \mathcal{S}(T')| = |\mathcal{S}(T') - \mathcal{S}(T^*)|.$$

The inequality stated in the lemma follows. ■

IV. Quartet distance

In this section we prove the following analogous result to Theorem 3.1 in which we replace the Robinson-Foulds distance d_{RF} by the quartet distance d_Q, another popular tree-comparison metric [5], [11], [19], [27].

We start with recalling the definition of the quartet distance. Let $Q(X)$ denote the set of all *quartets* of X, that is splits $A|B$ of subsets of X of size 4 with $|A| = 2 = |B|$. For brevity, we write $ab|cd$ rather than $\{a, b\}|\{c, d\}$ with $\{a, b, c, d\} \subseteq X$. For a tree T and a quartet $ab|cd$, we say that T *displays* $ab|cd$ if there exists some split $A|B \in \mathcal{S}(T)$ such that $a, b \in A$ and $c, d \in B$. Let $Q(T)$ denote the set of all quartets displayed by a tree T. Then for two trees T_1 and T_2 the *quartet distance* $d_Q(T_1, T_2)$ between T_1 and T_2 is defined as

$$d_Q(T_1, T_2) = |Q(T_1) - Q(T_2)| + |Q(T_2) - Q(T_1)|.$$

In contrast to the Robinson-Foulds distance, the quartet distance between any tree T and the optimal tree T^* can be directly estimated from the data. For example, the popular Quartet Puzzling algorithm [28], first estimates all quartets using maximum-likelihood based on the sequences corresponding to each of the taxa, and then builds a tree in a greedy way, trying to maximize the number of quartets being displayed by the inferred tree. Theorem 4.1 is thus related to the consistency of SPR-moves when the input is given in terms of quartets. In particular, assuming that these quartets exactly correspond to a phylogenetic tree T^*, it shows that we are

able to recover T^* starting from any tree T by simply applying SPR moves and using the quartet distance.

Theorem 4.1: If T^* is a fixed tree and T is any other tree, then there is a sequence of trees $T_0 = T, T_1, \ldots, T_k = T^*$, such that

1) tree T_{i+1} is obtained from T_i by a single SPR-operation, and

2) $d_Q(T_i, T^*) - d_Q(T_{i+1}, T^*) > 0$,

for all $0 \leq i \leq k-1$.

Theorem 4.1 is a direct consequence of the following lemma which is an analogue of Lemma 3.2.

Lemma 4.2: Let T and T^* be two trees with distinct topologies. Then there exists a tree T' such that $d_{SPR}(T, T') = 1$ and $d_Q(T^*, T') < d_Q(T^*, T)$.

Proof: Let B and D denote two disjoint subtrees of T and T^* such that $B \cup D$ is a subtree of T^* but not of T (which must exist by Lemma 2.1). We consider the following two trees: T' formed by pruning B and regrafting it adjacent to D, and T'' formed by pruning D and regrafting it adjacent to B.

For $R \in \{T, T', T''\}$ we consider a partition of the set $Q(R)$ of displayed quartets into four classes $Q_0^R, Q_1^R, Q_2^R, Q_3^R$ defined as follows.

$$\begin{aligned}
Q_0^R &= \{wx|yz \in Q(R) : \text{either } |\{w,x,y,z\} \cap B| > 1 \text{ or } |\{w,x,y,z\} \cap D| > 1 \\
&\qquad\qquad \text{or } |\{w,x,y,z\} \cap B| = 0 = \{w,x,y,z\} \cap D|\}, \\
Q_1^R &= \{wx|yz \in Q(R) : |\{w,x,y,z\} \cap B| = 1 \text{ and } |\{w,x,y,z\} \cap D| = 0\}, \\
Q_2^R &= \{wx|yz \in Q(R) : |\{w,x,y,z\} \cap B| = 0 \text{ and } |\{w,x,y,z\} \cap D| = 1\},
\end{aligned}$$

and

$$Q_3^R = \{wx|yz \in Q(R) : |\{w,x,y,z\} \cap B| = 1 = |\{w,x,y,z\} \cap D|\}$$

Note that

$$Q_0^T = Q_0^{T'} = Q_0^{T''}, \tag{2}$$

and

$$|Q_3^T \cap Q(T^*)| < |Q_3^{T'} \cap Q(T^*)| = |Q_3^{T''} \cap Q(T^*)|. \tag{3}$$

For $R \in \{T, T', T''\}$, a fixed leaf x, and $j \in \{0, 1, 2, 3\}$, let $Q_j^R(x)$ be the subset of Q_j^R consisting of quartets containing x. Now fix some $b \in B$. Observe that since B is a subtree of T^*,

$$|Q_1^T \cap Q(T^*)| = |B||Q_1^T(b) \cap Q(T^*)|.$$

Similarly, for a fixed leaf $d \in D$, we have

$$|Q_2^T \cap Q(T^*)| = |D||Q_2^T(d) \cap Q(T^*)|.$$

Moreover, since B and D are adjacent in T^* we can conclude that

$$|Q_1^{T''} \cap Q(T^*)| = \ |B||Q_1^T(b) \cap Q(T^*)| \text{ and } |Q_2^{T''} \cap Q(T^*)| = \ |D||Q_1^T(b) \cap Q(T^*)|.$$

Similarly, we can conclude that

$$|Q_1^{T'} \cap Q(T^*)| = \ |B||Q_2^T(d) \cap Q(T^*)| \text{ and } |Q_2^{T'} \cap Q(T^*)| = \ |D||Q_2^T(d) \cap Q(T^*)|.$$

Hence

$$\begin{aligned} & |(Q_1^{T''} \cup Q_2^{T''}) \cap Q(T^*)| - |(Q_1^T \cup Q_2^T) \cap Q(T^*)| \\ = \ & |D|(|Q_1^T(b) \cap Q(T^*)| - |Q_2^T(d) \cap Q(T^*)|), \end{aligned}$$

and

$$\begin{aligned} & |(Q_1^{T'} \cup Q_2^{T'}) \cap Q(T^*)| - |(Q_1^T \cup Q_2^T) \cap Q(T^*)| \\ = \ & |B|(|Q_2^T(d) \cap Q(T^*)| - |Q_1^T(b) \cap Q(T^*)|). \end{aligned}$$

Since these cannot both be negative, and by (2) and (3), either

$$|Q(T) \cap Q(T^*)| < |Q(T') \cap Q(T^*)|$$

or

$$|Q(T) \cap Q(T^*)| < |Q(T'') \cap Q(T^*)|$$

holds. The result now follows. ■

V. SPR MOVES AND THE BME TREE LENGTH

In this section we prove the main result of the paper (Theorem 5.2), from which it immediately follows that the BSPR algorithm is consistent with safety radius $\frac{1}{3}$. Note that for the rest of this section we assume that we are given a matrix δ of estimated distances on X, which corresponds in practise to estimated evolutionary distances between elements of X.

The key tool used in our proof is [6, Equation 10] which we now recall. First, for any tree R and for any two disjoint subtrees U and V of R, we define the *balanced average distance* δ^R_{UV} between the leaf sets of U and V recursively as follows. If U and V only contain a single taxa u and v, respectively, then δ^R_{UV} equals the estimated distance δ_{uv} between u and v. Moreover, if one of U and V, say V, is of the form $V = V_1 \cup V_2$ for disjoint subtrees V_1 and V_2 then

$$\delta^R_{UV} = \delta^R_{U(V_1 \cup V_2)} = \frac{1}{2}(\delta^R_{UV_1} + \delta^R_{UV_2}). \tag{4}$$

This definition is motivated by the observation that in biological studies a single isolated taxon often gives as much information as a cluster containing several remote taxa [24]. Also by placing less weight on pairs of taxa that are separated by numerous edges it addresses the problem that long evolutionary distances are poorly estimated (see [8, Section 1.2.7)] and [6] for more details).

Now, let T be the tree on the left in Fig. 1(a) and T' be the tree obtained from T by interchanging the subtrees B and C of T (i.e. T' is the tree depicted in the right of Fig. 1(a)). Then, with the total tree length as defined by (1) in the introduction, [6, Equation 10] states that

$$\hat{l}(T) - \hat{l}(T') = \frac{1}{4}[(\delta^T_{AB} + \delta^T_{CD}) - (\delta^T_{AC} + \delta^T_{BD})]. \tag{5}$$

As mentioned in the introduction, this formula allows a significant improvement of the efficiency of the BNNI algorithm [6].

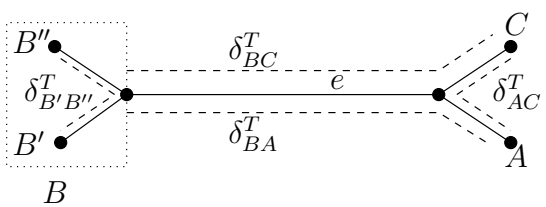

Fig. 4. Edge length estimation from average distance between subtrees using Equation (6).

Moreover, the balanced framework allows for simple edge length estimators [20] (see also [7]). Let e be the branch shown in Fig. 4, and assume B is composed of two disjoint subtrees B', B'', i.e. $B = B' \cup B''$. The estimated length of e is then equal to:

$$\hat{l}(e) = \frac{1}{2}(-\delta^T_{B'B''} + \delta^T_{BA} + \delta^T_{BC} - \delta^T_{AC}), \tag{6}$$

where the same formula holds if B is a leaf by defining $\delta^T_{B'B''} = 0$.

As a first step towards proving Theorem 5.2 we look at how a single SPR-operation applied to a tree T affects the total tree length of T.

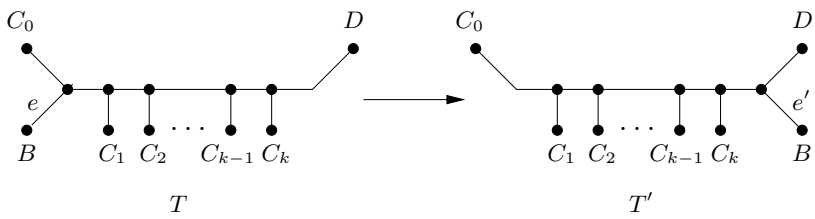

Fig. 5. The trees T and T' have SPR-distance 1; $C_0, \ldots, C_k$, B and D denote subtrees of T (or equivalently of T').

Lemma 5.1: Let T and T' be the trees given in Fig. 5, so that T' can be obtained from T by a single SPR operation in which subtree B is pruned and regrafted. Then $\hat{l}(T) - \hat{l}(T') =$

$$\left(\frac{1}{2} - \frac{1}{2^{k+1}}\right)\left(\delta^T_{C_0B} - \delta^T_{BD}\right) + \sum_{i=1}^{k}\left[\frac{1}{2^{k-i+2}}(\delta^T_{C_iD} - \delta^T_{C_iB}) - \frac{1}{2^{i+1}}(\delta^T_{C_0C_i} - \delta^T_{C_iB})\right].$$

Proof: We first provide a reformulation of (5), which gives the difference in tree length when performing one NNI operation. Let T and T' be the two trees in Fig. 1(a), in which T' is obtained from T by using a single NNI operation, and let e and e' be the edges connecting B in T and T', respectively. Using (4), (5) and (6) it follows that

$$\begin{aligned}
\hat{l}(e) - \hat{l}(e') &= \frac{1}{2}(-\delta^T_{B'B''} + \delta^T_{BA} + \delta^T_{B(C\cup D)} - \delta^T_{A(C\cup D)}) \\
&\quad -\frac{1}{2}(-\delta^T_{B'B''} + \delta^T_{BD} + \delta^T_{B(A\cup C)} - \delta^T_{D(A\cup C)}) \\
&= \frac{1}{4}(\delta^T_{AB} + \delta^T_{CD} - \delta^T_{AC} - \delta^T_{BD}) \\
&= \hat{l}(T) - \hat{l}(T').
\end{aligned}$$

In other words, the difference in tree length is simply the difference between the lengths of edges e and e'.

We now show that this property also holds for SPR moves. Let T and T' be the two trees shown in Fig. 5, and let e and e' denote the edges connecting B in T and T', respectively. Moreover, consider the series of trees $T = T_0, T_1, T_2, \ldots, T_k = T'$, where T_1 is obtained from T by one NNI move exchanging B and C_1, T_2 is obtained from T_1 by one NNI move exchanging B and C_2, ..., T' is obtained from T_{k-1} by one NNI move exchanging B and C_k. Let $e = e_i$ be the edge connecting B in T_i. Just as with the NNI move, we have

$$\hat{l}(T) - \hat{l}(T') = \sum_{i=0}^{k-1} \hat{l}(T_i) - \hat{l}(T_{i+1}) = \sum_{i=0}^{k-1} \hat{l}(e_i) - \hat{l}(e_{i+1}) = \hat{l}(e) - \hat{l}(e').$$

Using the equation above and Equations (4) and (6), it follows that

$$\begin{aligned}\hat{l}(T) - \hat{l}(T') &= \frac{\delta^T_{BC_0}}{2} + \sum_{i=1}^{k} \frac{\delta^T_{BC_i}}{2^{i+1}} + \frac{\delta^T_{BD}}{2^{k+1}} - \sum_{i=1}^{k} \frac{\delta^T_{C_0C_i}}{2^{i+1}} - \frac{\delta^T_{DC_0}}{2^{k+1}} \\ &\quad - \left(\frac{\delta^{T'}_{BD}}{2} + \sum_{i=1}^{k} \frac{\delta^{T'}_{BC_i}}{2^{k-i+2}} + \frac{\delta^{T'}_{BC_0}}{2^{k+1}} - \sum_{i=1}^{k} \frac{\delta^{T'}_{DC_i}}{2^{k-i+2}} - \frac{\delta^{T'}_{DC_0}}{2^{k+1}} \right).\end{aligned}$$

Since the topological structure within each labelled subtree of Fig. 5 is the same in T and T', we have $\delta^T_{UV} = \delta^{T'}_{UV}$ for all $U, V \in \{B, C_0, \ldots, C_k, D\}$. The lemma now follows by simplifying this formula. ■

We now prove our main result. Suppose T^* is a fixed edge-weighted phylogenetic tree on X and, for any edge e of T^*, denote the length of e in T^* by $l(e)$. In addition, let δ^* denote the distance on X defined by taking shortest paths between the leaves of T^* so that, in particular, δ^* is a binary tree-metric. Recall that we also have a matrix δ containing estimates of the distances given by δ^*.

Theorem 5.2: Let T be a tree having a different topology to T^*. Let B and D be disjoint subtrees in T such that B, D, and $B \cup D$ are subtrees of T^* but $B \cup D$ is not a subtree of T. Let T' be obtained from T by pruning the subtree B and regrafting it adjacent to D. Then provided that $|\delta_{ab} - \delta^*_{ab}| < \epsilon := \frac{1}{3} \min_{e \in E(T^*)} l(e)$ for all $a, b \in X$, we have $\hat{l}(T) - \hat{l}(T') > 0$.

Proof: Note that B and D are well defined by Lemma 2.1. Let $C_0, \ldots, C_k$ denote the subtrees depicted in Fig. 5, as in Lemma 5.1. For notational simplicity, for any two disjoint subtrees U, V of T we will write δ_{UV} for δ^T_{UV}, and for any subtree U of T and leaf $v \notin U$ we will write δ_{Uv} for $\delta^T_{U\{v\}}$. Let x be the parent vertex of subtrees B and D in T^*. Let e_x be the edge adjacent to x but not B or D, (see Fig. 6). Then for any subtree A in T^* disjoint with B

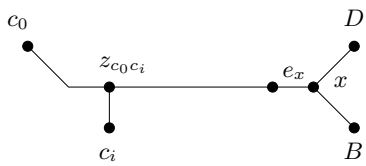

Fig. 6. Sketch illustrating the proof of Theorem 5.2

we have $\delta_{AB} = \sum_{b\in B} 2^{1-p_{xb}}\delta_{Ab}$, where p_{xb} is the number of edges in the path from x to b in T^*. Likewise $\delta_{AD} = \sum_{d\in D} 2^{1-p_{xd}}\delta_{Ad}$. Since $\sum_{b\in B} 2^{1-p_{xb}} = 1 = \sum_{d\in D} 2^{1-p_{xd}}$, Lemma 5.1 yields

$$\hat{l}(T) - \hat{l}(T') = \sum_{b\in B, d\in D} 2^{2-p_{xb}-p_{xd}} \left[\left(\frac{1}{2} - \frac{1}{2^{k+1}}\right)(\delta_{C_0b} - \delta_{bd}) + \sum_{i=1}^{k} \left[\frac{1}{2^{k-i+2}}(\delta_{C_id} - \delta_{C_ib}) - \frac{1}{2^{i+1}}(\delta_{C_0C_i} - \delta_{C_ib}) \right] \right]. \tag{7}$$

We now consider a specific pair $b \in B$ and $d \in D$ and examine its contribution to the summation over b and d in (7). To this end, we denote the sum of the lengths of the edges in the path P_{xb} between x and b in T^* by δ^*_{xb}, and similarly define δ^*_{xd}.

Since the path in T^* from any taxon in C_i to any taxon in B or D must pass through x, and the error in any estimated distance is at most ϵ, we have

$$\begin{aligned} \sum_{i=1}^{k} \frac{1}{2^{k-i+2}}(\delta_{C_id} - \delta_{C_ib}) &\geq \sum_{i=1}^{k} \frac{1}{2^{k-i+2}}(\delta^*_{C_id} - \delta^*_{C_ib} - 2\epsilon) \\ &= \left(\frac{1}{2} - \frac{1}{2^{k+1}}\right)(\delta^*_{xd} - \delta^*_{xb} - 2\epsilon). \end{aligned}$$

and also

$$\left(\frac{1}{2} - \frac{1}{2^{k+1}}\right)(-\delta_{bd}) \geq \left(\frac{1}{2} - \frac{1}{2^{k+1}}\right)(-\delta^*_{xd} - \delta^*_{xb} - \epsilon).$$

In addition

$$\left(\frac{1}{2} - \frac{1}{2^{k+1}}\right)\delta_{C_0b} = \sum_{i=1}^{k} \left[\frac{1}{2^{i+1}}\delta_{C_0b}\right].$$

Hence, (7) implies

$$\hat{l}(T) - \hat{l}(T') \geq \sum_{b\in B, d\in D} 2^{2-p_{xb}-p_{xd}} \left[\left(\frac{1}{2} - \frac{1}{2^{k+1}}\right)(-2\delta^*_{xb} - 3\epsilon) + \sum_{i=1}^{k} \left[\frac{1}{2^{i+1}}(\delta_{C_0b} - \delta_{C_0C_i} + \delta_{C_ib})\right] \right]. \tag{8}$$

Now consider the term $(\delta_{C_0b} - \delta_{C_0C_i} + \delta_{C_ib})$. For $c_0 \in C_0, c_i \in C_i$ let $z_{c_0c_i}$ be the vertex in T^* on the path between c_0 and c_i at which the path to the subtree $B \cup D$ is attached, (see Fig. 6). Then

$$\begin{aligned}
(\delta_{c_0b} - \delta_{c_0c_i} + \delta_{c_ib}) &\geq (\delta^*_{c_0b} - \delta^*_{c_0c_i} + \delta^*_{c_ib} - 3\epsilon) \\
&= (\delta^*_{c_0z_{c_0c_i}} + \delta^*_{z_{c_0c_i}b} - \delta^*_{c_0z_{c_0c_i}} - \delta^*_{z_{c_0c_i}c_i} + \delta^*_{c_iz_{c_0c_i}} + \delta^*_{z_{c_0c_i}b} - 3\epsilon) \\
&= 2\delta^*_{z_{c_0c_i}b} - 3\epsilon \\
&\geq 2l(e_x) + 2\delta^*_{xb} - 3\epsilon.
\end{aligned}$$

It follows that $(\delta_{C_0b} - \delta_{C_0C_i} + \delta_{C_ib}) \geq 2l(e_x) + 2\delta^*_{xb} - 3\epsilon$, and therefore (8) implies

$$\begin{aligned}
\hat{l}(T) - \hat{l}(T') &\geq \sum_{b\in B, d\in D} 2^{2-p_{xb}-p_{xd}} \left[\left(\frac{1}{2} - \frac{1}{2^{k+1}} \right) (-2\delta^*_{xb} - 3\epsilon) \right. \\
&+ \left. \sum_{i=1}^{k} \left[\frac{1}{2^{i+1}} (2l(e_x) + 2\delta^*_{xb} - 3\epsilon) \right] \right] \\
&= \sum_{b\in B, d\in D} 2^{2-p_{xb}-p_{xd}} \left[\left(\frac{1}{2} - \frac{1}{2^{k+1}} \right) 2(l(e_x) - 3\epsilon) \right] \\
&= (1 - 2^{-k})(l(e_x) - 3\epsilon) \\
&> 0.
\end{aligned}$$

This completes the proof. ■

We next show that our results imply that the safety radius of the BME principle itself is at least $1/3$. Recall that BSPR and BNNI are only heuristics for finding a tree of minimal tree length. The following corollary states that the tree that achieves the minimal tree length is the correct tree provided that the errors in the distance matrix are at most 1/3 the minimum edge length. In particular, this radius is independent of the method used to find the shortest tree.

Corollary 5.3: Suppose that $|\delta_{ab} - \delta^*_{ab}| < \epsilon := \frac{1}{3} \min_{e\in E(T^*)} l(e)$ for all $a, b \in X$, then the unique phylogenetic tree that minimises tree length relative to δ is T^*.

Proof: Suppose for contradiction that there is a tree T distinct from T^* which minimises tree length relative to δ, i.e. $\hat{l}(T) \leq \hat{l}(T')$ for all trees T'. Thus $\hat{l}(T)$ is minimal relative to δ. By Lemma 2.1 there exist disjoint subtrees B, D in T such that B, D, and $B \cup D$ are subtrees of T^* but $B \cup D$ is not a subtree of T. By Theorem 5.2 there exists a tree T' distinct from T such that $\hat{l}(T) - \hat{l}(T') > 0$, i.e. $\hat{l}(T) > \hat{l}(T')$, contradicting the minimality of $\hat{l}(T)$. ■

VI. Discussion

In this paper, we have shown that the BSPR algorithm is consistent. As noted in the introduction, SPR moves are more general than NNI moves in that any SPR move can be achieved through a sequence of NNI moves (Fig. 1). It would be interesting to know whether BNNI is also consistent.

In addition to consistency, we have shown that BSPR has safety radius of at least 1/3. Can this result be improved or extended to other variants of minimum evolution (ME) and to different search algorithms? We make the following observations.

1) As previously mentioned, no distance based method can have a safety radius greater than than 1/2 [2].
2) We have observed that our results imply that the safety radius of the BME principle itself is at least $1/3$. In particular, this radius is independent of the method used to find the shortest tree. We believe that the BME safety radius should be $1/2$ but a proof remains to be found.
3) Several variants of ME are discussed in the literature and are implemented within various computer programs. The most common, first proposed by Kid and Sgaramella-Zonta [18] and studied in depth by Rzhetsky and Nei [21], estimates tree edge lengths using ordinary least squares (OLS) and defines the tree length estimate to be the sum of the edge length estimates (including the negative ones). In [31], it is shown that this OLS version of ME has safety radius at most $1/4$ as the number of taxa grows large. Moreover, Gascuel and Guillemot [16] have recently shown that OLS-ME actually has safety radius converging to 0 as the number of taxa tends to infinity. These results could explain the poor accuracy of OLS-ME compared to BME, which has been observed in simulations (e.g. [6]). Moreover, it suggests that our approach to proving the consistency of the BSPR algorithm will not apply to the OLS-ME variant without significant modification.

In summary, there are a number of open problems in the context of using topological moves for inferring phylogenetic trees. We believe that this is an important direction for further research, and that such research should yield fundamental insights into the performance of some commonly used tree inference methods.

ACKNOWLEDGEMENT

The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, for hosting them in the context of their Phylogenetics Programme where part of the research presented in this paper was carried out. They would also like to thank the referees for their helpful comments, and Vincent Berry and Sylvain Guillemot for useful discussions on the topic discussed in this paper. MB was supported by an EPSRC postdoctoral fellowship (EP/D063574/1) and OG was supported by ModelPhylo project of ACI-IMPBIO. KTH and VM would like to thank LIRMM, Montpellier for hosting them during the first stages of this work.

REFERENCES

[1] B. Allen and M. Steel, "Subtree transfer operations and their induced metrics on evolutionary trees", *Annals Comb.*, vol. 5, pp. 1-13, 2001.

[2] K. Atteson, "The performance of the neighbor-joining methods of phylogenetic reconstruction", *Algorithmica*, vol. 25, pp. 251-278, 1999.

[3] W. J. Bruno, N. D. Socci, and A. L. ,Halpern, "Weighted Neighbor Joining: A likelihood based approach to distance-based phylogeny reconstruction", *Mol Biol Evol*, vol. 17 pp. 189-197, 2000.

[4] D. Bryant, "The splits in the neighbourhood of a tree", *Annals. Comb.*, vol. 8 pp.1-11, 2004.

[5] D. Bryant, J. Tsang, P. E. Kearney, and M. Li, "Computing the quartet distance between evolutionary trees", In *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms, N. Y., Jan. 9-11*, pp. 285-286, 2000.

[6] R. Desper and O. Gascuel, "Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle", *J. Comp. Biol.* vol. 9, pp. 587-598, 2002. Latest software available at `http://atgc.lirmm.fr/fastme/`.

[7] R. Desper and O. Gascuel, "Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting", *Mol. Biol. Evol.*, vol. 21, pp. 587-598, 2004.

[8] R. Desper and O. Gascuel, "The minimum-evolution distance based approach to phylogenetic inference", in *Mathematics of Evolution and Phylogeny*, Gascuel O., Ed., Oxford University Press, 2005.

[9] R. Desper and O. Gascuel, "Distance-based Phylogeny Reconstruction (Optimal Radius; **1999, Atteson; 2005, Elias and Lagergren)", in *Encyclopedia of Algorithms*, Ming-Yang Kao, Ed., Springer, in press.

[10] R. Desper, Lefort, Phan and O. Gascuel, manuscript in preparation.

[11] G. F. Estabrook, F. R. McMorris, and C. A. Meacham, "Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units", *Syst. Zool.*, vol. 34, pp. 192-200, 1985.

[12] J. Felsenstein, "PHYLIP - Phylogeny inference package (Version 3.2)", *Cladistics*, vol. 5, pp. 164-166, 1989.

[13] J. Felsenstein, "An alternating least-squares approach to inferring phylogenies from pairwise distances", *Syst. Biol.*, vol. 46, pp. 101-111, 1997.

[14] J. Felsenstein, *Inferring Phylogenies*. Sinauer Associates, Inc., 2004.

[15] O. Gascuel, "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data", *Mol. Biol. Evol.* vol.14, pp. 685-695, 1997.

[16] O. Gascuel and S. Guillemot, manuscript.

[17] W. Hordijk and O. Gascuel, “Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood”, *Bioinformatics* vol. 21(24), pp. 4338-4347, 2005.

[18] K. K. Kidd and L. A. Sgaramella-Zonta, “Phylogenetic analysis: concepts and methods”, *Am. J. Human Genet.* vol. 23, pp. 235-252, 1971.

[19] T. Mailund and C. N. S. Pedersen, “QDist — Quartet distance between evolutionary trees”, *Bioinformatics* vol. 20(10), pp. 1363-1637, 2004.

[20] Y. Pauplin, “Direct calculation of tree length using a distance matrix”, *J. Mol. Evol.* , vol. 51 pp. 66-85, 2000.

[21] A. Rzhetsky and M. Nei, “Theoretical foundation of the minimum-evolution method of phylogenetic inference”, *Mol. Biol. Evol.* vol. 10, pp. 1073-1095, 1993.

[22] D. Robinson and L. Foulds, “Comparison of phylogenetic trees”, *Math. Biosciences*, vol. 53 pp. 131-147, 1981.

[23] N. Saitou and M. Nei, “The neighbor-joining method: A new method for reconstructing phylogenetic trees”, *Mol. Biol. Evol.* , vol. 4, pp. 406-424, 1987.

[24] P. H. A Sneath and R. R. Sokal, “Numerical Taxonomy” pp. 230-234, W. K. Freeman and Company, San Francisco, CA.

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

[26] C. Semple and M. Steel, “Cyclic permutations and evolutionary trees”, *Adv. Appl. Math.* , vol. 32 pp.669-680, 2004.

[27] M. Steel and D. Penny, “Distributions of tree comparison metrics — some new results”, *Syst. Biology*, vol. 42(2) pp. 126-141, 1993.

[28] K. Strimmer and A. von Haeseler, “Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies”, *Mol. Biol. Evol.* , vol. 13 pp. 964-969, 1996.

[29] D. L. Swofford PAUP*, *Phylogenetic analysis using parsimony (* and other methods)*. Sinauer Associates, Sunderland, Massachusetts, 2003.

[30] L. S. Vinh, and A. von Haeseler, “Shortest triplet clustering: reconstructing large phylogenies using representative sets”, *BMC Bioinformatics*, vol. 6(92), pp. 1-14, 2005.

[31] S. J. Willson, “Minimum evolution using ordinary least squares is less robust than neighbor-joining”, *Bull. Math. Biol.* , vol. 67 pp. 261-279, 2005.