Hamiltonian Walks of Phylogenetic Treespaces

Kevaughn Gordon, Eric Ford, and Katherine St. John

Abstract—We answer Bryant's Combinatorial Challenge on minimal walks of phylogenetic treespace under the nearest neighbor interchange (NNI) metric. We show that the shortest path through the NNI-treespace of n-leaf trees is Hamiltonian for all n. That is, there is a minimal path that visits all binary trees exactly once, under NNI moves.

Index Terms—Phylogenetic Tree Construction, Graph Theory, Analysis of Algorithms.

1 Introduction

Phylogenetic trees depict the evolutionary relationships within a set of taxa, represented as leaf labels [1]. The trees may be rooted—in which case they illustrate the ancestry of the taxa—or unrooted. In this paper, we look at unrooted phylogenies.

Finding the tree that best fits the data, where the data is a set of taxa and ordered characters, is a central goal of evolutionary biology. However, the number of possible trees grows as an exponential function of the number of taxa, and finding the optimal tree under the criteria most used by biologists is NP-hard [2], [3]. Due to the size of the search space, exhaustive search is often not possible, so heuristic search is often used to discover the best tree. To systematically traverse the space, it is necessary that it be arranged in some manner. A common arrangement is to link trees that are a single move apart under some tree rearrangement operation; the resulting graph is often called a treespace.

Focusing on trees that differ by a single nearest neighbor interchange (NNI) move, David Bryant asked for the length of the shortest walk that visits all trees in NNI treespace [4]. It is known that using metrics that yield more neighbors than NNI (namely subtree prune and regraft (SPR) and tree bisection and reconnection (TBR)) have the shortest walk possible: a Hamiltonian path [5]. Previous to our current work, the best known NNI-walk of all binary trees visited every tree at most twice [5].

We show that, for all n, a Hamiltonian path exists on the space of all binary trees on n leaves under the NNI metric, settling Bryant's challenge. We follow the strategy of previous work in expanding Hamiltonian paths on n-leaf trees to the space of all binary trees on $(n+1)$ leaves [5]. This idea does not work directly for NNI-walks but can be employed with a subtle twist. Instead of developing walks on the expansion of a single n-leaf tree, we look at all trees that can be created from subsequent triples of n-leaf trees on the Hamiltonian path for the smaller space. Using the Hamiltonian path of the smaller space as a "backbone," we can then "glue" together the unions of the expansions to form a Hamiltonian path on the $(n+1)$-leaf trees. Since every NNI move can be simulated by an SPR or TBR move, this paper provides an alternative proof for the existence of Hamiltonian paths for the SPR and TBR treespaces.

- *K. Gordon is with the Department of Math & Computer Science, Lehman College, City University of New York, Bronx, New York, 10468;* *Supported by an undergraduate research fellowship from the Louis Stokes Alliance for Minority Participation in Research, NSF #07-03449.*
- *E. Ford is with the Department of Computer Science, the Graduate Center, City University of New York, New York, New York 10016.*
- *K. St. John is with the Department of Math & Computer Science, Lehman College and the Department of Computer Science, the Graduate Center, City University of New York;* *Corresponding author:* `stjohn@lehman.cuny.edu`; *Support provided by NSF Grant #09-20920.*

2 Background

We briefly define binary phylogenetic trees and the associated terms used in this paper. For a more detailed treatment, see Semple and Steel [1].

Phylogenetic trees depict evolutionary relationships between taxa placed at the leaves. Trees can be rooted, in which case they illustrate the ancestry of the taxa, or unrooted. We look at unrooted binary phylogenetic trees (hereafter referred to as trees). Formally, as defined by Robinson:

Definition 1: [6]: A (binary) *phylogenetic tree* is a graph G on collection of labelled nodes L (the taxa) and unlabelled interior vertices. The labelled nodes form the *leaves* of the tree and, therefore, have valency one, and each interior vertex has valency three.

We will use a well-known fact about the number of unrooted binary trees:

Lemma 1: [6]: For n taxa, there are $(2n-5)!! = (2n-5)(2n-3)\cdots 5\cdot 3\cdot 1$ possible unrooted trees.

Note that for all $n \geq 4$, $(2n-5)!!$ is divisible by three. We will use this characteristic of treespace to partition paths of n-leaf trees into triples. We will also examine pairs of leaves:

Definition 2: A *sibling pair*, or *cherry*, is a pair of leaves whose incident edges share a common vertex.

Difference measures on trees induce metric spaces on the set of n-leaf trees. We will focus on metrics that measure shape differences between trees (and ignore

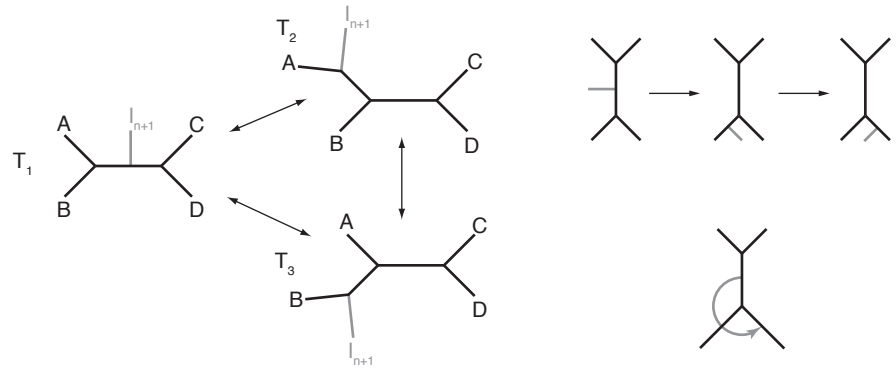

Fig. 1. The left side shows NNI transformations. To transform T_1 into T_2 or vice-versa, interchange the subtree B with the subtree containing the leaf l_{n+1}. To transform T_2 into T_3 or vice-versa, either interchange the subtree containing the leaf l_{n+1} with the subtree containing C and D or interchange the subtree containing A with the subtree containing B. T_1, T_2 and T_3 are all neighbors in NNI treespace. The right side illustrates how a path in NNI treespace will be represented in this paper. Note that the top series of moves is equivalent to moving from T_1 to T_2 to T_3, as in the left side of the figure. The bottom right represents the same moves with the curved edge representing the path of the $(n+1)^{st}$ leaf, l_{n+1}.

differences in the lengths of edges or branches). These metrics induce a discrete space that can be represented by a graph:

Definition 3: Given a set of trees $\mathcal{T} = \{T_1, T_2, \ldots, T_k\}$ with n leaves labelled by S, $\mathbf{G} = (\mathcal{T}, E)$, or *treespace*, is the graph $\mathbf{G}$ with vertices labelled by $\mathcal{T}$ and the edges E connecting vertices that are "neighbors"—distance one apart under a given metric.

Bryant's challenge focuses on nearest neighbor interchange (NNI). Other popular metrics include subtree prune regraft (SPR) and tree bisection and reconnection (TBR) [7].

Definition 4: [7]: A *nearest neighbor interchange (NNI)* swaps any two subtrees connected to opposite ends of an edge (see Figure 1). The NNI distance (d_{NNI}) between two trees is the minimum number of NNI operations that transforms one of the trees into the other.

We note that the NNI operation is symmetric in that any NNI tree rearrangement operation can be reversed. These moves define neighborhoods on the space:

Definition 5: Let d be a discrete tree metric. The set of all trees T_m where $d(T, T_m) \leq 1$ is the *1-neighborhood* (or simply *neighborhood*) of T.

An n-leaf tree has $n-3$ internal edges. Using an NNI move, a new tree can be formed by swapping one of the four subtrees on opposite sides of the internal edge (see Figure 1). Only two of these swaps will produce new trees, and thus every n-leaf tree has $2(n-3)$ neighbors in NNI treespace.

The challenge on which we focus is phrased in terms of "walks," we will use this term interchangably with the common term from graph theory: "paths."

Definition 6: [4]: An *NNI-walk* is a sequence $T_1, T_2, \ldots, T_k$ of unrooted binary phylogenetic trees where each consecutive pair of trees differ by a single NNI move.

Definition 7: [8]: A *Hamiltonian path* in a graph is a simple path that visits every node exactly once. This path can be represented as an ordered set of nodes, $v_1, v_2, \ldots, v_n$, where v_i is connected to v_{i+1} by an edge.

Determining whether an arbitrary graph has a Hamiltonian path is NP-hard [8]. However, for many classes of graphs (for example, complete graphs), Hamiltonicity can be determined easily in polynomial time.

3 Main Results

We prove that a Hamiltonian path exists through the set of n-leaf trees under the NNI metric, for all n. The proof constructs a Hamiltonian path of the $(n+1)$-leaf treespace from a Hamiltonian path of the n-leaf treespace (see Figure 2). This is done by taking subsequent triples from the path of the n-leaf treespace and contructing a path through all $(n+1)$-leaf trees that be created from those three trees (formally defined as the "expansion" of trees, below). Since every $(n+1)$-leaf tree belongs to exactly one such expansion of a triple, linking the paths of the expansions yields a path that visits every $(n+1)$-leaf tree exactly once.

Definition 8: Let T be an n-leaf tree and e an edge of T. The *expansion of an edge*, e, is the $(n+1)$-leaf tree, $T(e)$, generated when a new leaf is added to that edge. Let the *expansion* of an n-leaf tree, T, be the set of $(n+1)$-leaf trees that can be generated from expanding all edges of T (see Figure 3).

If two trees differ by only a single NNI move, then the edges of the trees are identical, except for a single edge, that we call the "edge of difference". Formally:

Definition 9: Consider two trees, T_1 and T_2, that differ by one NNI move. Let e_d, the *edge of difference*, be the single edge in the symmetric difference between the set of edges of T_1 and the set of edges of T_2.

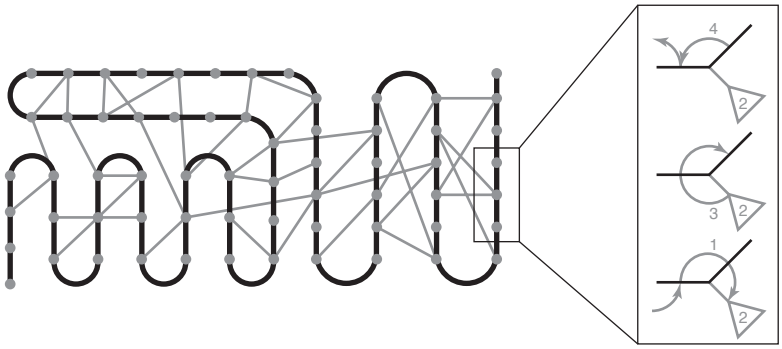

Fig. 2. The proof of the main result constructs Hamiltonian paths of the $(n+1)$-leaf treespace from Hamiltonian paths of the n-leaf treespace. This is done by using a Hamiltonian path of the smaller space (bold path) as a "backbone" for a path of the larger space. For every triple of trees in the path of the smaller space, a path is created through its expansion. An example of a path through a triple is boxed. The various trees in the expansions of the triple are visited in labeled order. These paths are linked to form the Hamiltonian path of the $(n+1)$-leaf treespace.

We note that the size of the expansion of an n-leaf tree is independent of the given tree topology and depends only on the number of internal edges. Likewise, in a binary tree, the number of internal edges is a function of the number of leaves. For a given tree T with n leaves, there are $2n-3$ trees with $n+1$ leaves contained in the expansion of T. We first prove several useful lemmas about expansions of edges:

Lemma 2: Let T be an unrooted binary tree, and let e_1 and e_2 be adjacent edges on T. $T(e_1)$ and $T(e_2)$ differ by one NNI move.

Proof: Let e_1 and e_2 be adjacent edges in an n-leaf tree. Let S be the subtree whose root edge is incident with e_1 and e_2. The addition of a new leaf, l_{n+1}, creates two new edges: e_3, which connects the new leaf node to the tree and e_4, which separates S and l_{n+1}. In $T(e_1)$, l_{n+1} is between e_1 and e_4 and S is between e_4 and e_2. The opposite occurs in the $T(e_2)$. In that case, S is between e_1 and e_4 and l_{n+1} is between e_4 and e_2. That is, $T(e_1)$ and $T(e_2)$ have the same tree topology save for the arrangement around e_4. Since the new taxon and the rooted subtree are on opposite sides of e_4, an internal branch, swapping them costs only one NNI move. Therefore, $T(e_1)$ and $T(e_2)$ differ by one NNI move. □

Lemma 3: Let T_1 and T_2 be two unrooted binary trees where T_1 and T_2 differ by one NNI move. Let e be an edge that is not the edge of difference, e_d. $T_1(e)$ and $T_2(e)$ differ by one NNI move.

Proof: Let A, B, C, and D be the four subtrees whose root edges are incident to e_d, between T_1 and T_2. By definition, the arrangement of the four subtrees in T_1 differ from their arrangement in T_2. Assume, without loss of generality, that e is an edge of A, and denote by A' the subtree created by the addition of the new leaf to e in A. Note that A' is identical in $T_1(e)$ and $T_2(e)$. The arrangement of A', B, C, and D around e_d is the only difference between the two trees. Therefore, $T_1(e)$ and $T_2(e)$ differ by one NNI move. □

When focusing on a triple of consecutive trees on a Hamiltonian path of n-space, there are many subtrees which are identical across all three trees. The next lemma shows how the expansions of these subtrees can be traversed such that each node in the expansions is visited only once (see Figure 4):

Lemma 4: Let T_1, T_2, and T_3 be three unrooted binary n-leaf trees where T_1 and T_2 are NNI neighbors and where T_2 and T_3 are NNI neighbors. Let S_i be some rooted subtree on T_i where $i = 1, 2,$ or 3. If $S_1 = S_2 = S_3$, the union of the expansions of the edges in S_1, S_2, and S_3 has a Hamiltonian path such that the walk starts on $T_i(p_i)$, where p_i is the root edge of S_i, and ends on $T_j(p_j)$, where p_j is the root edge of S_j and $i \neq j$.

Proof: We proceed by induction on the size of the subtree.

Base Case: The subtree has two leaves and three edges: p_i, which connects the root node to the internal node; l_i, which connects the internal node to a leaf node; and r_i, which connects the internal node to the other leaf node. All three edges are adjacent. By Lemma 2, the expansions of these edges are NNI neighbors. Since T_1 and T_2 are one NNI move apart, $T_1(p_1)$ and $T_2(p_2)$ are NNI neighbors by Lemma 3. The rest of the edges follow suit. That is, $T_z(y_z)$ and $T_{z+1}(y_{z+1})$ are NNI neighbors where $y \in \{p, l, r\}$ and $z \in \{1, 2\}$.

We can enumerate the possible walks that start on $T_i(p_i)$ and end at $T_j(p_j)$ where $i \neq j$ (see Figure 4). We identify the path through the $(n+1)$-leaf trees by the edge that is expanded:

- $p_1 \to l_1 \to r_1 \to r_2 \to r_3 \to l_3 \to l_2 \to p_2 \to p_3$,
- $p_1 \to l_1 \to r_1 \to r_2 \to r_3 \to p_3 \to l_3 \to l_2 \to p_2$,
- $p_2 \to r_2 \to r_1 \to p_1 \to l_1 \to l_2 \to l_3 \to r_3 \to p_3$.

We note that since the edges are not directed, each of the above three paths could be traversed in reverse. Thus, we have a Hamiltonian path of the expansions of the edges of the subtrees that begins on $T_i(p_i)$ and ends on

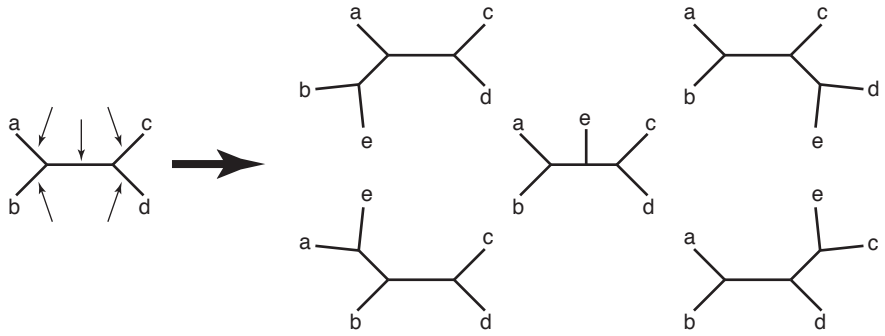

Fig. 3. Expansion of an unrooted tree on n leaves to $(n+1)$-leaf trees. When $n = 4$, there are five possible edges to which to attach a new leaf, resulting in five 5-leaf trees in the expansion of the initial tree.

$T_j(p_j)$ where $i \neq j$.

Inductive Step: Assume that the subtree, S_i, has three or more leaves and at least five edges: p_i, which connects the root node to an internal node; and edges l_i and r_i which are incident with p_i. By Lemma 2, the expansions of these edges are NNI neighbors. Further, since T_1 and T_2 are one NNI move apart, $T_1(p_1)$ and $T_2(p_2)$ are NNI neighbors by Lemma 3. The rest of the edges follow suit. That is, $T_z(y_z)$ and $T_{z+1}(y_{z+1})$ are NNI neighbors where $y \in \{p, l, r\}$ and $z \in \{1, 2\}$.

We show that a Hamiltonian path can start on $T_i(p_i)$ and end on $T_j(p_j)$ where $i \neq j$.

Without loss of generality, assume that l_i is the root edge of the inner subtree, C_i. Let $T(C_1, C_2, C_3)$ be the union of the expansions of all the edges in C_1, C_2, and C_3 except for the expansions of two of the root edges, l_i and l_j, whose visit we explicitly show. By the inductive hypothesis, a Hamiltonian path can start on $T_i(l_i)$, and end on $T_j(l_j)$ where $i \neq j$.

Case I: S_i is a complex rooted subtree with a leaf attached to the root. r_i connects the first internal node to a leaf (see Figure 4). The following are paths of the union starting at $T_i(p_i)$ and ending at $T_j(p_j)$, $i \neq j$:

- $p_1 \to r_1 \to l_1 \to T(C_1, C_2, C_3) \to l_3 \to r_3 \to r_2 \to p_2 \to p_3$,
- $p_1 \to r_1 \to l_1 \to T(C_1, C_2, C_3) \to l_2 \to r_2 \to r_3 \to p_3 \to p_2$,
- $p_2 \to p_1 \to r_1 \to r_2 \to l_2 \to T(C_1, C_2, C_3) \to l_3 \to r_3 \to p_3$.

Case II: S_i is a complex rooted subtree with another complex rooted subtree attached to its root. r_i connects the first internal node to another complex subtree, D_i. The following are paths of the union starting at $T_i(p_i)$ and ending at $T_j(p_j)$, $i \neq j$:

- $p_1 \to r_1 \to T(D_1, D_2, D_3) \to r_3 \to l_3 \to T(C_1, C_2, C_3) \to l_2 \to p_2 \to p_3$,
- $p_1 \to r_1 \to T(D_1, D_2, D_3) \to r_2 \to l_2 \to T(C_1, C_2, C_3) \to l_3 \to p_3 \to p_2$,
- $p_2 \to r_2 \to T(D_1, D_2, D_3) \to r_1 \to p_1 \to l_1 \to T(C_1, C_2, C_3) \to l_3 \to p_3$.

This completes the proof. □

Lemma 4 focuses on the union of the expansions of edges in rooted subtrees that are identical across a triple of subsequent trees in a path of the n-leaf treespace, giving multiple paths that can traverse that union from different starting and stopping points. The following lemma shows how these paths can be "glued" together to form a path for the unions of the expansions of the complete trees. The difficulty in the proof is the "lining up" of the endpoints of paths to create a single longer path. Since Lemma 4 provides paths from any root edge of identical subtrees, it suffices to show how to traverse the edges of difference in the expansions.

Lemma 5: Let T_1, T_2, and T_3 be three unrooted binary trees where T_1 and T_2 are NNI neighbors and where T_2 and T_3 are NNI neighbors. For any edge e of T_1, there exists a Hamiltonian path of the union of the expansions of T_1, T_2, and T_3 starting at $T_1(e)$.

Proof: Let T_1, T_2, and T_3 be three unrooted binary trees where T_1 and T_2 are NNI neighbors and where T_2 and T_3 are NNI neighbors. Let $e_{d_{12}}$ be the edge of difference between T_1 and T_2, and let $e_{d_{23}}$ be the edge of difference between T_2 and T_3. Let the expansion of a subtree, S, in an n-leaf tree, T, be the union of the expansions of the edges in S.

Let e be an edge in T_1. We will construct a walk that traverses, exactly once, every tree in the union of the expansions of T_1, T_2, and T_3. The location of e, whose expansion, $T_1(e)$, is the starting point, determines our strategy for building the Hamiltonian path. Denote by A and B the subtrees that result from removing e (but not its endpoints) from T_1. If one of the subtrees, say A, is identical across T_1, T_2, and T_3, then neither $e_{d_{12}}$ nor $e_{d_{23}}$ is in A, and by Lemma 4, there is a Hamiltonian path across the union of the expansions of A_1, A_2, and A_3, beginning at e in T_1 and ending at e in T_i where $i = 2$ or 3. So, assume that neither A nor B is identical across all trees. We proceed by cases on the relative locations of the edges of difference, $e_{d_{12}}$ or $e_{d_{23}}$, to e:

Case I: Assume that $e_{d_{12}}$ is on the path between e and $e_{d_{23}}$ (that is, $e_{d_{12}}$ is the "closer" edge of difference to e). Let $e_1, \ldots, e_m$ be the path between e and $e_{d_{12}}$ in T_1,

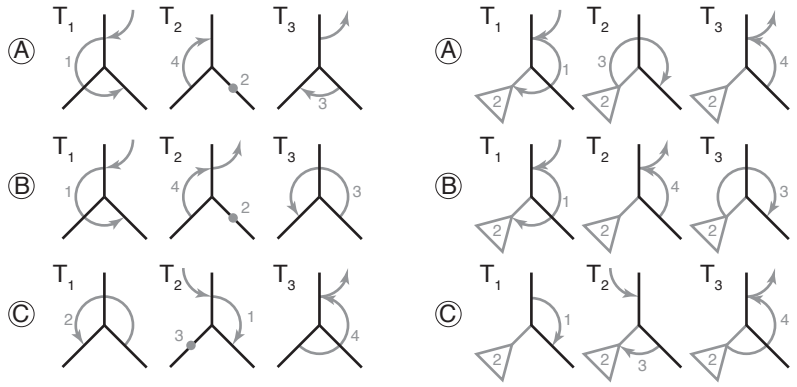

Fig. 4. Left: Base Case I for Lemma 4. The labels on the arrows and subtrees indicate the order of the traversal. S_x is a rooted cherry that is the same across T_1, T_2, and T_3. Note that the path through the expansions of the edges in S_x can be notated as a series of NNI moves, moving the leaf added in the expansions through the various edges in S_x Right: The three walks show that a Hamiltonian path can start at $T_i(p_i)$ and end on $T_j(p_j)$ where $i \neq j$.

and let $S_1, S_2, \ldots, S_{m-1}$ be the subtrees along the path. By hypothesis, the subtrees, $S_1, S_2, \ldots, S_{m-1}$, occur in all three trees. We apply Lemma 4 to each of the subtrees and link the resulting paths by visiting the the expansion of the subtree, S_i, followed by the expansion of the path edge, e_{i+1}, in T_1, T_2, and T_3, for each i, creating a path that extends to $T_2(e_{d_{12}})$ or $T_3(e_{d_{12}})$.

We may assume that our path thus far ends at $T_3(e_{d_{12}})$. Let C denote the subtree, identical in T_1, T_2 and T_3, that we have traversed thus far. The root edge of C is incident with $e_{d_{12}}$. Let D be the subtree whose root edge is incident with C in T_2 and T_3 (that is, C and D are on the "same side" of $e_{d_{12}}$ in T_2 and T_3) and let F and G be the remaining subtrees whose root edges are incident with $e_{d_{12}}$ in T_2 and T_3 (that is, F and G are on the "opposite side" of $e_{d_{12}}$ from C and D in T_2 and T_3). Since the root edges, r_C and r_D, of C and D are adjacent in T_3, we can move from $T_3(r_C)$ to $T_3(r_D)$ without having to traverse the expansion of $e_{d_{12}}$ (that traversal is optional). If D does not contain $e_{d_{23}}$, it is identical across all three trees and we can apply Lemma 4 to yield an extension of our path that traverses the union of the expansions of all three copies of D, ending in T_1.

T_1 differs from T_3 around $e_{d_{12}}$, so the root edge of C in T_1 is not incident with D, but with one of the other two subtrees, say F. To simplify the argument, we will assume that F does not contain the edge $e_{d_{23}}$ (if F does contain $e_{d_{23}}$, the proof follows by a slightly more complicated, but similar, argument). We can similarly extend our path to the union of the expansions of the subtree F in T_1, T_2, and T_3, ending in T_3. In addition, because D and F are on "opposite sides" of $e_{d_{12}}$ in T_1, our path must first visit $T_1(e_{d_{12}})$.

We now have a path that traverses the expansion of three out of four of the subtrees whose root edges are incident with $e_{d_{12}}$ and $T_1(e_{d_{12}})$. To reach the final of the four subtrees, G, we cross $T_3(e_{d_{12}})$, then $T_2(e_{d_{12}})$. From there, we visit the expansion of the root edge of G, which we traverse by an argument similar to the one above. We note that if D, above, had contained $e_{d_{23}}$, then we would use the argument for traversing $T_1(e_{d_{12}})$, $T_2(e_{d_{12}})$, and $T_3(e_{d_{12}})$ to traverse $T_1(e_{d_{23}})$, $T_2(e_{d_{23}})$, and $T_3(e_{d_{23}})$. Once we have traversed the expansion of the three copies of G, we have a path that visits all the edges of T_1, T_2, and T_3, and thus all trees in the union of the expansions of those trees.

Case II: We must also consider the case where e lies on the path between the edges of differences, $e_{d_{12}}$ and $e_{d_{23}}$. While the argument for this case is similar to that above, there is the additional difficulty of "starting in the middle." It is necessary to traverse the path from e to $e_{d_{12}}$ and still have unvisited edges upon which to return so that the "other side" of e can also be visited. This can be accomplished by traversing only the path edges (and none of the attached subtrees) in the tree T_1 until the edge of difference is reached. The path is then built, as above, but on the return, the subtrees on the path are linked by visits to the path edges in only T_2 and T_3. Once that section of the expansions of the trees has been visited, the remaining trees in the union of the expansions are visited (namely the union of the expansion of B where B is the subtree resulting from removing e and that contains $e_{d_{23}}$). The result is a Hamiltonian path of the union of the expansions of T_1, T_2, and T_3.

Case III: Note the special case where $e_{d_{12}} = e_{d_{23}}$. In this case, T_1 and T_2 are NNI neighbors, T_2 and T_3 are NNI neighbors, and T_1 and T_3 are NNI neighbors. Such a case is a simplified version of the previous cases, and thus is covered above. □

Theorem 1: For all n, there exists a Hamiltonian path through the n-leaf NNI treespace.

Proof: By induction on n, the number of leaves.

Base Case: When n = 4, $(2n-5)!! = 3$. Let the four leaves be a, b, c, and d. Then, without loss of generality, e_d in T_1 separates a, b from c, d; in T_2 e_d separates a, c from b, d; and in T_3 e_d separates a, d from b, c. T_1 and T_2 are NNI neighbors, and T_2 and T_3 are NNI neighbors. By

the previous lemma, there is a Hamiltonian path through the 4-leaf NNI treespace (see Figure 1).

Inductive Step: Assume there is a Hamiltonian path through the n-leaf NNI treespace. The walk visits the ordered set of trees, $T_1, T_2, T_3, \ldots T_{(2n-5)!!}$. By the definition of a Hamiltonian path, T_x and T_{x+1} are NNI neighbors where $1 \leq x < (2n-5)!!$.

By the previous lemma, the union of the expansions of the triplet T_y, T_{y+1}, and T_{y+2} has a Hamiltonian path where $y = 3z - 2$ and $1 \leq z \leq \frac{1}{3}(2n-5)!!$.

Because $(2n-5)!!$ is divisible by 3 when $n \geq 4$, there is an ordered set of successive triplets, $R_1, R_2, \ldots, R_{\frac{(2n-5)!!}{3}}$, where R_1 is the triplet of trees T_1, T_2, T_3, and $R_{\frac{(2n-5)!!}{3}}$ is the triplet $T_{(2n-5)!!-2}, T_{(2n-5)!!-1}, T_{(2n-5)!!}$. The unions of expansions on each of these triplets has a Hamiltonian path.

Consider T_y, the third tree in triplet R_v, and T_{y+1}, the first tree in R_{v+1}. T_y and T_{y+1} are NNI neighbors. Then, by Lemma 3, $T_y(e)$, where e is not e_d, is an NNI neighbor of $T_{Y+1}(e)$. The end of the Hamiltonian path through an expansion of a triplet can thus be connected to the beginning of the Hamiltonian path through the expansion of the succeeding triplet. As shown above, there is an ordered set of triplets which covers n-leaf treespace with a Hamiltonian path. The expansions of each of these triplets has a Hamiltonian path, and each of the walks can be linked by a single NNI move. Therefore, a Hamiltonian path exists through the $(n+1)$-leaf NNI treespace. □

4 Conclusion

We have shown that the shortest walk on the space of binary phylogenetic trees with n leaves under the NNI metric is a Hamiltonian path. Since visiting each node exactly once is the minimal path length possible, this answers Bryant's First Combinatorial Challenge on the length of the shortest walk of trees under the NNI metric. In addition, since every NNI move can be simulated by an SPR or TBR move, this also gives an alternative proof to the Hamiltonian paths of SPR and TBR treespace [5]. Our iterative approach to building a Hamiltonian path for the space of trees with $n+1$ leaves from a path of the smaller space of trees with n leaves does not yield an algorithm for producing a Hamiltonian path directly nor do we see an obvious way to do this.

5 Acknowledgments

The authors would like to thank the anonymous reviewers for their thoughtful and helpful comments, which greatly improved this paper. For spirited discussions, the authors thank the members of the Treespace Working Group at CUNY: Alan Joseph Caceres, Ann Marie Alcocer, Kadian Brown, Juan Castillo, Samantha Daley, John DeJesus, Michael Hintze, Daniele Ippolito, Jinnie Lee, 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).

References

[1] C. Semple and M. Steel, *Phylogenetics*, ser. Oxford Lecture Series in Mathematics and its Applications. Oxford: Oxford University Press, 2003, vol. 24.

[2] L. Foulds and R. Graham, "The Steiner problem in phylogeny is NP-complete," *Advances in Applied Mathematics*, vol. 3, no. 1, pp. 43–49, 1982.

[3] S. Roch, "A short proof that phylogenetic tree reconstruction by maximum likelihood is hard," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 3, no. 1, pp. 92–94, 2006.

[4] D. Bryant, "Penny Ante: A mathematical challenge," 2008, http://www.math.canterbury.ac.nz/bio/events/kaikoura09/penny.shtml.

[5] A. J. J. Caceres, J. DeJesus, M. Hintze, D. Moore, and K. St. John, "Walks in phylogenetic treespace," *Information Processing Letters*, vol. 111, pp. 600–604, 2011.

[6] D. Robinson, "Comparison of labeled trees with valency three," *Journal of Combinatorial Theory, Series B*, vol. 11, no. 2, pp. 105–119, 1971.

[7] B. L. Allen and M. Steel, "Subtree transfer operations and their induced metrics on evolutionary trees," *Annals of Combinatorics*, vol. 5, no. 1, pp. 1–15, 2001.

[8] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, 2nd ed. Cambridge, MA: MIT Press, 2001.

Brief Author Biographies

Kevaughn Gordon is pursuing a bachelors degree in computer science at Lehman College, City University of New York (CUNY). He received a bachelors degree in biochemistry from Binghampton University. His research interests include computational biology and natural language processing. Contact him at kevaughn.gordon@lc.cuny.edu.

Eric Ford is a doctoral candidate at the CUNY Graduate Center and a lecturer at Lehman College, CUNY. Ford holds a masters degree in linguistics from the Graduate Center. His research interests include computational linguistics and bioinformatics. Contact him at eford@gc.cuny.edu.

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.