JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 12, 2015
Mary Ann Liebert, Inc.
Pp. 1129–1142
DOI: 10.1089/cmb.2015.0146

Consequences of Common Topological Rearrangements for Partition Trees in Phylogenomic Inference

OLGA CHERNOMOR,[1,2] BUI QUANG MINH,[1] and ARNDT VON HAESELER[1,2]

ABSTRACT

In phylogenomic analysis the collection of trees with identical score (maximum likelihood or parsimony score) may hamper tree search algorithms. Such collections are coined phylogenetic terraces. For sparse supermatrices with a lot of missing data, the number of terraces and the number of trees on the terraces can be very large. If terraces are not taken into account, a lot of computation time might be unnecessarily spent to evaluate many trees that in fact have identical score. To save computation time during the tree search, it is worthwhile to quickly identify such cases. The score of a species tree is the sum of scores for all the so-called induced partition trees. Therefore, if the topological rearrangement applied to a species tree does not change the induced partition trees, the score of these partition trees is unchanged. Here, we provide the conditions under which the three most widely used topological rearrangements (nearest neighbor interchange, subtree pruning and regrafting, and tree bisection and reconnection) change the topologies of induced partition trees. During the tree search, these conditions allow us to quickly identify whether we can save computation time on the evaluation of newly encountered trees. We also introduce the concept of partial terraces and demonstrate that they occur more frequently than the original "full" terrace. Hence, partial terrace is the more important factor of timesaving compared to full terrace. Therefore, taking into account the above conditions and the partial terrace concept will help to speed up the tree search in phylogenomic inference.

Key words: nearest neighbor interchange, partial terraces, phylogenetic terraces, subtree pruning and regrafting, tree bisection and reconnection.

1. INTRODUCTION

In phylogenomics, one aims to reconstruct a phylogenetic *species* tree from multiple genes. One popular approach is to infer the trees from the concatenated gene alignment, the so-called supermatrix (Sanderson et al., 1998; De Queiroz and Gatesy, 2007). Here, if a gene sequence is not available for some taxon, it is represented by the sequence of unknown characters and is referred to as missing data. Several

[1]Max F. Perutz Laboratories, Center for Integrative Bioinformatics Vienna, University of Vienna, Vienna, Austria.
[2]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria.

studies (van der Linde et al., 2010; Pyron and Wiens, 2011; Pyron et al., 2011; Nyakatura and Bininda-Emonds, 2012; Springer et al., 2012; Hedtke et al., 2013) use quite sparse supermatrices in their analysis and the percentage of missing data sometimes constitutes up to 95% (Peters et al., 2011).

Recently, it has been shown that missing data can hamper the tree search via existence of phylogenetic terraces (Sanderson et al., 2011), a collection of trees with exactly the same likelihood or parsimony score. Terraces occur in the analysis with *partitioned data*, that is, when distinct blocks of a supermatrix are treated differently (e.g., when each gene corresponding to one block evolves under its own evolutionary model). Two trees are said to belong to one terrace if the collections of their *induced partition trees* are exactly the same. Here, the induced partition tree is obtained by pruning the taxa on species tree, which have no sequence for the corresponding partition block.

Since the number of trees on one terrace can be quite large (Sanderson et al., 2011), accounting for terraces in tree search algorithms can potentially save a lot of computation time. During the tree search, one explores the tree space by moving from one candidate tree to another by means of topological rearrangements. If the topological rearrangement does not change any of the induced partition trees, then the two trees belong to the same terrace and a recomputation of objective function (maximum likelihood or maximum parsimony) used in the tree search is not necessary in order to evaluate a new tree.

Here, we first specify the conditions under which the topological rearrangements applied to the species tree change the corresponding induced partition trees. Using these conditions, one can quickly identify whether it is necessary to recompute the objective function for a given partition or not as a consequence of one of the three widely used rearrangements: nearest neighbor interchange (NNI), subtree pruning and regrafting (SPR) and tree bisection and reconnection (TBR) (Felsenstein, 2004).

We further generalize the concept of terrace to *partial terrace*, which is even more useful in practical phylogenetic analysis. We analyze several published alignments by examining NNI neighborhoods of random trees and trees encountered during the tree search using IQ-TREE (Nguyen et al., 2015). We show that for large number of taxa partial terraces are mainly determined by the missing data and less dependent on the actual tree topology analyzed. By taking into account partial terraces, it will be possible to speed up the tree search algorithms even in the absence of terraces.

The outline of the article is the following. We first introduce the notations and then discuss the important features of NNI, SPR, and TBR. Next, we specify the conditions when these topological rearrangements do not change the topology of induced partition trees. We further elucidate why such conditions are helpful even in the absence of terraces and define the concept of partial terrace. We analyze several published alignments to point out that partial terraces do occur in practice. Finally, we discuss the additional practical advantages of using induced partition trees in the maximum likelihood framework.

2. BACKGROUND

2.1. Basic definitions and notations

In this section we provide basic definitions and notations used throughout the article. For a complete overview, see chapters 2, 3, and 6 in Semple and Steel (2003).

Definition 2.1. Let X be a taxon set. A *phylogenetic tree* T of X is a leaf-labeled tree with a bijection map from X into the set of leaves of T.

In the following, we work only with bifurcating phylogenetic trees; that is, all internal nodes have exactly three adjacent edges.

Definition 2.2. A *split*, denoted by $A|B$, is a bipartition of X into two nonempty, nonoverlapping sets A and B, where $A \cup B = X$.

Note that $A|B$ and $B|A$ are equivalent. Every edge of T is associated with a split. When cutting an edge e of T, we obtain two subtrees with leaf labels X_1 and X_2, and then a split corresponding to e is defined as $X_1|\ X_2$. We denote this with $e = X_1|\ X_2$

We denote by $\Sigma(T)$ a collection of all splits corresponding to edges of T.

The *symmetric difference* of two sets A and B, denoted $A\Delta B$, is given by $(A \setminus B) \cup (B \setminus A)$, or the union of taxa present in A but not B, and vice versa.

Definition 2.3. Let T_1 and T_2 be the two leaf-labeled trees with the same label set X, and $\Sigma(T_1)$ and $\Sigma(T_2)$ be the collections of splits of T_1 and, T_2, respectively. Then the Robinson–Foulds (RF) distance (Robinson and Foulds, 1981) between T_1 and T_2 is equal to $|\Sigma(T_1)\Delta\Sigma(T_2)|$.

If for two trees the RF distance between them is 0, then they have the same collection of splits, and from splits-equivalence theorem (Semple and Steel, 2003; p. 43), the trees are *equivalent*.

Definition 2.4. Let Y be a subset of X. An *induced subtree* of T, denoted by $T|Y$, is a leaf-labeled tree with the following collection of splits:

$$\Sigma(T|Y) = \{A \cap Y | B \cap Y : A|B \in \Sigma(T) \text{ and } A \cap Y \neq \emptyset, B \cap Y \neq \emptyset\}.$$

For a species tree T and a given partition with taxon set Y, a *partition tree* is an induced subtree $T|Y$.

2.2. Topological rearrangement operations

In this section we introduce the topological rearrangements on trees commonly used in phylogenetic inference.

The simplest possible operation that changes only one split on a tree is an NNI. It can only be applied to interior edges of the tree, since it requires the so-called quartet structure with an interior edge being the central edge of this structure (Fig. 1).

Let e be an interior edge of T and e_1, e_2, e_3, e_4 its four incident edges with A, B, C, D being the taxon sets leading from them, respectively (Fig. 1). An NNI on T around e is obtained by exchanging the subtrees below two nonincident edges from e_1, e_2, e_3, e_4. We denote a new tree by T_{NNI}.

For each interior edge e there are two possible NNIs obtained by exchanging a subtree below e_1 with a subtree below either e_3 or e_4 (note that this is equivalent to swapping the subtree below e_2 with either e_4 or e_3, respectively).

Let us assume that the NNI is applied to edge e by swapping e_1 and e_3. The splits corresponding to e_1, e_2, e_3, and e_4 stay unchanged:

$$e_1 = A|B \cup C \cup D,$$
$$e_2 = B|A \cup C \cup D,$$
$$e_3 = C|A \cup B \cup D,$$
$$e_4 = D|A \cup B \cup C.$$

This also holds true for the edges belonging to subtrees below e_1, e_2, e_3, and e_4 (Fig. 1). Here, if $e_1 = A|B \cup C \cup D$, *the subtree below* e_1 is a subtree with a leaf set A and not the union of sets. Hence, the splits corresponding to e_1, e_2, e_3, e_4 and edges below them will be shared by T and T_{NNI}.

The central edge e in terms of splits will be changed by the NNI from $A \cup B|C \cup D$ to $e^{NNI} = A \cup D|B \cup C$.

It follows from above that T and T_{NNI} are different only in one split; that is,

$$\Sigma(T) \,\Delta\, \Sigma(T_{NNI}) = \{A \cup B|C \cup D, A \cup D|B \cup C\}$$

and the RF distance between T and T_{NNI} is 2.

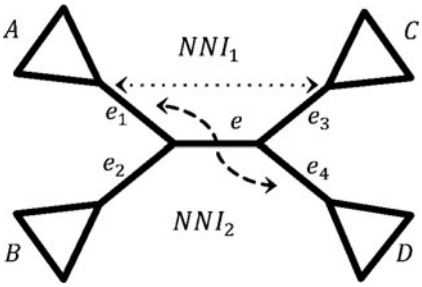

FIG. 1. Visualization of NNI. Species tree T and the two NNIs around central edge e. NNI_1 is obtained by exchanging subtrees below edges e_1 and e_3, while NNI_2 by exchanging subtrees e_1 and e_4. NNI, nearest neighbor interchange.

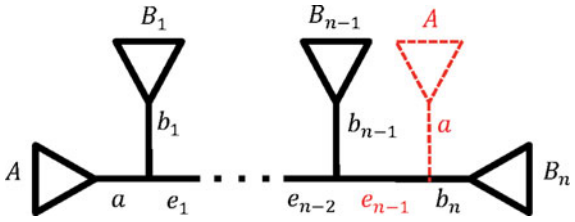

FIG. 2. Visualization of SPR. A new tree T_{SPR} is obtained by pruning the subtree A below edge a and regrafting it onto edge b_n (dashed red subtree). After SPR is applied, edges b_1 and e_1 are joined and edge b_n is split into e_{n-1} and b_n. SPR, subtree pruning and regrafting.

We now discuss SPR, a more general topological rearrangement that changes one or more splits of the tree.

An SPR on T is represented in Figure 2 (see also Hordijk and Gascuel, 2005). A new tree T_{SPR} is obtained from T by pruning the subtree below edge a and regrafting it onto edge b_n (we sometimes refer to such SPR as n-SPR). Note, that n is at least 3 and if $n=3$, an SPR is equivalent to an NNI obtained by swapping subtrees belonging to edges a and b_2. Let A, B_1, …, B_n denote the corresponding taxon sets leading from a, b_1, …, b_n, respectively (Fig. 2).

An SPR on T changes only the splits of the path edges, namely: for $\forall x \in \{1, \ldots, n-2\}$

$$e_x = A \cup B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n$$

is changed to

$$e_x^{SPR} = B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n \cup A,$$

where e_x^{SPR} is an edge that corresponds to e_x on a new tree T_{SPR}. Also, a new edge appears: $e_{n-1} = B_1 \cup \ldots \cup B_{n-1} | A \cup B_n$. The rest of splits remain unchanged and are shared by both trees. Hence, for T and T_{SPR} the symmetric difference $\Sigma(T) \Delta \Sigma(T_{SPR})$ consists of the following splits:

$$A \cup B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n, \quad \forall x \in \{1, \ldots, n-2\},$$
$$B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n \cup A, \quad \forall x \in \{2, \ldots, n-1\}.$$

The RF distance between T and T_{SPR} is equal to 2 $(n-2)$.

The last topological rearrangement we are going to discuss is the TBR. A TBR on T is shown in Figure 3, where a new tree T_{TBR} is obtained from T (Fig. 3, in black) by cutting edge e and reconnecting edges b_n and c_m with a new edge e^{TBR} (Fig. 3, red dashed line). Note that n or m must be greater than 2. W.l.o.g. assume that $m \leq n$. If $n=3$ and $m=2$, then a TBR corresponds to an NNI around edge e_1 by swapping subtrees below e and b_2. If $n>3$ and $m=2$, then a TBR corresponds to an SPR.

TBR only changes the splits corresponding to all path edges (e_i and z_j), but e. Namely,

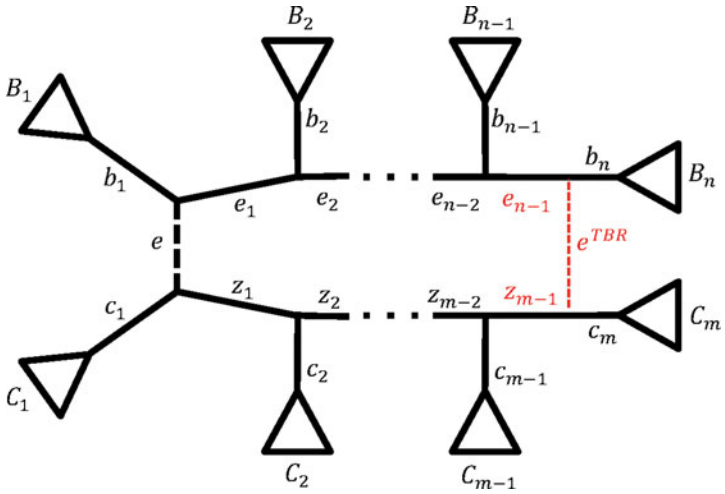

FIG. 3. Visualization of TBR. To obtain T_{TBR}, species tree T is cut into two parts (by removing edge e), which are further reconnected by joining edges b_n and c_m with e^{TBR}. Edge b_n is split into e_{n-1} and b_n, while c_m is split into z_{m-1} and c_m. Edges b_1 and e_1 are joined, as well as c_1 and z_1. TBR, tree bisection and reconnection.

$$e = B_1 \cup \ldots \cup B_n | C_1 \cup \ldots \cup C_m = e^{TBR},$$

while for $\forall x \in \{1, \ldots, n-2\}$

$$e_x = C_1 \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n$$

is changed to

$$e_x^{TBR} = B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_m$$

and for $\forall y \in \{1, \ldots, m-2\}$

$$z_y = B_1 \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_y | C_{y+1} \cup \ldots \cup C_m$$

is changed to

$$z_y^{TBR} = C_1 \cup \ldots \cup C_y | C_{y+1} \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_n.$$

Also two new edges appear

$$e_{n-1} = B_1 \cup \ldots \cup B_{n-1} | B_n \cup C_1 \cup \ldots \cup C_m,$$
$$z_{m-1} = C_1 \cup \ldots \cup C_{m-1} | C_m \cup B_1 \cup \ldots \cup B_n.$$

The remaining splits stay unchanged. Hence, for T and T_{TBR} the symmetric difference $\Sigma(T) \Delta \Sigma(T_{TBR})$ is a set consisting of the following splits

$$\begin{aligned}
&C_1 \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n, \quad \forall x \in \{1, \ldots, n-2\},\\
&B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_m, \quad \forall x \in \{2, \ldots, n-1\},\\
&B_1 \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_y | C_{y+1} \cup \ldots \cup C_m, \quad \forall y \in \{1, \ldots, m-2\},\\
&C_1 \cup \ldots \cup C_y | C_{y+1} \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_n, \quad \forall y \in \{2, \ldots, m-1\}.
\end{aligned}$$

Therefore, the RF distance between T and T_{TBR} is 2 $(n + m - 4)$.

3. CONSEQUENCES OF TOPOLOGICAL REARRANGEMENTS APPLIED TO A SPECIES TREE

In the following we discuss how the topological rearrangement of the species tree T influences the topology of the partition trees and start with the simplest operation, an NNI.

Proposition 1. *Let e be an interior edge and e_1, e_2, e_3, e_4 the four edges adjacent to e with A, B, C, D being the taxon sets leading from the corresponding edges (Fig. 1). Let a new tree T_{NNI} be obtained from T via NNI. For a partition with a taxon set Y, the topologies of $T|Y$ and $T_{NNI}|Y$ are different iff Y has at least one representative taxon in each subset A, B, C, D.*

Proof.

W.l.o.g. assume that T_{NNI} is obtained from T via swapping of subtrees below e_1 and e_3. Then $\Sigma(T) \Delta \Sigma(T_{NNI}) = \{A \cup B | C \cup D,\ A \cup D | B \cup C\}$ and as a consequence for corresponding partition trees we have

$$\Sigma(T|Y) \Delta \Sigma(T_{NNI}|Y) = \{(A \cup B) \cap Y | (C \cup D) \cap Y, (A \cup D) \cap Y | (B \cup C) \cap Y\}.$$

It is easy to show that if at least one set from $A \cap Y$, $B \cap Y$, $C \cap Y$, $D \cap Y$ were empty, then both splits $(A \cup B) \cap Y | (C \cup D) \cap Y$ and $(A \cup D) \cap Y | (B \cup C) \cap Y$ coincide with splits shared by $T|Y$ and $T_{NNI}|Y$ (e.g., see Fig. 4). Hence, $\Sigma(T|Y) \Delta \Sigma(T_{NNI}|Y) = \emptyset$ and the RF distance between these trees would be 0. Therefore, for $T|Y$ and $T_{NNI}|Y$ to have different topologies, all $A \cap Y$, $B \cap Y$, $C \cap Y$, $D \cap Y$ must be nonempty, meaning that Y has to have at least one representative in each subset A, B, C, D. ■

In simple words, if some intersections of A, B, C, D with Y are empty, then a partition tree does not have a corresponding quartet structure for the NNI to be applied to and edge e loses its centrality or interior

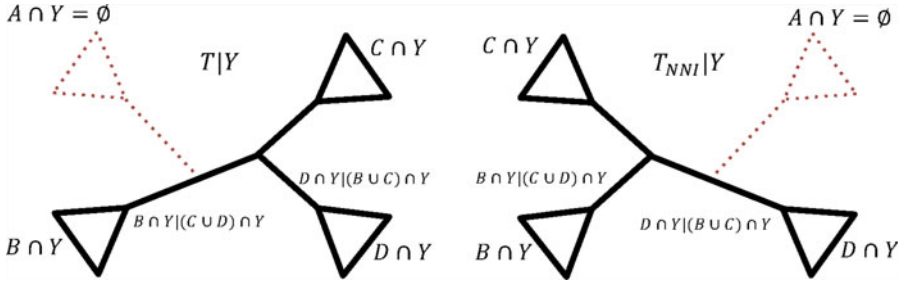

FIG. 4. An example when an NNI on T does not change the topology of $T|Y$. Solid lines correspond to two induced partition trees before ($T|Y$) and after ($T_{NNI}|Y$) the NNI was applied to edge e on T by swapping the subtrees below e_1 and e_3 (Fig. 1). In this case, Y does not have a representative in A (i.e., $A \cap Y = \emptyset$); therefore, $(A \cup B) \cap Y | (C \cup D) \cap Y = B \cap Y | (C \cup D) \cap Y$ and $(A \cup D) \cap Y | (B \cup C) \cap Y = D \cap Y | (B \cup C) \cap Y$. Since the splits $B \cap Y | (C \cup D) \cap Y$ and $D \cap Y | (B \cup C) \cap Y$ are shared by $T|Y$ and $T_{NNI}|Y$, then $\Sigma(T|Y) \Delta \Sigma(T_{NNI}|Y) = \emptyset$ and RF distance between $T|Y$ and $T_{NNI}|Y$ is 0.

feature (see, e.g., Fig. 4). When this happens, the topology of the partition tree $T|Y$ is not affected by the NNI applied to e on the species tree T.

We next specify the condition when an SPR changes the topology of partition tree.

Proposition 2. *Let tree T be in the form shown in Figure 2, and a new tree T_{SPR} is obtained with SPR by pruning subtree below edge a and regrafting it onto b_n.*

Then for a partition with a taxon set Y the following is true:

(i) the topologies of $T|Y$ and $T_{SPR}|Y$ are different, if Y has at least one representative in A and in at least another three subsets from $B_1, B_2, \ldots, B_n$;

(ii) this SPR will correspond to an SPR on $T|Y$ obtained by pruning the subtree below edge with a split $A \cap Y | (B_1 \cup \ldots \cup B_n) \cap Y$ and regrafting it onto edge with split $B_k \cap Y | \left(\cup_{i \in \{1, \ldots, n\} \setminus k} B_i \cup A\right) \cap Y$, where $k = \max_{1 \le i \le n} \{i | B_i \cap Y \neq \emptyset\}$.

Proof.

(i) The symmetric difference $\Sigma(T) \Delta \Sigma(T_{SPR})$ consists of the following splits

$$A \cup B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n, \quad \forall x \in \{1, \ldots, n-2\},$$
$$B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n \cup A, \quad \forall x \in \{2, \ldots, n-1\}.$$

As a consequence for the induced partition trees $T|Y$ and $T_{SPR}|Y$, the symmetric difference of $\Sigma(T|Y)$ and $\Sigma(T_{SPR}|Y)$ consists of

$$(A \cup B_1 \cup \ldots \cup B_x) \cap Y | (B_{x+1} \cup \ldots \cup B_n) \cap Y, \quad \forall x \in \{1, \ldots, n-2\},$$
$$(B_1 \cup \ldots \cup B_x) \cap Y | (B_{x+1} \cup \ldots \cup B_n \cup A) \cap Y, \quad \forall x \in \{2, \ldots, n-1\}.$$

It is easy to see that if $A \cap Y = \emptyset$, then all these splits would be shared by both partition trees, that is, $\Sigma(T|Y) \Delta \Sigma(T_{SPR}|Y) = \emptyset$ and the RF distance between $T|Y$ and $T_{SPR}|Y$ would be 0. Therefore, Y must have at least one representative in A.

For $T|Y$ and $T_{SPR}|Y$ to have different topologies, an SPR on T should correspond to at least an NNI on $T|Y$. Hence, $T|Y$ must have a corresponding quartet structure and together with A at least another three subsets from $B_1, B_2, \ldots, B_n$ should have at least one representative in Y. W.l.o.g. assume that together with A also B_m, B_h, B_k $(1 \le m < h < k \le n)$ have at least one representative in Y while $B_j \cap Y = \emptyset \ \forall j \in \{1, \ldots, n\} \setminus \{m, h, k\}$ (see, e.g., Fig. 5). Then

$$\Sigma(T|Y) \Delta \Sigma(T_{SPR}|Y) = \{(A \cup B_m) \cap Y | (B_h \cup B_k) \cap Y, \ (B_m \cup B_h) \cap Y | (B_k \cup A) \cap Y\}.$$

Thus, the RF distance between $T|Y$ and $T_{SPR}|Y$ is 2.

(ii) Let $I = \{i_1, .., i_k\}$ be the set of all indices, such that $\forall i \in I: \ B_i \cap Y \neq \emptyset$ and let $1 \le i_1 < \cdots < i_k \le n$. For edge $a = A | B_1 \cup \ldots \cup B_n$ its corresponding split on the partition tree $T|Y$ is equal to

$$A \cap Y | (B_1 \cup \ldots \cup B_n) \cap Y = A \cap Y | \cup_{i \in I} (B_i \cap Y).$$

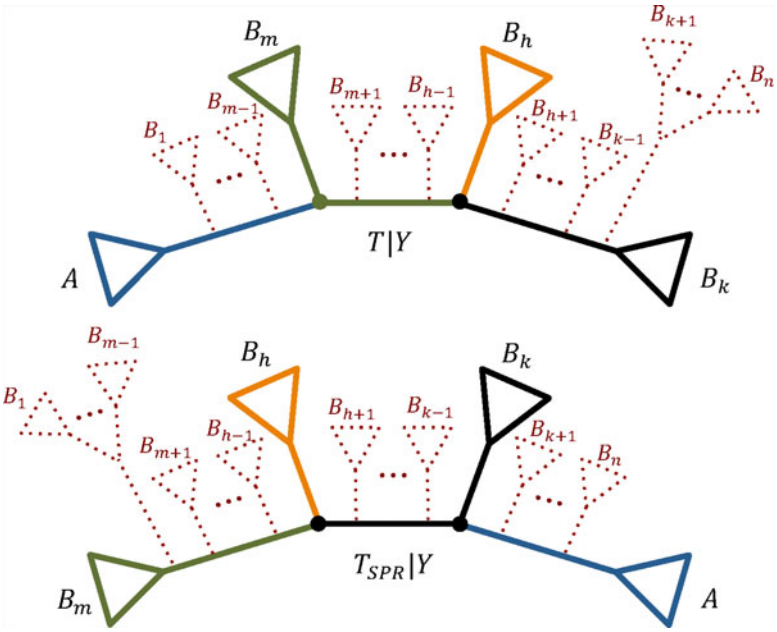

FIG. 5. An example when n-SPR on T is a 3-SPR (or NNI) on $T|Y$. There are two induced partition trees (solid lines): before ($T|Y$) and after ($T_{SPR}|Y$) an SPR was applied on T by pruning the subtree below edge a and regrafting it onto b_n (Fig. 2). The three dots denote all the subtrees between the corresponding pair of subtrees on the species trees T and T_{SPR}. Here, only A, B_m, B_h, and B_k have at least one representative in Y and $\forall j \in \{1, \ldots, n\} \setminus \{m, h, k\}$: B_j have no taxa in common with Y.

Similarly for e_{i_k-1} its corresponding split on $T|Y$

$$(A \cup B_1 \cup \ldots \cup B_{i_k-1}) \cap Y | (B_{i_k} \cup \ldots \cup B_n) \cap Y = (\cup_{i \in I \setminus i_k} B_i \cup A) \cap Y | B_{i_k} \cap Y,$$

and for $e^{SPR}_{i_k-1}$ its corresponding split on the partition tree $T_{SPR}|Y$

$$(B_1 \cup \ldots \cup B_{i_k-1}) \cap Y | (B_{i_k} \cup \ldots \cup B_n \cup A) \cap Y = (\cup_{i \in I \setminus i_k} B_i) \cap Y | (B_{i_k} \cup A) \cap Y.$$

The above means that an edge on $T|Y$ with split $(\cup_{i \in I \setminus i_k} B_i \cup A) \cap Y | B_{i_k} \cap Y$ was divided by an edge with split $A \cap Y | \cup_{i \in I} (B_i \cap Y)$ in two edges (see also Fig. 5, where $I = \{i_1, i_2, i_3\}$). Therefore, regrafting onto edge b_n on T corresponds to regrafting onto edge with a split $B_{i_k} \cap Y | \left(\cup_{i \in \{1, \ldots, n\} \setminus i_k} B_i \cup A\right) \cap Y$ on partition tree $T|Y$. And since $1 \leq i_1 < \cdots < i_k \leq n$, then $i_k = \max_{1 \leq i \leq n} \{i | B_i \cap Y \neq \emptyset\}$. ■

In other words, Proposition 2 states that an SPR on T changes the topology of $T|Y$ if the structure of T from Figure 2 corresponds to at least a quartet structure on $T|Y$ (e.g., Fig. 5). In this case, n-SPR on T is a 3-SPR (or NNI) on $T|Y$.

We now discuss TBR and the topological change of a partition tree as a consequence of TBR on species tree.

Proposition 3. *Let tree T be in the form shown in Figure 3 and a new tree T_{TBR} is obtained by cutting edge e and reconnecting b_n and c_m with a new edge.*

Then for a partition with a taxon set Y the following is true:

(i) the topologies of $T|Y$ and $T_{TBR}|Y$ are different if either of the following conditions is satisfied:

- Y has at least one representative in at least one subset from $B_1, B_2, \ldots, B_n$ and in at least another three subsets from $C_1, C_2, \ldots, C_m$
- Y has at least one representative in at least one subset from $C_1, C_2, \ldots, C_m$ and in at least another three subsets from $B_1, B_2, \ldots, B_n$

(ii) this TBR will correspond to a TBR on $T|Y$ obtained by cutting the edge with split $(B_1 \cup \ldots \cup B_n) \cap Y | (C_1 \cup \ldots \cup C_m) \cap Y$ and reconnecting edges with splits $B_k \cap Y | (\cup_{i \in \{1, \ldots, n\} \setminus k} B_i \cup C_1 \cup \ldots \cup C_m) \cap Y$ and $C_h \cap Y | (\cup_{j \in \{1, \ldots, m\} \setminus h} C_j \cup B_1 \cup \ldots \cup B_n) \cap Y$, where $k = \max_{1 \leq i \leq n} \{i \mid B_i \cap Y \neq \emptyset\}$ and $h = \max_{1 \leq j \leq m} \{j \mid C_j \cap Y \neq \emptyset\}$.

Proof.

(i) The symmetric difference $\Sigma(T) \Delta \Sigma(T_{TBR})$ consists of the following splits:

$$
\begin{aligned}
C_1 \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n, &\quad \forall x \in \{1, \ldots, n-2\},\\
B_1 \cup \ldots \cup B_x | B_{x+1} \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_m, &\quad \forall x \in \{2, \ldots, n-1\},\\
B_1 \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_y | C_{y+1} \cup \ldots \cup C_m, &\quad \forall y \in \{1, \ldots, m-2\},\\
C_1 \cup \ldots \cup C_y | C_{y+1} \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_n, &\quad \forall y \in \{2, \ldots, m-1\},
\end{aligned}
$$

As a consequence, the symmetric difference $\Sigma(T|Y)\Delta\Sigma(T_{TBR}|Y)$ consists of

$$
\begin{aligned}
(C_1 \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_x) \cap Y | (B_{x+1} \cup \ldots \cup B_n) \cap Y, &\quad \forall x \in \{1, \ldots, n-2\},\\
(B_1 \cup \ldots \cup B_x) \cap Y | (B_{x+1} \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_m) \cap Y, &\quad \forall x \in \{2, \ldots, n-1\},\\
(B_1 \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_y) \cap Y | (C_{y+1} \cup \ldots \cup C_m) \cap Y, &\quad \forall y \in \{1, \ldots, m-2\},\\
(C_1 \cup \ldots \cup C_y) \cap Y | (C_{y+1} \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_n) \cap Y, &\quad \forall y \in \{2, \ldots, m-1\},
\end{aligned}
$$

It is easy to see that if $\forall i \in \{1,..,n\}$: $B_i \cap Y = \emptyset$, then all these splits would be shared by both partition trees; that is $\Sigma(T|Y) \Delta \Sigma(T_{TBR}|Y) = \emptyset$ and the RF distance between $T|Y$ and $T_{TBR}|Y$ would be 0. Therefore, Y must have at least one representative in at least one from B_1, B_2, …, B_n. Similarly, Y must have at least one representative in at least one from C_1, C_2, … , C_m.

W.l.o.g. assume that $B_k \cap Y \neq \emptyset$ and $C_h \cap Y \neq \emptyset$, where $1 \leq k \leq n$ and $1 \leq h \leq m$.

Partition trees $T|Y$ and $T_{TBR}|Y$ will have different topologies if a TBR on T corresponds to at least an NNI on $T|Y$. Hence, the partition tree $T|Y$ must have a corresponding quartet structure and together with B_k and C_h at least other two subsets from the remaining B_i and C_j should have at least one representative in Y.

W.l.o.g. assume that together with B_k and C_h also C_p, C_q $(1 \leq p < q < h \leq m)$ have at least one representative in Y (Fig. 6, right panel). Then it is easy to show that

$$\Sigma(T|Y)\Delta\Sigma(T_{TBR}|Y) = \{(B_k \cup C_p) \cap Y | (C_q \cup C_h) \cap Y, (C_p \cup C_q) \cap Y | (C_h \cup B_k) \cap Y\}$$

and RF distance between $T|Y$ and $T_{TBR}|Y$ is 2.

Similarly, one can show that if together with B_k and C_h also B_p, B_q $(1 \leq p < q < k \leq n)$ have at least one representative in Y, then RF distance between $T|Y$ and $T_{TBR}|Y$ is also 2.

In contrast, if Y has at least one representative in B_k, C_h and also in B_p, C_q $(1 \leq p < k \leq n$ and $1 \leq q < h \leq m)$, then $\Sigma(T|Y)\Delta\Sigma(T_{TBR}|Y) = \emptyset$ and RF distance is 0 (Fig. 6, left panel).

(ii) Let $I = \{i_1,..,i_k\}$ be the set of all indices such that $\forall i \in I : B_i \cap Y \neq \emptyset$ and let $1 \leq i_1 < \ldots < i_k \leq n$. Similarly, let $J = \{j_1,..,j_h\}$ be the set of all indices such that $\forall j \in J : C_j \cap Y \neq \emptyset$ and let $1 \leq j_1 < \ldots < j_h \leq m$. Then for edge

$$e = B_1 \cup \ldots \cup B_n | C_1 \cup \ldots \cup C_m$$

the corresponding split on $T|Y$ is

$$\cup_{i \in I} B_i \cap Y | \cup_{j \in J} C_j \cap Y.$$

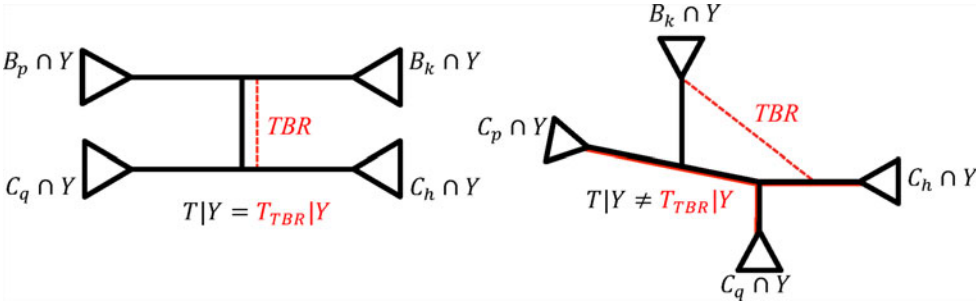

FIG. 6. Examples of corresponding TBRs on partition trees. Two partition trees with topologies before ($T|Y$, in black) and after ($T_{TBR}|Y$, in red) the TBR were applied to the species tree. For simplicity we do not show the pruned subtrees for which $B_i \cap Y = \emptyset$ and $C_j \cap Y = \emptyset$. On the left is an example case when the topology of partition tree remains unchanged after TBR. On the right is the simplest case when the TBR changes the topology of partition tree. In this case a TBR on species tree corresponds to an NNI on partition tree.

For edge

$$e_{i_k-1} = C_1 \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_{i_k-1} | B_{i_k} \cup \ldots \cup B_n$$

its corresponding split on tree $T|Y$ is

$$\begin{aligned}(C_1 \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_{i_k-1}) \cap Y | (B_{i_k} \cup \ldots \cup B_n) \cap Y = \\ = \left(\cup_{j\in J} C_j \cap Y\right) \cup (\cup_{i \in I \setminus i_k} B_i \cap Y) | B_{i_k} \cap Y.\end{aligned}$$

Similarly, for the corresponding edge on T_{TBR} $e^{TBR}_{i_k-1} = B_1 \cup \ldots \cup B_{i_k-1} | B_{i_k} \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_m$ its split on $T_{TBR}|Y$ is

$$\begin{aligned}(B_1 \cup \ldots \cup B_{i_k-1}) \cap Y | (B_{i_k} \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_m) \cap Y = \\ = (\cup_{i \in I \setminus i_k} B_i \cap Y) | (B_{i_k \cap Y}) \cup \left(\cup_{j\in J} C_j \cap Y\right).\end{aligned}$$

For edges

$$z_{j_h-1} = B_1 \cup \ldots \cup B_n \cup C_1 \cup \ldots \cup C_{j_h-1} | C_{j_h} \cup \ldots \cup C_m$$

and

$$z^{TBR}_{j_h-1} = C_1 \cup \ldots \cup C_{j_h-1} | C_{j_h} \cup \ldots \cup C_m \cup B_1 \cup \ldots \cup B_n$$

their corresponding splits on $T|Y$ and $T_{TBR}|Y$ are

$$(\cup_{i\in I} B_i \cap Y) \cup \left(\cup_{j \in J \setminus j_h} C_j \cap Y\right) | C_{j_h} \cap Y$$

and

$$\left(\cup_{j \in J \setminus j_h} C_j \cap Y\right) | \left(C_{j_h} \cap Y\right) \cap (\cup_{i\in I} B_i \cap Y)$$

respectively. The above means that edges on $T|Y$ with corresponding splits

$$\left(\cup_{j\in J}\ C_j \cap Y\right) \cup (\cup_{i \in I \setminus i_k} B_i \cap Y) | B_{i_k} \cap Y$$

and

$$(\cup_{i\in I}\ B_i \cap Y) \cup \left(\cup_{j \in J \setminus j_h} C_j \cap Y\right) | C_{j_h} \cap Y$$

were reconnected on $T_{TBR}|Y$ by $\cup_{i\in I}\ B_i \cap Y | \cup_{j\in J}\ C_j \cap Y$. Since $1 \le i_1 < \ldots < i_k \le n$ and $1 \le j_1 < \ldots < j_h \le m$, then $i_k = \max_{1\le i\le n}\{i \mid B_i \cap Y \neq \emptyset\}$ and $j_h = \max_{1\le j\le m}\{j \mid C_j \cap Y \neq \emptyset\}$. ■

4. PARTIAL TERRACES

4.1. Definition of partial terraces

In this section we discuss *partial terraces* that generalize the terrace concept (Sanderson et al., 2011), which we call *full terrace* for clarity. When comparing the two trees in a partitioned framework, we compare the sets of their induced partition trees. If the sets are identical, then the two trees belong to one full terrace. Sanderson et al. (2011) showed that the number of trees on one full terrace can be quite large. Large full terraces pose a problem in phylogenetic inference, since they may abort tree search prematurely or even if an optimal tree has been found, this tree is by no means unique. To reduce this problem, it is possible to reduce the terrace size by, for example, choosing a different partition scheme (Sanderson et al., 2015) or by excluding some taxa from the analysis.

Now, if two species trees T_1 and T_2 share only a subset of identical induced partition trees, then we say that they belong to the same *partial terrace*. The log-likelihoods and parsimony scores of identical partition trees $T_1|Y_i$ and $T_2|Y_i$ are the same. Obviously, partial terraces occur more frequently than full terraces (see below). Large partial terraces can be still problematic for tree search algorithms. On the other hand, partial terraces provide the potential to reduce computation time.

Table 1. Alignments Used to Study the Occurrence of Partial Terraces During the Tree Search

Type and ID	*No. of species*	*No. of genes*	*Missing data (%)*	*Source*
DNA1	128	32	30	Stamatakis and Alachiotis (2010)
DNA2	237	74	72	Nyakatura and Bininda-Emonds (2012)
DNA3	372	79	66	Springer et al. (2012)
DNA4	404	11	60	Stamatakis and Alachiotis (2010)
AA1	69	31	35	De Queiroz et al. (1995)
AA2	70	35	34	
AA3	72	51	35	

4.2. Occurrence of partial terraces in real data

In this section we evaluate how often partial terraces occur in real alignments. By no means do we intend to make a full exploration of potential computing time that may be saved since the performance of the particular software will depend on the data structures and particular implementation used for the tree space exploration.

To elucidate the occurrence of partial terraces and full terraces, we analyzed seven recently published alignments (Table 1). Alignments have different numbers of taxa ranging from 69 to 404 taxa. The number of partitions (here, genes) varies from 11 to 79.

For each alignment we performed a maximum likelihood tree search using IQ-TREE (Nguyen et al., 2015) under the edge-unlinked (EUL) partition model assuming a GTR+Γ (Lanave et al., 1984; Yang, 1994) model for all partitions. We collected all the intermediate trees encountered during the search. For each intermediate tree T, we explored all trees T_{NNI} in its NNI neighborhood. We examined partial terraces of each T_{NNI} and T by computing how many induced partition trees are shared between them.

Apart from intermediate trees collected during the tree search, we also analyzed NNI neighborhoods for 1000 random Yule–Harding (YH) trees (Harding, 1971) for each tested alignment.

We defined 12 bins based on the percentage of shared induced partition trees between T and T_{NNI} (Table 2) and counted how many T_{NNI} trees fall into each bin. Table 3 shows the mean percentage of T_{NNI} trees that fall into the corresponding bin for the intermediate trees. Figure 7 displays the boxplots for the first three alignments from Table 1 either for the IQ-TREE search trees (left column) or the random YH trees (right column) (see Supplementary Figs. S1–S4 for the remaining alignments; Supplementary Material is available online at www.liebertonline.com/cmb).

Intermediate and random trees have similar percentages of T_{NNI} trees across different bins (Fig. 7 and Supplementary Figs. S1–S4). This suggests that the general picture of partial terraces is mainly determined by the spread of missing data in the supermatrix and is less dependent on the actual tree topology. Moreover, increasing the number of taxa tends to decrease the variance of T_{NNI} percentage within each bin (for both intermediate and random trees).

Figure 8 integrates the information from Tables 2 and 3 and provides rough estimates of potential computational savings if accounting for partial and full terraces. The green bars reflect the average percentage of identical induced partition trees when T_{NNI} is compared to T. For example, for DNA1 there is no

Table 2. Partial Terrace Bins Based on the Percentage of the Shared Partition Trees Between T and T_{NNI}

Name	*Percentage of shared partition trees out of the total number of partition trees*
No partial terrace (PT)	$=0\%$, the topologies of all partition trees are pairwise different between T and T_{NNI}
PT1	(0%,10%]
PT2	(10%, 20%]
PT3	(20%, 30%]
…	…
PT9	(80%, 90%]
PT10	(90%, 100%)
Full terrace	$=100\%$, T and T_{NNI} belong to one terrace

Table 3. Mean Percentage of Trees from NNI Neighborhood of Intermediate Trees Falling into Corresponding Partial Terrace Bin

	No PT (%)	*PT1 (%)*	*PT2 (%)*	*PT3 (%)*	*PT4 (%)*	*PT5 (%)*	*PT6 (%)*	*PT7 (%)*	*PT8 (%)*	*PT9 (%)*	*PT10 (%)*	*Full terrace (%)*
DNA1	7.14	12.97	4.85	1.80	3.55	32.59	37.11	0	0	0	0	0
DNA2	0	0	0.02	0.63	1.82	5.07	11.31	8.69	18.19	10.77	41.75	1.75
DNA3	0	2.75	5.38	9.22	10.06	6.32	4.34	1.33	0.23	6.38	50.36	3.63
DNA4	0.35	0.26	1.88	4.06	5.20	6.56	8.77	11.34	16.48	23.37	17.68	4.05
AA1	12.11	10.64	7.47	8.10	6.35	15.42	11.28	8.20	10.50	7.18	2.76	0
AA2	8.73	11.90	6.77	9.10	11.22	9.08	16.27	10.95	11.63	2.92	1.44	0
AA3	12.25	11.62	4.07	7.15	3.04	15.47	15.43	10.40	7.55	7.92	4.85	0.26

full terrace, but we observe partial terraces that may lead to a reduction of about 38% (the percentage of green bars) in computation time.

There is a full terrace for DNA2, but it consists of only 1.75% of the NNI neighborhood, whereas partial terraces constitute the remaining 98.25% and lead to a potential reduction of computations of about 80% (the percentage of green bars). In fact, since no T_{NNI} tree falls into ''no PT'' bin, we can save some computation time for all the trees encountered during tree search. Similar trend is observed for DNA3 and DNA4 with the predicted timesaving of 71% for each alignment.

5. ADVANTAGES OF USING INDUCED PARTITION TREES IN MAXIMUM LIKELIHOOD INFERENCE

In maximum likelihood inference, after applying a topological rearrangement on T, one needs to optimize the edge lengths of a new tree T_{NEW}. Therefore, together with the topological changes of partition trees, it is important to consider how topological rearrangement on T influences edge length optimization.

In the following we discuss two partition models commonly used in likelihood inferences, EUL and edge-linked (EL), and the advantages of using induced partition trees for either model (Yang, 1996).

We start by considering the most general partition model, EUL. Given a species tree T, we first obtain the corresponding induced partition trees. Under the EUL model, the edge lengths of the partition trees are optimized separately. The edge lengths of T are then computed from the corresponding edges lengths inferred on the partition trees, for example, as mean edge length.

Therefore, if the topological rearrangement on T does not change the topology of a partition tree $T|Y$, no edge length optimization is necessary and, as a result, the optimal partition tree likelihood remains unchanged after such a topological rearrangement on T. Let T_{NEW} be a tree obtained from T by some topological rearrangement.

Under EUL partition model there is no need to optimize the edge lengths of partitioned trees shared between T and T_{NEW}. As a result, the log-likelihood of the corresponding partition trees is the same.

In contrast to the EUL model, the edges between T and partition trees are linked in the EL model. That means that there is only one set of edge lengths for T and partition trees with the possibility of rescaling edge lengths of each partition tree by a partition-specific evolutionary rate. Therefore, the optimization of edge lengths is done on the species tree. Even if a topological rearrangement on T does not change the topology of partition tree, it still affects the optimal partition tree likelihood via optimization of edge lengths. This is also the reason why full terraces cannot occur under the EL model (Sanderson et al., 2015). Theoretically, one would need to optimize each edge on the species tree, which would definitely influence the partition tree edge lengths and also the likelihood. But in practice, to save computations, one only optimizes those edges in the vicinity of topological changes (Stamatakis et al., 2005; Guindon et al., 2010; Nguyen et al., 2015). For example, for an NNI, one reoptimizes only five edge lengths (e, e_1, e_2, e_3, e_4) around the swap. Under the EL model, such a particular feature of practical optimization can take an advantage when considering the induced partition trees.

Given a partition tree with taxon set Y and an edge e on T with the corresponding split $A|B$, if $A \cap Y = \emptyset$ or $B \cap Y = \emptyset$ then the optimization of e does not affect the likelihood of $T|Y$.

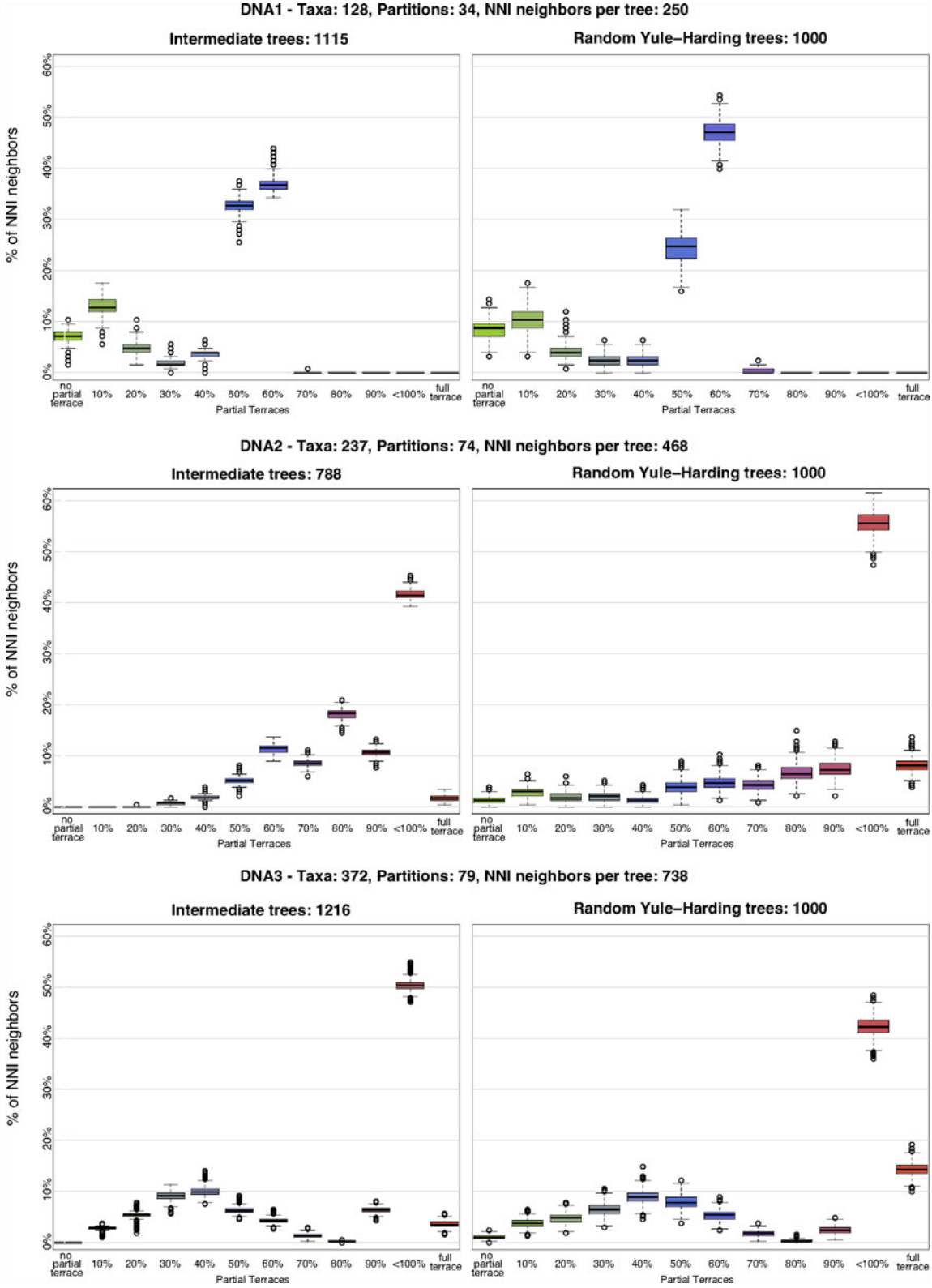

FIG. 7. NNI neighborhood analysis for alignments DNA1 (top), DNA2 (middle), and DNA3 (bottom).

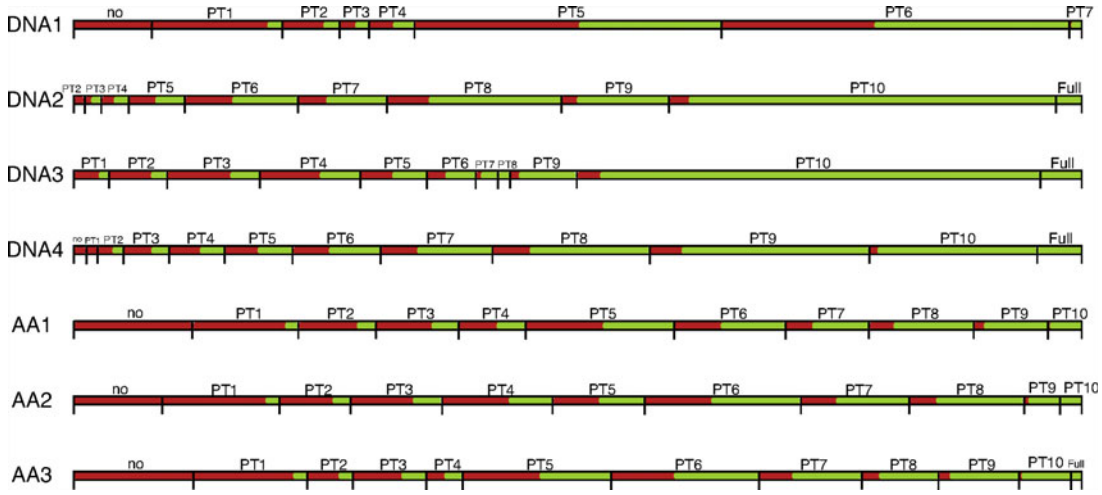

FIG. 8. Visualization of NNI neighborhoods and potential computational savings. Each horizontal line reflects the NNI neighborhood for each test alignment, that is 100% of T_{NNI} trees. These neighborhoods are divided into partial terrace bins (Table 2) and depicted here by horizontal segments. The length of each segment corresponds to the mean percentage of T_{NNI} trees falling into the bin (Table 3). Each segment is composed of a green and a red bar, corresponding to the fractions of partition trees that are shared and not shared between T and T_{NNI}, respectively. Basically, green bars indicate potential computational savings when accounting for partial and full terraces during the trees search.

In this case, a split $A|B$ does not have a corresponding split in $\Sigma(T|Y)$, and therefore edge e is not linked to any edge on $T|Y$. This observation can be exploited to save computing time.

6. DISCUSSION

We have shown that it is advantageous to identify and account for full and partial terraces during the tree search in phylogenomics. One main advantage is the saving of computation time. If two trees belong to the same full or partial terrace, then one needs to compute the objective function for the identical partition trees only once. The values of objective function will be the same for these partition trees. The larger the number of identical partition trees between species trees, the more computation time can be saved.

From the conditions discussed in the previous sections, the topological rearrangement that benefits the most from partial terraces is obviously NNI. It is intuitive that NNI applied to the species tree will not change the topology of partition trees more often than SPR or TBR. However, in tree searches one typically applies short SPR (e.g., RAxML); that is, the number of edges between the pruning and the regrafting edges are much smaller than the number of taxa. The same is true for TBR. And since one also expects short SPR and short TBR to result in no change of partition trees quite often for sparse supermatrices, partial terraces are also beneficial for these rearrangements.

Moreover, the use of induced partition trees has another advantage that long SPR or TBR on a species tree T, as a result of missing data, might correspond to a much shorter SPR or TBR on $T|Y$. This leads to computation saving even if SPR or TBR changes the topology of the induced partition trees.

Here, we elucidated the frequent existence of partial terraces in practice via NNI neighborhoods, showing that partial terraces are not only a theoretical concept, but also have practical implications in phylogenomics. The predicted timesaving for the examined real alignments is only the rough estimate, since we treated the alignment lengths per partition as equal. If the length of alignment corresponding to the shared partition trees is relatively large compared to the whole supermatrix, then one expects even more speed up.

Another important factor for timesaving is the actual implementation of search strategies in the particular software. We plan to implement efficient techniques to take full advantage of partial and full terraces in IQ-TREE. A more thorough analysis of such techniques will be presented elsewhere.

ACKNOWLEDGMENTS

This work was supported by the Austrian Science Fund (FWF, grant number I-1824-B22) to O.C.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

De Queiroz, A., Donoghue, M.J., and Kim J. 1995. Separate versus combined analysis of phylogenetic evidence. *Annu. Rev. Ecol. Syst.* 26, 657–681.

De Queiroz, A., and Gatesy, J. 2007. The supermatrix approach to systematics. *Trends Ecol. Evol.* 22, 34–41.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, MA.

Guindon, S., Dufayard, J.F., Lefort, V., et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321.

Harding, E.F. 1971. The probabilities of rooted tree shapes generated by random bifurcation. *Adv. Appl. Probability*. 3, 44–77.

Hedtke, S.M., Patiny, S., and Danforth, B.N. 2013. The bee tree of life: A supermatrix approach to apoid phylogeny and biogeography. *BMC Evol. Biol.* 13, 138.

Hordijk, W., and Gascuel, O. 2005. Improving the efficiency of SPR moves in phylogenetic tree search methods based on maximum likelihood. *Bioinformatics.* 21, 4338–4347.

Lanave, C., Preparata, G., Saccone, C., et al. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20, 86–93.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., et al. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274.

Nyakatura, K., and Bininda-Emonds, O.R.P. 2012. Updating the evolutionary history of Carnivora (Mammalia): A new species-level supertree complete with divergence time estimates. *BMC Biol.* 10, 12.

Peters, R.S., Meyer, B., Krogmann, L., et al. 2011. The taming of an impossible child: A standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. *BMC Biol.* 9, 55.

Pyron, R.A., Burbrink, F.T., Colli, G.R., et al. 2011. The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and comparison of support methods for likelihood trees. *Mol. Phylogenet. Evol.* 58, 329–342.

Pyron, R.A., and Wiens, J.J. 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Mol. Phylogenet. Evol.* 61, 543–583.

Robinson, D.F., and Foulds, L.R. 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147.

Sanderson, M.J., McMahon, M.M., Stamatakis, A., et al. 2015. Impacts of terraces on phylogenetic inference. *Syst. Biol*. 64, 709–726.

Sanderson, M.J., McMahon, M.M., and Steel, M. 2011. Terraces in phylogenetic tree space. *Science.* 333, 448–450.

Sanderson, M.J., Purvis, A., and Henze, C. 1998. Phylogenetic supertrees: Assembling the trees of life. *Trends Ecol. Evol.* 13, 105–109.

Semple, C., and Steel, M.A. 2003. *Phylogenetics*. Oxford University Press, New York, NY.

Springer, M.S., Meredith, R.W., Gatesy, J., et al. 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. *PLoS ONE* 7, e49521.

Stamatakis, A., and Alachiotis, N. 2010. Time and memory efficient likelihood-based tree searches on phylogenomic alignments with missing data. *Bioinformatics.* 26, i132–i139.

Stamatakis, A., Ludwig, T., and Meier, H. 2005. RAxML-III: A fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21, 456–463.

van der Linde, K., Houle, D., Spicer, G.S., et al. 2010. A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genet. Res.* 92, 25–38.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* 39, 306–314.

Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42, 587–596.

Address correspondence to:
Olga Chernomor
Max F. Perutz Laboratories
Center for Integrative Bioinformatics Vienna
Dr. Bohrgasse 9
A-1030 Vienna
Austria

E-mail: olga.chernomor@univie.ac.at