Annals of Combinatorics 8 (2004) 1-●●●
0218-0006/04/010001-11
DOI **********

Annals of Combinatorics

The Splits in the Neighborhood of a Tree

David Bryant

bryant@mcb.mcgill.ca
McGill Centre for Bioinformatics, 3775 University, Montréal, Québec, H3A 2B4, Canada

Received April 17, 2003

AMS Subject Classification: 68R10, 05C05, 68Q25, 92D15

Abstract. A *phylogenetic tree* represents historical evolutionary relationships between different species or organisms. The space of possible phylogenetic trees is both complex and exponentially large. Here we study combinatorial features of neighbourhoods within this space, with respect to four standard tree metrics. We focus on the *splits* of a tree: the bipartitions induced by removing a single edge from the tree. We characterize those splits appearing in trees that are within a given distance of the original tree, demonstrating close connections between these splits, the Whitney number of a tree, and the binary characters with a given parsimony length.

Keywords: Phylogenetic tree, splits, tree metric, Whitney numbers, parsimony

1. Introduction

The reconstruction of evolutionary trees is one of the primary objectives of phylogenetics. We search through tree space to find the tree that optimizes some objective function or, under a Bayesian approach, survey those trees with high posterior probability. These searches motivate the study of the local structure of tree space: Search algorithms search through neighborhoods; Trees with high likelihood tend to come in clusters; Validation requires a concept of a local "confidence region."

Tree space is, of course, somewhat complicated. It is huge — the number of possible phylogenetic trees with n leaves grows super-exponentially with n. The space is a mixture continuous and combinatorial aspects and has a complicated geometry [2].

Nevertheless, there are several combinatorial tools available to simplify our study of tree space. One of the most fundamental is the decomposition of phylogenetic trees into collections of splits. The splits correspond to 1-cuts of the tree: the bipartitions induced by removing single edges. We can reconstruct a tree from its splits. Indeed, splits are central to the geometry of tree space described by [2], where they correspond to dimensions in each orthant.

Considering trees as collections of splits has important algorithmic consequences. We have shown in [3–5] that constraining a tree search to trees with splits contained within a given set can make several NP-hard optimization problems polynomial time

solvable. Once we understand the splits in the neighborhood of a tree, we can use this information to design more efficient tree searching algorithms. In this paper we show that the splits in neighbouring trees have an elegant characterization.

The outline of the paper is as follows. After presenting basic terminology, we define the four tree metrics under study (Section 2). The first metric considered is the Robinson-Foulds (or partition) metric d_{RF}. In Section 3 we present a graphical characterization of the splits in trees T' with $d_{RF}(T,T') \leq r$, using this characterization to prove that the number of these splits is linear in the number of leaves for bounded r. We also characterize the split neighborhood for the weighted version of d_{RF}. These results are extended to the nearest neighbor interchange metric d_{NN} in Section 4. In Section 5 we discuss the subtree prune and regraft metric d_{SPR} and the tree bisection and reconnection metric d_{TBR}, proving that trees within distance r contain exactly those splits which correspond to binary characters of parsimony length at most $r+1$. This result leads to an exact formula for the size of the number of splits in the neighborhood of a tree, under these two metrics. It also yields an exact formula for the number of trees within SPR or TBR distance r of a tree containing a given split.

2. Preliminaries

2.1. Terminology

A *phylogenetic X-tree* T is a tree with leaf set X and no vertices of degree two. We use $\mathring{V}(T)$ and $\mathring{E}(T)$ to denote the interior vertices and edges of T. Let $\mathring{T}$ denote the subgraph of T formed from the edges $\mathring{E}(T)$ together with their incident vertices. If every interior vertex has degree three we say that T is *fully resolved* (or *binary*). Unless otherwise stated, the phylogenetic trees we discuss will be fully resolved.

A *split* $A|B$ of X is a partition of X into two non-empty blocks, A and B. Removing an edge e from a phylogenetic X-tree divides the tree into two connected components, thereby inducing a split of the leaf set X. We say that this is the split associated with e. The collection of all the splits associated with edges of T is called the *splits of* T and denoted $\Sigma(T)$. A given collection $\mathcal{S}$ of splits of X is *compatible* if it is contained within the splits of some tree T, which holds if and only if for every pair $A|B, C|D$ of splits in $\mathcal{S}$ at least one of the intersections $A\cap C, A\cap D, B\cap C, B\cap D$ is empty [6].

Let $UB(X)$ denote the set of binary phylogenetic X-trees and let d be a metric defined on $UB(X)$. The *r-neighborhood* of T with respect to d equals the set of trees

$$N_d(T,r) = \{T' \in UB(X) \colon d(T,T') \leq r\}.$$

The *split neigborhood* of T is the set of splits appearing in at least one of the trees in the r neighborhood of T:

$$\begin{aligned} S_d(T,r) &= \{A|B \colon \text{ there exists } T' \in N_d(T,r) \text{ such that } A|B \in \Sigma(T)\} \\ &= \bigcup_{T' \in N_d(T,r)} \Sigma(T'). \end{aligned}$$

In this paper we present characterizations of $S_d(T,r)$ for the four most widely used tree metrics. The characterizations lead to exact formulae for $|S_d(T,r)|$ in some cases,

and an asymptotic bound in others. The results draw on connections with Whitney numbers in trees and parsimony lengths of binary characters (defined in Section 5).

2.2. Robinson-Foulds Distance (RF)

The *Robinson-Foulds metric*, also called the *partition metric*, was first proposed by [15] and is one of the simplest metrics on trees. The distance between two X-trees T_1 and T_2 is defined

$$d_{RF}(T_1, T_2) = \tfrac{1}{2}\big|\Sigma(T_1) \triangle \Sigma(T_2)\big| = \tfrac{1}{2}\big|(\Sigma(T_1) - \Sigma(T_2)\big| + \tfrac{1}{2}\big|\Sigma(T_2) - \Sigma(T_1)\big|$$

The Robinson-Foulds metric between two phylogenetic X trees can be computed in $O(|X|)$ time [10].

The Robinson-Foulds metric may be extended to trees with weighted edges [14]. Suppose that every edge e in T_1 and T_2 has been assigned a non-negative length. For each split $A|B \in \Sigma(T_i)$, $i = 1, 2$, we let $w_i(A|B)$ denote the length of the edge corresponding to the split $A|B$. We set $w_i(A|B) = 0$ for all $A|B \notin \Sigma(T_i)$. The weighted Robinson-Foulds distance d_ω is then defined

$$d_\omega(T_1, T_2) = \sum_{A|B \in \Sigma(T_1) \cup \Sigma(T_2)} |w_1(A|B) - w_2(A|B)|.$$

2.3. Nearest Neighbor Interchange Metric (NNI)

For every fully resolved X-tree T with n leaves there are exactly $2(n-3)$ X-trees with Robinson-Foulds distance one from T. These correspond to trees obtained by swapping two subtrees in a tree that are adjacent to the same internal edge (Figure 1). The process of going from a tree to a neighbouring tree in this way is called a *nearest neighbor interchange*.

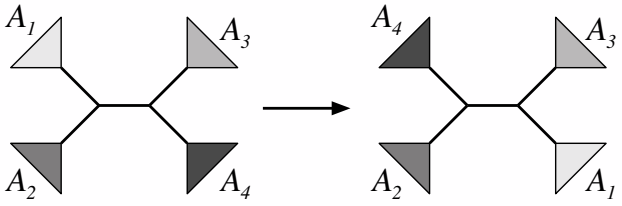

Figure 1: A Nearest Neighbor Interchange (NNI). A_1, A_2, A_3, A_4 represent subtrees. The positions of subtrees A_1 and A_4 are exchanged. The remainder of the tree is unchanged.

Every binary tree T_1 on n leaves can be reached from every other binary tree T_2 on n leaves by a sequence of nearest neighbor interchanges. The Nearest-Neighbor distance equals minimum number of nearest neighbor interchanges required to transform T_1 into T_2 and is denoted $d_{NN}(T_1, T_2)$ [13]. DasGupta et al. [9] proved that determining

$d_{NN}(T_1, T_2)$ is NP-hard. Since performing a nearest neighbor interchange on T_2 can decrease $d_{RF}(T_1, T_2)$ by at most one we always have $d_{NN}(T_1, T_2) \geq d_{RF}(T_1, T_2)$.

2.4. Subtree Prune and Regraft Distance (SPR)

A *subtree prune and regraft (SPR)* proceeds in three steps. We select and remove an edge $\{u, v\}$ of the tree, thereby dividing the tree into two connected subtrees T_u (containing u) and T_v (containing v). We then select and subdivide an edge of T_v, giving a new vertex w. Finally we connect u and w by an edge and suppress all vertices of degree two. The process is illustrated in Figure 2.

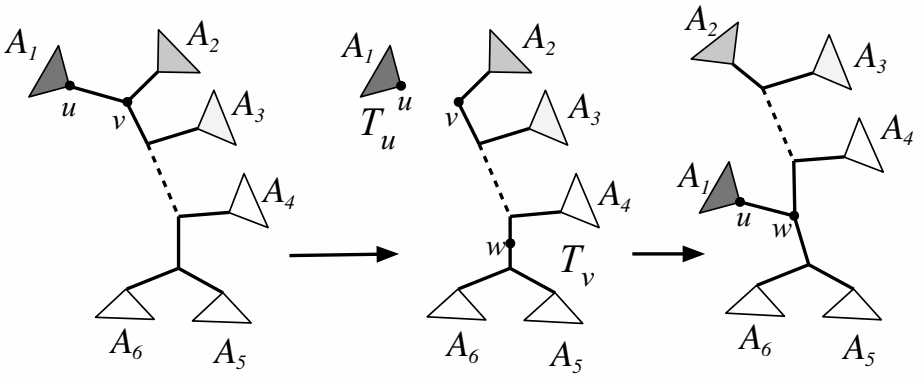

Figure 2: A Subtree Prune and Regraft (SPR). The edge $\{u, v\}$ is removed, giving two trees T_u and T_v. An edge of T_v is subdivided, giving a new vertex w. This is connected to u and the degree two vertex v is suppressed.

The subtree prune and regraft distance $d_{SPR}(T_1, T_2)$ is the number of SPRs required to transform T_1 into T_2, or vice versa. Every NNI is also an SPR, so we have $d_{SPR}(T_1, T_2) \leq d_{NNI}(T_1, T_2)$. Hein et al. [11] claimed that computing $d_{SPR}(T_1, T_2)$ is NP-hard, though an error in their proof was found by [1]. The computational complexity remains open.

2.5. Tree Bisection Reconnection Metric (TBR)

A *tree bisection and reconnection (TBR)* is similar to a SPR except that once we have removed $\{u, v\}$ we subdivide an edge from T_v *and* and an edge from T_u, connecting the two new vertices with an edge and suppressing vertices of degree two. If either of T_u or T_v consists of only a single vertex the TBR corresponds to a reattachment of this vertex to another part of a tree. The process is illustrated in Figure 3. We use $d_{TBR}(T_1, T_2)$ to denote the number of TBRs required to transform T_1 into T_2. Every SPR is a TBR, so $d_{TBR}(T_1, T_2) \leq d_{SPR}(T_1, T_2)$. Allen and Steel [1] proved that computing $d_{TBR}(T_1, T_2)$ is an NP-hard problem but showed that the problem is fixed parameter tractible (with parameter $d_{TBR}(T_1, T_2)$).

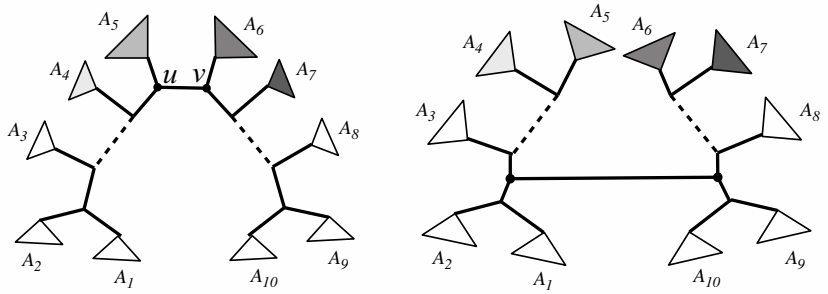

Figure 3: A Tree Bisection and Reconnection (TBR). The edge $\{u, v\}$ is removed, giving two trees. An edge from each tree is subdivided, and the new vertices are connected by an edge.

3. Splits in the Robinson-Foulds Neighborhood

Let T be a fully resolved phylogenetic X-tree and let $A|B$ be a split of X. If $A|B \in \Sigma(T)$ then $A|B$ is pairwise compatible with every split of $\Sigma(T)$. If $A|B \notin \Sigma(T)$ then $A|B$ is pairwise incompatible with some of the splits in $\Sigma(T)$, since $\Sigma(T)$ is itself a maximal pairwise compatible collection of splits. We say that these are the splits of T *conflicting* with $A|B$. The associated edges of T are the *edges conflicting with* $A|B$. Note that these edges will always be internal edges of T.

Lemma 3.1. *If T is a fully resolved X-tree and $A|B$ is a split of X then the edges of T conflicting with $A|B$ form a connected subgraph of T.*

Proof. The result holds vacuously if $A|B$ conflicts with fewer than two edges of T. Otherwise, suppose that e_1 and e_k are edges of T conflicting with $A|B$ and that $e_1, e_2, \ldots, e_k$ are the edges along the path connecting e_1 and e_k in T. We show that $e_2, e_3, \ldots, e_{k-1}$ also conflict with $A|B$.

For each $i = 1, 2, \ldots, k$ let $X_i|Y_i$ be the split associated to e_i, where $X_1 \subset X_2 \subset \cdots \subset X_k$. As $X_1|Y_1$ and $A|B$ are incompatible there is $a \in X_1 \cap A$ and $b \in X_1 \cap B$. Similarly there is $a' \in Y_k \cap A$ and $b' \in Y_k \cap B$. Hence for all $i = 1, 2, \ldots, k$ we have $a \in X_i \cap A$, $b \in X_i \cap B$, $a' \in Y_i \cap A$ and $b' \in Y_i \cap B$, and $A|B$ is incompatible with $X_i|Y_i$. ∎

There is a straightforward characterization of the set of all splits conflicting with exactly those edges in a given connected subset E' of $\mathring{E}(T)$. Let V' be the vertices incident with edges in E'. Let $\pi(E') = A_1|A_2|\cdots|A_k$ be the partition of X induced by the components of $T - V'$. We say that two blocks A_i, A_j of π are *adjacent* if they are contained in the same component of $T - E'$. The splits conflicting with exactly those edges in E' are the splits $A|B$ satisfying

(i) For all $i = 1, 2, \ldots, k$, either $A_i \subseteq A$ or $A_i \subset B$.

(ii) Adjacent blocks are on different sides $A|B$,

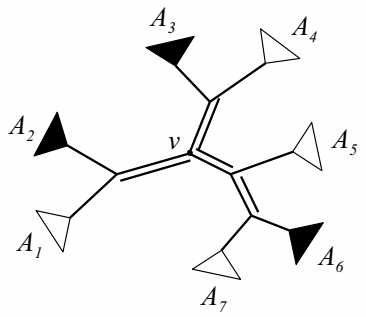

Figure 4: A set E' of conflicting edges and a split conflicting with exactly those edges. The edges in E' are marked with double lines. The blocks of $\pi(E')$ are labelled $A_1, \ldots, A_7$. The pairs $\{A_1, A_2\}$, $\{A_3, A_4\}$ and $\{A_6, A_7\}$ form adjacent pairs. One particular split conflicting with exactly those edges in E' is represented by the colouring of the blocks. Note that, for this set, v is a conflicting vertex (see Section 4).

as illustrated in Figure 4.

A simple induction proof based on this characterization gives

Lemma 3.2. *Let E' be a connected collection of k interior edges in T. There are exactly 2^k splits that conflict with exactly the edges in E'.*

We are now ready to characterize the splits in the Robinson-Foulds neighborhood.

Theorem 3.3. *Let T be a fully resolved phylogenetic X-tree. A split $A|B$ is in $S_{RF}(T, r)$ if and only if it conflicts with at most r edges of T.*

Proof. Suppose that $A|B \in \Sigma(T')$ and $d_{RF}(T, T') \leq r$. Then there are at most r splits in $\Sigma(T) - \Sigma(T')$. Since $A|B$ is compatible with all of the splits in $\Sigma(T')$ it is compatible with all but at most r splits in $\Sigma(T)$.

Conversely, suppose that $A|B$ conflicts with at most r edges of T. Let S be the associated set of conflicting splits. Then $(\Sigma(T) - S) \cup \{A|B\}$ is compatible, so there is a phylogenetic X-tree T' that is fully resolved, contains the splits $(\Sigma(T) - S) \cup \{A|B\}$, and therefore satisfies $d_{RF}(T, T') \leq r$. ∎

Recall that $\mathring{T}$ is the subgraph formed from the internal edges and vertices of T. Theorem 3.3 provides a direct connection between the splits in $S_{RF}(T, r)$ and the connected subgraphs of $\mathring{T}$. A connected subgraph of $\mathring{T}$ with k vertices (and therefore $k-1$ edges) is called a *k-subtree* of $\mathring{T}$, and the number of these subtrees is called the *Whitney number* of $\mathring{T}$ [12]. There is no general closed formula for the Whitney number of a tree, though Jamison [12] provides a recursive formula for the generating function for a fixed tree. The trees with the largest Whitney numbers have high degree nodes. Restricting our attention to fully resolved trees permits an upper bound on the Whitney number that is linear in n (for bounded k).

Lemma 3.4. *Let T be a fully resolved phylogenetic X-tree. The number of k-subtrees of $\mathring{T}$ is $O(nC_k)$, where $n = |X|$ and C_k is the kth Catalan number.*

Proof. Select an arbitrary vertex of $\mathring{T}$ and direct all of the edges in $\mathring{T}$ away from this vertex. This orientation turns all of the k-subtrees of $\mathring{T}$ into rooted directed subtrees. For every $v \in V(\mathring{T})$, the number of k-subtrees rooted at v is bounded above by C_{k+1}, the number of (ordered) binary trees with k internal nodes. There are $n-2$ vertices in $\mathring{T}$, giving an upper bound of $(n-2)C_k$ connected subtrees with k internal nodes. ∎

Lemma 3.2, Theorem 3.3 and Lemma 3.4 lead directly to

Corollary 3.5. *The number of splits in $S_{RF}(T, r)$ is linear in n for bounded r.*

For the weighted case, we can derive a characterization of $S_\omega(T, r)$ directly analogous to that for the unweighted case. Note that $A|B \in S_\omega(T, r)$ if there is a weighted tree T' such that $d_\omega(T, T') \leq r$ and $A|B \in \Sigma(T')$, even if the edge corresponding to $A|B$ in T' has length zero.

Theorem 3.6. *Let T be a fully resolved phylogenetic X-tree. A split $A|B$ is in $S_\omega(T, r)$ if and only if it conflicts with edges of T with total summed length at most r.*

Proof. Suppose that $T_1 = T$, $A|B \in \Sigma(T_2)$ and $d_\omega(T_1, T_2) \leq r$. Let E' denote the edges of T_1 conflicting with $A|B$. Every edge in E' corresponds to a split in $\Sigma(T_1) - \Sigma(T_2)$. Hence

$$\begin{aligned}
\sum_{e \in E'} (\text{length of } e) &\leq \sum_{C|D \in \Sigma(T_1) - \Sigma(T_2)} w_1(C|D) \\
&= \sum_{C|D \in \Sigma(T_1) - \Sigma(T_2)} \Big| w_1(C|D) - w_2(C|D) \Big| \\
&\leq d_\omega(T_1, T_2) \\
&\leq r.
\end{aligned}$$

Conversely, suppose that $A|B$ conflicts with edges of $T_1 = T$ with total summed edge length at most r. Let S be the associated set of conflicting splits. Then $(\Sigma(T) - S) \cup \{A|B\}$ is compatible, so there is a fully resolved X-tree T_2 that is fully resolved and contains the splits $(\Sigma(T) - S) \cup \{A|B\}$. Assign zero lengths to all edges of T_2 associated to splits not in $\Sigma(T_1)$ and leave all other edge lengths the same. We then have that $A|B \in \Sigma(T_2)$ and $d_\omega(T_1, T_2) \leq r$. ∎

4. Splits in the Nearest Neighbour Interchange Neighbourhood

We now extend the characterization of splits in the Robinson-Foulds neighborhood to the Nearest Neighbor Interchange neighborhood. For any two phylogenetic trees T_1, T_2 we have $d_{RF}(T_1, T_2) \leq d_{NN}(T_1, T_2)$. Consequently, $S_{NN}(T, r) \subseteq S_{RF}(T, r)$ and the number of splits in the NNI neighborhood is also linear for bounded r. These splits have an elegant graphical characterization.

Let v be an internal vertex in a fully resolved phylogenetic X-tree and let $A|B$ be a split of X. We say that v is a *vertex conflicting with* $A|B$ if every edge incident with v conflicts with $A|B$ (see Figure 4).

Theorem 4.1. *Let T be a fully resolved phylogenetic X-tree, let $A|B$ be a split of X and let E', V' be the edges and vertices of T conflicting with $A|B$. Then $A|B$ is in $S_{NN}(T,r)$ if and only if $|E'|+|V'| \leq r$.*

Proof. Suppose that $A|B \in \Sigma(T')$ and $d_{NN}(T,T') = s \leq r$. There is a sequence $T' = T_0, T_1, T_2, \ldots, T_s = T$ of phylogenetic X-trees such that for each i, T_{i+1} differs from T_i by one NNI. Let E'_i and V'_i denote the edges and vertices of T_i conflicting with $A|B$.

We claim that $|E'_i|+|V'_i| \leq i$ for all $i = 0, 1, 2, \ldots, s$, implying $|E'|+|V'| = |E'_s|+|V'_s| \leq s \leq r$. The bound clearly holds for $i = 0$. Suppose that it holds for all $i \leq j$ and that T_{j+1} is obtained from T_j by an NNI around the edge $\{u,v\}$. If u and v are both conflicting vertices then performing an NNI will not affect the number of conflicting edges nor the number of conflicting vertices. If one of u, v is a conflicting vertex, then performing an NNI will not change the number of conflicting edges, though it can increase the number of conflicting vertices by one. If neither u nor v are conflicting vertices then the NNI can increase the number of conflicting edges by one, but has no effect on the number of conflicting vertices. Thus $|E'_{j+1}|+|V'_{j+1}| \leq |E'_j|+|V'_j|+1 \leq j+1$. The result follows by induction.

Conversely, suppose that $A|B$ conflicts with the edges E' and vertices V' of T and that $|E'|+|V'| \leq r$. Choose an edge $\{u,v\}$ of E' such that u is adjacent to no other edges in E'. If v is a conflicting vertex then performing any NNI around $\{u,v\}$ gives a tree with one fewer conflicting vertices and the same number of conflicting edges. If v is not a conflicting vertex we can perform an NNI giving a tree with the same number of conflicting vertices and one fewer conflicting edge. Repeating the process $|E'|+|V'|$ times gives an X-tree T' that contains the split $A|B$ and satisfies $d_{NN}(T,T') \leq r$. ∎

5. Splits in the SPR and TBR Neighborhoods

Every NNI is an SPR and every SPR is a TBR, so for any two phylogenetic X-trees T_1, T_2 we have

$$d_{NN}(T_1,T_2) \geq d_{SPR}(T_1,T_2) \geq d_{TBR}(T_1,T_2)$$

with strict inequality in some cases. The split neighborhoods are therefore nested:

$$S_{NN}(T,r) \subseteq S_{SPR}(T,r) \subseteq S_{TBR}(T_1,T_2).$$

We show here that, in fact, the last two split neighborhoods are equal, and they are substantially larger than the NNI neighborhood. Our key result is a connection between the split neighborhoods for d_{SPR} and d_{TBR} and the parsimony length of a character.

A *binary character* for X is a function $\chi : X \longrightarrow \{0,1\}$. An *extension* of χ on a phylogenetic X-tree T is a function $\hat{\chi} : V(T) \to \{0,1\}$ such that the restriction of $\hat{\chi}$ to X equals χ. The *length* of $\hat{\chi}$ on T, denoted $\hat{l}_T(\hat{\chi})$, equals the number of edges $\{u,v\} \in E(T)$ for which $\hat{\chi}(u) \neq \hat{\chi}(v)$. The *parsimony length* of χ on T is then the minimum of $\hat{l}_T(\hat{\chi})$ over all extensions $\hat{\chi}$ of χ. We denote this length by $l_T(\chi)$.

Lemma 5.1. *Suppose that T' differs from T by one TBR operation. For any character χ we have $l_{T'}(\chi) \leq l_T(\chi) + 1$.*

Proof. Let $\hat{\chi}$ be a minimum length extension of χ on T, so $\hat{l}_T(\hat{\chi}) = l_T(\chi)$. A TBR is carried out in two stages. First we remove an edge $\{u, v\}$ of T and suppress any vertices of degree two. Suppose that $A|B$ is the split of T corresponding to the edge $\{u, v\}$. This first step gives two trees T_A and T_B, the first with leaf set A and the second with leaf set B. Let $\hat{\chi}_A$ and $\hat{\chi}_B$ be the restrictions of $\hat{\chi}$ to $V(T_A)$ and $V(T_B)$. We then have

$$\hat{l}_{T_A}(\hat{\chi}_A) + \hat{l}_{T_B}(\hat{\chi}_B) \leq \hat{l}_T(\hat{\chi})$$

since removing an edge and suppressing degree two vertices does not increase length.

The second stage of a TBR is to reconnect T_A and T_B. If one of T_A or T_B consists of a single vertex only, reconnecting this vertex to the other tree will produce an increase of at most one step. Otherwise we insert a new vertex x_A along an edge $\{u_A, v_A\}$ of T_A and a new vertex x_B along an edge $\{u_B, v_B\}$ of T_B. We then connect x_A and x_B with an edge, giving the tree T'. Setting $\hat{\chi}(x_A) = \hat{\chi}(u_A)$ and $\hat{\chi}(x_B) = \hat{\chi}(u_B)$ gives an extension of χ to the vertices of T' of length at most $l_T(\chi) + 1$. ∎

For any split $A|B$ of X we let $\chi_{A|B}$ denote the character

$$\chi_{A|B}(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 5.2. *Let T be a fully resolved phylogenetic X-tree and let $A|B$ be a split of X. The following three statements are equivalent*:

(i) $A|B \in S_{SPR}(T, r)$;

(ii) $A|B \in S_{TBR}(T, r)$;

(iii) $l_T(\chi_{A|B}) \leq r + 1$.

Proof. We have already established that (i) implies (ii).

Suppose that (ii) holds, that $A|B \in \Sigma(T')$ and $d_{TBR}(T, T') = s \leq r$. There is a sequence of phylogenetic X-trees $T' = T_0, T_1, T_2, \ldots, T_s = T$. Since $A|B \in \Sigma(T')$ we have $l_{T'}(\chi_{A|B}) = 1$. By Lemma 5.1 we have for each $i = 1, 2, \ldots, s$ that $l_{T_i}(\chi_{A|B}) \leq l_{T_{i-1}}(\chi_{A|B}) + 1$, so $l_T(\chi_{A|B}) \leq s + 1 \leq r + 1$.

Finally, suppose that (iii) holds and that $l_T(\chi_{A|B}) = s + 1 \leq r + 1$. If $s = 0$ then we are done, since $l_T(\chi_{A|B}) = 1$ if and only if $A|B \in \Sigma(T)$. Otherwise, let $\hat{\chi}$ be a minimum length extension of $\chi_{A|B}$. We can find three vertices u, v, w such that $\{u, v\} \in E(T)$, u lies on the path from v to w, and $\hat{\chi}(u) \neq \hat{\chi}(v) = \hat{\chi}(w)$. Perform an SPR by removing the edge $\{u, v\}$, inserting a new vertex x along an edge adjacent to w, and adding the edge v, x. Set $\hat{\chi}(x) = \hat{\chi}(v)$. The extension $\hat{\chi}$ will have length s in this new tree. Repeating the process we obtain a tree T' such that $A|B \in \Sigma(T')$ and $d_{SPR}(T, T') = s$. ∎

Charleston and Steel [8] present an exact formula for the number of characters of parsimony length k in a fully resolved phylogenetic tree. This result, together with Theorem 5.2, gives an exact formula for the number of splits in $S_{SPR}(T, r)$ and $S_{TBR}(T, r)$.

Corollary 5.3. *Let T be a fully resolved X-tree and let $n = |X|$. Then*

$$|S_{SPR}(T,r)| = |S_{TBR}(T,r)| = \sum_{k=1}^{r+1} \left(\binom{n-k}{k} + \binom{n-k-1}{k} \right) 2^k.$$

Hence the size of $S_{SPR}(T,r)$ and $S_{TBR}(T,r)$ grows much faster than the size of $S_{RF}(T,r)$ and $S_{NNI}(T,r)$. Suppose we are given a split $A|B$. Since $S_{SPR}(T,r)$ and $S_{TBR}(T,r)$ grow quickly, one would expect that the number of trees that are within SPR or TBR distance r of a tree containing $A|B$ will also grow. In fact we can derive an exact formula for the number of such trees, a result of importance when we consider the significance of finding a tree that *almost* contains a split.

Corollary 5.4. *Let $A|B$ be a split of X, let $n = |X|$ and $k = |A|$. The number of fully resolved X-trees T that are within SPR or TBR distance r of a tree containing $A|B$ equals*

$$\sum_{l=1}^{r+1} 2^l \frac{(2n-3l)(2k-l-1)!(2(n-k)-l-1)!(n-l)!}{(k-l)!(n-k-l)!(l-1)!(2n-2l)!}.$$

Proof. By Theorem 5.2, a tree T is within SPR or TBR distance r of a tree containing $A|B$ if and only if the length of the binary character $\chi_{A|B}$ on T is at most $r+1$. Carter et al. [7] showed that the number of fully resolved trees on which $\chi_{A|B}$ has length l equals

$$2^l \frac{(2n-3l)(2k-l-1)!(2(n-k)-l-1)!(n-l)!}{(k-l)!(n-k-l)!(l-1)!(2n-2l)!}$$

(see also [16]) . The result then follows by summing over l. ∎

Acknowledgments. Thanks to Mike Steel, Rachel Bevan and Trevor Bruen for helpful suggestions. This work was supported by NSERC grant number 238975-01 and FQRNT grant number 2003-NC-81840.

References

1. B.L. Allen and M. Steel, Subtree transfer operations and their induced metrics on evolutionary trees, Ann. Combin. **5** (1) (2001) 1–15.
2. L.J. Billera, S.P. Holmes, and K. Vogtman, Geometry of the space of phylogenetic trees, Adv. Appl. Math. **27** (4) (2001) 733–767.
3. D. Bryant, Hunting for trees in binary character sets, J. Comput. Biol. **3** (2) (1996) 275–288.
4. D. Bryant, Building trees, hunting for trees and comparing trees, Ph.D. Thesis, Department of Mathematics, University of Canterbury, 1997.
5. D. Bryant and M. Steel, Constructing optimal trees from quartets, J. Algorithms **38** (1) (2001) 237–259, Tenth Annual ACM-SIAM Symposium on Discrete Algorithms, Baltimore, MD, 1999.
6. P. Buneman, The recovery of trees from measures of dissimilarity, In: Mathematics in the Archaeological and Historical Sciences, F.R. Hodson, D.G. Kendall, and P. Tautu, Eds., Edinburgh University Press, Edinburgh, 1971, pp. 387–395.
7. M. Carter, M. Hendy, D. Penny, L.A. Székely, and N.C. Wormald, On the distributions of lengths of evolutionary trees, SIAM J. Discrete Math. **3** (1990) 38–47.

8. M. Charleston and M. Steel, Five surprising properties of parsimoniously colored trees, Bull. Math. Biol. **57** (2) (1993) 367–375.
9. B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang, On computing the nearest neighbor interchange distance, In: Discrete Mathematical Problems with Medical Applications, New Brunswick, NJ, 1999, DIMACS Ser. Discrete Math. Theoret. Comput. Sci., Vol. 55, Amer. Math. Soc., Providence, RI, 2000, pp. 125–143.
10. W.H.E. Day, Optimal algorithms for comparing trees with labeled leaves, J. Classification **2** (1985) 7–28.
11. J. Hein, T. Jiang, L. Wang, and K. Zhang, On the complexity of comparing evolutionary trees, Discrete Appl. Math. **71** (1-3) (1996) 153–169.
12. R.E. Jamison, Alternating Whitney sums and matchings in trees, I, Discrete Math. **67** (2) (1987) 177–189.
13. D.F. Robinson, Comparison of labeled trees with valency three, J. Combin. Theory, Ser. B **11** (1971) 105–119.
14. D.F. Robinson and L.R. Foulds, Comparison of weighted labelled trees, In: Combinatorial Mathematics, VI, Proc. Sixth Austral. Conf., Univ. New England, Armidale, 1978, Lecture Notes in Mathematics, Vol. 748, Springer-Verlag, Berlin, 1979, pp. 119–126.
15. D.F. Robinson and L.R. Foulds, Comparison of phylogenetic trees, Math. Biosci. **53** (1981) 131–147.
16. C. Semple and M. Steel, Phylogenetics, Oxford University Press, 2003.