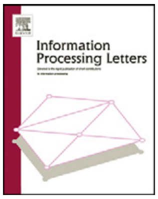
Information Processing Letters 111 (2011) 600–604

Contents lists available at ScienceDirect

Information Processing Letters

www.elsevier.com/locate/ipl



Walks in phylogenetic treespace

Alan Joseph J. Caceres [1,2], Samantha Daley [1,2], John DeJesus [1,2], Michael Hintze [1,2], Diquan Moore [1,2], Katherine St. John [*,1]

Dept. of Math & Computer Science, Lehman College, City University of New York, Bronx, NY 10468, United States

ARTICLE INFO

Article history:
Received 18 November 2010
Received in revised form 15 March 2011
Accepted 18 March 2011
Available online 23 March 2011
Communicated by J. Torán

Keywords:
Graph algorithms
Computational biology
Tree metrics
Phylogenetics

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 exist 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 [11]. These phylogenies, are often modeled by unrooted, leaf-labeled trees [15] and finding the optimal tree for biological data is NP-hard [9,14]. 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. The space of trees is huge; there are $(2n-5)!! = (2n-5)(2n-7)\cdots 7\cdot 5\cdot 3\cdot 1$ unrooted trees on n leaves which grows as $\frac{1}{2\sqrt{2}}(\frac{2}{e})^n n^{n-2}$ as $n \to \infty$ [15]. We make progress towards solving two conjectures of Bryant [6] on the shortest walk of the full treespace and of an 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 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 which visits every tree in the neighborhood exactly once.

Prior work has focused on the complexity of calculating the metrics (all three are NP-hard to compute [1,8,10,13]) as well as calculating the diameter of the space. By work of Li et al. and DasGupta et al. [8,13], 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. [2] examined the space of trees under NNI in terms of coherent algebras and spectrum. Motivated by the computationally expensive searches of treespace, Bryant [6] 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 [15] and Cormen, Leiserson, and Rivest [7].

---

[*] Corresponding author.
E-mail address: stjohn@lehman.cuny.edu (K.St. John).

[1] Support provided by NSF Grant #09-20920.

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

0020-0190/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ipl.2011.03.015

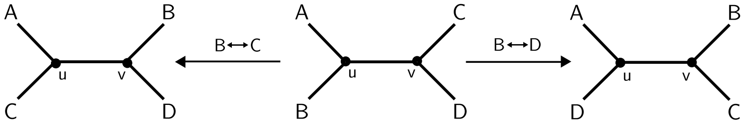
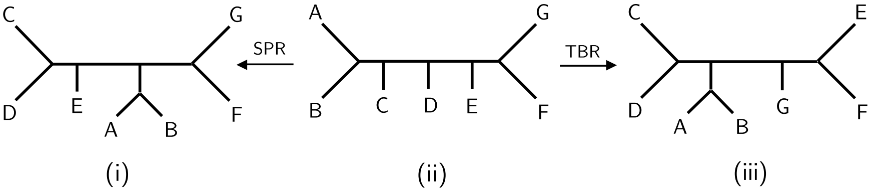


Fig. 1. The two possible Nearest Neighbor Interchange (NNI) operations on an internal edge (u, v).



(i) (ii) (iii)

Fig. 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.

2.1. Phylogenetic trees and distances

The evolutionary relationship between various biological organisms can be represented as a leaf-labeled 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 [9,14]. Biologically inspired operations are used to heuristically search the space of all trees. Additionally, these search heuristics often produce different evolutionary trees on the same 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. [8] and Li et al. [13]. Roughly, an NNI operation swaps two subtrees that are separated by an internal edge in order to generate a new tree [1].

Definition 1. (See Allen and Steel [1].) Any internal edge of an unrooted binary tree has four subtrees attached to it. A Nearest Neighbor Interchange (NNI) 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 Fig. 1.

Definition 2. 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 [1]. 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.

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 (see Fig. 2).

Definition 3. (See Allen and Steel [1].) A **subtree prune and regraft** (SPR) 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 Fig. 2).

Definition 4. 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.

The calculation of SPR distances has been proven NP-complete for both rooted and unrooted trees [5,10]. Approximation algorithms for SPR on rooted trees exist [3, 4]. A generalization of the SPR is the Tree-Bisection-Reconnection (TBR) operation. Roughly, a TBR move breaks an edge of the tree and then selects two edges on the resulting subtrees and connects the selected edges by a new edge.

Definition 5. (See Allen and Steel [1].) A tree bisection and reconnection (TBR) on a binary tree T is defined as removing any edge, giving two new subtrees, t_1 and t_2, which are then reconnected by creating a new edge between the midpoints of any edge in t_1 and any edge in t_2. Again forced contractions are applied to ensure the resulting tree is binary (see Fig. 2). In the case that one of the subtrees is a single leaf, then the edge connecting t_1 and t_2 is incident to the leaf.

Definition 6. 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.

We note that NNI $\subseteq$ SPR $\subseteq$ TBR [1]. 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, which we represent by a graph where two trees are connected if and only if they have distance 1 under the given metric. The trees that are a single rearrangement of a given tree form its 1-neighborhood. Since these spaces are defined both by the number of leaves in the underlying trees as well as the metric chosen, we will