Detecting Highways of Horizontal Gene Transfer

Mukul S. Bansal[1], J. Peter Gogarten[2], and Ron Shamir[1]

[1] The Blavatnik School of Computer Science, Tel-Aviv University, Israel
{bansal, rshamir}@tau.ac.il
[2] Department of Molecular and Cell Biology, University of Connecticut, Storrs, USA
gogarten@uconn.edu

Abstract. In a horizontal gene transfer (HGT) event a gene is transferred between two species that do not share an ancestor-descendant relationship. Typically, no more than a few genes are horizontally transferred between any two species. However, several studies identified pairs of species between which many different genes were horizontally transferred. Such a pair is said to be linked by a *highway of gene sharing*. We present a method for inferring such highways. Our method is based on the fact that the evolutionary histories of horizontally transferred genes disagree with the corresponding species phylogeny. Specifically, given a set of gene trees and a trusted rooted species tree, each gene tree is first decomposed into its constituent quartet trees and the quartets that are inconsistent with the species tree are identified. Our method finds a pair of species such that a highway between them explains the largest (normalized) fraction of inconsistent quartets. For a problem on n species, our method requires $O(n^4)$ time, which is optimal with respect to the quartets input size. An application of our method to a dataset of 1128 genes from 11 cyanobacterial species, as well as to simulated datasets, illustrates the efficacy of our method.

1 Introduction

Horizontal gene transfer (HGT) (also called lateral gene transfer) is an evolutionary process in which genes are transferred between two organisms that do not share an ancestor-descendant relationship. HGT plays an important role in bacterial evolution by allowing them to transfer genes across species boundaries. This transfer of genes between divergent organisms first became a research focus when the transfer of antibiotic resistance genes was discovered [1, 2]. Microbiologists soon realized that the sharing of genes between unrelated species resulted in evolutionary patterns very different from those found in multi-cellular animals. Since then, the problem of detecting horizontally transferred genes has been extensively studied; see, for example, [3] for a review.

An important problem in understanding microbial evolution is to infer the HGT events (i.e., the donor and recipient of each HGT) that occurred during the evolution of a set of species. This problem is generally solved in a comparative-genomics framework by employing a parsimony criterion, based on the observation that the evolutionary history of horizontally transferred genes does not agree with the evolutionary history of the corresponding set of species. This is illustrated in Fig. 1(b). More formally, given a gene tree and a species tree, the *HGT inference problem* is to find the minimum number of HGT events that can explain the incongruence of the gene tree with the species tree.

The HGT inference problem is known to be NP-hard [4, 5] and, along with some of its variants, has been extensively studied [5–12].

In general, one expects at most a few genes to have been horizontally transferred between any given pair of species. However, Beiko et al. [9] demonstrated that some pairs of species portray a multitude of horizontal gene transfer events. Such pairs are said to be connected by a *highway of gene sharing* [9]. Highways of gene sharing point towards major events in evolutionary history; well corroborated examples of this phenomenon are the uptake of endosymbionts into the eukaryotic host, and the many genes transferred from the symbiont to the hosts nuclear genome [13]. Recent proposals for evolutionary events that may be reflected in highways of gene sharing are the role of Chlamydiae in establishing the primary plastid in the Archaeplastida (red and green algae, plants and glaucocystophytes) [14], and the evolution of double membrane bacteria through an endosymbiosis between clostridia and actinobacteria [15]. Detecting these highways of gene sharing is thus an important biological problem and is crucial for inferring past symbiotic associations that shaped the evolution of organisms.

Given a rooted species tree, any two species (nodes) in it that are not related by an ancestor-descendant relationship define a *horizontal edge* connecting those two nodes. Any HGT event must take place along a horizontal edge in one of its two directions (see Fig. 1(a)). A horizontal edge along which an unusually large number of HGT events have taken place (say 10% of the genes) will be called a *highway of gene sharing* or simply a *highway*. The only existing method for detecting highways is the one employed originally by Beiko et al. [9]. That method takes as input a species tree and a set of gene (protein) trees, and computes, for each gene tree, the HGT events affecting that gene on the species tree. This is done by solving the HGT inference problem for each gene tree. The HGT events that are inferred in the HGT scenarios for a significant fraction of the gene trees are postulated as the highways. However, this approach suffers from several serious drawbacks. First, the HGT inference problem is NP-hard, and thus, difficult to solve exactly (and must often be solved using heuristics). Second, there may be multiple (in fact, exponentially many) alternative optimal solutions to the HGT inference problem [10]. And third, when the rate of HGT is relatively high, there is little reason to expect that the number of HGT events should be parsimonious; i.e., the HGT inference problem, even if solved exactly and yielding only one optimal solution, may not infer the actual HGT events. In this work we propose an alternative approach to detecting highways that does not rely on inferring individual HGT events. Moreover, our formulation allows exact solution of the problem in polynomial time. Our method thus avoids all of the aforementioned pitfalls.

As in [9], the input to our method is a trusted rooted species tree for some set of species, and a set of gene trees on genes taken from those species. Since it is often difficult to accurately root gene trees, we assume that the input gene trees are unrooted. Our method is based on the observation that highways, by definition, affect the topologies of many gene trees. Thus, the idea is to combine the phylogenetic signals for HGT events from all the gene trees and use the combined signal to infer the highways, thereby avoiding the need to infer individual HGT events. We achieve this by employing a quartet decomposition of the gene trees. In particular, our method decomposes each gene tree into its constituent set of quartet trees and combines the quartet trees from all the gene

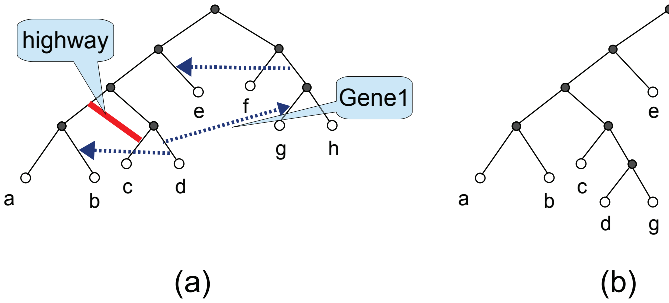


Fig. 1. Horizontal gene transfers and highways. (a) A species tree depicting three HGT events (dotted arcs) and a highway (bold red horizontal edge). The highway represents many individual HGT events all occurring between the same two (present-day or ancestral) species. (b) The corresponding gene tree for Gene1. Because of the HGT of Gene1 from species *d* into species *g*, the copy of that gene in *g* is most closely related to the one in *d*. Therefore, in the tree for Gene1, the species *g* appears next to *d*. (Here we assume that Gene1 was not transferred on the highway.)

trees to obtain a single weighted set of quartet trees. The intuition is that quartet trees that disagree with the species tree may indicate HGT events and thus the collective evidence from all quartet trees could pinpoint possible highways. The combined set of quartet trees is then analyzed against the given species tree to infer the highways of gene sharing. Decomposing the gene trees into quartet trees allows us to cleanly merge the phylogenetic signals for HGT events from all the different gene trees into a single summary signal, from which exact and efficient inference of the highways is possible.

To find highways, our method iteratively finds a horizontal edge that explains the largest fraction of inconsistent quartet trees. Essentially, for each (weighted) quartet tree inconsistent with the species tree, we identify the horizontal edges that can explain it by an HGT event (in either direction) along them. The horizontal edge that explains the most (normalized) inconsistency is proposed as a highway. (Normalization is needed since the structure of the species tree and the location of the horizontal edge in it influence the number of inconsistent quartet trees that edge may explain.) We give a dynamic programming algorithm that, given the weighted set of quartet trees, finds the best highway in $O(n^4)$ time. Since there may be $\Omega(n^4)$ input quartet trees, our algorithm is asymptotically optimal with respect to that input. In contrast, a naïve enumeration algorithm would require $O(n^6)$ time. Our efficient algorithms allow our method to be applied to fairly large datasets; for example, we can analyze a dataset of 1000 gene trees with 200 taxa within a day on a personal computer. We demonstrate the utility of our method on simulated data as well as on a dataset of 1128 genes from 11 cyanobacterial species [16], where its results match prior biological observations. For lack of space, proofs and some algorithmic details are omitted from this manuscript.

2 Basic Definitions and Preliminaries

Given a rooted or unrooted tree T, we denote its node set, edge set, and leaf set by $V(T)$, $E(T)$, and $Le(T)$ respectively. For the remainder of this paragraph, let T denote a rooted tree. Given T, the root node of T is denoted by *rt*(T). Given a node $v \in V(T)$, we denote the parent of v by *pa*$_T(v)$, its set of children by *Ch*$_T(v)$, and the subtree of T rooted at v by $T(v)$. We define $\leq_T$ to be the partial order on $V(T)$ where $u \leq_T v$ if v is a node on the path between *rt*(T) and u. Given a non-empty subset $L \subseteq$ *Le*(T), we denote by *lca*$_T(L)$ the least common ancestor (LCA) of all the leaves in L in tree T. Given a rooted tree T, a *horizontal edge* on T is a pair of nodes $\{u, v\}$, where $u, v \in V(T)$, such that $u, v \neq$ *rt*(T), $u \not\leq v$, $v \not\leq u$, and *pa*$_T(u) \neq$ *pa*$_T(v)$. We denote by $H(T)$ the set of all horizontal edges on T. Horizontal edges represent potential horizontal gene transfer events; the (directed) horizontal edge (u, v) represents the HGT event that transfers genetic information from the edge (*pa*$_T(u), u)$ to (*pa*$_T(v), v)$. Thus, the horizontal edge $\{u, v\}$ represents the HGT events (u, v) and (v, u). Also note that, while any particular HGT event is directional, we address the problem in which horizontal edges are undirected because highways can be responsible for transfer of genetic material in both directions. Throughout this work the term tree refers to a binary tree.

Our formulation and solution to the highway detection problem rely on the concept of quartets and quartet trees. A *quartet* is a four-element subset of some leaf set and a *quartet tree* is an unrooted tree whose leaf set is a quartet. The quartet tree with leaf set $\{a, b, c, d\}$ is denoted by $ab|cd$ if the path from a to b does not intersect the path from c to d. Given a rooted or unrooted tree T, let X be a subset of *Le*(T) and let $T[X]$ denote the minimal subtree of T having X as its leaf set. We define the *restriction* of T to X, denoted $T|X$, to be the unrooted tree obtained from $T[X]$ by suppressing all degree-two nodes (including the root, if T is rooted). We say that a quartet tree Q is *consistent* with a tree T if $Q = T|$ *Le*(Q), otherwise Q is *inconsistent* with T. Observe that, given any T and any quartet $X = \{a, b, c, d\}$ from *Le*(T), X induces exactly one quartet tree in T, that is, the quartet tree $T|X$. Also observe that this quartet tree must have one of three possible topologies: $ab|cd$, $ac|bd$, or $ad|bc$.

3 Detecting Highways

Our goal is to detect the highways of gene sharing in the evolutionary history of a set of species $\mathbb{S}$. To that end, we are given a set of unrooted gene trees $\{T_1, \ldots, T_m\}$, and a rooted species tree S showing the evolutionary history of $\mathbb{S}$. Thus, *Le*$(S) = \mathbb{S}$, and *Le*$(T_i) \subseteq \mathbb{S}$ for $1 \leq i \leq m$. The idea is to infer the highways by inspecting the differences in the topologies of the gene trees compared to the species tree. The *highway detection problem* can thus be stated as follows: Given a species tree S and a collection of gene trees, find the horizontal edges on S that correspond to highways.

Throughout this manuscript, S denotes the given species tree, and n denotes the number of species in the analysis, i.e., $n = |$ *Le*$(S)|$.

Our solution to the highway detection problem is based on decomposing each input gene tree T into its constituent set of $\binom{|Le(T)|}{4}$ quartet trees. To understand the intuition behind using quartet trees, consider the scenario depicted in Fig. 2. The tree on the left is a species tree on six species, along with two HGT events of two different genes. Consider the HGT event (C, E) that transfers Gene1. This HGT event causes the topology of the corresponding gene tree to deviate from the topology of the species tree. Essentially, according to the standard subtree transfer model of horizontal gene

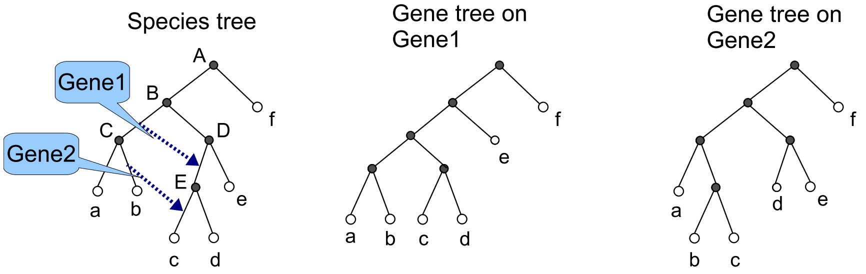


Fig. 2. The tree on the left is a species tree showing the evolutionary history of a set of six species. Two HGT events (C, E) and (b, c), shown by the dotted arcs, are also depicted on this species tree. The two other trees show the evolutionary histories of Gene1 and Gene2.

transfer (see, e.g., [17, 9, 8]), this HGT event causes the subtree rooted at node E to be pruned and then regrafted along the edge (B, C), as shown in the figure. Let us decompose both trees into their constituent set of quartet trees: Each tree generates $\binom{6}{4} = 15$ quartet trees. Note that four of the fifteen quartets induce different quartet trees in the two trees; in the gene tree, these appear as $ac|ef$, $ad|ef$, $bc|ef$ and $bd|ef$. In general, different HGT events produce gene trees with different sets of inconsistent quartet trees. Thus, given the species tree, and the set of the four inconsistent quartet trees from the gene tree on Gene1, we could have inferred the HGT event (C, E) that affected Gene1.

3.1 The Method in Detail

Our method proceeds iteratively, inferring one highway per iteration, as follows.

Step 1: Decompose each input gene tree T into its constituent set of $\binom{|Le(T)|}{4}$ quartet trees, and combine the quartet trees from the different gene trees into a single weighted set, Φ, of quartet trees.

Step 2: Remove from Φ all those quartet trees that are consistent with S.

Step 3: Compute the HGT score of each edge in $H(S)$. This HGT score for an edge is computed based on Φ, and is explained in detail below.

Step 4: Select the highest scoring horizontal edge as a highway.

Step 5: Remove from Φ all those quartet trees that are explained by the proposed highway, and go to Step 3 to start the next iteration.

The (raw) HGT score of a horizontal edge is simply the total weight of the quartet trees from Φ that are explained by a HGT along that edge (in either direction). Thus, this raw score of a horizontal edge captures the number of quartet trees from the input gene trees that support horizontal gene transfer along that edge. However, not all horizontal gene transfers affect the same number of quartet trees. Consider the example shown in Fig. 2. As seen previously, the HGT event (C, E) causes four of the quartet trees in the corresponding gene tree to become inconsistent. Consider the HGT event (b, c) that transfers Gene2. This HGT event causes ten of the quartet trees in the gene tree built on Gene2 (shown on the right in Fig. 2) to become inconsistent; these are $ad|bc$, $ae|bc$,

$af|bc$, $ac|de$, $ac|df$, $ac|ef$, $bc|de$, $bc|df$, $bc|ef$ and $de|cf$. Thus, considering only the raw scores of the horizontal edges would lead to overestimation of the quantity of HGT along certain horizontal edges and underestimation of this quantity for other horizontal edges, leading to incorrect inference of highways.

To overcome this bias we modify the score of each horizontal edge by dividing its raw score by a normalization factor: The maximum number of distinct quartet trees that could be explained by a horizontal gene transfer (in either direction) along that edge. More precisely, let Ψ be the set of all possible quartet trees on the leaf set *Le*(S). Given a horizontal edge $\{u, v\}$, let Q_1 denote the set of quartet trees in Ψ that become consistent due to the HGT event (u, v), and let Q_2 denote the set of quartet trees in Ψ that become consistent due to the HGT event (v, u). The normalization factor for $\{u, v\}$ is defined to be $|Q_1 \cup Q_2|$. After normalization, the HGT scores of all horizontal edges can be directly compared to one another.

The number of iterations in the method can either be fixed at the beginning or, preferably, be decided on the fly, based on the distribution of the horizontal edge scores computed in the current iteration.

4 The Highway Scoring Problem

This iterative quartet based method involves four computational steps: (i) Computing the initial set of weighted quartet trees from the gene trees, (ii) removing the quartet trees that are consistent with S, (iii) computing the (normalized) HGT score of each edge in $H(S)$, and (iv) identifying and removing those quartet trees that are explained by the proposed highway. It is relatively straightforward to show (details omitted for brevity) that step (i) can be executed in $O(mn^4)$ time, where m is the number of input gene trees, and steps (ii) and (iv) can be executed in $O(n^4)$ time. The main computational challenge here is (iii), i.e., computing the (normalized) HGT score of each horizontal edge. In this section we focus on this main problem.

Given a rooted species tree S and a set Φ of weighted quartet trees (that are inconsistent with S) on the leaf set *Le*(S), the *Highway Scoring (HS) problem* is to find the (normalized) HGT score of each edge in $H(S)$.

The naïve way to solve the HS problem would be to consider each edge in $H(S)$ one-at-a-time and to check which of the quartet trees from Φ are explained by that edge. Checking whether a quartet tree is explained by a horizonal edge can be accomplished in $O(1)$ time. Since there are $\Theta(n^2)$ candidate horizontal edges and $O(n^4)$ quartet trees in Φ, the complexity of computing just the raw score of each horizontal edge is still $O(n^6)$. In this section we show that the HS problem can be solved in $O(n^4)$ time. The time complexity of our algorithm is thus optimal.

Recall that each horizontal edge actually represents two HGT events. We denote the set of all these HGT events on S by $\overrightarrow{H}(S)$. Thus, for any horizontal edge $\{u, v\} \in H(S)$, there are two HGT events (u, v) and (v, u) in $\overrightarrow{H}(S)$.

Given a horizontal edge $\{u, v\}$, if Q_1 and Q_2 denote the sets of quartet trees that are explained by the HGT events (u, v) and (v, u) respectively, then, the raw score of $\{u, v\}$ is $|Q_1 \cup Q_2|$, which is $|Q_1| + |Q_2| - |Q_1 \cap Q_2|$. First, in Section 4.1, we show how to compute the raw score of each horizontal event (i.e., how to compute $|Q_1|$ and $|Q_2|$), and then, in Section 4.2, we show how to compute $|Q_1 \cap Q_2|$ and thus obtain the raw scores of horizontal edges. In Section 4.2, we also show how to reuse these algorithms to compute the normalization factor for each horizontal edge.

4.1 Computing the Raw Scores of HGT Events

For any given quartet tree $Q \in \Phi$, there may be several HGT events from $\overrightarrow{H}(S)$ that could explain Q; we denote this set of HGT events by $\overrightarrow{H}(S, Q)$. Since S is fixed, throughout the remainder of this work we will abbreviate $H(S)$, $\overrightarrow{H}(S)$ and $\overrightarrow{H}(S, Q)$ to H, $\overrightarrow{H}$ and $\overrightarrow{H}(Q)$ respectively. Our algorithm relies on an efficient characterization of the HGT events that can explain a given quartet. This characterization appears in the next lemma; but first, we need some additional definitions and notation.

Notation and Definitions. We denote the raw score of an HGT event $(u, v) \in \overrightarrow{H}$ by $RS(u, v)$. Given any two nodes $p, q \in V(S)$, let $p \rightarrow q$ denote the path between them in S, and let $V(p \rightarrow q)$ denote the set of nodes on this path (including p and q). A *subtree-path (SP) pair* on S is a pair $\langle S(v), p \rightarrow q\rangle$, where $v, p, q \in V(S)$, such that the subtree $S(v)$ and the path $p \rightarrow q$ are node disjoint and none of the nodes in $p \rightarrow q$ is an ancestor or descendant of v. Given an SP pair $\sigma = \langle S(v), p \rightarrow q\rangle$, the set of all HGT events (u, v) from $\overrightarrow{H}$ such that $u \in S(v)$ and $v \in V(p \rightarrow q)$ is denoted by $\overrightarrow{H}(\sigma)$. Similarly, a *subtree-complement-path (SCP) pair* on S is a pair $\langle S(v), p \rightarrow q\rangle$, where $v, p, q \in V(S)$, such that $V(p \rightarrow q) \subseteq V(S(v))$. We define $\overline{V}(S(v))$ to be the set $[V(S) \setminus V(S(v))] \cup \{v\}$. Given an SCP pair $\sigma = \langle S(v), p \rightarrow q\rangle$, the set of all HGT events (u, v) from $\overrightarrow{H}$ such that $u \in \overline{V}(S(v))$ and $v \in V(p \rightarrow q)$ is denoted, as before, by $\overrightarrow{H}(\sigma)$. If σ is an SCP pair, then we say that $S(v)$ is the *subtree-complement* of σ, and it refers to the subtree of S induced by $\overline{V}(S(v))$.

Lemma 1. *Given any quartet tree $Q \in \Phi$, there exist four SP/SCP pairs, denoted $\sigma_1, \sigma_2, \sigma_3, \sigma_4$, such that $\overrightarrow{H}(Q) = \overrightarrow{H}(\sigma_1) \cup \overrightarrow{H}(\sigma_2) \cup \overrightarrow{H}(\sigma_3) \cup \overrightarrow{H}(\sigma_4)$. Moreover, the four sets $\overrightarrow{H}(\sigma_1), \overrightarrow{H}(\sigma_2), \overrightarrow{H}(\sigma_3)$ and $\overrightarrow{H}(\sigma_4)$ are pairwise disjoint.*

In fact, after an initial $O(|\mathit{Le}(S)|)$ preprocessing step, we can compute the four SP/SCP pairs for any given quartet tree in $O(1)$ time. Our algorithm performs a nested tree traversal of S. Before we begin this nested tree traversal we (i) perform a preprocessing step, which precomputes certain values on the tree S, and (ii) perform a tree decoration step during which we decorate the nodes of S with information about the four SP/SCP pairs for each quartet tree in Φ. Next we describe these two steps in detail.

The preprocessing step. The first step in the algorithm is to preprocess the tree S so that, given any two nodes from $V(S)$, we can compute their LCA within $O(1)$ time [18]. This preprocessing step also allows us to label the nodes of S in such a way that given any two nodes $u, v \in V(S)$ we can check if $v \in V(S(u))$ in $O(1)$ time. We also associate with each $v \in V(S)$ a counter, denoted by $counter_v$, initialized to zero, and a set $path_v$ initialized to be empty.

Decorating the tree. For each quartet tree $Q \in \Phi$, we identify the four SP/SCP pairs $\sigma_1 = \langle S(v_1), p_1 \rightarrow q_1\rangle$, $\sigma_2 = \langle S(v_2), p_2 \rightarrow q_2\rangle$, $\sigma_3 = \langle S(v_3), p_3 \rightarrow q_3\rangle$, and $\sigma_4 = \langle S(v_4), p_4 \rightarrow q_4\rangle$. One of the end points of the path in each of these SP/SCP pairs must be a leaf node (see proof of Lemma 1). By convention, we let the q_is, for $i \in \{1, 2, 3, 4\}$, denote these leaf nodes. We mark these four paths on S as follows: For each $i \in \{1, 2, 3, 4\}$, if σ_i is an SP pair then add the triple (Q, v_i, SP) to the sets $path_{q_i}$ and $path_{\mathit{pa}(p_i)}$; if σ_i is an SCP pair, add the triple (Q, v_i, SCP) to the sets $path_{q_i}$ and $path_{\mathit{pa}(p_i)}$. Here SP/SCP is included as a binary label to indicate the type of the pair.

The tree decoration step, described above, marks the endpoints of the four paths in the SP/SCP pairs of any quartet. Our algorithm performs a post-order traversal of S and, at each node v, calls the procedure *Augment*(v) described below. This procedure marks the corresponding subtrees/subtree-complements for all the paths that appear in the set $path_v$, and computes a value val_u at each $u \in V(S) \setminus \{\mathit{rt}(S)\}$. This value val_u is the weight of all quartet trees Q from Φ such that (i) $(Q, x, \Gamma) \in path_v$ and (ii) if Γ is SP then $u \in V(S(x))$, and, if Γ is SCP then $u \in \overline{V}(S(x))$. The reason for computing these val_u's becomes clear in the context of Lemma 2.

```
Procedure Augment$(v)$    $\{v \in V(S)\}$
for each $x \in V(S)$ do
    Set $counter_x$ to $0$.
for each triple $(Q, y, \Gamma) \in path_v$ do
    if $\Gamma$ is SP then
        Increment $counter_y$ by the weight of $Q$.
    if $\Gamma$ is SCP then
        Increment $counter_{\mathit{rt}(S)}$ by the weight of $Q$ and, decrement $counter_{y_1}$ and $counter_{y_2}$ by the weight of $Q$, where $\{y_1, y_2\} = \mathit{Ch}(y)$.
for each $u \in V(S) \setminus \{\mathit{rt}(S)\}$ do
    Set $\mathit{val}_u$ to $\sum_{x \in V(\mathit{rt}(S) \rightarrow u)} counter_x$.
```

Our algorithm is based on the following key lemma.

Lemma 2. *Suppose S has been decorated and procedure Augment(v) has been executed for some $v \in V(S)$. Consider any $(u, v) \in \overrightarrow{H}$.*

1. *If $v \in \mathit{Le}(S)$, then $RS(u, v) = \mathit{val}_u$.*
2. *If $v \notin \mathit{Le}(S)$, then $RS(u, v) = RS(u, v_1) + RS(u, v_2) - \mathit{val}_u$, where $v_1, v_2 \in \mathit{Ch}(v)$.*

Nested tree traversal. Once the pre-processing and tree decoration steps have been executed, the algorithm performs a nested tree traversal of S and computes the raw score of each HGT event from $\overrightarrow{H}$ according to Lemma 2. More formally, the algorithm proceeds as follows:

```
Algorithm ComputeScores
for each $v \in V(S)$ in a post-order traversal of $S$ do
    Perform procedure Augment$(v)$.
    for each $u \in V(S) \setminus \{\mathit{rt}(S)\}$ do
        if $(u, v)$ is a valid HGT event, i.e., $(u, v) \in \overrightarrow{H}$, then
            if $v \in Le(S)$ then
                Set $RS(u, v)$ to be $\mathit{val}_u$.
            else
                Set $RS(u, v)$ to be $RS(u, v_1) + RS(u, v_2) - \mathit{val}_u$, where $v_1, v_2 \in \mathit{Ch}(v)$.
```

The preprocessing step, tree decoration step, and Algorithm *ComputeScores* require $O(n)$, $O(\Phi)$, and $O(n^2 + |\Phi|)$ time respectively. Thus, we have the following lemma.

Lemma 3. *The raw scores of all HGT events in* $\overrightarrow{H}$ *can be computed in* $O(n^2 + |\Phi|)$ *time.*

4.2 Raw Scores of Horizontal Edges and Normalization Factors

Our goal now is to compute the raw score of each horizontal edge in H. For any edge $\{u, v\} \in H$, let its raw score be denoted by $RS\{u, v\}$. Observe that $RS\{u, v\} = RS(u, v) + RS(v, u) - \textit{common}\{u, v\}$, where $\textit{common}\{u, v\}$ is the total weight of the quartet trees that are counted in both $RS(u, v)$ and $RS(v, u)$. A variant of the algorithm described above enables us to compute the value $\textit{common}\{u, v\}$ for each horizontal edge $\{u, v\} \in H$ in $O(n^2 + |\Phi|)$ time. Thus, we can compute the raw score of each horizontal edge in $O(n^2 + |\Phi|)$ time.

Recall that the normalization factor for a horizontal edge is simply the maximum number of distinct quartet trees that could be explained by that edge. Thus, we can reuse the algorithm that computes the raw scores of horizontal edges by running it on a set that contains all the possible $3 \times \binom{n}{4}$ quartet trees, each with weight 1. Thus, we have the following theorem.

Theorem 1. *The highway scoring problem can be solved in* $O(n^4)$ *time.*

5 Experimental Analysis

Cyanobacterial dataset. We first applied our method to a dataset of 1128 genes from 11 cyanobacterial species [16]. The existence of a highway on this set of species was postulated in [16, 19] and thus this dataset serves for method validation. Each of the 1128 gene trees had at least nine of the 11 species (see [16] for further details). As the trusted species tree, shown in Fig. 3, we used the rooted tree constructed on the 16S ribosomal RNA sequence from these species [20]. To account for uncertainty in the topologies of the gene trees, for each gene tree we used only those quartet trees that were present in at least 80% of the bootstrap replicates of that gene tree [16]. Our final weighted set had 799 different quartet trees with a total weight of 214,729. The total number of inconsistent quartet trees was 469 and their total weight was 23,042. There were 118 candidate horizontal edges. Fig. 4(a) shows the histogram of the normalized scores for these horizontal edges in the first application of the algorithm. The highest scoring edge is extremely well separated from the next candidate in terms of the scores (Fig. 4(a)). It is marked in Fig. 3. A priori, it is surprising that this highway connects two different genera that are distinguished by different light harvesting machineries, but the high rate of transfer between marine *Synechococcus* and *Prochlorococcus* has been previously observed and discussed [16, 19]. The discovered highway thus matches perfectly with prior biological observations.

We performed further analysis of this dataset with the aim of discovering other novel highways. In the second iteration, our method proposes the second highway shown

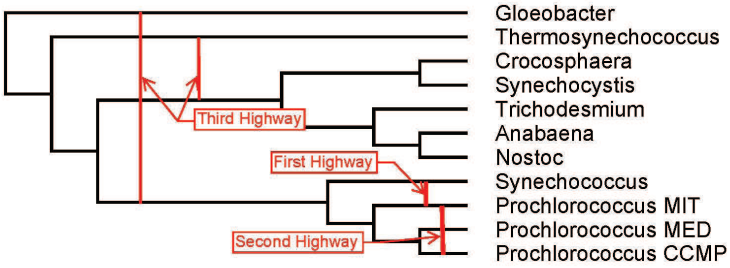
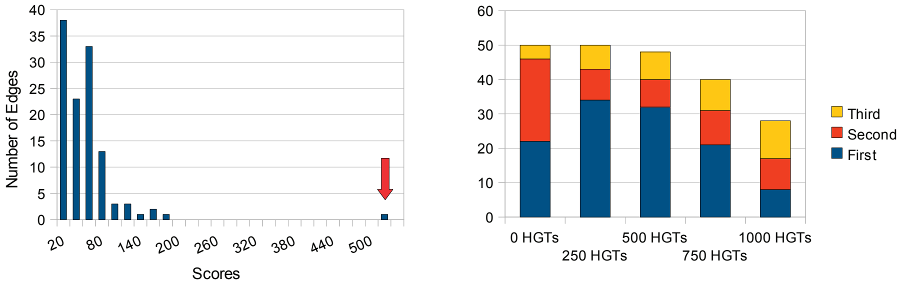


Fig. 3. The 16SrRNA tree on the 11 cyanobacterial species, with detected highways marked.



(a)

(b)

Fig. 4. Highway detection statistics. (a) histogram of edge scores for the first highway on the cyanobacterial dataset. (b) Simulation results: the number of times (out of 50) an implanted highway edge is detected in simulated datasets with varying levels of noise.

in Fig. 3. Though the normalized score of this highway is much smaller (179.4) than that of the first highway (508.6), it is significantly higher than the scores of the other edges (only two other edges have scores above 100). Like the first, this second highway also represents transfer between the small marine cyanobacteria, likely mediated by cyanophage. Further analysis also suggests the presence of a third highway (normalized score: 157.2, second-highest score: 97.7) along one of two possible horizontal edges, shown in Fig. 3. These two horizontal edges produce the same unrooted tree and are hence indistinguishable in our quartet-based model.

Simulated datasets. We used simulations to test the effect of HGT abundance on the ability to infer highways. Each simulated dataset consisted of a randomly generated species tree on 25 taxa and 1000 gene trees. For the experiment, we randomly chose a highway on the species tree, and randomly assigned 10% of the 1000 genes as having been transferred along this highway, with equal probability for each transfer direction. Next, we simulated varying levels of "noise" on the species tree in the form of random HGT events, each affecting a gene sampled at random without replacement (including the genes that were transferred on the chosen highway). We simulated noise at five different levels: 0 HGTs (i.e., no noise), 250 HGTs, 500 HGTs, 750 HGTs and 1000 HGTs. For each noise level, we created 50 different datasets (different species tree, highway, and random HGTs) and measured the number of times (out of 50) that the implanted highway is reported as one of the top three highest scoring edges by our method. As shown in Fig. 4(b), our method tends to identify the planted highway, even in datasets with high levels of noise; for instance, when there are 750 random HGTs (7.5 times the number of highway transfers) only 20% of the implanted highways were not included among the top three edges. By 1000 HGTs, performance has deteriorated. Interestingly, even when there is no noise in the data, the method does not always identify the implanted highway as its top-scoring edge. This is probably because of the way we normalize the scores. Our normalization factor is independent of direction, while the actual HGT events that take place along the highway are directed. This can cause some biases, which can make the normalized score of some nearby horizontal edges slightly higher than the score of the actual highway. Still, as the experiment demonstrates, even with relatively high levels of noise our algorithm usually brings to the top the correct highway, and further analysis of the top candidates can reveal the true highway.

6 Discussion

In this paper we addressed the problem of inferring highways of gene sharing, a fundamental problem in understanding the effects and dynamics of horizontal gene transfer, and crucial to inferring past symbiotic associations that shaped the evolution of organisms. Our new systematic approach and efficient algorithms for the highway detection problem facilitate accurate and in-depth analysis of relatively large datasets. The method detects the fingerprints of highways by looking at combined data from all the input gene trees summarized as quartet tree counts. We thus avoid the computational burden and uncertainty of inferring individual HGT events for each gene. Our experimental results demonstrate that our method is effective at detecting highways and is robust to noise in the data. We were able to identify the established highway in the cyanobacterial dataset, and our analysis identified two additional putative highways. As the experiments on the simulated data indicate, even in the presence of substantial noise our method reports the true highway among the few top-scoring edges.

While we demonstrate the effectiveness of our method, it still has some limitations. For example, if the dataset contains two highways that are closely related to one another then the method may only detect one of them (since many of the inconsistent quartet trees from one highway may also support the other highway). More generally, while the normalized scoring of the horizontal edges that we propose takes care of the variation in the number of candidate quartets of different edges, perhaps a better normalization could highlight the correct highways even more strongly. Similarly, the quartic running time is quite high and may be limiting for very large datasets. Further testing of the method in both simulations and on real datasets is also needed, and it might be instructive to compare it to alternative non-quartet-based methods. Finally, a statistical analysis of highway and HGT score distributions could provide more quantifiable significance, which we still lack.

Acknowledgements: MSB was supported in part by a postdoctoral fellowship from the Edmond J. Safra Bioinformatics program at Tel-Aviv University. JPG was supported in part by NSF grant DEB 0830024, the Edmond J. Safra Bioinformatics Program, and a fellowship from the Fulbright Program. RS was supported in part by the Israel Science Foundation (Grant 802/08) and by the Raymond and Beverly Sackler Chair in Bioinformatics.

References

1. Ochiai, K., Yamanaka, T., Kimura, K., Sawada, O.: Inheritance of drug resistance (and its transfer) between Shigella strains and Between Shigella and E.coli strains [In Japanese]. Hihon Iji Shimpor **1861** (1959) 34–46
2. Gray, G., Fitch, W.: Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. Mol Biol Evol **1**(1) (1983) 57–66
3. Zhaxybayeva, O.: Detection and Quantitative Assessment of Horizontal Gene Transfer. In: Horizontal gene Transfer: Genomes in Flux. Volume 532 of Methods in Molecular Biology. Humana Press (2009) 195–213
4. Bordewich, M., Semple, C.: On the computational complexity of the rooted subtree prune and regraft distance. Annals of combinatorics **8**(4) (2005) 409–423
5. Hallett, M.T., Lagergren, J.: Efficient algorithms for lateral gene transfer problems. In: RECOMB. (2001) 149–156
6. Boc, A., Makarenkov, V.: New efficient algorithm for detection of horizontal gene transfer events. In: WABI. (2003) 190–201
7. Nakhleh, L., Warnow, T., Linder, C.R.: Reconstructing reticulate evolution in species: theory and practice. In: RECOMB. (2004) 337–346
8. Nakhleh, L., Ruths, D.A., Wang, L.S.: Riata-hgt: A fast and accurate heuristic for reconstructing horizontal gene transfer. In: COCOON. (2005) 84–93
9. Beiko, R.G., Harlow, T.J., Ragan, M.A.: Highways of gene sharing in prokaryotes. Proc Natl Acad Sci U S A **102**(40) (2005) 14332–14337
10. Than, C., Ruths, D.A., Innan, H., Nakhleh, L.: Confounding factors in hgt detection: Statistical error, coalescent effects, and multiple solutions. Journal of Computational Biology **14**(4) (2007) 517–535
11. Jin, G., Nakhleh, L., Snir, S., Tuller, T.: Parsimony score of phylogenetic networks: Hardness results and a linear-time heuristic. IEEE/ACM Trans. Comput. Biology Bioinform. **6**(3) (2009) 495–505
12. Boc, A., Philippe, H., Makarenkov, V.: Inferring and Validating Horizontal Gene Transfer Events Using Bipartition Dissimilarity. Syst Biol **59**(2) (2010) 195–211
13. Gary, M.W.: Origin and evolution of organelle genomes. Curr Opin Genet Dev **3** (1993) 884–890
14. Huang, J., Gogarten, J.: Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biology **8**(6) (2007) R99
15. Lake, J.A.: Evidence for an early prokaryotic endosymbiosis. Nature **460** (2009) 967–971
16. Zhaxybayeva, O., Gogarten, J.P., Charlebois, R.L., Doolittle, W.F., Papke, R.T.: Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. Genome Research **16**(9) (2006) 1099–1108
17. Hein, J.: Reconstructing evolution of sequences subject to recombination using parsimony. Mathematical biosciences **98**(2) (1990) 185–200
18. Bender, M.A., Farach-Colton, M.: The LCA problem revisited. In: LATIN. (2000) 88–94
19. Zhaxybayeva, O., Doolittle, W.F., Papke, R.T., Gogarten, J.P.: Intertwined Evolutionary Histories of Marine *Synechococcus* and *Prochlorococcus marinus*. Genome Biol Evol **1** (2009) 325–339
20. Fournier, G.P., Gogarten, J.P.: Rooting the Ribosomal Tree of Life. Mol Biol Evol (2010) [Epub ahead of print] msq057