Inapproximability Results for the Lateral Gene Transfer Problem*

Bhaskar DasGupta[1], Sergio Ferrarini[2], Uthra Gopalakrishnan[1], and Nisha Raj Paryani[1]

[1] Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7053
{dasgupta, ugopalak, nparyani}@cs.uic.edu
[2] Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci 32 20133, Milano, Italy
sferrarini@gmail.com

Abstract. This paper concerns the *Lateral Gene Transfer Problem*. This minimization problem, defined by Hallet and Lagergren [6], is that of finding the *most parsimonious* lateral gene transfer scenario for a given pair of gene and species trees. Our main results are the following:

(a) We show that it is not possible to approximate the problem in polynomial time within an approximation ratio of $1+\varepsilon$, for some constant $\varepsilon > 0$ unless P=NP. We also provide explicit values of ε for the above claim.
(b) We provide an upper bound on the cost of any 1-active scenario and prove the tightness of this bound.

1 Introduction

A fundamental problem in the field of evolutionary molecular biology is that of inferring information on the evolutionary relationships between taxa from a given set of gene trees (*i.e.*, an evolutionary model for a set of gene families). The underlying assumption is that gene families evolve in the same way as species; therefore a gene tree should determine the species tree. Unfortunately, there are a number of biological events, such as *gene duplications*, *gene losses* and *lateral gene transfers* (also called horizontal gene transfers) (*e.g.*, see [5, 9]) that may occur during evolution and that generate "differences" between a gene and a species tree. For these reasons, a single gene tree is usually not sufficient to reliably build the species trees, but it is necessary to consider a set a gene families to perform the construction. Since the gene trees may contain contradictory information, a natural problem that arises is that of *reconciling* the different gene trees into a single species tree. Such a reconciliation process can be naturally formulated as an optimization problem where the goal is to *minimize* the number

* This research was supported by NSF grants CCR-0296041, CCR-0206795, CCR-0208749 and IIS-0346973.

M. Coppo et al. (Eds.): ICTCS 2005, LNCS 3701, pp. 182–195, 2005.

of biological events necessary to explain the "disagreements" between the gene trees and the species tree.

Several models have been proposed to solve the reconciliation problem. Each of these models is based on the assumption that only a restricted class of genomic events may occur. Here we focus on the so-called *lateral gene transfer model* defined by Hallett and Lagergren [6]. According to this model, all differences between the gene and the species trees are explained in terms of lateral gene transfer events. A lateral gene transfer is an event that causes some portion of the evolution represented by an arc in the gene tree to occur along one arc in the species tree, and the remaining portion of evolution to occur along another arc of the species tree. We say that the lateral transfer occurs between these two arcs of the species tree and involves the arc from the gene tree. Given a gene tree T and a species tree S (which we assume to be correct), an interesting optimization problem is that of identifying a *scenario* that is able to explain the differences between the two trees with the *minimum* number of lateral transfers. We refer to this problem as the Lateral Gene Transfer Problem, in short as the LGT Problem.

In this paper we investigate efficient approximability issues of the LGT Problem. We consider both the special case of activity level one and the general case of activity level some $\alpha \geq 1$, and establish hardness of efficient approximation for both cases. More specifically, we will prove that, unless P=NP, no algorithm can achieve an approximation ratio smaller than $1 + \varepsilon$ for some constant $\varepsilon > 0$. By easy calculations we also provide explicit values of ε for the above claim. We also show that an upper bound on the cost of *any* 1-active lateral transfer scenario is given by $n - 2$ where n is the number of leaves in the gene tree, and show that this upper bound is tight by explicitly giving a pair of species and gene trees with this cost.

1.1 Basic Definitions and Notations

In the remaining sections we consider just rooted binary trees, *i.e.*, trees where all vertices have out-degree at most two and all arcs are directed from the root to the leaves. Given a rooted tree T, we denote with $V(T)$ the set of vertices and with $A(T)$ the set of arcs. The leaves of T are denoted by $L(T)$ and the root by $r(T)$. We say that two distinct vertices v, v' are children of u in T if $\langle u, v\rangle$, $\langle u, v'\rangle \in A(T)$. We denote the left son of a vertex $u \in V(T)$ as $ls_T(u)$, the right son as $rs_T(u)$, and the parent of u as $p_T(u)$.

Let F be a *rooted forest*, that is, a union of disjoint rooted trees. If two vertices $u, v \in V(F)$ are connected by a directed path from u to v, then v is a *descendent* of u in F, and we write $v \leq_F u$ (note that every node is a descendent of itself). If $u \neq v$, then v is a *proper descendent* of u in F ($<_F$). Similarly, we can define *ancestors* ($\geq_F$) and *proper ancestors* ($>_F$). Moreover, let T be a rooted tree and $X \subset V(T)$. Then $T[X]$ denotes the forest of subtrees induced by X.

Let $\{t_i : 1 \leq i \leq n\}$ be a forest of non-empty rooted directed trees over a label set L. We use the notation $T = \prec t_1 \cdot t_2 \cdot \ldots \cdot t_n \succ$ to represent the tree built

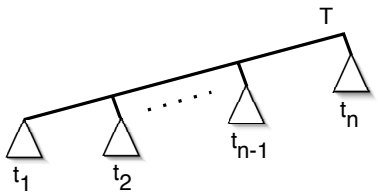

Fig. 1. The tree $T = \prec t_1 \cdot t_2 \cdot \ldots \cdot t_n \succ$

by connecting the subtrees t_i as shown in Figure 1. As a shorthand, we allow the notation $\prec \prod_{i=1}^{n} t_i \succ$ to mean $\prec t_1 \cdot t_2 \cdot \ldots \cdot t_n \succ$.

A *mixed graph* G is a graph where arcs may be both directed and undirected. We denote the set of directed arcs as $A(G)$ and the undirected arcs, or *edges*, as $E(G)$. $\varepsilon(A)$ indicates the set of edges underlying $A(G)$. Given a set of arcs A and a mixed graph G, we denote by $G \cup A$ the mixed graph with arcs $A(G) \cup A$, egdes $E(G)$, and vertices $V(G)$. Similarly is defined $G \cup E$, where E is a set of edges. A *directed mixed cycle* is a cycle on a mixed graph that may contain both arcs and edges, and where the cycle can be traversed so that the direction of the arcs that are part of the cycle is respected. Given a graph G and an arc $\langle x, y \rangle \in A(G)$, we say that we *subdivide* arc $\langle x, y \rangle$ if we replace it with a path from x to y that doesn't traverse any vertex in $V(G) \setminus \{x, y\}$. We say that a graph H is a *subdivision* of a graph G, if H is obtained from G by subdividing some of the arcs in $A(G)$ and adding new arcs between vertices in $V(H) \setminus V(G)$.

Finally, a $(1+\varepsilon)$-*approximate solution* (or simply an $(1+\varepsilon)$-approximation) of a minimization problem is a solution with an objective value no larger than $1+\varepsilon$ times the value of the optimum, and an algorithm achieving such a solution is said to have an *approximation ratio* of at most $1+\varepsilon$.

1.2 Basic Concepts of the Evolutionary Model

In this section we define some basic concepts for the evolutionary model we consider. We will first briefly introduce the concepts of a *gene tree* and a *species tree*. We will then present the concept of *least common ancestor mapping* and define a reconciliation model based on *lateral gene transfer events*.

Consider a set I of N biological taxa. The model for their evolutionary history is a rooted full binary tree S, where each of the N leaves is uniquely labeled by one element from I, and each internal node is unlabeled. Such tree S is called a *species tree*. An internal node in a species tree is equivalently treated as a subset (or *cluster*), which contains the labels of all leaves of the subtree rooted at that node. Thus, we can express the relation "m is a descendant of n" in set theory notation by $m \subset n$.

A *gene tree* T is a model for the evolution of a gene family. It is a rooted full binary tree where the internal nodes are unlabeled and the leaves are labeled by elements from I. As opposed to a species tree, labels in a gene tree may not be unique. In this case, internal nodes are represented by a multiset

$\{x_1^{i_1}, x_2^{i_2}, \ldots, x_m^{i_m}\}$, where i_j is the number of leaves labeled with x_j, among those reachable from the node. The *cluster* of an internal node is defined as the set $\{x_1, x_2, \ldots, x_m\}$.

Let Y be a rooted tree and $L(Y)$ the set of its leaf labels. The *least common ancestor (LCA)* of $X \subseteq L(Y)$, denoted by $lca_Y(X)$, is defined as the node $y \in Y$ such that $X \subseteq y$ and $X \not\subseteq w$ for every proper descendent w of y. Given a gene tree T and a species tree S such that $L(T) \subseteq L(S)$, we define $\lambda_{T,S} : V(T) \longrightarrow V(S)$ as a correspondence between nodes of the gene tree T and nodes of the species trees S. For any node $t \in T$, $\lambda_{T,S}(t)$ is the least common ancestor of t in S, *i.e.*, $\lambda_{T,S}(t) = lca_S(t)$. The function $\lambda_{T,S}$ is known as the *LCA mapping* from T to S.

1.2.1 The Lateral Gene Transfer Model

We are now ready to introduce the evolutionary model based on the concept of *lateral gene transfers*. This model was developed by Hallet and Lagergren [6]. It assumes a simplified evolutionary process where the only biological events that can occur are the so-called *lateral gene transfers*. In this framework, a natural problem is that of finding the most parsimonious *scenario* that explains in a biologically meaningful way how, via these events, the differences between the gene tree and the species tree arose. The definition of lateral transfer scenario is based on the concept of *lateral transfer scheme*.

Definition 1. *A* lateral transfer scheme *for a species tree S is a pair (S', A') where S' is a subdivision of S and $A' \subseteq \{\langle x, y\rangle : x, y \in V(S') \setminus V(S), x \neq y\}$ such that:*

1. *the mixed graph $S' \cup \varepsilon(A')$ does not contain a directed mixed cycle.*
2. *the tail of each arc in A' has in-degree 1 and out-degree 2 in $S' \cup A'$.*
3. *the head of each arc in A' has in-degree 2 and out-degree 1 in $S' \cup A'$.*

A lateral transfer scheme shows where the lateral transfers have occurred during evolution. The arcs in A' represent the *set of lateral transfers*. Note that the first condition in Definition 1 ensures that the scheme for a species tree S respects the partial order of evolution implied by S. Clearly, this is a required property for the model.

A lateral transfer scheme is meaningful when combined to the notion of *scenario*. A scenario is a mapping of a gene tree into a subdivision of a species tree. This mapping describes how the gene tree has evolved by showing at which point of evolution lateral gene transfers have occurred. In order for a scenario to be biologically meaningful, it must satisfy the conditions stated in the definition below.

An important parameter for this model is the *activity level* α. The parameter α measures the number of genes that are allowed to be simultaneously active in the genome of a taxa. Roughly speaking, an α-active scenario permits at most α copies of a gene to be mapped to the same ancestral taxon. In previous models, the presence of multiple copies of a gene was always assumed to be caused by a *gene duplication* event. The notion of acitivity level in a lateral transfer scenario

overcomes this restriction, by postulating that this multiplicity may be generated by lateral transfer events alone.

We will first give the definition for the special case of 1-activity, and then state the definition for the general α-active case where $\alpha \geq 1$.

Definition 2. *A* 1-active lateral transfer scenario *(or* 1-active scenario*) for a species tree S and a gene tree T is a triple (S', A', g) where (S', A') is a lateral transfer scheme for S and $g : V(S') \to V(T)$ is a function such that:*

1. *$g(r(S')) = r(T)$.*
2. *if v_1 and v_2 are distinct children of v_0 in T, then there exists x_0 with distinct children x_1 and x_2 in $S' \cup A'$ such that $v_i \in g(x_i)$ for $i \in \{0, 1, 2\}$, and x_i is the $\leq_{s'}$-maximal vertex such that $v_i \in g(x_i)$ for $i \in \{1, 2\}$.*
3. *for each $v \in V(T)$, the vertices $\{x \in V(S') : v \in g(x)\}$ induce a directed path in S'.*
4. *$g(l) = l$, for all $l \in L(S)$.*

The cost *of a 1-active scenario (S', A', g) w.r.t. T is given by $|A'|$.*

Definition 3. *A* lateral transfer scenario *(or* scenario*) for a species tree S and a gene tree T is a triple (S', A', g) where (S', A') is a lateral transfer scheme for S and $g : V(S') \to 2^{V(T)}$ is a function such that:*

1. *$T[g(r(S'))]$ is connected and $r(T) \in g(r(S'))$.*
2. *if v_1 and v_2 are distinct children of v_0 in T and $v_1, v_2 \notin g(r(S'))$, then there exists x_0 with distinct children x_1 and x_2 in $S' \cup A'$ such that $v_i \in g(x_i)$ for $i \in \{0, 1, 2\}$, and x_i is the $\leq_{S'}$-maximal vertex such that $v_i \in g(x_i)$ for $i \in \{1, 2\}$.*
3. *if v_1 and v_2 are children of v_0 in T, $v_1 \in g(r(S'))$ and $v_2 \notin g(r(S'))$, then there exists a child x of $r(S')$ in S' such that $v_2 \in g(x)$.*
4. *for each $v \in V(T)$, the vertices $\{x \in V(S') : v \in g(x)\}$ induce a directed path in S'.*
5. *$g(x)$ is a $\leq_T$-antichain for each $x \in V(S') \setminus \{r(S')\}$.*
6. *$g(l) = \{l\}$, for all $l \in L(S)$.*

A scenario (S', A', g) is α-active *iff $max_{x \in S'} |g(x)| = \alpha$. The* cost *of a α-active scenario (S', A', g) w.r.t. T is given by:*

$$\sum_{\langle x,y \rangle \in A'} |\{\langle u, v \rangle \in A(T) : u \in g(x), v \in g(y)\}| + |V(T[g(r(S'))]) \setminus L(T[g(r(S'))])|$$

Let (S', A', g) be a scenario for S and T. We say that an arc of T is *invloved* into a lateral transfer if it belongs to the following set:

$$F = \{\langle u, v \rangle \in A(T) : u \in g(x), v \in g(y), \text{ where } \langle x, y \rangle \in A'\}$$

1.3 Problem Definitions

In this paper we investigate the following optimization problems:

1-active LGT Problem
Instance: A species tree S and a gene tree T, such that $L(T) \subseteq L(S)$.
Goal: Find a 1-active lateral transfer scenario for S and T with minimum cost.

α-active LGT Problem
Instance: A species tree S and a gene tree T such that $L(T) \subseteq L(S)$, a *constant* $\alpha \geq 1$.
Goal: Find an α-active lateral transfer scenario for S and T with minimum cost.

Note that one can easily convert the above optimization problems into their *decision version* by having an extra integer τ as input and requiring the minimum cost to be $\leq \tau$. We call the decision versions of these problems the 1-active τ-LGT Problem and α-active τ-LGT Problem respectively. It was shown in [7] that the α-active τ-LGT Problem is NP-complete.

1.4 Inapproximability Reductions: Key Concepts and Results

In [10] Papadimitriou and Yannakakis defined the class of *MAX-SNP* optimization problems and a special approximation-preserving reduction, the so-called *L-reduction*, that can be used to show MAX-SNP-hardness of an optimization problem. The version of the L-reduction that we provide below is a slightly modified but equivalent version that appeared in [4].

Definition 4. [4, 10] *Given two optimization problems Π and Π', we say that Π* L-reduces *to Π' if there are three polynomial-time procedures T_1,T_2, T_3 and two constants a and $b > 0$ such that the following two conditions are satisfied:*

1. *For any instance I of Π, algorithm T_1 produces an instance $I' = f(I)$ of Π' generated from T_1 such that the optima of I and I', $OPT(I)$ and $OPT(I')$, respectively, satisfy $OPT(I') \leq a \cdot OPT(I)$.*
2. *For any solution of I' with cost c', algorithm T_2 produces another solution with cost c'' that is no worse than c', and algorithm T_3 produces a solution of I of Π with cost c (possibly from the solution produced by T_2) satisfying $|c - OPT(I)| \leq b \cdot |c'' - OPT(I')|$.*

An optimization problem is MAX-SNP-*hard* if any problem in MAX-SNP L-reduces to that problem. If this problem is also in MAX-SNP, then it is MAX-SNP-*complete*. The importance of proving MAX-SNP-hardness results comes from a result proved by Arora et al. [1] which shows that, assuming P$\neq$NP, for every MAX-SNP-hard problem there exists a constant $\varepsilon > 0$ such that no polynomial time algorithm can achieve an approximation ratio better than $1+\varepsilon$.

1.5 Precise Statements of Our Results

Theorem 1
(a) For some constant $\varepsilon > 0$, it is not possible to approximate in polynomial time the 1-active LGT PROBLEM within an approximation ratio of $1+\varepsilon$ unless P=NP.
(b) The constant ε in (a) is at least $(3/370024) - \kappa$ for any $\kappa > 0$.

Theorem 2
(a) For some constant $\varepsilon > 0$, it is not possible to approximate in polynomial time the α-active LGT PROBLEM within an approximation ratio of $1+\varepsilon$, where $\alpha \geq 1$, unless P=NP.
(b) The constant ε in (a) is at least $(3/378068) - \kappa$ for any $\kappa > 0$.

Lemma 1. *The minimum number of lateral transfers necessary to build a 1-active lateral transfer scenario for any pair of gene and species trees, uniquely labeled over the same set of labels L, is precisely $n-2$ where $n = |L|$. That is, there is a procedure to build a 1-active scenario of cost $n-2$ and there exists a pair of a gene tree and a species tree that require at least $n-2$ lateral transfers for any 1-active scenario.*

2 Hardness of Approximation of 1-Active LGT PROBLEM (Proof of Theorem 1(a))

In the following we will show that MAX-2SAT-B L-reduces to the 1-active LGT PROBLEM. MAX-2SAT-B is the variation of MAX-2SAT where the number of occurrences of each variable is bounded by a constant B. It is known from [2] that MAX-2SAT-B is MAX-SNP-complete for $B \geq 3$; thus the existence of an L-reduction will imply the result.

Let $X = \{X_1, \ldots, X_n\}$ be a set of n variables and let $\Phi = (C_1, \ldots, C_m)$ be a formula in 2-CNF, where each clause C_i is on two variables from X and where the number of occurrences of each variable is bounded by a constant B. We will refer to the j^{th} variable in the i^{th} clause as literal $C_{i,j}$. The goal of MAX-2SAT-B is to find a truth assignment on X that maximizes the number of satisfied clauses. Given an instance of MAX-2SAT-B, we will now exhibit how to build an instance of the 1-active LGT PROBLEM such that Conditions 1 and 2 of Definition 4 are satisfied. In other words, we will construct a gene tree T and a species tree S from Φ and prove that this transformation is an L-reduction.

Our construction of T and S from Φ is taken from the NP-completeness proof given in [7]. The only difference is that in our case the index j in $C_{i,j}$ ranges between 1 and 2, rather than 1 and 3 (their reduction is from 3SAT). A detailed description of this procedure is here omitted.

An important parameter used in the following proof is τ, defined as $\tau = 9m + 6k$, where

$$k = \left|\{\langle i, j, i', j'\rangle \mid C_{i,j} = C_{i',j'} \text{ or } C_{i,j} = \overline{C_{i',j'}}, 1 \leq i < i' \leq m, 1 \leq j, j' \leq 2\}\right|.$$

Notice that $k \leq \frac{nB(B-1)}{2}$, since the maximum number of occurrences is bounded by B for each variable in Φ; hence, $\tau \leq \gamma m$, where $\gamma = 9 + 6B(B-1)$. Also, let $\tau^+ = \tau + m + 1$. To ease our presentation, we will adopt the same notation used in [7] throughout the rest of the proof.

Let Ψ be a truth assignment on the variable set X, i.e. $\Psi : X_i \longmapsto \{\text{true, false}\}$, for every $i = 1, \ldots, n$. We show that there is a correspondence between truth assignments on Φ and scenarios for T and S. The proof of the following claim is omitted due to page limits.

Claim 1. *Given a truth assignment Ψ on Φ that satisfies ρ clauses, $\rho \leq m$, it is always possible to build in polynomial time a 1-active lateral transfer scenario for T and S with cost $\tau + (m - \rho)$.*

It is now easy to show that the first condition of Definition 4 is satisfied. Starting from an optimal truth assignment Ψ_{OPT} that satisfies OPT_Φ clauses from Φ, by Claim 1 we can build a 1-active scenario of cost $\tau + (m - OPT_\Phi)$. If we denote by $OPT_{T,S}$ the cost of the optimal scenario on T and S, then $OPT_{T,S} \leq \tau + (m - OPT_\Phi) \leq (\gamma + 1)m$. Moreover, it is not hard to see that a random truth assignment satisfies each clause with probability 3/4, and hence it is not hard to find (even deterministically) an assignment OPT_Φ that satisfies $3m/4$ clause (*e.g.*, see [8]). Thus, without loss of generality we may assume that $OPT_\Phi \geq 3m/4$. By combining the two inequalities we have $OPT_{T,S} \leq a \cdot OPT_\Phi$, where $a = \frac{4(\gamma+1)}{3}$. This completes the first part of the proof.

Lets now verify the second condition. Suppose we are given a 1-active lateral transfer scenario for T and S of cost c'. We can assume without loss of generality that $c' \leq \tau + m$, otherwise we could choose any scenario built from an arbitrary assignment to replace the given one. Observe that Claims 1-8 in [7] are true for the given scenario, while Claim 9 in [7] must be slightly modified to fit our construction. This is the modified result:

Claim 2. *[7] In any 1-active scenario for T and S, at least one element of $X_i = \{r(T_{i,j}) : 1 \leq j \leq 2\}$ is the tail of an arc involved in a lateral transfer, for every i. This requires $\geq m$ lateral transfers.*

Proof. Follows from the observation that, by the 1-activity conditions, only one element from X_i may be mapped to $r(B_{S_i})$. ❑

Note that Claims 1-8 in [7] together with Claim 2 imply that a lower bound on the cost of any lateral transfer scenario for T and S is equal to τ. We now show that, starting from the given scenario, it is always possible to build in polynomial time a new scenario of cost $\leq c'$, which induces a consistent truth assignment on Φ. *This new scenario and its induced truth assignment will satisfy Condition 2 of Definition 4 with $b = 1$.*

Let X_v be a variable from the variable set X and $\Omega_v = \{C_{i,j} \mid X_v$ appears in $C_{i,j}\}$, $\omega_v = |\Omega_v|$. Let $\tilde{T}_{i,j} = \prec\prec a_{i,j} \cdot c_{i,j} \succ \cdot \prec b_{i,j} \cdot d_{i,j} \succ\succ$ and F_v be a forest of subtrees of T defined as $F_v = \{\tilde{T}_{i,j} \mid C_{i,j} \in \Omega_v\} \cup \{\varepsilon'_{i,j,i',j'} \mid C_{i,j}, C_{i',j'} \in \Omega_v\}$. Moreover, let $k_v = |\{\langle i, j, i', j' \rangle \mid C_{i,j}, C_{i',j'} \in \Omega_v$ and $i < i'\}|$ and define $\tilde{\tau}_v = 2\omega_v + 4k_v$.

We say that a literal $C_{i,j}$ is *well-assigned* if the corresponding gene subtree $\tilde{T}_{i,j}$ has exactly two arcs involved in lateral transfers. This implies that either $lca_T(c_{i,j}, d_{i,j}) \in g(lca_S(c_{i,j}, d_{i,j}))$ or $lca_T(c_{i,j}, d_{i,j}) \in g(lca_S(a_{i,j}, b_{i,j}))$.

We also say that literal $C_{i,j}$ is *inconsistent* with respect to $C_{i',j'}$ if one of the two following conditions holds:

- $C_{i,j} = C_{i',j'}$, and i) $lca_T(c_{i,j}, d_{i,j}) \in g(lca_S(c_{i,j}, d_{i,j})$ and $lca_T(c_{i',j'}, d_{i',j'}) \notin g(lca_S(c_{i',j'}, d_{i',j'})$ or ii) $lca_T(c_{i,j}, d_{i,j}) \in g(lca_S(a_{i,j}, b_{i,j})$ and $lca_T(c_{i',j'}, d_{i',j'}) \notin g(lca_S(a_{i',j'}, b_{i',j'})$.
- $C_{i,j} = \overline{C_{i',j'}}$, and i) $lca_T(c_{i,j}, d_{i,j}) \in g(lca_S(c_{i,j}, d_{i,j})$ and $lca_T(c_{i',j'}, d_{i',j'}) \notin g(lca_S(a_{i',j'}, b_{i',j'})$ or ii) $lca_T(c_{i,j}, d_{i,j}) \in g(lca_S(a_{i,j}, b_{i,j})$ and $lca_T(c_{i',j'}, d_{i',j'}) \notin g(lca_S(c_{i',j'}, d_{i',j'})$.

Let $I_{i,j}$ be the set of all literals that are inconsistent w.r.t. $C_{i,j}$ and $i_{i,j} = |I_{I,J}|$.

Claim 3. *If no $C_{i,j} \in \Omega_v$ is well-assigned, then at least $\tilde{\tau}_v + \omega_v$ arcs in F_v are involved in lateral transfers. If there exists a well-assigned $C_{i,j} \in \Omega_v$, then at least $\tilde{\tau}_v + i_{i,j}$ arcs in F_v are involved in transfers.*

Proof. By Claims 7 and 8 in [7], $\tilde{\tau}$ is a lower bound on the number of arcs in F_v that are involved in lateral transfers. The first part of the claim follows immediately from the fact that for each non well-assigned literal $C_{i,j}$ at least three transfers are required for $\tilde{T}_{i,j}$, that is an additional transfer for every literal w.r.t. the minimum scenario.

Now suppose that $C_{i,j} \in \Omega_v$ is well-assigned, and assume that $lca_T(c_{i,j}, d_{i,j}) \in g(lca_S(c_{i,j}, d_{i,j}))$, i.e. $C_{i,j}$ is true. Consider a second literal $C_{i',j'} \in \Omega_v$ that is inconsistent w.r.t. $C_{i,j}$, and assume for example that $C_{i,j} = C_{i',j'}$. If $C_{i',j'}$ is not well-assigned, then $\tilde{T}_{i',j'}$ has at least three arcs involved in lateral transfers. Conversely, if $C_{i',j'}$ is inconsistent and well-assigned, it is straightforward to verify that $lca_T(c_{i',j'}, d_{i',j'}) \in g(lca_S(a_{i',j'}, b_{i',j'}))$, since any different mapping would require more than two arcs from $\tilde{T}_{i,j}$ involved in transfers. Moreover, the path from $p_T(b_{i',j'})$ to $a_{i',j'}$ in T blocks the path from $p_S(\beta_{i',j',i,j})$ to $\alpha_{i',j',i,j}$ in S, where by *blocks* we mean that at least one vertex belonging to the path on T is mapped to a vertex from the path on S. Equally, the path from $p_T(d_{i,j})$ to $c_{i,j}$ in T blocks the path from $p_S(\delta_{i,j,i',j'})$ to $\gamma_{i,j,i',j'}$ in S. Now, since both paths are blocked, for any valid scenario built under these assumptions at least three arcs in the subtree $\varepsilon'_{i',j',i,j}$ must be involved in lateral transfers. An example of this case is given in figure 2.

For the symmetric case where $lca_T(c_{i,j}, d_{i,j}) \in g(lca_S(a_{i,j}, b_{i,j}))$ and $lca_T(c_{i',j'}, d_{i',j'}) \in g(lca_S(c_{i',j'}, d_{i',j'}))$, i.e. $C_{i,j}$ is false and $C_{i',j'}$ is true, a similar argument shows that $\tilde{T}_{i',j'}$ or $\varepsilon'_{i,j,i',j'}$ require at least three lateral transfers.

We can therefore conclude that in both cases at least seven arcs from the subtrees $\tilde{T}_{i',j'}$, $\varepsilon'_{i,j,i',j'}$ and $\varepsilon'_{i',j',i,j}$ are involved in lateral transfers; that is, the given scenario requires on these subtrees at least one transfer more than the minimum scenario, which reaches the lower bound of six implied by Claims 7-8 in [7]. A similar reasoning establishes the same result for the cases where $C_{i,j} = \overline{C_{i',j'}}$.

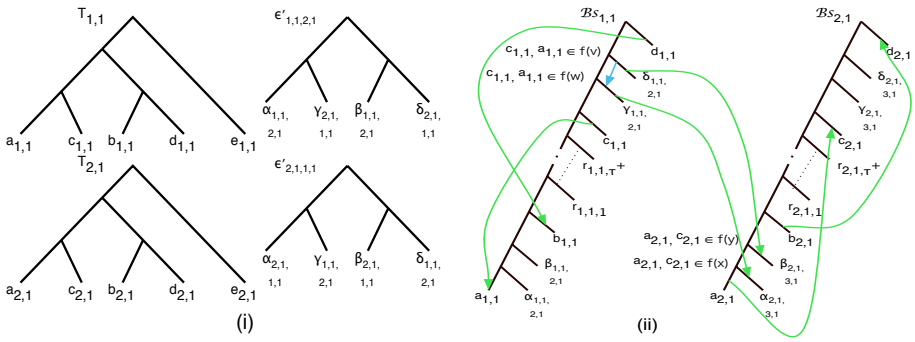

Fig. 2. Consider the formula $\Phi = (X_1 \vee \overline{X_2})(X_1 \vee X_2)$. (i) depicts the twister gadgets $(T_{1,1}, T_{2,1})$ and enforcement gadgets $(\varepsilon'_{1,1,2,1}, \varepsilon'_{2,1,1,1})$ corresponding to the literals $(C_{1,1}, C_{2,1})$ which are instances of $X_1 \in X$. (ii) represents a partial scenario where literals $C_{1,1}$ and $C_{2,1}$ are well-assigned and inconsistent. The dashed lines are the lateral transfers. Notice that both paths from $p_S(\delta_{1,1,2,1})$ to $\gamma_{1,1,2,1}$ and from $p_S(\beta_{2,1,1,1})$ to $\alpha_{2,1,1,1}$ are blocked by $\{c_{1,1}, a_{1,1}\}$,$\{a_{2,1}, c_{2,1}\} \in T$ respectively. Thus, an additional transfer from arc $\langle p_S(\delta_{1,1,2,1}), \delta_{1,1,2,1}\rangle$ to $\langle p_S(\gamma_{1,1,2,1}), \gamma_{1,1,2,1}\rangle$ is necessary.

Hence, for every literal not consistently assigned w.r.t $C_{i,j}$, at least one additional lateral transfer is required. Thus, $\geq \tilde{\tau_v} + i_{i,j}$ arcs of F_v are involved in transfers. ❑

We will now describe a simple procedure to build a new scenario of cost $\leq c'$ that is based on the given scenario as a starting point. Construct a truth assignment $\Psi : X \longrightarrow \{\text{true, false}\}$ on Φ in the following way. For each variable X_v, $v = 1, \ldots, n$, check if there exists (can be done in linear time) some literal $C_{i,j} \in \Omega_v$ which is well-assigned. If this is the case, assign to X_v the truth value read from $C_{i,j}$, i.e. $\Psi(X_v) = \text{true}$ if $lca_T(c_{i,j}, d_{i,j}) \in lca_S(c_{i,j}, d_{i,j})$ and $\Psi(X_v) = \text{false}$ if $lca_T(c_{i,j}, d_{i,j}) \in lca_S(a_{i,j}, b_{i,j})$. If no literal in Ω_v is well-assigned, then assign to X_v an arbitrary truth value. Now follow the procedure described in Claim 1 and build a scenario from Ψ. As a shorthand, call LTS_I the given scenario and LTS_F the new scenario built from Ψ.

We say that a clause C_i is *satisfied* by a lateral transfer scenario on T and S if exactly one element from the set $\{r(T_{i,j}) \mid 1 \leq j \leq 2\}$ is the tail of an arc involved in a lateral transfer. Clearly, this is true only if $\exists j$ s.t. $lca_T(c_{i,j}, d_{i,j}) \in lca_S(c_{i,j}, d_{i,j})$.

LTS_F has cost $c_a = \tau + (m - \rho)$, where ρ is the number of clauses that Ψ satisfies. In other words, LTS_F has a minimum number of transfers on all subtrees of S, except on the subtrees B_{S_i} corresponding to those clauses C_i that are not satisfied by Ψ, where an additional lateral transfer (w.r.t. the minimum cost scenario on this tree) is required. Claims 1-8 in [7] and Claim 2 establish that τ is a lower bound for any valid scenario, hence $c' \geq \tau$. Moreover, a clause that is not satisfied in LTS_F can be satisfied by LTS_I only by a non well-assigned literal $C_{i,j}$ (in the case where $C_{i,j} \in \Omega_v$ and X_v is arbitrarily assigned)

or by a literal which is inconsistent w.r.t. the chosen assignment $\Psi(X_v)$. By Claim 3, this implies that LTS_I has at least one lateral transfer more than the minimum scenario for each clause that is true in LTS_I and false in LTS_F. Hence, $c' \geq \tau + (m - \rho)$.

Therefore, given any scenario on S and T, we are able to build in polynomial time a new scenario of cost $\tau + (m - \rho) \leq c'$ which corresponds to a valid truth assignment on Φ that satisfies ρ clauses. The theorem follows.

3 Hardness of Approximation of the α-Active LGT PROBLEM (Proof of Theorem 2(a))

Once again, we L-reduce from MAX-2SAT-B. Let T^* and S^* respectively be the gene and species tree of the α-active LGT PROBLEM instance. Build T^* and S^* by following the procedure given in [7]. The details on this construction are here omitted. Starting from an optimal truth assignment Ψ_{OPT} that satisfies OPT_Φ clauses from Φ, first create a 1-active scenario on T and S by applying the construction described in Claim 1. Use this scenario to construct an α-active scenario for T^* and S^* as shown in [7]. The cost of this scenario is $\tau^* + (m - OPT_\Phi)$, where $\tau^* = \tau + (\alpha - 1)$, and hence $OPT_{T^*,S^*} \leq \tau^* + (m - OPT_\Phi)$. From Theorem 1, we know that $\tau \leq \gamma m$, where $\gamma = 9 + 6B(B-1)$. *For all sufficiently large values* of m, we have $\tau^* \leq (\gamma + 1)m$ and $OPT_{T^*,S^*} \leq (\gamma + 2)m$. Thus, the the first condition from Definition 4 is satisfied, with $a = \frac{4(\gamma+2)}{3}$.

Consider now an α-active scenario for T^* and S^* of cost c^*. We can assume w.l.o.g. that $c^* \leq \tau^* + m$; if this were not the case, any scenario built from an arbitrary truth assignment would do better, and we could use this scenario as a starting point. Notice that any α-active scenario requires at least $\alpha - 1$ lateral transfers for subtrees $T^1, \ldots, T^{\alpha-1}$ and S^*. Therefore, at most $\tau + m$ transfers are involved in the partial scenario for T and S^*. By Claim 11 of [7], the α-active scenario for T and S^* induces a 1-active scenario for T and S of cost $\leq \tau + m$. It has been shown that any 1-active scenario for T and S of cost c' induces a truth assignment on Φ that satisfies ρ clauses, where ρ is s.t. $c' \geq \tau + (m - \rho)$. It follows that $c^* = c' + (\alpha - 1) \geq \tau^* + (m - \rho)$. Thus, the the second condition of Definition 4 is satisfied with $b = 1$. This concludes our proof.

4 Hard Inapproximability Bounds (Proofs of Theorem 1(b) and Theorem 2(b))

Berman and Karpinski [3] proved that it is NP-hard to approximate MAX-2SAT-3 to within a factor $2012/2011 - \kappa$, for every $\kappa > 0$. The following result from [10] allows us to compute approximation ratios that are NP-hard to achieve for the 1-active and α-active LGT PROBLEM from that of MAX-2SAT-3.

Proposition 1. [10] *Let Π and Π' be two optimization problems. If Π L-reduces to Π' , and there is a polynomial time approximation algorithm for Π' with*

worst-case error ε, then there is a polynomial time approximation algorithm for Π with worst-case error $ab\varepsilon$, where a and b are the constants of the L-reduction.

Proposition 1 can be stated equivalently as follows: if approximating Π to within an approximation ratio smaller than $1+\varepsilon$ is NP-hard, then achieving an approximation ratio for Π' smaller than $1+\varepsilon/(ab)$ is also NP-hard.

Consider the L-reduction for the 1-active case. We have $a = \frac{40+24B(B-1)}{3}$, which implies $a = 184/3$ for $B = 3$, and $b = 1$. It follows that, for the 1-active LGT Problem it is NP-hard to achieve an approximation ratio of $370027/370024 - \kappa$, for every $\kappa > 0$.

Similarly, for the α-active case, $a = \frac{44+24B(B-1)}{3}$, which implies $a = 188/3$ for $B = 3$, and $b = 1$. This shows that it is NP-hard to approximate the α-active LGT Problem to within a factor of $378071/378068 - \kappa$, for every $\kappa > 0$.

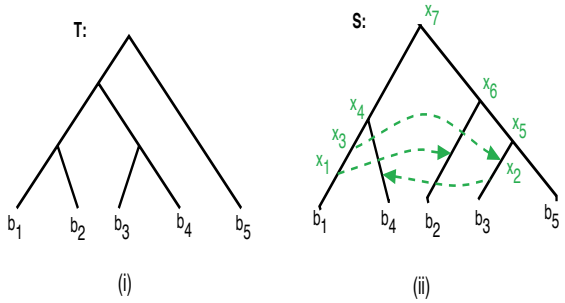

Fig. 3. An example of scenario built from the construction procedure illustrated below. (i) is the gene tree and (ii) the scenario built on the species tree. The dashed arcs are the lateral transfers. Here, $\{b_1b_2\} \in g(x_1)$, $\{b_3b_4\} \in g(x_2)$, $\{b_1b_2b_3b_4\} \in g(x_3)$ and $g(x_4)$, $\{b_5\} \in g(x_5)$ and $g(x_6)$, $r(T) \in g(r(S))$. Note that this isn't the minimum scenario.

5 Upper Bound of Cost of 1-Active Scenario (Proof of Lemma 1)

We first describe a procedure to build a 1-active scenario of cost $n-2$ for any given gene and species trees. Let T be a gene tree and S be a species tree that satisfy $L(T) = L(S)$. We order the *internal* vertices (i.e. all vertices except the leaves) of T, by imposing the ordering produced by a post-order traversal on the subtree of T containing the internal nodes only. Recall that a post-order traversal processes all vertices of a tree by recursively visiting all subtrees, then finally processing the root. Let $\{a_1, a_2, \ldots, a_{n-1}\}$ be the ordered sequence of internal vertices, where a_{n-1} is the root of T.

In the following description we will slightly abuse of notation, by applying the concepts of parent and children to nodes of S'. In this context, $p_{S'}(v)$, where $v \in V(S')$ has in-degree one, denotes the tail of v's unique incoming arc; $ls_{S'}(v)$ and $rs_{S'}(v)$, where $v \in V(S') \cap V(S)$, respectively refer to the nodes of S' that

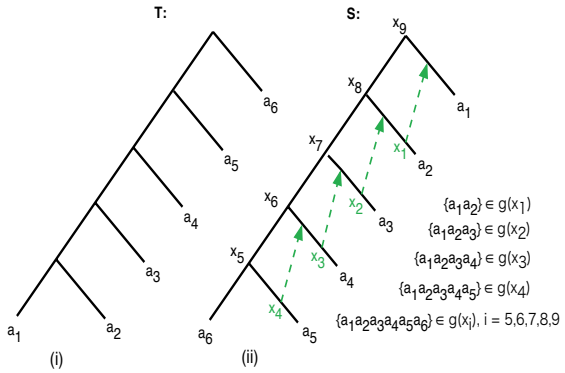

Fig. 4. An example of gene tree T (i) and the species tree S (ii) with $n = 6$. The dashed arcs in (ii) represent the lateral transfers. Note that 4 transfers are necessary for any scenario.

are first in the paths from v to $ls_S(v)$ and from v to $rs_S(v)$. We will now build a scenario (S', A', g) for T and S by creating the following lateral transfers:

– from arc $\langle p_{S'}(g^{-1}(ls_T(a_i))), g^{-1}(ls_T(a_i))\rangle$ to arc $\langle p_{S'}(g^{-1}(rs_T(a_i))), g^{-1}\ (rs_T(a_i))\rangle$, involving $\langle a_i, rs_T(a_i)\rangle$ from T,

for every i from 1 to $n-2$. Note that the post-ordering ensures that both children of a_i have been processed when vertex a_i is considered, hence $g^{-1}(rs_T(a_i))$ and $g^{-1}(rs_T(a_i))$ are defined. Also, observe that all lateral transfer in this scenario are incident on arcs of S that connect parents to leaves. Now, let $b = lca_{S'}(g^{-1}(ls_T(r(T)), g^{-1}(ls_T(r(T)))$, and for all vertices v_i belonging to the path connecting $ls_{S'}(b)$ to $g^{-1}(ls_T(r(T))$, place $ls_T(r(T) \in g(v_i)$. Similarly, for all vertices u_i belonging to the path connecting $rs_{S'}(b)$ to $g^{-1}(rs_T(r(T))$, place $rs_T(r(T) \in g(u_i)$. Finally, map all vertices in the path from the root of S to b, to the root of T.

It is straightforward to verify that the resulting scenario does not contain mixed cycles, since all transfers are non-intersecting by construction. In addition, all conditions from Definition 2 are satisfied, hence the procedure yields a valid 1-active scenario. It is also easy to see that the running time of this algorithm is linear in the size of the trees.

We now show that this upper bound is tight by giving a simple example of a gene tree and a species tree that require at least $n-2$ lateral transfers for any 1-active scenario.

Let L be the set of labels $\{a_i : 1 \leq i \leq n\}$, where n is a positive constant. Consider the gene tree T and species tree S shown in figure 4, both over the same set of labels L:

$$T = \prec \prod_{i=1}^{n} a_i \succ$$

$$S = \prec \prod_{i=n}^{1} a_i \succ$$

We will show that any 1-active scenario for T and S requires at least $n-2$ lateral transfers. The result follows from the following Claim.

Claim 4. *Let $X = \{p_T(a_i) : 1 < i < n\}$, that is, X is the set all internal nodes of T except the root. In any 1-active lateral transfer scenario for T and S, every element of X is tail of an arc involved in a lateral transfer.*

Proof. Assume that a node $x \in X$ is not tail of an arc of T involved in a lateral transfer. This means that $x \notin g(v)$, for all nodes $v \in V(S') \setminus V(S)$, which implies that there exists a node $y \in S$ such that $x \in g(y)$, where $y \geq_S lca_S(x)$. But $lca_S(x) = r(S)$ for all $x \in X$, hence $y = r(S)$. This is a contradiction, since $r(T) \in g(r(S))$, by Condition 1 of Definition 2, and $x \neq r(T)$. It follows that x must be involved in a lateral transfer. ❑

Therefore, any 1-active scenario for T and S has at least $|X| = n - 2$ lateral transfers.

References

1. Arora, S., Lund, C., Motwani, R. Sudan, M. and Szegedy M. (1998), Proof verification and hardness of approximation problems. *Journal of the ACM*, 45(3): 501-555.
2. Ausiello, G., Crescenzi, P., Gambosi, G., Kann, V., Marchetti Spaccamela, A., and Protasi, M. (1999), *Complexity and Approximation. Combinatorial Optimization Problems and their Approximability Properties*, Springer-Verlag, Berlin.
3. Berman, P., and Karpinski, M. (1998), On some tighter inapproximability results, further improvements, Technical Report TR98-065, *Electronic Colloquium on Computational Complexity* (ECCC); available online from `http://eccc.uni-trier.de/eccc-reports/1998/TR98-065/index.html`.
4. Berman, P. and Schnitger, G. (1992), On the complexity of approximating the independent set problem, *Information and Computation*, 96, 77-94.
5. Brown, J. R. (2003), Ancient horizontal gene transfer, *Nature Reviews, Genetics*, 4, 121-132.
6. Hallett, M. and Lagergren, J. (2001), Efficient algorithms for lateral gene transfer problems, Proc. 5^{th} Annual International Conference on Computational Molecular Biology (RECOMB), Montreal, Canada, 141-148.
7. Hallett, M. and Lagergren, J. (2004), Identifying lateral gene transfer events, submitted to *SIAM Journal of Computing* (available online from `http://www.mcb.mcgill.ca/~hallett/Lateral.pdf`)
8. Johnson, D.S. (1974), Approximation algorithms for combinatorial problems, *Journal of Computer and System Sciences*, 9, 256-278.
9. Page, R. D. M. and Charleston, M. A. (1997), From gene to organismal phylogeny: Reconciled tree and the gene tree/ species tree problem, *Molecular Phylogentics and Evolution*, 7, 231-240.
10. Papadimitriou, C. H. and Yannakakis, M. (1991), Optimization, approximation, and complexity classes, *Journal of Computer and System Sciences*, 43(3): 425-440.