Accuracy Guarantees for Phylogeny Reconstruction Algorithms Based on Balanced Minimum Evolution

Magnus Bordewich and Radu Mihaescu

***Abstract*—Distance based phylogenetic methods attempt to reconstruct an accurate phylogenetic tree from an estimated matrix of pair-wise distances between taxa. This paper examines two distance based algorithms (GREEDYBME and FASTME) which are based on the principle of minimising the balanced minimum evolution score of the output tree in relation to the given estimated distance matrix. This is also the principle that underlies the Neighbour-Joining (NJ) algorithm. We show that GREEDYBME and FASTME both reconstruct the entire correct tree if the input data is quartet consistent, and also that if the maximum error of any distance estimate is ϵ, then both algorithms output trees containing all sufficiently long edges of the true tree: those having length at least 3ϵ. That is to say, the algorithms have edge safety radius 1/3. In contrast, quartet consistency of the data is not sufficient to guarantee the NJ algorithm reconstructs the correct tree and, moreover, the NJ algorithm has edge safety radius of 1/4: only edges of the true tree of length at least 4ϵ can be guaranteed to appear in the output. These results give further theoretical support to the experimental evidence suggesting FastME is a more suitable distance based phylogeny reconstruction method than the NJ algorithm.**

***Index Terms*—Phylogenetics, Minimum Evolution, Safety radius, FASTME**

I. INTRODUCTION

A central problem in molecular phylogenetics is to reconstruct an accurate hierarchy of the evolutionary relationships between present day species, or *taxa*, based upon molecular sequence data. A phylogenetic tree is a formal representation of such a hierarchy of the evolutionary relationships, in which the leaves of the tree represent the sampled taxa and the internal nodes represent ancestral taxa. To be precise: a *phylogenetic tree* is a tree whose leaves are bijectively labelled by the elements of some finite set X. A *binary phylogenetic tree* is a phylogenetic tree in which every internal node has degree exactly three. The set X usually denotes a set of species or taxa, and the tree T represents the evolutionary relationships between them. In this paper we investigate two *distance-based methods* for reconstructing phylogenetic trees. A distance-based method is one in which the only input used in reconstruction is a matrix $\delta = [\delta_{ij}]$ whose entries are estimates of the evolutionary distance between each pair of sampled taxa. We analyse two algorithms for inferring binary phylogenetic trees from distance matrices, both based on the *Balanced Minimum Evolution (BME)* principle [5]. In each case the optimality criterion used is to minimize Pauplin's tree-length estimate [15] relative to the given distance matrix. For a binary phylogenetic tree T, given two distinct nodes i, j of T, we define p_{ij}^T to be the number of internal nodes of T which lie on the simple (closed) path between i and j in T. In particular, if i or j are internal, then they also contribute to p_{ij}^T. The *Balanced Minimum Evolution score* (BME score) of T relative to δ is the quantity

$$BME(\delta, T) = \sum_{i,j \in X} 2^{-p_{ij}^T} \delta_{ij}.$$

The minimization problem with objective function given by the BME score relative to a given estimated distance matrix is known as the Balanced Minimum Evolution Problem (BMEP) [4].

The algorithms we consider are GREEDYBME and FASTME. GREEDYBME is a constructive heuristic that greedily minimizes the objective function of the BMEP by adding at each iteration of the algorithm a taxon on a partial binary phylogenetic tree: we start with three (arbitrary) elements of X arranged in a star tree topology and iteratively add each remaining element of X, attaching each as a leaf pendant at the location on the current partial phylogenetic tree that minimises the BME score of the resulting tree (restricted to the elements of X inserted so far). It is interesting to note that Gascuel and Steel, in an excellent review [9], have shown that the Neighbor-Joining algorithm (NJ) of Saitou and Nei [16] is also a greedy heuristic for the BMEP, to be precise a hierarchical clustering heuristic: the NJ algorithm starts with a star topology on *all* taxa and iteratively chooses two nodes adjacent to the central high-degree node and agglomerates them (effectively regrafts the two nodes as a sibling pair attached to the central high degree node), where the two nodes are chosen to minimise the BME score of the resulting tree (see [9] for further details).

FASTME is a hill-climbing heuristic for minimizing the objective function of the BMEP. FASTME starts with a binary phylogenetic tree on X, typically the output of GREEDYBME, and iteratively searches through local

The first author was supported in part by EPSRC grant number EP/D063574/1. The work began while both authors were visiting the University of Canterbury, New Zealand. A preliminary version of this work appeared in the WABI'10 proceedings.

Magnus Bordewich is with the School of Engineering and Computing Sciences, Durham University, U.K. E-mail: m.j.r.bordewich@durham.ac.uk

During this work Radu Mihaescu was with the Department of Computer Science, U.C. Berkeley, U.S.A. E-mail: rmihaescu@gmail.com

topologies (those trees differing from the current tree by one topological rearrangement operation) and moves to the local topology that minimises the BME score. This approach is implemented in a software called FastME [5]. The two topological rearrangement operations available in the latest release of FastME are: the *Balanced Subtree Prune and Regraft (BSPR) algorithm* [10] and the *Balanced Nearest Neighbor Interchange (BNNI) algorithm* [5]. FastME has been shown experimentally by Desper and Gascuel [5], [6] to be a fast and accurate method for tree inference, compared to other popular distance-based methods such as NJ, BIONJ [8], FITCH [7] or WEIGHBOR [3]. The results in this paper provide further theoretical support for using this approach.

Atteson studied the NJ algorithm and gave a condition, the *safety radius* of the algorithm, for accurate reconstruction of the true tree [1]. Atteson showed that the NJ algorithm has safety radius of $1/2$, *i.e.* if the maximum error in the estimated distance matrix is at most half the minimum edge length in the true tree, then the NJ algorithm will correctly reconstruct the entire tree. Moreover, no distance based method can have safety radius greater than $1/2$. More recently Bordewich *et al.* [2] analysed FASTME and showed that it has safety radius at least $1/3$ (when using BSPR) and Shigezumi [18] has shown that GREEDYBME has safety radius $1/2$. Note that FASTME and GREEDYBME are heuristics for finding a tree that minimises BME score. Pardi *et al.* [14] have shown that the BME principle itself, or equivalently any algorithm returning the optimal solution to the BMEP, has safety radius $1/2$.

The results described above relate the minimum edge length in the entire tree to the maximum allowed error in the estimated distance matrix. Thus a single short edge in the true tree can greatly affect the permitted error across all estimated distances for the guarantee of correct reconstruction to hold. In contrast, in this paper we consider the *edge safety radius*, which guarantees that all sufficiently long edges, relative to the error, will be correctly reconstructed even if other edges of the true tree are very short. An algorithm that is guaranteed to output a tree topology containing all those edges of the true tree that have length at least l, whenever the the maximum error in the estimated distance matrix is less than rl, is said to have *edge safety radius* r. Atteson conjectured that the NJ algorithm has an edge safety radius of $1/4$, which has recently been proved [12]. The main result of this paper is to show that GREEDYBME and FASTME each have edge safety radius $1/3$. We also show that under a weaker condition than safety radius $1/2$, namely quartet consistency (which we will define below), GREEDYBME and FASTME will correctly reconstruct the true tree. Note that having maximum error at most $1/2$ the minimum edge length guarantees quartet consistency, but a distance matrix may be quartet consistent with the true tree while not satisfying the safety radius condition. In related work, it has been shown that the solution to the BMEP is guaranteed to be the true tree on quartet consistent inputs, but that an exact algorithm for the BMEP is strictly weaker than the two heuristic versions (GREEDYBME and NJ) in edge safety radius, having an asymptotic edge safety radius of $1/(2n)$ on n taxa [11].

Our results show a strict theoretical superiority of GREEDYBME over the NJ algorithm in two ways. It has been shown that quartet consistency is not a sufficient condition for the NJ algorithm to correctly reconstruct the true tree [12]. Thus GREEDYBME will correctly reconstruct the *whole* true tree under a weaker condition than NJ. Also, even when this condition does not hold GREEDYBME will correctly reconstruct all *edges* of the true tree having length at least 3 times the maximum error in the input matrix, whereas the NJ algorithm sometimes fails to reconstruct edges up to 4 times the maximum error [1]. It is intriguing to note that in the simulated tests of Desper and Gascuel [5] the NJ algorithm marginally outperformed GREEDYBME, particularly on small trees, and in turn FASTME (using BNNI) significantly outperformed the NJ algorithm.

II. Basics, definitions and notation

The notation and terminology largely follows Semple and Steel [17]. 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.

A matrix of pair-wise distances $\delta^* = [\delta^*_{ij}]$ is a *tree-metric* if there is a unique phylogenetic tree T^* with positive edge lengths l_e so that, for each $x, y \in X$, the distance δ^*_{xy} is the sum of the lengths of edges on the path between x and y in T^*. The input to our algorithms is an estimated pair-wise distance matrix $\delta = [\delta_{ij}]$, and the error ϵ of δ with respect to δ^* is $\max_{x,y \in X}(|\delta_{xy} - \delta^*_{xy}|)$.

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 phylogenetic tree on X. Each edge e of T corresponds to a split of $A|B$ of X, which may be obtained by deleting e and letting A be the leaf-label set of one of the resulting connected components and B be the leaf-label set of the other. We write $e = A|B$ to denote the edge and its corresponding split. The set of splits corresponding to edges of T are said to be the splits of T. A *clade* of T is any subset $C \subset X$ such that $C|X - C$ is a split of T. The clade is said to be rooted at the node c in T which is the end of the edge e which induces split $C|X - C$ closer to the leaves in C.

A quartet of T is a partial split $\{a, b\}|\{c, d\}$, where $a, b, c, d \in X$ and there is a split $A|D$ of T such that $a, b \in A$ and $c, d \in D$. For simplicity the quartet $\{a, b\}|\{c, d\}$ will be denoted by $ab|cd$. We say that an estimated distance matrix δ is *consistent with a quartet* $ab|cd$ if $\delta_{ab} + \delta_{cd} < \delta_{ac} + \delta_{bd}, \delta_{ad} + \delta_{bc}$. We say that δ is *consistent with an edge* $e = A|D$ of T, if δ is consistent with all quartets $ab|cd$ of T such that $a, b \in A$ and $c, d \in D$.

We say that δ is *quartet consistent* with a phylogenetic tree T if δ is consistent with all the quartets of T.

Given an estimated distance matrix δ and two disjoint clades A, B of a binary phylogenetic tree T, rooted at nodes a, b respectively, we define the *balanced average clade distance*, or *clade distance* for short, as follows. First, if A and B only contain a single taxa a and b, respectively, then δ_{AB} equals the estimated distance δ_{ab} between a and b. Now, if one of A and B, say B, is of the form $B = B_1 \cup B_2$ for disjoint subtrees B_1, B_2 of T, the roots of which are both children of b in the rooted subtree B, then

$$\delta_{AB} = \frac{1}{2}(\delta_{AB_1} + \delta_{AB_2}). \qquad (1)$$

The definition is extended recursively (see for example [13]) to yield:

$$\delta_{AB} = \sum_{i \in A, j \in B} \delta_{ij} 2^{-(p^T_{ia} + p^T_{jb})}.$$

Note that the clade distance thus only depends on the topologies of the rooted subtrees A and B and not on the entire topology T.

III. Results

We now formally state the main results of this paper. The first result concerns the algorithm GreedyBME, and gives sufficient conditions for accurate reconstruction of edges of the true tree.

Theorem 3.1: Let T^* be a binary phylogenetic tree with induced distance matrix δ^*. Let input matrix δ have error ϵ with respect to δ^*. Then the algorithm GreedyBME will return a binary phylogenetic tree T such that

1) T contains an edge with split $A|B$ for all edges $e = A|B$ in T^* with $l_e > 3\epsilon$, i.e. GreedyBME has edge-safety radius $1/3$. Furthermore, this bound is asymptotically tight.
2) if δ is quartet consistent with T^* then $T = T^*$.

The second of our results concerns the local topology search phase of FastME. We show that if a local search is conducted from a tree T that already contains certain edges from the true tree T^*, then the end result is guaranteed to also contain these edges of T^*. The two forms of local topology search considered are those which search over local topologies within one NNI and within one SPR operation of the current tree; for details of the definitions of NNI and SPR operations as used in FastME see, for example, [2].

Theorem 3.2: Let T^* be a binary phylogenetic tree with induced distance matrix δ^*. Let input matrix δ have error ϵ with respect to δ^*. Let T be a binary phylogenetic tree and let $e = A|B$ be an edge common to T and T^*. Then

1) if $l_e > 2\epsilon$ then for any T' that may be obtained in one NNI operation from T such that $BME(\delta, T') < BME(\delta, T)$, e must be an edge of T';
2) if $l_e > 3\epsilon$ and T' is the tree at most one SPR operation from T which minimises $BME(\delta, T')$ then e must be an edge of T'; and
3) if δ is consistent with e then for any T' that may be obtained in one NNI operation from T such that $BME(\delta, T') < BME(\delta, T)$, e must be an edge of T'.

Combining the above two theorems we obtain the following immediate corollary.

Corollary 3.3: FastME using an initial tree generated by GreedyBME and a local search based on NNI or SPR operations has edge safety radius 1/3.

Proof: Let T^* be a phylogenetic tree on X. Let δ be a matrix of pairwise distances that has error at most ϵ with respect to T^*. By Theorem 3.1, GreedyBME will return a tree containing all edges of T^* which have length greater than 3ϵ. By Theorem 3.2 the local topology phase of FastME, using NNI or SPR, will not destroy any of these edges, hence they are still present in the final output. Thus FastME has edge safety radius 1/3. ■

Note that in the local topology search using NNI operations, it would not matter if the algorithm jumped immediately to the first tree found with lower BME score than the current tree or completed the search of all neighbouring trees and moved to the one with lowest BME score. However for a local topology search using SPR operations we have the restriction that all neighbouring trees are checked, and the best of those (the one with lowest BME score) is selected for the next iteration.

IV. Proofs

A. GreedyBME*: proof of Theorem 3.1.*

Theorem 3.1 gives a condition under which the GreedyBME algorithm will correctly reconstruct edges of the true tree, namely that the maximum error in the distance matrix is less than $\frac{1}{3}$ of the edge length, and claims that this condition is tight. It also gives a second condition, quartet consistency, under which GreedyBME reconstructs the entire correct tree topology. In the proof of this theorem we shall make use of the following two lemmas. The first is Lemma 5.1 of [2], which gives a formula for the difference in BME score between two trees of certain structure. The second lemma gives a five point condition on the distance matrix δ, which is extended to clades.

Lemma 4.1 (Lemma 5.1 of [2]): Let T^A and T^B be the trees given in Fig. 1. Let

$$\Delta_i = \frac{1}{2^{t-i+1}}(\delta_{B_iB_t} - \delta_{B_iX_k}) - \frac{1}{2^{i+1}}(\delta_{A'B_i} - \delta_{B_iX_k}).$$

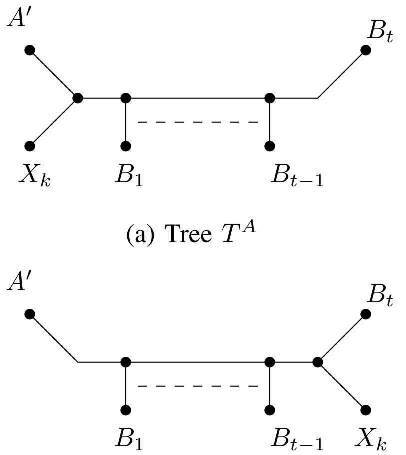
Then $BME(\delta, T^A) - BME(\delta, T^B) =$

$$\left(\frac{1}{2} - \frac{1}{2^t}\right)(\delta_{A'X_k} - \delta_{X_kB_t}) + \sum_{i=1}^{t-1} \Delta_i.$$

For any clades A, B, C, D, E, including single leaf clades, let us write $\Delta^{(*)}_{A,B,C,D,E}$ for

$$(\delta^{(*)}_{AB} + \delta^{(*)}_{AC} - \delta^{(*)}_{BC}) - (\delta^{(*)}_{AD} + \delta^{(*)}_{AE} - \delta^{(*)}_{DE}).$$

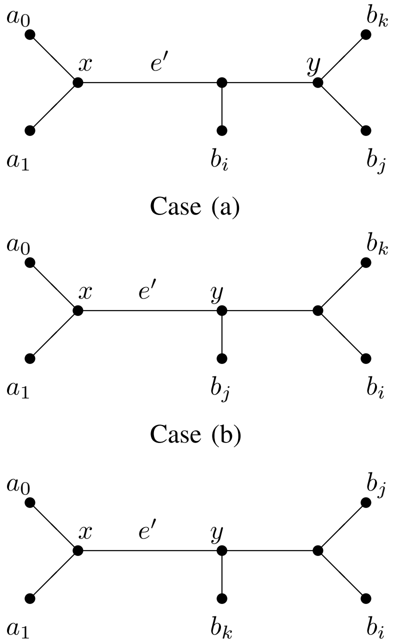
Lemma 4.2: Let T^* be a phylogenetic tree on X and let T^* have a split $e = A|B$ of length l_e. Let δ be a



(a) Tree T^A

(b) Tree T^B

Fig. 1. Two binary phylogenetic trees that differ only in the location of a single clade X_k.



Case (a)

Case (b)

Case (c)

Fig. 2. Possible true topologies relating the leaves a_0, a_1, b_i, b_j, b_k.

matrix of pairwise distances that has error at most $\epsilon < l_e/3$. Then for A_0, A_1 disjoint subsets of A, and B_i, B_j, B_k disjoint subsets of B, and for any tree T with clades A_0, A_1, B_i, B_j, B_k we have:

$$\Delta_{A_0A_1B_iB_jB_k} < 0.$$

Proof: Let a_0, a_1, b_i, b_j, b_k be leaves in clades A_0, A_1, B_i, B_j, B_k respectively. The true topology T^* restricted to these five leaves must be one of the three topologies shown in Fig. 2, where the edge e' depicted represents a path in T^* that contains edge e.

In each case the true distances satisfy

$$\Delta^*_{a_0,a_1,b_i,b_j,b_k} = -2\delta^*_{xy}$$
$$\le -2l_e,$$

where x and y are the corresponding internal nodes of T^*. Since each entry in the estimated distance matrix has an error of at most ϵ, we have

$$\Delta_{a_0,a_1,b_i,b_j,b_k} \le -2l_e + 6\epsilon$$
$$< 0.$$

Note that for any clade C with root r_c we have $\sum_{c \in C} 2^{-p^T_{cr_c}} = 1$. So we may sum all terms over the leaves in all clades. Thus,

$$\Delta_{A_0,A_1,B_i,B_j,B_k} = \sum p^T_{a_0r_0} p^T_{a_1r_1} p^T_{b_ir_i} p^T_{b_jr_j} p^T_{b_kr_k} [\Delta_{a_0,a_1,b_i,b_j,b_k}] < 0$$

where the summation is over all leaves a_0, a_1, b_i, b_j, b_k in A_0, A_1, B_i, B_j and B_k respectively, and r_0, r_1, r_i, r_j, r_k are the roots of the clades A_0, A_1, B_i, B_j and B_k in T respectively. ■

We now present the proof of Theorem 3.1, restated below.

Theorem 4.3 (Restatement of Theorem 3.1): Let T^* be a binary phylogenetic tree with induced distance matrix δ^*. Let input matrix δ have error ϵ with respect to δ^*. Then the algorithm GREEDYBME will return a binary phylogenetic tree T such that

1) T contains an edge with split $A|B$ for all edges $e = A|B$ in T^* with $l_e > 3\epsilon$, i.e. GREEDYBME has edge-safety radius $1/3$. Furthermore, this bound is asymptotically tight.
2) if δ is quartet consistent with T^* then $T = T^*$.

Proof: First we prove that the edge safety radius of the algorithm GREEDYBME is at least $1/3$, giving the first part of item 1. Let T^* be the true tree, and let δ be an estimated pairwise distance matrix with maximum error at most ϵ. Let $e = A|B$ be a split in T^* of length $l_e > 3\epsilon$. The proof is by induction on the size of X. If $|X| = 3$, then trivially GREEDYBME will return the true tree, since there is only one tree topology on three taxa.

Suppose now that the theorem holds for all trees on taxa sets of size at most $k-1$, and let $|X| = k$. Let x_k be the last taxa added by GREEDYBME, and consider the tree T'^* obtained from T^* by removing x_k and its pendent edge. Note that δ' obtained from δ by removing the row and column corresponding to x_k gives a pairwise distance matrix for T'^* with maximum error at most ϵ. Without loss of generality we assume $x_k \in A$. If $A = \{x_k\}$, then trivially GREEDYBME will construct a tree containing the split $A|B$. Now we assume $A' = A - \{x_k\} \neq \emptyset$. Then T'^* has split $e' = A'|B$ and the length of e' is at least l_e. By the inductive hypothesis GREEDYBME applied to δ' will construct a tree containing the split $e' = A'|B$. Thus GREEDYBME applied to δ will also construct a tree T' on $X - \{x_k\}$ containing the split $e' = A'|B$ after $k-1$ steps.

We must now show that after the addition of x_k in the final step of the algorithm the split $A|B$ is present in the resulting tree T. GREEDYBME will position x_k at the point which minimises $BME(\delta, T)$. Suppose for contradiction that there is some position in clade B of T' which minimises this, as depicted in Fig. 1(b), where $B = B_1 \cup B_2 \cup \ldots \cup B_t$ and $X_k = \{x_k\}$, resulting in tree T^B. We will show that tree T^A obtained by attaching x_k between the clades A' and B, as depicted in Fig. 1(a), must obtain a smaller BME score, giving the required contradiction.

By Lemma 4.1, we can express the difference in BME score between T^A and T^B as follows:

$$\begin{aligned} BME(\delta, T^A) - BME(\delta, T^B) &= \\ \left(\frac{1}{2} - \frac{1}{2^t}\right)(\delta_{A'x_k} - \delta_{x_kB_t}) &+ \sum_{i=1}^{t-1} \Delta_i \\ &= \sum_{i=1}^{t-1} \frac{1}{2^{i+1}} \Delta_{x_k,A',B_i,B_{t-i},B_t}. \quad (2) \end{aligned}$$

We may now apply Lemma 4.2 to each term in the summation (setting $A_0 = \{x_k\}, A_1 = A', B_i = B_i, B_j = B_{t-i}$ and $B_k = B_t$) to see that each term is less than zero, and hence $BME(\delta, T^A) < BME(\delta, T^B)$. This contradicts the assumption that x_k will be inserted in clade B, and completes the inductive step.

We now show that the edge safety radius of GreedyBME is no more than $1/3$, giving the second part of item 1. For contradiction assume the edge safety radius is $r > 1/3$. Consider a caterpillar tree T^* in which a', x_k is a cherry (a pair of sibling leaves) separated from the rest of the tree by an edge of length l, as depicted in T^A of Fig. 1(a), taking the clades $A', X_k, B_i, \ldots, B_t$ to be single leaves $a', x_k, b_1, \ldots, b_t$, and taking t to be odd. Let $\epsilon = rl$. We will assume the sum of all other edge lengths in the tree is at most ν for some very small $\nu > 0$. Define $\delta = [\delta_{xy}]$ as follows:

- for $i, j \in [1, \ldots, t-1]$, $\delta_{b_ib_j} = \delta^*_{b_ib_j}$;
- $\delta_{a'b_t} = \delta^*_{a'b_t}$;
- for $i = 1$ to $t-1$, $\delta_{a'b_i} = \delta^*_{a'b_i} - \epsilon$;
- for $i = 1$ to $t-1$, $\delta_{b_ib_t} = \delta^*_{b_ib_t} + \epsilon$;
- $\delta_{a'x_k} = \delta^*_{a'x_k} + \epsilon$;
- for $i = 1$ to $(t-1)/2$, $\delta_{x_kb_i} = \delta^*_{x_kb_i} + \epsilon$;
- and for $i = (t-1)/2 + 1$ to t, $\delta_{x_kb_i} = \delta^*_{x_kb_i} - \epsilon$;

Note that if we remove x_k then δ gives a tree metric for the tree T' with topology $T^* - x_k$, and in which the edge adjacent to a' has shrunk by ϵ and the edge adjacent to b_t has grown by ϵ, all other edge lengths remain the same. If, as assumed, GreedyBME has edge-safety radius r, then given any ordering of the taxa, GreedyBME should reconstruct a tree featuring the split $a'x_k|b_1 \ldots b_t$. We fix the ordering of the taxa to be any ordering in which x_k is the final element. Since δ restricted to $X - \{x_k\}$ is a tree metric, GreedyBME will correctly construct the topology $T^* - x_k$ when given δ and the leaves $a', b_1, \ldots, b_t$ in some order. From this position in the GreedyBME algorithm, we show that x_k can be inserted to form a cherry with b_t at lower BME score than in position forming a cherry with a'. This configuration is exactly as shown in T^B of Fig. 1(b), so again Lemma 4.1 gives

$$\begin{aligned} &BME(\delta, T^A) - BME(\delta, T^B) = \\ &\qquad \left(\frac{1}{2} - \frac{1}{2^t}\right)(\delta_{a'x_k} - \delta_{x_kb_t}) \\ &+ \sum_{i=1}^{t-1} \left[\frac{1}{2^{t-i+1}}(\delta_{b_ib_t} - \delta_{b_ix_k}) - \frac{1}{2^{i+1}}(\delta_{a'b_i} - \delta_{b_ix_k}) \right]. \end{aligned}$$

We evaluate this ignoring all quantities less than ν, which we can set as small as we like. Observing that

$$\left(\frac{1}{2} - \frac{1}{2^t}\right)(\delta_{a'x_k} - \delta_{x_kb_t}) = \left(\frac{1}{2} - \frac{1}{2^t}\right)(\epsilon - l + \epsilon)$$

and

$$\sum_{i=1}^{t-1} \left[\frac{1}{2^{t-i+1}} \delta_{b_ib_t} - \frac{1}{2^{i+1}} \delta_{a'b_i} \right] = \left(\frac{1}{2} - \frac{1}{2^t}\right) 2\epsilon - l$$

we obtain $BME(\delta, T^A) - BME(\delta, T^B) =$

$$\left(\frac{1}{2} - \frac{1}{2^t}\right)(4\epsilon - 2l) + \sum_{i=1}^{t-1} \left[(\frac{1}{2^{i+1}} - \frac{1}{2^{t-i+1}}) \delta_{b_ix_k} \right].$$

We now split the sum depending whether we over- or underestimated $\delta_{b_ix_k}$:

$$\begin{aligned} &\sum_{i=1}^{t-1} \left[(\frac{1}{2^{i+1}} - \frac{1}{2^{t-i+1}}) \delta_{b_ix_k} \right] \\ &= \sum_{i=1}^{(t-1)/2} \left[(\frac{1}{2^{i+1}} - \frac{1}{2^{t-i+1}})(l + \epsilon) \right] \\ &\qquad + \sum_{i=(t-1)/2+1}^{t-1} \left[(\frac{1}{2^{i+1}} - \frac{1}{2^{t-i+1}})(l - \epsilon) \right] \\ &= 2 \sum_{i=1}^{(t-1)/2} \left[\frac{1}{2^{i+1}} \epsilon \right] + 2 \sum_{i=(t-1)/2+1}^{t-1} \left[\frac{1}{2^{i+1}} (-\epsilon) \right] \\ &= 2 \left(\frac{1}{2} - \frac{1}{2^{(t-1)/2}} \left(1 - \frac{1}{2^{(t+1)/2}} \right) \right) \epsilon. \end{aligned}$$

Reinserting this into the main calculation gives $BME(\delta, T^A) - BME(\delta, T^B) =$

$$\begin{aligned} &\frac{(6\epsilon - 2l)}{2} - \frac{(4\epsilon - 2l)}{2^t} - \frac{2}{2^{(t-1)/2}} \left(1 - \frac{1}{2^{(t+1)/2}} \right) \epsilon \\ &= (3r - 1)l - \frac{1}{2^t}(4\epsilon - 2l) - \frac{1}{2^{(t+1)/2}} \left(1 - \frac{1}{2^{(t+1)/2}} \right) \epsilon. \end{aligned}$$

For any $r > 1/3$ we can choose t large enough (and ν small enough) that this is strictly positive, and hence $BME(\delta, T^A) > BME(\delta, T^B)$. Thus GreedyBME will insert x_k in the wrong position and so fail to reconstruct a tree with the required split corresponding to the long edge separating a' and x_k from the rest of the tree. This concludes the proof of part 1 of the theorem.

We now prove item 2 of the theorem: that if T is quartet consistent with T^* then GreedyBME returns the tree topology of T^*. Let T be a binary phylogenetic tree and let

δ be an estimated pairwise distance matrix which is quartet consistent with T. The proof is similar to that of item 1, and is again by induction on the size of X. If $|X| = 4$, then it is easy to check that GREEDYBME will return the true tree.

Suppose now that the theorem holds for all trees on taxa sets of size at most $k-1$, and let $|X| = k$. Let x_k be the last taxa added by GREEDYBME, and consider the tree T' obtained from T by removing x_k and its pendent edge. Note that δ' obtained from δ by removing the row and column corresponding to x_k, gives a pairwise distance matrix which is quartet consistent with the topology T'. Without loss of generality we assume the correct position for x_k to be inserted is as a pendent leaf regrafted to some edge $e = A'|B$ in T'. By the inductive hypothesis GREEDYBME applied to δ' will construct T' correctly. Thus GREEDYBME applied to δ will also construct T' on $X - \{x_k\}$ after $k-1$ steps.

GREEDYBME will position x_k at the point which minimises the score $BME(\delta, T)$. Suppose for contradiction that there is some position other than as a pendent leaf grafted to e which minimises BME score. Without loss of generality we may assume that the position is in clade B of T', as depicted in Fig. 1(b), where $X_k = \{x_k\}$ and $B = B_1 \cup B_2 \cup \ldots \cup B_t$, resulting in tree T^B. We will show that tree T, as depicted in Fig. 1(a), must obtain a smaller BME score, giving the required contradiction.

We will first assume that t is odd: a small adjustment will be needed if t is even. Using (2), we may express the difference in BME score between T^A and T^B as $BME(\delta, T^A) - BME(\delta, T^B) =$

$$\sum_{i=1}^{(t-1)/2} \frac{\Delta_{x_k,A',B_i,B_{t-i},B_t}}{2^{i+1}} + \frac{\Delta_{x_k,A',B_{t-i},B_i,B_t}}{2^{t-i+1}} \tag{3}$$

Note that in contrast to the proof of Theorem 3.1 part 1, in this case we know that the internal topology of the clade B is the same in T' as in T. For each set of leaves $a \in A', b_i \in B_i, b_{t-i} \in B_{t-i}$ and $b_t \in B_t$ $(i \leq (t-1)/2)$, we define

$$f(a, b_i, b_{t-i}, b_t) = \frac{\Delta_{x_k,a,b_i,b_{t-i},b_t}}{2^{i+1}} + \frac{\Delta_{x_k,a,b_{t-i},b_i,b_t}}{2^{t-i+1}}.$$

Since $i < t - i$, by quartet consistency we have

$$\delta_{b_{t-i}x_k} + \delta_{ab_i} > \delta_{ax_k} + \delta_{b_i b_{t-i}}$$

and

$$\delta_{x_k b_t} + \delta_{b_i b_{t-i}} > \delta_{b_i x_k} + \delta_{b_{t-i} b_t}.$$

Combining these conditions gives

$$(\delta_{ax_k} + \delta_{b_i x_k} - \delta_{ab_i}) - (\delta_{x_k b_t} + \delta_{b_{t-i} x_k} - \delta_{b_{t-i} b_t}) < 0.$$

I.e. $\Delta_{x_k,a,b_i,b_{t-i},b_t} < 0$. If in addition $\Delta_{x_k,a,b_{t-1},b_i,b_t} < 0$ then $f(a, b_i, b_{t-i}, b_t) < 0$. On the other hand, if $\Delta_{x_k,a,b_{t-1},b_i,b_t} \geq 0$ then $f(a, b_i, b_{t-i}, b_t)$ is less than

$$\begin{aligned} &\frac{1}{2^{i+1}} \Delta_{x_k,a,b_i,b_{t-i},b_t} + \frac{1}{2^{i+1}} \Delta_{x_k,a,b_{t-i},b_i,b_t} \\ &= \frac{1}{2^{i+1}} (2\delta_{ax_k} + \delta_{b_i b_t} + \delta_{b_{t-i} b_t} - 2\delta_{x_k b_t} - \delta_{ab_i} - \delta_{ab_{t-i}}) \\ &\qquad < 0, \end{aligned}$$

where the final inequality comes from the quartet conditions

$$\delta_{ab_i} + \delta_{x_k b_t} > \delta_{ax_k} + \delta_{b_i b_t}$$

and

$$\delta_{ab_{t-i}} + \delta_{x_k b_t} > \delta_{ax_k} + \delta_{b_{t-i} b_t}.$$

In either case $f(a, b_i, b_{t-i}, b_t) < 0$, hence (by summing over the leaves in each clade) we see that each term of the main summation (3) above is less than zero.

If t is even, then we need to adjust (3) as follows. $BME(\delta, T^A) - BME(\delta, T^B) =$

$$\sum_{i=1}^{t/2-1} \left[\frac{\Delta_{x_k,A',B_i,B_{t-i},B_t}}{2^{i+1}} + \frac{\Delta_{x_k,A',B_{t-i},B_i,B_t}}{2^{t-i+1}} \right] + \frac{\Delta_{x_k,A',B_i,B_{t/2},B_{t/2}}}{2^{t/2+1}}$$

The proof as for t odd shows that the summation is less than zero. The term $\Delta_{x_k,A',B_i,B_{t/2},B_{t/2}}$ is easily seen to be less than zero by quartet consistency, and so again $BME(\delta, T^A) - BME(\delta, T^B) < 0$

Thus $BME(\delta, T^A) < BME(\delta, T^B)$ which contradicts the placement of x_k in clade B. This completes the proof of Theorem 3.1. ■

B. Local Topology Search: proof of Theorem 3.2

Theorem 3.2 gave conditions under which a local topology search based upon NNI or SPR moves would preserve edges of the true tree. In particular, for any edge e common to the tree T and the true tree T^*, any NNI move taken from T will not remove the edge e if the estimated distance matrix δ is consistent with e or if the maximum error in δ is less than $l_e/2$, and the best available SPR move from T will not remove the edge e if the maximum error in δ is less than $l_e/3$. We restate, and then prove, the theorem below.

Theorem 4.4 (Restatement of Theorem 3.2): Let T^* be a binary phylogenetic tree with induced distance matrix δ^*. Let input matrix δ have error ϵ with respect to δ^*. Let T be a binary phylogenetic tree and let $e = A|B$ be an edge common to T and T^*. Then

1) if $l_e > 2\epsilon$ then for any T' that may be obtained in one NNI operation from T such that $BME(\delta, T') < BME(\delta, T)$, e must be an edge of T';
2) if $l_e > 3\epsilon$ and T' is the tree at most one SPR operation from T which minimises $BME(\delta, T')$ then e must be an edge of T'; and
3) if δ is consistent with e then for any T' that may be obtained in one NNI operation from T such that $BME(\delta, T') < BME(\delta, T)$, e must be an edge of T'.

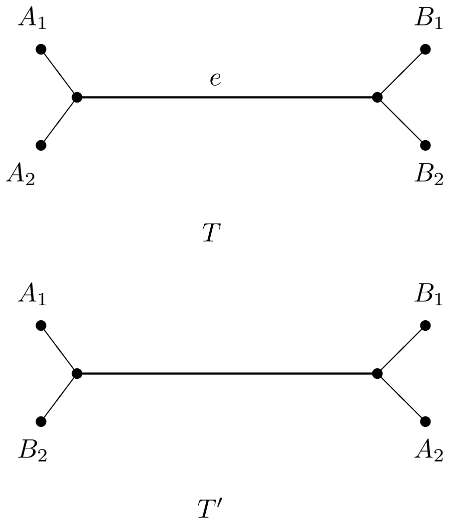


Fig. 3. An NNI operation that breaks split $e = A_1 \cup A_2 | B_1 \cup B_2$.

Proof: We first prove item 3: that an NNI based local topology search will never remove an edge e with which δ is consistent. From item 3 we will then obtain item 1 of the theorem, before proving item 2. Let T, T^* be binary phylogenetic trees with some common edge $e = A|B$. Let δ be a distance matrix consistent with e. If either A or B is a singleton (leaf), then every phylogenetic tree will contain the edge e and there is nothing to prove. Now assume A and B are not singletons. Consider an NNI move that could destroy the split $A|B$. Let A_1, A_2, B_1 and B_2 be the subclades of A and B obtained by dividing A and B at the point of attachment of e, as in Fig. 3. The only way e can be destroyed by an NNI is, without loss of generality, by swapping clades A_2 and B_2, as shown by T' in Fig. 3. By Lemma 4.1

$$\begin{aligned}&BME(\delta, T) - BME(\delta, T') = \\ &\qquad \frac{1}{4}[(\delta_{A_1A_2} + \delta_{B_1B_2}) - (\delta_{A_1B_2} + \delta_{B_1A_2})].\end{aligned}$$

For any leaves a_1, a_2, b_1, b_2 in A_1, A_2, B_1, B_2 we have

$$(\delta_{a_1a_2} + \delta_{b_1b_2}) - (\delta_{a_1b_2} + \delta_{b_1a_2}) \quad < \quad 0.$$

by the consistency of δ with any quartet spanning $A|B$. Summing over all leaves in A_1, A_2, B_1, B_2 we see that $BME(\delta, T) - BME(\delta, T') < 0$. Thus any NNI which removes e leads to an increase in BME score and will not be accepted. This gives item 3.

Now we may easily deduce that an NNI based local topology search will never remove an edge e which has length at least twice the maximum error, giving item 1 of the theorem. Let T, T^* be binary phylogenetic trees with some common edge e. Let δ have maximum error $\epsilon < l_e/2$ relative to δ^*. Let $A|D$ be the split corresponding to edge e. For any quartet $ab|cd$ such that $a, b \in A$ and $c, d \in D$ we have:

$$\begin{aligned}\delta_{ab} + \delta_{cd} &\leq \delta^*_{ab} + \delta^*_{cd} + 2\epsilon \\ &\leq \delta^*_{ac} + \delta^*_{bd} - 2l_e + 2\epsilon \\ &\leq \delta_{ac} + \delta_{bd} - 2l_e + 4\epsilon \\ &< \delta_{ac} + \delta_{bd},\end{aligned}$$

and similarly $\delta_{ab} + \delta_{cd} < \delta_{ad} + \delta_{bc}$. Hence δ is consistent with the edge e and by item 3 of this theorem, already proven, an NNI based local topology search will never remove the edge e.

Finally we prove item 2: we show that as long as we check all trees within one SPR operation of the current tree, and choose the best of these, we will never destroy an edge e which has length at least 3 times the maximum error. Let T, T^* be binary phylogenetic trees with some common edge e. Let δ have maximum error $\epsilon < l_e/3$ relative to δ^*. Let the split corresponding to e be $A|B$. For a SPR operation to break the edge e it would need to choose a subtree $A_0 \subseteq A$ (without loss of generality), and regraft it within the clade B. Let the position within clade B that minimises BME score be as depicted in Fig. 1(b) taking $X_k = A_0$. We will show that regrafting A_0 to the edge e, as depicted in Fig. 1(a) (taking $A' = A - A_0$ and $X_k = A_0$) gives a smaller BME score. Essentially this is same situation as in the proof that GREEDYBME has safety radius $1/3$, except A_0 is now a clade rather than the single taxon x_k. However since Lemma 4.2 holds for clades, the proof goes through unchanged. Thus we conclude that regrafting A_0 to e gives a lower BME score than regrafting it inside B.

This completes the proof of Theorem 3.2 ■

V. CONCLUSION

In this work we have shown that the algorithms GREEDYBME and FASTME are more robust methods of inferring a phylogeny than the Neighbor-Joining algorithm in two rigorous senses. Firstly, GREEDYBME and FASTME have edge safety radius of 1/3 and, secondly, GREEDYBME will correctly reconstruct the true tree given a distance matrix that is quartet consistent with the true tree. Both conditions are strict improvements over the Neighbor-Joining algorithm. Experimental evidence has already demonstrated that FASTME performs well compared to other distance based phylogenetic reconstruction algorithms [5]. The results in this paper provide further theoretical justification for using this approach.

The significance of proving bounds on the *edge* safety radius is that *any sufficiently long edge* is correctly reconstructed from the distance matrix (edges longer than three times the maximum error), even in the presence of very short edges elsewhere in the tree. In contrast, results on safety radius can only guarantee that the whole tree is correctly reconstructed if *all* edges are sufficiently long, otherwise it cannot guarantee anything.

Minimum Evolution, and in particular Balanced Minimum Evolution, has been proposed by several authors as a guiding principle for inferring phylogenies (for references and discussion see [5]). Moreover the underlying reason for the accuracy of certain phylogenetic algorithms, including the Neighbor-Joining algorithm, has been attributed to their relationship to the Balanced Minimum Evolution principle [9], [12]. It is therefore counterintuitive that a heuristic for minimising BME score, GREEDYBME, has an edge safety radius of 1/3, when the underlying principle (*i.e.* any algorithm that selects the tree of globally minimum BME score) has a weaker edge safety radius, which even approaches zero for large trees [11]. Further work on understanding this issue, as well as extending the robustness guarantees to more reasonable models of error in distance matrices, will help improve distance based phylogenetic inference in the future.

REFERENCES

[1] K. Atteson, "The performance of Neighbor-Joining methods of phylogenetic reconstruction," *Algorithmica*, vol. 25, no. 2, pp. 251–278, 1999.

[2] M. Bordewich, O. Gascuel, K. T. Huber, and V. Moulton, "Consistency of topological moves based on the Balanced Minimum Evolution principle of phylogenetic inference," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 1, pp. 110–117, 2009.

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

[4] D. Catanzaro, M. Labbé, R. Pesenti, and J. J. Salazar-González, "The Balanced Minimum Evolution Problem," *INFORMS Journal on Computing*, vol. 24, no. 2, pp. 276–294, 2012.

[5] R. Desper and O. Gascuel, "Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle," *Journal of Computational Biology*, vol. 19, no. 5, pp. 687–705, 2002.

[6] ——, "Theoretical foundation of the Balanced Minimum Evolution method of phylogenetic inference," *Molecular Biology and Evolution*, vol. 21, no. 3, pp. 587–598, 2004.

[7] J. Felsenstein, "An alternating least-squares approach to inferring phylogenies from pairwise distances," *Systematic Biology*, vol. 46, pp. 101–111, 1997.

[8] O. Gascuel, "BioNJ: An improved version of the NJ algorithm based on a simple model of sequence data," *Molecular Biology and Evolution*, vol. 14, pp. 685–695, 1997.

[9] O. Gascuel and M. A. Steel, "Neighbor-Joining revealed," *Molecular Biology and Evolution*, vol. 23, no. 11, pp. 1997–2000, 2006.

[10] W. Hordijk and O. Gascuel, "Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood," *Bioinformatics*, vol. 21, no. 24, pp. 4338–4347, 2005.

[11] R. Mihaescu, "Reliability results for the general balanced minimum evolution principle," 2012, private communication. Article in preparation.

[12] R. Mihaescu, D. Levy, and L. Pachter, "Why Neighbor-Joining works," *Algorithmica*, vol. 54, no. 1, pp. 1–24, 05 2009.

[13] F. Pardi and O. Gascuel, "Combinatorics of distance-based tree inference," *Proceedings of the National Academy of Sciences*, vol. 109, no. 41, pp. 16443–16448, 2012.

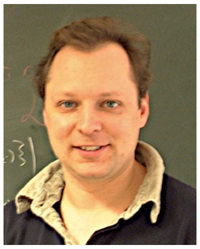
[14] F. Pardi, S. Guillemot, and O. Gascuel, "Robustness of phylogenetic inference based on Minimum Evolution," *Bulletin of Mathematical Biology*, vol. 72, no. 7, pp. 1820–1839, 2010.

[15] Y. Pauplin, "Direct calculation of tree length using a distance matrix," *Journal of Molecular Evolution*, vol. 51, pp. 66–85, 2000.

[16] N. Saitou and M. Nei, "The Neighbor-Joining method: A new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, pp. 406–424, 1987.

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

[18] T. Shigezumi, "Robustness of greedy type Minimum Evolution algorithms," in *International Conference on Computational Science (2)*, ser. Lecture Notes in Computer Science, V. N. Alexandrov, G. D. van Albada, P. M. A. Sloot, and J. Dongarra, Eds., vol. 3992. Berlin: Springer, 2006, pp. 815–821.



Magnus Bordewich received the MMath degree and the DPhil degree in mathematics from Oxford University in 1998 and 2003, respectively. He was a postdoctoral research fellow in the Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, and, then, in the School of Computer Science, Leeds University, Leeds, United Kingdom. In 2006, he joined Durham University, where he is currently a senior lecturer in the School of Engineering and Computing Sciences. From 2006 to 2009 he held an EPSRC postdoctoral fellowship in theoretical computer science, researching randomized algorithms and approximation in phylogenetics.

Radu Mihaescu received his Ph.D. degree in mathematics from The University of California at Berkeley in 2008. His thesis was titled "Distance Methods for Phylogeny Reconstruction" and was winner of the Bernard Friedman Memorial Prize for an outstanding thesis in applied mathematics. He is currently an Assistant vice-president at Knight Capital Group.