Efficient Algorithms for Lateral Gene Transfer Problems

M. T. Hallett*, J. Lagergren†

ABSTRACT

This paper develops a model for *lateral gene transfer events* (a.k.a. horizontal gene transfer events) between a set of gene trees $T_1, T_2, \ldots, T_k$ and a species tree S. To the best of our knowledge, this model possesses a higher degree of biological and mathematical soundness than any other model proposed in the literature. Among other biological considerations, the model respects the partial order of evolution implied by S. Within our model, we identify an *activity parameter* that measures the number of genes that are allowed to be simultaneously active in the genome of a taxa and show that finding the most parsimonious scenario that reconciles the disagreeing gene trees with the species tree is doable in polynomial time when the activity level and number of transfers are small, but intractable in general. To the best of our knowledge, all other models proposed in the literature assume implicitly that the activity is one. Finally, using a dataset of bacterial gene sequences from [4], our implementations found 5 optimal scenarios; one of which is the scenario proposed by the authors in [4].

1. INTRODUCTION

The completion of various bacterial sequencing projects has reinforced the view that evolutionary relationships between taxa (i.e. the species trees) cannot be inferred from a single gene family (i.e. a single gene tree) due to evolutionary events such as *gene duplication*, *gene loss*, *gene convergence*, and *horizontal gene transfer* (see [4, 5, 8, 9, 12, 13, 18]). Recent findings have also motivated the genomics community to determine how ubiquitous these events are throughout evolution. This paper is primarily concerned with answering this question w.r.t. lateral gene transfer events (a.k.a. *horizontal transfer* events) caused by, for example, the presence of *homing endonucleases* and *recombination*. Given a (hypothetically correct) species tree S and a set of (hypothetically correct) gene trees $T_1, \ldots, T_k$ not necessarily pairwise equal or equal to S, the goal is find the minimum number of lateral transfers necessary to explain this "disagreement". This allows us to identify edges of $T_1, \ldots, T_k$ that corresponds to lateral transfers, and where in S these transfers occur.

Early attempts to use generalized evolutionary models to allow for lateral transfers such as the *network model* [17, 8, 9], inspired several papers examining the *subtree transfer* operation on leaf labeled trees [2, 3, 10, 11]. In a subtree transfer operation, a subtree T' of a tree T and an edge e not contained in T' are chosen, a new vertex v is introduced by subdividing e, and T' is transferred so as to make its root the second child of v. Given two leaf labeled trees, the SUBTREE TRANSFER problem asks to find the minimum number of subtree transfer operations that transform one tree into the second. This problem is known to be NP-complete for unrooted binary trees but approximable to within a factor of 3 [10]. Weighted versions of the problem are also known to be approximable to within a logarithmic factor under a linear cost scoring scheme [2, 3].

This paper uses another approach that has received considerable attention over the past several years. It is similar to that applied to the DUPLICATION-LOSS problem, introduced in [5]. In this problem, the input consists of a distinguished tree called the *species tree* and a set of *gene trees*. Through the postulation of *gene duplication* and *gene loss* events, one builds a *mapping* of the gene trees into the species tree from which an evolutionary meaningful explanation or *reconciliation* can be derived. Such problems have been studied extensively (see [6, 7, 14, 15, 16] amongst others) and in several different contexts. For example, in [1] a mapping between host and parasite trees is developed which allows for both *host switching* and *parasite duplication*.

This paper formulates a model of lateral transfers which is biologically sound (it captures many impor-

*McGill Centre for Bioinformatics, McGill University, Montreal, Canada, Dept. of Computer Science, hallett@cs.mcgill.ca

†Stockholm Bioinformatics Center and Dept. of Numerical Analysis and Computer Science, KTH, Stockholm, Sweden, jensl@nada.kth.se

RECOMB 2001, Montreal, Canada

tant aspects of the biological reality) and also mathematically sound (the definition is precise). To the best of our knowledge previous models have failed with respect to at least one of these desirable properties. In our model the gene tree is mapped into the species tree as in [1, 5, 14] by a mapping satisfying certain biologically sound conditions. Furthermore, we address issues of intractability similar to those mentioned in [1], identify several key parameters in our model, and show that the reconciliation problem becomes tractable when these parameters are bounded.

When unrooted trees are used in the subtree transfer model, the only notion of "time" in the model is the condition that the chosen edge e not be in the chosen subtree. We use rooted trees and build our model in such a way that the partial order of evolution implied by S is respected. Also, any lateral transfer extends the partial order implied by S and, in our model, all such extensions are guaranteed to remain consistent. Consider, for example, the gene and species tree in Figure 1 (a), (b) resp. In a rooted version of the subtree transfer model, we can transform T into S via two transfers: subtree B to arc $\langle AC, A\rangle$, followed by a transfer of the subtree E to arc $\langle DF, D\rangle$. However, this reconciliation of T with S has no meaningful interpretation. In Figure 1 (b), one can see that the ancestor of A and B in S must have existed earlier than the ancestor of E and F in S and hence the ancestor of A, B and C in S must have existed previous to the ancestor of E and F. This in turn implies that the transfer from the ancestor of A, C, and E in T to E is a "forward" transfer.

Implicit in both the subtree transfer model and the models of [1] for co-speciation, there is an assumption that only one gene per gene tree (i.e. gene family) may exist in a genome at any point in the evolution of the taxa. When multiple copies of a gene are postulated to be present, this multiplicity is presumed to be the product of *gene duplication*. This need not be the case. Such multiplicity can occur through lateral transfer events alone. To capture this idea, we introduce the notion of *α-activity*. An α-activity scenario for a gene tree and species tree allows only α copies of a gene to simultaneously exist in the genome of an ancestral taxa. Figure 2(c) shows an example of a lateral transfer scenario that is 2-active.

The contributions of this paper are as follows. In Section 2, we introduce our notation. Section 3 gives our model for lateral gene transfer. In Section 4, we show that the α-activity, $\alpha \geq 1$, version of this problem remains NP-complete but can be solved in time $O(2^{2(\alpha+\tau)} \cdot (\alpha+\tau)^{\tau} \cdot \tau^{\tau} \cdot |L|^2)$, where τ is the number of transfers. In Section 5, we describe an algorithm for a fixed number of transfers τ for the special case of 1-activity, i.e. the version which allows for only one gene per gene family to exist in any taxon during the evolution. This algorithm generates a much smaller search tree than the general version and runs in time $O(2^{\tau} \cdot |L|^2)$. Lastly, we show that the 1-activity problem is, in the general case, NP-complete. Section 6 details the results of applying our algorithms to data from the literature. Section 7 gives a number of open problems and future directions related to this work.

2. DEFINITIONS

Throughout this paper, we will consider rooted directed trees where the arcs are directed from the root towards the leaves and a vertex has outdegree at most 2. We will call such a tree a *rooted* tree. By a rooted forest we mean a union of disjoint rooted trees. For a rooted forest F and a vertex v of F the number of outgoing arcs from v is denoted $d^+(v)$ and the number of ingoing arcs at v is denoted $d^-(v)$. For such a tree T, $V(T)$ denotes its set of vertices and $A(T)$ its set of arcs. The set of leaves of a rooted forest F is denoted $L(F)$ (i.e. the vertices without outgoing arcs). The *internal vertices* of T are $V(T) \setminus L(T)$. Both a *gene tree* and a *species tree* are binary directed trees. The *root* of a tree T is denoted $r(T)$. The parent of a vertex v in T is denoted $p_T(v)$. Let T be a directed binary tree. We will refer to the vertices u, u' such that $\langle v, u\rangle, \langle v, u'\rangle \in A(T)$ as the children of v in T. When convenient we will assume that each internal vertex $v \in V(T)\setminus L(T)$ has a left child denoted $c_l(v)$ and a right child denoted $c_r(v)$. For $u \in V(T)$, any vertex v reachable from u by a directed path is a *descendant* of u (this means that u is a descendant of u). We denote this by $v \leq_T u$. We also say that u is an *ancestor* of v ($u \geq_T v$). We say that v is a *proper descendant* (*proper ancestor*) of u, if $v \leq_T u$ ($v \geq_T u$) and $v \neq u$ and denote this relationship by $v <_T u$ ($v >_T u$). Two vertices $u, v \in V(T)$ are *incomparable* in T, if not $u \leq_T v$ and not $v \leq_T u$. We will sometimes denote a singleton set $\{x\}$ by x.

Let F be a rooted forest. For a vertex $u \in V(F)$, let F_u be the rooted subtree of F consisting of vertices of $V(F)$ reachable by directed paths from u. Let T be a rooted tree. For $X \subseteq L(T)$, the *least common ancestor* of X in T, written $lca_T(X)$, is defined as follows: if $X = \{v\}$, then $lca_T(X) = v$; otherwise, $lca_T(X)$ is the vertex v such that $X \subseteq L_T(v)$ but $X \not\subseteq L_T(u)$ for each proper descendant u of v. For $U \subseteq V(T)$, let T_U be the subtree of T rooted at $lca_T(U)$. We let $L_T(U)$ be the subset of $L(T)$ in T_U. For a pair x, y, we use $lca_F(x, y)$ to denote $lca_F(\{x, y\})$. For a set $U \subset V(T)$, $T[U]$ denotes the forest of subtrees induced by U. Let F be a forest and S a species tree such that $L(F) = L(S)$. The mapping $\lambda_{F,S} : V(F) \to V(S)$ is defined as follows: $\lambda_{F,S}(v) = lca_S(L_F(v))$.

A *mixed graph* G is a graph containing arcs as well as undirected edges. The arcs of G are denoted $A(G)$, the edges $E(G)$, and the vertices $V(G)$. If G is a mixed graph and A is a set of arcs, then $G \cup A$ is used to denote the mixed graph with arcs $A(G) \cup A$, $E(G)$, and vertices $V(G)$. For a set of edges E, $G \cup E$ is defined similarly. A *directed mixed cycle* is a mixed graph where each vertex has total degree 2, which contains arcs, and which can be traversed in the direction of all its arcs. If A is a set of arcs, then $E(A)$ denotes the underlying undirected edges, i.e. $E(A) = \{(u, v) : \langle u, v\rangle \in A\}$. For $u, v \in V(T)$, let $P^T_{u,v}$ be the unique (undirected) path

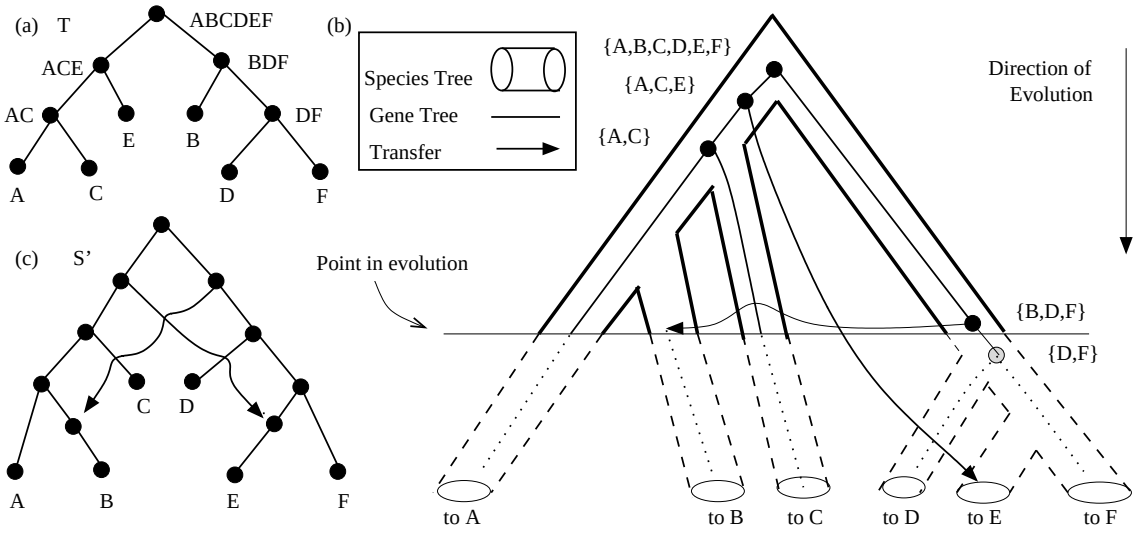


Figure 1: (a) The gene tree T. (b) An illegal scenario describing the evolution of the gene tree in a species tree. (c) The subdivision S' of S implied by the two lateral transfers in Figure 1(b). Note that the lateral transfer scheme does not satisfy the conditions in Definition 1 (Lateral Transfer Scheme) since it contains a directed mixed cycle.

between u and v in T. A path P in a graph G avoids a set $U \subseteq V(G)$ if and only if $V(P) \cap U = \emptyset$.

3. THE MODEL

Let S be a species tree and T be a gene tree. A lateral transfer effects two arcs $\langle x, y\rangle$, $\langle x', y'\rangle$ of S and one arc $\langle u, v\rangle$ from T; some portion of the evolution represented by $\langle u, v\rangle$ occurs along $\langle x, y\rangle$ in S, the transfer occurs, and the remaining portion of evolution occurs along $\langle x', y'\rangle$ in S. To model this, we subdivide both arcs in S and place an arc in the direction of the transfer between the newly created vertices. Figure 1(c) shows the subdivision S' of S implied by the lateral transfers in Figure 1(b).

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 $S' \cup E(A')$ does not contain a directed mixed cycle,*
2. *the tail of each arc in A' has indegree 1 and outdegree 2 in $S' \cup A'$, and*
3. *the head of each arc in A' has indegree 2 and outdegree 1 in $S' \cup A'$.*

If we demand that, at any point during the evolution represented by S, a taxa may only have one copy of any gene (i.e. the case of *1-activity*), then a lateral transfer scheme uniquely specifies a gene tree T such that $L(T) = L(S)$. When we allow $\alpha > 1$ genes to exist at any point during the evolution, this is no longer the case. We give a formal definition of α-activity below. For 1-activity, the unique tree corresponding to a lateral transfer scheme (S', A') for S with the same set of labels as S is the tree T obtained as follows. Let F be defined by $V(F) = V(S')$ and $A(F)$ is the set of arcs $\langle x, y\rangle \in A(S') \cup A'$ such that there is a $l \in L_{S'}(y)$ such that $P^{S'}_{y,l}$ avoids $H(A') \setminus \{y\}$ where $H(A') = \{y : \exists x, \langle x, y\rangle \in A'\}$. Finally, let T be the gene tree obtained from the connected component of F containing $r(S')$ by, recursively, short-cutting vertices of degree 2.

Given a gene tree T and a species tree S, we are interested in postulating the minimum number of lateral transfers that explains in an evolutionary meaningful way why the gene tree T is not equal to the species tree S. In order to be biologically meaningful, our *scenario* must satisfy the following constraints.

DEFINITION 2. *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 $g : V(S') \to 2^{V(T)}$ such that:*

1. *$T[g(r(S'))]$ is connected and $r(T) \in g(r(S'))$;*
2. *if v_1 and v_2 are children of v_0 in T and $v_1, v_2 \notin g(r(S'))$, then there exists x_0 with children x_1 and x_2 in $S' \cup A'$ (where $x_1 \neq x_2$) s.t. $v_i \in g(x_i)$, for $i = 0, 1, 2$ (moreover, the definition implies that x_i is the $\leq_{S'}$-maximal vertex such that $v_i \in g(x_i)$ for $i = 0, 1, 2$, and x_0 is the $\leq'_S$-minimal vertex such that $v_0 \in g(x)$);*
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 exist a child x of $r(S')$ in S' s.t. $v_2 \in g(x)$;*

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* $x \in V(S')\backslash\{r(S')\}$;
6. $g(l) = \{l\}$, *for all* $l \in L(S)$

We will refer to $g(x)$ as the *bag* of x. A lateral transfer scenario (S', A', g) is α*-active* if and only if $\max_{x \in S'} |g(x)| = \alpha$. The *transfer number* of (S', A', g) w.r.t. T is the sum over all $\langle x, y\rangle \in A'$ of

$$|\{\langle u, v\rangle \in A(T) : u \in g(x), v \in g(y)\}| + |L(T[g(r(S))])| - 1.$$

The α*-activity transfer number* of S and T, denoted $\tau_\alpha(S, T)$, is the minimum transfer number of any α-activity lateral transfer scenario (S', A', g) for S and T. A α-activity lateral transfer scenario (S', A', g) for S and T with transfer number at most τ is called a α, τ *lateral transfer scenario* for S and T.

We briefly describe each of the six conditions. Condition 1 guarantee that the root in the species tree corresponds to the root of the gene tree (one can have a more general model that does not demand this), but allows lateral transfers to have taken place "previous to the the root of S". Condition 2, 3 and 4 guarantee that the tree T exists within the species tree S in a connected manner and a manner that respects the direction of evolution implied by the arcs of T. The stipulation that $x_1 \neq x_2$ in Condition 2 together with Condition 4 disallows *gene duplication events* (see Figure 2.a). Consider a vertex v such that $v \in g(x)$ for some $x \in V(S')$. Condition 2 and 6 together with the definition of lateral transfer scheme forbid the case where there both outgoing arcs from a vertex in the gene tree correspond to lateral transfers. We believe this to be a valid restriction, since when interpretted in a biological sense, this means that a gene existed in the genome of a taxon at some point in the evolution represented by S but no descendant in the subtree of S' rooted by $g(x)$ has a copy of this gene (unless the gene was first transferred "away" and then transferred "back" to this subtree).

Lastly, our definition of *transfer number* also deserves some explanation. When $\alpha = 1$, only the root of T, $r(T)$, belongs to $g(r(S'))$ (by condition (1) and the definition of 1-activity). For $\alpha >$ 1-activity, this is no longer necessarily the case. Now, in essence, the vertices of a subtree of T may belong to $g(r(S'))$. This implies the existence of $|L(T[g(r(S))])| - 1$ transfers previous to the evolution described by S. We call this model the "external model" and use it throughout the remainder of the paper. An "internal model" is also possible where all transfer must take place within S. It is easy to verify that all of our results for the external model also hold in the internal model. The model can be generalized such that the lateral transfers that have happened previous to the root r, and which cause $g(r)$ not to be an antichain, also effect other vertices than the root.

4. ACTIVITY α AND τ TRANSFERS

We now examine the general version of the α-activity τ-transfer problem for $\alpha \geq 1$. Our first result shows this problem to be intractable in general. However, for small values of α and τ, the problem is fixed parameter tractable. In the next section, we give a faster algorithm for the special case of 1-activity.

THEOREM 1. *The decision version of the* α-ACTIVITY τ-TRANSFER *problem is NP-complete.*

PROOF. Omitted. □

However, for small values of α and τ where $\alpha \leq \tau$, the problem remains tractable. We give our algorithm and proof of correctness below:

α-ACTIVITY τ-TRANSFER problem

Input: A species tree S and a gene tree T with α-activity number $\leq \tau$.
Output: The minimum transfer number $\tau^* \leq \tau$ over all α-activity lateral transfer scenarios for S and T.

For the remainder of this section, we will assume that S is a species tree and that T is a gene tree over a label set L such that $\tau_\alpha(S, T) \leq \tau$. For any two trees T_1, T_2 and arc $\langle u, v\rangle \in A(T_2)$, we will abuse notation slightly and use $\lambda^{-1}_{T_1,T_2}(u, v)$ to denote the set $\{\langle x, y\rangle \in A(T_1) : \lambda_{T_1,T_2}(x) \geq_{T_2} u, \lambda_{T_1,T_2}(y) \leq_{T_2} v\}$.

A *board* is a pair (F, U) where $F \subseteq T$ is a directed forest without isolated vertices such that $d^+_F(v) \leq 1$ for all $v \in V(F)$ and $U \subseteq V(T)$ induces a subtree of T containing $r(T)$. (This means that F is the union of a number of disjoint paths.) The *cost* of (F, U) is $|A(F)| + |L(T[U])| - 1$. A *move* is an arc $e \in A(T)$. It is *valid* for (F, U) if and only if, for some $u \in V(T)$, $e = \langle u, c_r(u)\rangle$ and $\langle u, c_l(u)\rangle \notin F$ or $e = \langle u, c_l(u)\rangle$ and $\langle u, c_r(u)\rangle \notin F$. The *board obtain by making move* e *on* (F, U) is $(F \cup \{e\}, U)$. We will assume that all moves we make are valid.

Let (S', A', g) be a lateral transfer scenario for S and T. The *lateral transfers* of T w.r.t. (S', A', g) is the following set of arcs

$$\{\langle u, v\rangle \in A(T) : \exists \langle x, y\rangle \in A', u \in g(x) \text{ and } v \in g(y)\}.$$

Let (S', A', g) be a lateral transfer scenario for S and T. The board (F, U) is the *board* of T w.r.t. (S', A', g) iff F is the transfers of T w.r.t. (S', A', g) and $U = g(r(S'))$. Observe that if (S', A', g) has transfer number τ and B is the board of T w.r.t (S', A', g) then the cost of B equals τ. A board (F, U) is a α, τ-win iff there is an α, τ lateral transfer scenario (S', A', g) such that (F, U) is the board of T w.r.t. (S', A', g). The board (F, U) is cyclic iff there is no lateral transfer scenario (S', A', g) such that (F, U) is the board of T w.r.t. (S', A', g).

For two boards (F, U) and (F', U'), we write

$$(F, U) \prec (F', U')$$

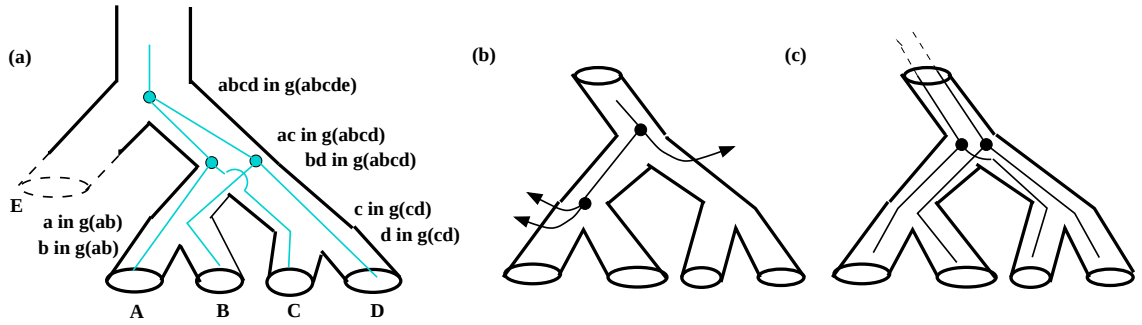


Figure 2: (a) An example of a scenario with a duplication. (b) An example of a scenario with two simultaneous transfers. Both scenarios are disallowed by Condition 2 and Condition 5 resp. (c) An example of a scenario with activity 2.

if and only if $U = U'$ and $F \subseteq F'$. The arcs $e_1, \ldots, e_l$ are *alternative forced moves* if and only if it is the case that the boards $B_1, \ldots, B_l$ obtained by making moves $e_1, \ldots, e_l$ on B satisfy the following: there is a τ, α-win W of cost t such that $B \prec W$ if and only if there is a τ, α-win W' of cost $\leq t$ such that $B_i \prec W'$ for some $i \in [l]$.

The following is an abstract version of the α-Activity τ-Transfer algorithm. A *trivial reduction* on a board (F, U) is performed by for a pairs of $u, v \in L(S)$ $(= L(T \setminus F))$ which are siblings in S as well as in $T \setminus F$ removing v and shortcutting the parent of u in S as well as in $T \setminus F$. Whenever a trivial reduction is possible in the algorithm it will be performed. Let C be the set of boards $B = (F, U)$ of cost at most $\tau - 1$ such that $U \subseteq \lambda^{-1}_{T,S}(r(S))$ and $F = \emptyset$. Let $t = \infty$. The output of the algorithm is t. We repeat the following until $C = \emptyset$.

1. Take $B \in C$ and let $B_1, \ldots, B_l$ be the boards obtained from B by making the alternative forced moves $e_1, \ldots, e_l$.
2. For each of $B_1, \ldots, B_l$, if B_i is an α, τ-win, let $C \leftarrow C \setminus \{B\}$ and let $t \leftarrow \min(t, 1 + c)$ where c is the cost of B_i.
3. else if $|B| = \tau - 1$, let $C \leftarrow C \setminus \{B\}$
4. else $C \leftarrow C \setminus \{B\} \cup \{B_1, \ldots, B_l\}$.

Via the lemmata given below, it is always possible to identify $\leq \alpha + \tau$ alternative forced moves in a board of cost $< \tau$. It also follows from the lemmata below that the number of ways to chose U is bounded by $2^{2(\alpha+\tau)}$. This implies that our algorithm runs in time $O(2^{2(\alpha+\tau)} \cdot (\alpha + \tau)^{\tau+1} \cdot |L|^2)$.

Lemma 1. *Let (S', A', g) be a lateral transfer scenario for S and T with lateral transfers F. For each v satisfying $d^+_{T \setminus F}(v) = 2$, it holds that $v \in g(\lambda_{T \setminus F, S}(v))$.*

Proof. Omitted. □

Lemma 2. *Let (F, U) be a board. Let M be the set of $\leq_T$-minimal vertices of $\lambda^{-1}_{T \setminus F, S}(x)$ and let $e_1, \ldots, e_l$ be the outgoing arcs from vertices of M. If $|M| > \alpha$, then $e_1, \ldots, e_l$ are alternative forced moves in (F, U).*

Proof. Assume that (F', U) is an α, τ-win such that $(U, F) \prec (F', U)$ and $e_1, \ldots, e_l \notin A(F')$. Since

$$e_1, \ldots, e_l \notin A(F'),$$

$\lambda_{T \setminus F', S}(v) = \lambda_{T \setminus F', S}(u)$, for any $u, v \in M$. Hence, by Lemma 1, the bag of $\lambda_{T \setminus F', S}(v)$ has cardinality $|M| > \alpha$ in any lateral transfers scenario of (F', U). This contradicts the assumption that (F', U) is an α, τ-win. □

Lemma 3. *For each $v \in V(S)$, the number of $\leq_T$-minimal vertices of $\lambda^{-1}_{T,S}(v)$ is $\leq \alpha + \tau$. Furthermore, $|\lambda^{-1}_{T,S}(v)| \leq 2(\alpha + \tau)$.*

Proof. Let (S', A', g) be a α, τ lateral transfers scenario for S and T. Let F be the lateral transfers of T w.r.t. (S', A', g) and note that $|F| \leq \tau$. Since (S', A', g) is α-active, the number of $\leq_T$-minimal vertices of $\lambda_{T \setminus F, S}(v)$ is $\leq \alpha$. It follows that the number of $\leq_T$-minimal vertices of $\lambda^{-1}_{T,S}(v)$ is $\leq \alpha + \tau$. Since T has no vertices of indegree and outdegree 1, $|\lambda^{-1}_{T,S}(v)| \leq 2(\alpha + \tau)$. □

Lemma 4. *Let (F, U) be a board and let $x \in V(S) \setminus \{r(S)\}$. Let v be a $\leq_T$-minimal vertex $v \in \lambda^{-1}_{T \setminus F, S}(x)$ such that there is a $u \in \lambda^{-1}_{T \setminus F, S}(x)$ with outdegree 2 in $T \setminus F$ for which $v <_T u$. If e_1 and e_2 are the outgoing arcs from the vertices of v, then e_1, e_2 are alternative forced moves in (F, U).*

Proof. Let $f_1, \ldots, f_r$ be the arcs of the path from u to v in $T \setminus F$ and let f be the other outgoing arc of u. Assume that (U', F') is an α, τ-win such that $(U, F) \prec (U', F')$, and $e_1, e_2 \notin A(F')$. Since $e_1, e_2 \notin A(F')$, it follows that $f, f_1, \ldots, f_r \notin A(F')$ (since $f_i \in A(F')$ with maximal i then f_i would correspond to lateral transfers between ancestor and descendant, also $f \in A(F')$

would correspond to lateral transfers between ancestor and descendant). Hence $\lambda_{T\setminus F',S}(v) = \lambda_{T\setminus F',S}(u)$. That is, $\lambda_{T\setminus F',S}(v)$ is not an $\leq_T$-antichain. This contradicts the assumption that (F', U) is an α, τ-win. □

LEMMA 5. *Let (F, U) be a board, let r be the root of T, and let $R = T \setminus F$. Let M be the vertices $u \in V(T)$ such that: (1) there is no directed path from r to u in R, (2) $d^+_R(u) = 2$, and (3) $L_R(u) = L(T')$ where T' is the connected component of R to which u belongs. Let $e_1, \ldots, e_l$ be the outgoing arcs in R of vertices in M. If (F, U) is cyclic then, then $e_1, \ldots, e_l$ are alternative forced moves in (F, U).*

PROOF. (Idea.) Notice that for any board (F', U) such that $(F, U) \prec (F', U)$ and $\{e_1, \ldots, e_l\} \cap A(F_0) = \emptyset$, it holds that $\lambda_{T\setminus F',S}(u) = \lambda_{T\setminus F,S}(u)$ for any $u \in M$. □

LEMMA 6. *Testing whether a given board B is cyclic or is a α, τ-win can be done in time $O((\tau)^\tau |L|^2)$ time.*

PROOF. Omitted. □

5. ACTIVITY 1

Unfortunately, even the 1-ACTIVITY τ-TRANSFER PROBLEM is intractable. However, there exists a faster fixed parameter algorithm for this special case. We describe it here.

THEOREM 2. *The decision version of the* 1-ACTIVITY τ-TRANSFER PROBLEM *is NP-complete.*

PROOF. Omitted. □

Let $T' \subseteq T$. Let x be an arbitrary vertex with children x_1 and x_2 in S. Let $I(x, x_i, T')$ be the set of all arcs $\langle u, v\rangle \in A(T')$ such that $\lambda_{T',S}(u) \geq_S x$, $\lambda_{T',S}(v) \leq_S x_i$, and u is $\leq_T$-minimal in $\lambda^{-1}_{T',S}(\lambda_{T',S}(u))$. Let $H(x, x_i, T')$ be the set of all arcs $\langle u, v\rangle \in A(T')$ such that $\lambda_{T',S}(u) = x$ and $\lambda_{T',S}(v) \leq_S x_i$. A vertex x is *H-fat* for T' iff $|H(x, x_1, T')| + |H(x, x_2, T')| \geq 3$; it is *$I$-fat* for T' iff $|I(x, x_i, T')| \geq 2$ and $I(x, x_i, T')$ contains an arc $\langle u, v\rangle$ such that $\lambda_{T',S}(u) = x$, for $i = 1$ or $i = 2$; and it is *fat* for T' iff it is H-fat for T' or I-fat for T'. Let $M(x, T')$ be the set of $\leq_T$-minimal vertices $v \in \lambda^{-1}_{T',S}(x)$ such that there is a $u \in \lambda^{-1}_{T',S}(x)$ with outdegree 2 in T' for which $v <_T u$. If x is H-fat for T', then two alternative H moves at x in T' is a pair of arcs e_1, e_2 where e_1 and e_2 are the two outgoing arcs from a vertex of $M(x, T')$. If x is I-fat for T', then two alternative I moves at x in T' is a pair of arcs $\langle u, v\rangle, \langle u', v'\rangle$ where (1) $\langle u, v\rangle, \langle u', v'\rangle \in I(x, x_i, T')$, for $i = 1$ or $i = 2$, (2) $\lambda_{T',S}(u) = x$, and (3) u and u' are $\leq_T$ minimal in x and $\lambda_{T',S}(u')$, respectively.

The following is our 1-ACTIVITY τ-TRANSFER algorithm. Again, whenever a trivial reduction is possible in the algorithm it will be performed. Let C be the set $\{B\}$ where $B = (\emptyset, r(T))$ is a board. Let $t = \infty$. After the computation the output is t. We repeat the following until $C = \emptyset$.

1. Take $B \in C$, assume that $B = (F, U)$
2. Let $T' \leftarrow T \setminus F$.
3. Pick a $\leq_S$ minimal vertex x which is fat in T'
4. If x is H-fat, let B_1, B_2 be the boards obtained from B by making the alternative H moves e_1, e_2.
5. else if x is I-fat, let B_1, B_2 be the boards obtained from B by making the alternative I moves e_1, e_2.
6. If B_i is a $1, \tau$-win, let $C \leftarrow C \setminus \{B\}$ and let $t \leftarrow \min(t, s_i)$ where s_i is the cost of B_i
7. else if $|B| = \tau - 1$, let $C \leftarrow C \setminus \{B, \}$
8. else $C \leftarrow C \setminus \{B\} \cup \{B_1, B_2\}$.

By the lemmata below we can perform H and I moves, without creating a cyclic board, until there no longer is any fat vertex. When there no longer is a fat vertex we have a $1, \tau$-win.

LEMMA 7. *Let (F, U) be a board. If e_1, e_2 are alternative H moves in $T \setminus F$, then e_1, e_2 are alternative forced moves.*

PROOF. Omitted. □

LEMMA 8. *Let (F, U) and (F', U) be two boards such that $(F, U) \prec (F', U)$. Assume that v_0 is a vertex of T with children v_1 and v_2. If $L_{T\setminus F}(v_1)$ is separated from $L_{T\setminus F}(v_2)$ in $T \setminus F'$, then $\langle v_0, v_1\rangle \in F'$ or $\langle v_0, v_2\rangle \in F'$.*

PROOF. Omitted. □

LEMMA 9. *Let (F, U) be a board. Assume that v_0 is a vertex of T with children v_1 and v_2 such that $\langle v_0, v_1\rangle \in F$, $\langle v_0, v_2\rangle \notin F$, and $L_{T\setminus F}(v_2)$ is separated from $L(T \setminus F) \setminus (L_{T\setminus F}(v_1) \cup L_{T\setminus F}(v_2))$ in $T \setminus F$. If $F' = (F \setminus \{\langle v_0, v_1\rangle\}) \cup \{\langle v_0, v_2\rangle\}$, then (1) the vertices of outdegree 2 in $T \setminus F'$ is exactly the the vertices of outdegree 2 in $T \setminus F$ and (2) for each such vertex u, $\lambda_{T\setminus F,S}(u) = \lambda_{T\setminus F',S}(u)$.*

PROOF. Omitted. □

LEMMA 10. *Let (F, U) be a board. If e_1, e_2 are alternative I-moves in $T \setminus F$, then e_1, e_2 are alternative forced moves.*

PROOF. Assume that (F', U) is a $1, \tau$-win such that $(F, U) \prec (F', U')$. Let $T'' = T \setminus F'$ and let $T' = T \setminus$

F. Let x be a vertex with children x_1 and x_2 in T'. Assume that $I(x, x_1, T')$ has cardinality ≥ 2. Assume that $\langle u_1, v_1\rangle, \langle u_2, v_2\rangle$ are two arcs of $I(x, x_1, T')$ and that $\lambda_{T',S}(u_1) = x$. By Lemma 1, for at least 1 value of i, (1) the set $L_{T'}(u_i) \setminus L_{T'}(v_i)$ is separated from $L_{T'}(v_i)$ in T'' and (2) the set $L_{T'}(u_i)$ is separated from $L(T') \setminus L_{T'}(u_i)$ in T''. W.l.o.g. assume that $i = 1$ is this value. By Lemma 8 and Lemma 9 we can assume that $\langle u_1, v_1\rangle \in F'$. Hence e_1, e_2 are alternative forced moves. □

Assume that e_1, e_2 are alternative forced moves at x. If B is the board obtained by making e_i on (F, U), then the cost of B equals the cost of (F, U) plus 1. That is, we make 2 branches and increase the cost by 1. Hence, our algorithm makes $O(2^\tau)$ branches.

Let $s\langle u_1, v_1\rangle, \ldots, \langle u_r, v_r\rangle$ and $s'\langle u'_1, v'_1\rangle, \ldots, \langle u'_s, v'_s\rangle$ be two sequences of arcs of T. Let $l = \min(r, s)$ and let $F = \cup_{1\leq i\leq l-1} e_i$. We write $s \prec_S s'$ iff (1) $\langle u_i, v_i\rangle = \langle u'_i, v'_i\rangle$ for each $1 \leq i \leq l-1$ and (2)

$$lca_S(\lambda_{T\setminus F,S}(u_l), \lambda_{T\setminus F,S}(v_l)) <_S lca_S(\lambda_{T\setminus F,S}(u'_l), \lambda_{T\setminus F,S}(v'_l)).$$

LEMMA 11. *In some branch of the algorithm, a sequence of moves $e_1, \ldots, e_\tau$ is made such that, for $1 \leq i \leq \tau$, (1) $(S \cup e(\{e_1, \ldots, e_i\}), r(T))$ is acyclic (where $r(T)$ is the root of T), and (2) $(\{e_1, \ldots, e_\tau\}, r(T))$ is a $1, \tau$-win.*

PROOF. Omitted. □

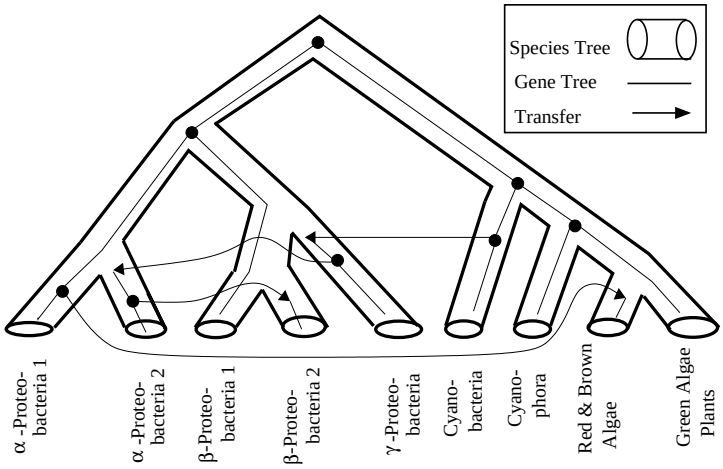
6. EXPERIMENTAL RESULTS

We have implemented our algorithms (they are available at `http://www.cs.mcgill.ca/ hallett/McKiTscH`) and tried them on several test sets. Figure 6 shows one of 5 optimal scenarios ($\tau = 4$) for the data from [4] we found via our implementations. The scenario reported in [4] is one of the 5 found by our algorithm but is not shown here.

Four of these 5 scenarios differ only on the order of the three transfers in the α-Proteobacteria-2, β-Proteobacteria-2, and γ-Proteobacteria lineages. The fifth scenario differs significantly: there is a transfer along the arc from the ancestor of α-Proteobacteria-1, red and brown algae, and β-Proteobactera-1 to β-Proteobacteria-1 but no transfer from the arc between the ancestor of $\{\alpha, \beta\}$-Proteobacteria-2, and γ-Proteobacteria to β, α-Proteobacteria-2. The program took on the order of 10 seconds to return all optimal scenarios for this particular dataset and on the order of a few minutes for (artificial) trees with as many as 25 leaves.

7. OPEN PROBLEMS

There are a number of future directions for this work. First, it would be interesting to incorporate weighted gene and species trees into our model (but still preserve tractability). A proper formulation has to address the fact that species trees are (close to) *ultrametric*, since the unit of measurement is *time* and gene trees are not. The most straightforward extensions to our model induce problems that are trivial to compute. It appears that either we must allow some "bounded error" in the estimations of time or we must use a new measure of evolution (such as possibly *genomic distance*). Second, currently both a species tree and gene trees are given as input. Do there exist fast algorithms that will find the species tree minimizing the number of lateral transfers, given only a set of gene trees? Third, we are planning to apply our algorithms to larger datasets for bacteria. It would be interesting to know if, in principal, this model will also work with, for example, viral phylogenies.



8. REFERENCES

[1] Charleston, M. A. (2000) Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *In Press. Math. Biosci.*.

[2] DasGupta, B., He, X., Jiang, T., Li, M., Tromp, J. and Zhang, L. (1997) On distances between phylogenetic trees. *Proc.* 8^{th} *Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '97*, pp. 427-436.

[3] DasGupta, B., He, X., Jiang, T., Li, M., Tromp, J. and Zhang, L. (1999) On the linear-cost subtree-transfer distance between phylogenetic trees. *Algorithmica* special issue on computational biology 25, pp. 176-195.

[4] Delwiche, C. F. and Palmer, J. D. (1996) Rampant horizontal transfer and duplication of Rubisco genes in Eubacteria and Plastids. *Mol. Biol. Evol.*, 13(6), pp. 873–882.

[5] Goodman, M. et al. (1979) Fitting the Gene Lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.*, 28.

[6] Guigó, R. et al. Reconstruction of ancient molecular phylogenies. *Molec. Phylogenet. and Evol.*, 6, 2, pp. 189-213.

[7] Hallett, M. and Lagergren, J. (2000) New Algorithms for the Duplication-Loss Model. 4^{th} *Annual RECOMB '00*, Tokyo, Japan, pp. 146-158.

[8] Hein. J. (1990) Reconstructing the evolution of sequences subject to recombination using parsimony. *Math. Biosci.*, 98, pp. 185-200.

[9] Hein. J. (1993) A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, 36, pp. 396-405.

[10] Hein, J., Jiang, T., Wang, L. and Zhang, K. (1995) On the complexity of comparing evolutionary trees. *Combinatorial Pattern Matching (CPM) '95*, LLNCS 937, pp. 177-190.

[11] Hein, J., Jiang, T., Wang, L. and Zhang, K. (1996) On the complexity of comparing evolutionary trees. *Discrete Applied Mathematics*, 71, pp. 153-169.

[12] Holmes, E. C., Worobey, M., and Rambaut, A. (1999) Phylogenetic evidence for recombination in Dengue virus. *Mol. Biol. Evol.*, 16(3), pp. 405–409.

[13] Olsen, G. J. and Woese, C. R. (1997) Archael genomics: an overview. *Cell*, 89, pp. 991-994.

[14] Page, R. D. M. (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms and areas. *Syst. Biol.*, 43, pp. 58-77.

[15] Page, R. D. M. (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, 14, pp. 819-820.

[16] Page, R. D. M. and Charleston, M. A. (1997) From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution*, 7, pp. 231-240.

[17] von Haseler, A. and Churchill, G. A. (1993) Network models for sequence evolution. *J. Mol. Evol.*, 37, pp. 77-85.

[18] Woese, C. (1987) Bacterial evolution. *Microbiol. Rev.*, 51, pp. 221-271.