DIMACS Technical Report 2005-14
May 2005

Inapproximability Results for the Lateral Gene Transfer Problem[1] (Preliminary Draft)

by

Bhaskar Dasgupta[2] Sergio Ferrarini[3] Uthra Gopalakrishnan[2]
Nisha Raj Paryani[2]

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

[2]Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607-7053. Email: {dasgupta,ugopalak,nparyani}@cs.uic.edu.

[3]Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza Leonardo da Vinci 32 20133, Milano, Italy. Email: sferrarini@gmail.com.

ABSTRACT

In this paper we establish some inapproximability results for the Lateral Transfer Problem. This optimization problem, which was defined by Hallet and Lagergren [3], is that of finding the most parsimonious lateral gene transfer scenario for a given pair of gene and species trees. We will prove that the Lateral Transfer Problem is MAX SNP-hard; thus *Polynomial Time Approximation Scheme* is not possible for it unless P = NP.

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* (a.k.a. horizontal gene transfers) that may occur during evolution and that generate "differences" between a gene and a species tree.

For these reasons, a single gene tree usually isn't 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. Interestingly, the reconciliation problem can be naturally formulated as an optimization problem, where the goal is to minimize the number 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. We consider the *lateral gene transfer model*, defined by Hallett and Lagergren in [3]. According to this model, all differences between the gene and the species trees are explained in terms of lateral gene transfer events. 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 TRANSFER PROBLEM.

In this paper we are interested in studying the approximability properties of the LATERAL TRANSFER PROBLEM. We will consider both the special case with activity level one and the general case with activity level α and will establish some inapproximability results. More precisely, we will prove that these problems are MAX SNP-hard, by finding a special approximation-preserving reduction from a problem in MAX SNP. This result will allow us to conclude that no *Polynomial Time Approximation Scheme* is possible for the LATERAL TRANSFER PROBLEM, unless P = NP.

2 Definitions

In the following 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 $\langle u, v \rangle$, $\langle u, v' \rangle \in A(T)$, and we denote the parent of a vertex $u \in V(T)$ 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.

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 $A(G) \cup A, E(G)$, and vertices $V(G)$. Similarly is defined $G \cup E$, where E is a set of vertices. 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.

3 Basic Concepts

In this section we shall define some basic concepts for the evolutionary model we will adopt. We will first briefly introduce the concepts of *gene tree* and *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 trees is treated as a subset (or *cluster*) which contains the labels of all leaves of the subtree rooted in the 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 species trees, 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 as $lca_Y(X)$, is defined as the node $y \in Y$ such that $X \subseteq y$ and $X \nsubseteq w$ 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.

3.1 Lateral Gene Transfer Model

We are now ready to introduce the evolutionary model based on the concept of *lateral gene transfer*. This model was developed by Hallet and Lagergren [3]. 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 arouse. 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* $S' \cup A'$*.*

Basically, 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, which clearly is a required property for the model.

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 α parameter 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 particular case of the 1-active scenario, and then state the definition for the general α-active case, where $\alpha \geq 2$.

Definition 2 *A* 1-active lateral transfer 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 = 0,1,2, and* x_i *is the* $\leq_{s'}$*-maximal vertex such that* $v_i \in g(x_i)$ *for* $i = 1, 2$*.*

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 *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* $i = 0, 1, 2$, *and* x_i *is the* $\leq_{S'}$*-maximal vertex such that* $v_i \in g(x_i)$ *for* $i = 1, 2$.
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:

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

F is the *set of lateral transfers* for the gene tree T w.r.t. the scenario (S', A', g).

3.2 Problem Definitions

In the next sections we are interested in studying the approximability properties of the following optimization problems:

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

α-Active Transfer Problem
Instance: A species tree S and a gene tree T such that $L(T) \subseteq L(S)$, a constant $\alpha \geq 1$.
Question: 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-Transfer Problem and α-Active, τ-Transfer Problem respectively.

4 Inapproximability Results

It has been shown in [4] that the α-Active, τ-Transfer Problem is NP-complete. We will now focus on the optimization version of such problem, the α-Active Transfer Problem, and study its approximability properties. The goal of this section is to establish some inapproximability results for such problem. In particular, we will prove that no polynomial time approximation algorithm can achieve an *approximation ratio* of $(1 + \varepsilon)$, for some positive constant ε, unless P = NP. Lets first introduce some preliminary concepts.

4.1 Polynomial Time Approximation Schemes, MAX SNP and L-Reductions

Let A be a minimization problem and OPT be the cost of its optimal solution. An algorithm achieves an *approximation ratio (or factor)* k, if, for every instance of A, it produces a solution of value at most $k\mathsf{OPT}$. A *polynomial time approximation scheme (PTAS)* for A is an algorithm that, for any given constant $\varepsilon > 0$, approximates the problem within a factor $(1 + \varepsilon)$ in time that is polynomial in the input size.

The class of *MAX SNP* problems was defined by Papadimitriou and Yannakakis [5]. They also defined an approximation-preserving reduction, the so-called *L-reduction*. This definition was then slightly modified in an equivalent version by Berman and Schnitger [2]. We will give this modified version, since it better represents the reduction we are performing in our proof.

Definition 4 [2] [5] *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 constants* α, $\beta > 0$ *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', $\mathsf{OPT}(I)$ *and* $\mathsf{OPT}(I')$, *respectively, satisfy* $\mathsf{OPT}(I') \leq \alpha\mathsf{OPT}(I)$.

2. *For any solution of* I' *with cost* c', *algorithm* T_2 *produces another solution with cost* $c'' \leq c'$, *and algorithm* T_3 *produces a solution of* I *of* Π *with cost* c *(possibly from the solution produced by* T_2*) satisfying* $|c - \mathsf{OPT}(I)| \leq \beta\,|c' - \mathsf{OPT}(I')|$.

An optimization problem is MAX SNP-*hard* if any problem in MAX SNP L-reduces to the problem. If this problem is also in MAX SNP, then it is MAX SNP-*complete*. The importance of showing MAX SNP-hardness for an optimization problem comes from a result proved by Arora et al. [1]. According to such result, all MAX SNP-hard problems do not have a PTAS, unless P = NP. In other words, for this class of problems there exists a constant $c > 1$ such that no polynomial time algorithm can achieve a better approximation ratio than c.

4.2 MAX SNP-hardness of the 1-ACTIVE TRANSFER PROBLEM

We will first consider the special case of finding the most parsimonious 1-active lateral transfer scenario, and then extend the result to the general model with activity level $\alpha \geq 1$.

Theorem 5 *The* 1-ACTIVE TRANSFER PROBLEM *is MAX SNP-hard.*

Proof. In the following we will show that MAX 3SAT-B L-reduces to the 1-ACTIVE TRANSFER PROBLEM. MAX 3SAT-B is a special version of MAX 3SAT where the number of occurrences of each variable is bounded by a constant B. It is known from [5] that MAX 3SAT-B is MAX SNP-complete, 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 3-CNF, where each clause C_i is on three variables from X and where the number of occurrences of each variable is bounded by a constant B. The goal ofMAX 3SAT-B is to find a truth assignment on X that maximizes the number of satisfied clauses. Given an instance of MAX 3SAT-B, we will now exhibit how to build an instance of the 1-ACTIVE TRANSFER PROBLEM s.t. conditions (1) and (2) from Definition 4 are both satisfied. In other words, we will construct a gene tree T and a specie tree S from Φ and prove that this transformation is an *L-reduction*.

The construction of G and S from Φ is taken from the NP-completeness proof given in [4]. To ease our presentation, we will adopt the same notation used in [4] throughout the proof. The details on how to build the two trees are here omitted.

In our case parameter τ^+ is defined as $\tau^+ = \tau + m + 1$, while the definition of τ is unchanged, i.e. $\tau = 14m + 6k$, where

$$k = \left|\{\langle i, j, i', j' \rangle \mid C_{i,j} = C_{i',j'} \text{ or } C_{i,j} = \bar{C}_{i',j'}, 1 \leq i < i' \leq m, 1 \leq j, j' \leq 3\}\right|$$

Since the maximum number of occurrences for each variable is bounded by B, $k \leq \frac{nB(B-1)}{2}$. Thus, $\tau \leq \gamma m$, where $\gamma = 14 + 9B(B-1)$. Notice that in this case τ does not represent the maximum cost for the scenario.

Let Ψ be a truth assignment on the variable set X, i.e. $\Psi : X_i \longmapsto \{\text{true}, \text{ false}\}$, for every $i = 1, \ldots, n$.

Claim 6 *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 $\tau + (m - \rho)$.*

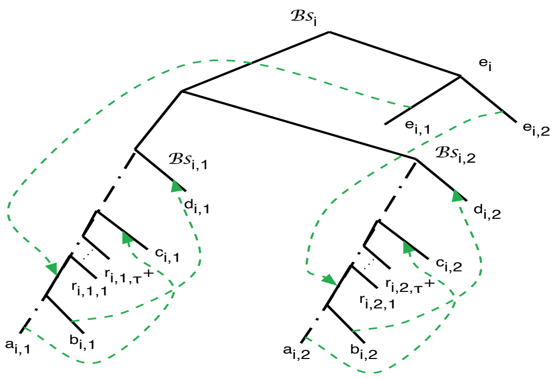

Figure 1: Lateral transfer scenario for subtree B_{S_i} when the corresponding clause C_i is not satisfied.

Proof. Apply the same construction adopted in [4] to build a scenario from an assigned Ψ. This construction assumes that Ψ satisfies every clause of Φ. Thus, a problem arises for those clauses which are not satisfied by Ψ. We will simply create an additional transfer for each non satisfied clause C_i to preserve the consistency of the scenario, while maintaining the rest of it unchanged.

For each non satisfied clause C_i, create for every $j = 1, 2, 3$ a lateral transfer from arc $\langle p_S(e_{i,j}), e_{i,j}\rangle$ to arc $\langle p_S(p_S(b_{i,j})), p_S(b_{i,j})\rangle$, involving arc $\langle p_T(e_{i,j}), lca_T(c_{i,j}, d_{i,j})\rangle$ from T (see figure 1). This requires three lateral transfers, as opposed to the two transfers necessary for each satisfied clause. Hence, for every non satisfied clause we need to create an additional transfer, leading to a scenario of 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 6 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 $OPT_{T,S} \leq \tau + (m - OPT_\Phi) \leq (\gamma + 1)m$. Since we can assume w.l.o.g. that $OPT_\Phi \geq m/B$, by combining the two inequalities we have $OPT_{T,S} \leq \alpha OPT_\Phi$, where $\alpha = (\gamma + 1)/B$. 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 w.l.o.g. that $c' \leq \tau + m$, otherwise we could choose any scenario built from an arbitrary assignment. Observe that Claims 1-9 in [4] are true for the given scenario. We will show that, starting from this 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 $\beta = 1$.

Let X_ν be a variable from the variable set X and $\Omega_\nu = \{C_{i,j} \mid X_\nu \text{ appears in } C_{i,j}\}$, $\omega_\nu = |\Omega_\nu|$. 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_ν be a forest of subtrees of T defined as $F_\nu = \{\tilde{T}_{i,j} \mid C_{i,j} \in \Omega_\nu\} \cup \{\varepsilon'_{i,j,i',j'} \mid C_{i,j}, C_{i',j'} \in \Omega_\nu\}$. Moreover, let $k_\nu = |\{\langle i, j, i', j'\rangle \mid C_{i,j}, C_{i',j'} \in \Omega_\nu \text{ and } i < i'\}|$ and define $\tilde{\tau}_\nu = 2\omega_\nu + 4k_\nu$.

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* w.r.t. $C_{i',j'}$ 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 $lca_T(c_{i',j'}, d_{i',j'}) \notin g(lca_S(a_{i',j'}, b_{i',j'})$

- $C_{i,j} = \bar{C}_{i',j'}$, 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 7 *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 [4], $\tilde{\tau}$ is a lower bound for 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 [4]. A similar reasoning establishes the same result for the cases where $C_{i,j} = \bar{C}_{i',j'}$.

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.

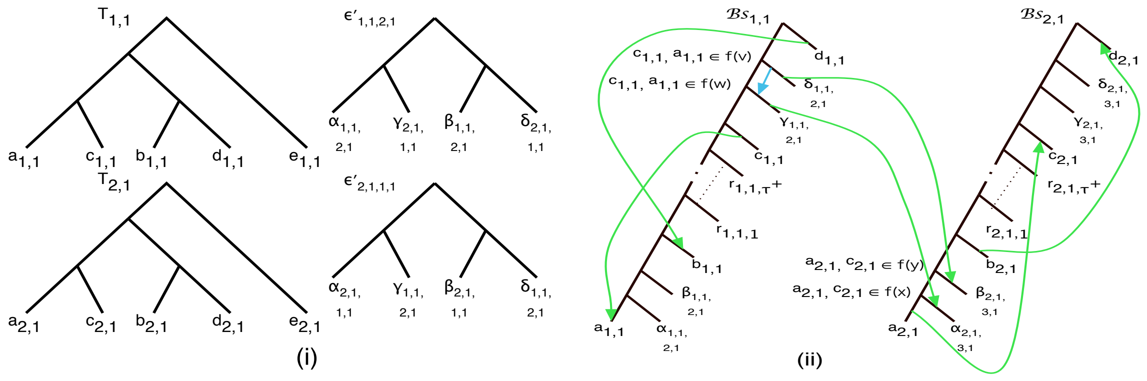

Figure 2: Consider the formula $\Phi = (X_1 \vee \bar{X}_2 \vee X_3)(X_1 \vee X_2 \vee \bar{X}_3)$. (i) depicts the twister gadgets $(T_{1,1}, T_{2,1})$ and enforcement gadgets $(\varepsilon'_{1,1,2,1}, \varepsilon'_{2,1,1,1})$ corresponding to the literals $(C_{1,1}, C_{2,1})$ which are instances of $X_1 \in X$. (ii) represents a partial scenario where literals $C_{1,1}$ and $C_{2,1}$ are well-assigned and inconsistent. 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.

❑

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 $\Psi : X \longrightarrow \{\text{true}, \text{false}\}$ on Φ in the following way. For each variable X_ν, $\nu = 1, \ldots, n$, check if there exists (can be done in linear time) some literal $C_{i,j} \in \Omega_\nu$ which is well-assigned. If this is the case, assign to X_ν the truth value read from $C_{i,j}$, i.e. $\Psi(X_\nu) = \text{true}$ if $lca_T(c_{i,j}, d_{i,j}) \in lca_S(c_{i,j}, d_{i,j})$ and $\Psi(X_\nu) = \text{false}$ if $lca_T(c_{i,j}, d_{i,j}) \in lca_S(a_{i,j}, b_{i,j})$. If no literal in Ω_ν is well-assigned, then assign to X_ν an arbitrary truth value. Now follow the procedure described in Claim 6 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 two elements from the set $\{root_T(T_{i,j}) \mid 1 \leq j \leq 3\}$ are tails of arcs involved in transfers. 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 1-9 in [4] 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 $C_{i,j} \in \Omega_\nu$ and X_ν is arbitrarily assigned) or by a literal which is inconsistent w.r.t. the chosen assignment $\Psi(X_\nu)$. By Claim 7, this implies that LTS_I has at least one 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.

❑

The following observation is based on the previous arguments and will be useful in the following sections.

Observation 8 *An optimal lateral transfer scenario for* T *and* S *has cost* $\tau + (m - \mathsf{OPT}_\Phi)$.

4.3 MAX SNP-hardness of the α-ACTIVE TRANSFER PROBLEM

Theorem 9 *The* α-ACTIVE TRANSFER PROBLEM *is MAX SNP-hard,* $\alpha \geq 1$.

Proof. Once again, we will L-reduce from MAX 3SAT-B. Let T^* and S^* respectively be the gene and specie tree of the α-ACTIVE TRANSFER PROBLEM instance. Build T^* and S^* by following the procedure given in [4]. 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 6. Use this scenario to construct an α-active scenario for T^* and S^* as shown in [4]. The cost of this scenario is $\tau^* + (m - \mathsf{OPT}_\Phi)$, where $\tau^* = \tau + (\alpha - 1)$. Hence, $\mathsf{OPT}_{T^*,S^*} \leq \tau^* + (m - \mathsf{OPT}_\Psi)$, and the first condition from Definition 4 follows.

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 [4], 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 $c^* = c' + (\alpha - 1) \geq \tau^* + (m - \rho)$. This concludes our proof. ❑

References

[1] Arora, S., Lund, C., Motwani, R. Sudan, M. and Szegedy M.. Proof Verification and Hardness of Approximation Problems. *Journal of ACM*, 45(3): 501-555, 1998.

[2] Berman, P. and Schnitger, G. (1992) On the Complexity of Approximating the Independent Set Problem, *Inform. and Comput.*, 96, pp. 77-94.

[3] Hallett, M. and Lagergren, J. (2001) Efficient Algorithms for Lateral Gene Transfer Problems. 5^th *Annual RECOMB '01*, Montreal, Canada, pp. 141-148.

[4] Hallett, M. and Lagergren, J. (2004) Identifying Lateral Gene Transfer Events. Full journal version. Submitted to *SIAM J. Comp.*. Available at *http://www.mcb.mcgill.ca/ hallett/Lateral.pdf*.

[5] Papadimitriou, C. H. and Yannakakis, M.. Optimization, approximation, and complexity classes. *Journal of Computer and System Sciences*, 43(3): 425-440, December 1991.