Walks in Phylogenetic Treespace

Alan Joseph Caceres*† Samantha Daley *† John DeJesus*†

Michael Hintze*† Diquan Moore*† Katherine St. John*

Abstract

We prove that the spaces of unrooted phylogenetic trees are Hamiltonian for two popular search metrics: Subtree Prune and Regraft (SPR) and Tree Bisection and Reconnection (TBR). Further, we make progress on two conjectures of Bryant on searching phylogenetic treespace: treespace under the Nearest Neighbor Interchange (NNI) metric has a 2-walk, and there exists SPR neighborhoods without complete NNI walks.

1 Introduction

Finding the evolutionary history, or phylogeny, for a set of species is a core activity in biology and is used for classifying species, building the "Tree of Life," designing the flu vaccine, and determining the origins of viruses such as HIV [?]. These phylogenies, are often modelled by unrooted, leaf-labelled trees [?] and finding the optimal tree for biological data is NP-hard [?, ?]. To improve these computationally expensive searches, we focus on the underlying space of trees, under metrics induced by popular operations used to search the space of all trees with n leaves. We make progress towards solving two conjectures of Bryant [?] on the shortest walk of the full treespace and of a SPR neighborhood under the NNI metric. We show that there is a Hamiltonian circuit of treespace for the SPR and TBR metrics and for the NNI metric for $n < 8$. Further, we show there exists a

*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.

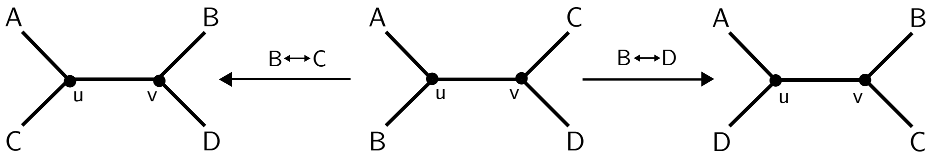

Figure 1: The two possible NNI operations on an internal edge (u, v).

2-walk for treespace under the NNI-metric. Towards the second conjecture of Bryant, we show that for every $n > 5$, there is an SPR neighborhood for which there is no NNI walk.

Prior work has focused on the complexity of calculating the metrics (all three are NP-hard to compute [?, ?, ?, ?]) as well as calculating the diameter of the space. By work of Li *et al.* and DasGupta *et al.* [?, ?], the diameter for NNI is $\Theta(n \lg n)$, while Allen and Steel showed that the diameter of the space for SPR and TBR is $\Theta(n)$. Bastert *et al.* [?] examined the space of trees under NNI in terms of coherent algebras and spectrum. Motivated by the computationally expensive searches of treespace, Bryant [?] posed the questions of how many NNI moves are needed to visit all trees, and similarly how many NNI moves are needed to traverse an SPR neighborhood.

2 Background

We briefly outline the background needed for this paper. For a more thorough overview, see Semple and Steel [?] and Corman, Leiserson, and Rivest [?].

2.1 Phylogenetic Trees and Distances

The evolutionary relationship between various biological organisms can be represented as a leaf-labelled tree where the tree is rooted if the evolutionary origin is known. We focus on binary (or "fully resolved"), unrooted trees. The leaf labels often represent DNA or protein sequences, and the internal nodes correspond to speciation events. Finding the optimal evolutionary tree for a set of species is NP-hard [?, ?]. Biologically inspires operations are used to heuristically search the space of all trees. These search heuristics often produce different evolutionary trees on the same

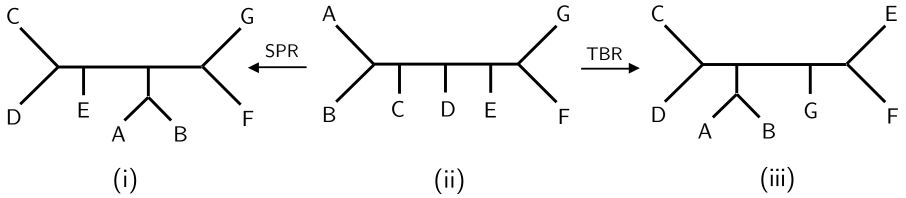

Figure 2: The trees on the left and center differ by a single SPR move. The tree on the right differs by a single TBR move from the center tree.

set of species. In order to compare phylogenies, several metrics for measuring distance have been defined in literature.

The *Nearest Neighbor Interchange* (NNI) is a distance metric introduced independently by Dasgupta *et al.* [**?**] and Li *et al.* [**?**]. An NNI operation swaps two subtrees that are separated by an internal edge in order to generate a new tree [**?**] (see Figure **??**).

Definition 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 one tree into the other.*

The complexity of computing the NNI distance was open for over 25 years, and was proven to be NP-complete by Allen and Steel [**?**]. For a tree with n uniquely labeled leaves, there are $n - 3$ internal branches. Thus, there are $2(n - 3)$ NNI rearrangements for any tree. The trees that are a single rearrangement of a given tree form its 1-neighborhood under the NNI metric.

The most popular move 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 (see Figure **??**).

Definition 2. *The* **SPR distance***, $D_{SPR} = (T_1, T_2)$, between two trees is the minimal number of SPR moves needed to transform the first tree into the second tree (see Figure* **??***).*

The calculation of SPR distances has been proven NP-complete for both rooted and unrooted trees [**?**, **?**]. Approximation algorithms for SPR on rooted trees exist [**?**, **?**]. A generalization of the SPR is the Tree-Bisection-Reconnnection (TBR) operation. Roughly, a TBR move breaks an edge

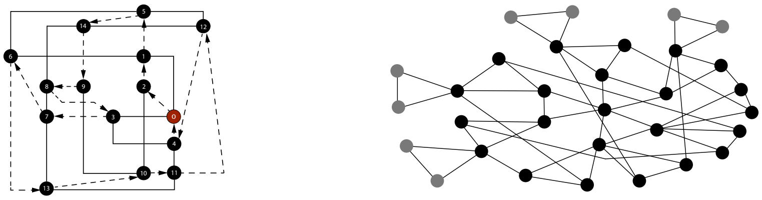

Figure 3: Walks of Treespace: Left: The solid edges are a 1-walk (Hamiltonian path) of 5-species treespace under the NNI metric. Right: For 6-species treespace, an SPR neighborhood with the NNI connections, illustrating that no NNI 1-walks exists.

of the tree and then selects two edges on the resulting subtrees and connects the selected edges by a new edge.

Definition 3. *The* **TBR distance**, *$D_{TBR} = (T_1, T_2)$, between two trees is the minimal number of TBR moves needed to transform the first tree into the second tree (see Figure* **??***).*

We note that NNI ⊆ SPR ⊆ TBR [**?**]. That is every NNI move is also an SPR move which is also a TBR move. We use this fact to extend our results from the SPR to TBR metrics. Each metric induces a discrete metric space on the set of n leaf trees. Since these spaces are defined both by the number of leaves in the underlying trees as well as the metric chosen, we will refer to these spaces as the ***n*-species treespace** under the said metric. The majority of this paper concentrates on the n-species treespace under the NNI metric and SPR metric. Figure **??** show the 5-species treespace under the NNI metric.

2.2 Walks

We focus on the shortest paths needed to visit all trees in a given treespace. In particular, we are interested in the length of the shortest path that visits all trees with the goal of bounding searches for the optimal phylogenetic tree.

Definition 4. *An undirected graph G has a* ***k*-walk** *if a walk along the edges can visit every vertex at most k times.*

The special case of $k = 1$ is often referred to as a Hamiltonian path:

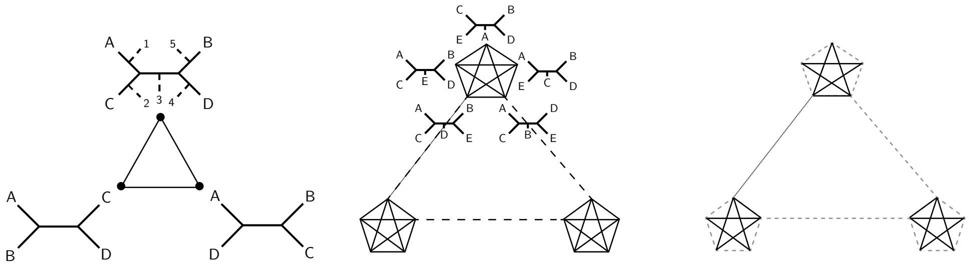

Figure 4: Theorem 1: Left: the circuit that visits all trees in 4-species space. Center: the expansion of each tree by a new leaf (the sets S_i) and the connection between them. Right: A Hamiltonian circuit of 5-species space under the SPR (and TBR) metric.

Definition 5. *An undirected graph G has a* **Hamiltonian path** *if a walk along the edges can visit every vertex exactly once. A* **Hamiltonian cycle** *is a walk along the edges of an undirected graph that visits every vertex exactly once and ends at the beginning vertex, also know as a Hamiltonian circuit.*

A Hamiltonian path is traceable and can be verified in polynomial time and finding a Hamiltonian cycle is NP-complete. [**?**]. A 2-walk is also traceable and can be verified in polynomial time. When a Hamiltonian path is unavailable, a 2-walk can be used to reduce the maximum number of times a single tree is evaluated by a search algorithm.

3 Walks of Treespace

We show that for every n, n-species treespace under the popular SPR and TBR metrics has a Hamiltonian circuit. Further, we show that under the NNI metric, for every n, n-species treespace has a 2-walk.

3.1 Hamiltonian Circuits under the SPR and TBR Metric

Since every SPR move is also a TBR move, it suffices to show that there is a Hamiltonian path for treespace under the SPR metric.

Theorems 1. *For every n, there exists an Hamiltonian circuits of the n-species treespace under the SPR and TBR metrics.*

Proof. Proof by induction on n. For $n = 4$, there are exactly three trees, and they form a complete graph under the SPR metric, and thus have a Hamiltonian circuit.

For $n > 4$: assume true for n-species treespace and show for $(n+1)$-species treespace. Let $t_1, t_2, \ldots, t_{(2n-5)!!}$ be the trees of the n-species treespace. Each $(n+1)$-species trees can be uniquely generated by adding a $n+1$ leaf to an edge of one of the n-species trees [?]. Let S_i be the set of $(n+1)$-species trees generated from t_i for $i = 1, 2, \ldots, (2n-5)!!$. We note that each S_i forms a complete graph under the SPR metric (see Figure 3). By inductive hypothesis, there exists a Hamiltonian circuit, H_n, on the n-species treespace which will form the "backbone" of a Hamiltonian circuit for the $(n+1)$-species treespace. To create a Hamiltonian circuit for $(n+1)$-species treespace, we replace each t_i with the set S_i. Since S_i is a complete graph, we can find a Hamiltonian path with any starting and ending points.

Define the Hamiltonian circuit H_{n+1} as follows: For each S_i, let h_i be the Hamiltonian path that begins with the tree with 1 and $n+1$ as sibling pairs and ends with the tree n and $n+1$ as sibling pairs. Begin the circuit of S_1 by traversing the trees in h_1. Attach the last tree in h_1 to the last tree in h_2. This is possible, since by construction, the new leaf is attached in identical positions, and these two trees can be connected by the first edge in H_n. Continue by following the path h_2 in reverse until reaching the first tree in its path. By similar construction, H_{n+1} can be extended to traverse the first $(2n-5)!! - 1$ sets of trees. Since the number of these sets is odd, we need to do something slighly different for the last two sets. For $S_{(2n-5)!!-1}$, choose $h_{(2n-5)!!-1}$ to be the Hamiltonian path from the tree with 1 and $n+1$ as a sibling pair and the tree with 2 and $n+1$ as a sibling pair. And for the last set, $S_{(2n-5)!!}$, choose $h_{(2n-5)!!}$ to be the Hamiltonian path from the tree with 2 and $n+1$ as a sibling pair and the tree with n and $n+1$ as a sibling pair. We continue the circuit H_{n+1} across $h_{(2n-5)!!-1}$ and $h_{(2n-5)!!}$, connecting the last tree in $h_{(2n-5)!!}$ to the first tree in h_1.

It is easy to check that this circuit visits every tree and is thus a Hamiltonian circuit of $(n+1)$-species space under the SPR metric. Since SPR $\subset$ TBR, the result also holds for TBR. □

3.2 2-Walk under the NNI Metric

For the NNI metric, we can generalize the argument of Theorem **??**. We rely on the fact that every set of $(n+1)$-leaf trees generated from a single n leaf trees (the sets S_i above) can be traversed by a 2-walk. We note that these sets S_i for NNI are not complete graphs (unlike for SPR and TBR) and further there exists sets for $n > 6$ where no Hamiltonian path is possible. Thus, our argument for a Hamiltonicity for treespaces under the SPR and TBR metrics will not extend to NNI, and the existence of a Hamiltonian walk through an n-species treespace under the NNI metric is currently unknown. We can however proof the existence of a 2-walk through any n-species treespace.

Theorem 1. *For every n, there exists a 2-walk of the n-species treespace under the NNI metric.*

Proof. Proof by Induction on n. For $n = 4$, there are 3 trees connected in a triangle and thus have a 2-walk.

For $n > 4$: Let W_n be a NNI 2-walk of the n-species treespace. As in Theorem 1, let S_i be the set of $(n+1)$-species trees generated from t_i for $i = 1, 2, \dots, (2n-5)!!$. We note that by construction, each tree in S_i corresponds to an edge of the tree t_i. Let P be a planar tree drawing of t_i and let C_i be a closed curve such that C_i is within ϵ of P for sufficiently small $\epsilon > 0$. C_i induces a 2-walk of the set T_i, by following the curve counter-clockwise around the nodes of the tree. Following a similar construction to the proof of Theorem 1, we can build a new 2-walk W_{n+1} of the $n+1$-species space that visits all the elements of the space. □

4 NNI-Walks on SPR Neighborhoods

Bryant [**?**] asked what is the shortest NNI walk that visits every tree in an SPR neighborhood. We make progress on this conjecture by showing that for every n, there exists an SPR neighborhood without a NNI 1-walk.

Theorem 2. *For every $n \geq 6$, there exists an SPR-neighborhood of the n-species treespace, that does not have a Hamiltonian path.*

Proof. Figure **??** shows an SPR-neighborhood of 6-species treespace where no NNI 1-walk exists since 8 nodes of degree 2 (the highlighed "triangles") only two of which could be traversed in any path that visits nodes only once. For larger n, we can find similar regions of the SPR neighborhood that preclude any Hamiltonian path of the SPR neighborhood.

For $n \geq 6$, let $T = (\ldots(((1,2),3),4),\ldots n-1),n)$, $T_1 = (\ldots(((n,1),2),\ldots n-2),n-1)$, $T_2 = (\ldots(((n,2),1),\ldots n-2),n-1)$, and $T_3 = (\ldots(((1,2),n),3),\ldots,n-2),n-1)$. T_1, T_2 and T_3 are each a single SPR move from T and thus part of the SPR neighborhood of T. We note that $d_{NNI}(T_1,T_2) = d_{NNI}(T_2,T_3) = d_{NNI}(T_3,T_1) = 1$ and, by analysis by cases, no other NNI neighbors of T_1 and T_2 occur in the SPR neighborhood. So, any path from T_1 and T_2 to the rest of the space must pass through T_3. We refer to T_1, T_2, and T_3 as isolated triangles in the NNI graph of the neighborhood. Note that 3 other isolated triangles exist– namely the trees resulting by moving the subtree consisting of $n-1$ to neighbor 1 and 2, those resulting by moving the subtree consisting of 1 to neighbor $n-1$ and n, and those resulting by moving the subtree consisting of 2 to neighbor $n-1$ and n. To visit all of these isolated triangles in the same path would require visiting nodes twice. Thus, no Hamiltonian path exists. □

5 Acknowledgments

We would like to thank the members of the Treespace Working Group at CUNY for helpful insights and conversations: Ann Marie Alcocer, Kadian Brown, Eric Ford, Kaitlin Hansen, Daniele Ippolito, Jinnie Lee, Ling Li, and Oliver Mendez.