Characterizing Local Optima for Maximum Parsimony

Ellen Urheim · Eric Ford · Katherine St. John*

the date of receipt and acceptance should be inserted later

Abstract Finding the best phylogenetic tree under the maximum-parsimony optimality criterion is computationally difficult. We quantify the occurrence of such optima for well-behaved sets of data. When Nearest Neighbor Interchange (NNI) operations are used, multiple local optima can occur even for "perfect" sequence data, which results in hill-climbing searches that never reach a global optimum. In contrast, we show that when neighbors are defined via the Subtree Prune and Regraft (SPR) metric, there is a single local optimum for perfect sequence data, and thus every such search finds a global optimum quickly. We further characterize conditions for which sequences simulated under the Cavender-Farris-Neyman and Jukes-Cantor models of evolution yield well-behaved search spaces.

Keywords: maximum-parsimony, compatibility, phylogenetic islands, attraction basins, tree searching.

1 Introduction

Finding the optimal phylogeny, or evolutionary history, for a set of species is a central goal of biology. The maximum-parsimony optimality criterion for a phylogenetic tree is often used due to its simplicity in capturing evolutionary change and its ease of computation [15]. Given traits or characters for a set of taxa, the maximum-parsimony criterion seeks the tree which has the minimal number of changes of character states across its edges. For a given tree with n taxa and associated character sequences of length k, this score can be calculated in linear time in n and k [15]. However, to exactly find the optimal tree (that is, the tree with smallest maximum-parsimony score) is computationally difficult [17]. The size of the search space of trees grows super-exponentially in the number of taxa [32], and exact algorithms (e.g. [21,22]) are thus only effective for small sets of taxa. Therefore, heuristic search methods are often used (e.g. [19, 31,39]). The underlying structure of these methods follows the local search paradigm: at each step, a "neighbor" of the current tree in the search space is chosen to be the new tree, and this is repeated until a local optimum is reached or time is exhausted. In the simplest case, a "hill-climbing" approach is used, where the chosen tree at each step is the best-scoring neighbor, though more sophisticated techniques have been explored, such as simulated annealing [11] and Monte Carlo methods [31] (see [40] for a survey of popular approaches).

We examine two popular operations to define neighbors of trees and their effectiveness for hill-climbing search with well-behaved data. Roughly, the Nearest Neighbor Interchange (NNI) operation "swaps" subtrees around an internal branch of the tree [32], and the Subtree Prune and Regraft (SPR) breaks off a subtree and regrafts it to a different branch [20]. These operations are metrics and can be used to define a search space of trees. For a given metric, one can view the space of all binary n-leaf trees as a graph with vertices labeled by the trees. Neighbor vertices are connected by an edge when they differ by a single application of the relevant operation, and a *neighborhood* of a vertex is a set of all the vertices

* Corresponding Author: stjohn@lehman.cuny.edu

Departments of Mathematics and Chemistry, Johns Hopkins University, Baltimore, MD; Department of Computer Science, Graduate Center, City University of New York, New York, NY; Department of Mathematics and Computer Science, Lehman College, City University of New York, Bronx, NY; and Invertebrate Zoology, American Museum of Natural History.

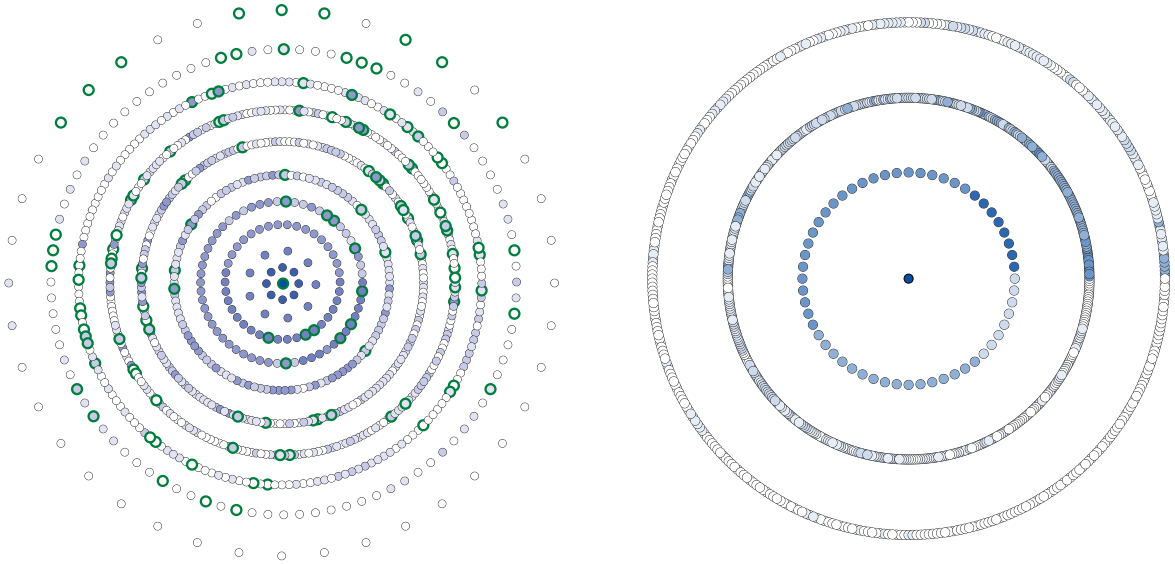

Fig. 1 The center tree is an unrooted caterpillar tree, T, on 7 leaves. Each node is a 7-leaf tree, colored by its parsimony score with respect to a character sequence compatible only with T (darker is more optimal). Nodes are arranged in concentric circles by their distance from T. Local minima are outlined with a thicker green rule. *Left*: NNI space has multiple local minima, even among the worst-scoring trees. *Right*: SPR space has a single attraction basin and one local minimum, T.

that share an edge with that vertex. Different metrics yield graphs that differ in both neighborhood size and diameter (the maximum distance between two vertices in the graph), suggesting very different results for heuristic search. For example, an unrooted tree with n leaves has $2n-6$ neighbors in the search space with NNI operations and $(2n-6)(2n-7)$ neighbors in the search space with SPR operations [1]. The maximal distance between any two trees using NNI operations is $\Theta(n \lg n)$ [26], while the maximal distance under SPR is $O(n)$. Each search space can be divided into *attraction basins*: all trees whose search would lead to the same local optimum under the greedy algorithm. While commonly used for analyzing for steepest ascent algorithms and real-valued functions optimality functions [9,18], it can be extended to our graph search space by using hill-climbing as the greedy search algorithm.

Empirical studies suggest that the search spaces resulting from these two operations differ greatly. Maddison [27] defined an *island* as a connected set of trees with scores better than some number L. For multiple biological data sets, he found that there were multiple islands for all tree-manipulation operations he studied, including NNI and SPR. Kirkup and Kim [25] also investigated the search space of trees under NNI and SPR, observing that as the number of taxa increased, so did the number of local optima found via heuristic search. Money and Whelan [28] examined local optima for another popular optimality criterion: the maximum-likelihood criterion [16]. They analyzed the well-studied yeast data set of Rokas *et al.* [33], which consists of 106 genes for 8 taxa of yeast. While the search spaces for all of the genes under NNI yielded multiple local optima and multiple attraction basins, under SPR only one gene had a search space with multiple local optima under the GTR+Γ model of evolution, and only five had multiple optima under the simpler Jukes-Cantor model of evolution.

We rigorously show the difference between NNI and SPR search spaces for maximum-parsimony with "perfect data"—sequences of compatible characters that are displayed by exactly one tree (a "perfect phylogeny"). For any such set of sequences, the SPR search space has only one attraction basin and one local optimum; whereas, there exist perfect data for which the NNI search space has multiple attraction basins and multiple local optima. We further show that if character sequences are generated, with suitable parameters, under the Jukes-Cantor or Cavender-Farris-Neyman models, then hill climbing search with SPR neighbors will, with high probability, find the model tree, echoing previous results [2,4]. We include experimental results on both biological and simulated data.

2 Methods

We use the precise, but terse, notation of Semple and Steel [35], since we rely heavily on it in the proofs of the results. See [23,40] for more biologically inspired versions.

Throughout, n is the number of taxa, and $T_n = (\{1,\ldots,n\}, E)$ will refer to a binary phylogenetic tree (referred to here as a *binary tree* for simplicity) on n leaves with ρ denoting the root, when present. Two distinct leaves of a tree are a *sibling pair* if they are adjacent to a common vertex. Trees with the minimal number of sibling pairs are called *caterpillars*. A *balanced tree* of height $h \geq 0$ is a rooted binary tree with $n = 2^h$ leaves, each of which is separated from the root by exactly h edges, and an *unrooted balanced tree* can be formed by simply deleting the root. A tree can be augmented by a real-valued function $w : E \to \mathbb{R}$ that assigns *weights* to the edges. Following Atteson's notation [2], we extend those weights to any two leaves of the tree, i and j by assigning the sum of the weights on the unique, minimal path between them. A *distance matrix*, D^T for tree (T, w), has entries $d_{i,j}$, the weighted path distance between i and j in (T, w).

A *character* on a set $\mathcal{L}$ is a function, χ, from $\mathcal{L}$ into a set C of character states. We assume characters are defined on all leaves, and exclude the more general possibility that a character can be defined on subset of leaves (and undefined on the remaining leaves). A character χ is *convex* on a tree $T = (V, E)$ if there exists $\bar{\chi} : V \to C$ such that $\bar{\chi}$ extends χ and for each $\alpha \in C$, the subgraph of T induced by $\{v \in V : \bar{\chi}(v) = \alpha\}$ is connected. A collection of characters on $\mathcal{L}$ is said to be *compatible* if there exists a tree on which all the characters are convex.

The *changing set* of a function f on V is $Ch(f) = \{\{u, v\} \in E : f(u) \neq f(v)\}$ and the *changing number* of f, $ch(f)$ is the cardinality of $Ch(f)$. The *parsimony score*, $\ell(\chi, T)$, of χ on T is the minimum value of $ch(\bar{\chi})$ over all extensions $\bar{\chi}$ of χ to T. We refer to such extensions as *minimal*. Let $\mathcal{S} = \chi_1, \chi_2, \ldots, \chi_k$ be a sequence of characters on $\mathcal{L}$. The *parsimony score*, $\ell(\mathcal{S}, T)$ of $\mathcal{S}$ on T is the sum of the individual parsimony scores of the characters on the tree:

$$\ell(\mathcal{S}, T) = \sum_{i=1}^{k} \ell(\chi_i, T)$$

When restricting to a single edge of a tree, the parsimony score, $\ell(\mathcal{S}, T|_{\{u,v\}})$, is the changing number, $ch(\mathcal{S}, T)$. That is, it is the amount of change between the labellings of u and v. We note that the notation can be extended to specific extensions of the sequence of characters. If we wish to denote the changing number with respect to a specific extension to the internal vertices $\mathcal{U}$, we will write $\ell(\mathcal{S},\mathcal{U}, T)$ for the parsimony score on the tree $\ell(\mathcal{S},\mathcal{U}, T|_{\mathcal{U}})$ or $\ell(\mathcal{S},\mathcal{U}, \{u, v\})$ for the number of characters in the seqeunce S for which the extension (specified by U) require a change on the edge $\{u, v\}$. The parsimony score, $\ell(\mathcal{S}, T)$, can be computed in linear time [13,15]. For unrooted trees, a root can be arbitrarily added for the computation of the score. The rough idea (often called "Fitch's algorithm") is to assign labelings to the internal nodes in two passes: the "first pass" starts at the leaves, and assigns tentative labels starting with parents of leaves, and then parents of other labeled nodes, continuing until the root is labeled. The tentative label at each node is assigned character by character: if the children have overlapping labels for a given character, then the parent is also assigned the labels common to both children. If the children have different labels for a given character, then those union of the labels is assigned. The "second pass" starts at the root and resolves any unambiguities in the tentative labeling and then to the children of nodes with resolved labelings until all leaves have been visited. It was shown that this results in a minimal labeling from which the parsimony score can be computed. We will assign scores to trees based on their parsimony score with respect to character sequences assigned to the leaves. $\ell(\mathcal{S}, T)$ refers to the parsimony score of tree T with respect to the sequence of characters $\mathcal{S}$. We will call sequences that are compatible with some tree *perfect*. If a tree is compatible with a set of sequences, then it is *optimal*, and often called a *perfect phylogeny*. We will assume that perfect phylogenies are binary trees.

We focus on two tree rearrangement operations: Nearest-Neighbor Interchange (NNI) and Subtree Prune and Regraft (SPR) and their associated metrics (see Figure 2). A (discrete) *treespace* is a graph, $G = (\mathcal{T}_n, E)$, where $\mathcal{T}_n$ is all n-leaf trees and there is an edge between two trees T_1 and T_2 (i.e. $\{T_1, T_2\} \in E$) if and only if they differ by a tree rearrangement operation. The *neighbors* or *neighborhood* of a tree T are all trees that differ from T by a single tree rearrangement operation. Maddison [27] defines an *island* as a connected set of trees in treespace with scores better than some number L. For a set of taxa,

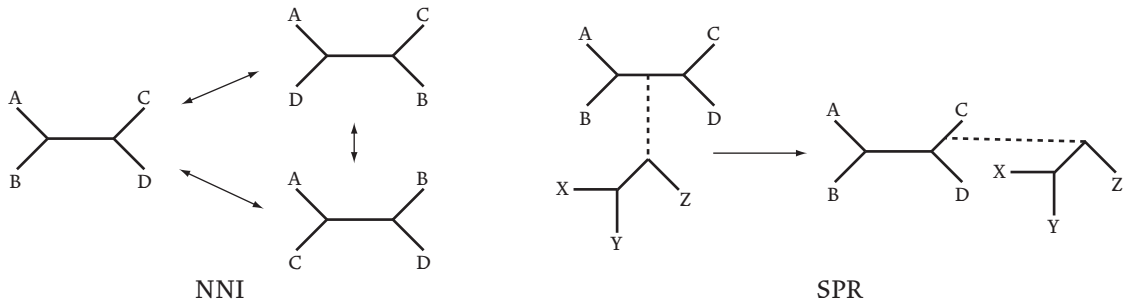

Fig. 2 *Left:* The two NNI tree rearrangements possible from the tree with an edge separating the subtrees A and B from the subtrees C and D. *Right:* A possible SPR tree rearrangement where the subtree containing subtrees X, Y, and Z is pruned and regrafted on the edge neighboring subtree, C.

a score can be assigned to each tree measuring how well each tree agrees with a fixed set of character sequences [15,16].

We will examine character sequences from biological data sets as well as those generated under standard model of evolution: the 4-state Jukes-Cantor (JC) model [24] and the 2-state Cavender-Farris-Neyman (CFN) model [7,14,30]. Roughly, in each of these models, at each edge and for each character there is an equal probability of that character changing to another character state (for more details, see Semple & Steel or Wheeler [35,40]). Heuristic search for optimal trees often follows a *hill climbing* approach, where at each step the neighborhood of the current tree is examined and the best scoring tree is chosen, and the process repeated (see [10]). This algorithm can be stymied by the presence of *local optima*. We define a *local optimum* as a group of connected trees, all with equal scores, and all having better scores than the scores of any of the neighbors of the group. Note that this definition is more strict than Maddison's islands. This local optimum may, indeed, be a global optimum, but it is possible that it is not. In that case, if it is encountered during a hill-climbing search, the search algorithm may stop at the local optimum, and therefore report an incorrect tree or group of trees as the best tree in the space. We extend the defintion of *attraction basin* [18] to our discrete search space to be the set of all trees who lead to the same local optimum under a hill-climbing search.

3 Theoretical Results

We show that SPR-treespace has only a single local optimum, and thus a single attraction basin, for the maximum-parsimony criterion evaluated with perfect sequence data on binary trees.

Phrased in terms of search, we show that every binary tree has either a neighbor with a better parsimony score or is the optimal tree (the trivial case), thus yielding a search space where the naive hill-climbing routine always finds the optimal tree. To show this, we first look at sibling pairs that occur in the optimal tree. If a sibling pair from the optimal tree is missing from the current tree, we prove that performing the SPR operation that re-joins the sibling pair results in a tree with a better parsimony score than that of the current tree. We note that there could be other trees in the SPR neighborhood with even better parsimony scores, but we need only that there is some tree with a better score in the neighborhood (not necessarily the best score) to show that the current tree is not a local optimum.

Theorem 1 *Let $\mathcal{S}$ be a sequence of binary characters on leaf set $\mathcal{L}$ that is compatible with exactly one unrooted binary tree, T, on leaf set $\mathcal{L}$. Suppose $a, b \in \mathcal{L}$ are a sibling pair in T. If T' is any other binary tree in which a and b are not a sibling pair, then either of the two SPR moves that join the sibling pair a, b will result in a new tree with better parsimony score than that of T'.*

Proof: Let $\mathcal{S} = \{\chi_1, \chi_2, \ldots, \chi_k\}$ be a sequence of characters on $\mathcal{L}$ that is compatible with exactly one binary tree, T. Suppose $a, b \in \mathcal{L}$ are a sibling pair in T with parent node p. Let T' be another binary tree in which a and b are not a sibling pair. We claim that moving a from T', via an SPR move, to be next to b decreases the parsimony score. In T', label the parent of a as q. Since q is an internal vertex

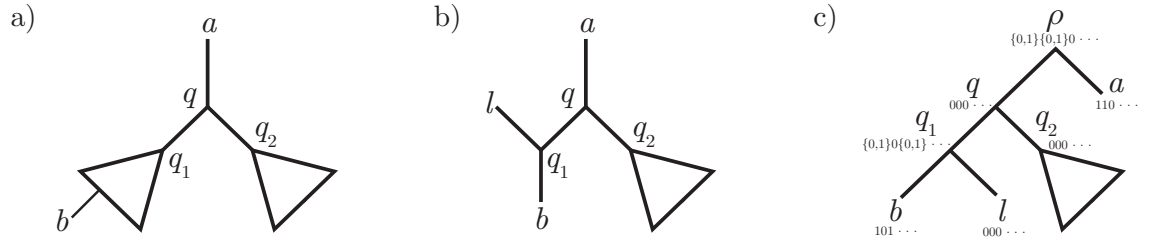

Fig. 3 a) A possible topology for T'; b) The most difficult topology for T' considered in the proof of Theorem 1; c) The tree T'_ρ with optimal labels on internal nodes.

and T' is a binary tree, q has two other neighbors besides a, which we call q_1 and q_2. By assumption, q is not the parent node of b, so without loss of generality, suppose that b is a leaf of the subtree rooted at q_1 (as in Figure 3a). In order to find the parsimony score, we root T', place the root ρ between q and a and call this rooted tree T'_ρ.

We classify the set $\mathcal{S}$ of sequence of characters into the following subsets:

- $\mathcal{S}_a = \{\chi \in S \mid \chi(a) \neq \chi(b), \chi \text{ is constant on } \mathcal{L} \setminus a\}$.
- $\mathcal{S}_b = \{\chi \in S \mid \chi(a) \neq \chi(b), \chi \text{ is constant on } \mathcal{L} \setminus b\}$.
- $\mathcal{S}_{a\cup b} = \{\chi \in S \mid \chi(a) = \chi(b), \forall l \in \mathcal{L} \setminus \{a \cup b\}, \chi(a) \neq \chi(l), \chi \text{ is constant on } \mathcal{L} \setminus \{a \cup b\}\}$.
- $\mathcal{S}^* = \mathcal{S} \setminus (\mathcal{S}_a \cup \mathcal{S}_b \cup \mathcal{S}_{a\cup b})$.

By hypothesis, $\mathcal{S}$ are compatible with T. (We note that every character either agrees on a and b or is "non-informative" in that it provides no information for determining the relationships between the leaves since it has the same value on all leaves or on all but one leaf.) Since a and b are a sibling pair of T, the last set is constant on a and b. Let $\lambda_a = |\mathcal{S}_a|$, $\lambda_b = |\mathcal{S}_b|$, $\lambda_{a\cup b} = |\mathcal{S}_{a\cup b}|$, and $\lambda = \lambda_a + \lambda_b + \lambda_{a\cup b}$.

Without loss of generality, we may assume that a and b have the following labellings:

$$a = 1^{\lambda_{a\cup b}} 1^{\lambda_a} 0^{\lambda_b} \chi_{\lambda+1}(a) \cdots \chi_k(a)$$
$$b = 1^{\lambda_{a\cup b}} 0^{\lambda_a} 1^{\lambda_b} \chi_{\lambda+1}(b) \cdots \chi_k(b)$$

where $\chi_i(a) = \chi_i(b)$ for $i = \lambda + 1, \ldots, k$. Most importantly, as a result of this labeling, any other leaf $\ell \in \mathcal{L}$ must have the labelling $\ell = 0^\lambda \chi_{\lambda+1}(\ell) \cdots \chi_k(\ell)$ by hypothesis.

Assume that $\overline{\mathcal{S}}$ is a minimum extension for T' computed via Fitch's algorithm. By construction, either q_2 is a leaf and its labelling must begin 0^λ, or it is the root of some subtree of T'. However, since the labelling of the leaves of this subtree must all begin 0^λ, any minimum extension $\overline{\mathcal{S}}$ will have the labelling of q_2 beginning 0^λ in order to minimize changes.

Next, we show that there is a minimal extension $\overline{\mathcal{S}}$ in which q has labelling 0^λ. We will consider the most difficult case (see Figure 3b), in which b is a direct descendant of q_1, and show that this is the case. Suppose the subtree rooted at q_1 is merely a sibling pair, where b is one leaf, and the other leaf is ℓ. In this case, the labelling of q_1 must start with a sequence from $\{0,1\}^{\lambda_{a\cup b}} 0^{\lambda_a} \{0,1\}^{\lambda_b}$, where the choice $\{0,1\}$ will be fixed by a second pass.

Since q_2 begins 0^λ and q_1 begins with a sequence from $\{0,1\}^{\lambda_{a\cup b}} 0^{\lambda_a} \{0,1\}^{\lambda_b}$, to minimize change q must begin $0^{\lambda_{a\cup b}+\lambda_a+\lambda_b}$. As a begins $1^{\lambda_{a\cup b}} 1^{\lambda_a} 0^{\lambda_b}$, the root, ρ, must begin with a sequence from $\{0,1\}^{\lambda_{a\cup b}} \{0,1\}^{\lambda_a} 0^{\lambda_b}$, and there is a minimal extension, $\overline{\mathcal{S}}$, in which ρ begins $0^{\lambda_{a\cup b}+\lambda_a+\lambda_b}$. Thus we have the following labellings:

$$\rho = 0^{\lambda_{a\cup b}} 0^{\lambda_a} 0^{\lambda_b} \chi_{\lambda+1}(\rho) \cdots \chi_r(\rho)$$
$$a = 1^{\lambda_{a\cup b}} 1^{\lambda_a} 0^{\lambda_b} \chi_{\lambda+1}(a) \cdots \chi_r(a).$$

where $\lambda = \lambda_p + \lambda_a + \lambda_b$.

Therefore, $l(\mathcal{S}, \mathcal{A}, \{\rho, a\}) \geq \lambda_p + \lambda_a$. Removing a must decrease the overall parsimony score by at least $l(\mathcal{S}, \mathcal{A}, \{\rho, a\})$, even if we retain the same labellings for the internal vertices; the minimum extension on this new tree must either be the same as that of the old tree, or decrease the parsimony score even more.

(We note that the simpler case where q and the parent of b are not the only nodes on the path between a and b follows by a similar argument; the parent of b will have the labelling $\{0,1\}^{\lambda_p}0^{\lambda_a}\{0,1\}^{\lambda_b}$, but with other nodes between the parent of b and q_1, q_1 must have a labelling of 0^{λ} even after just the first pass.)

Finally, we show that reattaching a to the same branch as b increases the parsimony score by exactly λ_a. If r is the parent node of b, then attach a to the parent edge of b, and label the new parent node of a and b as p'.

If we root T' and complete a first pass, we find that the parent node, p, of a, b in T has the representation from $1^{\lambda_{a\cup b}}\{0,1\}^{\lambda_a}\{0,1\}^{\lambda_b}\chi_{\lambda+1}(p)\cdots\chi_r(p)$, as a and b only differ at the second and third subsequences of characters. However, since every other leaf in T' has a labelling beginning with $0^{\lambda_{a\cup b}+\lambda_a+\lambda_b}$, after the first pass the direct ancestor of p will have a labelling beginning with $\{0,1\}^{\lambda_{a\cup b}}0^{\lambda_a}0^{\lambda_b}$. For an optimal labelling, p will have labelling $1^{\lambda_{a\cup b}}0^{\lambda_a}0^{\lambda_b}\chi_{\lambda+1}(p)\cdots\chi_r(p)$.

Thus, the SPR move that moves a to the position of b reduces the overall parsimony score by at least $\lambda_{a\cup b}$. A similar argument shows that the SPR move that moves b to the position of a strictly reduces the overall parsimony score. □

Theorem 1 shows that joining leaves to form sibling pairs of the optimal tree reduces the parsimony score. By similar argument, joining subtrees to form larger subtrees of the optimal tree also reduces the parsimony score:

Corollary 1 *Let $\mathcal{S}$ be a sequence of binary characters on leaf set $\mathcal{L}$ that is compatible with exactly one unrooted binary tree, T, on leaf set $\mathcal{L}$. Suppose T_L, T_R, (T_L, T_R) are subtrees of T. If T' is any other tree which contains T_L and T_R as subtrees but does not contain the subtree (T_L, T_R), then either of the two SPR moves that join the subtree (T_L, T_R) will lower the parsimony score of T'.*

From this, it follows that there is single local optimum for SPR operations for perfect data:

Theorem 2 *Let $\mathcal{S}$ be a binary sequence of characters on leaf set $\mathcal{L}$ that is compatible with exactly one binary tree, T. Then, for all binary trees, T', on $\mathcal{L}$, if $T' \neq T$, there exists a binary tree T'' that is an SPR neighbor of T' and $l(\mathcal{S}, T'') < l(\mathcal{S}, T')$.*

Proof: Assume $T' \neq T$. Then there exist subtrees T_L and T_R such that T_L, T_R, and (T_L, T_R) are subtrees of T; T_L and T_R are subtrees of T'; and (T_L, T_R) is not a subtree of T'. Let T'' be the result of pruning the subtree T_R from T' and regrafting it on the edge for the subtree T_L. By construction, T'' contains the subtrees T_L, T_R, and (T_L, T_R) and is one SPR operation away from T'. By Corollary 1, $l(\mathcal{S}, T'') < l(\mathcal{S}, T')$. □

As a corollary, we note that this implies that in the SPR search space there is only one local optimum (and thus a unique global optimum) and the hill-climbing algorithm always finds it for perfect data:

Corollary 2 *Let $\mathcal{S}$ be a sequence of binary characters on leaf set $\mathcal{L}$ that is compatible with exactly one binary tree, T. Then for all binary trees, T', on $\mathcal{L}$, a hill-climbing search that starts at T' will always end at T.*

The same result does not hold when neighbors are defined via NNI operations. Figure 1 provides a counterexample in the case of a 7-taxon tree. In addition, we show that, in general, there exists a sequence of characters that is compatible with exactly one tree on n leaves for which multiple local optima exist in the search space with neighbors defined by the NNI operation. To show this, we construct a sequence that is compatible with exactly one tree (and thus has a global optimum) but whose search space has multiple attraction basins. For each edge of a balanced tree, we include a character that changes only on that edge. This creates a sequence which is only compatible with the initial balanced tree. We then construct a new balanced tree where, roughly, each sibling pair has a leaf from the left subtree of the initial tree and a leaf from the right subtree of the initial tree. We then show that this construction will have the same parsimony score as all of its neighbors, leading to a search space with multiple attraction basins. The last part of the proof handles the case where the number of leaves $n \neq 2^m$ for some m.

Theorem 3 *For $n \geq 5$, there exists a sequence, $\mathcal{S}$, of binary characters on $\mathcal{L}$ that is compatible with exactly one binary tree, T, and for which there are multiple local optima for NNI search space.*

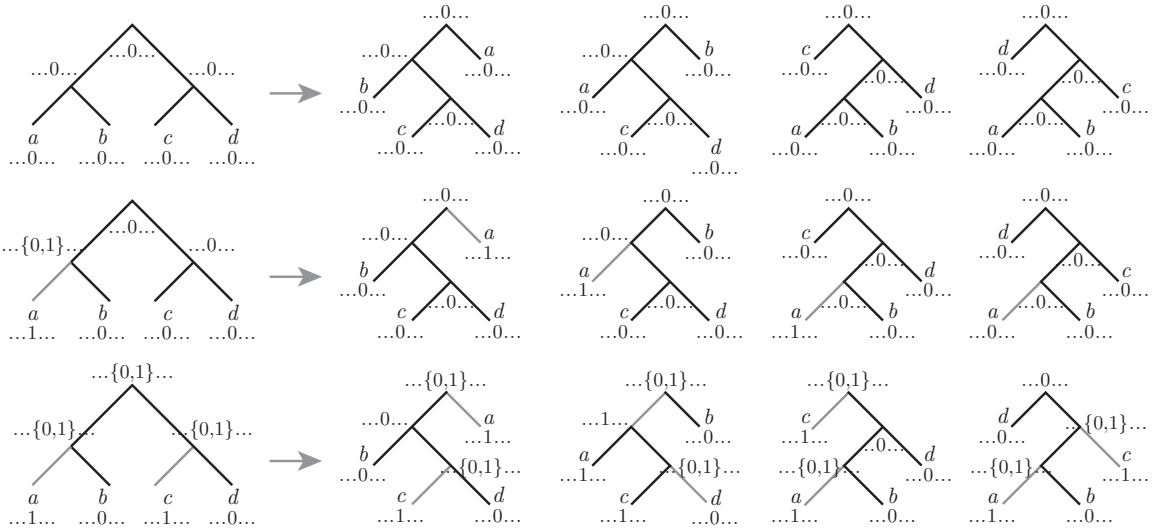

Fig. 4 Parsimony changes on a rooted subtree with four taxa: Original rooted subtrees are on the left and all possible rooted subtrees after a single NNI move are on the right. Grey edges are edges across which a single character changes, and that therefore add to the parsimony score. Note that the number of grey edges (thus the parsimony score) does not change in any row from the left to the right of the image.

Proof: The cases of $n = 5, 6, 7$ are straightforward (we include an illustration of $n = 7$ in Figure 1). For $n > 7$, we first show the result explicitly for $n = 2^m$ for $m \geq 3$ and then show how sequences can be constructed for all n that are not a power of two.

First, we set up our counterexample. Suppose we have an unrooted balanced binary tree T with 2^m leaves, for some $m \geq 3$. We can find the edge e that splits T into two identical topologies, each with 2^{m-1} leaves; call these subtrees T_1 and T_2. Note that e is incident with ρ_1, the root of T_1, and ρ_2, the root of T_2. We then give T a set of compatible character sequences $\mathcal{S}$ as follows: If T has 2^m leaves, then it has $2 \cdot 2^m - 3$ edges, thus we let $\mathcal{S} = \{\chi_1, \ldots, \chi_{2 \cdot 2^m - 3}\}$ be $2 \cdot 2^m - 3$ binary sequences. Let χ_1 be the character that is 0 on ρ_1 and 1 on ρ_2, let the next $2^m - 2$ characters in $\mathcal{S}$ be those that change across the edges of T_2, and let the final $2^m - 2$ characters be those that change across the edges of T_1. Note that assigning states to the characters in this manner means that the remaining characters on the leaves will be given 1's or 0's corresponding to the subtrees. That is, by construction, if ℓ_1 is any leaf of T_1, then it must have $\chi_1(\ell_1) = \cdots = \chi_{2^m-1}(\ell_1) = 0$ and if ℓ_2 is any leaf of T_2 then it must have $\chi_{2^m}(\ell_2) = \cdots = \chi_{2 \cdot 2^m - 3}(\ell_2) = 0$, and $\chi_1(\ell_2) = 1$.

Next, we construct another balanced binary tree T', in the same space as T, that has the same parsimony score as all of its NNI neighbors. To do this, we keep the topology of the original tree T, but we permute the leaves such that every sibling pair consists of one leaf from T_1 and one leaf from T_2. If we root T' on its center edge (creating a rooted balanced tree) and perform a first pass, then every internal vertex will have a labelling consisting entirely of zeros or unresolved character states. Since any characters that are left unresolved at the root can be chosen to be zero, this means that there is a minimal extension $\overline{\mathcal{S}}$, where every internal vertex has a labelling consisting entirely of zeros. This follows by the construction for this character, and without loss of generality, choosing a from T_1 and b from T_2, since then $\chi_i(a) = 0$ and $\chi_i(b) \in \{0, 1\}$ for $1 < i \leq 2^m - 1$, and $\chi_i(a) \in \{0, 1\}$ and $\chi_i(b) = 0$ for $2^m < i \leq 2^{m+1} - 3$.

Due to the construction of T', any NNI move that swaps subtrees (i.e. does not break any of the sibling pairs) will yield a new tree with the *same* minimal extension of $\mathcal{S}$—that is, with all internal vertices having labellings that consist only of zeros—hence the same parsimony score as T'. Thus, the only NNI moves we consider are the ones that break apart a sibling pair. Consider the subtree of T' in Figure 4. We perform the NNI move diagrammed and call this new tree T''. For any arbitrary character χ_i, we show that the parsimony score on this character is unchanged; that is, $l(\chi_i, T') = l(\chi_i, T'')$.

Suppose the leaves of this subtree are $\{a, b, c, d\}$, where a, c are leaves of T_1 and b, d are leaves of T_2. Let $i \in \{1, 2, \ldots, 2 \cdot 2^m - 3\}$. Without loss of generality, there are three possible cases:

1. $\chi_i(a) = \chi_i(b) = \chi_i(c) = \chi_i(d) = 0$

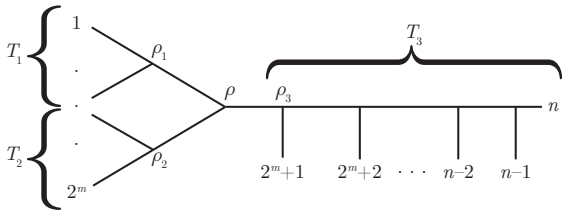

Fig. 5 Proof of Theorem 1: the topology for the case where $n > 9$ is not a power of 2. Let m be the largest number such that $2^m < n$. The tree consists of a balanced tree with 2^m leaves connected to a rooted catepillar with $n - 2^m$ leaves.

2. $\chi_i(a) = 1$, $\chi_i(b) = \chi_i(c) = \chi_i(d) = 0$
3. $\chi_i(a) = \chi_i(c) = 1$, $\chi_i(b) = \chi_i(d) = 0$

If we only look at the i^{th} character, then in the first two cases, the NNI move does not change the labellings of the internal vertices in the minimal extension, so $l(\chi_i, T') = l(\chi_i, T'')$ trivially. In the third case, it is easy to score both the original subtree and its neighbor; one finds that both have the same parsimony score (see Figure 4).

Thus, since $l(\chi_i, T') = l(\chi_i, T'')$ for any arbitrary character χ_i, we find that $l(\mathcal{S}, T') = l(\mathcal{S}, T'')$ for any NNI neighbor T'' of T', meaning that T' is in a non-trivial attraction basin with score $l(T')$. We note, that, by construction, this is the worst case for a greedy local search strategy (see the conclusion section for a more detailed discussion).

The extension to the case $n \neq 2^m$ follows from the above argument. Suppose $n \geq 9$, and $n \neq 2^m$ for any integer m. We choose a specific topology for T, as follows: Let m instead denote the maximal integer such that $2^m < n$. We construct the unrooted balanced tree of 2^m leaves as before, with congruent subtrees T_1, T_2 connected by a center edge $e = \{\rho_1, \rho_2\}$. The remaining $n - 2^m$ leaves of T are arranged in a rooted caterpillar T_3, whose root ρ_3 is attached to e, between ρ_1 and ρ_2. Label this new vertex (the one adjacent to ρ_1, ρ_2, ρ_3) as ρ' (see Figure 5).

Now, we choose the sequence $\mathcal{S}$ of compatible sequences on $\mathcal{L}$ in a similar manner as before, with the end result that the sequences of characters $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ that change across T_1, T_2, and T_3 respectively, are pairwise disjoint.

Finally, we construct a new tree T''' in the same space as T, whose neighbors have parsimony scores greater or equal to that of T'. The construction is the same as before; we rearrange the leaves of T_1 and T_2 such that the resulting tree T''' has the same topology as T, with the exception that each sibling pair in the two balanced subtrees (which we call T_1''', T_2''') consists of one leaf from T_1 and one leaf from T_2. We leave the leaves of T_3 untouched. T''' is formed by joining the roots of T_1''', T_2''', and T_3. Note that completing a first and second pass, taking any unresolved characters at the root to be zero, results in an internal labelling of T_1''', T_2''' in which all of the characters are zero, and an internal labelling of T_3 in which each edge witnesses one character change.

By the earlier argument, any NNI move that swaps leaves or subtrees within T_1''' and T_2''' will result in a new tree with equal parsimony score, and as T_3 is left in its optimal arrangement, any NNI moves that swaps leaves or subtrees within T_3 can only result in a tree with a higher parsimony score. We choose to root T''' between ρ' and ρ_2, and thus it remains to show that any NNI move that swaps subtrees or leaves between T_1''' and T_3 will either raise the parsimony score or leave it fixed. But this follows easily from the construction of the sequence of compatible characters $\mathcal{S}$, when one considers the three NNI moves that are possible: swapping a subtree of T_1''' with T_3, swapping a leaf of T_3 with T_1''', and swapping a subtree of T_3 with T_1'''. (Note that there is no option to swap a leaf of T_1''' with T_3, as $n \geq 9$ and T_1' must have at least 4 leaves, arranged in the balanced topology.) However, the last two NNI moves are equivalent, because T_3 is a caterpillar, so the leaf and the subtree of T_3 that could be swapped with T_1 will have the same labelling. It is easy to see that swapping a subtree of T_1''' with T_3 will result in a tree with the same parsimony score as T''', since any minimal internal labelling will have any internal vertices of T_1''' with a labelling of all zeros, and ρ_3 with a single character equal to 1 (the character that changed from ρ' to ρ_3)—see Figure 4, row 2). By similar arguments to the balanced case, the remaining case, that of swapping the available leaf of T_3 with the subtree T_1''', increases the parsimony score by 1.

Thus, we find that $l(\mathcal{S}, T') \leq l(\mathcal{S}, T''')$ for any NNI neighbor T''' of T', meaning that T' is an attraction basin, as desired. □

3.1 Local Optima for Simulated Sequences

While NNI-treespaces can have multiple local optima even for sequences that are compatible with exactly one tree, SPR-treespaces have a single local optimum. Following the work of Atteson on the Neighbor Joining method [2,34], we give bounds on how far sequences can be from compatible and still yield a search space with a single optimum under the SPR metric. We then use this bound to determine, with high probability, when simulated sequences yield an SPR-search space with a single local optimum. To simplify the statement of the theorem, we use the "nearly additive" concept [2] that captures when a set of sequences is extremely close to a set that is compatible with exactly one tree: A distance matrix D is *nearly additive* within a factor of c with respect to a weighted binary tree (T, w) if

$$||D - D_T||_\infty < \frac{1}{c} \min_{e \in E(T)} w(e)$$

We can extend this notion to a set of sequences of characters, $\mathcal{S}$, by defining the distance matrix, $D_{T,\mathcal{S}}$ to have entries, $d_{i,j}$, that are the sum of the changing numbers on the path from i to j with respect to an optimal extension of $\mathcal{S}$ to the vertices of T. We say that $\mathcal{S}$ is nearly additive for T if there exists a weighting w on T such that $D_{T,\mathcal{S}}$ is nearly additive with respect to (T, w).

We can extend Theorem 1 to give similar results for nearly additive sequences of characters with factor of 3. Roughly, the factor of 3 results from the three edges that surround the subtree moved by the SPR move. Each one could contribute less than $\frac{x}{3}$ to the overall bound yielding the desired decrease of x.

Theorem 4 *Let T be a binary tree with leaf set $\mathcal{L}$. Let $\mathcal{S}$ be a sequence of characters on $\mathcal{L}$ that is nearly additive with a factor of 3 for T. Suppose T_L, T_R, (T_L, T_R) are subtrees of T. If T' is any other binary tree which contains T_L and T_R as subtrees but does not contain the subtree (T_L, T_R), then either of the two SPR moves that re-form the subtree (T_L, T_R) will lower the parsimony score of T'.*

Proof: Let T be a binary tree on n taxa and let $\mathcal{S}$ be a set of sequences that is nearly additive for T. Then there exists a weighting, w on T such that

$$||D_{T,\mathcal{S}} - D_T|| < \min_{e \in E(T)} \frac{w(e)}{3}$$

Suppose T_L, T_R, (T_L, T_R) are subtrees of T. Assume that binary tree T' contains T_L and T_R as subtrees but does not contain the subtree (T_L, T_R). Let ρ_L be the root of the subtree T_L in T, ρ_R be the root of the subtree T_R in T, p be the vertex connected to ρ_L and ρ_R in T, and q be the remaining neighboring vertex of p in T. Set $x = \min_{e \in E(T)} w(e)$. By hypothesis, the difference between any entries in $D_{T,\mathcal{S}}$ and D_T is less than $\frac{x}{3}$. Thus, $|w(\{\rho_L, p\}) - ch(\mathcal{S}, \{\rho_L, p\})| < \frac{x}{3}$ and $|w(\{p, q\}) - ch(\mathcal{S}, \{p, q\})| < \frac{x}{3}$. By similar argument to Theorem 1, the removal of the subtree, T_L from T' lowers the score by more than $ch(\mathcal{S}, \{\rho_L, p\}) + ch(\mathcal{S}, \{p, q\})$ which is larger than $w(\{\rho_L, p\}) - \frac{x}{3} + w(\{p, q\}) - \frac{x}{3}$. Joining T_L next to T_R will raise the score of the resulting tree by less than $w(\{\rho_L, p\}) + \frac{x}{3}$. Thus, the score of the resulting tree is lower by more than: $w(\{\rho_L, p\}) - \frac{2x}{3} + w(\{p, q\}) - (w(\{\rho_L, p\}) + \frac{x}{3}) = w(\{p, q\}) - x \geq 0$ □

As a corollary, Atteson's analysis [2] for Neighbor joining applies, yielding a lower theoretical, though quite large bound on sufficient sequence length to guarantee convergence:

Corollary 3 *[2] There is a single local optimum for SPR-treespace with probability at least $1 - \delta$ if the sequences generated under the Cavender-Farris-Neyman model for weighted n-leaf tree (T, w) have sequence length at least:*

$$\frac{8 \ln(n^2/\delta)}{(1 - e^{-x/3})^2} e^{4M}$$

where x is the shortest branch and M is longest distance between leaves of (T, w).

4 Experimental Results

We examine the local search paradigm with both empirical and simulated sequence data. For empirical data, we analyzed the number of attraction basins for the 106 genes of the yeast dataset of Rokas *et al.* [33]. To examine the effects of the mutation rate on the smoothness of the search space, we simulated data sequences, varying the parameters. The results are detailed in the subsections below.

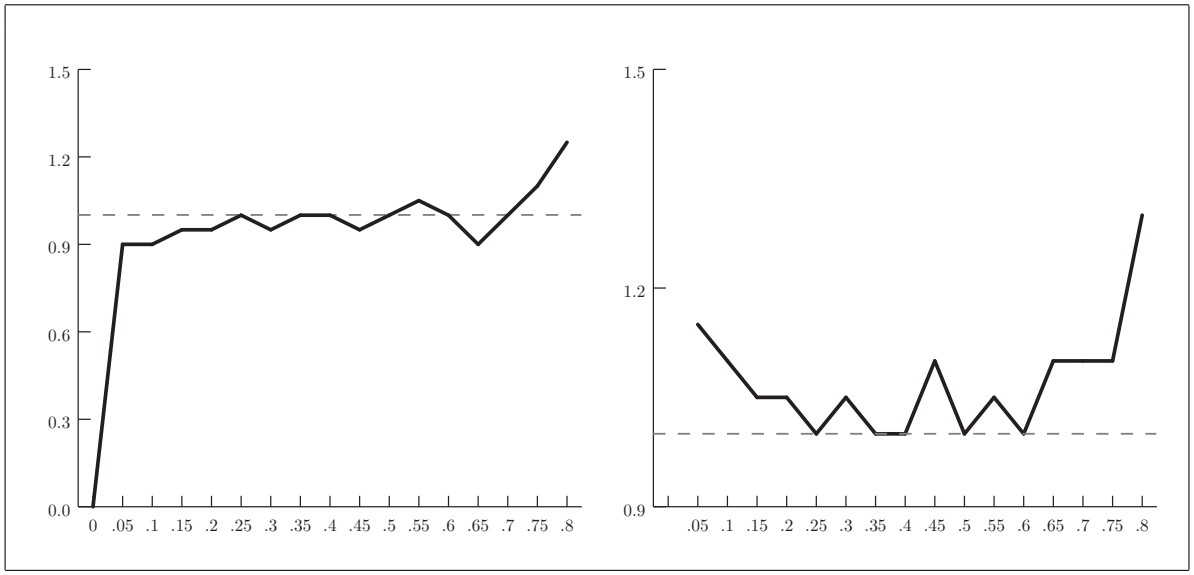

Fig. 6 Graph of average number of optima (vertical axis) versus edge weight, w, (horizontal axis) for 8-taxon SPR space. 20 simulated genes created on two topologies were tested for local optima. *Left:* attraction basins with constant value were counted as local optima and genes with multiple globally optimal trees were given a score of 0 (see discussion in text). *Right:* global optima and nonglobal local optima were summed equally, but the leftmost point was removed, as it dominated the graph. All genes with $w = 0.001$ had multiple globally optimal trees, and no trees with $w < 0.55$ had multiple nonglobal local optima: all optima were global.

4.1 Empirical Results on Yeast Data Set

Rokas *et al.* [33] systematically studied the full genomes of seven *Saccharomyces* species (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii* and *S. kluyveri*) as well as an outgroup: the fungus species, *Candida albicans*. They focus on 106 genes distributed on all chromosomes of S. cervisiae and totaling 127,026 nucleotides. Their analyses on the separate genes yielded 20 different topologies with varying levels of support, while the concatenated analysis yielded a single tree with maximal bootstrap values on every branch for the methods of analysis (maximum parsimony and maximum likelihood searches using PAUP [38]). Money and Whelan [28] examined the search space for the maximum likelihood criteria under both the NNI and SPR neighborhoods. For SPR, they found that 101 of 106 genes using Jukes Cantor paramters for maximum likelihood and 105 of 106 under the GTR+Γ parameters had single SPR attraction basin. We explored the maximum-parsimony optimality criterion on the yeast data set [33] and found similar empirical results for the maximum parsimony criteria to those for the maximum likelihood criteria of Money and Whelan [28]: most—75 of 106 of the genes—yield SPR-spaces with a single optimum. An additional 15 of the genes have a search space where multiple trees achieve the optimal score and those trees form an attraction basin.We further performed a combined analysis of the data, scoring all possible 8-taxon trees with a concatenated sequence of all 106 genes. For this combined analysis, the SPR search space has a single optimum, echoing the results of Rokas *et al.* [33].

4.2 Empirical Results on Simulated Sequences

We generated simulated data under the more general 4-state model of evolution due to Jukes and Cantor and evaluated the number of optima found for SPR-space. By generating aligned data on a small space, we were able to score the neighborhoods of every tree searching for local optima. Using simulated data allowed us to vary the amount of change in each tree, giving us the ability to examine the relationship between the sequence-length-to-edge-weight ratio and the topography of the search space.

As the size of $n = 8$ treespace is fairly small, we were able to evaluate the parsimony score on all possible trees on $n = 8$. We therefore simulated sequences on 8-leaf trees, and plotted the edge weight, w vs. the number of local optima. For $w \in (0.001, 0.05, 0.1, 0.15, \ldots, 0.8)$ we created two trees, a caterpillar and a balanced tree. We then evolved sequences down those trees using DAWG[6]. DAWG takes a tree

with specified edge weights, creates an initial sequence, then evolves the sequence down the tree, using the weights of the edges in the tree to nondeterministically determine the number of character changes at each step. The character sequences were of length 1000, had no gaps, and used the Jukes-Cantor model of evolution. The DAWG commands used were:
`Treescale = 1.0, Length = 1000, Model="GTR", Freqs = {0.25, 0.25, 0.25, 0.25},`
`Params = {1.0, 1.0, 1.0, 1.0, 1.0, 1.0}, GapParams = {0, 0, 0}`. The meanings of these parameters are: *a*) *Treescale* how much to scale the given edge weights; *b*) *Length* the length of the initial sequence; *c*) *Model* the model of evolution to use. Here, generalized time reversible is specified; *d*) *Freqs* the frequencies that each base, A, C, T, G appears; *e*) *Params* the substitution rates, {A-C, A-G, A-T, C-G, C-T, G-T}. These rates are even, which is equivalent to the Jukes-Cantor model; and *f*) *GapParams* in this case we did not specify a model of evolution with gaps.

For each tree we ran 10 simulations, giving us a total of 340 simulated genes (see Figure 6). In order to highlight the "sweet spot," wherein hill-climbing search will both work and is likely to discover a significant number of the optimal trees, we distinguished between global optima, which are necessarily also local optima, and nonglobal local optima, which act as spurious basins of attraction. As no genes had both multiple global and spurious basins of attraction, we considered a gene with multiple global optima to have 0 nonglobal local optima, but a gene with a single global optimum to have one local optimum. Next, we simply summed the number of local optima. As global optima are also local optima, the counts of the two were summed.

For trees with $w = 0.001$, every tree had multiple global optima. For other values of w less than 0.55, there were either single or multiple global optima. The average number of nonglobal local optima did not exceed 1 until $w > 0.7$. Note that in the case of multiple global optima, hill-climbing search will still work, but will find only one of multiple optimal trees.

5 Conclusions

We rigorously show what has been hinted at by empirical studies [8,25,28]: NNI searches can have many more local optima than SPR even for "well-behaved" data. Further, for perfect data and data simulated under a range of parameters of the Jukes-Cantor model of evolution, the SPR search space has a single local optimum, and simple hill-climbing searches are guaranteed to find the optimal tree. Computing SPR neighborhoods is costly, so many approaches use heuristics to capture the neighborhood efficiently [19,36,39]. Our results show that tree rearrangements based solely on NNI moves do not approximate SPR moves well, due to the multiple optima of NNI. Caceres *et al.* [5] showed that enumerating an SPR neighborhood via efficient-to-compute NNI operations is computationally expensive. While searching with neighborhoods based on NNI rearrangements (such as the "lazy SPR" operation [36]) are attractive due to their linear size, our results show that the resulting search space becomes much more complex for even the simplest case of perfect data. Instead of a "smooth" basin, the space becomes quite "rugged" with many plateaus of non-global optima (explored empirically by Bastert *et al.* [3] for NNI search spaces). This suggests that the trade-off between efficiently computed search neighborhoods should be weighed against the increase in non-global optima of the resulting search spaces.

We examine the efficiency of hill climbing on data simulated under the Jukes-Cantor model of evolution, echoing the work of Atteson on the efficiency of Neighbor Joining [2]. Both his and our work show that sequences that are close to the model tree (the difference in leaf distances is less than a third of the shortest branch in the model tree) will perform efficiently, and follow work of Moulton and Steel [29] and Farach and Kannan [12] of identifying how much the distance matrix can be perturbed and still yield the correct tree [35, §7.7]. Further interesting questions about the efficiency of searching treespace include the behavior of hill-climbing search on data simulated by more general models of evolution (such as General Time Reversible (GTR) Markov models), as well as what happens to the search space when parsimony is replaced by the maximum-likelihood criterion [16]. The latter poses interesting challenges to proving results rigorously, given that branch weights must be chosen for each tree topology to calculate the score, and it is possible to have several different branch weight yield local optima for a single tree topology [37].

We evaluated the effectiveness of hill-climbing search with SPR neighborhoods on data simulated under the Jukes-Cantor model of evolution. Hill-climbing search did surprisingly well, as almost all search spaces yielded single local optima. The exceptions occurred when the rate of change was exceedingly small

($w \leq 0.05$) and the resulting sequences yielded taxa with identical character sequences, making those taxa indistinguishable. Similarly, when the change along branches was exceedingly high ($w \geq 0.55$), the sequences had so many mutations as to appear random. For the majority of the values ($0.05 < w < 0.55$), a larger range than what we showed will perform well with high probability), there was a single local optimum, thus hill-climbing succeeded. While our sample space is small ($n = 8$) to allow for exhaustive evaluation, it is likely that we will see a similar phenomenon of ease of search for simulated data on larger size data sets, exceeding what is predicted by the theoretical bounds.

We show that even for the simplest case of character sequences displayed by a single tree, the choice of which tree-rearrangement operation employed to explore the space of trees can have profound effect on the complexity and success of the search algorithm. While our theoretical results focus on the worst case for a neighborhood-based search that greedily progresses, this behavior is also seen for empirical data both for the maximum parsimony criteria (as shown here) and for the maximum likelihood criteria [28]. The NNI search space can have multiple non-global optima and multiple attraction basins that leads to failed searches. While for the SPR search space, perfect data and simulated data that is nearly-additive have hill climbing searches that always find the optimal phylogenetic tree. While the simplicity of the computing NNI neighbors makes it tempting to use, the resulting search space is quite "rugged" leading to many searches that are "trapped" at non-global optima. On a positive note, our work on SPR search spaces suggest that the space is "smoother" but at the expense of greater computation to computing SPR neighbors. The proof of the SPR result suggests that search heuristics that allow moves that can join arbitrary subtrees (like SPR or TBR) or allow a random search (to escape non-global optima of NNI) yield search spaces with fewer non-global optima to derail searches.

6 Acknowledgements

The authors would like to thank Mike Charleston, Kevaughn Gordon, Barbara Holland, and Ward Wheeler for insightful and helpful discussion and the anonymous referees and Associate Editor Mike Steel for their comments which greatly improved this paper. We would also like to thank the American Museum of Natural History for hosting the first author as a visiting summer student. Partial funding was provided by the US National Science Foundation (#0920920 to KS) and the Simons Foundation.

References

1. Benjamin L. Allen and Mike Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5(1):1–15, 2001.
2. K. Atteson. The performance of the neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25:251–278, 1999.
3. Oliver Bastert, Dan Rockmore, Peter F. Stadler, and Gottfried Tinhofer. Landscapes on spaces of trees. *Applied Mathematics and Computation*, 131(2–3):439 – 459, 2002.
4. Magnus Bordewich, Olivier Gascuel, Katharina T. Huber, and Vincent Moulton. Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 6(1):110–117, 2009.
5. Alan Joseph J. Caceres, Juan Castillo, Jinnie Lee, and Katherine St. John. Walks on SPR neighborhoods. *IEEE/ACM Transactions on Compututational Biology and Bioinformatics*, 10(1):236–239, January 2013.
6. Reed A Cartwright. Dna assembly with gaps (dawg): simulating sequence evolution. *Bioinformatics*, 21(Suppl 3):iii31–iii38, 2005.
7. James A Cavender. Taxonomy with confidence. *Mathematical Biosciences*, 40(3):271–280, 1978.
8. Michael A. Charleston. Toward a characterization of landscapes of combinatorial optimization problems, with special attention to the phylogeny problem. *Journal of Computational Biology*, 2(3):439–450, 1995.
9. Stephen Chen, James Montgomery, Antonio Bolufé-Röhler, and Yasser Gonzalez-Fernandez. Invited paper: A review of thresheld convergence. *GECONTEC: Revista Internacional de Gestión del Conocimiento y la Tecnología*, 3(1):1–13, 2015.
10. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to algorithms*. MIT Press, Cambridge, MA, second edition, 2001.
11. Andreas Dress and Michael Krüger. Parsimonious phylogenetic trees in metric spaces and simulated annealing. *Advances in Applied Mathematics*, 8(1):8–37, 1987.
12. Martin Farach and Sampath Kannan. Efficient algorithms for inverting evolution. *Journal of the ACM (JACM)*, 46(4):437–449, 1999.
13. J. S. Farris. A method for computing wagner trees. *Systematic Zoology*, 19:83–92, 1970.
14. James S. Farris. A probability model for inferring evolutionary trees. *Systematic Biology*, 22(3):250–256, 1973.
15. Walter M. Fitch. Towards defining the course of evolution: Minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.
16. Walter M Fitch, Emanuel Margoliash, et al. Construction of phylogenetic trees. *Science*, 155(760):279–284, 1967.

17. L.R. Foulds and R.L. Graham. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics*, 3(1):43–49, 1982.
18. Josselin Garnier and Leila Kallel. Efficiency of local search with multiple local optima. *SIAM Journal on Discrete Mathematics*, 15(1):122–141, 2001.
19. Pablo A. Goloboff, James S. Farris, and Kevin C. Nixon. TNT, a free program for phylogenetic analysis. *Cladistics*, 24(5):774–786, 2008.
20. Jotun Hein. Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical biosciences*, 98(2):185–200, 1990.
21. M.D. Hendy, L.R. Foulds, and D. Penny. Proving phylogenetic trees minimal with l-clustering and set partitioning. *Mathematical Biosciences*, 51(1):71–88, 1980.
22. M.D. Hendy and David Penny. Branch and bound algorithms to determine minimal evolutionary trees. *Mathematical Biosciences*, 59(2):277 – 290, 1982.
23. D.M. Hillis, B.K. Mable, and C. Moritz. *Molecular Systematics*. Sinauer Assoc., Sunderland, Mass., 1996.
24. T.H. Jukes, C.R. Cantor, and H.N. Munro. Mammalian protein metabolism. *Evolution of protein molecules*, 3:21–132, 1969.
25. B. Kirkup and Junhyong Kim. From rolling hills to jagged mountains: Scaling of heuristic searches for phylogenetic estimation. Molecular Biology and Evolution (In revision), 2003.
26. Ming Li, John Tromp, and Louxin 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.
27. David R. Maddison. The discovery and importance of multiple islands of most-parsimonious trees. *Systematic Zoology*, 40(3):315–328, September 1991.
28. Daniel Money and Simon Whelan. Characterizing the phylogenetic tree-search problem. *Systematic Biology*, 61(2):228–239, 2012.
29. Vincent Moulton and Mike Steel. Retractions of finite distance functions onto tree metrics. *Discrete Applied Mathematics*, 91(1):215–233, 1999.
30. Jerzy Neyman. Molecular studies of evolution: a source of novel statistical problems. *Statistical decision theory and related topics*, pages 1–27, 1971.
31. Kevin C. Nixon. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15(4):407–414, 1999.
32. D.F. Robinson. Comparison of labeled trees with valency three. *Journal of Combinatorial Theory, Series B*, 11(2):105–119, 1971.
33. Antonis Rokas, Barry L. Williams, Nicole King, and Sean B. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425:798–804, 2003.
34. N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.
35. Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.
36. Alexandros Stamatakis, Filip Blagojevic, Christos D. Antonopoulos, and Dimitrios S. Nikolopoulos. Exploring new search algorithms and hardware for phylogenetics: RAxML meets the IBM cell. *Journal of VLSI Signal Processing Systems*, 48(3):271–286, 2007.
37. M.A. Steel. The maximum likelihood point for a phylogenetic tree is not unique. *Systematic Biology*, 43(4):560–564, 1994.
38. D.L. Swofford. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4*. Sinauer Associates, Sunderland, Massachusetts, 2002.
39. Andrés Varón, Le Sy Vinh, and Ward C. Wheeler. POY version 4: phylogenetic analysis using dynamic homologies. *Cladistics*, 26(1):72–85, 2010.
40. Ward C Wheeler. *Systematics: A Course of Lectures*. Wiley-Blackwell, 2012.