Walks on SPR Neighborhoods

Alan Joseph J. Caceres*† Juan Castillo*† Jinnie Lee* Katherine St. John*

October 22, 2012

Abstract

A nearest-neighbor-interchange (NNI)-walk is a sequence of unrooted phylogenetic trees, $T_1, T_2, \ldots, T_k$ where each consecutive pair of trees differs by a single NNI move. We give tight bounds on the length of the shortest NNI-walks that visit all trees in a subtree-prune-and-regraft (SPR) neighborhood of a given tree. For any unrooted, binary tree, T, on n leaves, the shortest walk takes $\Theta(n^2)$ additional steps more than the number of trees in the SPR neighborhood. This answers Bryant's Second Combinatorial Challenge from the Phylogenetics Challenges List, the Isaac Newton Institute, 2011, and the Penny Ante Problem List, 2009.

Index Terms: Analysis of Algorithms and Problem Complexity; Biology and Genetics; Trees; Graphs and Networks.

1 Introduction

Evolutionary histories, or phylogenies, are essential structures for modern biology [12]. Finding the optimal phylogeny is NP-hard, even when we restrict to tree-like evolution [9, 15]. As such, heuristic searches are used to search the vast set of all trees. There are many search techniques used (see [19] for a survey), but most rely on local search. That is, at each step in the search, the next tree is chosen from the "neighbors" of the current tree. A popular way to define neighbors is in terms of the subtree-prune-and-regraft (SPR) metric (defined in Section 2). Current techniques for computing SPR neighborhoods are computationally intensive. Finding an efficient way to traverse these neighborhoods would have significant impact on the running time of searches for optimal phylogenetic trees. The second "Walks on Trees" challenge of Bryant [5, 17] focuses on efficiently traversing this neighborhood via the nearest-neighbor-interchange (NNI) tranformations (defined in §2). Bryant asks:

> An NNI-walk is a sequence $T_1, T_2, \ldots, T_k$ of unrooted binary phylogenetic trees where each consecutive pair of trees differs by a single NNI.
>
> ii. [Question] Suppose we are given a tree T. What is the shortest NNI-walk that passes through all the trees that lie at most one SPR (subtree-prune-and-regraft) move from T?

Bryant's challenges were posed as part of the New Zealand Phylogenetic Meetings' Penny Ante Problems [5] as well as the Challenges problems from the most recent Phylogenetics Programme at the Isaac Newton Institute [17]. We prove that the shortest walk takes $\Theta(n^2)$ more steps than the theoretical minimum that visits every tree exactly once (that is, a Hamiltonian path). This builds on past work [7] that showed that such a Hamiltonian path was not possible.

*Dept. of Math & Computer Science, Lehman College, City University of New York, Bronx, New York, 10468; Corresponding author: `stjohn@lehman.cuny.edu`; Support provided by NSF Grant #09-20920.

†Supported by an undergraduate research fellowship from the Louis Stokes Alliance for Minority Participation in Research, NSF #07-03449.

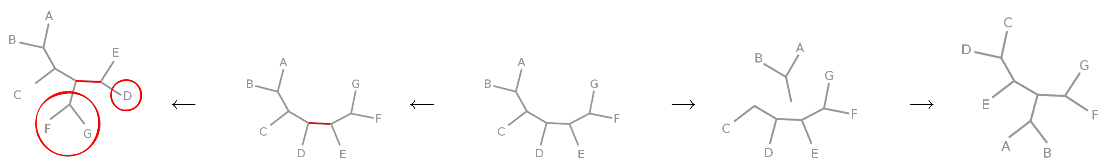

Figure 1: The trees on the left and center differ by a single NNI move. The tree on the right differs by a single SPR move from the center tree.

2 Background

This section includes definitions and results that we use from Allen and Steel [1]. For a more detailed background, see Semple and Steel [16].

Definition 1. *An* **unrooted binary phylogenetic tree** *(or more briefly a tree) is a tree whose leaves (degree 1 vertices) are labelled bijectively by a (species) set S, and such that each non-leaf vertex is unlabelled and has degree three. We let $UB(n)$ denote the set of such trees for $S = \{1, \ldots, n\}$.*

Each internal edge, e, of a tree $T \in UB(n)$ yields a natural bipartition, or **split** of the leaves. We write $A \mid B$ if there is an edge which partitions the leaf set, S, into the two sets A and B. T_A refers to the smallest subtree of T containing leaves only from A, and $E(T)$ refers to the edges of T . A **sibling pair** consists of two leaves that have the same parent. A "caterpillar tree" refers to the unrooted tree with exactly 2 sibling pairs.

The nearest-neighbor-interchange (NNI) distance was introduced independently by DasGupta *et al.* [8] and Li *et al.* [14]. Roughly, an NNI operation swaps two subtrees that are separated by an internal edge.

Definition 2. *Allen and Steel [1]: Any internal edge of an unrooted binary tree has four subtrees attached to it. A* **nearest-neighbor-interchange** *(NNI) move occurs when one subtree on one side of an internal edge is swapped with a subtree on the other side of the edge, as illustrated in Figure 1. The* **NNI distance**, *$d_{NNI}(T_1, T_2)$, between two trees T_1 and T_2 is defined as the minimum number of NNI operations required to change T_1 into T_2.*

The complexity of computing the NNI distance was open for over 25 years and was proven to be NP-complete by Allen and Steel [1]. For a binary tree with n uniquely labeled leaves, there are $n - 3$ internal branches. Thus, there are $2(n - 3)$ NNI rearrangements for any tree.

One of the most popular moves used to search treespace is the subtree-prune-and-regraft (SPR). Roughly, an SPR move prunes a selected subtree and then reattaches it on an edge selected from the remaining tree.

Definition 3. *Allen and Steel [1]: A* **subtree-prune-and-regraft** *(SPR) move on a phylogenetic tree T is defined as cutting any edge and thereby pruning a subtree, t, and then regrafting the subtree by the same cut edge to a new vertex obtained by subdividing a pre-existing edge in T–t. We also apply a forced contraction to maintain the binary property of the resulting tree (see Figure 1). The* **SPR distance**, *$d_{SPR}(T_1, T_2)$, between two trees is the minimal number of SPR moves needed to transform T_1 into T_2.*

For trees, T_1 and T_2, we will say that T_1 has a **unique** SPR move from T_2 if and only if there is exactly one subtree t that can be pruned from T_2 and regrafted to form T_1. Computing the SPR distance is NP-complete [4, 11]. Approximation algorithms for calculating the SPR distance on rooted trees exist [2, 3].

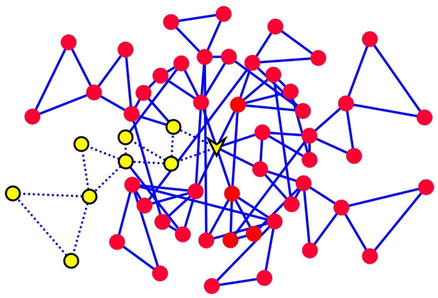

Figure 2: The SPR neighborhood of a 7-leaf caterpillar tree, indicated by the center triangle. There is an edge between two trees if they differ by a single NNI move. The lighter (yellow) nodes show the trees in the orbit that prunes a leaf from one of the sibling pairs.

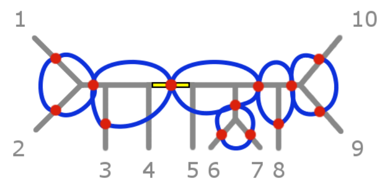

Figure 3: The orbit of the edge, $e = 1\ 2\ 3\ 4 \mid 5\ 6\ 7\ 8\ 9\ 10$, for an unrooted 10-leaf tree. The tree is shown in the background with edge e highlighted. The trees (red dots) are shown relative to the target edge in the initial tree with blue lines indicating trees that differ by an NNI move. The edges adjacent to e yield the initial tree when used as the target edge.

Definition 4. *Let T_0 be an unrooted, binary tree. Define $N_{SPR}(T_0)$ to be the* **SPR neighborhood** *of T_0; namely,*

$$N_{SPR}(T_0) = \{T \mid d_{SPR}(T_0, T) \leq 1\}$$

When the tree is obvious, we will drop the argument and call the neighborhood N_{SPR}.

Definition 5. *Let T_0 be an unrooted, binary tree and $S \subset N_{SPR}(T_0)$. Define $N_{NNI}(S, T_0)$ to be the* **NNI-neighbors** *of S; namely,*

$$N_{NNI}(S, T_0) = \{T \mid \exists T' \in S, d_{NNI}(T, T') \leq 1 \text{ and } d_{SPR}(T_0, T) \leq 1\}$$

Definition 6. *An NNI-***walk** *is a sequence, $T_1, T_2, \ldots, T_k$ of unrooted binary phylogenetic trees where each consecutive pair of trees differs by a single NNI move. An NNI-walk of a set S that visits only elements of S and visits each element at least once and at most k times, it is called a NNI k-walk of S. An NNI 1-walk is also called a* **Hamiltonian path***.*

3 Results

We give tight bounds on the shortest NNI-walk of any SPR neighborhood, improving on previous work [7] that showed that there exist trees for which the shortest NNI-walks are not Hamiltonian. We introduce the new concept of an orbit of an edge, e; roughly, it is all the trees that result from regrafting the pruned edge, e, in either direction. More formally:

Definition 7. *Define for each edge e of the tree T_0, the* **orbit** *of e, O_e, to be all the trees that are one SPR move from T_0 where the edge broken by the SPR move is e.*

As in the definition of the SPR move, we allow the "empty move" of regrafting to an edge adjacent to the pruned edge, yielding the starting tree (see Figure 2). Allen and Steel [1] characterized some properties of the SPR neighborhood:

Theorem 8. *Allen and Steel [1]: Let T_0 be an unrooted phylogenetic tree on n leaves and let N_{SPR} be all trees that are at most a single SPR move from T_0.*

1. *The size of the SPR neighborhood is $|N_{SPR}| = 2(n-3)(2n-7)+1$.*
2. *The trees in $N_{SPR} \setminus \{T_0\}$ that are not a unique SPR move from T_0 are exactly those from the $2n-6$ NNI transformations.*
3. *The number of trees in $N_{SPR} \setminus \{T_0\}$ that can be obtained by a unique SPR move from T_0 is $4(n-3)(n-4)$.*

From this theorem, we observe:

Observation 9. *Let T_0 be an unrooted phylogenetic tree on n leaves:*

1. *Every tree $T \in N_{SPR}(T_0)$ belongs to some orbit O_e, where e is an edge of T_0.*
2. *Each orbit contains T_0.*
3. *Excluding T_0, there are exactly $2n-6$ trees that are included in at least two orbits.*
4. *The number of orbits is $2n-3$.*
5. *The size of each orbit is $2n-7$.*

The structure of the orbits echos that of the underlying tree, since two trees are neighbors in an orbit exactly when the target edges of the moves that created them are adjacent. Formally:

Lemma 10. *Let T_0 be an unrooted phylogenetic tree on n leaves. Let $T_1, T_2 \in N_{SPR}(T_0)$ such that $\exists e \in E(T_0), T_1, T_2 \in O_e$. Let e_i be the target edge of the move that created T_i for $i = 1, 2$ (that is, T_1 is formed by grafting some pruned subtree of T_0 to e_1 and T_2 is the result of grafting a pruned subtree to e_2).*

Then, T_1 and T_2 differ by at most a single NNI move if and only if e_1 and e_2 have a common endpoint in $T_0 \setminus \{e\}$.

Proof. $\Longleftarrow$: Assume that e_1 and e_2 have a common endpoint in $T_0 \setminus \{e\}$. Let M be the set of leaves of the subtree pruned by the SPR move that creates T_1. Without loss of generality, let the split induced by e_1 in T_0 be $ABC \mid DEM$ and the split induced by e_2 in T_0 be $AB \mid CDEM$, where A,B,C,D,E, and M are sets of leaves of subtrees of T_0. Let T_X refer to the subtree with leaves only from the set X.

Since T_M is pruned to create T_1, we have that T_1 contains the splits: $ABCM \mid DE$ and $ABC \mid MDE$. If T_M is also pruned to create T_2, then we have that T_2 contains the splits: $ABM \mid CDE$ and $AB \mid CMDE$. Thus, T_1 and T_2 differ by a single NNI move (swapping T_C and T_M), and the hypothesis holds.

So, assume that T_M is not pruned to create T_2, but instead that e is pruned in the other direction. Let $N = S \setminus M$, where S is the set of leaves of T_0 and T_N is pruned to create T_2. By assumption, e_1 is the target of T_M and thus an edge in T_N, while e_2 is the target of T_N and thus an edge in T_M, contradicting that e_1 and e_2 have a common endpoint in $T_0 \setminus \{e\}$.

$\Longrightarrow$: Assume that T_1 and T_2 differ by a single NNI move. Then, there exists an edge $e' \in E(T_1)$ that when removed (along with its endpoints and adjacent edges), breaks T_1 into 4 distinct subtrees, T_A, T_B, T_C, T_D with leaf sets, A, B, C, D. The split $AB \mid CD$ belongs to T_1 while $BC \mid AD$ belongs to T_2. Since both T_1 and T_2 are in the same orbit, the same edge e is pruned to create both. Let e_i be the target edge of the move that created T_i for $i = 1, 2$. Let M be the set of leaves of the subtree that is pruned to form T_1.

Case 1: T_M is properly contained in one of T_A, T_B, T_C, T_D. Without loss of generality, assume $T_M \subsetneq T_A$, and let $A' = A \setminus M$. Since T_1 is formed from T_0 by moving T_M, we have $T_0|_{A' \cup B \cup C \cup D} = T_1|_{A' \cup B \cup C \cup D}$. T_2 is formed by moving T_M or $T_{A' \cup B \cup C \cup D}$ which implies that $T_0|_{A' \cup B \cup C \cup D} = T_2|_{A' \cup B \cup C \cup D}$. Thus, $A'B \mid CD$ and $BC \mid A'D$ both belong to T_0 which is a contradiction.

Case 2: T_M properly contains one of T_A, T_B, T_C, T_D. Without loss of generality, assume $T_M \supsetneq T_A$. If $M \subset A \cup B$, then let $B' = (A \cup B) \setminus M$. Since T_2 is formed by only moving T_M or $T_{B' \cup C \cup D}$ and $BC \mid AD$ belongs to T_2, either $T_M = T_{A \cup B'}$ is a subtree of $T_{B \cup C}$ or a subtree of $T_{A \cup D}$ which is a contradiction. The subcase where $M \supseteq A \cup B$ follows by similar argument.

Case 3: T_M is one of T_A, T_B, T_C, T_D. Without loss of generality, assume $T_M = T_A$. Since T_1 and T_2 are in the same orbit, we must have that T_A or $T_{B \cup C \cup D}$ is the subtree pruned to form T_2. Since the split $BC \mid AD$ belongs to T_2, T_A is pruned to form T_2. Further, since $AB \mid CD$ belongs to T_1 while $BC \mid AD$ belongs to T_2, e_1 corresponds to the split $B \mid CD$ in $T_{B \cup C \cup D}$ while e_2 corresponds to the split $BC \mid D$ in $T_{B \cup C \cup D}$. So, e_1 and e_2 share a common endpoint, namely the intersection point of T_B,T_C, and T_D. □

The above lemma shows that neighboring trees in an orbit correspond to adjacent target edges, implying that the structure of the orbit echoes the tree structure (see Figure 2). We can further characterize the adjacent trees in each orbit:

Corollary 11. *Let T_0 be an unrooted, binary tree on n leaves. Let $e \in E(T_0)$ and $T \in O_e$. Let $N = \{T' \in O_e \mid T \text{ and } T' \text{ differ by an NNI-move}\}$. Then $|N|$ is 2 or 4. If $T \neq T_0$, then there exists $T_1, T_2 \in N$ such that $d_{NNI}(T_0, T_1) + 1 = d_{NNI}(T_0, T_2) = d_{NNI}(T_0, T)$. Further, if $|N| = 4$, there exists $T_3, T_4 \in N$ such that $d_{NNI}(T_0, T) + 1 = d_{NNI}(T_0, T_3) = d_{NNI}(T_0, T_4)$.*

Proof. By Lemma 10, the trees that differ by a single NNI move from T are those whose target edges are adjacent to the edge e. Since T is binary, the number of such adjacent trees is either 2 or 4. Assume that T corresponds to a target edge that is part of a sibling pair. Then, let T_1 be the tree corresponding to the target edge that is the other part of the sibling pair, and T_2 be the tree corresponding to the only edge adjacent to the sibling pair. Then, $d_{NNI}(T_0, T_1)+1 = d_{NNI}(T_0, T_2) = d_{NNI}(T_0, T)$.

Assume that T corresponds to a target edge, e_T, that is not part of a sibling pair. By Lemma 10, e has 4 adjacent edges. Let e_1 refer to the unique edge of T_0 on the path from e to e_T and e_2 to the edge that shares the common endpoint of e_1 and e_T. Let e_3 and e_4 be the edges that share the other endpoint of e_T. Let T_i be the tree that corresponds to the target edge e_i for $i = 1, \ldots, 4$. By the underlying tree structure of T, we have the desired properties. □

We can immediately give an upper bound on the length of an NNI-walk of an SPR neighborhood. The underlying idea is to traverse each orbit separately, and then link these paths to form a traversal of the entire SPR neighborhood:

Lemma 12. *For every unrooted, binary tree, T_0, on n leaves, every NNI-walk of its SPR neighborhood, $N_{SPR}(T_0)$, has length at most $|N_{SPR}(T_0)| + O(n^2)$.*

Proof. We will break the NNI-walk of the SPR neighborhood into NNI-walks of the orbit of each edge in T_0.

It suffices to show that there is a 2-walk of each orbit O_e for $e \in E(T_0)$. Each tree, $T \in O_e$, is created by pruning the edge e in T_0 and regrafting the pruned subtree to another edge in T_0 (see Figures 2 and 2). Every tree in the orbit corresponds to an edge in T_0 (namely, the target edge), and the trees in the orbit are connected exactly when their target edges share an endpoint in T_0 by Lemma 10. Thus, the orbit can be traversed by at most $2(2n-7)$ steps by starting at T_0 and following a depth-first-search of the tree (each tree in the orbit is visited at most once on the way "down" the search and once on the way "up" the search).

Since each orbit contains the initial tree T_0, we can glue together the walks of the orbits to make a walk of the entire space. Since each orbit contains at most $2n-7$ trees, the 2-walk of each of the $2n-3$ orbits yields a walk where the number of steps is bounded by $2(2n-7)(2n-3) = |N_{SPR}(T_0)| + O(n^2)$. □

To show the lower bound takes more work and relies on the fact that the orbits in an SPR neighborhood are, surprisingly, mostly disjoint:

Lemma 13. *Let T_0, T_1, T_2 be unrooted binary trees with $T_1, T_2 \in N_{SPR}(T_0)$, $d_{NNI}(T_1, T_2) \leq 1$, and T_1 and T_2 are in different orbits. Then $d_{NNI}(T_0, T_1), d_{NNI}(T_0, T_2) \leq 2$.*

Proof. Assume that there exists $e_1, e_2 \in E(T_0)$, $T_1 \in O_{e_1}$, $T_1 \notin O_{e_2}$, $T_2 \notin O_{e_1}$ and $T_2 \in O_{e_2}$. Let M_1 be the set of leaves of the subtree pruned with e_1 from T_0 to create tree T_1. Since T_1 and T_2 are a single NNI move apart, by definition, there exists a split in T_1, $AB \mid CD$ that is rearranged in T_2: $BC \mid AD$. We will argue, by cases, that both T_1 and T_2 are within 2 NNI moves of T_0. Without loss of generality, we will assume that $M_1 \cap A \neq \emptyset$.

Case 1: $M_1 \subsetneq A$. Then, let $A' = A \setminus M_1$. We have that T_1 contains the split $A'M_1B|CD$, and T_2 contains the split $BC \mid A'M_1D$. Since

T_1 is only one SPR move from T_0, the structure of the 2 trees is identical without M_1; that is, $T_1|_{A' \cup B \cup C \cup D} = T_0|_{A' \cup B \cup C \cup D}$, and T_0 includes an edge corresponding to the split $A'B|CD$. Since T_2 does not contain such an edge, the move that creates it must prune one of T_{M_1}, $T_{A'}$, T_B, T_C, or T_D. Pruning T_{M_1} is not possible since T_1 and T_2 are in different orbits. Pruning $T_{A'}$ is only possible if T_0 contains the split $M_1 D|A'BC$. T_0 can be transformed into T_1 by NNI moves that interchange the neighbor subtrees T_{M_1} and T_C, followed by T_{M_1} and T_B. We can similarly transform T_0 into T_2 and T_1 into T_2 with 2 NNI moves. Thus, $d_{NNI}(T_0, T_1), d_{NNI}(T_0, T_2) \leq 2$ and the lemma holds.

Pruning T_B to create T_2 implies that T_0 contains the split $A'BM_1|CD$ and either $T_0 = T_2$ or $d_{NNI}(T_0, T_2) = 2$, implying $d_{NNI}(T_0, T_1) = 2$. Lastly, pruning T_C, or T_D is only possible if $T_0 = T_1$, in which case, $d_{NNI}(T_0, T_1) = 0, d_{NNI}(T_0, T_2) = 1$.

Case 2: $M_1 = A$. We have that T_1 contains the split $M_1 B|CD$ and T_2 contains the split $BC \mid M_1 D$. We have three possibilities for T_0; namely, it could contain one of the following three splits: $M_1 B \mid CD$, $BC \mid M_1 D$, or $BD \mid M_1 C$. We note that these are the three possible NNI rearrangements for this edge, so, we have $d_{NNI}(T_0, T_1), d_{NNI}(T_0, T_2) \leq 1$ and the lemma holds.

Case 3: $M_1 \supsetneq A$. If $M_1 \cap B = \emptyset$, then $M_1 = A \cup C \cup D$ (else pruning T_{M_1} would not yield a connected tree). The argument is similar to Case 2.

If $M_1 \cap B \neq \emptyset$, then $B \subseteq M_1$. If $M_1 = A \cup B$, then the target edges in T_1 and T_2 must separate C and D, and are identical. Similarly, if $A \cup B \subsetneq M_1$, M_1 must contain all of C or all of D, and the target edges in T_1 and T_2 must preserve the rooting of the remaining subtree and are identical. Thus, $d_{NNI}(T_0, T_1), d_{NNI}(T_0, T_2) = 0$. □

We say that $U \subseteq O_e$ is **connected** if for any two trees $T_1, T_2 \in U$ there exists $U_1, \ldots, U_k \in U$ such that $U_1 = T_1$, $U_k = T_2$, and $U_1, \ldots, U_k$ is an NNI-walk. We call any NNI-walk that begins and ends at the same tree an **NNI-circuit**.

Lemma 14. *Let T_0 be an unrooted binary tree, $e \in E(T_0)$, and O_e its orbit. Let $U \subseteq O_e$ be a connected set consisting of trees more than 2 NNI moves from T_0. Then any NNI-circuit of U takes at least $\frac{3}{2}(|U| - 1)$ steps.*

Proof. By induction on the size of $|U|$.

For $|U| = 1$: Then any circuit takes $1 \geq \frac{3}{2}(|U| - 1) = 0$ steps.

For $|U| > 1$, choose $x \in U$ closest to T_0. By Lemma 10, two trees are neighbors in O_e (that is, are a single NNI move apart) if and only if there target edges have a common endpoint in the initial tree T_0. Since T_0 is binary, each tree in O_e can have at most 4 possible neighbors.

If x has one neighbor in U, then a circuit of U must traverse the same edge from x to its neighbor twice, and the number of steps needed is at least two more than the number of steps needed for the smaller set $|U| - \{x\}$. By inductive hypothesis, this smaller set takes at least $\frac{3}{2}(|U - \{x\}| - 1)$ steps. So, the number of steps for U is:

$$\frac{3}{2}(|U - \{x\}| - 1) + 2 \geq \frac{3}{2}(|U| - 1)$$

If x has two neighbors, x_1 and x_2 in U, then $U - \{x, x_1, x_2\}$ is disconnected in U by Corollary 11. Let U_1 and U_2 be the components of $U - \{x, x_1, x_2\}$ such that x_1 is adjacent to some element of U_1 and x_2 is adjacent to some element of U_2. If $d_{NNI}(x_1, x_2) = 1$, then it takes 3 steps to visit x in a circuit of x, U_1, and U_2. If they are not connected, it takes 4 steps. Thus, by inductive hypothesis, the number of steps needed is:

$$\frac{3}{2}(|U_1| - 1) + \frac{3}{2}(|U_2| - 1) + 3 \geq \frac{3}{2}(|U| - 1)$$

If x has 3 neighbors in U, then by similar argument, we have the lower bound. If x has 4 neighbors in U, then it is not the closest element of U to T_0, giving a contradiction. □

From the last two lemmas, we have that the orbits are mostly isolated; the only trees that have neighbors from outside their orbits are within 2 steps of T_0. An NNI-walk of these isolated regions takes many extra steps. This yields our lower bound:

Lemma 15. *Every NNI-walk of $N_{SPR}(T_0)$ has length $|N_{SPR}(T_0)| + \Omega(n^2)$.*

Proof. Let $e \in E(T_0)$ and O_e its orbit. By Lemma 13, every orbit, O_e, has $\Omega(n)$ trees that have no neighbors in other orbits. It follows from Lemma 10, these trees are in at most two connected sets. By the Pigeonhole Principle, one set has at least $\Omega(n)$ trees. By Lemma 14, it takes $\Omega(n)$ steps to visit the larger connected set. By Theorem 8, there are $2n-3$ orbits, and any NNI-walk of N_{SPR} must take $\geq (2n-3)\Omega(n) = \Omega(n^2)$ extra steps. □

The above lemmas immediately show that $\Theta(n^2)$ extra steps are needed to traverse the neighborhood:

Theorem 16. *For any unrooted binary tree, T_0, on n leaves, an NNI-walk of $N_{SPR}(T_0)$ takes $|N_{SPR}(T_0)| + \Theta(n^2)$ steps.*

4 Discussion

Finding optimal phylogenetic trees is a computationally expensive process given the hardness of the preferred optimality criteria [9, 15]. Searches of treespace often step from tree to tree, looking for the optimal tree. A popular way to determine the next tree is by examining the SPR neighborhood of the current tree (a standard option in many popular software packages: MrBayes [13], PAUP [18], and TNT [10]). Unlike NNI moves which make only local rearrangements to a tree, SPR moves can move large sections of trees far away from their original location. As such, NNI neighborhoods are efficient to calculate, while the calculation of an SPR neighborhood can be quite time-consuming. Bryant's Second Challenge asks how efficiently can an SPR neighborhood be traversed by NNI moves. We show that any NNI-walk will need extra steps proportional to the size of the SPR neighborhood ($\Theta(n^2)$), implying that an NNI-walk does not provide an efficient alternative. Bryant [6] suggests that NNI-walks might provide an efficient way to traverse another popular tree neighborhood: tree-bisection-reconnection (TBR).

5 Acknowledgments

We thank the anonymous reviewers for their thoughtful and helpful comments that greatly improved this paper. For spirited discussions, we thank the members of the Treespace Working Group at CUNY: Ann Marie Alcocer, Kadian Brown, Samantha Daley, John DeJesus, Eric Ford, Michael Hintze, Daniele Ippolito, Joan Marc, Oliver Mendez, Diquan Moore, and Rachel Spratt.

This work was supported by grants from the US National Science Foundation programs in mathematical biology and computational mathematics (NSF #09-20920) and the New York City Louis Stokes Alliance for Minority Participation in Research (NSF #07-03449). Some of this work was completed at the Isaac Newton Institute at Cambridge University, and we gratefully acknowledge travel funding to the last author provided by the institute and the Lehman College-IBM Research travel fund.

References

[1] B. Allen and M. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Ann. of Combinatorics*, 5:1–13, 2001.

[2] M.L. Bonet, K. St. John, R. Mahindru, and N. Amenta. Approximating subtree distances between phylogenies. *J. of Computational Biology*, 13(8):1419–1434 (electronic), 2006.

[3] M. Bordewich, C. McCartin, and C. Semple. A 3-approximation algorithm for the subtree distance between phylogenies. *J. of Discrete Algorithms*, 6(3):458–471, 2008.

[4] M. Bordewich and C. Semple. On the computational complexity of the rooted subtree prune

and regraft distance. *Ann. of Combintorics*, 8:409–423, 2004.

[5] D. Bryant. Annual New Zealand Phylogenetics Meeting (Kaikoura 2009) Penny Ante prize problems: A mathematical challenge. http://www.math.canterbury.ac.nz/bio/events/kaikoura09/penny.shtml, 2009.

[6] D. Bryant, 2012. Personal communication.

[7] A.J.J. Caceres, S. Daley, J. DeJesus, M. Hintze, D. Moore, and K. St. John. Walks in phylogenetic treespace. *Information Processing Letters*, 111:600–604, 2011.

[8] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang. On computing the nearest neighbor interchange distance. In D.Z. Du, P.M. Pardalos, and J. Wang, editors, *Proceedings of the DIMACS Workshop on Discrete Problems with Medical Applications*, volume 55, pages 125–143. Am. Math. Soc., 2000.

[9] L.R. Foulds and R.L. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Math.*, 3(1):43–49, 1982.

[10] P.A. Goloboff, J.S. Farris, and K.C. Nixon. Tnt, a free program for phylogenetic analysis. *Cladistics*, 24:774–786, 2008.

[11] G. Hickey, F. Dehne, A. Rau-Chaplin, and C. Blouin. SPR distance computation for unrooted trees. *Evolutionary Bioinformatics*, 4:17–27, 2008.

[12] D.M. Hillis, B.K. Mable, and C. Moritz. *Molecular Systematics*. Sinauer Assoc., Sunderland, Mass., 1996.

[13] J.P. Huelsenbeck and F. Ronquist. MrBayes: Bayesian inference of phylogeny, 2001.

[14] M. Li, J. Tromp, and L. Zhang. Some notes on the nearest neighbour interchange distance. In *COCOON '96: Proceedings of the Second Annual International Conference on Computing and Combinatorics*, pages 343–351, London, UK, 1996. Springer-Verlag.

[15] S. Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans. on Computational Biology and Bioinformatics*, 3(1):92–94, 2006.

[16] C. Semple and M. Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.

[17] M. Steel. Challenges and Conjectures: Isaac Newton Institute Phylogenetics Program 2011. http://www.newton.ac.uk/programmes/PLG/phylogenetics_challenges.pdf.

[18] D.L. Swofford. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts, 2002.

[19] S. Whelan. New approaches to phylogenetic tree search and their application to large numbers of protein alignments. *Systematic Biology*, 56(5):727–740, 2007.

Brief Author Biographies

Alan Joseph J. Caceres is an intern at the IBM T.J. Watson Research Center, Hawthorne, New York. His research interests include computational biology, data mining, and ubiquitous computing. Caceres received a bachelors degree in computer science from Lehman College, City University of New York (CUNY) in May 2012. Caceres plans to pursue a doctoral degree in computer science at the University of Notre Dame. He is a member of the ACM. Contact him at `alan.j.caceres@gmail.com`.

Juan Castillo is a masters candidate in computer science at Lehman College, CUNY. His research interests include computational biology and analysis of algorithms. He completed his undergraduate computer science major at Lehman College, CUNY in May 2012. Contact him at `jcastillo0525@hotmail.com`.

Jinnie Lee is an adjunct lecturer at Lehman College, CUNY. Lee received a bachelors degree with a double major in mathematics and art from Lehman College and is currently pursuing masters courses at City College of New York, CUNY. Her research interests include discrete mathematics and computational biology. Contact her at `parangnarae@gmail.com`.

Katherine St. John is a professor of mathematics and computer science at Lehman College, CUNY and holds appointments to the doctoral faculties of anthropology and computer science at the Graduate Center of CUNY, as well as the invertebrate zoology and paleontology divisions of the American Museum of Natural History. St. John received her doctoral degree from UCLA. Her research interests include computational biology, random structures, and algorithms. She is a member of ACM, AMS, and SIAM. Contact her at `stjohn@lehman.cuny.edu`.